Twenty-One Day Online Training Manual
on

Advanced Statistical and Machine Learning Techniques for Data Analysis Using Open-Source Software for Abiotic Stress Management in Agriculture

16 July – 5 August 2025

Training Manual

Volume 04

Spatial & Agricultural System Modeling for Abiotic Stress Management

Edited by

Dr. Santosha Rathod
Dr. Nobin Chandra Paul
Ms. Ponnaganti Navyasree
Mr. K Ravi Kumar
Dr. Prabhat Kumar

Organised by:

School of Social Science and Policy Support
ICAR-National Institute of Abiotic Stress Management, Baramati, Maharashtra - 413115

Spatial & Agricultural System Modeling for Abiotic Stress Management

Editors

Santosha Rathod

Nobin Chandra Paul

Ponnaganti Navyasree

K Ravi Kumar

Prabhat Kumar

2025

School of Social Science and Policy Support
ICAR-National Institute of Abiotic Stress Management, Baramati – 413115
Maharashtra, India

Title: Spatial & Agricultural System Modeling for Abiotic Stress Management

Editors: Santosha Rathod, Nobin Chandra Paul, Ponnaganti Navyasree, K Ravi Kumar, Prabhat Kumar

Published by: ICAR-National Institute of Abiotic Stress Management, Malegaon Khurd, Baramati – 413115, Maharashtra, India.

Edition: I

Volume: 4

ISBN: 978-81-985897-1-2

Citation:

Rathod, S., Paul, N. C., Ponnaganti, N., Kumar, K. R., & Kumar, P. (Eds.). (2025). *Spatial & Agricultural System Modeling for Abiotic Stress Management: Training manual of the twenty-one-day online training programme on "Advanced statistical and machine learning techniques for data analysis using open-source software for abiotic stress management in agriculture"* (Vol. 4). ICAR-National Institute of Abiotic Stress Management. ISBN 978-81-985897-1-2.

CONTENTS

Introduction to Remote Sensing & GIS

Nobin Chandra Paul, Ponnaganti Navyasree, K. Ravi Kumar, Prabhat Kumar and Santosha Rathod
ICAR-National Institute of Abiotic Stress Management, Baramati, Pune-413115
Email: ncp375@gmail.com

1. Introduction

Remote sensing involves collecting data or information about objects, regions, or phenomena from a distance, typically using satellites or aircraft, without direct physical contact. It relies on detecting and analyzing electromagnetic radiation that is reflected, emitted, or scattered from the Earth's surface or atmosphere. This method enables researchers and experts to study and monitor the Earth's features and atmospheric conditions across different spatial, spectral, and temporal dimensions, offering critical insights for a wide range of applications.

2. Principles of Remote Sensing

Remote sensing is based on the interaction between electromagnetic radiation (EMR) and various features on the Earth's surface. When radiation-typically from a natural source like the sun-strikes the Earth's surface, it may be reflected, absorbed, or transmitted depending on the material it encounters. Sensors on remote sensing platforms detect the reflected or emitted radiation, which is then analyzed to derive information about surface characteristics.

Different surface features exhibit distinct spectral signatures due to their unique physical and chemical properties. For example, vegetation tends to reflect strongly in the near-infrared portion of the spectrum, while water bodies primarily absorb incoming radiation. By detecting these variations in reflectance or emission, remote sensing technologies can effectively differentiate among surface types such as vegetation, water, and bare land.

Stages in Remote Sensing

- Generation of electromagnetic radiation (EMR), either from a natural source like the sun or emitted by the object itself.
- Propagation of this energy through the atmosphere, during which it may be absorbed or scattered by atmospheric particles.

- Interaction of the radiation with the Earth's surface, where it is reflected, absorbed, or re-emitted based on the surface properties.
- Travel of the reflected or emitted energy from the surface back through the atmosphere to the sensor.
- Detection and recording of the energy by the remote sensing instrument.
- Transmission of collected data to ground stations, followed by processing, interpretation, and analysis

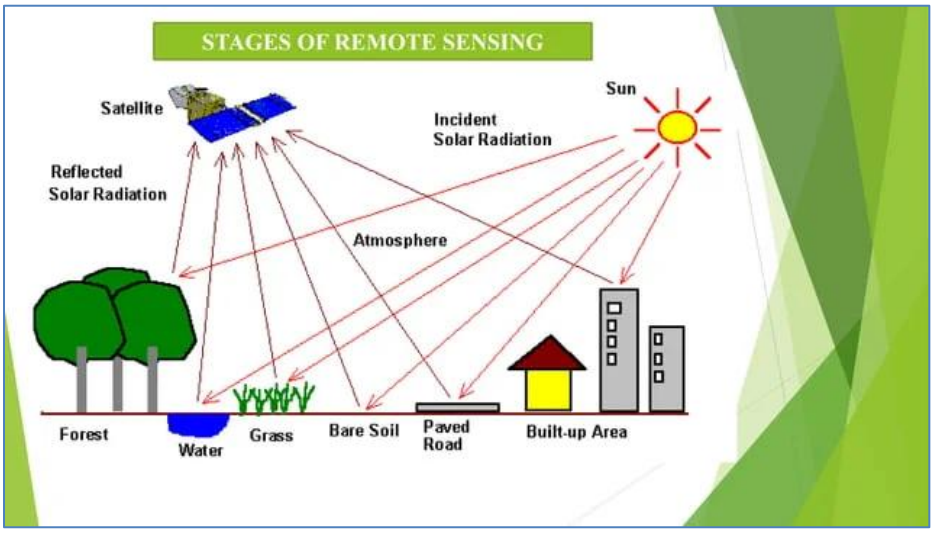


Figure 1. Remote sensing stages (***source:*** *https://sl.bing.net/emkyDb8t988*)

3. Types of Remote Sensing

Remote sensing systems are typically categorized based on the origin of the energy used and the nature of the sensors employed.

1. **Passive Remote Sensing:** These systems depend on natural energy sources, most commonly sunlight, to observe the Earth's surface. They detect the radiation that is either reflected or emitted by features on the ground. Because they rely on solar energy, passive systems are limited to daytime operations and can be influenced by atmospheric conditions like cloud cover. Well-known examples include satellite sensors on platforms such as Landsat and MODIS.

2. **Active Remote Sensing:** In contrast, active systems generate their own source of energy to scan the target area. The emitted energy interacts with the Earth's surface, and the reflected signal is then captured by the sensor. This approach allows for data collection regardless of time of day or weather conditions. Radar and LiDAR are common examples of active remote sensing technologies.

4. Electromagnetic Spectrum in Remote Sensing

Remote sensing makes use of different parts of the electromagnetic spectrum (EMS) to gather data. The EMS ranges from gamma rays to radio waves, with each type of radiation having unique properties and interactions with Earth's surface. The most commonly used regions of the EMS in remote sensing are:

- **Visible Light**: This is the portion of the spectrum that can be seen by the human eye, ranging from 0.4 to 0.7 micrometers (μm).
- **Infrared Radiation**: The near-infrared (0.7-1.5 μm), mid-infrared (1.5-5 μm), and thermal infrared (5-15 μm) are used to detect heat and moisture levels in vegetation, soil, and water bodies.
- **Microwaves**: These are used in radar systems to penetrate clouds and gather data on surface roughness, elevation, and moisture content, independent of weather conditions.

Each of these regions provides different types of information about Earth's surface, making them useful for a variety of applications, including environmental monitoring, land use mapping, and agriculture.

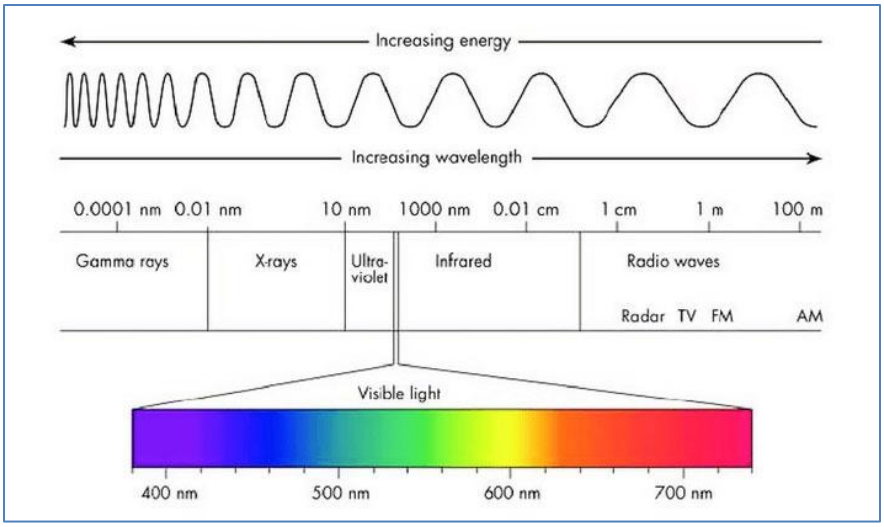


Figure 2. Electromagnetic radiation spectrum with bands *(**source:** https://sl.bing.net/bQc8kSJheqy)*

5. Platforms and Sensors

Remote sensing platforms can be divided into two main categories: **spaceborne** and **airborne**.

- **Spaceborne Platforms**: Satellites orbiting the Earth capture images and data from a wide area. They are equipped with sensors to measure reflected sunlight or emitted radiation across different parts of the electromagnetic spectrum. Examples include the Landsat, Sentinel, and WorldView satellites.
- **Airborne Platforms**: These platforms, including aircraft and drones, are used for more localized or high-resolution remote sensing. Airborne sensors, such as those used for LiDAR or hyperspectral imaging, provide more detailed and targeted data over smaller areas than spaceborne systems.

Sensors on both spaceborne and airborne platforms vary in terms of their spatial resolution (the level of detail in an image), spectral resolution (the ability to capture data across different parts of the EMS), and temporal resolution (how frequently the sensor revisits a given area).

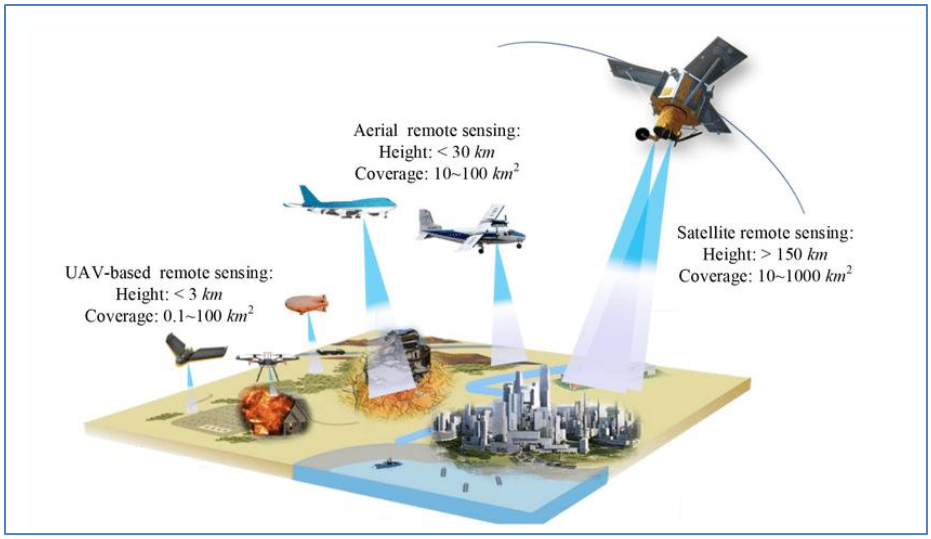


Figure 3: Platforms in remote sensing ***(source:*** *https://sl.bing.net/jNwEQbw7wOq)*

6. Types of Resolution in Remote Sensing

Remote sensing data is often described in terms of three key types of resolution: ***spatial resolution, temporal resolution, and spectral resolution***. Each of these resolutions influences the quality and usefulness of the data captured by remote sensing systems, and they are essential for understanding how satellite and aerial imagery can be used for various applications.

6. 1. Spatial Resolution

Spatial resolution, also known as ground resolution, refers to the smallest area on the ground that can be distinguished in an image. In other words, it is the size of a pixel in the remote sensing image, which represents a specific geographic area. The spatial resolution of a sensor determines the level of detail in the images it produces.

For instance, the Landsat Thematic Mapper (TM) sensor has a spatial resolution of **30 meters**, meaning each pixel in the image corresponds to a 30-meter by 30-meter square on the ground. Higher spatial resolution means more detailed images, which are especially useful for identifying smaller features, such as individual buildings or roads.

- **Low spatial resolution**: Sensors with low spatial resolution capture larger areas, but with less detail. For example, weather satellites might have resolutions larger than 1 kilometer per pixel, providing a broader view with less fine detail.
- **High spatial resolution**: Some satellites or systems can capture images at much finer resolutions, such as less than one meter per pixel. These systems are often military satellites or expensive commercial satellite platforms.

In general, higher spatial resolution comes at the cost of reduced coverage, meaning higher-resolution images cover smaller areas but provide more detail.

6. 2. Temporal Resolution

Temporal resolution refers to the frequency with which a satellite or sensor revisits the same location on the Earth's surface. This is crucial for applications such as monitoring changes in the environment or urban development, as it determines how often updated data can be obtained for a specific area.

- **High temporal resolution**: Satellites with high temporal resolution have frequent revisit times, sometimes capturing images multiple times per day. This is typical for weather satellites or certain earth observation satellites that monitor rapidly changing phenomena, such as storms or crop growth. These satellites can revisit the same spot several times a day to capture time-sensitive data.
- **Low temporal resolution**: Satellites with low temporal resolution revisit the same area only a few times per year. For example, the Landsat TM satellite revisits a given area about **8–20 times per year**, which is suitable for applications like land-use and land-cover change detection where frequent monitoring is not essential.

The frequency of satellite pass depends on factors such as satellite orbit, its operational design, and its specific mission objectives. In many cases, satellites in **polar orbits** (which pass over the Earth from pole to pole) can offer consistent temporal resolution, ensuring wide global coverage over time.

6. 3. Spectral Resolution

Spectral resolution refers to the ability of a sensor to capture data in multiple bands across the electromagnetic spectrum. The more spectral bands a sensor has, the higher its spectral resolution. Each spectral band represents a specific range of wavelengths of electromagnetic radiation, such as visible light, near-infrared, or thermal infrared. By capturing data in multiple spectral bands, remote sensing sensors can identify and analyze different surface materials and features based on their unique spectral signatures.

- **Low spectral resolution**: Sensors with low spectral resolution might have only a few bands, such as just one band for visible light. These systems produce simple images that represent the Earth's surface in terms of broad categories, similar to black-and-white photographs.
- **High spectral resolution**: Sensors with high spectral resolution capture data across many narrow spectral bands, including those outside the visible range, such as near-infrared or thermal infrared. The **Landsat TM sensor**, for example, has **seven spectral bands**, including visible light and several infrared bands. This allows the sensor to distinguish between different land cover types, such as water, vegetation, and soil, and provides a richer dataset for environmental monitoring, agriculture, and other applications.

Sensors with high spectral resolution can distinguish subtle differences in surface materials that may not be apparent to the human eye. This is especially useful in applications like vegetation health monitoring, mineral exploration, and land cover classification.

7. Applications of Remote Sensing

Remote sensing has a wide range of applications across various fields, including:

- **Environmental Monitoring**: Remote sensing plays a vital role in tracking changes in the environment, such as deforestation, land degradation, and water quality.
- **Agriculture**: It helps monitor crop health, estimate yields, and detect issues like water stress, pests, and diseases.

- **Urban Planning**: Remote sensing data is used for land use/land cover mapping, urban growth monitoring, and disaster management.
- **Natural Resource Management**: This includes monitoring forests, wetlands, and water bodies, as well as supporting mineral and petroleum exploration.
- **Climate Change**: Remote sensing is key in studying atmospheric conditions, sea level rise, temperature variations, and other phenomena associated with climate change.

- Forestry Applications: Satellite imagery maps tree species, monitors forest health for diseases or stressors, and tracks forest extent, aiding conservation and environmental management.
- Vegetation Health: Changes in spectral reflectance, driven by chlorophyll and stressors like drought or pests, allow early detection of vegetation health issues using narrow bands (0.4–0.9 μm).
- Biodiversity: Satellite-derived vegetation data, combined with GIS, maps ecosystems and supports biodiversity assessments for conservation planning.
- Change Detection: Imagery tracks vegetation cover changes due to clearing, wildfires, or water scarcity, enabling timely environmental management.
- Geology: Remote sensing aids mineral/petroleum exploration, geomorphological mapping, and volcanic monitoring using spectral and thermal data.
- Land Degradation: Imagery identifies degraded areas from saline soils or overgrazing, supporting restoration efforts.
- Meteorology: Satellites map cloud types and temperatures, supporting weather forecasting and climate monitoring.

Geographic Information System (GIS): An Introduction

A Geographic Information System (GIS) is a powerful tool for collecting, managing, analyzing, and visualizing spatial (location-based) and attribute (descriptive) data. It integrates hardware, software, data, people, and methods to support spatial analysis and decision-making.

Unlike traditional mapping, GIS allows users not only to visualize maps but also to ask complex spatial questions, analyze patterns, model changes over time, and integrate data from multiple sources. In simple terms, it helps us answer questions like:

- *Where is this happening?*
- *What is near this location?*
- *How is this place changing over time?*

GIS is used across domains including agriculture, forestry, urban planning, environmental monitoring, disaster management, public health, and transportation.

Vector and Raster Data Models

GIS data are represented in two primary formats:

- **Vector Data Model**: Represents spatial features as discrete objects (points, lines, polygons) using x-y coordinates. Points (e.g., wells) have location; lines (e.g., roads) have length; and polygons (e.g., land parcels) have area and perimeter. Vector models are precise for discrete features and support topological relationships, ideal for mapping infrastructure or boundaries.

- **Raster Data Model**: Uses a grid of cells (pixels) to represent continuous phenomena, where each cell holds a value reflecting characteristics like elevation or temperature. Raster is suited for continuous data (e.g., satellite imagery, elevation models) but may lose detail at lower resolutions.

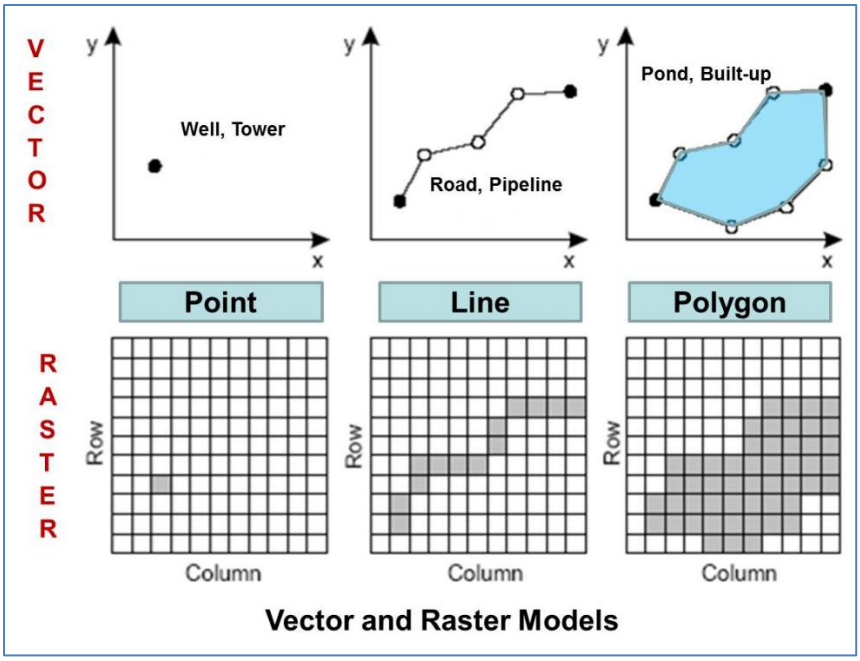


Figure 4: Vector & raster data model (source: https://sl.bing.net/izbrtPTsLPE)

DN Values, Spatial Resolution, and Pixel

- **DN Values (Digital Numbers)**: In raster data, particularly from remote sensing, DN values represent the intensity of electromagnetic radiation recorded by a sensor for each pixel. These values, often ranging from 0 to 255 in 8-bit imagery, correspond to reflectance or emittance and are used to derive information like land cover or vegetation health.
- **Spatial Resolution**: The size of the ground area represented by a single pixel, determined by the sensor's instantaneous field of view (IFOV). For example, Landsat Thematic Mapper has a 30 m resolution, meaning each pixel covers a 30 m x 30 m area. Higher resolution (smaller pixel size) provides finer detail but increases data volume.
- **Pixel**: The smallest addressable unit in a raster image, representing a specific ground area. Pixel size directly relates to spatial resolution, influencing the level of detail visible in GIS analyses.

Components of GIS

A functional GIS comprises **five core components**, each of which plays a crucial role in the successful operation of the system:

- ✓ **Hardware:** Refers to physical devices required for GIS operations:
  - Computers (desktops, laptops, servers)
  - Data storage devices (external hard drives, cloud storage)
  - GPS units for field data collection
  - Printers and plotters for map outputs
  - High-performance computing enables handling of large geospatial datasets.
- ✓ **Software:** GIS software enables users to input, manipulate, analyze, and display geospatial data.

- Examples:
  - **Commercial**: ArcGIS, ERDAS Imagine
  - **Open-source**: QGIS, ILWIS
- Functions include spatial analysis, geo-statistics, 3D modeling, remote sensing integration, and web-based mapping.
  - ✓ **Data:** The core of any GIS project; can be spatial or non-spatial.
    - **Spatial data**: Geographic coordinates representing physical features.
    - **Attribute data**: Descriptive information about spatial features (e.g., population, land-use type).
- Data sources include:
  - Satellite imagery, aerial photographs
  - Survey data, GPS recordings
  - Scanned maps and digitized vector layers
  - ✓ **People:** Users who manage, analyze, interpret, and apply GIS outputs.
- Categories:

- GIS analysts and technicians
- Developers and system administrators
- Decision-makers and domain experts

- Their expertise shapes how GIS is used in real-world applications.

  ✓ **Methods (or Procedures):** Refer to the **workflows**, **standards**, and **documented protocols** used in GIS projects.

- Include:
  - Data collection and quality control methods
  - Spatial analysis techniques
  - Metadata documentation
  - Reproducibility protocols

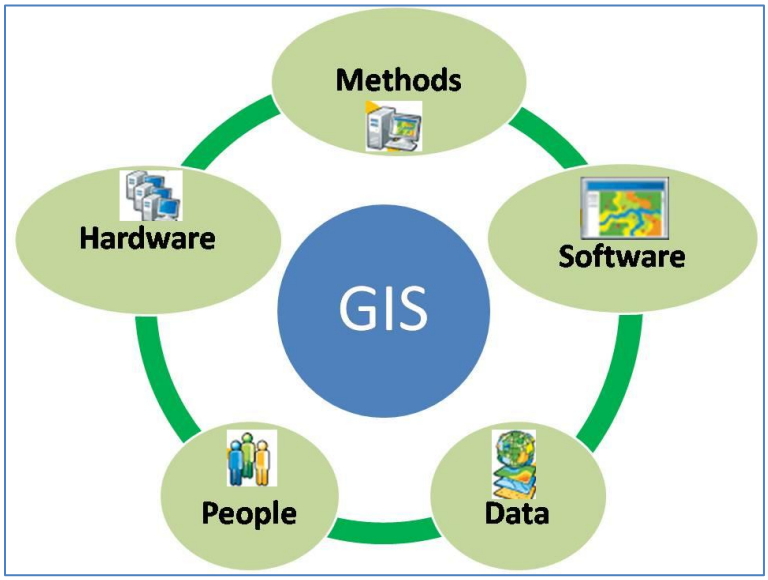


Figure 5. GIS components ***(source:*** *https://sl.bing.net/bmWaz6tl1Eq)*

Applications of GIS

GIS supports a wide range of applications by integrating and analyzing spatial data:

- **Urban Planning**: Maps infrastructure, land use, and population density to guide city development and zoning.
- **Environmental Management**: Monitors land cover changes, deforestation, and biodiversity, supporting conservation efforts.
- **Agriculture**: Analyses soil properties and crop health *i.e.*, digital soil mapping.
- **Disaster Management**: Assesses flood or drought impacts by overlaying thematic layers (e.g., topography, water resources) to predict inundation or identify vulnerable areas.
- **Transportation**: Optimizes routes through network analysis, such as determining efficient bus routes.
- **Public Health**: Maps disease spread or resource allocation, using spatial queries to identify high-risk areas.

Practical Application: **Mapping and Assessment of Abiotic Stresses by Integrating Remote Sensing and Machine Learning Techniques**

Abiotic stresses such as drought, heatwaves, and soil degradation significantly hinder agricultural productivity, especially in hot semi-arid regions like Pune district in western India. These stresses, driven by non-living environmental factors such as temperature extremes, erratic rainfall, and poor soil conditions, affect the growth, yield, and health of crops, making farming riskier and less predictable (Minhas and Reddy, 2017). In this context, accurate mapping and identification of abiotic stress-prone areas become essential for better planning, risk mitigation, and improved agricultural management. This study introduces a novel integrated approach that combines the Analytical Hierarchy Process (AHP) with machine learning (ML) models to map and assess abiotic stress zones. The methodology incorporates multiple satellite data as raster layers, including climate, soil, terrain, and vegetation indices such as soil depth, soil pH, land use/land cover (LULC), NDVI, SAVI, VHI, slope, land surface temperature (LST), and mean annual rainfall (MAR). These factors were weighted using AHP based on expert input and combined through a weighted sum method to generate the final abiotic stress map. The use of AHP ensures that expert knowledge is systematically incorporated, while ML models like Random Forest and Support Vector Regression enhance the accuracy of classification and prediction. The stress map was validated using high-resolution Google Earth imagery at randomly selected locations, ensuring

reliability and real-world relevance. From a practical perspective, the results revealed that tehsils such as Purandar, Baramati, Indapur, and Daund are highly stress-prone due to shallow soils and low rainfall levels (<550 mm). The findings are in line with those from the Central Ground Water Board (CGWB), reinforcing the credibility of the approach. This methodology enables the identification of high-risk areas where targeted interventions can be planned. For instance, farmers in stress-prone zones can be guided to adopt drought-tolerant crop varieties, implement soil moisture retention practices like mulching, and invest in efficient water management techniques such as drip irrigation and rainwater harvesting. These location-specific strategies can help farmers reduce crop losses, optimize resource use, and increase resilience to climatic uncertainties.

The study's relevance extends beyond individual farmers. For government agencies, agricultural planners, and development organizations, the abiotic stress maps serve as decision-support tools for policy formulation, infrastructure development, and allocation of subsidies or schemes. Local and regional stakeholders can prioritize investments in areas identified as high stress zones, while researchers and agronomists can use the methodology to further study the interactions between environmental variables and crop performance. Moreover, the methodology is scalable. Although developed for Pune district, it can be extended to other parts of India with similar agro-climatic conditions. In the future, this approach can help develop a national-level multiple abiotic stress index, robust, replicable, and practical framework for mapping abiotic stresses.

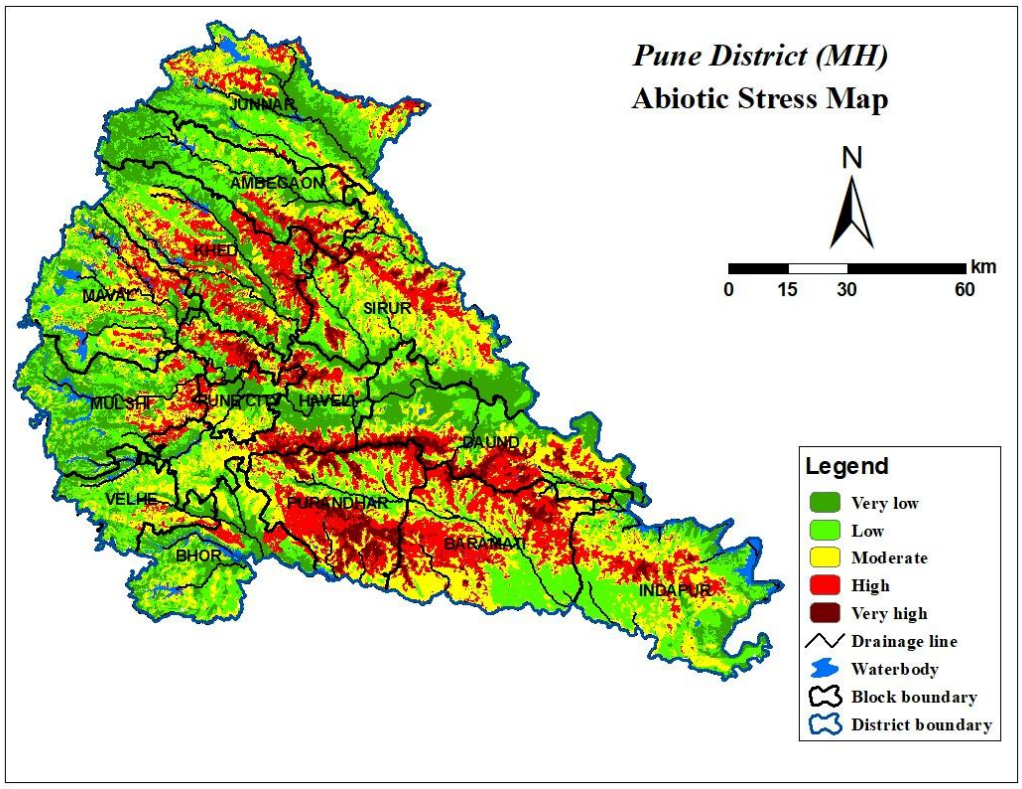


Figure 6. Abiotic stress map of Pune district.

Conclusions

Remote sensing and Geographic Information Systems (GIS) are complementary and powerful technologies that together enable a deeper understanding of the Earth's surface and dynamic processes. Remote sensing provides the means to continuously observe and collect large volumes of spatial data using satellite and airborne sensors, capturing valuable information across different spectral, spatial, and temporal resolutions. This data is essential for monitoring environmental changes, assessing natural resources, and responding to global challenges such as climate change, deforestation, and urban expansion. Together, remote sensing and GIS empower researchers, planners, and decision-makers to make informed, timely, and sustainable decisions. Their integration is vital for applications in urban planning, agriculture, disaster management, environmental conservation, and many other fields, paving the way for more resilient and data-driven solutions in the face of global environmental and societal challenges.

References

Cressie, N.A.C. 1991. Statistics for spatial data. John Wiley & Sons, Inc. USA. p.900.

ICAR-NIASM. (2024). Mitigating abiotic stress in agriculture: Promising technologies, pp. 25-28.

Lillesand, T.M., Kiefer, R.W., & Chipman, J.W. (2015). Remote Sensing and Image Interpretation (7th ed.). Wiley.

Minhas, P.S. and Reddy, G.P.O. (2017). Edaphic stresses and agricultural sustainability: an Indian perspective. *Agricultural Research*, 6, 8-21.

Paul, N. C., Reddy, G. O., Kumar, N., Reddy, K. S., Gaikwad, B. B., et al. (2025). Mapping and assessment of abiotic stresses in hot semi-arid ecosystem of western India using analytical hierarchy process and machine learning models. *Environmental Earth Sciences*, 84(10), 276.

Introduction to QGIS

Nobin Chandra Paul, Ponnaganti Navyasree, K. Ravi Kumar, Prabhat Kumar and Santosha Rathod
ICAR-National Institute of Abiotic Stress Management, Baramati, Pune-413115
Email: ncp375@gmail.com

1. Introduction

A Geographic Information System (GIS) serves as a comprehensive toolset for collecting, managing, analyzing, and visualizing spatial or geographic data. It combines hardware, software, and data to handle information that is tied to specific locations on the Earth's surface. Among various GIS tools, QGIS is a widely used open-source desktop application known for its ability to view, edit, and analyze geospatial datasets. Its adaptability, user-friendly interface, and cost-free accessibility have made it a preferred choice among researchers, urban planners, students, and policymakers alike.

2. What is QGIS?

QGIS (Quantum GIS) is an open-source, freely available Geographic Information System designed for viewing, editing, and analyzing geospatial data. Launched in 2002, it has grown significantly through global community contributions. QGIS supports numerous raster and vector formats, leveraging libraries like GDAL and OGR. Its flexible, plugin-based architecture enables customization, making it ideal for applications in fields such as agriculture, forestry, urban planning, hydrology, and disaster management.

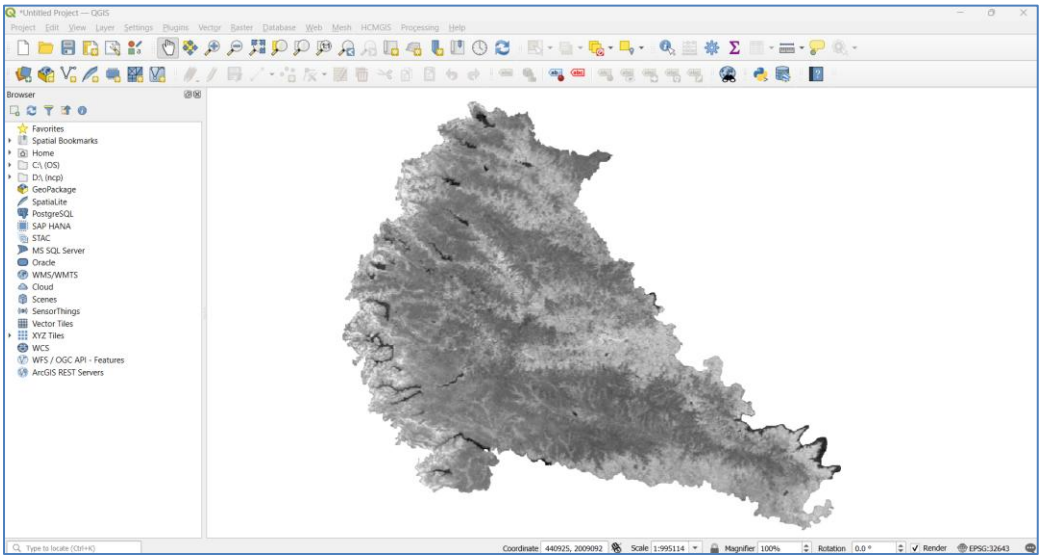
3. Graphical User Interface of QGIS

The QGIS Graphical User Interface (GUI) is designed for ease of use, comprising five main components: the Menu Bar, Toolbars, Panels, Map View, and Status Bar. The Menu Bar provides hierarchical access to functions like project management, layer operations, and geoprocessing tools, with customizable keyboard shortcuts (Settings > Keyboard Shortcuts). Toolbars, such as Project, Map Navigation, and Attributes, offer quick access to common tasks like zooming, panning, or editing. Panels, including the Layers and Layer Order Panels, allow users to manage datasets and their display order. The Map Canvas is the central canvas for visualizing spatial data,

while the Status Bar displays coordinates and scale information. Users can customize the interface by rearranging toolbars and panels to suit their workflow, enhancing efficiency. The QGIS interface is user-friendly and customizable. It consists of several key components:

- **Menu Bar**: Provides access to all major functionalities and tools.
- **Toolbars**: Quick-access buttons for commonly used tools.
- **Panels**: Including Layers, Browser, and Map Legend.
- **Map Canvas**: Where spatial data is displayed and interacted with.
- **Status Bar**: Shows coordinate positions, scale, and rendering status.

The interface is fully customizable. As users become more experienced, they can add or hide panels and toolbars according to their workflow.



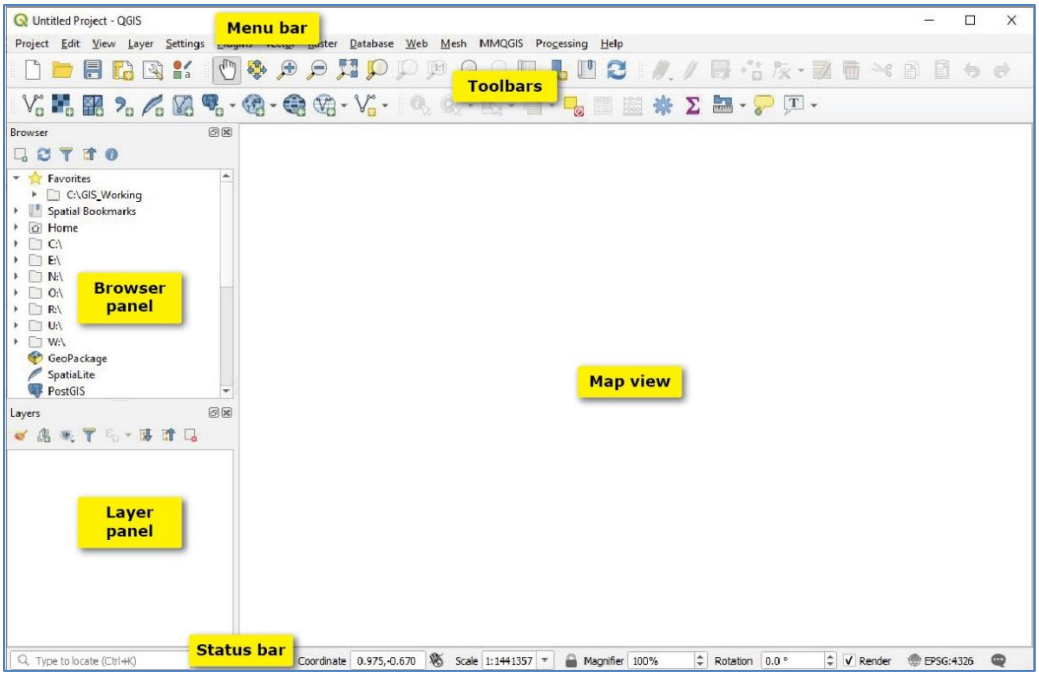


Figure 1: QGIS main GUI.

Download and Installation

Follow these steps to download and install QGIS from the official website (https://qgis.org/download/):

(a) **Visit the QGIS Website**:

- Open your web browser and navigate to https://qgis.org/download/.
- This page provides download options for various operating systems (Windows, macOS, Linux, etc.).

(b) **Choose the Appropriate Version**:

- **Latest Release (Stable)**: Recommended for most users. Select the latest stable version (e.g., QGIS 3.4 or newer) for reliability and regular updates.
- **Long-Term Release (LTR)**: Suitable for users needing maximum stability, typically updated annually.

(c) **Select Your Operating System**:

- **Windows**:
  - Click the link for the Windows installer *(e.g.*, "QGIS Standalone Installer").

        - Choose between 64-bit (recommended for modern systems) or 32-bit versions.
        - Download the .exe file (e.g., QGIS-OSGeo4W-<version>-setup-x86_64.exe).
    - **macOS**:
        - Click the macOS download link, typically a .dmg file for the latest version.
        - Ensure compatibility with your macOS version (e.g., macOS High Sierra or later).

(d) **Download the Installer**:

- Click the download link for your chosen version and operating system.
- Save the installer file to your computer (e.g., Desktop or Downloads folder).

(e) **Run the Installer**:

- **Windows**:
    - Double-click the .exe file.
    - Follow the installation wizard, accepting the license agreement and selecting the installation directory (default is recommended).
    - Choose components (e.g., QGIS, GRASS, SAGA) if prompted.
    - Wait for the installation to complete (requires ~2 GB disk space and 4 GB RAM minimum).
- **macOS**:
    - Open the .dmg file and drag the QGIS icon to the Applications folder.
    - Follow any additional prompts for dependencies (*e.g.*, Python or GDAL).

(f) **Verify Installation**:

- Launch QGIS from your applications menu or desktop shortcut.
- Ensure the interface loads correctly and check for optional plugins (*e.g.*, GRASS, SAGA) in the Plugins menu.

(g) **Optional: Install Plugins**:

- Open QGIS, go to Plugins > Manage and Install Plugins.
- Search for and install plugins like Python consol or Google Earth Engine.

(h) **Keep Updated**:

- Regularly check https://qgis.org for updates to ensure you have the latest features and security patches.

When beginning a new project, one of the first steps is to set the Coordinate Reference System (CRS), which aligns your digital map with real-world geography. You can configure CRS and preferred units (distance, area, angle) through Settings > Options > CRS. This setup is crucial for data accuracy and interoperability, especially when combining datasets from different sources.

4. Data Types in QGIS: Raster and Vector

GIS systems primarily handle two types of spatial data:

- **Vector data**: Includes **points** (*e.g.*, tree locations), **lines** (*e.g.*, roads), and **polygons** (*e.g.*, land parcels). Each feature is stored in a **layer** along with its **attributes** (*e.g.*, type of land use, population).
- **Raster data**: Composed of a matrix of cells or pixels, often used to store aerial imagery, elevation models, or satellite data.

QGIS allows you to visualize, edit, and analyze both data types with great flexibility. One of the most common vector data formats is the **Shapefile**, which must include multiple components (.shp, .shx, .dbf, and .prj) to be usable.

5. Key Menu Options and Tools in QGIS

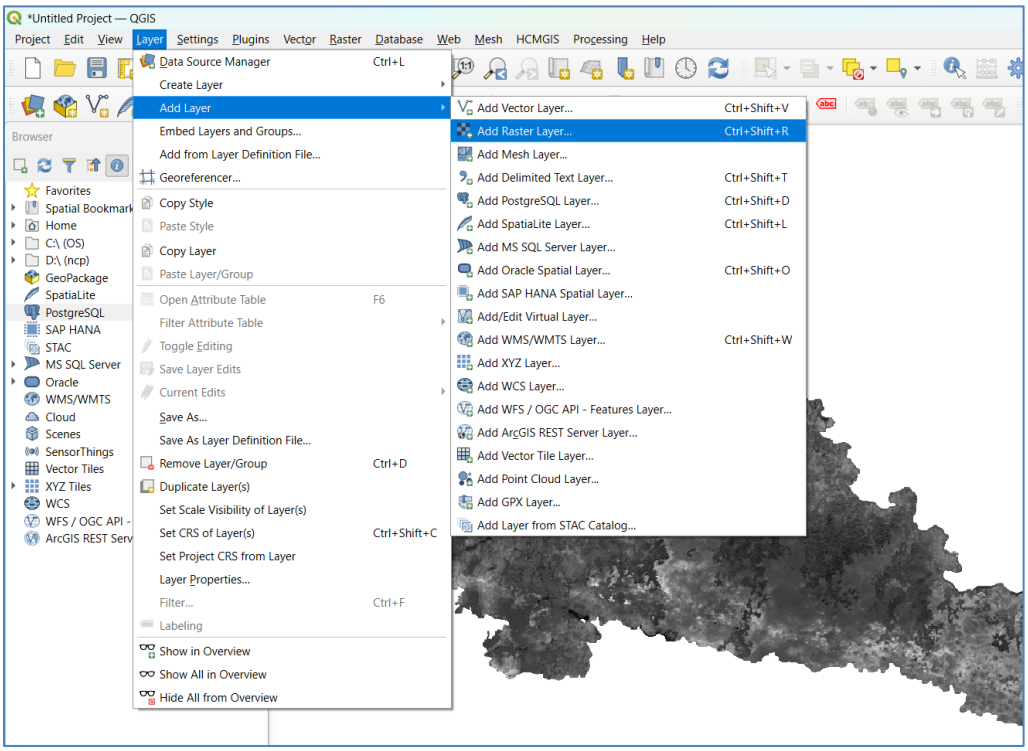
- **Project Menu**: Create, open, save, or export projects; manage project properties.
- **Layers Menu**: Add new data, style layers, label features, and manage symbology.
- **Vector Menu**: Access geoprocessing tools such as **Buffer**, **Clip**, **Dissolve**, and **Intersect**.
- **Data Management Tools**: Merge vector layers, transform projections, and edit fields.
- **Plugins Menu**: Access and manage plugins, such as **QuickMapServices** for basemap integration (Google Maps, OSM, etc.).

6. Data Import and Export in QGIS

Adding Layers:

1. Open **QGIS**.
2. Go to **Project > New** to start a new project.
3. To add data, go to **Layer > Add Layer**, and select the appropriate type:
    - **Add Vector Layer** (e.g., shapefiles, GeoJSON)
    - **Add Raster Layer** (e.g., TIFF, JPEG)

- **Add Layer from Database** (e.g., PostGIS, SpatiaLite)
- **Add Layer from Web Services** (e.g., WMS, WFS)



Exporting Data:

- Right-click on a layer and select **Export > Save Features As...**
- Choose the format (e.g., Shapefile, GeoPackage, KML, CSV)
- Set CRS and file location, then click **OK** to export.

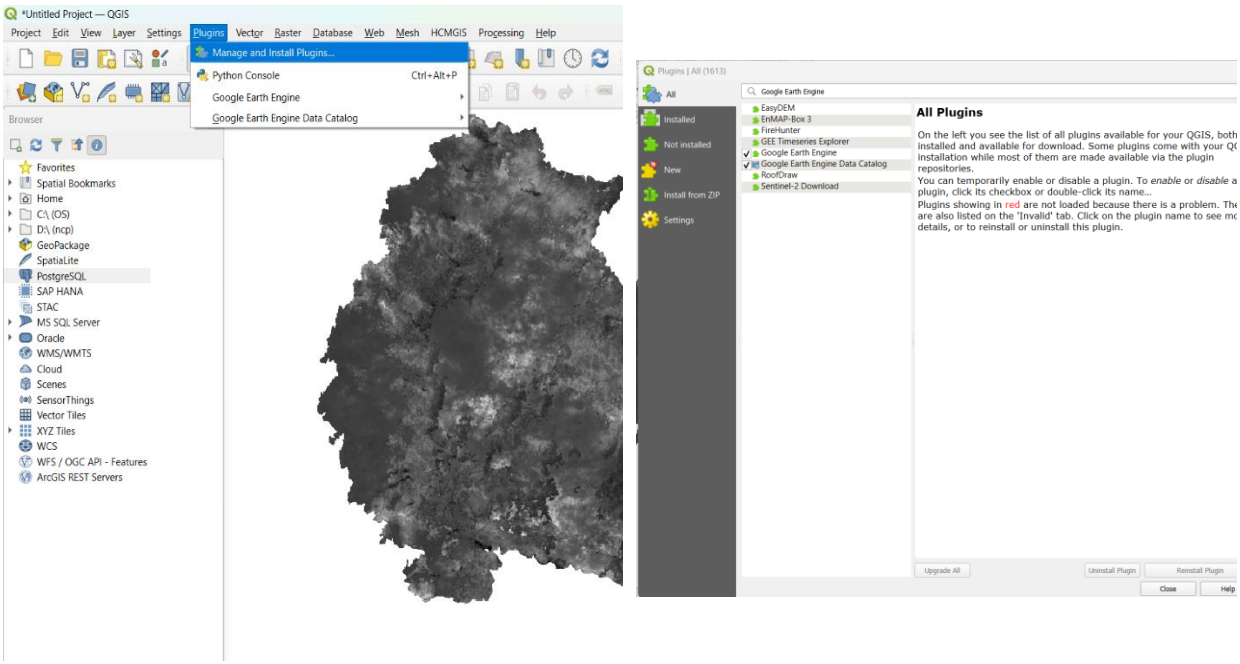
7. Installing Plugins in QGIS

(a) Go to **Plugins > Manage and Install Plugins**.

(b) Under the **All** tab, search for the plugin by name or description.

(c) Select the desired plugin and click **Install Plugin**.

(d) If required, go back to **Plugins > Manage and Install Plugins** and check the plugin to activate it.



8. Adding Raster Layers in QGIS

(a) Go to **Layer > Add Layer > Add Raster Layer**.

(b) In the dialog box, click **Browse** and locate the raster file (e.g., .tif).

(c) Select the file and click **Add**.

(d) The raster will appear as a new layer in the **Layers Panel** and **Map Canvas**.

9. QGIS Analysis Tools (Processing Toolbox)

QGIS provides a suite of spatial analysis tools available under the **Processing Toolbox**:

- **Buffer**: Creates buffer zones around vector features.
- **Clip**: Extracts input layer features within the bounds of another layer.
- **Dissolve**: Merges adjacent features based on a shared attribute.
- **Intersect**: Returns overlapping areas from multiple layers.
- **Union**: Combines geometries and attributes from multiple layers.
- **Field Calculator**: Performs calculations using field values or custom expressions.
- **Raster Calculator**: Performs cell-by-cell calculations on raster layers.
- **Zonal Statistics**: Calculates statistics (mean, sum, etc.) of a raster within vector zones.

These tools can be accessed via **Processing > Toolbox**, and support both **vector** and **raster** data analysis.

10. Creating Maps in QGIS

Maps are essential communication tools. QGIS allows users to design professional maps through the **Print Layout** interface. Key map elements include:

- **Scale bar**
- **North arrow**
- **Legend**
- **Title and metadata**
- **Locator or inset maps** for context

After designing, maps can be exported in various formats like **JPEG**, **PNG**, or **PDF**. This makes it easy to share outputs with stakeholders or include in reports and presentations.

11. Essential GIS Vocabulary for Beginners

To help users get familiar with GIS, understanding basic terms is crucial:

- **Attributes**: Descriptive information linked to spatial features.
- **CRS**: Coordinate Reference System used to geo-locate features.
- **Projection**: Method to represent Earth's surface in 2D.
- **Features**: Mappable objects (points, lines, polygons).
- **Layer**: A dataset containing geographic features.
- **Plugin**: Extensions that add functionality to QGIS.
- **Raster/Vector Data**: Two primary data formats in GIS.
- **Shapefile**: A common vector data format (.shp, .shx, .dbf, .prj).

12. Applications of QGIS in Agriculture

QGIS, an open-source GIS, enhances agricultural management by integrating spatial data for informed decision-making. Its tools and plugins, like SCP and GRASS, support efficient farming practices.

- **Crop Health Monitoring:** Analyzes spectral data to detect stress from drought or pests for timely interventions.
- **Land Suitability Analysis:** Assesses soil, slope, and climate to match crops to optimal areas.
- **Spatial Analysis:** Maps soil types and land use to guide crop selection and planting strategies.

- **Climate-Smart Agriculture:** Integrate climate data to identify risk zones and plan adaptive strategies for drought or flood-prone areas.
- **Precision Agriculture:** Uses satellite imagery to target fertilizers and pesticides, reducing costs and environmental impact.
- **Land Parcel Management:** Map farm boundaries and land parcels for accurate area measurement, legal documentation, and input planning.
- **Yield Prediction:** Combines historical yield, soil, and weather data to forecast crop yields for better planning.

Conclusion

QGIS stands as a cornerstone of open-source desktop GIS platforms, offering extensive functionality for spatial data visualization, analysis, and cartographic output. With its user-friendly interface and rich suite of tools-from geoprocessing operations like buffering and clipping to spatial statistics and network analysis-QGIS is accessible to both beginners and advanced users. Its robust plugin ecosystem and support for a wide range of data formats further expand its capabilities, making it suitable for diverse applications such as environmental planning, resource management, and urban development. The active global community and wealth of learning resources reinforce its standing as a premier GIS solution. QGIS empowers users to take full control of their geospatial data in an offline environment, making it a vital tool for localized and customized spatial analysis.

References

Gray, J. (2008). Quantum GIS: the open-source geographic information system. Linux Journal, 2008(172), 8.

QGIS Project. (n.d.). *A gentle introduction to GIS*.

Moyroud, N., & Portet, F. (2018). Introduction to QGIS. QGIS and generic tools, 1, 1-17.

https://docs.qgis.org/3.4/en/docs/gentle_gis_introduction.

https://docs.qgis.org/3.4/en/docs/user_manual/index.html

Introduction to Google Earth Engine

Nobin Chandra Paul, Ponnaganti Navyasree, K. Ravi Kumar, Prabhat Kumar and Santosha Rathod
ICAR-National Institute of Abiotic Stress Management, Baramati, Pune-413115
Email: ncp375@gmail.com

What is Google Earth Engine?

Google Earth Engine (GEE) is a cloud-based platform developed by Google for large-scale geospatial and environmental data analysis. It offers access to an extensive collection of satellite imagery and geospatial datasets, paired with robust cloud computing tools for processing data at a planetary scale. After the Landsat series became freely available in 2008, Google archived these datasets and integrated them with its cloud computing infrastructure for open-source use. The platform's data archive now includes imagery from various satellites, as well as GIS-based vector datasets, social, demographic, weather, digital elevation models, and climate data layers.

GEE enables users to perform a wide range of processing tasks on its servers, from cloud correction to machine learning-based analyses, allowing regional-scale data processing in minutes. It also facilitates seamless handling of time-series data. This section will explore how to access and utilize satellite data within the Google Earth Engine editor. With its intuitive and accessible interface, GEE provides an ideal environment for interactive algorithm and data development. Users can upload and manage their own datasets while leveraging Google's cloud resources for processing. This empowers scientists, independent researchers, and nations to analyze vast datasets for tasks like change detection, trend mapping, and resource quantification on Earth's surface in ways previously unattainable. Notably, GEE eliminates the need for high-end computers or advanced software, leveling the playing field so researchers in resource-limited regions can conduct analyses on par with those in more developed countries.

The programming interface of GEE enables users to develop and execute custom algorithms, with computations distributed across multiple processors to significantly accelerate processing times. This allows for efficient global-scale analysis compared to traditional desktop computing. Users can export images from GEE in formats like GeoTIFF, including raw or processed images, map tiles, tables, and videos, to destinations such as Google Drive, Google Cloud Storage, or as new Earth Engine assets. Google Cloud Storage, a paid service, requires users to set up a project, enable billing, and create a storage bucket. Users can also upload their own datasets and choose whether

to share their data or scripts with others. This vast array of multi-temporal, local-to-global data provides researchers with unprecedented opportunities to conduct cost-effective studies with minimal equipment.

GEE's cloud computing capabilities allow for the processing of petabytes of satellite imagery and vector data within its cloud environment, eliminating the need to store or analyze large datasets on local computers. This reduces the demand for high-performance computers with substantial storage capacity. While specialized remote sensing software like ENVI or ERDAS Imagine may still be required for specific tasks (e.g., object-based image analysis) not supported by GEE, the platform removes the need to download satellite imagery—a significant advantage in areas with limited internet speeds. However, an internet connection is still necessary to access and use GEE. The GEE significantly advances the ability to transform vast amounts of Big Data into actionable insights for addressing environmental challenges. It is specifically engineered to handle large datasets, overcoming a key obstacle for researchers working with satellite imagery. GEE offers the added benefit of providing numerous pre-processed data layers, including those corrected for cloud cover, converted to top-of-atmosphere or surface reflectance, and georeferenced. The platform fosters collaboration by enabling users to share code, reducing the need for advanced proficiency in JavaScript or Python, and is supported by a vibrant online community. With its speed, versatility, and accessibility, GEE opens up substantial opportunities for the research community to leverage it for earth observation studies. It addresses common barriers faced by researchers in developing countries, such as limited access to data, funding for hardware and software, and resource constraints. By offering a robust solution, GEE empowers these researchers to conduct impactful studies. The hope is that more researchers in such regions will embrace this platform, leading to the development of innovative applications to better manage our planet's increasingly scarce resources.

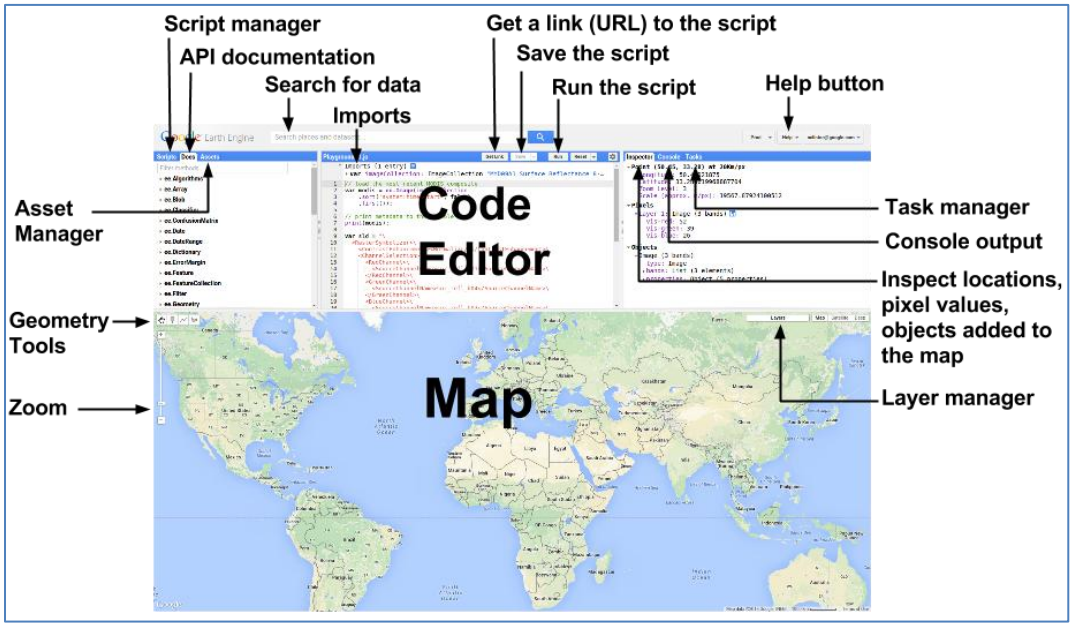
Importance of Google Earth Engine

- **Massive Satellite Archive:** GEE provides direct access to petabytes of satellite data from Landsat, MODIS, Sentinel, and many others without the need for manual downloads.
- **Cloud Computing:** Unlike desktop GIS software like QGIS or ArcGIS, all computations in GEE are performed on Google's cloud infrastructure, saving local computing resources.
- **Real-Time and Historical Analysis:** Users can analyze recent and historical data seamlessly using JavaScript or Python APIs.

- **Supports Environmental Monitoring:** Applications range from deforestation mapping and agricultural monitoring to urban expansion and disaster response.

Why GEE is Advantageous Over Desktop GIS (QGIS/ArcGIS)

Feature	Google Earth Engine	QGIS / ArcGIS Desktop
Data Access	Built-in cloud-based satellite data catalog	Manual download and local storage
Processing Power	High-performance cloud computing	Depends on local hardware
Ease of Sharing	Easily share scripts and outputs	Limited to file-based sharing
Real-time Analysis	Supports near real-time updates	Less efficient for real-time use
Free Access	Free for research & education	ArcGIS is commercial



GEE interface.

Case Study: NDVI Calculation and Classification for Pune District

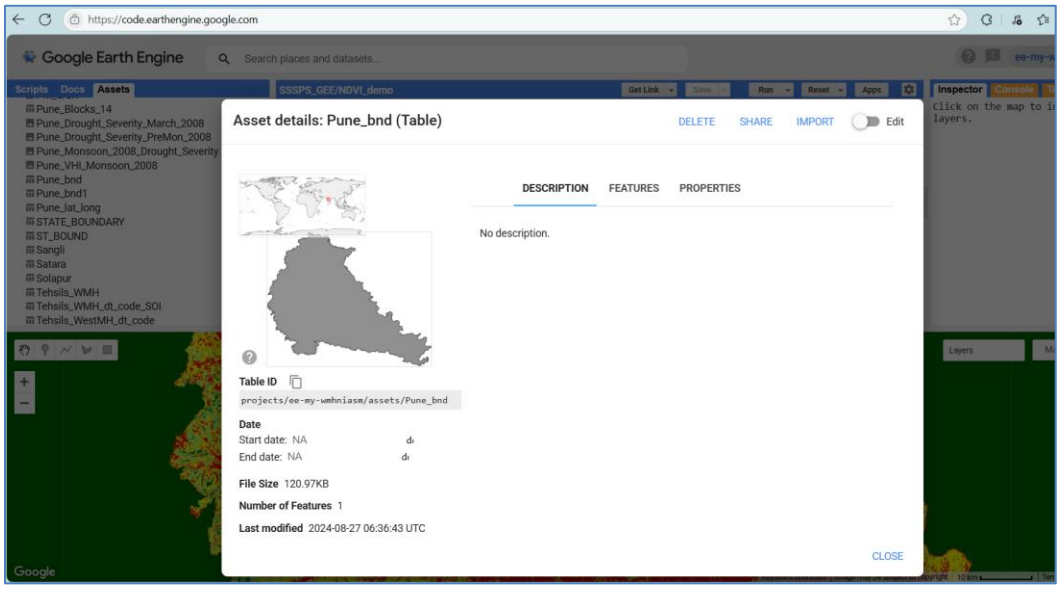
This case study demonstrates how to compute the Normalized Difference Vegetation Index (NDVI) using Sentinel-2 satellite data for the Pune district using Google Earth Engine.

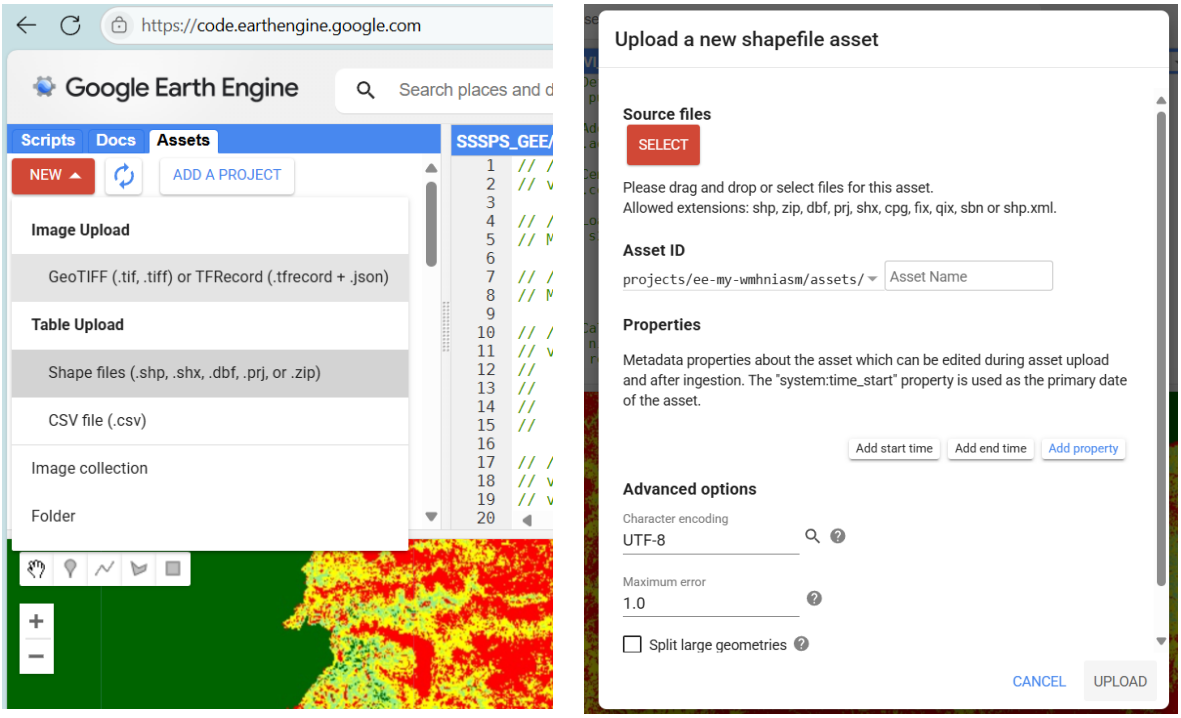
To begin with, create an account using your Gmail ID at code.earthengine.google.com. Once you're inside the Google Earth Engine Code Editor, you can begin working with satellite data and your region of interest.

This chapter uses **Sentinel-2 multispectral imagery** to analyze vegetation health using **NDVI (Normalized Difference Vegetation Index)** over the **Pune district** in Maharashtra, India.

1. **Importing the Study Area**:

   First, you need to upload the shapefile of the study area. Go to the **"Assets"** tab on the left, click on **"New"**, and choose **"Shapefile"**. Upload the zipped shapefile of **Pune district**. Once uploaded, GEE will ingest the file and display the status under the **"Tasks"** tab on the right. After successful upload, refresh the **"Assets"** tab to find your shapefile listed. Import it into your script by clicking on it—by default, it may be named table, but you can rename it (*e.g.*, **pune or aoi**) for clarity.

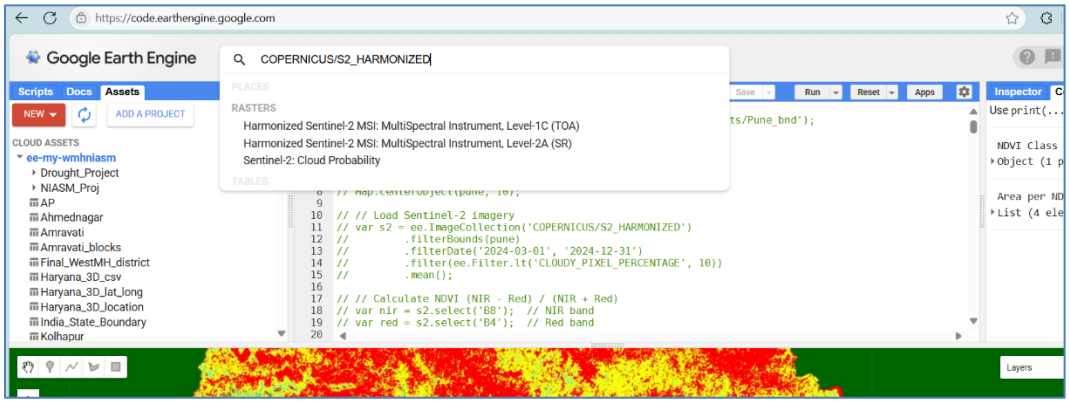




Importing the Pune shapefile to asset

2. Importing Sentinel-2 Satellite Data:

Sentinel-2 data is part of the Copernicus Earth observation program. You can import this data either by coding directly or using the **"Search"** tab to browse the data catalog. In this study, we use the dataset **COPERNICUS/S2_HARMONIZED**, which provides harmonized and atmospherically corrected Sentinel-2 imagery.



3. Filtering the Dataset:

In the script, we filter the Sentinel-2 image collection by:

- **Date** range (March 1, 2024 to July 7, 2024)

- **Cloud cover** (less than 10% cloudiness)
- **Spatial extent**, limited to the Pune district (filterBounds(pune))

After filtering, we use .mean() to create a composite image that averages all selected images over the time period. This helps in reducing noise and minimizing cloud cover.

4. NDVI Calculation:

NDVI is calculated using the Near-Infrared (NIR) band (B8) and the Red band (B4) of Sentinel-2 using the formula: $NDVI = (NIR - Red)/(NIR + Red)$. The resulting NDVI layer gives a continuous index where higher values indicate denser vegetation.

5. Visualizing the Data:

Visualization parameters are defined using a color palette ranging from blue (low NDVI) to dark green (high NDVI). The NDVI layer is then displayed on the GEE map viewer.

6. Classification and Analysis:

The NDVI values are classified into four vegetation classes (Low, Moderate, Dense, Very Dense), and a separate visualization layer is added for this classification. We also compute:

- **Histogram** of NDVI class frequency using zonal statistics
- **Area (in sq. km)** of each NDVI class using pixel area calculations

7. Exporting Results:

Finally, both the raw NDVI image and the classified NDVI map are exported as **GeoTIFF** files to Google Drive for further analysis or use in other GIS platforms.

8. Execution and Output:

Always remember that GEE is case-sensitive. After entering or pasting the code, click the **"Run"** button at the top. Outputs from **print()** statements will appear in the **Console** tab on the right. You can click on image layers and expand their metadata to inspect bands, projection, and tile-specific information.

Step-by-Step Explanation of GEE NDVI Code

This code processes Sentinel-2 satellite imagery to calculate NDVI, classify it into four vegetation density classes, compute zonal and area statistics, and export the results as GeoTIFF files for the Pune district. Below is a detailed breakdown of each step, explaining the purpose and implementation.

1. Load the Region of Interest (ROI)

```
var pune = ee.FeatureCollection('projects/ee-my-wmhniasm/assets/Pune_bnd');
```

- Loads Pune district boundary from your GEE asset storage.

```
js

Map.addLayer(pune, {}, 'pn');
Map.centerObject(pune, 10);
```

- Adds the Pune boundary to the map and centers the view.

2. Load and Filter Sentinel-2 Data

```
var s2 = ee.ImageCollection('COPERNICUS/S2_HARMONIZED')
        .filterBounds(pune)
        .filterDate('2024-03-01', '2024-07-07')
        .filter(ee.Filter.lt('CLOUDY_PIXEL_PERCENTAGE', 10))
        .mean();
```

- Loads Sentinel-2 imagery over Pune between March and July 2024.
- Filters for <10% cloud coverage.
- `.mean()` computes average pixel values across all filtered images to reduce noise and cloud cover.

3. NDVI Calculation

```
var nir = s2.select('B8');  // Near-Infrared Band
var red = s2.select('B4');  // Red Band

var ndvi = nir.subtract(red).divide(nir.add(red)).rename('NDVI').clip(pune);
```

- NDVI formula: **(NIR - Red) / (NIR + Red)**
- Clipped to Pune boundary to reduce unnecessary data processing.

4. Visualize NDVI

```
var vispars = {min: -0.1, max: 0.9, palette: ['blue', 'darkgreen',
'green', 'brown', 'lightyellow']};
Map.addLayer(ndvi, vispars, 'NDVI');
```

5. Classify NDVI into 4 Categories

```
var ndviClasses = ndvi.expression(
  'NDVI < 0.2 ? 1 : NDVI < 0.4 ? 2 : NDVI < 0.6 ? 3 : 4',
  {'NDVI': ndvi}
).rename('NDVI_Class');
```

- Converts continuous NDVI into four classes:
  - **Class 1** (Red): Low vegetation (< 0.2)
  - **Class 2** (Yellow): Moderate vegetation (0.2–0.4)
  - **Class 3** (Light Green): Dense vegetation (0.4–0.6)
  - **Class 4** (Dark Green): Very dense vegetation (> 0.6)

```
var classPalette = ['red', 'yellow', 'lightgreen', 'darkgreen'];
Map.addLayer(ndviClasses, {min: 1, max: 4, palette: classPalette}, 'Pune NDVI Classified');
```

6. Compute Frequency Distribution (Zonal Stats)

```
var zonalStats = ndviClasses.reduceRegion({
  reducer: ee.Reducer.frequencyHistogram(),
  geometry: pune.geometry(),
  scale: 250,
  maxPixels: 2e9,
  bestEffort: true
});
print('NDVI Class Distribution (Histogram)', zonalStats);
```

- Calculates how many pixels fall into each NDVI class.
- Outputs a histogram showing class distribution.

7. Compute Area per NDVI Class

```
var classAreas = ee.Image.pixelArea().divide(1e6).addBands(ndviClasses).rename(['Area', 'NDVI_Class']);

var areaStats = classAreas.reduceRegion({
  reducer: ee.Reducer.sum().group({
    groupField: 1,
    groupName: 'NDVI_Class'
  }),
  geometry: pune.geometry(),
  scale: 250,
  maxPixels: 2e9,
  bestEffort: true
});

```

- Calculates area in square kilometres for each NDVI class using pixel area.
- Output gives area distribution per class.

8. Export NDVI Maps as GeoTIFFs

```
Export.image.toDrive({
  image: ndvi,
  description: 'ndvi_2024_Pune',
  scale: 250,
  region: pune.geometry(),
  crs: 'EPSG:32643',
  maxPixels: 2e9
});
```

- Exports raw NDVI image as GeoTIFF to Google Drive.

```
Export.image.toDrive({
  image: ndviClasses,
  description: 'ndvi_classes_2024_Pune',
  scale: 250,
  region: pune.geometry(),
  crs: 'EPSG:32643',
  maxPixels: 2e9
});
```

- Exports the classified NDVI image as a GeoTIFF for visualization or further GIS analysis.

Complete GEE Code of NDVI:

```
// Define the region of interest (Pune district from asset)
var pune = ee.FeatureCollection('projects/ee-my-wmhniasm/assets/Pune_bnd');

// Add Pune boundary to the map
Map.addLayer(pune, {}, 'pn');

// Center the map on the centroid of the Pune AOI
Map.centerObject(pune, 10);

// Load Sentinel-2 imagery
var s2 = ee.ImageCollection('COPERNICUS/S2_HARMONIZED')
        .filterBounds(pune)
        .filterDate('2024-03-01', '2024-07-07')
        .filter(ee.Filter.lt('CLOUDY_PIXEL_PERCENTAGE', 10))
        .mean();

// Calculate NDVI (NIR - Red) / (NIR + Red)
var nir = s2.select('B8');  // NIR band
var red = s2.select('B4');  // Red band
var ndvi = nir.subtract(red).divide(nir.add(red)).rename('NDVI').clip(pune);

// Define visualization parameters for NDVI
var vispars = {min: -0.1, max: 0.9, palette: ['blue', 'darkgreen', 'green', 'brown', 'lightyellow']};

// Add NDVI layer to the map
Map.addLayer(ndvi, vispars, 'NDVI');

// Classify NDVI into 4 classes
var ndviClasses = ndvi.expression(
  'NDVI < 0.2 ? 1 : ' +  // Class 1: Low vegetation (0 to 0.2)
  'NDVI < 0.4 ? 2 : ' +  // Class 2: Moderate vegetation (0.2 to 0.4)
  'NDVI < 0.6 ? 3 : ' +  // Class 3: Dense vegetation (0.4 to 0.6)
  '4',                    // Class 4: Very dense vegetation (>0.6)
  {'NDVI': ndvi}
).rename('NDVI_Class');
```

```
// Define visualization parameters for NDVI classes
var classPalette = ['red', 'yellow', 'lightgreen', 'darkgreen'];
var classVis = {min: 1, max: 4, palette: classPalette};

// Add classified NDVI layer to GEE map
Map.addLayer(ndviClasses, classVis, 'Pune NDVI Classified');

// Compute zonal statistics (frequency histogram for NDVI classes)
var zonalStats = ndviClasses.reduceRegion({
  reducer: ee.Reducer.frequencyHistogram(),
  geometry: pune.geometry(),
  scale: 250,
  maxPixels: 2e9,
  bestEffort: true
});

// Print zonal statistics to console
print('NDVI Class Distribution (Histogram)', zonalStats);

// Calculate area for each NDVI class
var areaImage = ee.Image.pixelArea().divide(1e6); // Convert to square kilometers
var classAreas = ee.Image.pixelArea().divide(1e6).addBands(ndviClasses).rename(['Area', 'NDVI_Class']);

var areaStats = classAreas.reduceRegion({
  reducer: ee.Reducer.sum().group({
    groupField: 1, // Group by NDVI_Class (second band)
    groupName: 'NDVI_Class'
  }),
  geometry: pune.geometry(),
  scale: 250,
  maxPixels: 2e9,
  bestEffort: true
});
```

```
// Print area statistics to console with error handling
areaStats.evaluate(function(result) {
  if (result && result.groups) {
    // Format the output for clarity
    var formattedAreas = result.groups.map(function(group) {
      return {
        NDVI_Class: group.NDVI_Class,
        Area_sqkm: group.sum // The sum of the area for each class
      };
    });
    print('Area per NDVI Class (sq km)', formattedAreas);
  } else {
    print('Error: Area statistics computation failed or returned no data.');
  }
});

// Export the NDVI image as a GeoTIFF
Export.image.toDrive({
  image: ndvi,
  description: 'ndvi_2024_Pune',
  scale: 250,
  region: pune.geometry(),
  crs: 'EPSG:32643',
  maxPixels: 2e9
});
```

```
// Export the classified NDVI image as a GeoTIFF
Export.image.toDrive({
  image: ndviClasses,
  description: 'ndvi_classes_2024_Pune',
  scale: 250,
  region: pune.geometry(),
  crs: 'EPSG:32643',
  maxPixels: 2e9
});
```

Conclusion

Google Earth Engine (GEE) has revolutionized geospatial analysis by offering a cloud-based platform capable of handling massive datasets, including satellite imagery and climate records. Its ability to process petabytes of data on the fly enables users to perform complex computations-such as NDVI analysis, land cover classification, and change detection-without needing high-end hardware. GEE's integration of machine learning algorithms, global datasets, and time-series analytics makes it an indispensable tool for researchers, environmentalists, and policymakers aiming to monitor and understand Earth system dynamics at scale. Moreover, its collaborative coding environment and export functionalities streamline sharing and reproducibility. In essence,

GEE democratizes access to powerful geospatial analysis, bridging the gap between data and decision-making.

References

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote sensing of Environment, 202, 18-27.

Paul, N. C., Ponnaganti, N., Gaikwad, B. B., Sammi Reddy, K., & Nangare, D. D. (2025). Optimized soil adjusted vegetation index mapping of Pune district using Google Earth Engine. Remote Sensing Letters, 16(7), 728-736.

Paul, N. C., Ponnaganti, N., Reddy, K. S., & Nangare, D. D. (2025). Modified Normalized Difference Water Index Mapping of Pune District Using Google Earth Engine. National Academy Science Letters, 1-6.

Spatial Interpolation and Regression Techniques

Nobin Chandra Paul[1], Ponnaganti Navyasree[1], Pradip Basak[2], K. Ravi Kumar[1], Prabhat Kumar[1] and Santosha Rathod[1]

[1]ICAR-National Institute of Abiotic Stress Management, Baramati, Pune-413115

[2]Uttar Banga Krishi Viswavidyalaya, Cooch Behar, West Bengal

Email: ncp375@gmail.com

1. Introduction

Spatial interpolation and spatial prediction are cornerstone techniques in geospatial analysis, enabling the estimation of values at unsampled locations based on known data points. These methods are critical in fields like agriculture, environmental science, meteorology, urban planning, and geology. This chapter provides a step-by-step explanation of spatial interpolation and prediction, covering definitions, methods (*e.g.,* Inverse Distance Weighting, Kriging, Geographically Weighted Regression), spatial sampling, spatial statistics approaches, advantages, and real-world applications.

2. Spatial Data and Spatial Statistics

Spatial data, also known as geospatial or geo-referenced data, refers to any dataset that includes spatial coordinates associated with each observation, indicating its specific location on the Earth's surface. This type of data encapsulates both locational information (such as latitude and longitude) and attribute information, which describes the characteristics of the spatial feature being represented, for example, a road, lake, mountain, or built-up area. Geospatial data conveys the location, shape, and extent of physical or man-made objects on Earth. The attributes linked to these spatial features provide essential descriptive information, such as type, condition, or value. To manage, visualize, and analyze such data, specialized software tools known as Geographical Information Systems (GIS) are widely used. These platforms support a variety of spatial operations, including mapping, querying, spatial modeling, and data integration. Conceptually, spatial data can be viewed as a realization of a random variable distributed over a two-dimensional space. The field of statistics that addresses such data is termed spatial statistics (Cressie, 1993). Spatial statistics rests on the fundamental principle that observations located closer together in space tend to exhibit more similarity than those farther apart-a concept known as spatial autocorrelation. The increasing availability of spatial data and the growing recognition of spatial context in research have significantly advanced the application of spatial statistical methods,

especially in the agricultural sciences. These methods enable researchers to estimate spatial patterns, such as crop yields, soil nutrient levels (*e.g.*, exchangeable potassium), or soil water infiltration rates, using a relatively small number of samples taken from strategically selected locations (Gupta, 2007).

3. Spatial Interpolation and Spatial Prediction Techniques

Spatial interpolation estimates values of a variable at unsampled locations using known data points with spatial coordinates. It relies on the principle of spatial autocorrelation, where closer points are more likely to have similar values. Its main purpose is to create continuous surfaces (*e.g.*, maps) from discrete data points, such as temperature or soil properties. For example: Estimating rainfall at an unsampled location based on nearby weather station data.

Spatial prediction is a broader term encompassing interpolation and other methods (e.g., regression) to predict values at unsampled locations, often incorporating covariates or spatial statistical models. It includes both deterministic and stochastic approaches. Its purpose is to model spatial phenomena, account for trends or covariates, and provide uncertainty estimates (in stochastic methods). Interpolation focuses on estimating values based solely on observed spatial data. Prediction may include external variables (*e.g.*, elevation, land use) and often quantifies uncertainty. In general, spatial prediction refers to any prediction method that takes into account spatial dependence. Experts in various fields frequently use the terms "estimation" and "prediction" interchangeably. The term "estimation" refers to inferences about the value of fixed but unknown parameters, whereas "prediction" refers to inferences about the value of random variables (Cressie, 1993). Spatial statistical prediction differs from classical prediction in its use of spatial models, which account for spatial relationships, while classical methods do not. Spatial statistics assumes that the variable of interest varies randomly across space, driven by one or more random processes. These processes are captured by models that form the basis for predictions. Kriging, often referred to as optimal prediction, is a spatial prediction technique that minimizes mean-squared error. It typically relies on the second-order properties of a random process, enabling inferences about unobserved values based on the spatial model.

Spatial prediction involves two steps

(i) First model the covariance or semi-variance of the spatial process.

(ii) Use this dependence model to solve the kriging system at a specified set of spatial points, resulting in predicted values and associated standard errors.

Ordinary kriging is a widely used spatial interpolation technique that generates both predicted values and their associated standard errors. It requires a fully defined model of spatial dependence, which describes the spatial process and is characterized by the distance between pairs of locations within the study area. This dependence is typically modeled using a covariance or semi-variogram function. The spatial prediction process involves two key steps: first, defining the covariance or semi-variogram model by selecting its mathematical form and determining the parameter values; second, applying this model to solve the kriging system at specific spatial points, producing predicted values along with their standard errors.

The aim of kriging is to estimate the value of random variable Z at one or more unsampled points say Y_i for the random variable $Z(x_i), i = 1,2,\dots,N$ at nearby locations $\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_N}$. The data can be distributed in one, two, or three dimensions, though most applications in agricultural sciences are two-dimensional. It assumes that the mean is unknown. The value of Z can be estimated at an unknown point x_0 by $\hat{Z}(x_0)$, using the following equations.

$$\hat{Z}(\mathbf{x}_0) = \sum_{i=1}^{N} \lambda_i Z_i$$

where, λ_i are weights subject to condition $\sum_1^N \lambda_i = 1$. The estimate is unbiased, and the expected error is $E\left[\hat{Z}(\mathbf{x}_0) - Z(\mathbf{x}_0)\right] = 0$. The estimation variance is given by

$$V\left[\hat{Z}(\mathbf{x}_0)\right] = E\left[\left\{\hat{Z}(\mathbf{x}_0) - Z(\mathbf{x}_0)\right\}^2\right]$$

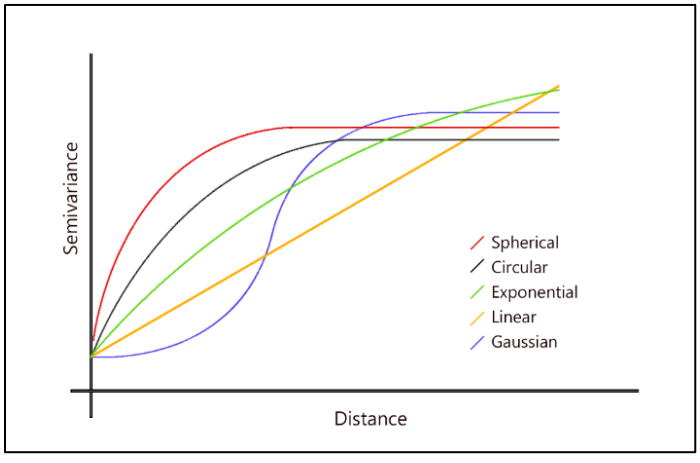


Figure 2. Different theoretical variogram models

4. Types of Interpolation and Spatial Prediction Techniques

Interpolation techniques can be broadly categorized into **deterministic** and **stochastic (geostatistical)** methods, based on whether or not they incorporate the statistical structure and uncertainty of spatial data.

Deterministic methods estimate unknown values using mathematical formulas that are solely based on the distances between sampled and unsampled points, without accounting for the underlying spatial structure or correlation among observations. One of the most widely used deterministic methods is **Inverse Distance Weighting (IDW)**. This technique assumes that points closer to the prediction location exert a stronger influence than those farther away. The predicted value at an unsampled location is calculated as a weighted average of nearby known values, where the weights are inversely proportional to the distance raised to a power. Specifically, the formula is given by:

$$Z(x_0) = \frac{\sum_{i=1}^{n} w_i Z(x_i)}{\sum_{i=1}^{n} w_i}, \text{ where, } w_i = \frac{1}{d_i^p}$$

Here, $Z(x_0)$ is the predicted value at location x_0, $Z(x_i)$ are the observed values at known points, d_i is the distance from the known point i to the prediction location, and p is the power parameter that controls how quickly the influence of a point decreases with distance. Although IDW is simple and computationally efficient, it does not account for spatial autocorrelation or provide any measure of prediction uncertainty.

In contrast, **stochastic or geostatistical methods** not only use distance but also incorporate the **spatial autocorrelation** among observations by modeling the spatial structure statistically. These methods rely on the concept of a **variogram**, which quantifies how the similarity between observations decreases with increasing distance. The **empirical variogram** is computed from the data as:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2$$

where $\gamma(h)$ represents the semi-variance for a given lag distance h, and $N(h)$ is the number of observation pairs at that distance. A **theoretical variogram model**-such as **spherical**, **exponential**, or **Gaussian**-is then fitted to the empirical variogram. These models are characterized by key parameters: the **nugget** (which represents measurement error or microscale variability), the **sill** (total variance), and the **range** (the distance beyond which spatial autocorrelation becomes negligible).

Once the variogram is established, various forms of **kriging**-a family of advanced geostatistical interpolation methods-can be applied. **Simple Kriging (SK)** assumes a known and constant mean throughout the study area and uses the variogram to compute optimal weights for interpolation. **Ordinary Kriging (OK)**, the most commonly used kriging method, relaxes this assumption and considers the mean to be unknown but locally constant. **Universal Kriging (UK)** goes a step further by assuming that the mean varies systematically over space (*e.g.*, following a linear trend), and it models the residuals using kriging techniques. Lastly, **Regression Kriging (RK)** combines a multiple linear regression model with kriging of the residuals, enabling the incorporation of auxiliary variables such as elevation, land use, or vegetation indices (*e.g.*, NDVI). This method is especially useful in scenarios where additional explanatory data are available and can enhance the accuracy of spatial predictions. Overall, stochastic methods like kriging offer the dual benefits of modeling spatial structure and providing estimates of prediction uncertainty, making them a powerful choice for spatial analysis in environmental and agricultural sciences.

Deterministic vs. Stochastic Approaches

Aspect	**Deterministic Methods**	**Stochastic Methods**
Definition	Predict unknown values from known points based on mathematical or geometric rules	Use statistical models to incorporate spatial autocorrelation and quantify uncertainty
Examples	Inverse Distance Weighting (IDW), Radial Basis Function	Kriging (Simple, Ordinary, Universal), Regression Kriging
Assumptions	No random error or spatial uncertainty; relies on distance or fixed trends	Assumes spatial correlation and includes a stochastic component

Weighting Mechanism	Based on inverse distance or functions of distance	Derived from semi-variogram/covariance models
Error Estimation	No formal method for quantifying prediction error	Provides a kriging variance (prediction uncertainty)
Data Requirement	Requires only the variable of interest	Requires variogram modeling and may include auxiliary variables
Best Use Case	When data is dense and uniform; exploratory visualization	When spatial dependence is present and error quantification is important

5. Geographically Weighted Regression

In agricultural surveys, parameters of interest are often tied to specific geographic locations, making them spatial data. Due to the homogeneity of neighboring units, this data exhibits spatial autocorrelation, meaning it is not independent. Traditional survey estimation methods for populations rarely incorporate this geographic information. However, spatially referenced data can be integrated to provide more detailed insights. In model-based prediction for survey sampling, the finite population total (or mean) is estimated using a regression model with specific assumptions. In a linear regression model-based approach, when the parameter is geographic, the regression coefficient varies across space, a phenomenon known as spatial non-stationarity. This violates the assumption of independent population units when observations are spatially correlated, as location is not accounted for in standard modeling. To address spatial non-stationarity, Brunsdon et al. (1996) introduced geographically weighted regression (GWR), a local spatial statistical method that models spatially varying relationships. Unlike ordinary least squares (OLS), where coefficients are fixed, GWR coefficients vary by spatial location. This approach effectively handles spatial non-stationarity, leading to improved parameter estimates. Let, $k_i(\text{latitude}_i, \text{longitude}_i)$ denotes the geographical location of i^{th} unit in space. We can define a GWR model as

$$y_i = \beta_o(k_i) + \sum_{l=1}^{p} \beta_l(k_i)\, x_{il} + e_i \quad ; i = 1, 2, ..., N \quad ; \; l = 1, 2, ..., \; p$$

where, y_i is the dependent variable at location 'k_i', $\beta_0(k_i)$ is the intercept parameter at location point 'k_i', $\beta_l(k_i)$ represents the coefficient of l^{th} independent variable at location 'k_i', x_{il} is the value of l^{th}auxiliary variable at location 'k_i' and e_i is the independent and identically distributed random error term with mean '0' and constant variance σ^2.

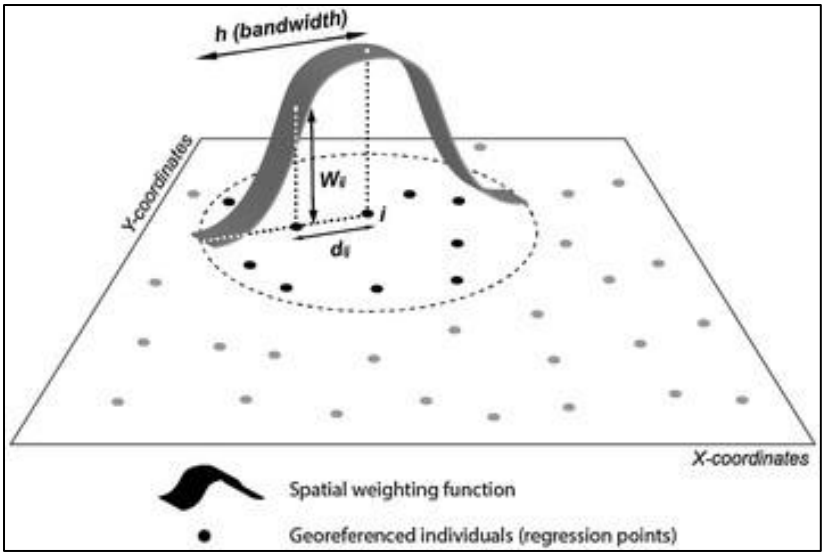


Figure 4. Schematic representation of the GWR model along with its spatial parameters. Source: Adapted from Feuillet et al. (2015)

6. Applications of Spatial Interpolation & Prediction Techniques

6. 1. Agriculture

- **Soil Property Mapping:** Spatial interpolation helps map soil parameters like pH, nitrogen, phosphorus, and potassium using kriging and regression kriging.
    - Benefit: Enables site-specific nutrient management (precision agriculture).
    - Tools: GWR, Regression Kriging with remote sensing data (e.g., NDVI, LST).
- **Yield Prediction:**
    - Combine historical yield data with remote sensing covariates.
    - Use regression kriging to predict crop yield at fine scales.

6. 2. Hydrology and Groundwater Studies

- **Groundwater Level Interpolation:**
    - Kriging is widely used to interpolate water table depth across wells.
    - Ordinary Kriging and Universal Kriging help identify potential zones of water scarcity.
- **Water Quality Mapping:**
    - Parameters like salinity, nitrate, TDS can be spatially interpolated.

- Supports aquifer vulnerability assessments and sustainable groundwater management.

6.3. Meteorology and Climatology

- **Rainfall Estimation:**
  - Kriging and co-kriging used for interpolating rainfall between weather stations.
  - Regression Kriging can integrate elevation, temperature, and wind data for improved results.
- **Temperature Surface Modeling:**
  - GWR or Universal Kriging models are used to generate spatial temperature maps incorporating altitude and land cover.

6.4. Ecology and Environmental Monitoring

- **Species Distribution Modeling:**
  - Interpolating biodiversity indices, species richness using spatial statistics.
  - IDW and kriging both used depending on scale and species behavior.
- **Air Quality Monitoring:**
  - Mapping pollutants (PM2.5) with sparse sensor data.
  - Kriging and GWR incorporate meteorological and traffic data to enhance estimation.

6.5. Urban and Regional Planning

- **Land Surface Temperature (LST) Mapping:**
  - Using satellite-derived LST and regression kriging to understand urban heat islands.
- **Noise Pollution Mapping:**
  - Spatial interpolation of sound levels across cities using IDW or OK.
  - Helps in zoning and regulatory planning.
- **Population Density Estimation:**
  - Interpolation based on census point data or mobile phone data using GWR or co-kriging.

6.6. Public Health and Epidemiology

- **Disease Incidence Mapping:**

- Interpolation of disease cases (e.g., dengue, malaria, COVID-19) across health centers.
- GWR used to assess how socioeconomic and environmental factors influence disease spread.

6.7. Geology and Natural Resource Management

- **Mineral Prospecting:**
  - Kriging and co-kriging used to model ore grade or mineral concentration over regions.
  - Integrates field samples with geophysical or geochemical covariates.
- **Soil Erosion Estimation:**
  - Regression kriging with terrain, land use, and rainfall data to estimate erosion risk zones.

Practical: R Code: Spatial Interpolation & GWR

```
## Load necessary packages ##
install.packages(c("sp", "rgdal", "gstat", "automap", "GWmodel", "terra", "caret"))
library(sp)
library(rgdal)
library(gstat)
library(automap)
library(GWmodel)
library(terra)
library(caret)
# Step 1: Load your SOC data and prepare the spatial data
# Replace 'your_data.csv' with your actual data file path
your_data <- read.csv("your_data.csv")
coordinates(your_data) <- ~Longitude + Latitude
proj4string(your_data) <- CRS("+proj=longlat +datum=WGS84")

# Step 2: Create a prediction grid for the study area
grid <- expand.grid(
  Longitude = seq(min(your_data$Longitude), max(your_data$Longitude), by = 0.01),
  Latitude = seq(min(your_data$Latitude), max(your_data$Latitude), by = 0.01)
)
coordinates(grid) <- ~Longitude + Latitude
proj4string(grid) <- CRS("+proj=longlat +datum=WGS84")

# Step 3: Fit the Geographically Weighted Regression (GWR) model
```

```
# Determine optimal bandwidth for GWR
bw_gwr <- bw.gwr(SOC ~ elevation + slope + land_use, data = your_data, approach = "AICc", adaptive = TRUE)
# Fit the GWR model
gwr_model <- gwr.basic(SOC ~ elevation + slope + land_use, data = your_data, bw = bw_gwr, adaptive = TRUE)
# Make predictions using GWR
gwr_pred <- predict(gwr_model, newdata = grid)
# Step 4: Calculate residuals for GWR-Kriging
your_data$residuals <- gwr_model$SDF$gwr.e
# Fit a variogram model to the residuals
vgm_residuals <- autofitVariogram(residuals ~ 1, your_data)

# Perform Ordinary Kriging on the residuals
residuals_kriging <- krige(residuals ~ 1, locations = your_data, newdata = grid, model = vgm_residuals$var_model)

# Combine GWR predictions and kriged residuals to get GWR-Kriging predictions
gwr_kriging_pred <- gwr_pred$SDF$pred + residuals_kriging$var1.pred

# Step 5: Ordinary Kriging
vgm_model <- autofitVariogram(SOC ~ 1, your_data)
kriging_pred <- krige(SOC ~ 1, your_data, grid, model = vgm_model$var_model)

# Step 6: Regression Kriging
reg_krige_model <- krige(SOC ~ elevation + slope + land_use, your_data, grid, model = vgm_model$var_model)

# Step 7: Inverse Distance Weighting (IDW)
idw_model <- idw(SOC ~ 1, your_data, grid, idp = 2)

# Step 8: Performance Metrics Calculation
# Assuming validation_data has true SOC values and predicted SOC values for each model
# Replace with your actual validation data

calculate_metrics <- function(true, pred) {
  rmse <- RMSE(pred, true)
  r2 <- R2(pred, true)
  mae <- MAE(pred, true)
  return(c(RMSE = rmse, R2 = r2, MAE = mae))
}
# Placeholder for validation data (replace with actual validation data)
```

```
validation_data <- your_data # Use a separate validation dataset in practice

metrics_kriging <- calculate_metrics(validation_data$SOC, kriging_pred$var1.pred)
metrics_gwr <- calculate_metrics(validation_data$SOC, gwr_pred$SDF$pred)
metrics_gwr_rk <- calculate_metrics(validation_data$SOC, gwr_kriging_pred)
metrics_reg_kriging <- calculate_metrics(validation_data$SOC, reg_krige_model$var1.pred)
metrics_idw <- calculate_metrics(validation_data$SOC, idw_model$var1.pred)

# Step 9: Create Prediction Maps and Save as TIFF files
# Create SpatialPixelsDataFrame for each prediction
grid@data$kriging <- kriging_pred$var1.pred
grid@data$gwr <- gwr_pred$SDF$pred
grid@data$gwr_rk <- gwr_kriging_pred
grid@data$reg_kriging <- reg_krige_model$var1.pred
grid@data$idw <- idw_model$var1.pred

# Convert to raster layers
kriging_raster <- raster(grid, layer = "kriging")
gwr_raster <- raster(grid, layer = "gwr")
gwr_rk_raster <- raster(grid, layer = "gwr_rk")
reg_kriging_raster <- raster(grid, layer = "reg_kriging")
idw_raster <- raster(grid, layer = "idw")

# Save as .tiff files
writeRaster(kriging_raster, "kriging_SOC.tiff", format = "GTiff", overwrite = TRUE)
writeRaster(gwr_raster, "gwr_SOC.tiff", format = "GTiff", overwrite = TRUE)
writeRaster(gwr_rk_raster, "gwr_rk_SOC.tiff", format = "GTiff", overwrite = TRUE)
writeRaster(reg_kriging_raster, "reg_kriging_SOC.tiff", format = "GTiff", overwrite = TRUE)
writeRaster(idw_raster, "idw_SOC.tiff", format = "GTiff", overwrite = TRUE)

# Step 10: Compare and Visualize Results
comparison_metrics <- data.frame(
  Method = c("Ordinary Kriging", "GWR", "GWR-Kriging", "Regression Kriging", "IDW"),
  RMSE = c(metrics_kriging[1], metrics_gwr[1], metrics_gwr_rk[1], metrics_reg_kriging[1],
metrics_idw[1]),
  R2 = c(metrics_kriging[2], metrics_gwr[2], metrics_gwr_rk[2], metrics_reg_kriging[2],
metrics_idw[2]),
  MAE = c(metrics_kriging[3], metrics_gwr[3], metrics_gwr_rk[3], metrics_reg_kriging[3],
metrics_idw[3])
)
# Print the comparison metrics
print(comparison_metrics)
```

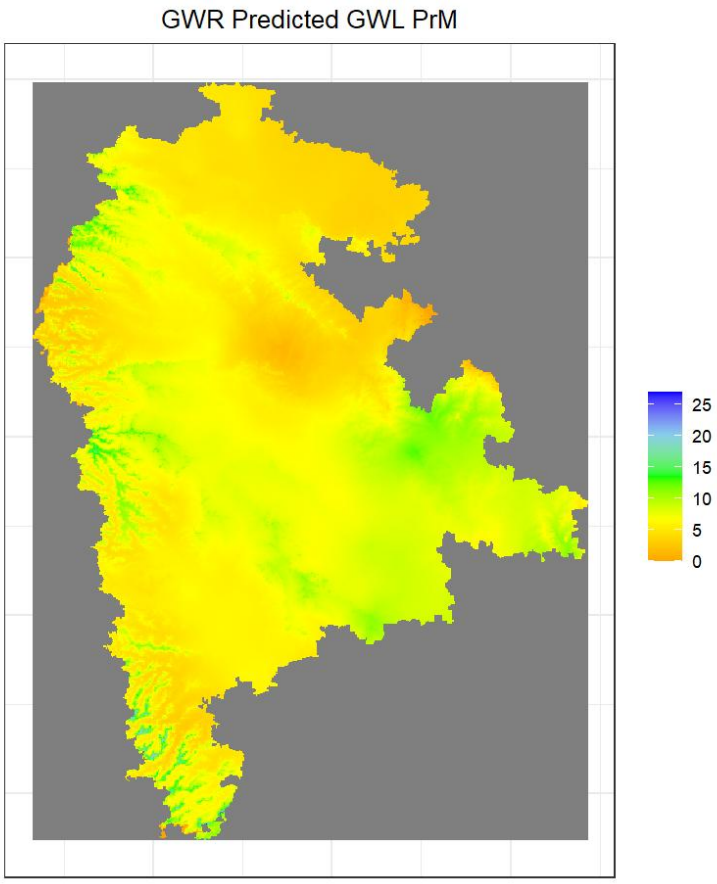


Figure 5. Prediction of ground water level using GWR.

Conclusion

Spatial sampling, interpolation, and Kriging are pivotal for geospatial analysis, enabling accurate predictions at unsampled locations. Spatial sampling optimizes data collection by capturing spatial variability, ensuring representative datasets for reliable modeling. Interpolation, such as IDW, creates continuous surfaces from discrete data, while Kriging (Simple, Ordinary, Universal, Regression) enhances predictions by modeling spatial autocorrelation and providing uncertainty estimates. These methods support critical applications like digital soil mapping, crop health monitoring, environmental monitoring, meteorology, and urban planning. Challenges include designing cost-effective sampling, handling assumptions (*e.g.*, stationarity), and computational complexity. Future advancements in machine learning and remote sensing promise enhanced precision and real-time applications, solidifying their role in data-driven spatial decision-making.

References:

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial non-stationarity. *Geographical Analysis,* **28**, 281-298.

Brunsdon, C., Fotheringham, S., & Charlton, M. (1998). Geographically weighted regression-modelling spatial non-stationary. *The Statistician*, **47(3)**, 431-443.

Cochran, W.G. (1977). *Sampling Techniques*. Third edition. New York, N, Y., John Wiley and Sons, Inc.

Cressie, N. A. C. (1993). *Statistics for spatial data*. Wiley, New York.

Feuillet, T., Charreire, H., Menai, M. *et al*. (2015). Spatial heterogeneity of the relationships between environmental characteristics and active commuting: towards a locally varying social ecological model. *International Journal of Health Geographics*, **14(1)**, 1-14.

Gupta, N. K. (2007). On Spatial Prediction Modelling. *Ph.D. Thesis*, ICAR-IARI, New Delhi.

Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & geosciences, 32*(9), 1378-1388.

Moran, P. A. P. (1948). The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society*, Series B (Methodological), **10(2)**, 243-251.

Moran., P.A.P. (1950). Notes on Continuous Stochastic Phenomena. *Biometrika*, **37**, 17-23.

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers and Geosciences*, **30(7)**, 683-691.

https://dickbrus.github.io/SpatialSamplingwithR/

12 Spatial Interpolation – Spatial Data Science

Chapter 13 Spatial interpolation methods | Spatial Statistics for Data Science: Theory and Practice with R.

Introduction to Sampling Theory and Spatial Sampling Strategy

Pradip Basak[1] and Nobin Chandra Paul[2]
[1]Uttar Banga Krishi Viswavidyalaya, Cooch Behar, West Bengal
[2]ICAR-National Institute of Abiotic Stress Management, Baramati, Pune-413115
Email: pradipbasak.99@gmail.com

1. Introduction

Sample surveys are the indispensable tool to know the characteristics of a poulation. A scientifically designed sample survey is a true represebtattive of the poulation. Thus, sample survey is a data collection technique by which a portion (sample) of the larger group (population) is selected to infer the characteristics of that group. Sample surveys are cost-effective, time-saving, and often more practical than complete enumeration (census), especially when dealing with large populations.

2. Definitions and Basic Concepts

2.1. Population

Population is a collection of all units of a specified type in a given region at a particular point of time. Population is also referred to as universe and denoted as U, and the size of the population is denoted by N. Example - A population of farmers, livestocks in a region or a population of trees or birds in a forest etc. depending on the objectives of the survey.

2.2. Population Parameter

Population parameter is a real valued function of the population observations. Let, a finite population $U=(U_1, U_2,\ldots,U_N)$ consists of N units and Y_i be the value of characteristic under study, y corresponding to i^{th} unit of the population, $\forall\ i=1, 2, \ldots, N$. The unit may be a farm household and the characteristic under study may be the income of that household. The parameters that are of major interest in sample surveys, defined as follows:

Population mean, $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$,

Population proportion, $P=N_1/N$,

Population mean square, $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2$,and

Coefficient of variation, $\text{CV}=\frac{S}{\bar{Y}}$.

where, N_1 being the number of population units possessing a particular attribute, say A, out of N population units when the characteristcs of interest is binary in nature.

2.3. Sample

A sample is a true representative of the population. It is used for inferring about the population parameters through estimators. A single element or group of elements considered for selection is

known as sampling unit. The sample mean is the unbiased estimator of population mean whereas sample proportion is the unbiased estimator for population proportion. These estimators are defined as follows:

Sample mean, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$,

Sample proportion, $p=n_1/n$,

Sample mean square, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, and

Estimator of coefficient of variation, $\hat{CV} = \frac{s}{\bar{y}}$.

where, n_1 being the number of sample units possessing a particular attribute, say A, out of n sample units. Sometimes, CV of the population is known from previous surveys or past records. However, such type of information may not be always readily available. In this situation, a small preliminary sample (n_1) is to be drawn from the population to estimate CV by the ratio of the square root of sample mean square (s^2) to the sample mean ($\bar{y}$).

2.4. Sampling Frame

Sampling frame is a list of all the sampling units present in the population along with their identification particulars. Example: a list of the farm households in a village constitutes a sampling frame when objective of the survey is to estimate the income of farmers in that village. The sampling frame should be up to date, and deletion and duplication of sampling units should not be there.

2.5. Estimator and statistic

An estimator is a real valued function of the sample observations that is used for estimating population parameter, e.g. suppose a sample of size n is selected from a population of size N, then, sample mean, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is an estimator of population mean. Any real valued function of the sample observations is known as statistic. It may or may not be an estimator, e.g. sample median is statistic but not an estimator for population mean. All estimators are statistic but all statistic are not estimators. The value of the estimator differs from sample to sample. The particular value of the estimator for a given sample is known as estimate.

3. Objectives of Sample Survey

a) To collect data that represent the characteristics of a population.
b) To estimate population parameters (e.g., mean income, unemployment rate).
c) To make decisions based on sample data with acceptable confidence and precision.
d) To track changes over time (trend analysis) in repeated surveys.

4. Types of Sample Surveys

- Cross-sectional Surveys: Data is collected from the sampling units at a particular point of time.
- Longitudinal or Panel Surveys: Data is collected from the same sampling unit repeatedly over a period of time.

5. Types of Sampling Methods

5.1. Probability Sampling

When probability is associated with the selection of the sampling units, then it is referred to as probability or random sampling. Example - Simple Random Sampling (SRS), Systematic Sampling, Two-stage Sampling etc.

5.2. Non-Probability Sampling

When selection of sampling units is not governed by laws of probability, then it is referred to as non-probability or non-random sampling. Example - Snowball sampling, Convenience sampling etc. Non-probability sampling does not gurantee valid estimates of the population parameters as probability is not associated with the selection of the sampling units.

6. Survey Design Components

1. Defining Objectives and Population
2. Questionnaire Design
3. Mode of Data Collection
4. Training of Enumerators
5. Supervision and Quality Control

7. Errors in Sample Surveys

7.1. Sampling Error

The difference between the estimate and true value of population parameter is known as sampling error or sampling fluctuation. The sampling error is present in every sampling design. The sampling design with lower sampling error is preferable and it is inversely proportional to the sqaure root of sample size.

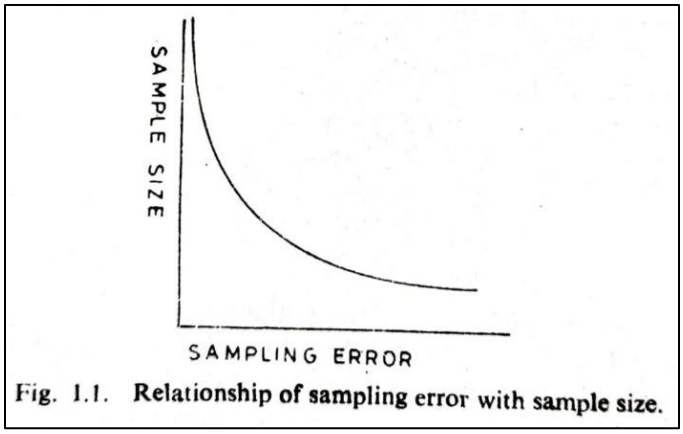


Fig. 1.1. Relationship of sampling error with sample size.

7.2. Non-Sampling Error

The errors in the estimate other than the sampling error, grouped together, are known as non-sampling errors. The major sources of non-sampling errors are: response error, non-response error, measurement error, errors introduced by typing mistakes and editing. The non-sampling error is directly proportional to the sample size.

8. Advantages of Sample Surveys

a) Sample surveys are generally more cost-effective and time-efficient compared to census surveys, enabling quicker collection of the required information.

b) However, cost savings are not the sole reason for choosing a sample survey.

c) Ensuring an acceptable level of accuracy in the results is equally crucial.

d) Sometimes, sample surveys are conducted to validate findings from a census.

e) In certain cases, a well-designed and properly executed sample survey can yield results that are even more precise than those from a full census..

9. Sample Size Determination

Determining the appropriate sample size is one of the most crucial aspects of survey planning. The goal is to select a sample size that ensures the resulting estimates are both reliable and efficient. An excessively large sample may lead to unnecessary expenditure of time, effort, and resources, making the survey economically and operationally inefficient. On the other hand, a sample that is too small may yield estimates that lack reliability. This challenge can be addressed through the application of sample survey theory. The precision of an estimate is typically assessed by the allowable margin of error and the confidence level desired to ensure that the estimate falls within this margin. Several methods exist for calculating the sample size, depending on the availability

of prior information, such as the population's coefficient of variation (CV) and the sampling design employed.

The sampling error in estimation of population mean is defined as $(\bar{y} - \bar{Y})$ whereas for population proportion it is (*p*-*P*). The permissible margin of error in the estimation of population parameters is denoted by β. In the estimation of population mean, $\beta = |\bar{y} - \bar{Y}|$ whereas for population proportion, $\beta = |p - P|$. When permissible margin of error (β) is expressed as a fraction of true population parameter value i.e. says 10% (or 5%) of population mean or population proportion, then it is referred to as relative error (ε) and it is defined as,

$$\varepsilon = \left| \frac{\bar{y} - \bar{Y}}{\bar{Y}} \right| \text{ (for population mean), and}$$

$$\varepsilon = \left| \frac{p - P}{P} \right| \text{ (for population proportion).}$$

When a permissible margin of error is specified, say, 150, it means that this amount must be added to and subtracted from the estimated value of a population parameter to construct an interval where the true value is expected to lie. For instance, if a survey estimates wheat productivity in a region at 3,500 kg/ha with a margin of error of 150, the actual productivity is likely to fall between 3,350 and 3,650 kg/ha. Similarly, for population proportions, the margin of error indicates the range around the estimated proportion. If a survey estimates that 60% of people in a region consume rice, and the margin of error is 5%, then the true proportion is likely to lie between 55% and 65%.

Besides the margin of error (also called relative error), determining the appropriate sample size also depends on the confidence level, which reflects how certain we are that the estimated interval contains the true population parameter. The confidence level is expressed as (1 − α), where α is the level of significance. The level of significance (α) is the probability of incorrectly rejecting a true null hypothesis and is commonly set at 1%, 5%, or 10%, depending on the desired precision. A confidence level of 95%, for example, indicates a 95% probability that the population parameter falls within the estimated range derived from the sample. While higher confidence levels provide greater assurance, they also require larger sample sizes, and thus, leading to increased cost of survey. For this reason, confidence levels of 90% or 95% are typically used in practice.

9.1. Methodology of sample size determination

The availability of prior information about the population coefficient of variation (CV) plays a key role in determining the appropriate approach for sample size determination. In cases where no prior information on the population CV is available, a small preliminary sample of size n_1 is first selected to estimate the CV. This estimated CV is then used to calculate the required final sample

size n. If the final sample size n exceeds the preliminary sample size n_1, then an additional $m=n-n_1$ units are selected from the remaining population to augment the sample, ensuring the total sample size equals n. Conversely, if the preliminary sample size n_1 is larger than the final sample size n, then the initial sample of size n_1 is retained and treated as the final sample.

9.2 Sample Size Determination for Estimating the Population Mean under SRSWOR Sampling Design

Case 1: Population CV is not known

In this case, a preliminary sample of size n_1 is drawn using simple random sampling without replacement (SRSWOR) design and the study variable, y, is observed on n_1 units. The estimated CV from the preliminary sample is then used for determining the final sample size. For determination of sample size, the following information are required: (a) Confidence Level (1-α) (in %), (b) Relative error (0 - 1), (c) Population Size (N), (d) Size of the preliminary sample (n_1), and (e) y-values obtained from the preliminary sample drawn using SRSWOR. Let, y_1, y_2, …, y_{n_1} be the values of the study variable y in the preliminary sample. Then,

Sample mean, $\bar{y}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$,

Sample mean square, $s_{n_1}^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(y_i - \bar{y}_{n_1}\right)^2$, and

Estimator of CV, $\hat{\text{CV}} = \frac{s_{n_1}}{\bar{y}_{n_1}}$.

The resulting sample size is obtained as follows:

$$n = \frac{N\left(\hat{\text{CV}}\right)^2}{\frac{N\varepsilon^2}{t^2_{\alpha/2,\, n_1-1}} + \left(\hat{\text{CV}}\right)^2}.$$

Example 1

An investigator is interested in estimating the average number of tractors per village in a particular community development block of a district. The block consists of 70 villages. A sample of 10 villages is selected in the preliminary sample using SRSWOR and the number of

tractors is obtained as 14, 8, 15, 39, 9, 11, 19, 12, 18, and 22. Determine the sample size needed to estimate the average number of tractors per village with relative error of 0.20 with 95% confidence level.

Solution:

Given, $N = 70$, $n_1 = 10$, $\alpha = 0.05$, $\varepsilon = 0.20$.

Here,

Sample mean, $\bar{y}_{n_1} = 16.70$,

Sample mean square, $s^2_{n_1} = 81.34$,

Estimate of CV, $\hat{\text{CV}} = \dfrac{\sqrt{81.34}}{16.70} = 0.54$, and

$t_{(0.025,\ 9)} = 2.262$.

The resulting sample size is obtained as follows:

$$n = \frac{70 \times (0.54)^2}{\dfrac{70 \times 0.20^2}{(2.262)^2} + (0.54)^2},$$

$$n = 24.33 \approx 24.$$

Since 10 villages are already selected in the preliminary sample, therefore, 14 additional villages need to be selected from the remaining 60 villages using SRSWOR design to get the desired final sample of 24 villages.

Case 2: Population CV is known

In this case, it is assumed that an approximate value of the population CV is known either from past surveys or from prior records and this value is being used for determining the sample size. For determination of sample size, the following information are required: (a) Confidence Level (1-α) (in %), (b) Relative error (0 - 1), (c) Population Size (*N*), and (d) CV. The sample size is determined as,

$$n = \frac{N(\text{CV})^2}{\dfrac{N\varepsilon^2}{Z^2_{\alpha/2}} + (\text{CV})^2}.$$

Example 2

Suppose one investigator is interested to estimate the average value of output from a group of 1000 sugar mills in a region so that the sample estimate lies within 0.10 of the true value with a confidence level of 95%. Determine the sample size required in this case assuming SRSWOR design. The population coefficient of variation is known to be 0.60.

Solution:

Given, N=1000, α = 0.05, ε = 0.10, CV=0.60, and $Z_{0.025}$=1.96.

The resulting sample size is obtained as follows:

$$n=\frac{1000\times(0.60)^2}{\frac{1000\times 0.10^2}{(1.96)^2}+(0.60)^2},$$

$$n=121.49\approx 121.$$

9.3 Sample Size Determination for Estimating the Population Mean under SRSWR Sampling Design

Case 1: Population CV is not known

Here, a preliminary sample of size n_1 is selected using simple random sampling with replacement (SRSWR) design. The estimated CV is then used for determining the sample size. For determination of sample size, the following information are required: (a) Confidence Level (1-α) (in %), (b) Relative error (0 - 1), (c) Size of the preliminary sample (n_1), and (d) y-values obtained from the preliminary sample. Let, y_1, y_2, …, y_{n_1} be the values of the study variable y in the preliminary sample. Then,

Sample mean, $\bar{y}_{n_1}=\frac{1}{n_1}\sum_{i=1}^{n_1} y_i$,

Sample mean square, $s_{n_1}^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(y_i-\bar{y}_{n_1}\right)^2$, and

Estimator of CV, $\hat{CV}=\frac{s_{n_1}}{\bar{y}_{n_1}}$.

The resulting sample size is obtained as follows:

$$n=\frac{\left(\hat{\text{CV}}\right)^2 t^2_{\alpha/2,\, n_1-1}}{\varepsilon^2},$$

$$\Rightarrow n=\left(\frac{\hat{\text{CV}}\times t_{\alpha/2,\, n_1-1}}{\varepsilon}\right)^2.$$

Example 3

An investigator is interested in estimating the average number of fertilizers selling shops per village in a particular district. A preliminary sample of 10 villages is selected using SRSWR design. The values of number of fertilizers selling shops in the selected villages are: 12, 14, 8, 5, 36, 24, 18, 17, 6, 9. Determine the sample size required to estimate the average number of fertilizers selling shops per village with relative error of 0.15 with 95% confidence level.

Solution:

Given, $n_1 = 10$, $\alpha = 0.05$, and $\varepsilon = 0.15$.

Here, Sample mean, $\bar{y}_{n_1} = 14.90$,

Sample mean square, $s^2_{n_1} = 90.10$,

Estimate of CV, $\hat{\text{CV}}=\frac{\sqrt{90.10}}{14.90} = 0.64$, and $t_{(0.025,\, 9)} = 2.262$.

The resulting sample size is obtained as follows:

$$n=\left(\frac{0.64\times 2.262}{0.15}\right)^2,$$

$$n=93.15\approx 93.$$

Since 10 villages are already selected in the preliminary sample, therefore, 83 additional villages need to be selected from the remaining villages in the district using SRSWR design to get the desired final sample size.

Case 2: Population CV is known

In this case, some approximate value of the population *CV* is known and this value is being used for determination of the sample size. For determination of sample size, the following information are required: (*a*) Confidence Level (1-α) (in %), (*b*) Relative error (0 - 1), and (*c*) CV. The sample size is obtained as,

$$n=\frac{(\mathrm{CV})^2\left(Z_{\alpha/2}\right)^2}{\varepsilon^2},$$

$$\Rightarrow n=\left(\frac{\mathrm{CV}\times Z_{\alpha/2}}{\varepsilon}\right)^2.$$

Example 4

An investigator is interested in estimating the average number of animal farms per village in a particular district. The population coefficient of variation is known to be 0.50. Determine the sample size needed to estimate the average number of animal farms per village with relative error of 0.20 with 95% confidence level assuming SRSWR design.

Solution:

Given, $CV = 0.50$, $\varepsilon = 0.20$, $\alpha = 0.05$, $Z_{0.025} = 1.96$.

The resulting sample size is obtained as follows:

$$n=\left(\frac{0.50\times 1.96}{0.20}\right)^2,$$

$$n=24.01\approx 24.$$

9.4 Sample Size Determination for Estimating the Population Proportion under SRSWOR Sampling Design

Case 1: Population proportion is not known

Here, the population proportion is assumed to be unknown and it is estimated by drawing a preliminary sample on the study variable, *y*, using SRSWOR design. The estimated population proportion is then used for determination of the sample size. For determination of sample size, the following information are required: (a) Confidence Level (1-α) (in %), (b) Relative error (0 - 1),

(c) Population Size (*N*), (d) Size of the preliminary sample (n_1), and (e) *y*-values in terms of 0's and 1's in the preliminary sample. Let a preliminary sample of size n_1 is drawn from a population of size *N* and n_1^* be the number of units possessing an attribute, say *A*, in the preliminary sample. Then, $p_1 = \frac{n_1^*}{n_1}$, denotes the proportion of units possessing the attribute and $q_1 = 1 - p_1$, denotes the proportion of units not possessing the attribute in the preliminary sample, respectively. Thus, final sample size is determined as,

$$n = \frac{N\left(t^2_{\alpha/2,\, n_1-1} q_1 + \varepsilon^2 p_1\right)}{\left(t^2_{\alpha/2,\, n_1-1} q_1 + \varepsilon^2 p_1 N\right)}.$$

Example 5

A researcher is interested in estimating the proportion of rice eaters in a particular village. A preliminary sample 10 villagers are taken from a total of 500 villagers using SRSWOR. The response is denoted by 1 if the villager is rice eater; otherwise, 0. Determine the sample size needed to estimate the proportion of rice eaters in that village with relative error of 0.20 and 95% confidence level. The preliminary sample data is provided below:

Villagers	1	2	3	4	5	6	7	8	9	10
Response	1	0	1	1	1	0	0	0	1	1

Solution:

Here, $p_1 = \frac{6}{10} = 0.6$, $q_1 = 1 - 0.6 = 0.4$, ε=0.20, α=0.05, *N*=500, n_1=10, and $t_{0.025,\, 9}$=2.262.

The final sample size is,

$$n = \frac{500\left(2.262^2 \times 0.4 + 0.20^2 \times 0.6\right)}{\left(2.262^2 \times 0.4 + 0.20^2 \times 0.6 \times 500\right)},$$

n=73.71 ≈ 74 .

Since, 10 villagers are already selected in the preliminary sample, therefore, 64 additional villagers need to be selected using SRSWOR design from the remaining 490 farmers.

9.5 Sample Size Determination for Estimating the Population Proportion under SRSWR Sampling Design

Case 1: Population proportion is not known

In this case also a preliminary sample is drawn on study variable, *y*, using SRSWR design to estimate the population proportion and then, this estimated value is being used for sample size determination. The following information is required to determine of sample size: (a) Confidence Level (1-α) (in %), (*b*) Relative error (0 - 1), (*c*) Size of the preliminary sample (n_1), and (*d*) *y*-values in terms of 0's and 1's in the preliminary sample. Thus, final sample size is determined as,

$$n=\frac{t^2_{\alpha/2,\, n_1-1} q_1 + \varepsilon^2 p_1}{\varepsilon^2 p_1}.$$

Example 6

A farmer is interested to see whether submerging of seeds will rot them. A preliminary sample of 15 seeds is taken by the farmer using SRSWR. A value 1 is recorded if the seed rot; otherwise, 0. Suppose the farmer wishes to estimate the proportion of seeds that rot with 95% confidence level and with a relative error of 0.20, then how many seeds should he submerge? The preliminary sample data is given below:

Seeds	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Response	1	0	1	1	0	1	1	1	0	0	0	1	1	1	1

Solution:

Here, $p_1 = \frac{10}{15} = 0.67$, $q_1 = 1 - 0.67 = 0.33$, ε =0.20, α =0.05, n_1 =15 and $t_{0.025,\, 14}$ =2.145.

The final sample size is,

$$n=\frac{2.145^2\times 0.33+0.20^2\times 0.67}{0.20^2\times 0.67},$$

$$n=57.65\approx 58.$$

Since, 15 seeds are already selected in the preliminary sample, therefore, 43 additional seeds need to be selected using SRSWR design to estimate the proportion of seeds that rot with 95% confidence level and with a relative error of 0.20.

R codes for determination of sample size

```
install.packages("SscSrs")
library(SscSrs)
# Sample Size Determination for Estimating the Population Mean under SRSWOR when population CV is known
SscSrsMean(TRUE, FALSE, 0.05, 0.2, 100, NA)
# Sample Size Determination for Estimating the Population Mean under SRSWOR when population CV is unknown
preliminary_sample =c(12, 14, 8, 5, 36, 24, 18, 17, 6, 9)
SscSrsMean(FALSE, FALSE, 0.05, 0.2, 100, preliminary_sample)
# Sample Size Determination for Estimating the Population Mean under SRSWR when population CV is known
SscSrsMean(TRUE, TRUE, 0.05, 0.2, NA, NA)
# Sample Size Determination for Estimating the Population Mean under SRSWR when population CV is unknown
preliminary_sample =c(12, 14, 8, 5, 36, 24, 18, 17, 6, 9)
SscSrsMean(FALSE, TRUE, 0.05, 0.2, NA, preliminary_sample)
# Sample Size Determination for Estimating the Population Proportion under SRSWOR
preliminary_sample=c(1,0,1,1,1,0,0,0,1,1)
SscSrsProp(FALSE, 0.05, 0.2, 500, preliminary_sample)
```

```
# Sample Size Determination for Estimating the Population Proportion under SRSWR
preliminary_sample=c(1,0,1,1,0,1,1,1,0,0,0,1,1,1,1)
SscSrsProp(TRUE, 0.05, 0.2, NA, preliminary_sample)
```

Spatial Sampling Strategy

10. Introduction to Spatial Sampling

Spatial sampling is used to estimate attributes of a geographically distributed population, presenting unique challenges related to the selection of sample locations. Two primary approaches are employed: design-based and model-based sampling. Kotz and Johnson (1986) describe spatial sampling as a survey sampling method focused on two-dimensional spaces, such as agricultural fields. In agriculture, parameters like crop yield, soil properties, and climatic variables are inherently spatial, making it advantageous to select samples from diverse geographic locations using spatial sampling techniques. Spatial autocorrelation measures how similar observations at specific coordinate points are to each other. A widely used statistic for assessing spatial autocorrelation is Moran's I (Moran, 1948), which evaluates the relationship between a variable's values at a given location (y) and the average values of that variable at neighboring locations (w_y). The null hypothesis for Moran's I assumes that the variable is randomly distributed across space. Moran (1950) proposed the following measure to calculate spatial autocorrelation (I):

$$\mathrm{I} = \frac{n}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(y_i - \bar{y}\right)\left(y_j - \bar{y}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}$$

where, y_i is the observed value at location i, n is the number of locations, the weighting function w_{ij} is used to assign weights to every pair of locations in the study area, $w_{ij} = 1$; if i and j are neighbour or 0 ; otherwise . The range of Moran's autocorrelation varies from -1 to $+1$, a positive sign indicates positive spatial autocorrelation and zero indicates no spatial autocorrelation. Moran's scatterplot in Figure 3 shows the correlation between observed values of the study variable y at each location and the average value of y at neighbouring locations (w_y). The plot is segmented into four quadrants. The upper-right quadrant (IV) represents cases where both the value of the

study variable y and its local average value are above the overall mean, indicating positive spatial autocorrelation. Conversely, the lower-left quadrant (II) shows cases where both the value of y and its local average are below the overall mean, also indicating positive spatial autocorrelation. The remaining two quadrants (I and III) represent negative spatial autocorrelation, where the value of y and its local average diverge relative to the overall mean.

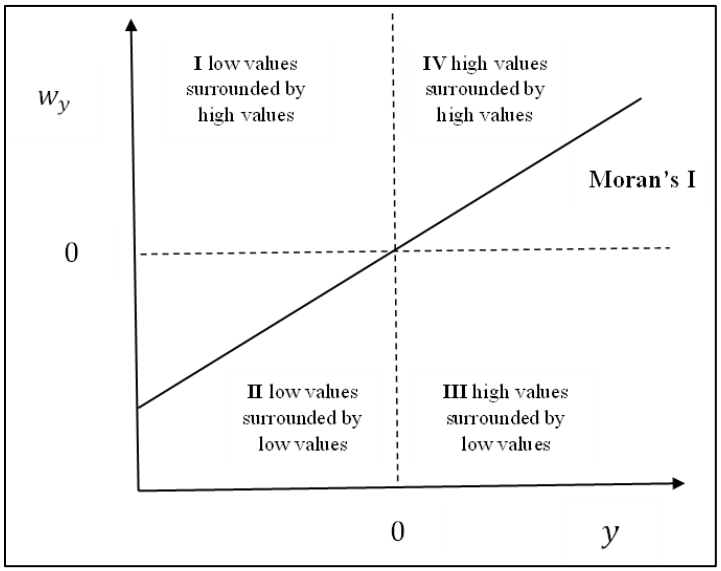


Figure 1. Moran's scatterplot divided into 4 quadrants based on values of the variable y and its local average value w_y at neighbouring locations

10.1 Concept and Motivation

Spatial sampling is typically employed to estimate the characteristics of a spatial population without observing it in its entirety. For example, to estimate the proportion of land covered by a specific land use, it is often more feasible to take a well-distributed sample than to conduct a complete census. Here, the focus is on the current spatial pattern rather than assuming a hypothetical underlying superpopulation. A basic estimator in such cases is the sample mean:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y(i)$$

where $y(i)$ equals 1 if the sample site i exhibits the land use of interest and 0 otherwise. The efficiency of this estimator depends on the sampling design chosen, which must account for spatial autocorrelation- the tendency of nearby locations to have similar values. Spatial autocorrelation

influences the choice of sampling design, as it suggests that sample points should be spaced to avoid redundant information.

11. Design-based and Model-based approach in sampling

In the design-based approach, the variable's value at each location is considered fixed, and the aim is to estimate properties like the mean or total using a probability-based sampling design. Conversely, the model-based approach views the value at each location as a realization of a spatial random process, with the goal of predicting properties of this realized surface or estimating parameters of the process model, such as the spatial covariance structure.

12. Types of Spatial Sampling Designs

Spatial sampling is commonly carried out using techniques such as random sampling, stratified random sampling, and systematic sampling. In addition to these, methods like cluster sampling, nested sampling, and fixed interval point sampling are also utilized. Each of these approaches offers specific advantages and drawbacks, which largely depend on the characteristics of the target variable and the spatial arrangement of the data.

12.1 Random Sampling

Random sampling selects sample points without a specific pattern, which can lead to clustering in certain areas. This design is the least efficient for spatial data because it does not ensure even coverage, potentially resulting in redundant information due to spatial autocorrelation.

12.2 Stratified Random Sampling

In stratified random sampling, the population is divided into mutually exclusive and exhaustive subpopulations, or strata, such as soil mapping units or administrative areas. Within each stratum, a probability sample is selected, often using simple random sampling, resulting in stratified simple random sampling. This design ensures better coverage than random sampling by distributing points across strata. However, sample points may still be adjacent across strata boundaries, potentially reducing efficiency due to spatial autocorrelation.

12.3 Systematic Sampling

Systematic sampling selects points at regular intervals, ensuring a minimum separation distance between points. This design is often the most efficient for spatial data because it minimizes redundancy by spreading points evenly across the study area. According to Dunn and Harrison (1993), systematic sampling outperforms random and stratified random sampling when spatial autocorrelation is strong, as it avoids clustering and ensures comprehensive coverage.

12.4 Cluster Random Sampling

Cluster random sampling entails selecting groups (clusters) of population units, with all units within a chosen cluster being surveyed. This approach is referred to as one-stage cluster sampling. In contrast, two-stage sampling involves selecting a subset of units from within each chosen cluster. A frequently used cluster configuration is a transect, which is particularly practical for fieldwork, especially before the widespread use of GPS technology. The process typically begins by randomly selecting a starting unit, followed by identifying the rest of the cluster based on a predetermined rule—such as an east-west transect with units spaced 100 meters apart (de Gruijter *et al.*, 2006). This procedure is repeated until the required number of clusters is obtained. Cluster sampling is particularly advantageous in extensive study regions where minimizing travel time is essential; however, it can lead to reduced precision if the units within clusters are highly similar.

12.5 Conditioned Latin Hypercube Sampling

The **Conditioned Latin Hypercube Sampling (cLHS)** approach is widely recognized for its ability to generate spatially distributed samples that capture the full variability of environmental covariates. By ensuring that every combination of input variables is adequately represented, cLHS produces statistically robust samples and minimizes sampling bias. Previous studies, such as that of Minasny and McBratney (2006), have shown that cLHS can closely replicate the original distribution of environmental variables even with relatively small sample sizes, making it an efficient and cost-effective strategy for environmental monitoring and modeling. To evaluate the representativeness of different sample sizes, **box plots** and **density plots** are often employed. Box plots visually compare the central tendency and spread (interquartile range) of sampled values against the full dataset for each covariate. Minimal differences in the medians and interquartile ranges between the samples and the original data indicate that the samples effectively represent the underlying distribution. As the sample size increases, the box plots tend to cluster more tightly around the median, reflecting reduced variability. However, beyond a certain threshold, typically between 100 and 150 samples, additional increases in sample size yield diminishing returns in terms of representational accuracy. Similarly, **density plots** provide insights into how well the sampled data reproduces the shape of the original distribution for each covariate. When the density curves of the sampled subsets align closely

with those of the full dataset, it suggests that the samples have captured the overall distribution effectively. Overall, these visual and statistical assessments suggest that a well-chosen sample size within this range can strike an ideal balance between representativeness and sampling efficiency. This reinforces the utility of cLHS as a powerful sampling design tool for spatial studies, particularly in applications where field data collection is expensive or logistically challenging.

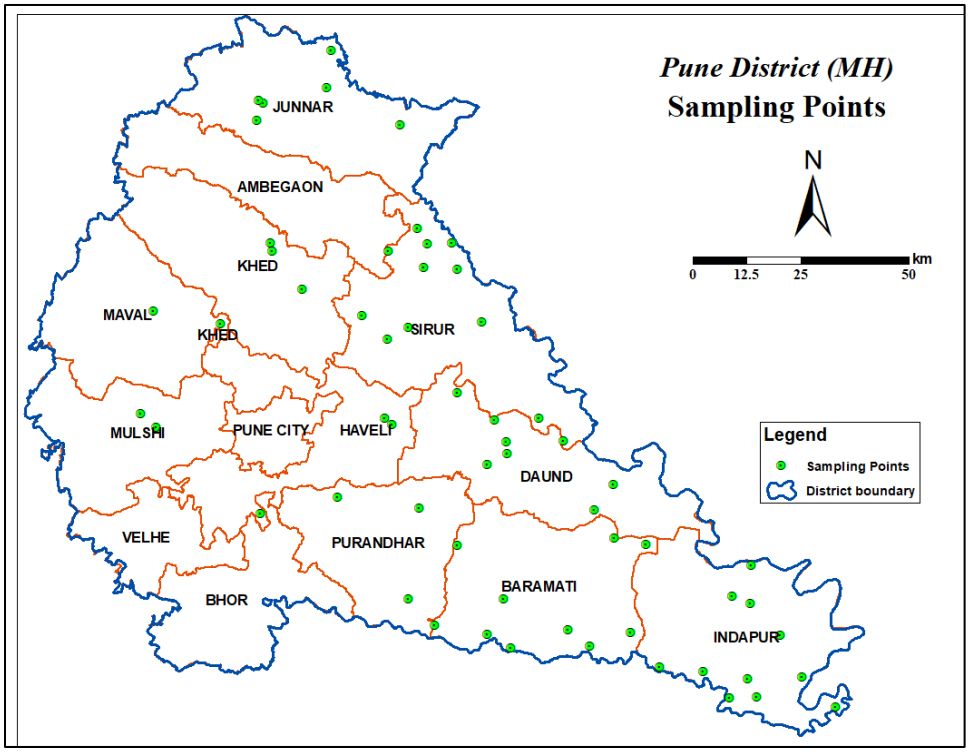


Figure 2. Determination of optimal sampling points of the Pune District using cLHS sampling strategy

13. Spatial Autocorrelation and Sampling Efficiency

Spatial autocorrelation significantly affects sampling design efficiency. When autocorrelation is strong, nearby locations provide similar information, making it critical to space sample points to avoid redundancy. Systematic sampling and stratified random sampling with geographical strata outperform random sampling by ensuring better coverage. In large study areas, cluster random sampling may be preferred to reduce travel time, despite potential precision trade-offs due to similarity within clusters.

14. Advantages of Spatial Sampling

- **Representativeness**: Captures spatial variability, reducing bias.
- **Efficiency**: Optimizes resources by targeting key areas.
- **Accuracy**: Enhances interpolation and prediction quality.

- **Flexibility**: Adapts to specific objectives.

15. Applications in Agriculture

Soil sampling and fertility mapping

Land use/land cover classification

Monitoring abiotic stresses (e.g., salinity, drought)

Yield estimation and model calibration

Precision farming and variable input management

16. Conclusion

Spatial sampling is essential for studying geographically distributed populations, requiring careful consideration of spatial dependence. By choosing an appropriate design-random, stratified random, systematic, cluster random, or random grid-researchers can optimize the accuracy and efficiency of their estimates. A critical consideration in spatial sampling is the presence of spatial autocorrelation, which is inherent in most spatial data. Positive spatial autocorrelation means that observations located close to one another are more likely to have similar values. This affects the information content of the sample, and therefore the efficiency and precision of the resulting estimates. The choice depends on the strength of spatial autocorrelation, study area size, and practical constraints. Understanding these factors ensures that spatial sampling effectively captures population characteristics while minimizing redundancy and maximizing precision.

References

Brus, D. J. (2022). Spatial sampling with R. Chapman and Hall/CRC.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd Edition. New York: John Wiley & Sons, Inc.

Delmelle, E. (2009). Spatial sampling. The SAGE handbook of spatial analysis, 183-206.

Dumelle, M., Kincaid, T., Olsen, A. R., & Weber, M. (2023). spsurvey: spatial sampling design and analysis in R. *Journal of Statistical Software*, *105*, 1-29.

Gruijter, J. J., Bierkens, M. F., Brus, D. J., & Knotters, M. (2006). Sampling for natural resource monitoring (pp. xiii+-332). Springer-Verlag Berlin Heidelberg.

Mukhopadhyay, P. (2011). Theory and Methods of Survey Sampling (2nd ed.). PHI Learning Pvt. Ltd.

Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & geosciences, 32(9), 1378-1388.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rao, J. N. K. (2000). Essentials of survey sampling. New Delhi: Narosa Publishing House.

Singh, D. and Chaudhary, F.S. (1986). *Theory and Analysis of Sample Survey Designs*. New York: John Wiley & Sons, Inc.

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory of Surveys with Applications*. Iowa State University Press. (USA).

Wang, J. F., Stein, A., Gao, B. B., & Ge, Y. (2012). A review of spatial sampling. Spatial Statistics, 2, 1-14.

Yang, L., Zhu, A. X., Qi, F., Qin, C. Z., Li, B., & Pei, T. (2013). An integrative hierarchical stepwise sampling strategy for spatial sampling and its application in digital soil mapping. International Journal of Geographical Information Science, 27(1), 1-23.

Introduction to Small Area Estimation and Its Applications in Food Security, Poverty & Inequality Estimation: A Practical Approach Using R

Saurav Guha
Department of Statistics, Mathematics & Computer Application
Bihar Agricultural University, Sabour, Bhagalpur 813210
Email: saurav@bausabour.ac.in; saurav.iasri@gmail.com

1. Introduction

In light of the growing socio-economic and demographic challenges, researchers, policymakers, and experts have given a great deal of emphasis on the development of reliable indicators and obtaining of high-quality data on living conditions at not only the national and state levels but also at the district and further down the geographic levels. As a result, the focus of agricultural and socio-economic planning has moved from the macro to the micro level, and the small area estimation approach is currently of great interest to the researchers and planners in the country. Existing data from large-scale social-economic and agricultural surveys provide representative estimates at the national and state levels, but they cannot be readily used to generate reliable disaggregate, micro or local level statistics. Eventually, the absence of reliable, representative and robust statistics at the local (and disaggregate) level often put restrictions for designing target-oriented interventions and policy development related to reducing the inequality and disparity among the deprived. The disaggregate level statistics is inevitably needed in India because Government is committed to "leave no one behind" agenda for sustainable development goals. For example, such disaggregate level statistics is unavoidable for planning, implementation and evaluation of several programmes and schemes of Government of India such as Doubling Farmers Income, Pradhan Mantra Fasal Bima Yojan and many more. The need for micro or local level statistics (also known as small domains, small areas, such as district, blocks or district classified by age-sex-race or several other groups) where programs are designed and executed is unavoidable for efficient policy formulation and planning, resources allocation, monitoring and evaluation. Censuses provide such local level statistics but Censuses are less regular and information are very limited whereas national level large scale surveys are frequent and collect a diverse range of data. However, samples at the disaggregate level are not big enough to produce "direct sample estimates

(estimates that use only the data on the target variable from the domain of study and time period of interest)" with acceptable precision.

In general, sample surveys are structured to yield accurate estimates for broader geographical areas such as the national or state levels. The sample sizes are typically determined to ensure that direct estimates from surveys are reliable for these larger domains. However, in many real-world scenarios, there is a growing need to estimate parameters for smaller geographic domains where the number of sampled units is insufficient. These smaller domains are commonly referred to as "small areas," where the limited sample sizes result in imprecise or unstable direct estimates. For example, surveys conducted by the National Statistical Office (NSO) in India are well-designed to generate precise estimates at the national and state levels. However, they fall short at more granular levels like districts or blocks due to higher sampling variability and smaller sample sizes. Expanding these surveys to cover smaller areas directly would require significant time and financial investment. Consequently, estimates derived directly from such sparse data can be highly unreliable at the district or sub-district levels. To overcome this limitation, statisticians rely on **model-based estimation techniques**-also known as **indirect methods**—which use statistical models and auxiliary data (such as census or administrative records) to improve precision in small area estimates. This domain of research is known as **Small Area Estimation (SAE)**. SAE provides a framework to "borrow strength" from related areas by linking data through statistical models, thereby producing more reliable estimates even when direct data is limited (Rao and Molina, 2015).

SAE methods are generally categorized into two types depending on the availability of auxiliary information:

1. **Area-level models**, such as the well-known **Fay-Herriot (FH) model** (Fay & Herriot, 1979), are used when auxiliary data is only available in aggregated form at the area level.

2. **Unit-level models**, such as the **nested error regression model** by Battese et al. (1988), are used when auxiliary variables are available for individual units.

The FH model is particularly popular because it effectively accounts for complex survey designs while utilizing aggregate-level contextual data. This model is based on a **linear mixed model (LMM)** framework where areas are treated as random effects (Guha and Chandra, 2022). Several researchers have further extended the FH model. For instance, **Prasad and Rao (1990)** worked on estimating the mean squared error for FH estimates, with subsequent advancements by **Datta and Lahiri (2000)**, **Gonzalez-Manteiga et al. (2010)**, and **Datta et al. (2011)**. Spatial and temporal extensions of the FH model have also been explored. For example, **Marhuenda et al. (2013)** proposed a spatiotemporal FH model, while **Morales et al. (2015)** extended it to multivariate contexts. The **Multivariate Fay-Herriot (MFH) model**, initially proposed by **Fay (1987)** and later expanded by **Datta et al. (1991)** and others (e.g., Benavent & Morales, 2016; Guha & Chandra, 2024), allows for modeling multiple correlated variables by incorporating flexible covariance structures between them. Bayesian approaches have also found increasing application in SAE. For instance, **Arima et al. (2017)** introduced Bayesian estimators to address functional measurement errors. **Esteban et al. (2020)** focused on estimating proportions under an area-level LMM framework, while **Benavent and Morales (2021)** developed a temporal bivariate LMM. **Burgard et al. (2021, 2022)** proposed models for handling missing data and provided optimal predictors under bivariate FH frameworks. Other advancements include methods for confidence region estimation and parametric bootstrapping under the MFH model (Ito & Kubokawa, 2021; Saegusa et al., 2022). Incorporating **spatial correlations** into SAE has also gained traction. The **Simultaneous Autoregressive (SAR) model** is commonly used in the frequentist framework to model spatial dependency in area-level random effects (Anselin, 1992; Cressie, 1993). This has led to multiple attempts to generalize the FH model by incorporating SAR-based spatial effects (e.g., Molina et al., 2009; Pratesi & Salvati, 2008, 2009; Singh et al., 2005). In contrast, Bayesian SAE models often utilize **conditionally autoregressive (CAR)** structures (Besag et al., 1991; Leroux et al., 1999). In this study, we apply the **MFH model** to jointly estimate three food insecurity-related indicators at the **district level** for **rural areas of Uttar Pradesh**, India. Uttar Pradesh, the most populous state in India-and globally among subnational units-accounts for approximately **16.16% of India's population** and **2.9% of the world's population**, spread across an area of **243,290 sq. km**, which constitutes about **6.88%** of India's total area.

Despite being one of the country's largest state economies, Uttar Pradesh faces significant poverty. As per **NITI Aayog (2019)**, nearly **29.43%** of the state's population lives below the poverty line—substantially higher than the **national average of 21.92%**. According to the **Global Hunger Index (2020)**, India ranked 94th out of 132 countries with a score of **27.2**, while Uttar Pradesh ranked **24th out of 28 states**, scoring **31**, which is below the national average. Given that a majority of the state's population is engaged in agriculture and nearly **78% live in rural areas** (Census 2011), rural Uttar Pradesh is an appropriate case study for applying SAE to estimate food insecurity indicators. This study uses SAE techniques to provide reliable estimates at the **district level**, which are otherwise not feasible through direct survey methods due to small sample sizes.

2. Data

This section outlines the primary datasets used in the multivariate small area estimation (SAE) analysis. The study relies on two major sources: the **Household Consumer Expenditure Survey (HCES) 2011-12** conducted by the **National Sample Survey Office (NSSO)**, and the **Population Census 2011**. These data sources are employed to generate district-level estimates of food insecurity indicators for rural areas in **Uttar Pradesh (UP)**. The **2011-12 HCES** is the most recent round of consumer expenditure data widely referenced for policy-making in India. While this dataset is not openly accessible online, it can be requested from the **Ministry of Statistics and Programme Implementation (MoSPI)**, Government of India (http://mospi.nic.in). The NSSO conducts HCES at regular intervals under its ongoing survey "rounds," typically covering one year, using a scientifically designed sampling method. A representative set of households is randomly selected and surveyed through personal interviews. Over time, the entire geographical landscape of India is included through such rotational surveys.

For the 2011-12 HCES, a **stratified multi-stage random sampling** approach was used. Here, districts served as **strata**, villages as **first-stage units**, and households as **second-stage units**. While the survey is designed to produce reliable statistics at the **state and national levels**, it falls short when applied to smaller domains like districts, due to the limited number of sampled households per district. Consequently, direct district-level estimates from HCES suffer from large standard errors and lack sufficient precision. This is particularly concerning given that district-level statistics are crucial for localized development planning. However, India currently lacks

nationally representative surveys that generate reliable district-level data, creating a gap in the evidence base for district-specific policy interventions. In Uttar Pradesh, the 2011-12 HCES covered **5915 rural households across 71 districts**. Sample sizes varied significantly across districts, ranging between 32 and 128 households, with an **average sample size of 83 households**. The average sampling fraction was notably low, around **0.00023**. This highlights the challenge of obtaining reliable direct estimates from such small samples (see Chandra et al., 2011; Rao and Molina, 2015). To address these limitations, this paper applies a **multivariate small area estimation technique** that leverages auxiliary variables drawn from the **2011 Population Census**. By combining the strengths of survey and census data, the SAE method helps enhance the precision of district-level estimates. To quantify food insecurity, food consumption reported by households in HCES is converted into energy intake using a **nutritional conversion chart** primarily based on the **Indian Council of Medical Research (ICMR)** standards (Gopalan et al., 1991). The conversion expresses food quantities in terms of kilocalories (Kcal). However, it is important to acknowledge that actual nutrient absorption may vary based on cooking methods and food processing practices used by households (Government of India, 2014). Despite this limitation, calorie intake serves as a commonly accepted metric for assessing food security. One inherent constraint of the HCES data is that food consumption is recorded at the **household level**, not at the individual level. As a result, the data cannot capture **intra-household disparities** in food intake. For the purpose of this study, all estimates are computed as **per capita averages** at the household level. The analysis focuses on three key indicators of food insecurity, defined at the household level and derived from the 2011–12 HCES data, following the framework of **Foster et al. (1984)** and later **Hossain et al. (2020)**: **Y1: Food Insecurity Prevalence (FP), Y2: Food Insecurity Gap (FG) and Y3: Food Insecurity Severity (FS).** In rural India, the **average daily dietary energy requirement per person** is set at **2400 Kcal**, as specified by the **Ministry of Health and Family Welfare**. In this paper, we aim to produce **district-level estimates** of food insecurity in rural Uttar Pradesh by **jointly modeling the three indicators-FP, FG, and FS-using a multivariate SAE framework**.

3. Practical Applications in HCES data using R

Data

This is an R Markdown document which includes some essential R-codes for statistical analysis for Small Area Estimation of food insecurity.We use the dataset "HCES" throughout the article. This dataset is imported by by:

```
mydata <- read.csv("C:\\Users\\saura\\Downloads\\NIASM 31 July\\Direct-Estimate-2023-24-Bihar-FI.csv"
# load the dataset and renamed it mydata
```

Below a preview of this dataset and its structure:

```
head(mydata) # first 6 observations
```

```
##   X Area         R         U       var.R       var.U n.R n.U     N.R    N.U
## 1 1    1 0.6037925 0.5429720 0.000773405 0.002243862 360 126 2725455 285349
## 2 2   10 0.5736163 0.4297064 0.000601543 0.002262363 450 144 3348325 341872
## 3 3   11 0.5255657 0.5463706 0.000882834 0.006058825 324  54 2354178 236082
## 4 4   12 0.5473594 0.3449485 0.000841823 0.004236411 324  72 2561289 167965
## 5 5   13 0.5892157 0.6187967 0.000522836 0.002224136 504 126 3780678 519857
## 6 6   14 0.5712604 0.4065157 0.000539513 0.003044739 522 144 3876827 502295
```

```
str(mydata) # structure of dataset
```

```
## 'data.frame':    38 obs. of  10 variables:
##  $ X    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Area : int  1 10 11 12 13 14 15 16 17 18 ...
##  $ R    : num  0.604 0.574 0.526 0.547 0.589 ...
##  $ U    : num  0.543 0.43 0.546 0.345 0.619 ...
##  $ var.R: num  0.000773 0.000602 0.000883 0.000842 0.000523 ...
##  $ var.U: num  0.00224 0.00226 0.00606 0.00424 0.00222 ...
##  $ n.R  : int  360 450 324 324 504 522 396 450 540 576 ...
##  $ n.U  : int  126 144 54 72 126 144 18 108 36 72 ...
##  $ N.R  : int  2725455 3348325 2354178 2561289 3780678 3876827 2990943 3013328 3675530 4339930 ..
##  $ N.U  : int  285349 341872 236082 167965 519857 502295 43429 336280 47217 222997 ...
```

Summary

You can compute the minimum, first quartile, median, mean, third quartile and the maximum for all numeric variables of a dataset at once using summary():

Correlation

The correlation measures the linear relationship between two variables, and it can be computed with the cor() function:

```
cor(mydata$R, mydata$U)
```

```
## [1] NA
```

```
# computing correlation matrix
cor_data <- cor(mydata[, 1:4])

print("Correlation matrix")
```

```
## [1] "Correlation matrix"
```

```
print(cor_data)
```

```
##              X      Area         R  U
## X    1.0000000 0.2340519 0.2083191 NA
## Area 0.2340519 1.0000000 0.6619989 NA
## R    0.2083191 0.6619989 1.0000000 NA
## U           NA        NA        NA  1
```

```
#Visualize a Correlation Matrix in R
#install.packages("corrplot")

# Correlogram in R
# required packages
library(corrplot)
```

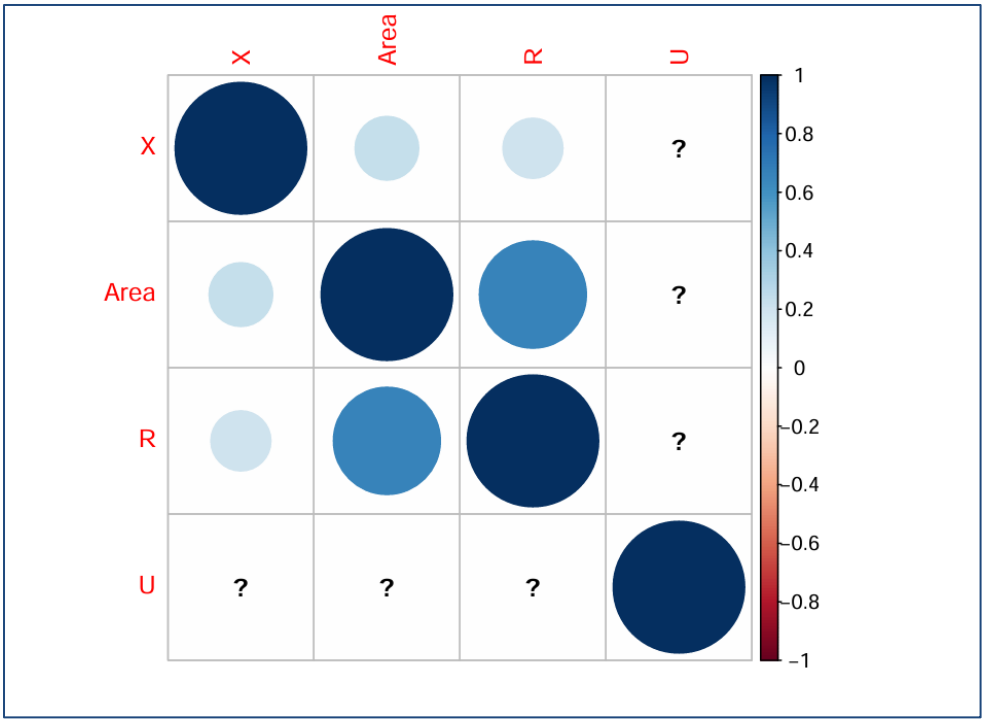
```
## Warning: package 'corrplot' was built under R version 4.4.3
```

```
## corrplot 0.95 loaded
```

```
# correlation matrix
M<-cor(mydata[, 1:4])
head(round(M,2))
```

```
##         X Area    R  U
## X    1.00 0.23 0.21 NA
## Area 0.23 1.00 0.66 NA
## R    0.21 0.66 1.00 NA
## U      NA   NA   NA  1
```

```
corrplot(M, method<-"circle") # visualizing correlogram as circle
```



variable selction and PCA for rural area

```
data.rural<-read.csv("C:\\Users\\saura\\Downloads\\NIASM 31 July\\Biharruralfarmerdata2.csv", header
x11<-cbind(data.rural$M_LIT,data.rural$F_LIT,data.rural$TOT_WORK_M,
          data.rural$TOT_WORK_F)
x12<-cbind(data.rural$MAINWORK_M,data.rural$MAINWORK_F,data.rural$MAIN_CL_M,
          data.rural$MAIN_CL_F,data.rural$MAIN_AL_M,data.rural$MAIN_AL_F)
x13<-cbind(data.rural$MARG_CL_M,data.rural$MARG_CL_F,data.rural$MARG_AL_M,
          data.rural$MARG_AL_F)
## PCA
pc11<-prcomp(x11)
pc12<-prcomp(x12)
pc13<-prcomp(x13)
summary(pc13)
```

```
## Importance of components:
##                             PC1       PC2       PC3       PC4
## Standard deviation     79225.032 2.823e+04 1.057e+04 2.191e+03
## Proportion of Variance     0.873 1.108e-01 1.553e-02 6.700e-04
## Cumulative Proportion      0.873 9.838e-01 9.993e-01 1.000e+00
```

```
## Stepwise regression
lm1 <- lm(mydata$R ~ data.rural$P_SC+data.rural$P_LIT+data.rural$P_ILL
```

```
          +data.rural$TOT_WORK_P+data.rural$MAINWORK_P+data.rural$MAIN_CL_P
          +data.rural$MAIN_AL_P+data.rural$MARGWORK_P+data.rural$MARG_CL_P
          +data.rural$MARG_AL_P+ pc11$x[,1]+pc12$x[,1]+pc12$x[,2]+pc13$x[,1]
          +pc13$x[,2])
slm1 <- step(lm1)
```

variable selction and PCA for urban area

```
data.urban<-read.csv("C:\\Users\\saura\\Downloads\\NIASM 31 July\\Biharurbanfarmerdata2.csv",header =
data.urban<-data.urban[1:34,]
x21<-cbind(data.urban$M_LIT,data.urban$F_LIT,data.urban$TOT_WORK_M,
         data.urban$TOT_WORK_F)
x22<-cbind(data.urban$MAINWORK_M,data.urban$MAINWORK_F,data.urban$MAIN_CL_M,
         data.urban$MAIN_CL_F,data.urban$MAIN_AL_M,data.urban$MAIN_AL_F)
x23<-cbind(data.urban$MARG_CL_M,data.urban$MARG_CL_F,data.urban$MARG_AL_M,
         data.urban$MARG_AL_F)
## PCA
pc21<-prcomp(x21)
pc22<-prcomp(x22)
pc23<-prcomp(x23)
summary(pc23)
```

```
## Importance of components:
##                             PC1       PC2       PC3       PC4
## Standard deviation     3562.0718 533.35287 321.00566 122.81126
## Proportion of Variance    0.9692   0.02173   0.00787   0.00115
## Cumulative Proportion     0.9692   0.99098   0.99885   1.00000
```

```
## Stepwise regression
lm2 <- lm(mydata$U[1:34] ~ data.urban$P_SC+data.urban$P_LIT+data.urban$P_ILL+
            data.urban$TOT_WORK_P+data.urban$MAINWORK_P+data.urban$MAIN_CL_P+
            data.urban$MAIN_AL_P+data.urban$MARGWORK_P+data.urban$MARG_CL_P+
            data.urban$MARG_AL_P+pc21$x[,1]+pc22$x[,1]+pc23$x[,1])
slm2 <- step(lm2)
```

Small Area Estimation

First install the required package

```
library(sae)
```

```
## Loading required package: MASS
```

```
## Loading required package: lme4
```

```
## Loading required package: Matrix
```

```
FH.R<-mseFH(mydata$R~data.rural$P_SC+data.rural$TOT_WORK_P+data.rural$MAINWORK_P+
             data.rural$MARG_CL_P,mydata$var.R)
FH.U<-mseFH(mydata[1:34,]$U~data.urban$P_ILL+data.urban$MARG_CL_P,
           mydata[1:34,]$var.U)
Result.fh.R<-matrix(c(FH.R$est$eblup,FH.R$mse),38,2)
Result.fh.U<-matrix(c(FH.U$est$eblup,FH.U$mse),34,2)

## Model fitting summary
FH.R$est$fit
```

```
## $method
## [1] "REML"
##
## $convergence
## [1] TRUE
##
## $iterations
## [1] 4
##
## $estcoef
##                                 beta    std.error     tvalue       pvalue
## X(Intercept)            7.095673e-01 4.362521e-02 16.2650735 1.746574e-59
## Xdata.rural$P_SC       -5.870352e-08 1.289636e-07 -0.4551945 6.489693e-01
## Xdata.rural$TOT_WORK_P  3.119063e-07 3.076919e-07  1.0136969 3.107274e-01
## Xdata.rural$MAINWORK_P -4.597903e-07 4.047716e-07 -1.1359253 2.559878e-01
## Xdata.rural$MARG_CL_P  -1.361655e-06 1.512607e-06 -0.9002042 3.680116e-01
##
## $refvar
## [1] 0.01120336
##
## $goodness
##   loglike       AIC       BIC       KIC
##  32.73670 -53.47340 -43.64789 -47.47340
```

```
FH.U$est$fit
```

```
## $method
## [1] "REML"
##
```

```
## Saving the output

write.csv(Result.fh.R,"FH.R_output2.csv",sep = TRUE,col.names = TRUE)
```

```
## Warning in write.csv(Result.fh.R, "FH.R_output2.csv", sep = TRUE, col.names =
## TRUE): attempt to set 'col.names' ignored

## Warning in write.csv(Result.fh.R, "FH.R_output2.csv", sep = TRUE, col.names =
## TRUE): attempt to set 'sep' ignored
```

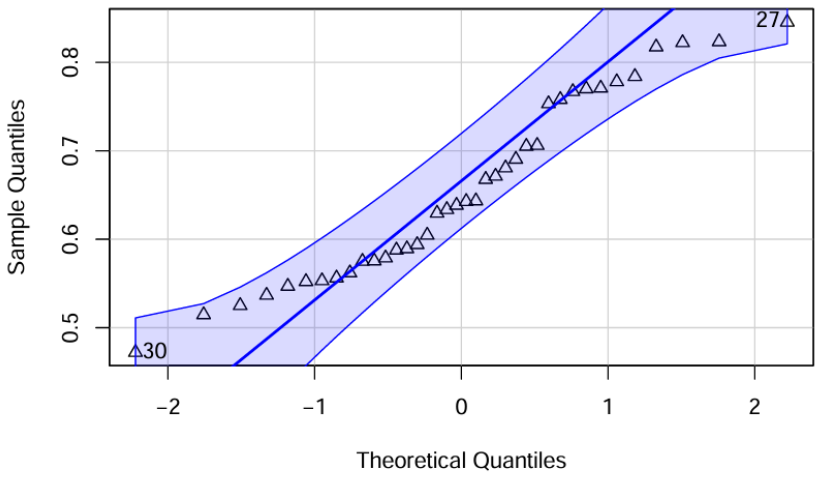
```
write.csv(Result.fh.U,"FH.U_output2.csv",sep = TRUE,col.names = TRUE)
```

```
## Warning in write.csv(Result.fh.U, "FH.U_output2.csv", sep = TRUE, col.names =
## TRUE): attempt to set 'col.names' ignored

## Warning in write.csv(Result.fh.U, "FH.U_output2.csv", sep = TRUE, col.names =
## TRUE): attempt to set 'sep' ignored
```

Validation and Diagnostics

```
Result=read.csv("C:\\Users\\saura\\Downloads\\NIASM 31 July\\final result 08.07.2025.csv",header = TRUE

## Normality
library(car)
```

```
qqPlot(Result$Estimate.rural,ylab = "Sample Quantiles",xlab =
         "Theoretical Quantiles",pch = 2)
```



```r
## Normality test
shapiro.test(Result$Estimate.rural)
```

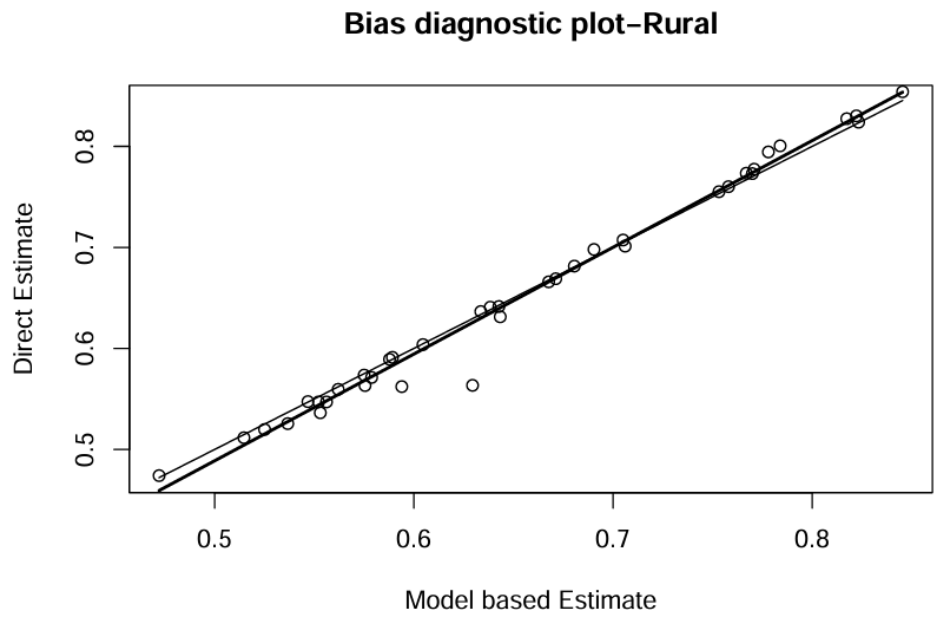
```
##
##  Shapiro-Wilk normality test
##
## data:  Result$Estimate.rural
## W = 0.9482, p-value = 0.07769
```

```r
shapiro.test(Result$Estimate.urban[1:34])
```

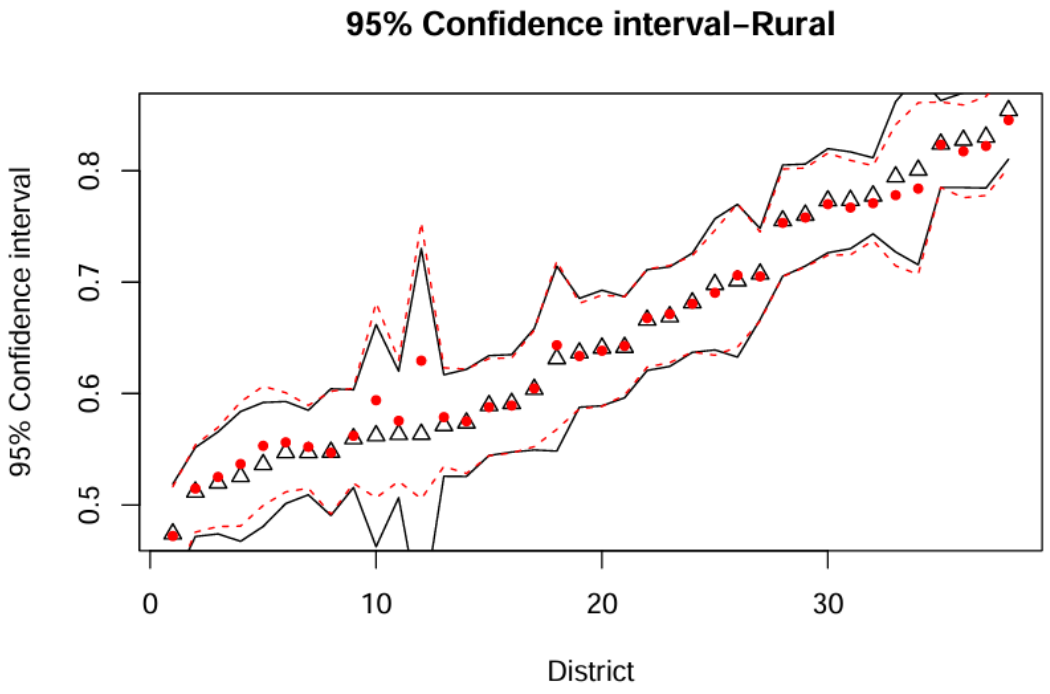
```
##
##  Shapiro-Wilk normality test
##
## data:  Result$Estimate.urban[1:34]
## W = 0.98208, p-value = 0.8357
```

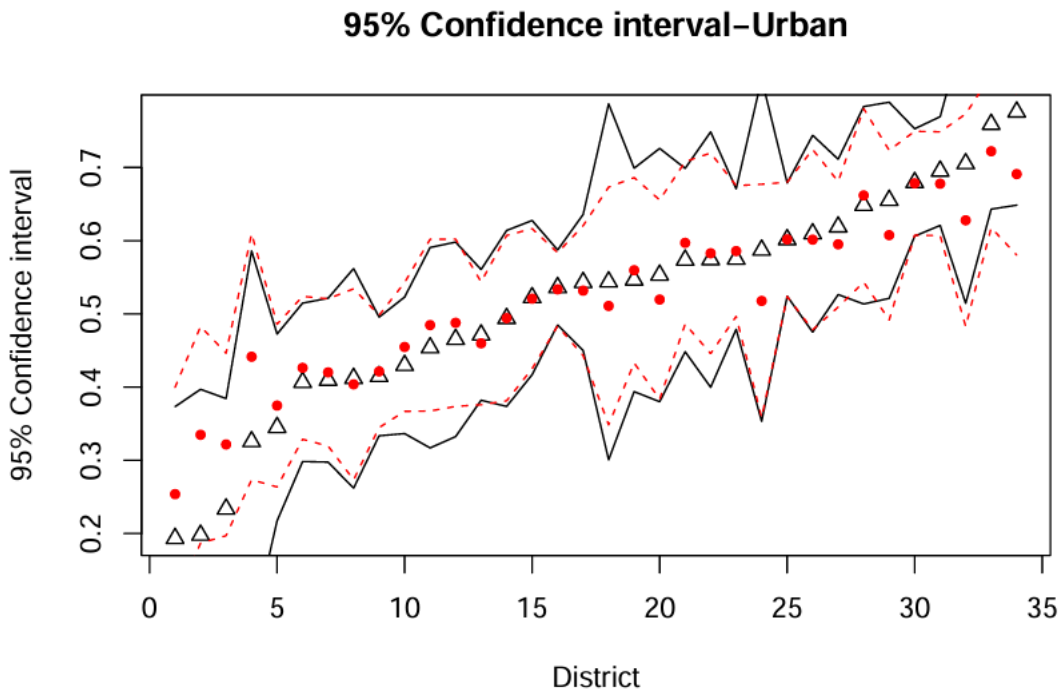
```r
## Bias diagnostic plot
A=lm(Result$direct.R~Result$Estimate.rural,Result)
summary(A)
```

```r
plot(Result$Estimate.rural,Result$Estimate.rural,
     main="Bias diagnostic plot-Rural",
     ylab="Direct Estimate",xlab="Model based Estimate",col="black",
     type = "l",lty = 1)
lines(Result$Estimate.rural,A$fitted.values,col="black",type = "l",
      lty = 1, lwd = 2)
lines(Result$Estimate.rural,Result$direct.R,
      col="black",pch = 1,type = "p")
```

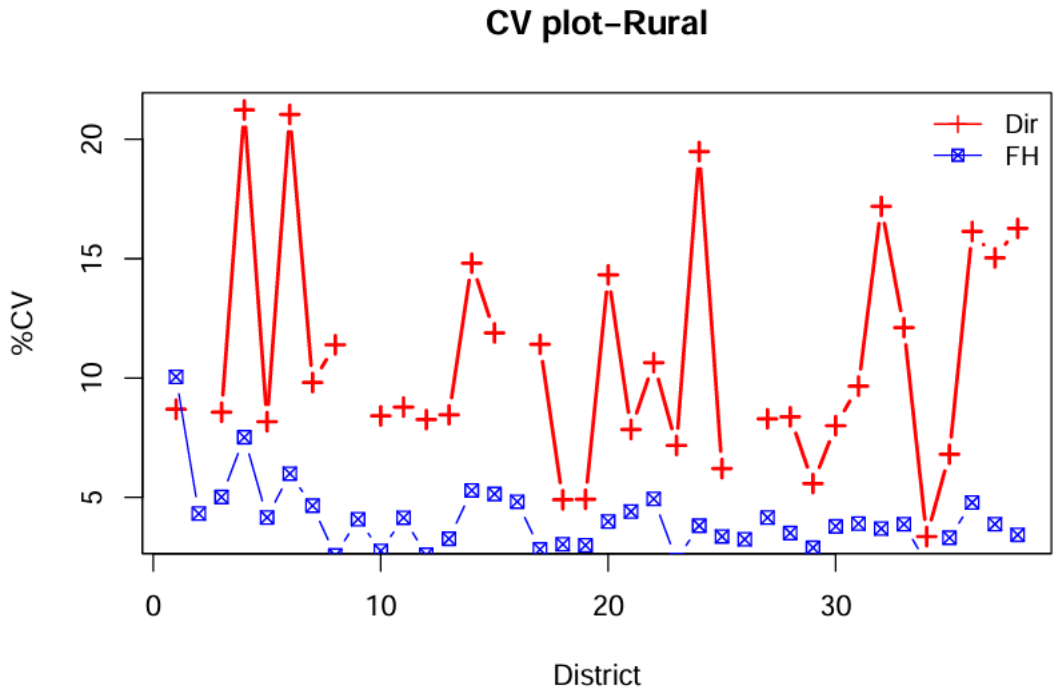


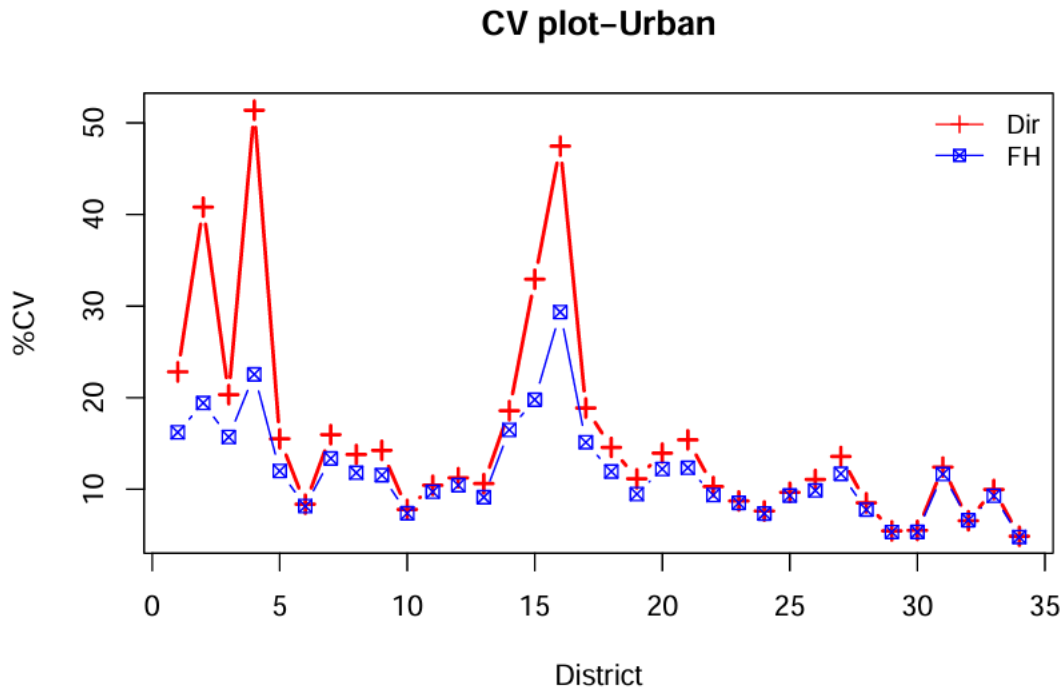

```
A=lm(Result$direct.U[1:34]~Result$Estimate.urban[1:34],Result)
summary(A)
```






```
##CV plot

data_rmse <- Result[1:38,] %>% arrange((n.R))
D=38
District=1:D
par(mfrow=c(1,1))
plot(District,sqrt(data_rmse$dir.var.U)*100/data_rmse$direct.R,
     main="CV plot-Rural",ylab="%CV",col="red",pch = 3,type = "b",lwd=2)
lines(District,sqrt(data_rmse$Var.rural)*100/data_rmse$Estimate.rural,
      col="blue",pch = 7,type = "b",lwd=1)
legend("topright", legend = c("Dir","FH"), col = c("red","blue"), pch = c(3,7),
       lty = 1, cex = 0.9,  bty = 'n', box.lty =1, box.lwd =0.2, merge = TRUE,
       trace = TRUE)
```

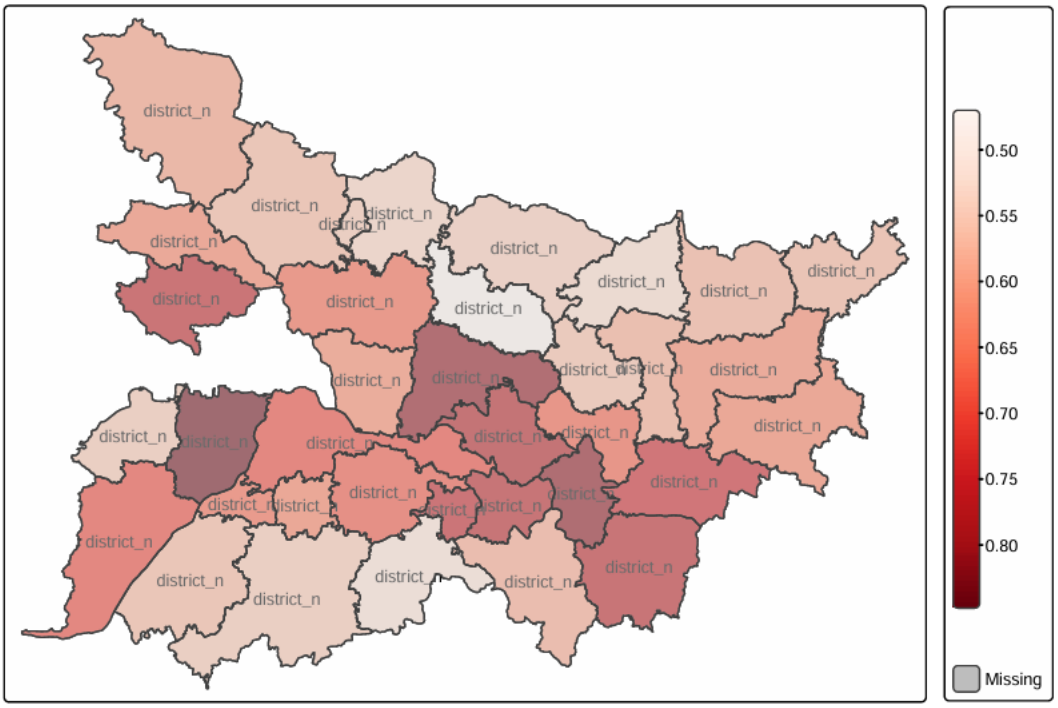

CV plot–Rural
Dir
FH
%CV
5
10
15
20
0
10
20
30
District


CV plot–Urban
Dir
FH
%CV
10
20
30
40
50
0
5
10
15
20
25
30
35
District

Spatial mapping of food insecurity

```
library(sf)
```

```
Bihar=read_sf(dsn ="C:\\Users\\saura\\Downloads\\NIASM 31 July\\Census_2011\\2011_Dist.shp")
library(tidyverse)

Bihar = Bihar %>% filter(ST_NM == "Bihar")

real.data <- read.csv("C:\\Users\\saura\\Downloads\\NIASM 31 July\\bivariatemap.csv")

OA.Census <- merge(Bihar, real.data, by.x="DISTRICT", by.y="District")

A=tm_shape(OA.Census) + tm_polygons(c("FI_rural"),
                                    style = "cont",n=6,
                                    palette = "Reds",
                                    title = c(" "))+
  tm_text("district_n",size = 0.55,col ="black")+
  tm_borders("black",alpha=.5)+
  tm_layout(legend.outside=TRUE)
```



4. Concluding remarks

In recent years, **Small Area Estimation (SAE)** has gained significant traction, with strong theoretical foundations. However, its practical applications, especially in agriculture and social sciences, remain limited. In India, **Census data** is sparse and infrequent, while **NSSO surveys** offer regular socioeconomic data but are only reliable for national and state-level estimates due to small sample sizes at lower administrative levels. This study applies a **Multivariate Fay-Herriot (MFH) model** to estimate three food insecurity indicators-**Prevalence (FP)**, **Gap (FG)**, and **Severity (FS)**-at the **district level** in **rural Uttar Pradesh** using the **2011–12 NSSO HCES** and **2011 Census** data. Given the limited district-level samples in the HCES (average sample size: 83 households), direct estimates are unreliable. To address this, we integrate auxiliary variables from the Census through **Principal Component Analysis (PCA)** for dimension reduction and covariate selection. By jointly modeling the correlated indicators using **multivariate SAE**, the approach improves estimate precision over univariate methods, as shown through reductions in **Mean Squared Error (MSE)** and **Coefficient of Variation (CV)**. Spatial maps highlight district-level disparities in food insecurity, identifying vulnerable districts such as **Gonda, Basti, and Gorakhpur**, and more secure areas like **Lucknow and Jhansi**. This analysis underscores the advantages of SAE in producing **reliable, cost-effective local estimates** with confidence intervals, supporting **evidence-based planning** at the district level. It also aligns with **SDG Indicator 2.1.2** on food insecurity severity and can serve as a blueprint for applying SAE to other health indicators like stunting or undernutrition. The findings have direct relevance for policymakers at national and international levels, aiding targeted interventions for food security and rural development.

References

Anselin L. (1992). *Spatial econometrics methods and models*. Kluwer Academic. https://doi.org/10.1007/978-94-015-7799-1

Arima S., Bell W., Datta G., Franco C., & Liseo B. (2017). Multivariate Fay–Herriot Bayesian estimation of small area means under functional measurement error. *Journal of the Royal statistical Society, Series A*, 180(4), 1191–1209. https://doi.org/10.1111/rssa.12321

Benavent R., & Morales D. (2016). Multivariate Fay–Herriot models for small area estimation. *Computational Statistics and Data Analysis*, 94, 372–390. https://doi.org/10.1016/j.csda.2015.07.013

Benavent R., & Morales D. (2021). Small area estimation under a temporal bivariate area-level linear mixed model with independent time effects. *Statistical Methods and Applications*, 30(1), 195–222. https://doi.org/10.1007/s10260-020-00521-x

Besag J., York J., & Mollié A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1), 1–20. https://doi.org/10.1007/BF00116466

Burgard J., Esteban M., Morales D., & Pérez A. (2021). Small area estimation under a measurement error bivariate Fay–Herriot model. *Statistical Methods and Applications*, 30(1), 79–108. https://doi.org/10.1007/s10260-020-00515-9

Burgard J., Morales D., & Wölwer A. (2022). Small area estimation of socioeconomic indicators for sampled and unsampled domains. *ASTA Advances in Statistical Analysis*, 106(2), 287–314. https://doi.org/10.1007/s10182-021-00426-4

Census (2011). *Primary census abstracts*, Registrar General of India, Ministry of Home Affairs, Government of India. https://censusindia.gov.in/.

Chandra H., Salvati N., Chambers R., & Tzavidis N. (2012). Small area estimation under spatial nonstationarity. *Computational Statistics & Data Analysis*, 56(10), 2875–2888. https://doi.org/10.1016/j.csda.2012.02.006

Cressie N. (1993). *Statistics for spatial data*. John Wiley & Sons.

Datta G., Fay R., & Ghosh M. (1991). Hierarchical and empirical Bayes multivariate analysis in small area estimation. In *Proceedings of bureau of the census 1991 annual research conference* (pp. 63–79). U.S. Dept. of Commerce, US Bureau of the Census, Washington, DC.

Datta G., Kubokawa T., Molina I., & Rao J. (2011). Estimation of mean squared error of model-based small area estimators. *TEST*, 20(2), 367–388. https://doi.org/10.1007/s11749-010-0206-2

Datta G., & Lahiri P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10(2), 613–627. http://www.jstor.org/stable/24306735

Esteban M., Lombardía M., López-Vizcaíno E., Morales D., & Pérez A. (2020). Small area estimation of proportions under area-level compositional mixed models. *TEST*, *29*(3), 793–818. https://doi.org/10.1007/s11749-019-00688-w

Fay R. (1987). Application of multivariate regression of small domain estimation. In R. Platek, J. Rao, C. Särndal, & M. Singh (Eds.), *Small area statistics* (pp. 91–102). John Wiley & Sons.

Fay R., & Herriot R. (1979). Estimates of income for small places: An application of James stein procedures to census data. *Journal of the American Statistical Association*, *74*(366a), 269–277. https://doi.org/10.1080/01621459.1979.10482505

González-Manteiga W., Lombardía M. J., Molina I., Morales D., & Santamaría L. (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay–Herriot model. *Computational Statistics and Data Analysis*, *52*(12), 5242–5252. https://doi.org/10.1016/j.csda.2008.04.031

González-Manteiga W., Lombardía M. J., Molina I., Morales D., & Santamaría L. (2010). Small area estimation under Fay–Herriot models with nonparametric estimation of heteroscedasticity. *Statistical Modelling*, *10*(2), 215–239. https://doi.org/10.1177/1471082X0801000206

Guha S., & Chandra H. (2021a). Measuring and mapping disaggregate level disparities in food consumption and nutritional status via multivariate small area modelling. *Social Indicator Research*, *154*(2), 623–646. https://doi.org/10.1007/s11205-020-02573-8

Guha S., & Chandra H. (2021b). Measuring disaggregate level food insecurity via multivariate small area modelling: Evidence from rural districts of Uttar Pradesh. *Food Security*, *13*(3), 597–615. https://doi.org/10.1007/s12571-021-01143-1

Guha S., & Chandra H. (2022). Measuring and mapping micro level earning inequality towards addressing the sustainable development goals: A multivariate small area modelling approach. *Journal of Official Statistics*, *38*(3), 823–845. https://doi.org/10.2478/jos-2022-0036

Guha S., Das S., Baffour B., & Chandra H. (2023). Multivariate small area modelling of undernutrition prevalence among under-five children in Bangladesh. *The International Journal of Biostatistics*, *19*(1), 191-215. https://doi.org/10.1515/ijb-2021-0130

Guha, S., Chandra, H. (2024) Small area estimation under a spatially correlated multivariate area level model. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 187(1), 60-82. https://doi.org/10.1093/jrsssa/qnad079

Ito T., & Kubokawa T. (2021). Corrected empirical Bayes confidence region in a multivariate Fay–Herriot model. *Journal of Statistical Planning and Inference*, *211*, 12–32. https://doi.org/10.1016/j.jspi.2020.05.008

Spectral Modelling in Agriculture

Naveena K
Land and Water Management Research Group, Centre for Water Resources Development and
Management (CWRDM), Kunnamangalam, Kozhikode - 673571, Kerala, India.
Email: naveenak@cwrdm.org

Introduction

Spectral modelling has emerged as a cornerstone of modern precision agriculture, offering a rapid, non-destructive, and cost-effective method for assessing soil and crop characteristics. This technique leverages the principles of spectroscopy, analyzing how light across the electromagnetic spectrum is reflected or absorbed by a material. The resulting spectral signature a unique fingerprint for that material can be used to infer a wide range of physical and chemical properties, such as nutrient content, organic matter, moisture levels, and texture. This chapter provides a comprehensive, step-by-step protocol for implementing a robust spectral modelling workflow in an agricultural context. It is designed to guide scientists, students, and technicians through the entire process, from systematic field sampling and laboratory analysis to advanced data pre-processing and comparative machine learning. By combining rigorous field methodology with powerful data science techniques, this guide aims to empower researchers to generate accurate, reproducible, and actionable insights for enhancing soil health, optimizing resource management, and advancing sustainable agriculture.

2. Materials, Equipment, and Software

A successful spectral modelling project requires specific tools for field, laboratory, and computational work.

2.1. Field Equipment

- Soil Sampling: Power auger or trowel labelled and airtight sample bags, permanent markers.
- Georeferencing: GPS device for recording sampling locations.
- Layout: Measuring tape for establishing sampling grids.
- Documentation: Field notebook or digital data sheets.

2.2. Laboratory Equipment

- Processing: Drying trays for shade drying, 2 mm mesh sieve, weighing balance.

- Chemical Analysis: Standard laboratory glassware and reagents for nutrient analysis (e.g., Kjeldahl, Olsen methods).

2.3. Spectral Data Collection

- Instrumentation: Handheld spectroradiometer (typically covering the 350–2500 nm range).
- Calibration: White reference panels (e.g., PTFE, Halon) for radiometric calibration.
- Stability: Tripod or stand to ensure consistent measurement geometry.

3. Data Acquisition Protocol

The quality of a predictive model is fundamentally dependent on the quality of the data used to train it. This section outlines a systematic protocol for data acquisition.

3.1. Field Sampling Procedure

- Define Study Area & Grid Design: Clearly demarcate the study boundaries. Overlay a regular grid (e.g., 50 m × 50 m) to ensure unbiased spatial coverage.
- Mark Sampling Points: Use a GPS device to navigate to the center of each grid cell. This ensures the sampling is representative of all major land uses and soil types within the area.
- Composite Soil Collection: At each point, clear surface debris. Collect 8-10 subsamples of topsoil (0–15 cm depth) from random locations within a small radius. Combine these subsamples in a clean container and mix thoroughly to create a single, homogenized composite sample.
- Sample Reduction (Quartering): To obtain a manageable lab sample, spread the composite soil on a clean sheet, divide it into four equal parts, and discard two opposite quarters. Mix the remaining two and repeat the process until approximately 250 g of soil remains.
- Labelling and Storage: Place the final sample in a clearly labelled, airtight bag. Record the sample ID, GPS coordinates, date, time, and any relevant field observations.

3.2. Spectral Data Collection

Instrument Preparation: Power on the spectroradiometer and allow it to warm up and stabilize according to the manufacturer's guidelines.

- Calibration: Before the first measurement and periodically thereafter (especially if lighting changes), perform a white reference calibration using the reference panel. This converts raw digital numbers into reflectance values.
- Measurement Protocol: Place the processed (dried and sieved) soil sample in a shallow container (e.g., petri dish). Position the sensor perpendicular to the surface at a fixed height (e.g., 0.45 m). Conduct measurements during peak daylight hours (9:00 AM–1:00 PM) under clear, cloud-free conditions.

- Data Acquisition: For each soil sample, collect multiple readings (e.g., 7-10) from slightly different positions on the sample. Average these readings to create a single, representative spectrum with a high signal-to-noise ratio.

3.3. Laboratory Soil Nutrient Analysis

Following standard laboratory protocols, analyze each soil sample for the target nutrients of interest. Common methods include Kjeldahl for Nitrogen, Olsen for Phosphorus, and flame photometry for Potassium. Meticulous record-keeping is crucial to ensure each lab result is correctly matched with its corresponding sample ID and spectral file.

4. Data Processing and Pre-processing

Raw spectral data contains noise and artifacts that can obscure the relationship between reflectance and soil properties. Pre-processing is a critical step to clean the data and enhance the relevant spectral features.

4.1. Spectral Pre-processing Techniques

- Splice Correction: Correct for discontinuities at the detector overlap regions within the spectroradiometer.
- Smoothing: Apply algorithms like the Savitzky-Golay filter to reduce high-frequency instrumental noise while preserving the shape of key absorption features.
- Scatter Correction: Use transformations like Standard Normal Variate (SNV) or Multiplicative Scatter Correction (MSC) to remove variations caused by light scattering due to particle size differences, which are unrelated to chemical composition.
- Baseline Correction & Detrending: Remove additive and multiplicative baseline effects in the spectra.
- Continuum Removal: Normalize spectra by fitting a convex hull over the data. This technique isolates and enhances individual absorption features, making them more comparable across different samples.

5. Predictive Modelling with Machine Learning

With clean data, the next step is to train machine learning models to predict nutrient concentrations from the spectral data.

5.1. Data Preparation and Splitting

Merge the pre-processed spectral data with the laboratory nutrient data using the unique sample ID. Split this final dataset into a training set (typically 70-80% of the data) and a testing set (the remaining 20-30%). The model will be developed on the training set and its final performance will be evaluated on the unseen testing set.

5.2. Model Selection and Training

Two widely used algorithms for analyzing spectral data are Partial Least Squares Regression (PLSR) and Random Forest. It is recommended to train and compare multiple models for optimal results.

- Partial Least Squares Regression (PLSR): This linear regression method excels in handling datasets with more variables (*e.g.*, spectral bands) than samples, particularly when variables are highly correlated (collinear), as is common in hyperspectral data. PLSR reduces the variables into a smaller set of uncorrelated components and performs regression on these components.
- Random Forest: A non-linear, ensemble learning approach that constructs numerous decision trees during training. It is resistant to overfitting and effectively captures complex, non-linear relationships between spectral data and soil properties. Additionally, it provides a direct assessment of variable importance.

5.3. Model Validation and Comparison

Model robustness is evaluated through validation techniques. Cross-Validation: Employ k-fold cross-validation (e.g., k=10) during training to optimize model hyperparameters (such as the number of components in PLSR) and obtain a reliable estimate of performance.

Performance Metrics: Assess the model on a held-out test set using the following metrics:

- Coefficient of Determination (R^2): Measures the proportion of variance in the observed data explained by the model, with values closer to 1 indicating a better fit.
- Root Mean Square Error (RMSE): Calculates the square root of the average squared differences between observed and predicted values, providing error in the same units as the target variable.
- Mean Absolute Error (MAE): Computes the average absolute differences between observed and predicted values, offering a metric less sensitive to large outliers compared to RMSE.

6. Interpretation and Application: A predictive model is only useful if its results can be interpreted and applied.

6.1. Identifying Key Spectral Bands: Both PLSR and Random Forest models can provide insights into which spectral bands are most influential in predicting the target nutrient.

- PLSR Variable Importance in Projection (VIP): A weighted sum of the squared correlations between the PLS components and the original variable. Bands with a VIP score greater than 1 are generally considered important.
- Random Forest Feature Importance: Calculated by measuring how much the prediction error increases when a specific band's data is randomly shuffled.
- Identifying these bands helps validate the model against known soil science (e.g., are the bands related to organic matter or clay minerals?) and can inform the development of simpler, more targeted multispectral sensors in the future.

6.2. Generating Prediction Maps

By applying the validated model to new spectral data (collected in the field or from airborne/satellite sensors), one can generate high-resolution spatial maps of soil nutrient

concentrations. These maps are powerful tools for precision agriculture, enabling farmers to apply fertilizers and soil amendments with high spatial precision, leading to reduced costs, lower environmental impact, and improved crop yields.

7. Practical Implementation: An R-Based Workflow

The following R script provides an end-to-end, enhanced workflow for the methods described in this chapter. It includes data loading, pre-processing, training and comparison of PLSR and Random Forest models, validation, and visualization.

```
# Required packages
required_packages <- c("pls", "ggplot2", "dplyr", "prospectr", "randomForest", "gridExtra")
# Install and load packages
invisible(lapply(required_packages, function(pkg) {
  if (!require(pkg, character.only = TRUE)) install.packages(pkg, dependencies = TRUE)
  library(pkg, character.only = TRUE)
}))
# User Configuration
SPECTRAL_FILE <- "spectral_data.csv"          # Spectral reflectance data
NUTRIENT_FILE <- "nutrient_data.csv"          # Nutrient reference data
TARGET_VARIABLE <- "Nitrogen"                  # Target nutrient: Nitrogen, Phosphorus, etc.
TEST_SET_RATIO <- 0.3                         # Test set proportion (e.g., 0.3 = 30%)
MAX_PLSR_COMPONENTS <- 20                     # Max number of PLS components to test
RANDOM_SEED <- 123                            # Seed for reproducibility
# DATA LOADING & PREPROCESSING
# Load and merge data
tryCatch({
  spectra <- read.csv(SPECTRAL_FILE)
  nutrients <- read.csv(NUTRIENT_FILE)
  full_data <- merge(spectra, nutrients, by = "SampleID")
}, error = function(e) {
  stop(" Error: Unable to load or merge datasets. Please check file paths and 'SampleID' column.")
})
# Prepare predictors and response
X <- as.matrix(full_data %>% select(starts_with("X")))
Y <- full_data[[TARGET_VARIABLE]]
# --- Optional Spectral Preprocessing ------------------------
# Uncomment one or more lines below if preprocessing is desired
# X <- savitzkyGolay(X, p = 2, w = 11, m = 1)  # Savitzky-Golay smoothing
# X <- standardNormalVariate(X)               # Standard Normal Variate correction
# TRAIN-TEST SPLIT
set.seed(RANDOM_SEED)
train_indices <- sample(seq_len(nrow(full_data)), size = (1 - TEST_SET_RATIO) * nrow(full_data))
```

```
X_train <- X[train_indices, ]; Y_train <- Y[train_indices]
X_test  <- X[-train_indices, ]; Y_test  <- Y[-train_indices]

# --- 4. MODEL TRAINING ---------------------------------------
# --- 4.1 Partial Least Squares Regression (PLSR) -------------
cat("\n🔧 Training PLSR model...\n")
pls_model <- plsr(Y_train ~ X_train, ncomp = MAX_PLSR_COMPONENTS, validation
= "CV", method = "oscorespls")
opt_ncomp <- which.min(RMSEP(pls_model)$val[1, 1, ])
cat("Optimal number of PLS components:", opt_ncomp, "\n")
Y_pred_pls <- as.vector(predict(pls_model, newdata = X_test, ncomp = opt_ncomp))
pls_results <- data.frame(Observed = Y_test, Predicted = Y_pred_pls)
# --- 4.2 Random Forest Regression ----------------------------
cat("\n Training Random Forest model...\n")
rf_model <- randomForest(x = X_train, y = Y_train, ntree = 500)
Y_pred_rf <- predict(rf_model, newdata = X_test)
rf_results <- data.frame(Observed = Y_test, Predicted = Y_pred_rf)
# --- 5. MODEL EVALUATION -------------------------------------
calculate_metrics <- function(observed, predicted) {
  data.frame(
    R2   = round(cor(observed, predicted)^2, 4),
    RMSE = round(sqrt(mean((observed - predicted)^2)), 4),
    MAE  = round(mean(abs(observed - predicted)), 4)
  )
}
cat("\n Model Performance Comparison:\n")
comparison_df <- rbind(
  PLSR = calculate_metrics(pls_results$Observed, pls_results$Predicted),
  `Random Forest` = calculate_metrics(rf_results$Observed, rf_results$Predicted)
)
print(comparison_df)
```

8. Conclusion and Future Directions

This chapter has detailed a comprehensive workflow for spectral modelling of soil nutrients, integrating best practices in field sampling, data processing, and machine learning. The ability to generate rapid and accurate predictions of soil properties from spectral data represents a significant leap forward for agricultural management. Future research will likely focus on several key areas: data fusion (combining spectral data with other data sources like topography or climate data), deep learning models (like Convolutional Neural Networks) for more automated feature extraction, and the development of global, open-source spectral libraries to make these techniques more accessible to researchers and farmers worldwide. By continuing to refine these methods,

spectral modelling will play an increasingly vital role in ensuring global food security and environmental sustainability.

References

Krishnamurti, T. N., Hardiker, V. M., Bedi, H. S., & Ramaswamy, L. (2006). An introduction to global spectral modeling. New York, NY: Springer New York.

Krishna, G., Sahoo, R. N., Singh, P., Bajpai, V., Patra, H., Kumar, S., ... & Sahoo, P. M. (2019). Comparison of various modelling approaches for water deficit stress monitoring in rice crop through hyperspectral remote sensing. Agricultural water management, 213, 231-244.

Peddle, D. R., & Smith, A. M. (2005). Spectral mixture analysis of agricultural crops: endmember validation and biophysical estimation in potato plots. International Journal of Remote Sensing, 26(22), 4959-4979.

Sahoo, R. N., Ray, S. S., & Manjunath, K. R. (2015). Hyperspectral remote sensing of agriculture. Current science, 848-859.

Quantification of carbon sequestration of agroforestry systems: Allometry approach

Sangram B Chavan[1], Nobin Chandra Paul[1], Uthappa AR[2], Keerthika A[3]
[1]ICAR-National Institute of Abiotic Stress Management, Baramati, Maharashtra-413115
[2]ICAR-Central Coastal Agricultural Research Institute, Ela, Goa
[3]ICAR-Central Arid Zone Research Institute, RRS, Pali-Marwar, Rajasthan- 306 401
Email: sangramc8@gmail.com

1. Introduction

Carbon sequestration is a vital strategy for mitigating the impacts of climate change. It is a process which entails capturing atmospheric CO_2 and storing it in long-term biological reservoirs. Agroforestry systems hold considerable potential for carbon storage, primarily through their aboveground and below-ground biomass, which act as key carbon sinks (Fig. 1). These systems facilitate the uptake of atmospheric CO_2 during the process of photosynthesis, capturing and storing fixed carbon in trees, detritus, and soil, thus contributing to a safer environment. Agroforestry's prominence is evident from its widespread adoption across the globe. With approximately 1 billion hectares of agricultural land practicing agroforestry, benefitting around 1.5 billion farmers, its significance in enhancing sustainability and resilience is increasingly recognized. Additionally, vast unproductive croplands and grasslands hold the potential for future agroforestry expansion, further increasing carbon sequestration capacity. In India, agroforestry covers roughly 25 million hectares, representing 8.2% of the country's declared geographical area. Trees are present not only on farmland but also within agricultural lands, making agroforestry an integral component of the country's land-use strategy. Different agroforestry practices store varying amounts of carbon, which depends on factors such as system type, species composition, soil characteristics, and climate conditions. Recognizing the significance of agroforestry in climate change mitigation, India has enacted comprehensive set of policies, missions, and national action plans to promote tree-based farming systems. The Green India mission, National Agroforestry Policy, and National Agroforestry and Bamboo Mission are some notable initiatives aimed at increasing tree cover and offsetting greenhouse gas emissions.

Quantifying carbon stocks in agroforestry systems remains challenging, but it is essential for monitoring and reporting carbon storage accurately. Ground-based estimations, aided by allometric equations, provide valuable reference points for regional biomass carbon mapping. These equations establish precise relationships between tree attributes, facilitating reliable biomass estimation. However, most existing allometric equations are developed for forests or sole plantations and may underestimate biomass carbon in agroforestry due to specific planting geometries, local environmental conditions, and tree management practices. Under such context, this study aims to develop and refine allometric

equations tailored to agroforestry trees, improving carbon stock estimations in this critical land-use system. By deepening our understanding on potential to sequester carbon, we can contribute to national and global initiatives aiming to combat climate change, achieve climate targets, and create sustainable, adaptable and resilient agroforestry landscapes.

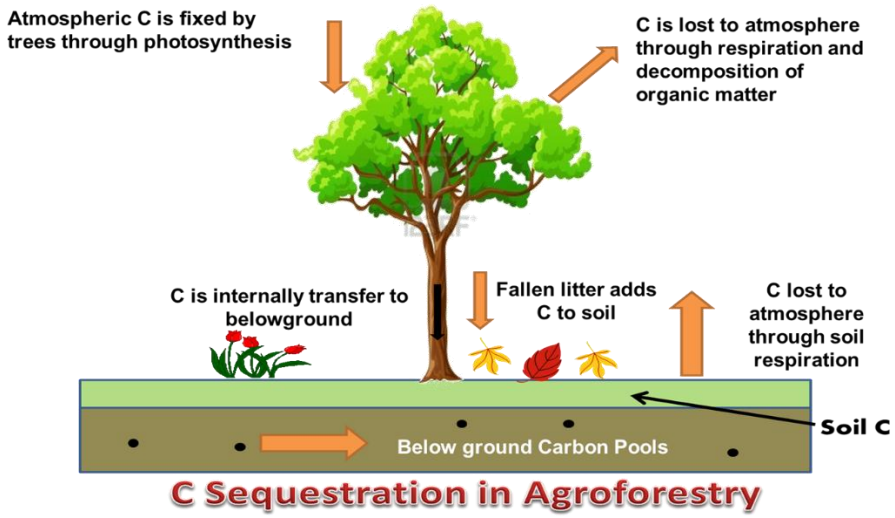


Fig1 : Process and carbon pools in agroforestry

2. Carbon pools in Agroforestry

Carbon pools in agroforestry refer to the different components of the agroforestry system where carbon is stored. These pools play an important role in carbon sequestration, facilitating to remove carbon dioxide from the atmosphere and store it in long-lived resilient forms. The major carbon pools in agroforestry include:

1. Aboveground Biomass: This pool comprises the living components of trees, including stems, branches, leaves, and fruits. Aboveground biomass is a significant carbon store, as trees absorb CO_2 from the atmosphere during the process of photosynthesis and converting it into organic carbon in their tissues.
2. Belowground Biomass: Belowground biomass includes tree roots and associated soil organisms. Roots contribute to carbon storage by accumulating organic matter, and soil microorganisms help decompose organic materials, converting them into stable soil carbon.
3. Soil Organic Carbon: Agroforestry practices enhance soil carbon sequestration due to the presence of the trees and diverse vegetation or species composition. Trees shed leaves, branches, and other organic matter, contributing to soil organic carbon content. Additionally, the root systems of trees enhance soil structure and facilitate the accumulation of carbon in the soil.

4. Litter and Detritus: Litter and detritus refer to dead plant material, such as fallen leaves, twigs, and fruits, as well as decomposed organic matter on the forest floor. These materials gradually decompose, releasing CO_2 back into the atmosphere, but a significant portion can also be incorporated into the soil carbon pool.
5. Dead Wood: Dead wood includes standing or fallen dead trees and branches. This carbon pool contributes to carbon storage in the ecosystem as it decomposes slowly, depending on factors like climate and microorganisms present.

Each of these carbon pools plays a vital role in the overall carbon sequestration capacity of agroforestry systems. The management practices, tree species composition, soil conditions, and climate of the specific agroforestry system influence the distribution and dynamics of these carbon pools. By understanding and optimizing these pools, agroforestry can serve as a valuable climate-smart land-use option, playing an important role in climate change mitigation and sustainable agricultural practices. The protocol to estimate carbon stock in different agroforestry systems in provided in fig 2.

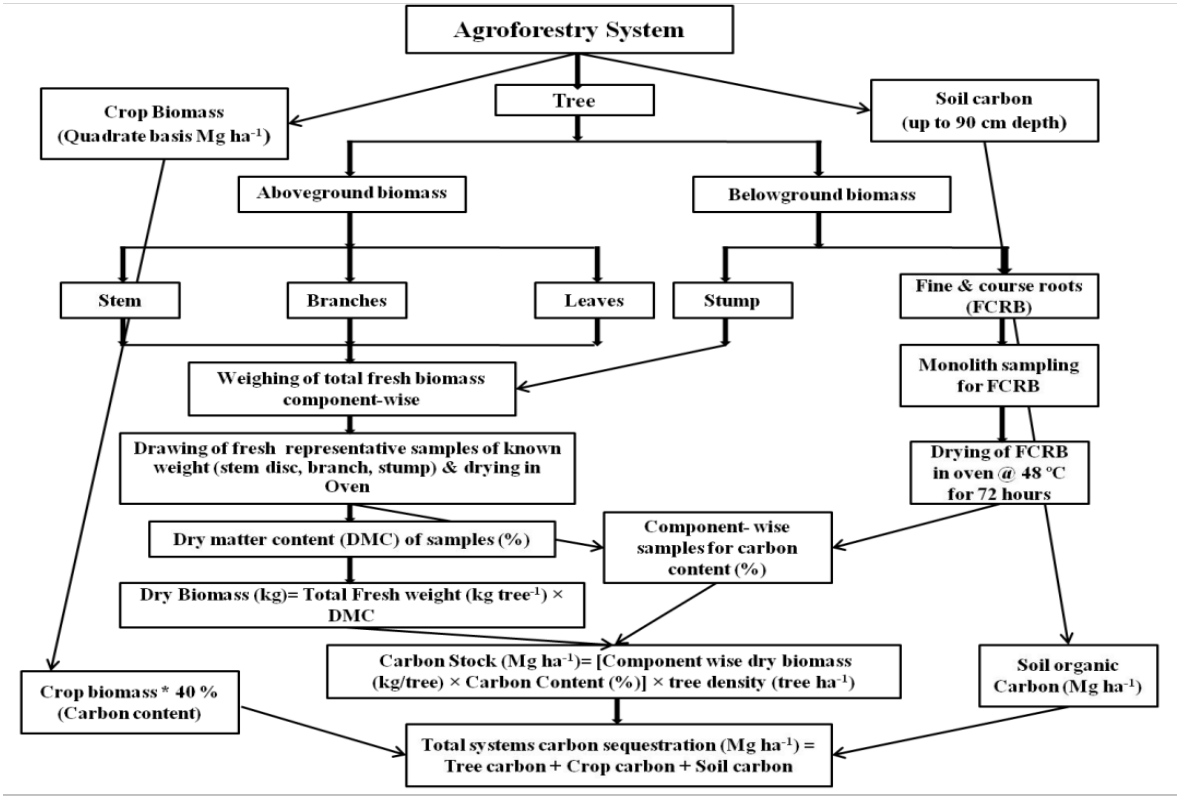


Fig 2: Flowchart of Measurement of CSP of agroforestry in Haryana (Chavan, 2019).

3. Allometry equations

An allometric relation isa type of relationship in which one measured parameter serves as a reliable indicate of another unmeasured parameter. To estimate biomass, researcher can

create allometric equation that allows for the calculation of a tree's mass based on few morphological parameters. The developed equation are practically used to estimate biomass in trees within forest. The concept allometry is defined as "the measure and study of relative growth of a part in relation to an entire organism or to a standard". Galileo Galilee first described this principle in 1630's, emphasizing the proportions of an organism must change as it grows. Allometric equations, relating biomass with one or more tree dimensions, are frequently used to compute average tree biomass. Non-destructive techniques for component-wise biomass estimation are preferred, although they often require harvesting various tree parts, including felling entire trees, to develop equations for biomass estimation. Allometric equations, which link tree diameter at breast height (1.37 m) to attributes such as standing carbon stock and leaf area, serve as vital tools in both ecological research and commercial applications. These equations represent the primary method for estimating above-ground forest dry matter or carbon.

If we call biomass B and diameter D, this second definition means that there is a coefficient a such that:

$$\frac{dB}{b} = a\frac{dD}{D}$$

which integrates to a power relationship: $B = b \times D^a$ The parameter 'a' represents the allometric coefficient, reflecting the proportionality between relative increases, while parameter 'b' denotes the proportionality between cumulative variables. In some cases, it may be necessary to include a y-intercept in the equation, resulting in the form $B = c + bD^a$, where c represents the biomass of an individual tree before it reaches the height at which diameter is measured (e.g., 1.30 m if D is measured at that height).

In simple terms, regression is a statistical method used to create mathematical models that define the relationship between variables. These models can be used to estimate, interpolate, or extrapolate values. For example, to estimate tree biomass (weight), factors such as height, diameter at breast height (DBH), wood density, form factor, and tree management practices (like pruning or thinning) are important. In this case, biomass is the dependent (predicted) variable, while height, DBH, wood density, form factor, and tree management are the independent (predictor) variables.

3.1 Assumptions of Regression analysis

Regression analysis makes several major assumptions to the validity and reliability of the results. Let's discuss these assumptions in the context of a regression analysis with tree diameters as independent parameters and tree biomass as the dependent parameter:

- Linearity: The relationship between the independent variable (tree diameters) and the dependent variable (tree biomass) should be linear. In other words, the effect of tree diameters on tree biomass should be constant and proportional. In a linear regression analysis, we assume that an increase in tree diameters will lead to a

proportional increase in tree biomass. For instance, if the diameter of a tree doubles, the biomass should also double.

- Independence: The observations in the regression analysis should be independent of each other. Each data point's value should not be influenced by or dependent on the value of another data point. Example: In a dataset with tree diameters and biomass measurements, each tree's diameter and biomass should be measured independently, without being influenced by other trees' measurements.
- Homoscedasticity: The residuals (the differences between the observed tree biomass and the predicted tree biomass based on diameters) should have constant variance across all levels of tree diameters. Example: Homoscedasticity means that the spread of the residuals around the regression line should be consistent for all tree diameters. In other words, the variability of the errors should not change as tree diameters increase or decrease.
- Normality of Residuals: It must follow a normal distribution with a mean of zero. This assumption ensures that the errors are unbiased and normally distributed. Example: When plotting the residuals, they should roughly follow a bell-shaped curve around zero.
- No Perfect Multicollinearity: There should be no perfect linear relationship between the independent variables (tree diameters). Multicollinearity occurs when two or more independent variables are highly correlated with each other, making it difficult to separate the individual effects on the dependent variable. Example: If tree diameters are highly correlated with each other (e.g., trees of similar sizes tend to grow together), this could lead to multicollinearity issues.
- No Autocorrelation: The residuals should not exhibit any patterns or correlations among themselves. Autocorrelation occurs when the residuals at one observation are correlated with the residuals at nearby observations. Example: In a time series analysis of tree growth, if the residuals from one year's biomass estimation are correlated with the residuals from the next year's estimation, it indicates the presence of autocorrelation.

It is essential to assess and validate these assumptions before interpreting the results of the regression analysis. Violation of these assumptions may lead to biased estimates and affect the validity of the regression model. Various diagnostic tools and tests are available to check the assumptions' validity and address any issues that may arise during the analysis. Developing allometry equations for biomass estimation involves several steps, including data collection, statistical analysis, and model validation.

- Step 1: Data Collection

  Collect a dataset that includes measurements of tree biomass and easily measurable tree attributes, like diameter at breast height (DBH), height, or volume. Ideally, the dataset should cover a wide range of tree species and sizes to ensure the generalizability of the allometry equation.

- Step 2: Data Exploration and Visualization
  Plot the relationship between tree biomass and the selected tree attribute(s) using scatter plots. This helps identify the general trend between biomass and the predictor variable and any potential nonlinearities.
- Step 3: Data Transformation (if needed)
  If the relationship between predictor variable and biomass appears nonlinear, consider applying data transformations. Common transformations include logarithmic, power, or square root transformations. Plot the transformed data to check if it better fits a linear relationship.
- Step 4: Model Selection
  Select the appropriate model for biomass estimation. For most cases, a linear model is used, but in some cases, nonlinear models may be more suitable. Linear models are of the form: biomass = $\beta 0 + \beta 1 *$ predictor, where $\beta 0$ and $\beta 1$ are coefficients to be estimated.
- Step 5: Regression Analysis
  Perform regression analysis to estimate the model coefficients ($\beta 0$ and $\beta 1$). Use the least squares method to minimize the sum of squared differences between the observed biomass and the predicted biomass from the model.
- Step 6: Model Evaluation
  Evaluate the quality of the allometry equation using statistical metrics such as R-squared (coefficient of determination), root mean square error (RMSE) or mean absolute error (MAE). These metrics help to assess how well the model fits the data.
- Step 7: Graphing the Allometry Equation
  Once you have developed the allometry equation, plot the observed biomass and the predicted biomass based on the equation on the same graph. This graph allows you to visualize how well the equation represents the actual data points.
- Step 8: Validate the Allometry Equation
  Validate the allometry equation using an independent dataset. If possible, collect additional data from different locations or tree species to verify the equation's accuracy and applicability.
- Step 9: Interpretation and Application
  Interpret the coefficients of the allometry equation to understand the relationship between tree attributes and biomass. The equation can also be used to determine the tree biomass in other locations or to project future biomass changes.
- Step 10: Sensitivity Analysis (Optional)
  This is conducted to assess the impact of uncertainties or variations in the data on the allometry equation's performance.

Throughout the process, data visualization using graphs is essential to understand the relationship between variables, identify potential issues, and assess the model's

performance. Scatter plots, line plots of the regression model, and residual plots are common graphical tools used in the development and validation of allometry equations.

3.2 Model Diagnostics for Biomass Estimation

Once the model has been optimized, the final phase of development of model involves rigorous statistical testing to ensure the model's reliability for predictive purposes, whether it be interpolation or extrapolation. In addition to the previously mentioned parameters (RSS, F-ratio, and R2), we employ residual diagnostic plots for further evaluation (Fig 3). These plots include probability plots of residuals, autocorrelation plots of residuals, plots of residuals against their expected values, and plots of residuals against the independent variable (tree diameter).

1. To verify the normal distribution of residuals, we plot them against their expected values for normality assessment.

2. To check for independence and no correlation among residuals, we create an autocorrelation plot of the residuals.

3. To validate that the residuals exhibit constant variance, we plot them against the independent variable (tree diameter).

4. To verify that the residuals are not consistently over- or under-estimated, we plot them against the independent or explanatory variable.

In addition to these classical procedures, we employ recent statistical model validation techniques to develop robust models. These include paired t-tests between observed and predicted values and fitting linear regressions between them. As an independent dataset was unavailable for validation, we randomly divided the original datasets into two mutually exclusive and pseudo-independent datasets containing 80% and 20% observations, respectively, for model estimation and validation (Ajit et al., 2011).

The following methodology was implemented:

1. Calculate the mean of the residuals (r= observed-predicted) and the 95% confidence interval.
2. Use paired-t tests to test the unbiasedness of the model (Neter et al., 1996).
3. Fit a linear regression between observed values (y-obs) and estimated or predicted values (y-pred), i.e., (observed) = a + b * (predicted). For a perfect model, we expect 'a' to be zero and 'b' to be one. The adjusted R2 must also be one (Amaro, 1998).

By diligently employing these validation procedures, we ensure the model's accuracy and credibility, making it a valuable tool for precise biomass estimation in forest and agroforestry studies.

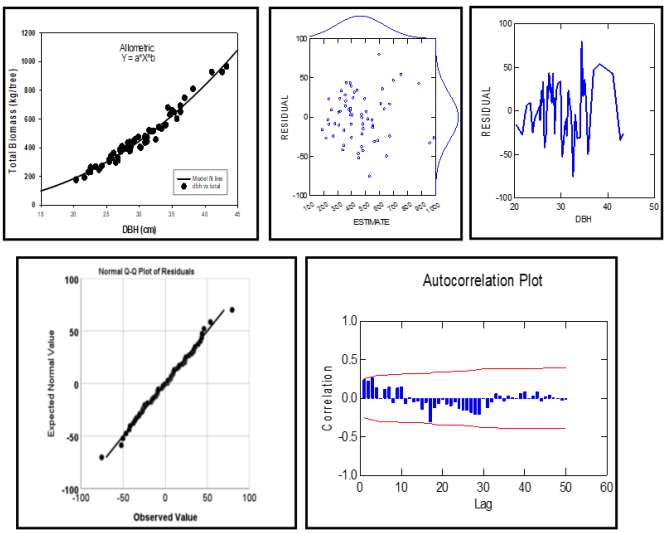


Fig 3: Plots of residuals against the values of predicted and explanatory variate of total biomass

4. Protocol for estimation of biomass:

To construct a regression equation for allometry, it is necessary to destructively harvest trees and measure their components individually for biomass and volume.

A) Biomass Measurement

1. Live standing trees with a diameter at breast height (DBH) of 5 cm and above within the sample plots are selected for measurement. The following data were collected:

   i) Tree species ii) DBH of individual tree and iii) Total tree height. The methodology for destructive sampling to estimate tree biomass were adopted from FAO (2012). The key steps involved in the measurement process are:

   Clearing ground vegetation for tree access especially to make tree identification easier is the first step before measuring DBH.
2. Marking measuring position for DBH using a 1.3 m pole (See box 1)
3. Record all necessary details, including the number of stumps, buttress diameter, buttress height, and note any irregularities observed in the tree form or structure.
4. Group the trees into DBH classes using 10 cm intervals and enter in data sheet. The DBH classes include: 5–15 cm, 15–25 cm, 25–35 cm, 35–45 cm, 45–55 cm, 55–65 cm and 65–75 cm
5. Randomly select sample trees from each DBH class within the sample plots.
6. After selecting the sample trees, fell each tree at its base using a chainsaw, following appropriate logging and safety procedures.
7. Once a sample tree is felled, measure the following parameters accurately: a. Diameter at stump and DBH at 1.37 m; b. Total tree height (from the stump to the top of the crown), c. Length of the tree bole (from the stump to the first main branch), d. Length of the bole from the stump to the point where the

diameter reduces to 10cm and e. For trees with buttresses, record the diameter and height of the buttress

8. Separate the felled tree into different components such as bole, branches, and leaves.
9. Immediately measure the fresh weight of each component—stem, branches, leaves, and buttress (if present).
10. Record all observations and measurements related to the destructive sampling of each tree on the designated data sheet.
11. The complete overview of the tree selection process and the component-wise measurement of biomass is given in Figure 4.

12. Tree diameter (cm)

Girth at breast height (GBH) was measured at 1.37 m above the ground (over bark) using a measuring tape and recorded in centimetres (cm). Additional observations includes collar girth (measured 15 cm above ground), girth at breast height (1.37 m), mid-girth, and top girth recorded in centimetres. For stem logs, base girth, mid-girth, and top girth were measured separately to facilitate accurate volume estimation of the tree (in cubic metres, m^3). The girth was converted to diameter by using following formula:

$$DBH\ (cm) = \frac{GBH\ (cm)}{\pi}$$

Where, DBH: diameter at breast height & GBH: girth at breast height at 1.37 m on ground

Tree height was measured in metres (m) using a measuring tape. The height was determined from the base of the harvested tree (stump) to the longest tip of the felled tree laid on the ground.

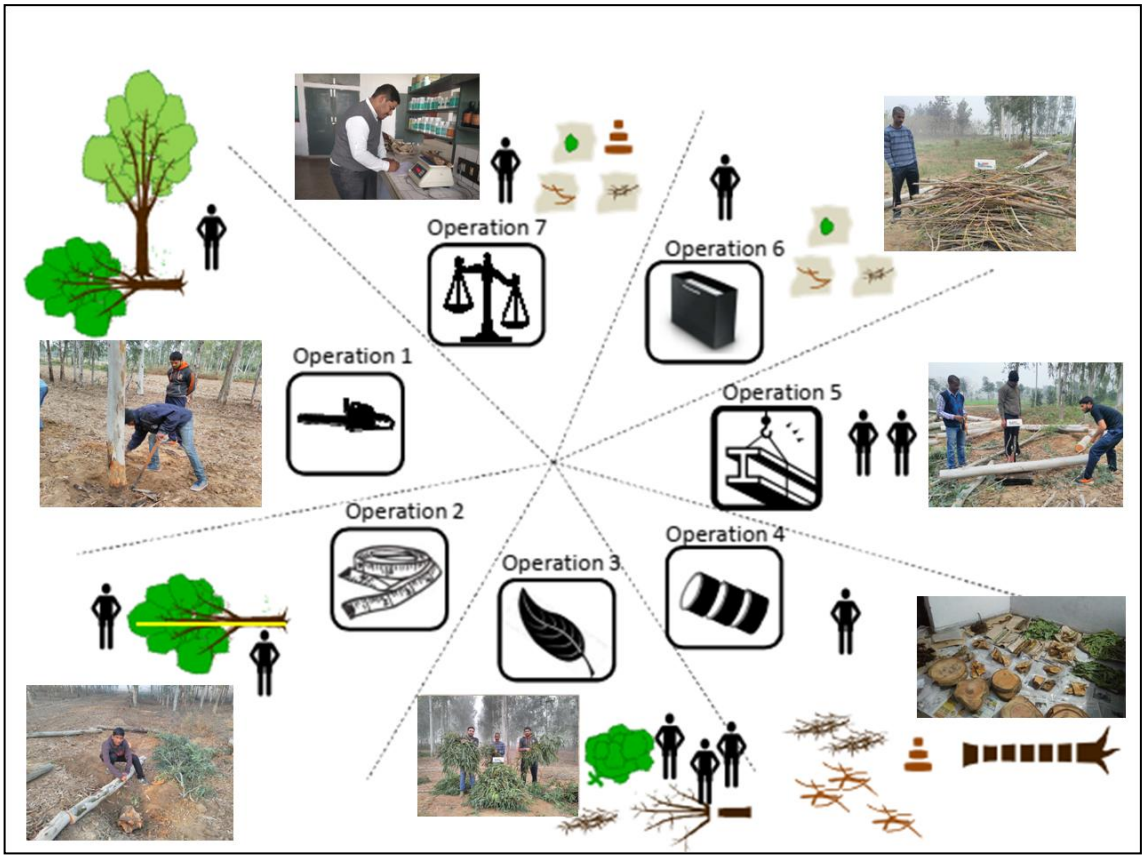


Figure 4: *Destructive sampling of biomass & their measurement in 7 different felling rules. Rule 1: site preparation and tree felling; Rule 2: measurement of felled trees stem profile, marking for cross-cutting; Rule 3: stripping of leaves;Rule 4: component wise biomass separation; operation Rule 5: weighing of logs; Rule 6, sampling of branches; & operation Rule 7: sample weighing and oven drying in laboratory*

Sampling for dry mass analysis should be conducted immediately after recording the fresh weight of each tree component. The following procedure is recommended:

13. Sample Collection for Dry Mass Analysis

Three representative samples per tree were collected from the stem, branches, and leaves. It is crucial that each sample accurately represents the variability within the respective component. The sampling guideline includes **Stem Samples:** Take two-to-three-disc sections from different positions along the stem. If the discs are too large, radial sections can be extracted instead. The total sample weight should represent approximately 0.2% of the total stem fresh weight. **Branch Samples:** Collect about 0.5 to 1.0 kg of material by selecting four small disc sections from different branches. **Leaf Samples:** Select leaves from various parts of the crown to ensure representation of both sun-exposed and shaded foliage. Place each collected sample (stem, branches, and leaves) into polyethylene bags. The recommended weight of each sample is: **0.5 – 1.0 kg** for stem and branch samples and **0.3 – 0.5 kg** for leaf samples

14. Wood Density Analysis

To determine wood density, four-disc samples were taken from the bole of the tree. Mark and collect samples from the following standardized positions along the length of the tree bole: Stump level (0.0 m),1/4 of the total bole length,1/2 of the total bole length and 3/4 of the total bole length. At each marked position, extract one wood disc (or a radial section, if the bole is too large to handle). The thickness of each wood disc should be between 5 to 10 cm. after collection, tie the bags tightly to prevent moisture loss due to evaporation during transport or temporary storage.Clearly label the bag using a permanent marker. The information like i) Plot code ii) Tree name iii) DBH size and iv) Sample type (e.g., stem, branch, or leaves) were written on the bag. The fresh weight of each dry mass sample must be measured either on-site or off-site, but must be completed within the same day to ensure accuracy and prevent moisture loss. All collected samples (for both dry mass and wood density analysis) must be properly packed and sent to qualified laboratory. Timely processing is essential to ensure the accuracy and reliability of biomass and wood density data.

15. Laboratory analysis

1) Total dry weight (TDW) for each part of the sample tree:

The total dry weight (TDW) for each part of the sample tree is calculated by combining the total fresh weight (TFW) measured in the field with the dry-to-fresh weight ratio obtained from laboratory analysis of samples. The formula is:

$$TDW = TFW \frac{SDW}{SFW}$$

Where: *TDW* is total dry weight; *TFW* is total fresh weight; *SDW* is absolute dry sample weight and *SFW* is fresh sample weight.

1) Wood density:

Wood density of every wood disc for each sample tree species were analyzed in the laboratory and is calculated by the following formula:

$$WD = \frac{SDWc}{SV}$$

Where: *WD* is wood density in g/cm^3; *SDWc* is dry weight of sample cube and *SV* is volume of sample cube.

A) Volume measurement:

For volume equations, only commercial tree portion is considered in case of teak, dalbergia, sal, gmelina etc species.

Tree volumes can be estimated in different ways. Most method estimate tree volume using different formulas. After felling of trees, stem is further cut in to standard logs of 1-3 m as per the industrial requirement (Figure 5). A girth of each log measures at three points i.e. base of log (dl), mid of log (dm) and top of the log (du) to calculate merchantable volume of the tree by Newton's formula (Husch *et al.*, 1982) and the sum of all the logs of trees calculated as tree volume (m^3 tree^{-1}). The girth (cm) was converted to diameter (cm).

Here; d_l: denotes for Lower end, (d_m) for middle and (d_u) for upper end diameters of log.

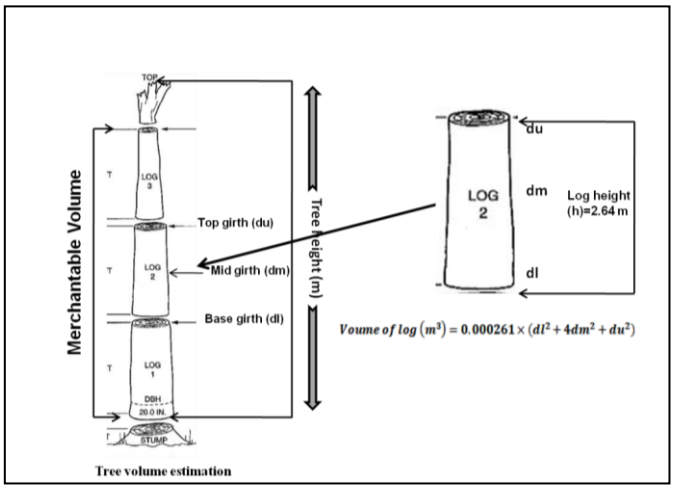
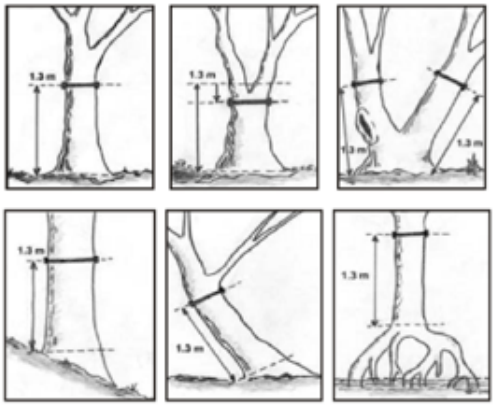


Figure 5: Sectional pieces of the main bole for computation of volume

Box 1. DBH measurement guideline

1. Sloping Ground: Measure the distance from the uphill side of the stem.
2.Leaning Trees on Level Ground: Measure from the under-side of the tree, parallel to the axis of the stem
3. Trees Forked Below Breast Height: Treat as two individual trees (double stems) and measure each separately.
4. Trees Forked Above Breast Height: Treat as a single stem and measure according to the tree's position on the ground or slope.
5. Trees Forking at or Just Above Breast Height: Measure at the point of minimum diameter below the fork.
6. Coppice Growth: Measure from ground level, not from the stool.
7. Surface Materials: Remove vines, moss, loose bark, and any other loose materials at breast height before measuring.
8. Fixed Height: Use a fixed height (bh) stick to ensure measurement is consistently taken at 1.37 m.
9. Measuring Technique: Measure at right angles to the stem axis and keep the tape taut for accuracy.



Data sheets for recording the tree harvesting data

Species Name		Place		Age	
Tree No.		Spacing			
Tree height (m)		GBH (cm)			
Log number	Log length (m)	Log girth			
		Lower end	Middle end	Top end	Fresh Weight
1					
2					
3					
4					
5					
6					
7					
8					
Fresh weight of branch (kg)					
Fresh weight of leaves (kg)					
Sample	Fresh weight	Dry weight			
Bole					
Leaves					
Branch					

References

Chavan, S. B., Dhillon, R. S., Ajit, Rizvi, R. H., Sirohi, C., Handa, A. K., ... & Kumari, S. (2022). Estimating biomass production and carbon sequestration of poplar-based agroforestry systems in India. Environment, Development and Sustainability, 24(12), 13493-13521.

Chavan, S. B., Dhillon, R. S., Sirohi, C., Uthappa, A. R., Jinger, D., Jatav, H. S., ... & Rajput, V. D. (2023). Carbon sequestration potential of commercial agroforestry systems in Indo-Gangetic Plains of India: Poplar and eucalyptus-based agroforestry systems. Forests, 14(3), 559.

Nair, P. K. R. (2012). Carbon sequestration studies in agroforestry systems: a reality-check. Agroforestry systems, 86(2), 243-253.

Abiotic Stress Management Using Crop Simulation Modelling

Sarath Chandran M.A.
ICAR – Central Research Institute for Dryland Agriculture, Hyderabad, Telangana – 500059
Email: sarath.crida@yahoo.com

Introduction

Agriculture is a complex system which involves interactions between biotic and abiotic factors. Abiotic stresses, such as drought, heat, salinity, and nutrient deficiency, are major constraints to crop productivity globally, particularly under changing climate conditions. Managing these stresses requires a detailed understanding of plant-soil-climate interactions, which can be effectively addressed using crop simulation models. Crop simulation models (CSMs) are valuable decision-support tools that simulate crop growth, development, and yield as a function of genotype, environment, and management (G × E × M) interactions. These models help in evaluating the impacts of abiotic stresses, testing adaptation strategies, and optimizing resource use.

The necessity of models in agriculture

Traditionally, research in agricultural sciences is undertaken by conducting field experiments. Experimental designs and treatments are used to conduct the field experiment, which in turn, will test the hypothesis. This approach requires a lot of land, labour, and input, which ultimately results in increased cost of conducting field experiments. Models emerged as an alternative to this approach. A scientific model is a physical and/or mathematical and/or conceptual representation of a system of ideas, events or processes. This helps to avoid the requirement for conducting field experiments year after year, saving costs required for land, labour and input. Moreover, models can evaluate multiple scenarios, which, with the traditional approach, would take more time to conclude.

Crop model

A model serves as a mathematical depiction of a real-world system, incorporating the latest scientific insights from diverse fields such as crop physiology, plant breeding, agronomy, agrometeorology, soil physics, soil chemistry, soil fertility, plant pathology, entomology, and more. It is grounded in an understanding of how plant genetics, soil, and crop management practices interact. These models are process-oriented, simulating crop growth, development, and yield based on factors like genetics, weather patterns, soil characteristics, and farming techniques. A flow diagram illustrating the input, process, and output components of crop modeling is shown in Figure 1.

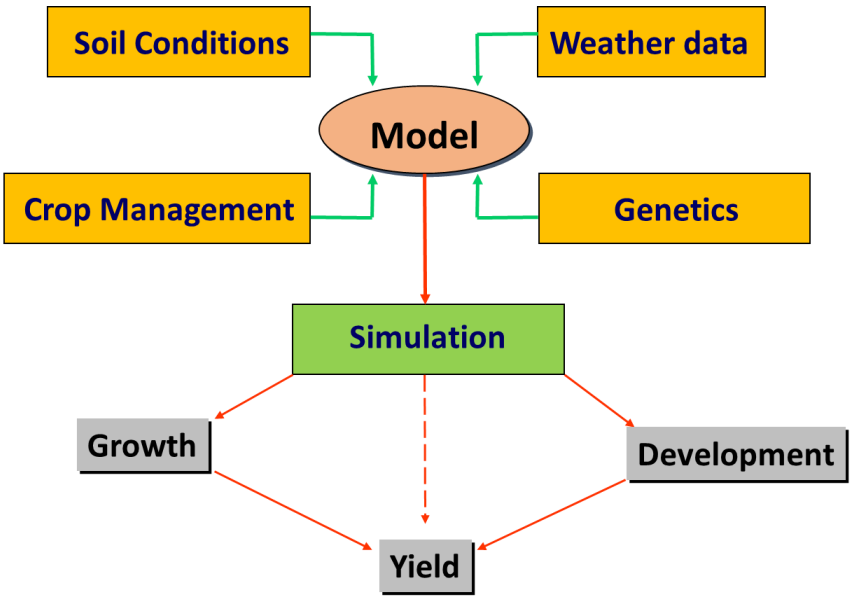


Fig. 1 Flow diagram depicting input, process and output of crop modelling

Applications of crop models

One key use of crop models is in conducting yield gap analysis. The yield gap for a crop in a specific location and cropping system refers to the difference between the yield obtained under ideal management conditions and the typical yield achieved by farmers. It offers the framework for determining which crop is most crucial, as well as the soil and management variables affecting present farm yields and better ways to bridge the yield gap. In precision agriculture, crop models are used to diagnose the factors causing yield variations and recommend location-specific crop management practices. It is also applied

in water and nutrient management, plant breeding, yield prediction, etc. The other areas of application include climate risk assessment, climate change impacts and adaptation, soil carbon sequestration studies, etc. Many popular crop simulation models are applied worldwide for these applications.

Steps in crop modelling for abiotic stress management

Calibration: This step involves model parameterization, where the genetic coefficients of the models are calibrated. These coefficients are mathematical representations crafted to replicate the phenotypic effects of genes across various environments, encompassing processes such as the life cycle (including fractional allocation across phases), photosynthesis, vegetative growth, reproductive processes, and root development. Specific to each cultivar, these coefficients are derived from experimental data and are vital in predicting how a particular crop variety responds to weather, soil, and management practices. Precise genetic coefficients are critical for ensuring the model delivers accurate simulations.

Validation: The performance of the calibrated genetic coefficients is tested using an independent dataset from the same locations. Model evaluation metrics such as the index of agreement, normalized root mean square error, mean bias error, etc., are used to compare the observed versus simulated phenology, yield, etc. If the metrics are within acceptable limits, we can say calibration is robust.

Sensitivity analysis: This step is employed to evaluate the model's sensitivity to different input parameters. If a slight adjustment in the input parameters leads to a significant change in yield or other outputs, the model is considered highly sensitive to that parameter.

Climate change impact assessment: The effectiveness of the crop model is assessed by comparing its performance under future climate conditions against the current climate (baseline). This process utilizes daily climate projections from general circulation models (GCMs). Prior to their integration into crop models, these GCM data undergo bias correction to adjust for systematic errors and enhance their alignment with observed data. The underlying assumption is that soil parameters and crop management practices remain

unchanged under future climate scenarios. Subsequently, the yield and other simulated outputs for the future are compared with those from the baseline period to measure the changes.

Identification of adaptation options: If the simulated yield in the future period is less than that of the baseline, we can check the effectiveness of adaptation options to reduce the magnitude of yield reduction through crop models. The following adaptation options can be tested in the existing crop models:

- Adjusting the sowing time: early or late sowing
- Fertilizer application: changing dosage and number of splits
- Irrigation: changing timing, number of applications, quantity
- Combinations of fertilizer + irrigation options
- Altered genetic traits

Current approaches to model single stressors in crops using crop simulation models

One of the stressors in crop simulation models (CSMs) that has been studied the most is water stress or drought. Water stress is incorporated into models like as APSIM, DSSAT, EPIC, and CropSyst through functions that decrease processes including photosynthesis, root growth, leaf expansion, and assimilate partitioning to reproductive organs. These models use soil water dynamics, evapotranspiration, and precipitation to calculate the daily water balance. When soil moisture drops below predetermined thresholds, stress coefficients are applied to inhibit physiological processes. To ensure that the simulated reactions closely resemble observed plant behaviour under drought conditions, APSIM and DSSAT, for instance, use multiplicative stress functions to modify plant development processes based on water availability. The effects of drought on crops like maize, wheat, and legumes under various environmental circumstances have been successfully simulated using these models.

The main focus of heat stress modelling in CSMs is how temperature affects phenological development and reproductive functions. Since the majority of models, including DSSAT,

APSIM, STICS, and CropSyst, model crop development rates as functions of temperature, higher temperatures may shorten the growth cycle and lower yield by accelerating phenology. By altering processes including pollen viability, grain set, and kernel growth, heat stress surrounding blooming and grain filling is replicated. In certain models, yield components are directly penalised by high-temperature thresholds. For instance, reductions in grain size or quantity are implemented if temperatures rise above a key threshold during flowering. While most models use temperature-based stress functions, some advanced modeling systems integrate vapor pressure deficit (VPD) to capture the combined effects of heat and atmospheric dryness on plant water status and yield.

Salinity stress has also been incorporated into some CSMs, although its representation varies across models. AquaCrop is widely recognized for its detailed treatment of salinity stress. It simulates both osmotic effects and ion toxicity by reducing canopy development and crop transpiration based on salt concentration in the root zone. The model computes salinity build-up over time, affecting water uptake and crop growth. On the other hand, DSSAT uses simpler salinity routines, often applying a reduction factor directly to evapotranspiration or photosynthesis without dynamically modelling in-season salt accumulation. Similarly, models like CropSyst and EPIC use empirical functions that relate soil salinity levels to reductions in transpiration and yield. Salinity modelling remains challenging, especially in regions with fluctuating irrigation water quality and groundwater salinity, but CSMs like AquaCrop are increasingly applied for these environments due to their relatively simpler calibration and predictive capacity.

Another environmental constraints that is modelled in a number of CSMs is waterlogging or excessive water stress. The detrimental effects of soil saturation on root aeration and plant function are simulated by models like APSIM, DSSAT, WOFOST, and AquaCrop. Waterlogging mostly lowers root respiration and hinders nutrient uptake, which lowers biomass accumulation and photosynthesis. When the soil water content surpasses field capacity for prolonged periods, models replicate these effects by decreasing root growth and applying stress factors to transpiration and photosynthesis. For example, the decrease in oxygen availability in the root zone, which causes stomatal closure and slower growth, is specifically modelled by APSIM. However, because of the intricate relationships

between soil hydrology, oxygen diffusion, and plant responses, modelling waterlogging is still comparatively less advanced than modelling drought and salinity.

Crop modelling for optimizing genetic traits as an adaptation option

Crop models provide a robust and logical method to enhance and expedite the creation of new crop varieties. They can be utilized to develop crop ideotypes, which are conceptualized plants designed to produce higher yields or improved quality of grain, oil, or other valuable products when cultivated as varieties (Donald, 1968). In crop models, an ideotype is defined by a specific set of cultivar parameters, which can be refined by optimizing yield across diverse environmental conditions. Examples of how crop models are applied to design crop ideotypes are outlined in Table 1.

Table 1 Applications of crop models in designing crop ideotypes

Crop	**Region/ Country**	**Model used**	**Purpose**	**Reference**
Groundnut	India	CROPGRO	To evaluate genetic traits for improving productivity	Singh et al. (2012)
Wheat	Europe (2 sites)	Sirius	Optimized high-yielding wheat ideotypes for 2050	Semenov and Stratonovitch (2013)
Cotton	Sub-Saharan Africa	CSM-CROPGRO-Cotton	To identify best cultivar for cotton adapted to future climate	Loison et al. (2017)
Barley	Europe	Used 8 models	Designed climate-resilient barley ideotypes	Tao et al. (2017)
Wheat	Australia	APSIM	Determined the traits of rainfed wheat ideotypes to suit future climate conditions.	Wang et al. (2019)
Maize	North China Plain	APSIM	To design high-yielding maize ideotypes to adapt to a changing climate	Xiao et al. (2020)

Summary

Crop models are process-driven and aim to virtually mirror field conditions, serving as an effective tool for analyzing G x E x M interactions (genotype, environment, and management). They are particularly valuable for assessing climate change impacts and identifying optimal adaptation strategies to address abiotic stresses. Recent advancements suggest using crop modeling to design ideotypes tailored to future climates. Integrating traditional crop simulation with modern breeding techniques and genetic modeling offers significant potential to hasten the development of future cereal varieties suited to diverse environments, capable of tolerating multiple abiotic stresses and delivering higher yields. The reliability of model-assisted ideotype design can be further strengthened by refining simulation models to more accurately reflect the impact of extreme conditions. However, relying solely on a single crop or climate model is unwise due to inherent uncertainties. Employing a combination of multiple crop and climate models can provide a more reliable prediction of crop performance under future climates, along with an estimate of uncertainty.

References

Donald, C.T., 1968. The breeding of crop ideotypes. Euphytica, 17, pp.385-403.

Hoogenboom, G., C.H. Porter, K.J. Boote, V. Shelia, P.W. Wilkens, U. Singh, J.W. White, S. Asseng, J.I. Lizaso, L.P. Moreno, W. Pavan, R. Ogoshi, L.A. Hunt, G.Y. Tsuji, and J.W. Jones. 2019. The DSSAT crop modeling ecosystem. In: p.173-216 [K.J. Boote, editor] Advances in Crop Modeling for a Sustainable Agriculture. Burleigh Dodds Science Publishing, Cambridge, United Kingdom (https://dx.doi.org/10.19103/AS.2019.0061.10). Jones, J.W., G. Hoogenboom, C.H. Porter, K.J. Boote, W.D. Batchelor, L.A. Hunt, P.W. Wilkens, U. Singh, A.J. Gijsman, and J.T. Ritchie. 2003. The DSSAT cropping system model. European Journal of Agronomy 18:235-265 (https://doi.org/10.1016/S1161-0301(02)00107-7).

Loison, R., Audebert, A., Debaeke, P., Hoogenboom, G., Leroux, L., Oumarou, P., Gérardeaux, E., 2017. Designing cotton ideotypes for the future: reducing risk of

crop failure for low input rainfed conditions in Northern Cameroon. Eur. J. Agron. 90, 162–173

Semenov, M.A., Stratonovitch, P., 2013. Designing high-yielding wheat ideotypes for a changing climate. Food and Energy Security 2, 185–196.

Singh, P., Boote, K.J., Kumar, U., Srinivas, K., Nigam, S.N., Jones, J.W., 2012. Evaluation of genetic traits for improving productivity and adaptation of groundnut to climate change in India. J. Agron. Crop Sci. 198, 399–413

Tao, F., Rötter, R.P., Palosuo, T., Gregorio Hernández Díaz-Ambrona, C., Mínguez, M.I., Semenov, M.A., Kersebaum, K.C., Nendel, C., Cammarano, D., Hoffmann, H., Ewert, F., Dambreville, A., Martre, P., Rodríguez, L., Ruiz-Ramos, M., Gaiser, T., Höhn, J.G., Salo, T., Ferrise, R., Bindi, M., Schulman, A.H., 2017. Designing future barley ideotypes using a crop model ensemble. Eur. J. Agron. 82, 144–162.

Wang, B., Feng, P.Y., Chen, C., Liu, D.L., Waters, C., Yu, Q., 2019a. Designing wheat ideotypes to cope with future changing climate in South-Eastern Australia. Agric. Syst. 170, 9–18.

Xiao, D., Li Liu, D., Wang, B., Feng, P. and Waters, C., 2020. Designing high-yielding maize ideotypes to adapt changing climate in the North China Plain. Agricultural Systems, 181, p.102805.

CMIP6 Models for Agriculture and Water Resources

Rahul Patil
Department of Soil and Water Conservation Engineering
College of Agriculture Engineering, UAS, Raichur
Email: rahul1235110@gmail.com

Introduction

The Coupled Model Inter-comparison Project Phase 6 (CMIP6) represents the latest and most ambitious international collaborative effort in climate science. It brings together leading climate modeling centers from around the globe to conduct standardized experiments and generate a vast, publicly accessible archive of climate model outputs. These outputs are not merely academic exercises; they are indispensable tools for advancing our understanding of the Earth's climate system, projecting future climate conditions, and critically, for informing societal responses to the escalating challenges of climate change. In an era defined by intensifying climate impacts, the data and insights derived from CMIP6 are particularly crucial for understanding and addressing the vulnerabilities and opportunities in two fundamental sectors: **hydrology** and **agriculture**. These sectors are intrinsically linked to climate, and their sustainability directly impacts human well-being, food security, and economic stability worldwide.

This document aims to provide a clear and concise overview of CMIP6. It will explain the project's intricate structure, detail the types of data it produces, and, most importantly, illustrate its practical applications. By highlighting CMIP6's profound relevance for developing sustainable farming practices and ensuring robust water resource management, this document underscores its role as a cornerstone for climate change adaptation and resilience building in a rapidly changing world.

1. CMIP6 in Hydrology and Agriculture: An Overview

Climate change significantly affects food security and water availability, altering **precipitation patterns**, raising **temperatures**, and increasing **extreme weather events**. Robust climate projections are essential for informed adaptation.CMIP6, the latest and most comprehensive climate modeling project, provides critical data for:

- **Hydrology:** Insights into future changes in **precipitation**, **evapotranspiration**, **soil moisture**, **runoff**, and **temperature trends**, all vital for water availability, flood/drought risks, and reservoir management.
- **Agriculture:** Projections for **crop yield changes**, **climate-induced risks** to farming, and optimizing **agricultural practices** like irrigation and pest management under future climates.

CMIP6 is a foundational tool for climate change impact assessment, vulnerability analysis, and developing adaptation and mitigation strategies in both sectors.

2. CMIP6 Structure and Key Components

CMIP6 is organized into three main tiers:

- **DECK (Diagnostic, Evaluation, and Characterization of Klima):** Foundational experiments to evaluate model performance and understand basic climate system characteristics.
- **Historical Simulations:** Replicate observed climate from 1850 to recent decades (e.g., 2014) using historical forcings (GHGs, aerosols, etc.). These validate models and provide initial conditions for future projections.
- **ScenarioMIP (Shared Socioeconomic Pathways):** Projects future climate under various **socioeconomic development pathways** (SSPs) and their associated greenhouse gas emissions.

Prominent Global Climate Models (GCMs) in CMIP6 come from various centers, including:

- **GFDL-ESM4 (USA):** Known for coupled carbon-cycle and climate dynamics.
- **MPI-ESM1.2-HR (Germany):** Features detailed land-atmosphere interactions.
- **MRI-ESM2-0 (Japan):** Focuses on atmospheric and oceanic feedbacks.
- **UKESM1-0-LL (UK):** A fully coupled Earth system model with biogeochemistry.

These models generate essential outputs like **precipitation (pr)**, **near-surface air temperature (tas)** and **surface latent heat flux (evspsbl)** typically at daily or monthly temporal resolution.

3. Understanding Shared Socioeconomic Pathways (SSPs) and Warming Projections

SSPs are narratives describing plausible future socioeconomic developments that drive greenhouse gas emissions and land-use changes. CMIP6 uses core SSPs, each with distinct global warming projections:

- **SSP1-1.9: "Very Low GHG Emissions – Sustainable Development"**
  - **Narrative:** Focus on sustainability, reduced inequality, and strong environmental protection; CO2 net zero around 2050.
  - **Estimated Warming (2041–2060):** 1.6°C
  - **Estimated Warming (2081–2100):** 1.4°C (Very Likely Range: 1.0–1.8°C)
  - **Implications:** Most optimistic. Lower frequency of extreme heat, more manageable changes in precipitation, and fewer severe agricultural droughts, favoring adaptation.
- **SSP1-2.6: "Low GHG Emissions – Sustainable Development"**
  - **Narrative:** Similar to SSP1-1.9 but slightly less aggressive mitigation; CO2 net zero around 2075.
  - **Estimated Warming (2041–2060):** 1.7°C
  - **Estimated Warming (2081–2100):** 1.8°C (Very Likely Range: 1.3–2.4°C)
  - **Implications:** Relatively contained warming, allowing for effective climate adaptation in agriculture and water management.
- **SSP2-4.5: "Intermediate GHG Emissions – Middle of the Road"**
  - **Narrative:** Trends follow historical patterns with moderate growth and fragmented progress; CO2 emissions around current levels until 2050, then falling but not net zero by 2100.
  - **Estimated Warming (2041–2060):** 2.0°C
  - **Estimated Warming (2081–2100):** 2.7°C (Very Likely Range: 2.1–3.5°C)
  - **Implications:** Moderate warming with more pronounced impacts: longer heatwaves, increased precipitation variability (dry spells and intense downpours), and growing water stress. Adaptation becomes critical and challenging.
- **SSP3-7.0: "High GHG Emissions – Regional Rivalry – A Rocky Road"**
  - **Narrative:** Fragmented world with nationalism, conflicts, slow tech progress, high population growth; CO2 emissions double by 2100.
  - **Estimated Warming (2041–2060):** 2.1°C

- **Estimated Warming (2081–2100):** 3.6°C (Very Likely Range: 2.8–4.6°C)
- **Implications:** Severe challenges with substantial temperature increases, erratic precipitation, and major water scarcity. Adaptation is extremely difficult due to the magnitude of change and limited cooperation.

- **SSP5-8.5: "Very High GHG Emissions – Fossil-Fueled Development – Taking the Highway"**
- **Narrative:** Rapid, fossil-fueled economic growth with high energy demands and limited climate mitigation; CO2 emissions triple by 2075.
- **Estimated Warming (2041–2060):** 2.4°C
- **Estimated Warming (2081–2100):** 4.4°C (Very Likely Range: 3.3–5.7°C)
- **Implications:** Worst-case scenario with unprecedented heat, potentially unviable agricultural practices, and immense pressure on water resources, leading to chronic stress and severe losses. Requires enormous, potentially disruptive, transformational adaptation.

Using diverse SSPs allows CMIP6 to simulate a wide range of futures, enabling comprehensive stress-testing of hydrological and agricultural systems for robust planning.

4. Key CMIP6 Variables for Hydrology and Agriculture

CMIP6 provides a rich set of variables critical for water resources and farming:

- **pr (Precipitation rate):** Essential for water balance, runoff, groundwater recharge, and flood/drought risk. (e.g., kg/m²/s, convertible to mm/day)
- **tas (Near-surface air temperature):** Drives evapotranspiration, influences crop growth and phenology. (Kelvin, convertible to Celsius)
- **hurs (Relative humidity):** Affects evapotranspiration rates and crop health. (%)
- **rsds (Downward shortwave radiation):** Crucial for photosynthesis and surface energy balance. (W/m²)

These variables, often daily or monthly, form the basis for climate change impact assessments.

5. Accessing CMIP6 Data

Several platforms facilitate access to CMIP6 data, crucial for applying its outputs:

- **Earth System Grid Federation (ESGF):** The primary global network for raw CMIP6 data. Offers comprehensive access but requires specialized tools for processing.

- **Copernicus Climate Data Store (CDS):** Provides user-friendly access, often with pre-processed (regridded, bias-corrected) CMIP6 data, especially useful for regional studies.
- **NASA NEX-GDDP:** Offers statistically downscaled and bias-corrected CMIP6 data at 0.25-degree global resolution, reducing user processing needs for regional studies.
- **Google Earth Engine (GEE):** A cloud-based platform for large-scale analysis, processing, and visualization of selected CMIP6 variables, eliminating local storage and download needs.

Choosing a source depends on user expertise, required resolution, and processing needs.

6. Downscaling CMIP6 Data for Local Use

GCMs' coarse resolution isn't suitable for local hydrological or agricultural applications. **Statistical downscaling** bridges this gap by transforming coarse GCM outputs into finer-resolution climate information using **local observational data**.

6.1 Statistical Downscaling Techniques

These methods establish statistical relationships between large-scale GCM outputs and **local observed climate variables** from a historical period. These relationships are then applied to future GCM projections.

- **Regression Models:** Simple models (linear/non-linear) relating GCM predictors to local variables (e.g., predicting local rainfall from GCM atmospheric pressure).
- **Weather Generators:** Statistically characterize observed local weather patterns and perturb them based on GCM changes to simulate future local sequences.
- **Analog Methods:** Find historical large-scale patterns similar to future GCM patterns in **local observational datasets**, using associated historical local weather as downscaled output.
- **Machine Learning (ML) Algorithms:** Advanced techniques (ANNs, SVMs, Random Forests) learn complex, non-linear relationships between GCMs and **local observations**.

Pros: Computationally efficient, cost-effective, and leverage local historical observations.

Cons: Assume stationarity (historical relationships remain valid in future), don't explicitly resolve physical processes, and require high-quality local observations.

6.2 Empirical Statistical Downscaling (ESD)

A broad category that combines statistical downscaling with bias correction, ensuring downscaled data aligns with **observed local climatology** while reflecting GCM-projected changes.

6.3 Bias Correction and Statistical Downscaling (BCSD)

A common two-step approach using **local observational data**:

1. **Bias Correction:** Adjusts systematic GCM biases to match **observed local climatology** at the GCM grid scale (e.g., correcting mean precipitation or temperature).
2. **Spatial Downscaling:** Distributes the bias-corrected GCM data to a finer resolution by applying **observed local historical spatial patterns** (e.g., distributing monthly GCM precipitation daily using historical local daily patterns).

BCSD is ideal for driving hydrological models like SWAT, providing essential daily, gridded, bias-corrected inputs leveraging **local observational data**.

7. Bias Correction: Adjusting CMIP6 Data for Hydrologic Relevance

GCMs have systematic errors or "biases" compared to **observed local climate data**. **Bias correction** adjusts GCM outputs to match the statistical characteristics (**mean, variance, extremes**) of **observed local climatology**, making projections more reliable for local applications.

7.1 Common Techniques (Utilizing Local Data)

- **Linear Scaling:** Simple adjustment of GCM mean (and sometimes variance) to match **observed local mean**. Effective for average biases but less so for extremes.
- **Delta Change (Perturbation Method):** Applies the *projected change* from the GCM to the **observed local historical data**. Preserves local variability but doesn't correct absolute GCM biases.
- **Quantile Mapping (QM):** Aligns the Cumulative Distribution Function (CDF) of GCM data with the CDF of **observed local data** over a historical period, correcting biases across the entire distribution, including extremes. More robust but requires sufficient **local observed data**.
- **Empirical Quantile Mapping (EQM):** A variant of QM using empirical (observed) quantiles directly, without fitting specific distributions. Robust but needs large **local observed datasets**.

- **LOCI (Local Intensity Scaling):** For precipitation, corrects both occurrence and intensity biases by comparing GCM outputs to **local precipitation observations**.

7.2 Bias Correction Workflow (Emphasizing Local Data)

1. **Gather Data:** Obtain high-quality, long-term **local observational climate data** and corresponding historical CMIP6 GCM simulations.
2. **Define Overlapping Period:** Establish a common "calibration" period where both datasets are available.
3. **Analyze Biases:** Compare raw GCM data with **local observations** to identify systematic errors.
4. **Apply Correction:** Calculate and apply correction parameters (from chosen method) to historical and future GCM data.
5. **Validate:** Crucially, validate corrected data against an independent **local observational dataset** to ensure effectiveness and generalization.

7.3 Relevance to Hydrology

Bias correction is **essential** for reliable hydrological simulations:

- **Reliable Forecasts:** Uncorrected biases lead to significant errors in simulated runoff, impacting flood and drought forecasts. Correction aligns GCM inputs with **local observed characteristics**, yielding more accurate hydrological outputs.
- **Extreme Event Correction:** GCMs often misrepresent local extreme events. Bias correction (especially QM) improves the representation of floods and droughts by aligning distribution tails with **local observations**.
- **Improved Agricultural Planning:** Accurate daily precipitation and temperature from bias-corrected data (calibrated with **local station data**) are vital for precise crop water requirements, irrigation scheduling, and accurate local drought indices.

Bias correction transforms raw CMIP6 outputs into locally relevant, hydrologically consistent data using **local observational data**, making projections actionable for planning.

8. Integrating CMIP6 Data into Hydrologic and Agricultural Models

Processed CMIP6 data is fed into specialized impact models to simulate how climate change affects water resources and food production at regional and local scales.

8.1 Hydrologic Models

Simulate water movement through watersheds.

- **SWAT (Soil and Water Assessment Tool):** Physically-based model for water balance, sediment, and nutrient transport. Requires daily precipitation, max/min temperature, solar radiation, humidity, wind. Outputs: streamflow, ET, soil moisture.
- **VIC (Variable Infiltration Capacity):** Grid-based model for water and energy balances. Requires gridded daily/sub-daily precipitation, max/min temperature, wind. Outputs: ET, runoff, soil moisture.
- **HEC-HMS (Hydrologic Engineering Center – Hydrologic Modeling System):** Rainfall-runoff model for flood forecasting and water planning. Requires event-based or continuous precipitation. Outputs: hydrographs, peak flows.
- **WEAP (Water Evaluation and Planning):** Integrated water resource management tool. Requires monthly/daily climate data, water demands, infrastructure details. Outputs: water allocation, unmet demands.

8.2 Agricultural Models

Simulate crop growth and yield under environmental conditions and management.

- **DSSAT (Decision Support System for Agrotechnology Transfer):** Suite of crop models linking climate, soil, genotype, and management to predict growth and yield. Requires daily precipitation, max/min temperature, solar radiation. Outputs: crop yield, biomass, phenology.
- **AquaCrop:** FAO water-driven model focused on crop water productivity. Requires daily precipitation, max/min temperature, reference ET. Outputs: biomass, yield, soil water content.
- **CropSyst:** Cropping systems model simulating daily crop growth, water balance, and nitrogen dynamics. Requires daily weather, soil, and management data. Outputs: yield, water use efficiency.

8.3 General Inputs and Outputs

Most models need daily **precipitation (pr)**, **minimum/maximum temperature (tasmin/tasmax)**, **solar radiation (rsds)**, **relative humidity (hurs)**, and **wind speed (sfcwind)**. These bias-corrected and downscaled variables drive the models.

Outputs include:

- **Runoff and discharge trends:** Changes in river flow.

- **Crop yield and phenology:** Predicted crop performance and growth shifts.
- **Irrigation demand forecasts:** Future water needs for agriculture.
- **Adaptation option assessments:** Evaluation of strategies to mitigate impacts.

This integrated approach is fundamental for resilient water and agriculture planning.

9. Practical Applications in Agricultural and Water Resource Planning

Outputs from CMIP6-driven models provide critical information for proactive climate adaptation.

9.1 Agriculture

- **Assess Yield Risks:** Quantify potential yield changes for crops under different SSPs, identifying vulnerable areas and prioritizing research for resilient varieties.
- **Evaluate Planting/Harvesting Shifts:** Identify altered optimal farming calendars due to changing climate, helping farmers adjust schedules to maximize yields.
- **Estimate Water Stress:** Quantify days of crop water stress under future climates, crucial for understanding drought impacts and designing efficient irrigation.
- **Forecast Irrigation Demand:** Project future irrigation needs based on changing evapotranspiration and precipitation, aiding water allocation and infrastructure planning.

9.2 Hydrology

- **Drought Risk Management:** Calculate SPI and SPEI using CMIP6 data to project future drought frequencies, durations, and severities, enhancing monitoring and early warning.
- **Flood Prediction:** Integrate CMIP6 precipitation extremes into hydrological models to project future flood magnitudes and frequencies, informing flood protection and preparedness.
- **Groundwater Recharge Identification:** Use projected soil moisture and runoff changes to identify future groundwater recharge zones, essential for sustainable aquifer management.

9.3 Water Management and Policy

- **Climate-Resilient Reservoir Planning:** Re-evaluate reservoir capacities and operating rules using CMIP6 streamflow and precipitation variability projections to manage water supply and flood risks effectively.
- **Scenario-Based Water Allocation:** Develop water allocation plans robust across various SSPs, stress-testing existing schemes and proposing adaptive measures.

- **Input for Climate Action Plans:** Downscaled and bias-corrected CMIP6 data provides scientific basis for national and regional climate action plans, informing localized adaptation strategies.

In essence, CMIP6 models, when processed with **local observational data** for downscaling and bias correction, empower data-driven decisions. This leads to more effective planning and policies, enhancing resilience in agriculture and water resources, and contributing to global food and water security.

References

Bao, Q., Ding, J., Wang, J., Han, L., & Tan, J. (2025). Utilizing CMIP6-SSP scenarios with the VIC model to enhance agricultural and ecological water consumption predictions and deficit assessments in arid regions. *Computers and Electronics in Agriculture*, *232*, 110083.

Kumari, P., Jaiswal, R. K., & Singh, H. P. (2024). Assessing climate change impacts on irrigation water requirements in the Lower Mahanadi Basin: A CMIP6-based spatiotemporal analysis and future projections. *Journal of Water and Climate Change*, *15*(7), 3328-3345.

Karan, K., Singh, D., Singh, P. K., Bharati, B., Singh, T. P., & Berndtsson, R. (2022). Implications of future climate change on crop and irrigation water requirements in a semi-arid river basin using CMIP6 GCMs. *Journal of Arid Land*, *14*(11), 1234-1257.

Kumar, N., Singh, V. G., Singh, S. K., Behera, D. K., & Gašparović, M. (2023). Modeling of land use change under the recent climate projections of CMIP6: a case study of Indian river basin. *Environmental science and pollution research*, *30*(49), 107219-107235.

Crop Modelling for sustainable agriculture in Context of Climate Change

Nishant K Sinha and M. Mohanty
ICAR-Indian Institute of Soil Science, Bhopal, 462038
Email: nishant.sinha76211@gmail.com

1. Introduction

Traditionally, agronomic research emphasized maximizing crop yields. However, growing environmental concerns have shifted the focus toward sustainable practices that balance productivity with ecological health. This shift necessitates strategies that preserve soil quality, reduce environmental degradation, and enhance crop profitability.

Achieving sustainable crop production requires understanding the complex interactions among soil, water, atmosphere, and crops. With advancements in computing, these systems can now be integrated and modeled to simulate and predict crop growth and yield more accurately. Crop models use mathematical equations to represent biological systems and, once validated, help forecast the impacts of environmental changes and management practices on crop performance.

Crop growth models serve three main purposes:

1. **Analyzing** crop responses to improve research focus,
2. **Simulating** plant growth for education,
3. **Forecasting** outcomes for management and decision-making (Rimmington & Charles-Edwards, 1987).

These models require interdisciplinary expertise across agronomy, soil science, crop physiology, and agrometeorology. They are also valuable in participatory research, enabling collaboration between farmers, researchers, and advisors.

Decision Support Systems (DSS), such as **DSSAT** and **APSIM**, integrate crop models with economic and environmental data to guide farm-level decisions. DSSAT helps match crop needs with land conditions, while APSIM considers a broader range of growth factors.

This chapter covers the classification of crop models (simple and complex), steps in model development, essential datasets for calibration and validation, and the relevance of CGSMs in the context of climate change. Key model concepts and a conceptual crop model example are introduced in the first section (see Figure 1).

1.1. Terminology

1.1.1 System

A system is a collection of interacting components grouped to study a particular part of the real world. The components included depend on the study's objectives and simplify

complex realities. Systems have defined spatial and temporal boundaries and are influenced by external factors. For example, in crop models, the crop and root zone are considered as system components influenced by weather and management, while other models may focus on smaller units like leaves or cells.

1.1.2 Environment and System Boundary

The environment includes all elements outside the system's components. The system boundary separates the system from its environment, often based on cause-and-effect relationships. While the environment can influence the system (e.g., temperature affecting crop growth), the system typically does not alter the environment. For instance, crops are influenced by sunlight and temperature but have minimal impact on these factors.

1.1.3 Simulation

Simulation is the process of running a model to mimic real-world system behavior. It involves designing logic, writing code, and executing it on a computer to generate results.

1.1.4 Inputs and Outputs

Inputs are external factors affecting the system but are not affected by it, such as rainfall, temperature, and light in crop systems. Outputs are measurable responses of the system, like crop biomass or soil moisture, which are of interest in modeling.

1.1.5 Parameters and Constants

Constants are fixed values that do not change across experiments (e.g., gravitational constant, molecular weight). Parameters are model-specific values that remain unchanged during simulation but may vary across systems or studies (e.g., photosynthetic response to light, degree-days to flowering).

1.1.6 State Variables

State variables describe the condition of system components at a specific time and change dynamically as the system evolves. In crop models, common state variables include soil moisture and plant biomass.

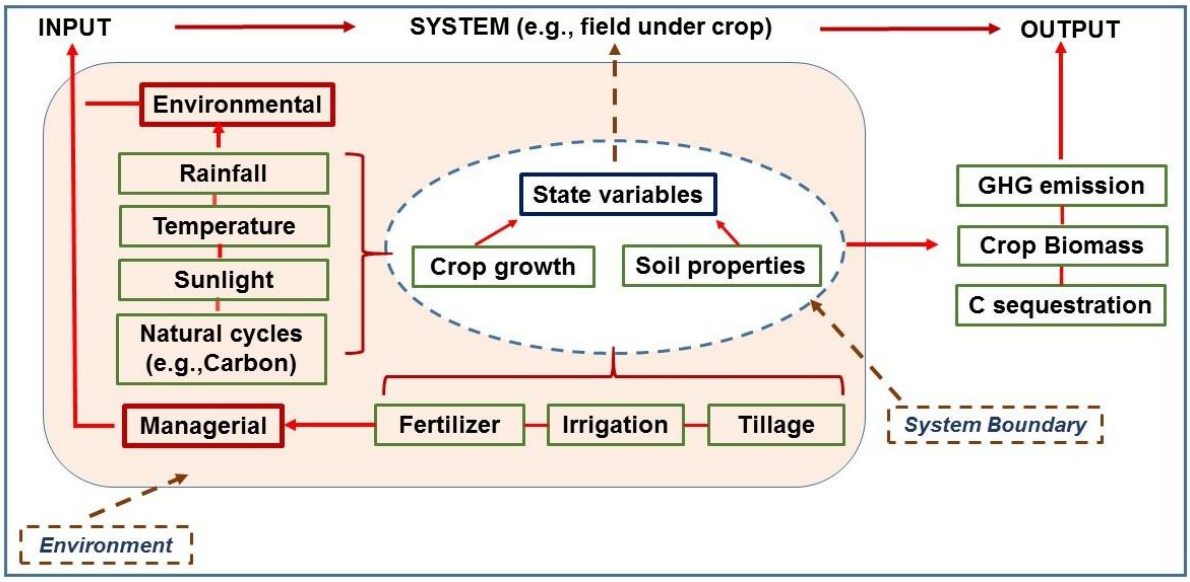


Figure 1. Diagrammatic representation of a model.

2. Types of models

Computer models are mathematical representations of real-world systems. In agriculture, simulation models are used to mimic natural processes—like crop growth or water using computer programs based on mathematical, statistical, or empirical relationships. These models help estimate crop yields based on factors such as weather, soil, and management practices. They typically use differential equations to calculate changes in key variables (e.g., growth rate, biomass) from planting to harvest.

Crop Model Types:

Descriptive vs. Explanatory:
Descriptive models summarize observed data without explaining underlying mechanisms (e.g., estimating crop weight from past measurements). Explanatory models, on the other hand, describe how processes like leaf growth or tillering drive crop development using quantified system interactions.

Deterministic vs. Stochastic:
Deterministic models produce the same output for a given input, with no randomness. Stochastic models introduce variability, giving a range of possible outcomes with associated probabilities.

Static vs. Dynamic:

Static models don't consider time and assume constant variables. Dynamic models incorporate time, simulating how variables change over days or seasons to reflect real-time crop behavior.

Mechanistic vs. Empirical:
Mechanistic models are process-based, built on scientific understanding of system functions (e.g., how light and nutrients affect growth). Empirical models rely on observed data patterns and statistical techniques like regression or correlation.

Steps in Model Development:

Model development is iterative and involves three main activities: conceptualization (theoretical design), coding (building the model), and testing (experimentation). These steps often overlap and require adjustments based on feedback and performance.

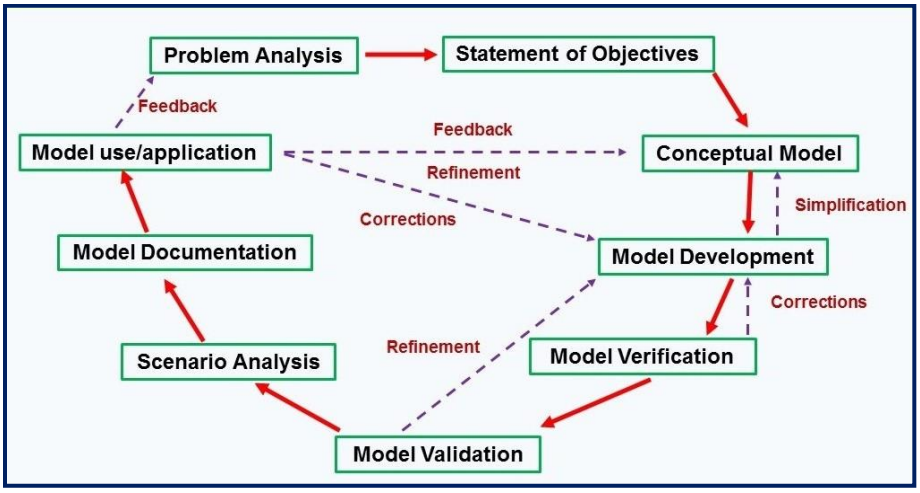


Figure 2. The modelling cycle

Steps in Model Development

3.1 Problem Analysis – Defining the Purpose (Why?)

The first step is to clearly state the purpose of the model. In agriculture, models may be used for understanding crop growth dynamics, forecasting yield, or supporting decisions related to climate change, variety evaluation, and crop planning. Key considerations include:

- The hypothesis being tested
- Time frame (e.g., forecasting, long-term projections)
- Required accuracy
- Target users (e.g., researchers, planners)

3.2 Conceptual Model – Defining the System (What?)

3.2.1 Modeling Framework

This involves identifying:

- System components (e.g., crop, soil, weather)
- System boundaries
- Driving variables (e.g., temperature, rainfall)
- Desired outputs
- Temporal and spatial scale
- Available resources
- Flexibility and scope of the model

3.2.2 Conceptualization

The system is conceptualized using diagrams and prior knowledge. Key variables are chosen based on their importance to the outputs, and less significant ones may be simplified or removed to keep the model practical.

3.2 Model Development – Translating to Equations (How?)

Here, model features are selected according to the type (e.g., deterministic, stochastic) and level of detail needed. Model structure includes relationships among components, choice of spatial units, and process equations. The model may integrate various modules (e.g., climate, hydrology, irrigation). Calibration involves adjusting model parameters and structure to best reflect real-world behavior.

3.4 Verification, Calibration, and Sensitivity Analysis

- **Verification** checks whether the model is correctly implemented as per its conceptual design.
- **Calibration** fine-tunes parameters so outputs match observed data.
- **Sensitivity analysis** examines how changes in input variables affect outputs, helping to identify which inputs most influence the model.

3.5 Model validation and scenario analysis

In contrast to model verification, model validation seeks to answer the question "*has the right model been constructed*?" Answer to this question is usually obtained by comparing model results with data from real system (field experiments). Thus, validation is the process of testing the accuracy of a model with respect to the system being modelled. The data used for validation should not have been used during model development, and should normally contain data from multi-location (e.g., from different agro-ecological zones with a wide range of conditions), multi-year experiments. Sensitivity analysis helps determine what data is required and level of detail in data required, *i.e.*, what is important to measure and how often it must be measured. Usually, one hopes to come within 95% confidence interval during validation. However, this is often difficult to accomplish with biological systems. Validity is never absolute, therefore cases for which the model is valid need to be stated. Once the model is successfully validated, one can proceed to conduct scenario analysis, to answer '*what if*?' questions, for example, what if temperature or CO_2 concentration rises?

Such studies provide insight into the effect of various processes and parameters on the model output, and can help to explore divergent management options. This process of 'deriving insight into reality through analyses of model output for different conditions' is also simply called 'simulation' (Refsgaard and Henriksen 2004).

3.6 Model documentation and maintenance

Documentation of a model is important for its utility, and usually contains the following components, which parallel the steps of the modeling process:

3.6.1. Analysis of the problem: The purpose for which the model was developed and the underlying objectives must be clearly outlined.

3.6.2. Model design and development: The amount of detail with which a model is explained depends on the situation. In a comprehensive technical report, the source code for the programmes is often included. The simplifying assumptions and the rationale for employing them are also explained. Clearly labeled diagrams of the relationships among variables and submodels are usually very helpful in understanding the model.

3.6.3. Model verification and validation: The detailed procedure including the type of data required for running the model and calibrating the model parameters are to be laid down. Once verified, the procedure for validating the model with examples may be provided. For a written report, appendices may contain more detail, such as source code of programs and additional information about the solutions of equations.

3.6.4. Model utility (and revision): The model report should include results, interpretations, implications, recommendations, and conclusions in the final part. Suggestions for future work may also be mentioned. Once the model is used by end users who provide valuable feedback, it may be necessary to make corrections, improvements, or enhancements. In this case, it becomes imperative for the model developer to go through the modeling process to develop a revised version.

4. Crop growth simulation models

Crop Growth Simulation Models are mathematical tools used to simulate the interaction between crops and their environment. Originating in the 1960s, these models integrate knowledge of crop growth processes to better understand and predict crop performance under varying environmental and management conditions.

Unlike empirical models, CGSMs operate at finer time scales (daily or even hourly), simulating processes such as phenological development, canopy formation, photosynthesis, assimilate partitioning, and soil–plant–atmosphere interactions involving water and nutrient dynamics. These models enable the exploration of crop behavior in ways not feasible through field experiments alone. CGSMs serve two main purposes: (1) advancing understanding of crop growth dynamics, and (2) supporting decision-making in agriculture. Despite their value, all models simplify reality and contain limitations due to

the complexity of biological systems and data requirements. Most models blend empirical and mechanistic approaches, depending on the level of detail needed for a specific purpose. Well-known CGSMs include **DSSAT** and **APSIM**, widely used in research and agricultural planning. There are many other models, but these two are more popular. Some crop models reported in recent literature are listed below:

Table 1. Description of different models used for crop simulation study.

Model	Details	References
SLAM II	Forage harvesting operation	Buck et al., 1988
SPICE	Whole plant water flow	Cruiziat and Thomas,
REALSOY	Soybean	Meyer and Curry, 1986
MODVEX	Model development and validation system	McKinion, 1986
IRRIGATE	Irrigation scheduling model	Tscheschke et al., 1978
COTTAM	Cotton	Jackson et al. 1990
APSIM	Modelling framework for a range of crops	McCown et al., 1996
GWM	General weed model in row crops	Wiles et al., 1996
MPTGro	*Acacia spp.and Leucaena Spp.*	Harrington and Fownes,
GOSSYM-	Cotton	McKinion et al., 1996
CropSyst	Wheat & other crops	Stockle et al., 1994
SIMCOM	Crop (CERES crop modules) & economics	Bootes et al., 1996

LUPINMOD	Lupin	Fernandez et al., 1996
TUBERPRO	Potato & disease	Nemecek et al., 1996
SIMPOTATO	Potato	Rosenzweig et al., 1996
WOFOST	Wheat & maize, Water and nutrient	Supit et al., 1997
WAVE	Water and agrochemicals	Vanclooster et al., 1994
SUCROS	Crop models	Spitters et al., 1988
ORYZA1	Rice, water	Kropff et al. 1994
SIMRIW	Rice, water	Horie, 1987
SIMCOY	Corn	Place and Brown, 1987
CERES-Rice	Rice, water	Alocilja and Richie, 1998
GRAZPLAN	Pasture, water, lamb	Moore et al., 1997
EPIC	Erosion Productivity Impact Calculator	Williams et al., 1984
CERES	Series of crop simulation models	Jones et al., 1984
DSSAT	Framework of crop simulation models including modules of CERES, CROPGRO and CROPSIM	Tsuji et al., 1994
QCANE	Sugarcane, potential conditions	Liu and Kingston, 1995
AUSCANE	Sugarcane, potential & water stress conds.,	Jones et al., 1989
CANEGRO	Sugarcane, potential & water stress conds	Inman-Bamber, 1995
APSIM-	Sugarcane, potential growth, water and nitrogen	Keating et al., 1999

5. Crop Production Levels for simulation

Simulation of crop growth can be done at four levels of crop production, which have been defined based on the major growth limiting factors viz. water and nutrients, as depicted in Table 2.

Table 2. Different levels of crop production system and factors affecting it.

Production level	Conditions of crop growth and development	Factors
Level 1 (Potential production)	This is an ideal situation in which the crop does not suffer from any shortage of water or nutrients and without pests and diseases in the entire growing season.	Crop characteristics, solar radiation and temperature
Level 2 (Water limited production)	The growth of a crop is limited by shortage of water for at least some part of the growing season. This situation frequently occurs in semi-arid regions and also in areas where the rainfall is inadequate and/or poorly distributed.	Crop characteristics, solar radiation, temperature and rainfall
Level 3 (Nitrogen-limited production)	Nitrogen shortage during the growing season limits the crop growth at this production level. Water shortage or adverse weather conditions occur in the remaining part of the season. This condition is very frequent in the agricultural systems all over the world.	Crop characteristics, solar radiation, temperature, rainfall and soil nitrogen
Level 4 (Phosphorus limited production)	Crop growth is restricted by low levels of phosphorus and other mineral nutrients in the soil during the growing season.	Crop characteristics, solar radiation, temperature, rainfall, soil nitrogen and phosphorus

6. Level of Simplicity in a CGSM

The optimal simplicity of a Crop Growth Simulation Model (CGSM) depends on balancing precision, data availability, and computational power. Complex models require detailed input data and smaller time steps, which may not be practical due to data limitations. For example, while detailed water-flow processes can be simulated using Richard's equation, limited rainfall data often necessitates simpler water-balance models. Most CGSMs use a daily time step and link plant water uptake to leaf area, climate, and soil moisture. While

more detailed models include processes like leaf water potential, they demand intensive computation and finer time steps. Simple models are easier to apply but may lack accuracy or be too site-specific. A practical approach is to simplify a comprehensive model for a specific application and calibrate it with field data.

6.1. Model Calibration

Calibration involves adjusting model parameters so that simulated results match observed data. It's essential, especially when adapting existing models to new environments. Effective calibration requires well-controlled experiments with detailed crop, soil, and climate data. Data from conventional agronomic studies are often insufficient. Accurate yield prediction depends on appropriate model structure, reliable parameters, and quality input data.

6.2. Model Validation

Validation tests how well a model represents real-world systems by comparing predictions with independent field data. It ensures the model's reliability under diverse conditions and identifies its limitations. Validation helps refine model structure and guides further research. Only after rigorous validation can a model serve as a practical decision-support tool in agriculture.

6.3. Minimum Dataset for Model Calibration and Validation

Crop simulation models require specific datasets to simulate growth accurately. The level of detail depends on the model's purpose and complexity. Essential data categories include:

6.3.1 Soil Data

Models need detailed soil profile data up to 1.5 m depth, including texture, bulk density, horizon thickness, nutrient levels (N, P, K), pH, organic matter, and hydraulic properties. Surface residue and soil-water balance parameters are also essential.

6.3.2 Management Data

Includes planting date and density, cultivar type, fertilizer use (amount, type, application method), irrigation details, tillage operations, and pest control measures. These reflect existing or simulated management practices.

6.3.3 Crop Data

Covers:

- **Photosynthesis and Respiration**: Leaf area, photosynthetic rate, carbohydrate requirements.
- **Biomass Partitioning**: Dry matter distribution across plant parts, growth rates, and yields.
- **Phenology**: Crop development stages (e.g., V1, R1), recorded at 2–3-day intervals.
- **Root Development**: Root depth, growth rate, and stress sensitivity.

6.3.4 Weather Data

Key inputs include daily solar radiation, max/min temperatures, humidity, rainfall, and wind speed. These are critical for modeling water and nutrient-limited scenarios. When direct data are unavailable, weather generators can create statistically valid datasets using historical summaries.

7. Conclusion

Crop growth modeling is a powerful tool for informed decision-making in agriculture. It enables analysis of agricultural systems by simulating natural processes. However, gaps in our understanding of the complex soil–plant–atmosphere interactions limit model accuracy. Hence, research should prioritize improving this understanding over relying solely on empirical models. Calibration and validation of crop models play a key role in refining these insights and guiding research strategies. Despite their value, many agronomists lack training in crop modeling and systems research, highlighting the need for capacity building. The development of reliable models requires an integrated research approach with clearly defined objectives and comprehensive datasets. Such efforts can significantly advance sustainable agriculture to meet global food demands.

Suggested Readings

Boote, K.J., Jones, J.W., and Pickering, N.G., 1996. Potential uses and limitations of crop models. *Agronomy Journal*, **88**, 704-716.

Boughton, W., 2004. The Australian water balance model. *Environmental Modelling and Software*, **18**, 943-956.

Connolly, R.D., Bell, M., Huth, N., Freebairn, D.M., Thomas, G., 2002. Simulating infiltration and the water balance in cropping systems with APSIMSWIM. Australian Journal of Soil Research 40 (2), 221-242.

Cook, F.J., Fitch, P., Thorburn, P.J., Charlesworth, P.B., Bristow, K.L., 2006. Modelling trickle irrigation: Comparison of analytical and numerical methods for estimation of wetting front position with time. Environmental Modelling and Software 21 (9), 1353-1359.

CRC Catchment Hydrology, 2000. General approach to modelling and practical issues to model choice Melbourne, Australia. www.toolkit.net.au/ modelchoice.

Cruiziat, P., and Thomas, R., 1988. SPICE – a circuit simulation program for physiologists.

France, J., and J. H. M. Thornley. 1984. Mathematical Models in Agiculture. London: Butteworths.

Mohanty, M., Sinha, N. K., Painuli, D. K., Bandyopadhyay, K. K., Hati, K. M., Sammi Reddy, K., & Chaudhary, R. S. (2015). Modeling soil water contents at field capacity and permanent wilting point using artificial neural network for Indian soils. *National Academy Science Letters*, *38*, 373-377.

Sinha, N. K., Mohanty, M., Somasundaram, J., Chaudhary, R. S., Patra, H., Hati, K. M., ... & Prabhakar, M. (2021). Maize productivity analysis in response to climate

change under different nitrogen management strategies. *Journal of Agrometeorology*, *23*(3), 279-285.

Somasundaram, J., Sinha, N. K., Mohanty, M., Chaudhary, R. S., Hati, K. M., Singh, R. K., ... & Patra, A. K. (2018). Soil hydro-thermal regimes as affected by different tillage and cropping systems in a rainfed vertisol. *Journal of the Indian society of soil science*, *66*(4), 362-369.

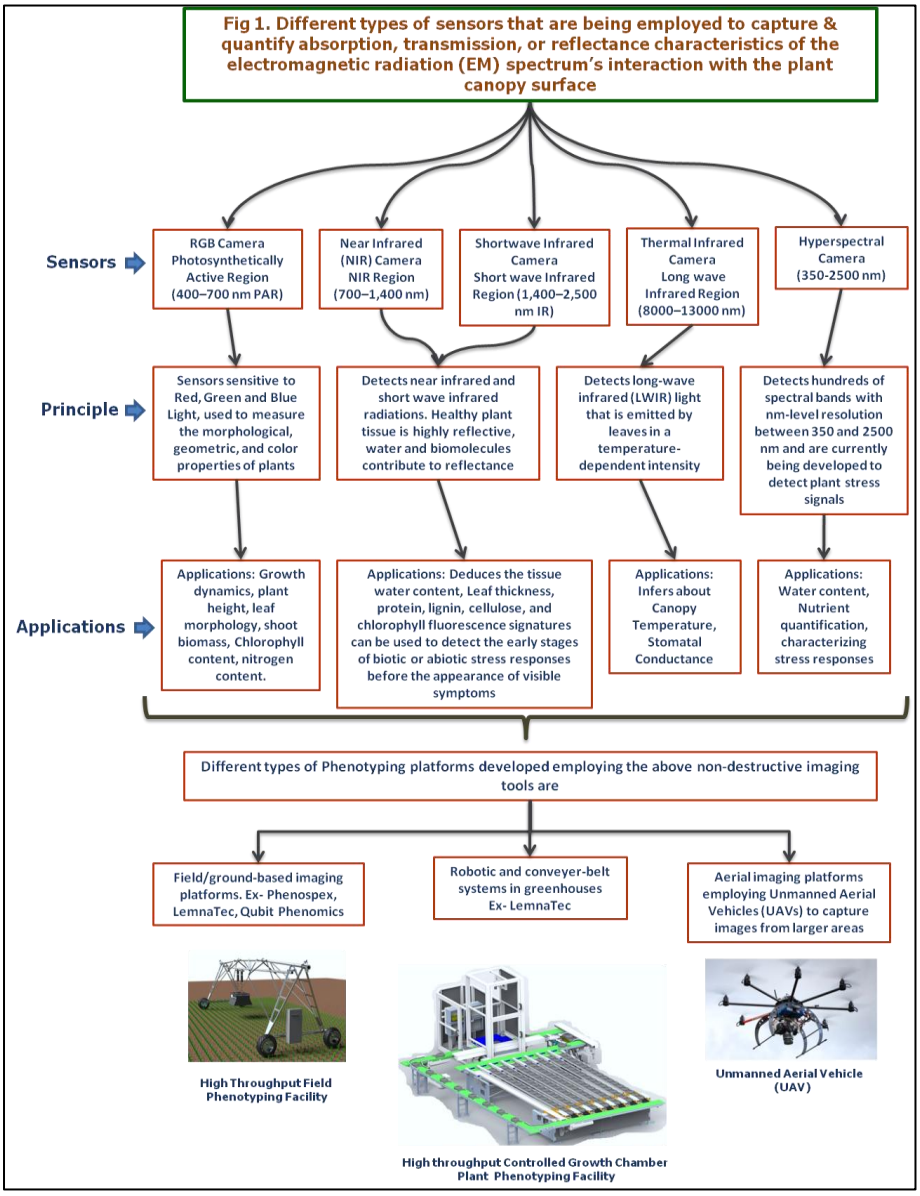
High Throughput Phenotyping for Abiotic Stress Management

Dr Prashantkumar S Hanjagi
ICAR-National Institute of Abiotic Stress Management, Baramati-413115
Email: psh7160@gmail.com

1. Introduction

Under the changing climate scenario, various abiotic stresses pose significant threat to global agricultural productivity and food security. Faced with extreme climate changes such as drought, high temperatures, flooding stresses, crop production must be maintained and improved while using less input resources. The complex multigenic and quantitative nature of abiotic stress tolerance mechanisms had previously made it difficult to fully understand the underlying molecular processes (Collins et al., 2008). However, the recent advancements in genomics and gene technology over the past decade and a half have increased confidence in the ability to find solutions, resulting in a wealth of genomic data. To make the most of this data for developing climate-resilient crop varieties, it is crucial to develop innovative methods for identifying quantitative phenotypes and understanding the genetic basis of agriculturally important traits using high-throughput approaches. However, the rapid advancement of genotyping has outpaced phenotyping capabilities, creating a "genotype-to-phenotype gap." To bridge this gap, it is necessary to develop comprehensive high-throughput phenotyping frameworks like High Throughput Phenotyping (HTP) Applications. High Throughput Phenotyping (HTP) technologies rely on non-invasive automated sensing methods capable of capturing plant responses to external stimuli by utilizing a wide variety of electromagnetic radiation wavelength bands perceived by camera through the image processing process. These non-destructive methods quantify the absorption, transmission, or reflectance characteristics of the electromagnetic radiation (EM) spectrum as it interacts with the plant canopy surface (Figure 1). When compared to stressed/infected plants, healthy/normal plants interact (absorb, reflect, transmit, and fluoresce) differently with electromagnetic radiation. HTP facilities/tools are extremely beneficial in a wide range of applications spanning from crop management to crop improvement. Monitoring crop health, performance, and growth phases at regular intervals aids in the efficient management of inputs such as water and nitrogen. Sensors attached on unmanned aerial vehicles (UAVs) can be used for large scale high throughput phenotyping to monitor plant growth status across a broader area, assisting agricultural management. Advanced integrated phenotyping technologies that combine molecular methodologies and non-invasive sensors help to accelerate the advancement of high-throughput crop improvement research, particularly in challenged conditions. This advanced research allows for the monitoring of high throughput phenotypic features and how they change in response to environment and genetics. This will also improve our

ability to carefully test a large number of germplasm for diverse abiotic stress tolerances. The found tolerant genotypes can be used in crop development programmes to generate climate resilient crop varieties. This lecture note provides an overview of high throughput phenomics research at multiple scales, focusing on technological and platform aspects, including microscopic, ground-based, and aerial phenotyping, and phenotypic data analysis. Recent uses of high-throughput phenotyping technologies for abiotic/biotic stress and yield evaluation will be described. Finally, this paper explores current issues and speculates on the future of phenomics research in abiotic stress management.



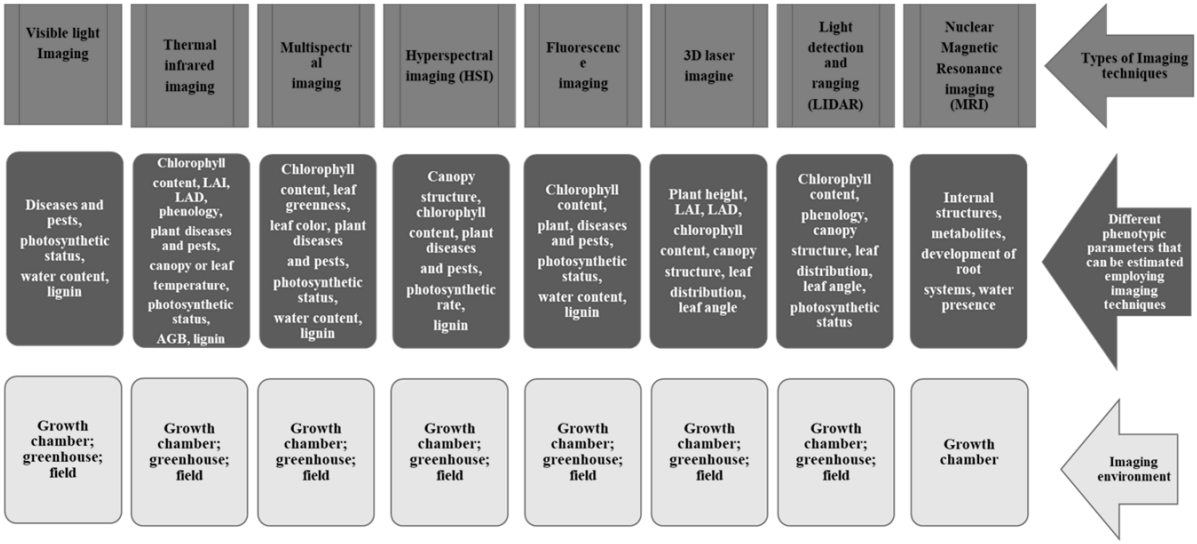
2. High Throughput Phenotyping Tools

The desire to improve the capacity to explore a large number of characteristics in a non-destructive and accurate manner has emerged as a significant goal in breeding for abiotic stress tolerance. The labor-intensive and costly aspect of phenotyping has encouraged the scientific community over the last decade to create automated systems for high-throughput plant phenotyping (HTPP). This has resulted in significant advances in the dissection of a wide variety of qualitative, agronomical, morphological, and physiological variables, as well as the examination of features associated to biotic and abiotic stressors. These technologies, which depend on automated non-invasive sensing methods, collect plant properties through picture analysis. Several technologies are currently available, including standard RGB/CIR cameras, thermal infrared systems, spectroscopy (hyperspectral and multispectral remote sensing), fluorescence and tridimensional (3D) imaging, and magnetic resonance imagers (MRI). These next-generation image sensors have been designed to enhance accuracy, resolution, and throughput, each with its own set of benefits and drawbacks. Depending on the aims of each phenotyping experiment, planned objectives, and outcomes, several image sensors can be used in phenotyping. Non-destructive optical imaging, in contrast to traditional approaches based on visual grading, strives for quick and contact-free evaluation of features in crop morphology and physiology. Figure 2 highlights the most commonly used optical imaging sensors for plant phenotyping in various conditions. The features and potentialities of these techniques have been thoroughly reported by Fritsche-Neto and Borém (2015). We will briefly discuss the implications (benefits and limits) that have emerged from their use in crop phenotyping.

RGB cameras/ colour infrared (CIR) cameras are widely used to assess plant canopy and biomass, which can be estimated by integrating footage from red, blue, and green light (RGB) and CIR(. The "leaf area index" (LAI) or "normalized difference vegetation index" (NDVI) can be calculated from the derived indices (NDVI). Multispectral (MS) and hyperspectral (HS) analysis are more focused on plant physiological features (e.g., photosynthetic efficiency, water and nutrient content, , etc.) by employing a series of images that spans the entire spectrum of radiation from visible (400-765 nm: VIS) to near-infrared (765 to 3200 nm: NIR). Thermal imaging is more adapted to evaluating (i) the stage of pathogen infections or disease spread in crops, and (ii) reactions to abiotic stressors (e.g., drought and heat tolerance). It is based on observations of the fluctuation in leaf surface temperature, which is connected to variances in stomatal conductance caused by different stressors that influence water balance and transpiration (Liew et al., 2008). Fluorescence, which captures excitation and emission spectra during chlorophyll absorption at shorter wavelengths, can determine plant metabolic status (Li et al., 2014). 3D imaging can estimate vegetation canopy structure and topographic maps, whereas MRI detects nuclear resonance signals from isotopes to provide pictures of the plant's interior components. These devices can be used alone or in conjunction to provide various data results. Depending on the image resolution, HTP platforms may be categorized into three

groups: microscopic, ground-based, and aerial phenotyping platforms. which enable the identification of phenotypic features at the tissue, organ, individual plant, plot, and field levels. (Figure 2).

Figure 2: A summary of most commonly employed image sensors under high throughput plant phenotyping



Many commercial platforms have been developed for a wide range of species, comprising of small plants such as Arabidopsis and the primary cereal crops (Table 1). High-throughput Rice Phenotyping Facility (HRPF), has a capacity of accommodating 5472 rice plants, including colour imaging, X-ray computed tomography (X-ray CT), automated controls, and an image processing pipeline that is capable of monitoring at least 15 agronomic variables in populations of up to 1920 rice plants (Yang et al., 2014). Furthermore, some image-based phenotypic traits that are difficult to analyze manually, such as leaf rolling and stay-green traits in drought tolerant rice genotypes can also be assessed (Duan *et al*., 2018). The Nanaji Deshmukh Plant Phenomics Centre, ICAR-IARI, New Delhi has been developed for precise HTP of the different crop genotypes throughout the life cycle in controlled stress conditions to identify genotypes for developing climate resilient crop varieties (Dwivedi et al., 2022).

Table 1: High throughput HTP platforms for phenotyping a wide range of crop species in different parts of the world

HTP platforms of the world	Facility placed	HTPA for Abiotic stresses	Reference
HTPheno	Leibniz Institute of Plant Genetics and Crop Plant Research, Germany	Drought	Hartmann et al., 2011
High-throughput Rice Phenotyping Facility (HRPF)	HAU, Wuhan, China	Drought	Yang et al., 2014
Bellwether Phenotyping Platform	Donald Danforth Plant Science Center (St. Louis, MO, USA)	Drought	Fahlgren etal.,2015
The Plant Accelerator	Australian Plant Phenomics Facility, Adelaide, Australia	Drought, Salinity and Nitrogen deficiency	Hairmansis et al., 2014; Neilson et al., 2015; Atieno et al., 2017
Lemnatec Scanalyzer	ICAR-NIASM, Baramati	Drought, Salinity and high temperature	Rane et al., 2021
Nanaji Deshmukh Plant Phenomics Centre	ICAR-IARI, New Delhi	Drought and Salinity	Dwivedi et al., 2022

3. Role of HTP in Abiotic Stress Management:

Abiotic stress management requires precise phenotypic screening to identify and characterize stress-tolerant genotypes. HTP enables early detection of stress responses, monitor temporal changes in plant performance; quantify stress effects on physiological traits and to screen large populations under abiotic stress conditions. HTP enables large-scale, objective, and time-series evaluation of plants in response to environmental stimuli. Abiotic stresses such as heat, drought, salinity, waterlogging drastically affect plant growth, yield, and food security. Accurate phenotyping is critical to understand plant responses under stress, identify stress-tolerant genotypes, to support breeding and management strategies and to accelerate crop improvement. HTP plays a pivotal role in each of these steps by allowing researchers to capture real-time physiological, biochemical, and morphological responses to abiotic stress.

4. Applications of HTP in Abiotic Stress Management:

HTP has emerged as a powerful tool to address challenges due to various abiotic stresses such as drought, heat, salinity, and waterlogging, etc. Its integration into crop science allows researchers to measure plant traits rapidly, accurately, and non-destructively across large populations, facilitating the development of stress-resilient varieties. Table 2 summarizes different physiological traits that can be measured by employing HTP sensors/tools. While the importance of physiological traits is being summarized in Table 3. Below are key applications of HTP in managing abiotic stress:

Table 2: Physiological Traits Measured by different HTP sensors/tools

Physiological Trait	Description	HTP Tools/Sensors Used	Relevance in Abiotic Stress Management
Canopy Temperature	Surface temperature of plant canopy, influenced by transpiration rate.	Infrared Thermography, IR Cameras, FLIR systems	Elevated temperature indicates stomatal closure under drought stress. Helps detect early water stress.
Chlorophyll Fluorescence (Fv/Fm, ΦPSII)	Measures photosystem II efficiency and photochemical quenching.	PAM Fluorometers, Fluorescence Cameras (e.g., FluorCam, MultispeQ)	Early indicator of photoinhibition due to heat, drought, or salinity. Helps assess photosynthetic performance.
Chlorophyll Content / Greenness Index (SPAD, NDVI)	Indicates pigment concentration and plant health.	SPAD Meter, Multispectral Sensors, NDVI Cameras, MultispeQ	Reflects stress-induced senescence or nutrient deficiency. Useful under drought, salinity, and heat stress.
Leaf Water Content / Relative Water Content (RWC)	Degree of cellular hydration.	NIR Spectroscopy, Gravimetric HTP platforms, Thermal Imaging (indirect)	Decline under water deficit; early marker of drought tolerance.
Stomatal Conductance	Rate of gas exchange through stomata.	Porometers, MultispeQ, Open Gas Exchange Systems	Reduced under water deficit; correlates with yield under drought.
Canopy Reflectance / Spectral Indices (NDVI, PRI)	Measures reflectance at specific wavelengths.	Hyperspectral and Multispectral Sensors, UAV-mounted cameras	NDVI tracks biomass and greenness; PRI indicates light use efficiency and stress response.
Plant Height / Growth Dynamics	Monitors elongation, biomass gain over time.	LiDAR, RGB imaging, Stereo cameras, Structure-from-Motion	Affected by nutrient and moisture stress. Tracks developmental response to stress over time.
Biomass Estimation	Non-destructive assessment of total aboveground biomass.	RGB-based 3D reconstruction, LiDAR, Multiview Imaging	Useful under drought or nutrient stress to track yield potential.
Canopy Architecture (Leaf Area Index, Leaf Angle, Compactness)	Structure and orientation of canopy affecting light interception.	LiDAR, 3D Imaging, Stereo Vision	Determines photosynthetic capacity, radiation use efficiency under abiotic stress.

Thermal Imaging for Heat Stress Tolerance	Detects temperature variation within the plant body.	IR Cameras, Thermal UAVs	Identifies genotypes with efficient cooling mechanisms (transpiration efficiency).
Root Traits (length, volume, angle)	Root system traits related to water/nutrient uptake.	Rhizotrons, MRI, X-ray CT, WinRHIZO, shovelomics + imaging	Crucial for understanding drought, waterlogging, and nutrient stress tolerance.

Table 3: Physiological traits and their importance under abiotic stresses

Physiological Traits	Importance under Stress
Chlorophyll Fluorescence (Fv/Fm)	Indicator of PSII efficiency; early stress signal
(NDVI	Proxy for green biomasSs and photosynthetic capacity
Canopy Temperature (CT)	Proxy for transpiration and stomatal behaviour
SPAD Index	Leaf greenness, linked to chlorophyll/nitrogen content
Stomatal Conductance (gs)	Indicator of water use and gas exchange
Plant Height and Leaf Area	Biomass estimation and growth rate

4.1 Early Detection of Stress Symptoms

Thermal cameras detect canopy temperature depression (CTD), indicating stomatal conductance and early water stress. Chlorophyll fluorescence sensors monitor photosynthetic efficiency (Fv/Fm) under heat or salinity. Multispectral imaging detects leaf discoloration, senescence, and nutrient stress.

4.2 Quantifying Physiological Responses: HTP platforms can measure photosynthetic rate and transpiration rate using gas exchange systems, relative chlorophyll content, NDVI and PRI from reflectance sensors, canopy architecture through 3D imaging and LiDAR. These are key indicators of plant function under drought, salinity, and nutrient stress.

4.3 Understanding the root architecture and function: Rhizotrons, MRI, and X-ray CT integrated with ML algorithms allow non-invasive analysis of root length, root surface area, root depth and angle. This is critical for selecting genotypes with improved water and nutrient uptake under stress.

4.4 Mapping traits and computing stress Indices: HTP-derived traits are used to compute stress indices like Stress Tolerance Index (STI) and Drought Susceptibility Index (DSI). These indices support mapping of quantitative trait loci (QTL) and genome-wide association studies (GWAS) for stress tolerance.

4.5 Temporal Dynamics under Stress: Time-series data from UAVs/drones help track growth recovery, leaf senescence, and flowering shifts under intermittent stress that can capture the kinetics of stress response, not just end-point measurements.

4.6 HTP Applications (HTPA) in Plant Breeding

The assessment of crop phenotypic traits in both controlled environments and the field has significantly increased over the past 20 years because of advances in non-destructive sensing and imaging tools. Phenomics is becoming increasingly important in genetic research and precision agriculture because it provides a precise definition of various features in crops and enables a better understanding of phenotypic changes caused by underlying heritable genetic variation.

4.6.1 Breeding assisted by HTPA:

Image-based phenotyping technologies can accelerate crop breeding by quickly and precisely evaluating characteristics like biotic/abiotic resistance, grain yield, and grain quality. This can be demonstrated by contrasting a conventional long-term breeding experiment with a more modern, technologically-supported version. Hopkins initiated long-term directional recurrent selection for oil concentration in maize in Illinois in 1896; after about 100 generations, the oil content increased from 4.69% to 20.37% (Dudley and Lambert, 2004). Song et al. (1999) developed synthetic populations in China where the oil content increased from 4.71% to 15.55% in just 18 cycles of selection (Song and Chen, 2004). This relative acceleration can be chiefly attributed to NIR and MRI spectroscopy.

4.6.2 Genetic Mapping and High-Throughput Phenotyping

Genetic research has greatly been benefited directly from automated phenotyping applications. Different automated platforms that have been successfully utilized in developing climate resilient crop varieties. Numerous loci regulating yield and its constituent parts in diverse crops have been identified as a result of the numerous HTP platforms, technologies like sequencing, GWASs (Huang et al., 2010), and statistical methods for genetic mapping (van Eeuwijk et al., 2010). (Table 4).

Table 4: Employment of different HTP Platforms in Trait and Genotype Variation Discovery.

Sl. No	Automated platforms (HTPA)	Traits studied	Genotyping Methods employed	Reference
1	Rhizoscope phenotyping platform	Root system architecture, tiller number, and root/shoot biomass	Genome Wide Association Studies (GWAS)	Courtois et al.,2013
2	3D root imaging, GiA Roots 2D and 3D image analysis platform	2D and 3D root system architecture traits	Linkage analysis (LA)	Topp et al., 2013
3	High-throughput rice phenotyping facility (HRPF)	Plant morphological traits, biomass, and yield-related traits	GWAS	Yang et al.,2014
4	LemnaTec 3D Scanalyzer system	32 salinity-responsive fluorescence color classes	GWAS	Campbell et al.,2015
5	High-throughput leaf scoring (HLS)	29 leaf traits: leaf size, shape, and color traits	GWAS	Yang et al.,2015
6	PANorama	49 panicle phenotypes	GWAS & LA	Crowell et al.,2016
7	Australian Plant Phenomics Facility, The Plant Accelerator	Relative plant growth rate, transpiration rate and transpiration use efficiency, and select salinity tolerance traits	GWAS	Al-Tamimi et al.,2016
8	High-throughput hyperspectral imaging system	1540 hyperspectral indices, chlorophyll content	GWAS	Feng et al.,2017
9	Tractor-based high throughput phenotyping	Canopy height, canopy temperature depression, and three reflectance ratios	LA	Tanger et al.,2017
10	HRPF and HLS	51 image-based traits (i-traits), 10 leaf color-related traits,11 yield traits to reflect drought response	GWAS & LA	Guo et al.,2018
11	High-throughput micro-CT-RGB imaging system	74traits: tiller traits, tiller growth traits, biomass, shoot morphological and shoot growth traits	GWAS	Wu et al.,2019
12	Nanaji Deshmukh Plant Phenomics Centre	Projected shoot area (PSA), water use (WU), transpiration rate (TR), and near-infrared values (NIR)	GWAS	Dwivedi et al., 2022

4.7 Integration with Artificial Intelligence/Machine Learning (AI/ML) for Better Decision Making:

High-throughput phenotyping (HTP) is rapidly progressing to cater the growing demand for data-driven solutions to crop stress management, especially in the face of climate variability. Emerging technologies such as robotics, IoT, cloud computing, and artificial intelligence are now being harnessed to enhance the efficiency, accuracy, and scalability of phenotyping platforms.

HTP datasets are multi-dimensional, making machine learning essential for pattern recognition in stress response, genotype classification under stress, predictive modeling of yield or stress resilience and multivariate trait selection for breeding. For ex: ML models combining spectral + thermal + physiological data can accurately predict plant performance under drought stress.

Image-based phenotyping technologies hold immense promise for accelerating crop breeding by enabling rapid and accurate assessment of traits related to resistance against biotic and abiotic stresses, as well as grain yield and quality. The impact of such technological advancements becomes evident when comparing traditional long-term breeding efforts with more modern, technology-driven approaches. For example, Hopkins initiated a directional recurrent selection program for increasing oil content in maize in Illinois back in 1896, and it took around 100 generations to elevate oil concentration from 4.69% to 20.37% (Dudley and Lambert, 2004). In contrast, Song et al. (1999) in China developed synthetic populations where oil content improved from 4.71% to 15.55% in just 18 selection cycles (Song and Chen, 2004). This remarkable improvement in efficiency is largely due to the use of tools like Near-Infrared (NIR) and Magnetic Resonance Imaging (MRI) spectroscopy.

High-throughput phenotyping (HTP) platforms enable efficient screening of plant responses to abiotic stresses by accommodating large populations and automating data collection through non-destructive technologies that generate high-resolution data (Ubbens and Stavness, 2017). Phenomics facilities employ multidimensional imaging—including visible, infrared (IR), near-infrared (NIR), and hyperspectral bands—which facilitates time-series assessment of various physiological traits (Peuelas and Filella, 1998; Tester and Langridge, 2010). For instance, stomatal behavior under drought, a key indicator of stress tolerance, is commonly evaluated using thermal imaging and canopy temperature (Jones et al., 2009; Rischbeck et al., 2017). Continuous crop monitoring over time enables a more precise understanding of stress responses, improving trait selection accuracy. Kim et al. (2020) demonstrated that RGB imaging can be used to assess plant area, color, and compactness; NIR imaging to estimate plant water content; IR imaging to detect plant temperature and fluorescence; and gravimetric systems like DroughtSpotter R to measure water use efficiency (WUE), transpiration rate, and water loss dynamics across different crop stages. **4.7.1 IoT Integration for Real-Time Remote Sensing**

The Internet of Things (IoT) allows the integration of sensors and devices across field and greenhouse environments for continuous data acquisition and remote monitoring.

Advantages of IoT integration:

- ✓ Real-time tracking of soil moisture, canopy temperature, and microclimate conditions.
- ✓ Automated alerts for early stress detection (e.g., drought, heat).
- ✓ Integration with mobile and web dashboards for precision agriculture.

Applications:

- ✓ IoT-enabled leaf temperature and chlorophyll sensors.
- ✓ Wireless data transmission for irrigation scheduling.

4.7.2 Robotic Phenotyping in Field and Greenhouses

Robotics offers automation in data collection, minimizing human error and increasing throughput. Autonomous vehicles or gantry systems equipped with multispectral, thermal, and 3D sensors can be employed for image acquisition that will enable standardized and repeatable data collection.

Examples:

- ✓ Robotic carts in greenhouses monitoring plant height and leaf count.
- ✓ Field robots performing daily stress scoring and growth measurements.

4.7.3 Cloud-Based Analytics for Cross-Institutional Data Sharing

Cloud platforms enable centralized data storage, access, and analysis, facilitating collaboration and integration across research institutions allowing scalability for large datasets (e.g., multi-season and multi-location trials) with an added advantage of common standards and metadata formats for interoperability.

Platforms: CyVerse, Plant Phenomics Network, Elixir Europe, etc

4.7.4 Edge Computing and Artificial Intelligence for On-Site Analysis

Edge computing allows real-time data processing at the site of data collection, reducing dependency on internet connectivity and cloud access. Edge computing brings computational power directly to the sensor or local device (e.g., drone, handheld scanner) instead of sending data to distant servers. This allows instant data processing, even in remote agricultural fields. This will allow for on-the-fly image analysis i.e., real-time processing and analysis of images as they are being captured, without needing to store them first or pause data collection enabling adaptive responses to plant needs and trait detection. Integration with AI models for decision support (e.g., stress indexing, irrigation triggers).

Applications:

- ✓ Deep learning models for predicting canopy temperature trends or NDVI-based yield estimations under water or heat stress.
- ✓ AI-driven segmentation of root and shoot structures from complex backgrounds in field imagery.

Example Use Case: A drone-mounted multispectral camera with onboard AI model to identify water-stressed sugarcane in real time and mark them for irrigation scheduling.

4.7.5 Integration Synergy: Toward a Unified Platform

The prospects of HTP lie in the synergistic integration of these components. The next generation of phenotyping systems will be built on the synergy of IoT, robotics, cloud, edge computing, and AI. A truly next-generation phenotyping system will collect high-resolution multi-sensor data using IoT and robotics, process it rapidly on-site using edge AI and upload results to cloud platforms for predictive modelling, visualization, and breeding decisions. The k**ey** capabilities of such systems will be real-time, multi-sensor data fusion, rapid feedback loops to breeding programs and remote trial monitoring and management.

5. Conclusions and future perspectives

The rapid advancements in sensors, image processing, data analytics, and phenotyping technologies have enabled high-throughput phenotyping (HTP) both in the field and under controlled environments across various scales—ranging from microscopic setups to ground-based and aerial platforms. Equipped with diverse sensors, modern phenotyping systems can effectively evaluate traits related to yield potential and stability, supporting crop improvement efforts. The integration of cutting-edge technologies such as machine vision, automation, 5G networks, cloud computing, and artificial intelligence (including deep learning) has significantly enhanced the power of phenotyping in advancing both basic crop science and breeding. High-speed, high-throughput phenotyping is critical for scaling up breeding programs, particularly in integrating with genomic selection pipelines to accelerate genetic gain. Moreover, these innovative approaches are key to developing climate-resilient crop varieties, helping ensure future food security amid increasing climate variability and water scarcity. By leveraging these technologies, HTP platforms can dramatically improve the speed, accuracy, and cost-effectiveness of trait evaluation under abiotic stress conditions.

References

Al-Tamimi, N., Brien, C., Oakey, H., Berger, B., Saade, S., Ho, Y.S., Schmockel, S.M., Tester, M., and Negrao, S. (2016). Salinity tolerance loci revealed in rice using high-throughput non-invasive phenotyping. Nat. Commun. 7:13342

Atieno, J., Li, Y., Langridge, P., Dowling, K., Brien, C., Berger, B., Varshney, R.K., and Sutton, T. (2017). Exploring genetic variation for salinity tolerance in chickpea using image-based phenotyping. Sci. Rep. 7:1300.

Brachi, B.; Morris, G.P.; Borevitz, J.O. Genome-wide association studies in plants: The missing heritability is in the field. Genome Biol. 2011, 12, 232.

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, Goodman MM, Harjes C, Guill K, Kroon DE, Larsson S, Lepak NK, Li H, Mitchell SE, Pressoir G, Peiffer JA, Oropeza Rosas M, Rocheford TR, Romay MC, Romero S, Salvo S, Sanchez Villeda H, da Silva HS, Sun Q, Tian F, Upadyayula N, Ware D, Yates H, Yu J, Zhang Z, Kresovich S, McMullen MD (2009) The genetic architecture of Maize flowering time. Science 325(5941):714–718. doi:10.1126/science.1174276

Campbell, M.T., Knecht, A.C., Berger, B., Brien, C.J., Wang, D., and Walia, H. (2015). Integrating image-based phenomics and association analysis to dissect the genetic architecture of temporal salinity responses in rice. Plant Physiol. 168:1476–1489.

Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Phil Trans R Soc B 363:557–572

Collins NC, Tardieu F, Tuberosa R (2008) Quantitative trait loci and crop performance under abiotic stress:where do we stand? Plant Physiol 147:469–486

Courtois, B., Audebert, A., Dardou, A., Roques, S., Ghneim-Herrera, T., Droc, G., Frouin, J., Rouan, L., Goze, E., Kilian, A., et al. (2013). Genome-wide association mapping of root traits in a japonica rice panel. PLoS One 8:e78037.

Crowell,S., Korniliev, P., Falcao,A.,Ismail, A.,Gregorio, G.,Mezey,J., and McCouch, S. (2016). Genome-wide association and highresolution phenotyping link Oryza sativa panicle traits to numerous trait-specific QTL clusters. Nat. Commun. 7:10527.

Dudley, J.W., and Lambert, R.J. (2004). 100 generations of selection for oil and protein in corn. Plant Breed. Rev. 24:79–110.

Fahlgren, N., Feldman, M., Gehan, M.A., Wilson, M.S., Shyu, C., Bryant, D.W., Hill, S.T., McEntee, C.J., Warnasooriya, S.N., Kumar, I., et al. (2015). A versatile phenotyping system and analytics platform reveals diverse temporal responses to water availability in Setaria. Mol. Plant 8:1520–1535.

Feng, H., Guo, Z., Yang, W., Huang, C., Chen, G., Fang, W., Xiong, X., Zhang, H., Wang, G., Xiong, L., et al. (2017). An integrated hyperspectral imaging and genome-wide association analysis platform provides spectral and genetic insights into the natural variation in rice. Sci. Rep. 7:4401.

Foolad MR, Panthee DR (2012) Marker-assisted selection in Tomato breeding. Crit Rev Plant Sci 31:93–123.

Fritsche-Neto, R.; Borém, A. Phenomics: How Next-Generation Phenotyping Is Revolutionizing Plant Breeding; Springer: Dordrecht, Switzerland, 2015; pp. 1–142.

Furbank, R.T.; Tester, M. Phenomics—Technologies to relieve the phenotyping bottleneck. Trends Plant Sci. 2011, 16, 635–644.

Guo, Z., Yang, W., Chang, Y., Ma, X., Tu, H., Xiong, F., Jiang, N., Feng, H., Huang, C., Yang, P., et al. (2018). Genome-wide association studies of image traits reveal genetic architecture of drought resistance in rice. Mol. Plant 11:789–805.

Hairmansis, A., Berger, B., Tester, M., and Roy, S.J. (2014). Image based phenotyping for non-destructive screening of different salinity tolerance traits in rice. Rice 7:16.

Hartmann, A., Czauderna, T., Hoffmann, R., Stein, N., and Schreiber, F. (2011). HTPheno: an image analysis pipeline for high-throughput plant phenotyping. BMC Bioinformatics 12:148.

Hirschhorn, J.N.; Daly, M.J. Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 2005, 6, 95–108.

Huang X, Paulo MJ, Boer M, Effgen S, Keizer P, Koornneef M, van Eeuwijk FA (2011) Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. Proc Natl Acad Sci 108:4488–4493

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42:961–967.

Jin L, Lu Y, Shao Y, Zhang G, Xiao P, Shen S, Corke H, Bao J (2010) Molecular marker assisted selection for improvement of the eating, cooking and sensory quality of Rice (Oryza sativa L.). J Cereal Sci 51:159–164

Jones, H., Serraj, R., Loveys, B. R., Xiong, L., Wheaton, A., and Price, A. H. (2009). Thermal infrared imaging of crop canopies for the remote diagnosis and quantification of plant responses to water stress in the field. Funct. Plant Biol. 36, 978–989. doi: 10.1071/FP09123

Kim, S. L., Kim, N., Lee, H., Lee, E., Cheon, K. S., Kim, M., et al. (2020). High throughput phenotyping platform for analyzing drought tolerance in rice. Planta 252:38. doi: 10.1007/s00425-020-03436-9

Li, L.; Zhang, Q.; Huang, D. A review of imaging techniques for plant phenotyping. Sensors 2014, 14, 20078–20111.

Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, Harriman J, Glaubitz JC, Buckler ES,

Kresovich S (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. PNAS 110(2):453–458. doi: 10.1073/pnas.1215985110

Neilson, E.H., Edwards, A.M., Blomstedt, C.K., Berger, B., Moller, B.L., and Gleadow, R.M. (2015). Utilization of a high-throughput shoot imaging system to examine the dynamic phenotypic responses of a C4 cereal crop plant to nitrogen and water deficiency over time. J. Exp. Bot. 66:1817–1832.

Paux E, Faure S, Choulet F, Roger D, Gauthier V, Martinant JP, Sourdille P, Balfourier F, Le Paslier MC, Chauveau A (2010) Insertion site-based polymorphism markers open new perspectives for genome saturation and marker-assisted selection in wheat. Plant Biotechnol J 8:196–210

Peñuelas, J., and Filella, L. (1998). Visible and near-infrared reflectance techniques for diagnosing plant physiological status. Trends in Plant Sci. 3, 151–156. doi: 10.1016/S1360-1385(98)01213-8

Poorter, H., Fiorani, F., Stitt, M., Schurr, U., Finck, A., Gibon, Y., Usadel, B., Munns, R., Atkin, O.K., Tardieu, F., et al. (2012). The art of growing plants for experimental purposes: a practical guide for the plant biologist. Funct. Plant Biol. 39:821–838.

Rane J, Raina SK, Govindasamy V, Bindumadhava H, Hanjagi P, Giri R, Jangid KK, Kumar M and Nair RM (2021) Use of Phenomics for Differentiation of Mungbean (Vigna radiata L. Wilczek) Genotypes Varying in Growth Rates Per Unit of Water. Front. Plant Sci. 12:692564. doi: 10.3389/fpls.2021.692564

Rischbeck, P., Cardellach, P., Mistele, B., and Schmidhalter, U. (2017). Thermal phenotyping of stomatal sensitivity in spring barley. J. Agron. Crop Sci. 203, 483–493. doi: 10.1111/jac.12223

Robbins MD, Massud Mohammed AT, Panthee DR, Gardner RG, Francis DM, Stevens MR (2010) Marker-assisted selection for coupling phase resistance to Tomato spotted wilt virus and Phytophthora infestans (late blight) in Tomato. Hort Sci 45:1424–1428

Song, T.M., Kong, F., Li, C.J., and Song, G.H. (1999). Eleven cycles of single kernel phenotypic recurrent selection for percent oil in Zhongzong no. 2 maize synthetics. J. Genet. Breed. 53:31–35.

Assessment of Extreme Weather Events in India

Dr Ram Narayan Singh and Dr Sudhir Kumar Mishra
ICAR-National Institute of Abiotic Stress Management, Baramati-413115
Email: ranjay.singh@icar.org.in

Introduction

Extreme weather refers to unusual or unexpected weather events that deviate from the typical range of conditions for a particular area or time of year. These events can be characterized by unusual intensity, duration, or spatial extent. While related, it's important to differentiate extreme weather from severe weather, with severe weather being defined as any weather posing a risk to life or property. Extreme weather, on the other hand, encompasses unusual weather events that are at the extremes of the historical distribution for a given area. These events include heatwaves, cold waves, heavy rainfall, floods, droughts, cyclones, lightning strikes, and hailstorms. Their impact spans across agriculture, water resources, health, infrastructure, and livelihoods, particularly affecting the most vulnerable populations. Extreme weather events are increasingly impacting the world, driven by a combination of natural climate variability and human-induced climate change. Understanding their causes and consequences is essential for developing effective adaptation and mitigation strategies to reduce the risks to human societies and the environment.

The global human population is projected to reach approximately 9 billion by 2050, necessitating a 70% increase in food production to meet future demands (Alletto et al., 2010), along with a substantial rise in the demand for fibre and fuel (Asgedom et al., 2011; Beedy et al., 2010). Concurrently, the global energy requirement is anticipated to grow nearly tenfold compared to the levels at the beginning of the twentieth century (Boyle, 2004). These challenges are further compounded by climate change, which continues to exert mounting pressure on agricultural systems. The IPCC (2022) stresses that to limit global warming to 1.5°C above pre-industrial levels and prevent the most devastating effects of climate change, global CO_2 emissions must be halved by 2030 and reach net-zero by 2050. Moreover, achieving this goal requires the removal of 2 to 10 billion metric tons (Gt) of CO_2 annually. Beyond rising temperatures and intensifying heatwaves, climate

change also brings altered precipitation patterns, increased frequency and severity of storms, extended drought periods, and a rising risk of wildfires. According to NASA's Goddard Institute for Space Studies (GISTEMP, 2023), the global mean surface temperature has already increased by approximately 1.1°C since 1880, with heatwaves becoming more intense and prolonged, underscoring the growing impact of anthropogenic climate change on weather extremes.

In recent years, heat-waves and extended periods of extreme heat have markedly increased the frequency, durability, intensity and severity of hotter days and warmer nights, as a result of climate change. Chronic heat-waves combined with high humidity, pose a serious threat to living organisms in tropical and subtropical regions of the world. A rapid rise in the scale, intensity, frequency, and duration of extreme heat events has been observed globally, with a notable increase in concurrent heat-waves and droughts over the past century. Now days, Temperatures exceeding 40°C have become common in many parts of the world, and in some regions, the frequency of temperatures above 50°C is steadily increasing. According to IPCC (2023), global surface temperature has increased by 0.99 °C during 2001–2020 and by 1.09 °C during 2011–2020, compared to the pre-industrial baseline of 1850–1900. Under 1.5°C temperature warming would cause 150 or more days above 35°C per year in 67 cities while, a 3°C elevated temperature may potentially affect up to 197 cities (Mackres et al., 2023). Projection confirmed that by 2050, the number of urban poor exposed to extreme heat conditions could rise by 700%, with the most significant increases expected in regions like West Africa and Southeast Asia (UCCRN, 2018). If the current trend continues until 2050, almost every child below the age of 18 (nearly 2.2 billion) are expected to be suffered with heatwave which was just 24 per cent of children in 2020 (UNICEF, 2022). Similarly, mortality of people over the age of 65 years due to heat also increased nearly 85 % during 2000-2004 over 2018-2022 (LANCET 2023)

Causes and contributing factors

Extreme weather events are a result of complex interactions between natural variability and human influence. The natural factors like the El Niño-Southern Oscillation (ENSO) or the North Atlantic oscillation (NAO) can significantly influence weather patterns worldwide.

Climate Change is A major driver of extreme weather, human-caused global warming increases the frequency and intensity of events like heatwaves, droughts, heavy precipitation, and strengthens hurricanes. The anthropogenic Activities such as Deforestation, urbanization, and the destruction of wetlands can worsen the impacts of extreme weather. Poor urban planning can exacerbate flood risks and contribute to the formation of urban heat islands.

Table 1: Causes of Extreme Weather Events

Cause	Explanation
Climate Change	Rising global temperatures increase atmospheric instability, altering weather patterns.
Greenhouse Gas Emissions	Trapping of heat causes thermal stress and enhances storm frequency and intensity.
Deforestation & Urbanization	Disrupts local climate balance, increases surface temperatures and water runoff.
Oceanic Changes (El Niño, La Niña)	Disrupt monsoon patterns, cause droughts or floods.
Global Warming of Oceans	Intensifies cyclones and rainfall variability.
Anthropogenic Activities	Unsustainable agriculture, mining, and industrial emissions destabilize local climate.

Characteristics and Potential Impacts of Extreme weather event

Extreme weather events are characterized by their rarity and deviation from normal weather patterns in terms of magnitude, location, timing, or extent. These events are becoming more frequent and intense globally, largely due to climate change. Following Table (Table 2) provides a comprehensive overlook about the specific features of different common extreme weather events.

Table 2: Characteristics and Potential Impacts of Extreme weather event

Extreme weather event	Characteristics and Potential Impacts
Heatwave	Extended period of abnormally high temperatures. Can lead to heatstroke, dehydration, increased energy demand (e.g., for air conditioning), and exacerbation of wildfires
Cold Wave	Prolonged period of unusually low temperatures. Can cause hypothermia, frostbite, strain energy grids (e.g., for heating), and damage infrastructure and crops.
Drought	Prolonged period of deficient rainfall or moisture availability. Leads to water scarcity, crop failure, increased risk of wildfires, and impacts on agriculture and ecosystems.
Heavy Precipitation & Floods	Excessive rainfall over a short or extended period, leading to flooding. Can include flash floods, river floods, and coastal floods (storm surge). Results in damage to property and infrastructure, displacement, and health risks.
Tropical Cyclone	Powerful rotating storm systems over oceans with low-pressure centers, strong winds, and heavy rainfall. Also known as hurricanes or typhoons depending on the region. Can cause destructive winds, high waves, storm surge, coastal and inland flooding.

Tornado	Violently rotating columns of air extending from a thunderstorm to the ground. Capable of immense destruction in a localized area.
Severe Winter Weather	Includes events like blizzards (heavy snowfall, strong winds, low visibility) and ice storms (freezing rain accumulating as ice). Can disrupt transportation, cause power outages, and lead to injuries.
Wildfire	Uncontrolled fires spreading through vegetation, often intensified by dry conditions and high temperatures. Can cause widespread damage, air pollution, and force evacuations.
Hailstorm	Characterized by the presence of large quantities or size of hailstones. Can damage crops, vehicles, buildings, and livestock.
Thunderstorm	A rain shower with lightning and thunder. Can be severe, producing hail, high winds, and even tornadoes. Associated with phenomena like derechos (widespread straight-line winds) and downbursts.
Dust Storm	Strong winds carry large amounts of dust and sand, reducing visibility and causing respiratory problems.

Impacts of extreme weather

Human Health: Extreme weather events pose direct threats to human health, increasing the risk of heatstroke, injuries, and the spread of waterborne diseases.

Ecosystems: Habitat loss, species migration, and ecosystem disruption are significant consequences.

Social and Economic Impacts: Extreme weather can cause property damage, infrastructure destruction, displacement, and economic hardship.

In recent decades, Extreme weather events have increased in frequency, intensity, and geographical spread due to climate change and environmental degradation. According to IMD and Council on Energy, Environment and Water (CEEW), over 75% of Indian districts are now classified as extreme weather hotspots. Between 2005 and 2022, extreme rainfall events surged by more than 50%. In 2022 alone, India faced 314 days of extreme

weather, resulting in over 3,000 deaths and major economic losses. The Arabian Sea has seen a sharp rise in cyclone frequency and intensity. Additionally, the duration and severity of heatwaves have tripled in recent decades. Some significant global examples of the extreme weather events and their impacts are illustrated in following Table 3.

Table 3: Description of extreme weather events and their impacts

Event Type	Region & Year	Description	Impacts
Heatwave	Western Canada, 2021	Record-breaking heatwave with Lytton, British Columbia reaching 49.6°C, the highest temperature ever recorded in Canada.	Hundreds of deaths, wildfires, significant environmental damage.
Heatwave	Europe, 2003	Record-breaking heat caused tens of thousands of deaths, particularly in France, Spain, and the UK. Temperatures exceeded 40°C in some areas.	Thousands of deaths, heat-related illnesses, strain on infrastructure.
Heatwave	Asia, 2023	Record-high temperatures across Bangladesh, India, Thailand, and	Potential for heatstroke, dehydration, and increased energy demand.

Tropical Cyclone	USA, 2005	Hurricane Katrina, a major Category 3 hurricane, caused catastrophic flooding in New Orleans as levees failed.	1800+ deaths, over a million displaced, an estimated $125 billion in damage.
Tropical Cyclone	USA, 2021	Hurricane Ida, a Category 4 storm, made landfall in Louisiana and moved up the East Coast, causing devastating floods in the Mid-Atlantic and Northeast.	Fifth-costliest weather disaster in world history ($65 billion).
Tropical Cyclone	India & Bangladesh, 2020	Super Cyclone Amphan caused massive devastation and ranked as one of the year's most costly climate disasters.	Widespread damage and economic losses.
Drought	Siberia, 2021	One of the world's coldest regions experienced its driest summer in 150 years.	Record number of wildfires, including a mega-wildfire consuming over a million and a half hectares.
Flood	Germany, 2021	Continuous rains in western Germany, particularly North Rhein-Westphalia and Rheinland-Palatinate, led to widespread destruction.	Hundreds dead and missing, infrastructure damage.
Flood	Libya, 2023	Storm Daniel brought torrential rain, with Al Bayda receiving 414mm in 24 hours. The deluge broke dams near Derna, destroying neighborhoods.	Thousands killed, critical infrastructure damaged.
Flood	China, 2021	The city of Zhengzhou experienced rainfall equivalent to one year's precipitation in a single day.	Evacuation of around 200,000 people, widespread flooding.

Wildfire	Amazon, 2023	The Brazilian Amazon basin experienced its lowest cumulative rainfall since 1980.	Lowest Rio Negro water level since 1902, affecting ecosystems and livelihoods.

Severe weather is a particular type of extreme weather which poses risks to life and property. In general, any dangerous meteorological phenomenon which can make damage, disrupt or loss of for society and human life (WMO, 2004). Severe weather encompasses a variety of dangerous meteorological events that can significantly impact human life, infrastructure, and ecosystems. Understanding the different types and the factors contributing to their occurrence is crucial for effective preparedness, risk management, and adaptation strategies. Severe weather events encompass a wide range of meteorological phenomena:

Thunderstorms: These can be categorized into various types, including single-cell, multi-cell, squall line, and the strongest, supercell thunderstorms. They pose risks due to phenomena like lightning, strong winds (including downbursts and derechos), hail, and heavy rainfall that can lead to flash flooding.

Tornadoes: Violently rotating columns of air that extend from a thunderstorm to the ground, capable of immense destruction.

Tropical Cyclones: Known as hurricanes, typhoons, or cyclones depending on the region, these are powerful, rotating storm systems characterized by low-pressure centers, strong winds, and heavy rainfall. They can cause destructive winds, high waves, and storm surges, resulting in extensive coastal and inland flooding.

Extratropical Cyclones: These storms, also called European windstorms or Nor'easters, are powered by horizontal temperature differences and can cause hurricane-force winds, heavy precipitation (rain or snow), coastal flooding, and erosion.

Heavy Rainfall and Flooding: Excessive precipitation can lead to various types of flooding, including river flooding, flash flooding, and coastal flooding exacerbated by sea-

level rise and storm surges. Floods can damage property and infrastructure, endanger lives, and pose public health risks.

Droughts: Prolonged periods of unusually dry weather that can deplete water resources, cause crop failure, and increase the risk of wildfires.

Heatwaves: Extended periods of excessively hot weather that can pose serious health risks, particularly for vulnerable populations, and increase demand for electricity.

Severe Winter Weather: This includes heavy snowfall, blizzards, and ice storms. Blizzards are strong snowstorms with low visibility and powerful winds, making travel dangerous and posing risks of frostbite and hypothermia. Ice storms involve the gradual accumulation of ice on surfaces, which can be highly destructive to trees, power lines, and infrastructure.

Dust Storms: Characterized by strong winds carrying large quantities of dust and sand, reducing visibility and posing respiratory health risks.

Factors contributing to severe weather

A combination of factors can contribute to severe weather events:

Atmospheric conditions: These include moisture content, instability, wind shear (changes in wind speed or direction with height), and the presence of cold air aloft.

Geographical features: Latitude, altitude, and topography play a role in determining the types and severity of weather that a region experiences.

Climate Variability: Natural climate patterns like El Niño or La Niña can influence the occurrence of extreme weather events.

Climate Change: A significant factor contributing to the increasing frequency and intensity of severe weather events globally. Rising global temperatures can exacerbate heatwaves and droughts, increase atmospheric water vapor leading to heavier precipitation, and potentially intensify hurricanes and winter storms.

Extreme weather versus severe weather: a comparison

While often used interchangeably, the terms "extreme weather" and "severe weather" have distinct definitions in meteorology. Understanding the differences is crucial for accurate

communication and effective preparedness In summary, all severe weather is extreme because it's dangerous and outside typical conditions, but not all extreme weather is severe. An unusual snowfall in a desert is extreme due to rarity but not necessarily severe unless it causes harm. Similarly, a severe thunderstorm may be both severe (hazardous) and extreme (a significant deviation from normal weather).

Table 4: Comparison between extreme weather versus severe weather

Feature	Extreme Weather	Severe Weather
Definition	Unusual weather events at the extremes of historical distribution.	Dangerous phenomena with potential to cause harm.
Focus	Statistical rarity or deviation from the norm.	Potential for hazards and impacts.
Relationship	Severe weather is a subset of extreme weather.	Not necessarily rare.
Examples	Prolonged heatwaves, unseasonal cold snaps, tropical cyclones in unusual locations.	Thunderstorms with tornadoes or large hail, flash floods, blizzards, strong winds.
Duration	Can be prolonged.	Often short-term, but can be prolonged.

Classification of Extreme Weather Events

The India Meteorological Department (IMD) classifies weather phenomena, including temperature, rainfall, and cyclones, using standardized thresholds to issue alerts and forecasts. The key IMD classifications of extreme weather events are described below (Table 5 and Table 6).

Table 5: Specific conditions for different Extreme Weather Events

1. Rainfall Classification (Daily Rainfall)	
Category	**Rainfall (mm/day)**
No Rain	0
Very Light Rain	0.1 – 2.4
Light Rain	2.5 – 15.5
Moderate Rain	15.6 – 64.4
Heavy Rain	64.5 – 115.5
Very Heavy Rain	115.6 – 204.4
Extremely Heavy Rain	≥204.5
2. Heatwave Classification	
Criteria	**Condition**

2. Heatwave Classification	
Criteria	**Condition**
Heatwave	Departure of ≥4.5°C to 6.4°C from normal OR temperature ≥45°C
Severe Heatwave	Departure of >6.4°C from normal OR temperature ≥47°C
3. Cold Wave Classification	
Criteria	**Condition**
Cold Wave	Departure of -4.5°C to -6.4°C from normal OR minimum temperature ≤4°C
Severe Cold Wave	Departure of > -6.4°C OR minimum temperature ≤2°C

3. Cold Wave Classification	
Criteria	**Condition**
Cold Wave	Departure of -4.5°C to -6.4°C from normal OR minimum temperature ≤4°C
Severe Cold Wave	Departure of > -6.4°C OR minimum temperature ≤2°C
4. Cyclone Classification (Based on Wind Speed)	
Category	**Wind Speed (km/h)**
Depression	31 – 49
Deep Depression	50 – 61
Cyclonic Storm	62 – 88
Severe Cyclonic Storm	89 – 117
Very Severe Cyclonic Storm	118 – 165
Extremely Severe Cyclonic Storm	166 – 220
Super Cyclonic Storm	≥221

5. Drought Classification (Standardized Precipitation Index - SPI)	
Category	**SPI Value**
Near Normal	-0.99 to 0.99
Mild Drought	-1.0 to -1.49
Moderate Drought	-1.5 to -1.99
Severe Drought	≤ -2.0

Table 6: Analysis of meteorological events based on multi-variant indices:

ID	Indicator Name	Indicator Definitions	Units
TXx Max	Tmax	Monthly maximum value of daily max temperature	°C
TNx Max	Tmin	Monthly maximum value of daily min temperature	°C
TXn Min	Tmax	Monthly minimum value of daily max temperature	°C
TNn Min	Tmin	Monthly minimum value of daily min temperature	°C
TN10p	Cool nights	Percentage of time when daily min temperature < 10th percentile	%
TX10p	Cool days	Percentage of time when daily max temperature ≤10th percentile	%
TN90p	Warm nights	Percentage of time when daily min temperature > 90th percentile	%
TX90p	Warm days	Percentage of time when daily max temperature > 90th percentile	%
DTR	Diurnal temperature range	Monthly mean difference between daily max and min temperature	°C

GSL(Growing season length)	Annual (1st Jan to 31st Dec in Northern Hemisphere, 1st July to 30th June in Southern Hemisphere)	count between first span of at least 6 days with TG > 5 °C and first span after July 1 (January 1 in SH) of 6 days with TG < 5 °C days	
FD0	Frost Days	Annual count when daily minimum temperature < 0 °C	Days
SU25	Summer Days	Annual count when daily maximum temperature > 25 °C	Days

TR20	Tropical Nights	Annual count when daily minimum temperature > 20 °C	Days
WSDI	Warm Spell Duration Indicator	Annual count of periods with ≥6 consecutive days of max temperature > 90th percentile	Days
CSDI	Cold Spell Duration Indicator	Annual count of periods with ≥6 consecutive days of min temperature < 10th percentile	Days
RX1day		Monthly maximum 1-day precipitation amount	mm
RX5day		Monthly maximum 5-day consecutive precipitation	mm

CDD	Consecutive Dry Days	Maximum number of consecutive days with precipitation <1 mm	Days
	CWD Consecutive Wet Days	Maximum number of consecutive days with precipitation ≥1 mm	Days
R95p	Very Wet Days	Annual total precipitation from days > 95th percentile	mm
R99p	Extremely Wet Days	Annual total precipitation from days > 99th percentile	mm
PRCPTOT	Annual Wet-Day Total	Annual total precipitation from days with precipitation ≥1 mm	mm

Types of Extreme Weather Events in India

India, due to its diverse geography and climate, experiences a wide range of extreme weather events, which are increasingly being influenced by climate change. Indian regions prone to specific type of extreme weather are discussed below (Table 7).

Table 7: Extreme weather prone regions of India

Event Type	Description	Regions Most Affected
Heatwaves	Unusually high temperatures, often exceeding 45°C	Northern, Central, and Northwestern India (e.g., Rajasthan, Delhi, UP)
Cold waves	Sudden drop in minimum temperatures during winter	North and Central India (e.g., Punjab, Bihar)
Heavy rainfall	Short bursts of intense rainfall leading to flash floods and landslides	Western Ghats, Northeast India, Himalayan states
Floods	Riverine, urban, and flash floods affecting croplands, settlements	Assam, Bihar, Kerala, Maharashtra
Droughts	Extended dry spells affecting crops, drinking water, and livestock	Maharashtra, Karnataka, Telangana, Rajasthan
Cyclones	Intense tropical storms from the Bay of Bengal and Arabian Sea	Eastern Coast (Odisha, Andhra Pradesh, West Bengal), Western Coast (Gujarat)
Lightning	Sudden electrical discharges during thunderstorms	Bihar, Jharkhand, Odisha, West Bengal
Hailstorms	Frozen precipitation damaging crops and property	Maharashtra, Madhya Pradesh, Chhattisgarh

Impact of Extreme weather events on Agriculture and Allied Sectors

Impact of Extreme weather events on Agriculture and Allied Sectors

The impacts of extreme weather events on agriculture and allied sectors are undeniable and multifaceted. Addressing these challenges requires a concerted effort involving policymakers, researchers, extension workers, and farming communities to build

a more resilient and sustainable agricultural future in the face of climate change. Likely impacts of the specific extreme weather conditions have been categorized as under:

A. Crop Growth & Yield

Extreme weather events (EWEs) pose a significant and growing threat to agriculture worldwide, particularly in regions like India with a high dependence on farming. Impact of the extreme weather events on crop growth and yield has been illustrated in Table 8.

Table 8: Impact of the extreme weather events on crop growth and yield

Extreme Event	Impact on Crops
Heatwaves	Spike in evapotranspiration; pollen sterility in cereals; reduced grain filling
Droughts	Soil moisture stress; stunted growth; higher pest attack risk
Floods/Heavy Rain	Root anoxia, nutrient leaching, crop lodging, fungal infections
Cold Waves	Delayed germination, frost injury, reduced photosynthesis
Hailstorms	Physical damage to standing crops, especially fruits and vegetables
Unseasonal Rains	Disruption in harvesting, post-harvest losses, fungal diseases

B. Overall Agricultural Production

- **Yield loss** leads to food insecurity.
- Regional imbalance in **crop suitability zones**.
- Decline in **agricultural GDP**, especially in rain-fed areas.
- **Increased dependency on irrigation** and input-intensive methods.

5. Impact on Farmers' Livelihoods

- **Crop failure** → loss of income → rising farm debt.
- **Distress migration** due to unviable farming conditions.
- Mental health issues including **farmer suicides**.

- **Reduced capacity to invest** in future farming cycles.
- Dependency on **government compensation** and relief programs.

6. Impact on National Economy

Agriculture contributes ~15-18% of India's GDP, but supports **over 50% of the population**. Extreme events may cause:

- Disruption in supply chains (mandis, transport, exports).
- Inflation in food prices due to supply-demand mismatch.
- Increase in government expenditure on disaster relief, rehabilitation, and subsidies.
- Insurance payouts surge under schemes like PMFBY.
- Negative impact on rural consumption and agribusiness sectors.

Response and Adaptation Measures

Addressing the growing threat of extreme weather events to Indian agriculture requires a combination of immediate responses and long-term adaptation measures. By implementing these comprehensive response and adaptation measures, India can build a more climate-resilient and sustainable agricultural sector, safeguarding food security and ensuring the well-being of its farming communities in the face of escalating extreme weather events.

A. Responses

Early Warning Systems: Improved weather forecasting by agencies like the India Meteorological Department (IMD) helps farmers make timely decisions about planting, harvesting, and protecting their crops and livestock from sudden weather changes.

Contingency Crop Planning: Developing and disseminating region-specific contingency plans for different Extreme weather events allows farmers to switch to more suitable crops or adjust farming practices in response to weather anomalies like droughts, floods, or heatwaves.

Emergency Relief: Providing immediate financial assistance to farmers affected by Extreme weather events through subsidies, low-interest loans, or grants helps them recover from losses and restart their livelihoods.

Post-Harvest Management: Investing in resilient infrastructure for storing, transporting, and marketing agricultural produce helps minimize post-harvest losses and ensures food security during times of crisis.

B. Adaptation measures

Climate-Resilient Crop Varieties: Research and development in collaboration with institutions like the Indian Council of Agricultural Research (ICAR) focuses on developing and promoting crop varieties that are tolerant to drought, heat, flood, and salinity, as well as resistant to pests and diseases.

Sustainable Water Management: Promoting efficient irrigation methods such as micro-irrigation (drip and sprinkler systems), rainwater harvesting, and watershed management can help farmers manage water resources more effectively, especially in regions facing water scarcity or erratic rainfall patterns.

Soil Conservation Practices: Encouraging practices like zero-tillage farming, crop rotation, and the use of organic fertilizers helps improve soil health, conserve moisture, and enhance resilience to climate variability.

Agroforestry: Integrating trees and shrubs into farming systems can provide multiple benefits like carbon sequestration, improved soil fertility, enhanced biodiversity, and additional income streams for farmers.

Livelihood Diversification: Promoting integrated farming systems that combine crops with livestock, horticulture, or fisheries can help reduce farmers' reliance on single crops and provide alternative sources of income, making them more resilient to Extreme weather events.

Technology and Digital Interventions: Leveraging digital platforms, AI-based tools, and mobile applications can provide farmers with real-time weather updates, expert advice on

crop management, pest and disease alerts, and market information, enabling data-driven decision-making.

Financial Inclusion and Insurance: Schemes like the Pradhan Mantri Fasal Bima Yojana (PMFBY) provide insurance coverage and financial support to farmers in case of crop losses due to Extreme weather events, offering a safety net during challenging times.

Capacity Building and Awareness: Providing training and educational programs to farmers on climate-smart agricultural practices and risk management strategies enhances their knowledge and enables them to adopt new technologies and practices more effectively.

Policy Support and Governance: Developing and implementing supportive policies and initiatives like the National Mission for Sustainable Agriculture (NMSA) and the National Innovations in Climate Resilient Agriculture (NICRA) project are crucial for creating an enabling environment for climate adaptation and mitigation in the agricultural sector.

References

Alletto L, Coquet Y, Benoit P, Heddadj D, Barriuso E (2010) Tillage management effects on pesticide fate in soils: A review. Agron Sustain Dev. 30 (2): 367-400.

Asgedom H, Kebreab E (2011) Beneficial management practices and mitigation of greenhouse gas emissions in the agriculture of the Canadian Prairie: a review. Agron Sustain Dev. 31 (3):433–451.

Beedy TL, Snapp SS, Akinnifesi FK, Sileshi GW (2010). Impact of Gliricidia sepium intercropping on soil organic matter fractions in a maize-based cropping system. *Agric Ecosyst Environ*. 138(3–4): 139–146.

Boyle G. (2004). *Renewable Energy: Power for a sustainable future*. 2nd edn. Milton Keynes, UK: Oxford University Press.

GISTEMP (2023). *GISS Surface Temperature Analysis (GISTEMP).* version 4. NASA Goddard Institute for Space Studies. https:// data.giss.nasa.gov/gistemp/

IPCC (2022). Climate Change 2022: impacts, adaptation, and vulnerability. *In*: H-O Po¨rtner, DC Roberts, M Tignor, ES Poloczanska, K Mintenbeck, A Alegrıa, M Craig, S. Langsdorf, S. Lo¨schke, V. Mo¨ller, et al., eds, Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK and New York, NY, USA, pp 3056

UNICEF (2022). United Nations Children's Fund, The Coldest Year of the Rest of their Lives: Protecting children from the escalating impacts of heatwaves, UNICEF, New York, October 2022. https://www.unicef.org/media/129506/file/UNICEF-coldest-year-heatwaves-and-children-EN.pdf

IPCC (2023). Summary for Policymakers. Figure SPM.1. In: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland, pp. 1-34, doi: 10.59327/IPCC/AR6-9789291691647.001

LANCET (2023). Lancet Countdown: Heat-related Mortality. 2023. https://www.lancetcountdown.org/data-platform/health-hazards-exposures-and-impacts/1-1-persons

Mackres E, Wong T Null S, Campos R and Mehrotra S (2023). The Future of Extreme Heat in Cities: What We Know — and What We Don't. https://www.wri.org/insights/future-extreme-heat-cities-data

UCCRN (2018) Urban Climate Change Research Network. (February 2018). How Climate Change Could Impact the World's Greatest Cities: UCCRN Technical Report. C40 Cities. 1789_Future_We_Dont_Want_Report_1.4_hi-res_120618.original-compressed.pdf

Climate Change Indices

Dr. Gopalakrishnan B
ICAR-National Institute of Abiotic Stress Management, Baramati-413115
Email: leogopi@yahoo.co.in

Introduction

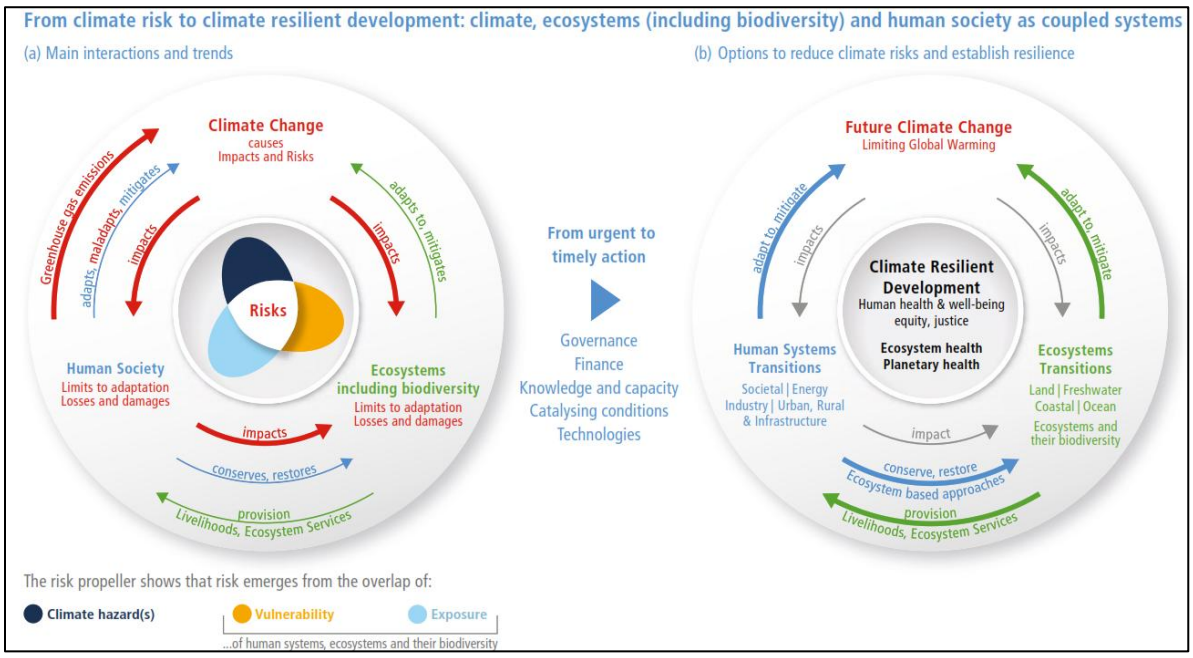
Climate change refers to long-term alterations in temperature, rainfall patterns, wind systems, and other elements of the Earth's climate. These changes are primarily driven by human actions, including the burning of fossil fuels, large-scale deforestation, and emissions from industrial activities. Such actions have significantly increased the concentration of greenhouse gases like carbon dioxide and methane in the atmosphere, resulting in global warming. The impacts of climate change are widespread—manifesting as more frequent and intense heatwaves, rising sea levels, melting glaciers, shifting ecosystems, and an increase in the frequency and severity of extreme weather events such as droughts and hurricanes.

In recent years, the urgency to respond to climate change has intensified, as it poses serious threats to biodiversity, food and water security, human health, and economic stability. Tackling this global challenge requires coordinated international action and the adoption of sustainable development practices. To monitor and manage the effects of climate change, a range of indices have been developed. These indices help quantify aspects such as risk, exposure, vulnerability, and hazards associated with climate change. This document provides an in-depth overview of these climate change indices, detailing their significance, definitions, and the mathematical methods used in their calculation.

Climate risk and IPCC

The Intergovernmental Panel on Climate Change (IPCC) in its Sixth Assessment Report (AR6) gave the risk propeller diagram, a conceptual model to visually represent the components of climate risk in the context of climate change. It helps illustrate how hazard, exposure, and vulnerability interact to create risk, and it highlights the role of adaptation and mitigation in reducing this risk. Adaptation (usually depicted as a stabilizing force) refers to adjustments in systems or practices to reduce vulnerability and exposure. This

includes early warning systems, climate-resilient infrastructure, improved health care access, etc. Mitigation (often presented as a reduction of hazard potential refers to efforts to reduce greenhouse gas emissions and slow climate change. This includes activities like switching to renewable energy, enhancing carbon sinks, and improving energy efficiency. The diagram is used to guide climate risk assessments and design climate-resilient development policies. It illustrates that reducing one or more of the propeller blades (hazard, exposure, vulnerability) can reduce overall risk, emphasizing integrated action that combines adaptation and mitigation efforts.



Climate Risk Assessment (CRA)

The Intergovernmental Panel on Climate Change (IPCC), in its Sixth Assessment Report (AR6), employs a thorough risk assessment framework that combines scientific data with socio-economic factors to analyze the effects of climate change. This method highlights climate risk as a product of three main elements: hazard (climate-driven physical events or trends), exposure (the presence of people, ecosystems, or assets in vulnerable areas), and vulnerability (the degree to which those exposed are susceptible to harm and their ability to adapt and recover).

AR6 introduces a more nuanced, systems-based framework that considers compound and cascading risks, reflecting how multiple climate hazards can interact and amplify one

another. It also incorporates risk thresholds and limits to adaptation, highlighting areas where adaptation may no longer be feasible. This methodology enables a more holistic understanding of climate impacts across regions, sectors, and timeframes, supporting better-informed decision-making and more effective climate resilience strategies.

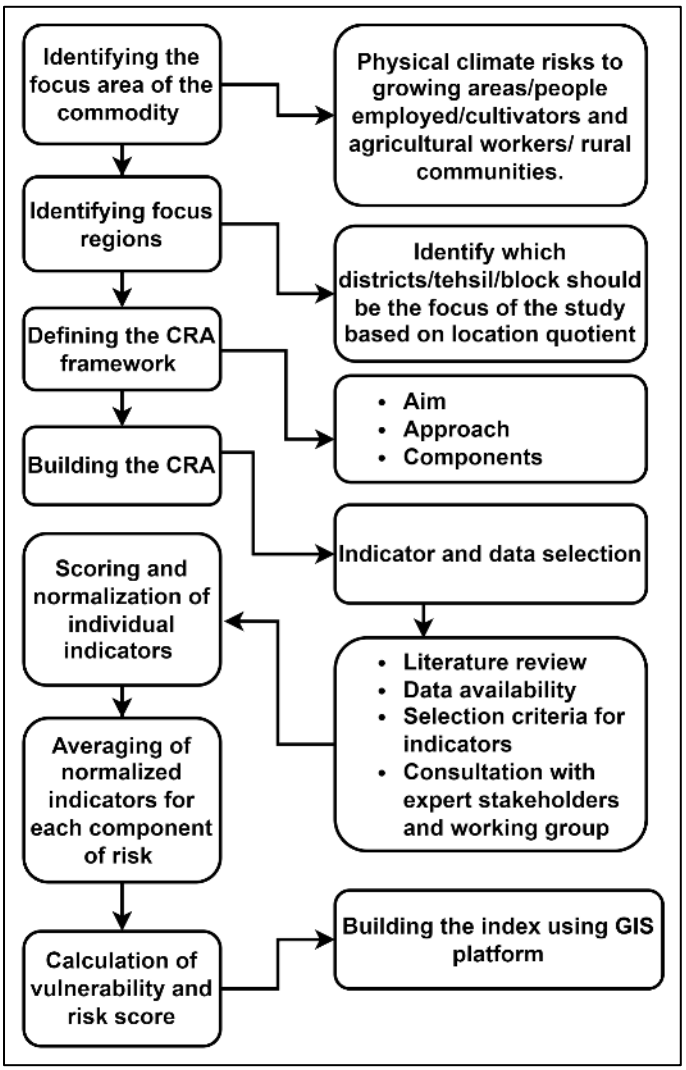


Figure 1. Risk assessment methodology (IPCC)

CRA Framework

The IPCC's Climate Risk Assessment (CRA) framework is a structured approach to evaluating the potential impacts of climate change on various systems and identifying strategies for adaptation and mitigation. It involves assessing vulnerabilities, identifying potential risks, and developing strategies to manage those risks. The IPCC's CRA framework provides a systematic approach for understanding the potential impacts of

climate change, identifying vulnerabilities and risks, and developing strategies to manage those risks and adapt to a changing climate.

Key Components of the IPCC's CRA Framework

Vulnerability Assessment

This involves evaluating the susceptibility of a system (e.g., a community, ecosystem, or infrastructure) to climate change impacts. It considers factors like exposure, sensitivity, and adaptive capacity.

Risk Assessment

This step identifies the potential hazards and their likelihood of occurrence, as well as the potential consequences of those hazards. It considers both the hazards associated with climate change and the potential risks of the project or activity increasing climate vulnerability.

Adaptation and Mitigation Strategies

Based on the vulnerability and risk assessments, the framework guides the development of strategies to reduce vulnerability and manage risks. Adaptation strategies aim to reduce the negative impacts of climate change, while mitigation strategies aim to reduce greenhouse gas emissions.

The framework also includes:

Screening: Identifying potential climate risks.

Scoping: Defining the boundaries of the assessment.

Report Generation: Documenting the findings of the assessment.

Climate Risk Management Plan: Developing strategies for managing identified risks.

Climate Change Indices

At the core of IPCC's risk assessment is the following conceptual formula:

Risk = *f* (Hazard, Exposure, Vulnerability)

This is a qualitative functional expression, meaning that risk is not the product of these terms in a strict mathematical sense but rather a result of their interaction.

Hazard (H)

A climatic hazard is a weather or climate-related event that has the potential to cause harm to people, property, ecosystems, or economies. These hazards include both extreme events—such as hurricanes, heatwaves, floods, droughts, and wildfires—and slow-onset

changes, like rising sea levels or shifting rainfall patterns. Climatic hazards can occur naturally, but their frequency and intensity are increasingly influenced by human-induced climate change. They become particularly dangerous when they intersect with vulnerable populations or poorly prepared systems, leading to significant social, environmental, and economic impacts.

$H = P(E) \times I$

Where:

$P(E)$ is the probability of the extreme climate event; I is the intensity of the event

Climate Hazard Index (CHI)

It measures the potential occurrence and intensity of climate-related hazards such as droughts, floods, heatwaves, and storms.

$$CHI = \sum_{i=1}^{n} (H_i \times W_i)$$

Where: -

H_i = Hazard value for event; W_i = Weight of the event based on severity and frequency; n = Number of hazard events considered. A higher CHI indicates a greater level of climatic hazard.

Exposure (E)

In the context of climate risk, exposure refers to the presence of people, ecosystems, infrastructure, economic assets, or cultural resources in places that could be adversely affected by climate-related hazards. It describes what is at risk and where, without necessarily considering how vulnerable those elements are. For example, a coastal city is exposed to sea-level rise and storm surges, while a farming community in a semi-arid region may be exposed to drought. Exposure is a key component of climate risk assessment, alongside hazard and vulnerability, and helps determine the potential scale and scope of climate impacts.

Exposure Index (EI)

The Exposure Index quantifies the extent to which people, property, systems, or other elements are present in areas that could be adversely affected by climate hazards.

$$EI = \frac{\sum_{j=1}^{m}(P_j \times A_j)}{T}$$

Where:

Pj = Population or asset value in area; Aj = Area exposed to climate hazard; T = Total population or asset value; m = Number of areas considered. A higher EI denotes a higher degree of exposure to climate hazards.

Vulnerability

It refers to the degree to which a system, community, or individual is susceptible to harm from climate-related hazards. It is shaped by a range of factors, including socio-economic conditions, governance, access to resources, health, infrastructure, and social inequalities. High vulnerability means greater sensitivity to climate hazards and limited capacity to respond effectively. Low-income communities may be more vulnerable to floods due to inadequate housing and a lack of access to emergency services. Vulnerability, together with exposure and hazard, is a key component in assessing climate risk, and addressing it is essential for building resilience and reducing the impacts of climate change.

Vulnerability Index (VI)

The index reflects the susceptibility of a system or population to harm due to exposure to climate hazards.

$$VI = \frac{(S + AC + SES)}{3}$$

Where:

S = Sensitivity component (e.g., age, health status); AC = Adaptive capacity component (e.g., access to healthcare, infrastructure); SES = Socioeconomic status. Higher values of VI imply a greater vulnerability.

Sensitivity

It refers to the degree to which a system or population is affected, positively or negatively, by climate-related hazards or changes. It reflects how responsive a system is to changes in

climate conditions such as temperature, precipitation, or extreme weather events. For instance, crops are highly sensitive to changes in rainfall and temperature, while elderly populations may be more sensitive to heat waves due to health vulnerabilities. Sensitivity is closely related to vulnerability, but while vulnerability also includes the capacity to cope or adapt, sensitivity focuses specifically on how strongly the system reacts to climate stimuli. Understanding sensitivity helps in identifying which systems or groups are more at risk and informs targeted adaptation strategies.

Adaptive Capacity

It describes the capacity of a system, community, or individual to adapt to climate-related hazards by minimizing potential harm, seizing possible benefits, or managing the resulting impacts. This capacity plays a vital role in determining both vulnerability and resilience, and is shaped by various factors including financial resources, access to education, infrastructure quality, technological advancement, governance structures, and the strength of social networks.

High adaptive capacity enables better preparation for and response to climate impacts, reducing overall risk. For example, a country with strong institutions, effective early warning systems, and access to financial and technical resources will generally have a higher adaptive capacity than one lacking these elements. Enhancing adaptive capacity is essential for climate change adaptation and long-term sustainable development.

Adaptive Capacity Index (ACI)

This index measures the ability of a system, community, or individual to adjust to potential damage, to take advantage of opportunities, or to respond to consequences.

$$ACI = \frac{E + I + H + T}{4}$$

Where:

E = Education level; I = Income level; H = Health index; T = Technology access. Higher ACI values indicate greater adaptive capacity.

Resilience

The ability of a system, community, or society to anticipate, prepare for, respond to, and recover from climate-related hazards and stresses, while maintaining essential functions and adapting to change. It involves not only withstanding shocks such as floods, droughts, or storms but also learning from these events to improve future responses and reduce long-term vulnerability. Resilience is strengthened through adaptive capacity, social cohesion, effective governance, access to resources, and infrastructure that can endure or quickly recover from disruption.

Climate Resilience Index (CRI)

CRI assesses the capacity of a system to anticipate, prepare for, respond to, and recover from climate-related disruptions.

$$CRI = \frac{ACI}{1 + VI}$$

Where:
ACI: Adaptive Capacity Index; VI: Vulnerability Index. Higher CRI values indicate more resilience to climate impacts.

Climate Risk

It refers to the potential for adverse impacts resulting from climate-related hazards, which arise from the interaction between hazards (such as extreme weather events or slow-onset changes like sea-level rise), exposure (the presence of people, assets, or ecosystems in harm's way), and vulnerability (the susceptibility and capacity to cope or adapt). Climate risk can be both current and future, and it encompasses a wide range of physical, social, economic, and environmental consequences.

Climate Risk Index (CRI)

The risk index estimates the overall potential impact of climate hazards, combining hazard, exposure, and vulnerability. It typically evaluates factors like loss of life, economic

damage, and the frequency and severity of extreme weather events, offering a way to gauge a country's or region's exposure and vulnerability to climate hazards. CRI helps highlight areas most at risk, supports climate adaptation planning, and draws attention to the urgent need for climate action, especially in vulnerable developing nations. It is a comprehensive indicator used to prioritize areas for climate adaptation and mitigation interventions.

$$CRI = CHI \times EI \times VI$$

Social Vulnerability Index (SVI)

SVI is a measure used to assess the susceptibility of communities to harm from external stresses, such as natural disasters or climate change, based on social, economic, and demographic factors. It evaluates aspects like poverty, age, disability, housing conditions, access to transportation, education, and minority status to determine how well a population can prepare for, respond to, and recover from hazards. A high SVI indicates that a community is more likely to experience greater impacts due to limited resources and adaptive capacity. SVI is widely used in disaster planning, public health, and climate resilience efforts to prioritize aid and develop targeted support strategies for at-risk populations.

$$SVI = \sum_{k=1}^{p} (Z_k \times W_k)$$

Where:

Zk = Normalized value of vulnerability indicator; Wk = Weight of the indicator; p = Number of indicators. Communities with higher SVI values are more socially vulnerable.

Environmental Sensitivity Index (ESI)

It is an index developed to identify areas that are ecologically sensitive and more likely to be affected by climate change or disasters. It integrates physical, biological, and human-use information to map and assess which areas are most vulnerable to damage and which should be prioritized for protection in the event of an environmental disaster. ESI maps typically include three key components: habitat type (e.g., rocky, sandy, marshy), biological resources (such as species), and human-use resources (like recreational areas,

cultural sites, or water intakes). By providing a detailed understanding of the environmental and socio-economic value of the regions, the ESI supports effective emergency response planning, resource management, and environmental impact assessments.

$$ESI = \frac{R + B + H}{3}$$

Where:

R = Resource dependency; B = Biodiversity index; H = Habitat fragility. High ESI scores signal high environmental sensitivity.

Growing Degree Days (GDD)

It is a climate-based index used to estimate the growth and development of plants and insects during the growing season. It measures the accumulation of heat above a certain base temperature, which is the minimum temperature required for the growth of a particular organism.

$$\text{GDD} = \frac{T_{\text{max}} + T_{\text{min}}}{2} - T_{\text{base}}$$

Where:

T_{max} = daily maximum temperature; T_{min} = daily minimum temperature; T_{base} = base temperature (commonly 10°C for many crops). If the daily mean temperature is below the base temperature, the GDD for that day is set to zero.

Palmer Drought Severity Index (PDSI)

This index relies on easily accessible temperature and precipitation data to assess relative dryness. It is standardized, typically ranging from -10 (extremely dry) to +10 (extremely wet). While agencies like NOAA often present values within a -4 to +4 range, more extreme readings can occur. The Palmer Drought Severity Index (PDSI) has proven effective in measuring long-term drought conditions. By incorporating temperature data and a physical water balance model, it also reflects the influence of global warming on drought through shifts in potential evapotranspiration.

PDSI value	classification
4.0 or more	extremely wet
3.0 to 3.99	very wet
2.0 to 2.99	moderate wet
1.0 to 1.99	slightly wet
0.5 to 0.99	incipient wet spell
0.49 to -0.49	near normal
-0.5 to -0.99	incipient dry spell
-1.0 to -1.99	mild drought
-2.0 to -2.99	moderate drought
-3.0 to -3.99	severe drought
-4.0 or less	extreme drought

Heat Index (HI)

The Heat Index (HI), often referred to as the "apparent temperature," is a measure of how hot it feels to the human body when relative humidity is combined with air temperature. High humidity reduces the body's ability to cool itself through sweating, making the temperature feel hotter than it is.

$$\mathrm{HI} \approx T + 0.33 \times e - 0.70 \times ws - 4.00$$

Where:

T = air temperature in °C; e = water vapor pressure (in hPa)

$$e \approx \frac{\mathrm{RH}}{100} \times 6.105 \times e^{\left(\frac{17.27 \times T}{237.7 + T}\right)}$$

ws = wind speed in m/s

Climate Change Performance Index (CCPI)

CCPI is an independent monitoring tool developed by Germanwatch, NewClimate Institute, and the Climate Action Network (CAN) to evaluate and compare the climate protection performance of countries. It assesses how well countries are performing on the global goal of limiting climate change to well below 2°C, as outlined in the Paris Agreement. It helps to increase transparency in international climate politics by tracking

country progress in reducing greenhouse gas emissions, promoting peer pressure among nations, and highlighting leaders and laggards in climate action.

CCPI Score $=0.4\times Score_{GHG} + 0.2\times Score_{RE} + 0.2\times Score_{EU} + 0.2\times Score_{Policy}$

Where: $Score_{GHG}$: Greenhouse Gas Emissions Score; $Score_{RE}$: Renewable Energy Score; $Score_{EU}$: Energy Use Score; $Score_{Policy}$: National and international climate policy score (based on expert assessments)

Conclusion

Understanding and calculating climate change indices are essential for effective disaster risk reduction, urban and regional planning, resource allocation, and policymaking. These indices offer critical insights into where and how to implement mitigation and adaptation strategies. In the future, climatic indices will play an increasingly vital role in enhancing our understanding and management of climate-related risks. They will aid in climate modeling, impact assessments, and the development of evidence-based adaptation and mitigation strategies across sectors such as agriculture, water resources, public health, and urban planning. Moreover, by translating complex climate data into accessible and actionable insights, climatic indices will empower policymakers, researchers, and communities to make informed decisions, build resilience, and navigate the uncertainties of a rapidly changing climate with greater preparedness and precision.

References

Forchhammer, M. C., & Post, E. (2004). Using large-scale climate indices in climate change ecology studies. *Population Ecology*, *46*, 1-12.

Edmonds, H. K., Lovell, J. E., & Lovell, C. A. K. (2020). A new composite climate change vulnerability index. *Ecological Indicators*, *117*, 106529.

Dash, S., & Maity, R. (2019). Temporal evolution of precipitation-based climate change indices across India: contrast between pre-and post-1975 features. Theoretical and Applied Climatology, 138(3-4), 1667-1678.

Mukherjee, S., Mishra, A., & Trenberth, K. E. (2018). Climate change and drought: a perspective on drought indices. *Current climate change reports*, *4*, 145-163.

Water Footprint Assessment for Mitigating Water Stress in Agriculture

Dr Sudhir Kumar Mishra
ICAR-National Institute of Abiotic Stress Management, Baramati-413115
Email: sudhirmet@yahoo.com

Introduction

Water is an indispensable resource for ensuring global food security, sustaining ecosystem services, and achieving long-term sustainable development goals. Agriculture, being the largest water-consuming sector, accounts for nearly 70% of global freshwater withdrawals, thereby exerting immense pressure on water resources worldwide (Hoekstra et al., 2011). This pressure is expected to intensify as the global population approaches 9.7 billion by 2050 and dietary preferences shift toward water-intensive food products such as meat and dairy (Zeng et al., 2020). Concurrently, climate change is altering global hydrological cycles, increasing the frequency and severity of droughts, and undermining the stability of water supply systems, particularly in arid and semi-arid regions (Vanham et al., 2022). In this context, quantifying and managing water use in agriculture has become imperative for mitigating water stress and ensuring sustainable agricultural production. Conventional water use metrics, which primarily quantify volumetric water withdrawals, often fail to capture the broader environmental implications of water consumption and pollution.

The concept of the Water Footprint (WF), first proposed by Hoekstra and Hung (2002), offers a comprehensive perspective on freshwater use by accounting for both direct and indirect water consumption across the entire life cycle of a product or activity. The WF is composed of three main elements: green, blue, and grey water (Hoekstra et al., 2011). Green water denotes rainwater retained in the soil that is utilized by plants through evapotranspiration, which is particularly significant in rain-fed agriculture. Blue water encompasses surface and groundwater extracted for agricultural, industrial, or domestic purposes. When this water is not returned to its original source, it contributes to freshwater depletion. Grey water refers to the volume of freshwater needed to assimilate pollutants and maintain water quality standards, thus serving as an indicator of water pollution.

These three components form the basis of Water Footprint Assessment (WFA), a standardized methodology developed by the Water Footprint Network. WFA enables the measurement and evaluation of water use and contamination throughout the production process of goods and services at both local and global scales. Closely related to this is the concept of virtual water, which describes the unseen water embedded in traded commodities, especially in agriculture and industry. For instance, the export of water-intensive crops like rice effectively transfers large quantities of virtual water-mainly blue water-from one region to another. Consumptive water use refers to water that is removed from the system through evaporation, plant uptake, or incorporation into products, making it unavailable for reuse. This contrasts with non-consumptive use, where water is returned to the source. Return flows, such as treated wastewater or runoff, re-enter water bodies but can affect water quality and contribute to grey water volumes if polluted. Together, these concepts are vital for evaluating the sustainability, efficiency, and fairness of water use. WFA serves as a valuable tool for identifying water consumption patterns, assessing environmental impacts, and informing strategies for improved water governance, particularly in agriculture and international trade.

The escalating global demand for food, feed, and bioenergy presents formidable challenges to water security. In India, the home of 18% of the global population endowed with only 4% of the world's freshwater resources, faces acute water stress in arid and semi-arid zones. Overexploitation and mismanagement of blue water sources, such as rivers, lakes, wetlands, and aquifers, have already led to significant depletion (Falkenmark and Molden, 2008; Wada and Bierkens, 2014), thereby constraining the productivity of irrigated agriculture (Davis et al., 2017). In contrast, green water i.e., moisture derived from precipitation stored in the soil, constitutes a substantially larger share of agricultural water use globally. Despite its significance, green water utilization has been underrepresented in agricultural water management strategies. Although green water contributes four to five times more than blue water in global food production (Hoff et al., 2010), yet focus of water policies has predominantly been on blue water. Numerically, rainfed agriculture accounts for over 60% of India's net sown area, therefore, optimizing green water use is crucial for enhancing food security and climate resilience, especially in vulnerable regions

(Falkenmark and Rockström, 2006). Water productivity, defined as the crop yield per unit of evapotranspiration, is a vital indicator of sustainable water use in both irrigated and rainfed systems. Enhancing water productivity not only reduces the need to expand agricultural land but also minimizes the pressure on limited water resources (Molden et al., 2010). On a global scale, trade in agricultural commodities from regions with high water productivity to regions with lower productivity facilitates net water savings and contributes to alleviating local water stress (Chapagain et al., 2006; Fader et al., 2011). In India, augmenting green water productivity holds considerable potential for reducing water stress in both water-scarce states like Rajasthan, Gujarat, and Maharashtra, and relatively water-abundant states such as West Bengal, Assam and Meghalaya.

Water footprint assessment has demonstrated significant utility in identifying inefficiencies and environmental burdens associated with water use in agriculture. Empirical studies have revealed that certain crops such as rice, sugarcane, and cotton exhibit disproportionately high blue and grey water footprints, particularly when cultivated in water-stressed regions (Mekonnen and Hoekstra, 2011). Spatially explicit analyses have also highlighted stark regional disparities in water productivity and pollution levels, often revealing zones where water use exceeds renewable supply or where pollutant loads surpass the assimilative capacity of local ecosystems (Vanham et al., 2022; Zeng et al., 2020). Such insights are invaluable for informing policy decisions related to water governance, especially in contexts marked by competing demands from agricultural, industrial, and domestic sectors. In recent years, Water footprint assessment has gained increasing prominence in agricultural policy and planning under climate-resilient agriculture. By linking water use to environmental, economic, and social outcomes, Water footprint assessment facilitates the development of adaptive strategies such as crop diversification, modernization of irrigation infrastructure, promotion of deficit irrigation, and water quality management (Jalilov et al., 2018; Karandish et al., 2020). The concept of virtual water trade has also emerged as a strategic tool for conserving local water resources, enabling water-scarce nations to import water-intensive goods while mitigating internal resource depletion (Hoekstra and Chapagain, 2008). Nevertheless, the application of water footprint assessment is challenged by methodological discrepancies, data gaps,

and the complex interplay of biophysical and socio-economic variables (Chouchane et al., 2015). Amidst escalating global water crises and the centrality of agriculture in water management, there is a compelling need to expand the water footprint assessment as a robust decision-support system. This study examines the water footprint of cotton, wheat and cotton-wheat production systems to develop appropriate strategies for water stress mitigation. Specifically, the study focuses on identifying high-impact crops, assessing the sustainability of water use across agro-ecological zones, and proposing context-specific interventions to enhance water-use efficiency and reduce pollution. Through the integration of recent methodological advancements and high-resolution spatial datasets, the study seeks to inform evidence-based policy decisions that bolster the sustainability and resilience of agricultural water management under the dual pressures of climate change and resource scarcity. The water footprint components are briefly discussed in following points.

Footprinting

Footprinting is a method used to assess the environmental impact of human activities, especially in terms of resource consumption and pollution. In water footprinting, it quantifies the amount of water used, directly and indirectly, throughout the life cycle of a product, service, or activity. It helps in identifying water-scarce regions and improving sustainable water use strategies. Traditional irrigation-based approaches are increasingly being replaced with holistic concepts such as ‘virtual water’ and ‘water footprint’.

Virtual water

Virtual water refers to the hidden or embedded water used to produce a product or service, which is not visible in the final item. For example, the water used to grow wheat, process it, and bake bread is considered its virtual water. When goods are traded, virtual water is also transferred between countries or regions. This concept is essential for understanding global water dependencies and the impact of trade on water resources. Virtual water refers to the total volume of water used across the entire production chain of

a commodity (Allan, 1997; 2003), while water footprint provides an estimate of the direct or indirect water involved in crop production (Hoekstra et al., 2011).

Water footprint

Water footprint is a measure of the direct and indirect total volume of freshwater used to produce goods, services, or by individuals and communities. It includes green (rainwater), blue (surface/groundwater), and grey (polluted water) components. The concept helps track water consumption along the supply chain and informs sustainable water management. It can be applied at the individual, product, regional, or national level.

Green water footprint

Green water footprint is the amount of rainwater consumed during the production of crops or biomass. It includes water that is stored in the soil and evaporated or transpired by plants. This component is crucial for rain-fed agriculture and forestry and does not deplete surface or groundwater sources.

Blue water footprint

Blue water footprint measures the volume of surface and groundwater consumed in production processes. This includes irrigation water for crops or industrial and domestic water withdrawals that are not returned to the source. It is particularly important in irrigated agriculture and can contribute to freshwater depletion if overused.

Grey water footprint

Grey water footprint indicates the volume of freshwater required to dilute pollutants to meet water quality standards. It reflects the level of water pollution caused by agricultural runoff, industrial discharges, or domestic waste. A high grey water footprint suggests poor water quality and the need for better waste management. Total water footprint is the sum of green, blue, and grey water footprints. It represents the complete water impact of a product, process, or region.

Total water footprint:

Total water footprint is the sum of green, blue, and grey water footprints used in the production and consumption of a product, service, or by an individual, sector, or nation. It provides a comprehensive view of the total freshwater resources consumed and polluted across the entire supply chain. This metric helps assess the full environmental impact of water use and supports better planning for sustainable water management. Total water footprint can also be used to compare the water intensity of different activities or products

Agricultural water footprint

Agricultural water footprint refers to the amount of water used in crop and livestock production, including irrigation, rainwater use, and water pollution. It is a major part of the global water footprint, as agriculture consumes over 70% of freshwater resources. Managing this footprint is vital for food security and water sustainability. Global and national water footprint studies have been analyzed on various crops such as mango (Australia), banana (Peru), citrus (South Africa), cocoa (Colombia), and date palm (Israel and Iran) (Shtull-Trauring et al., 2016; Mekonnen and Hoekstra, 2011). The below table represents computation of Water footprint components of conventional and drip fertigation

Eqn. no.	Parameters	Formula	Reference
(1)	Adjusted time	2.2 * Time of irrigation (min) / Discharge of dripper (lph)	Anonymous (2018)
(2)	Crop evapotranspiration (ETc)	Kc * ETo	Singh et al. (2018)
(3)	Water balance	(Peff + I) – (R + D +ΔS - Fx dt)	Dar et al. (2017)
(4)	Reference evapotranspiration (ETo)	$\frac{[0.408 * \Delta * (Rn - G) + \gamma * (900 / (T + 273)) * u2 * (es - ea)]}{[\Delta + \gamma * (1 + 0.34 * u2)]}$	Allen et al. (1998)
(5)	Soil moisture changes (ΔS)	$\frac{(M_2 - M_1) \times \rho d \times d_1}{100}$	Dar et al. (2017)
(6)	Effective precipitation (P_{eff})	P * (125 - 0.2 * P) / 125 for P <= 250 mm; 125 + 0.1 * P for P > 250 mm	USDA-SCS (1993)
(7)	Green (P_{eff}) water footprint (WF_{green})	$\frac{10 * \{Rainfall - (DP + RO)\}}{Yield}$	Luan et al. (2018)
(8)	Blue (irrigation) water footprint (WF_{blue})	$\frac{10 * \{IR - (DP + RO)\}}{Yield}$	Filmon et al. (2021)
(9)	Grey water footprint (WF_{grey})	(1000×α×AR×A) / (Y(Cmax–Cnat))	
(10)	Total water footprint (WF_{total})	$WF_{blue}+WF_{green}+WF_{grey}$	

Where, ETc = crop evapotranspiration (mm day^{-1}); Kc crop coefficient; ETo = reference evapotranspiration (mm day^{-1}); Rn = net radiation at the crop surface (MJ m^{-2} day^{-1}); G = soil heat flux density (MJ m^{-2} day^{-1}); Ta = mean daily air temperature (°C); u2 = wind speed at 2 m height (m s^{-1}); es = saturation vapor pressure (kPa); ea = actual vapor pressure (kPa); es - ea is saturation vapor pressure deficit (kPa); Δ = slope of saturation vapor pressure curve (kPa °C^{-1}); γ = psychrometric constant (kPa °C^{-1}); 900 = the conversion factor; P = rainfall (mm); ETc = crop evapotranspiration (mm.day-1); Peff = effective rainfall (mm); I = irrigation (mm); ΔS = soil moisture changes in storage (mm); D = soil water drainage (mm); Fx = vertical flux (mm.day-1); dt = change in time duration. The factor 10 is the conversion coefficient, from millimeter (mm) to cubic meter per hectare (m^3 ha-1); A is the cropped area (ha); AR = area of nitrogen fertilizer usage (kg ha-1); α = leaching factor (%), which is set at 10%; Cmax = maximum concentration of nitrogen for a given water body (mg l^{-1}), which is set at 10; Cnat = natural background on concentration of nitrogen (mg l^{-1}), which is set at 0.

Results

Case Study 1: Water footprint assessment of surface and subsurface drip fertigated cotton-wheat cropping system – A case study under semi-arid environments of Indian Punjab

Journal of Cleaner Production 365 (2022) 132735

Contents lists available at ScienceDirect

Journal of Cleaner Production

journal homepage: www.elsevier.com/locate/jclepro

Water footprint assessment of surface and subsurface drip fertigated cotton-wheat cropping system – A case study under semi-arid environments of Indian Punjab

Kulvir Singh [a,*], Sudhir Kumar Mishra [b], Manpreet Singh [c], Kuldeep Singh [d], Ajmer Singh Brar [e]

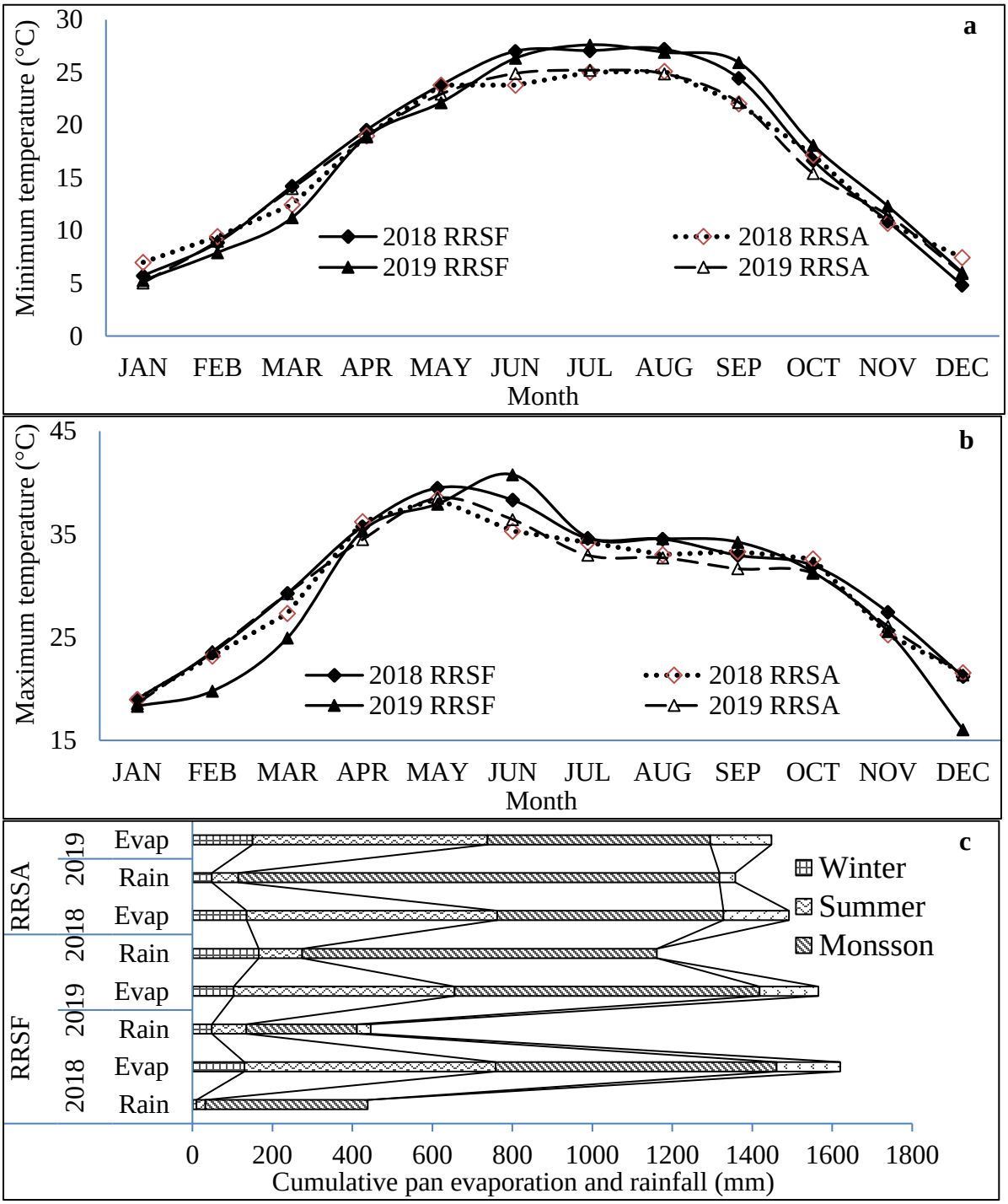


Fig.1: Spatial variation in weather conditions; a: Minimum temperature; b: Maximum temperature; c: Cumulative pan evaporation and rainfall (mm)

RRSA: Regional Research Station, Abohar; RRSF: Regional Research Station, Faridkot

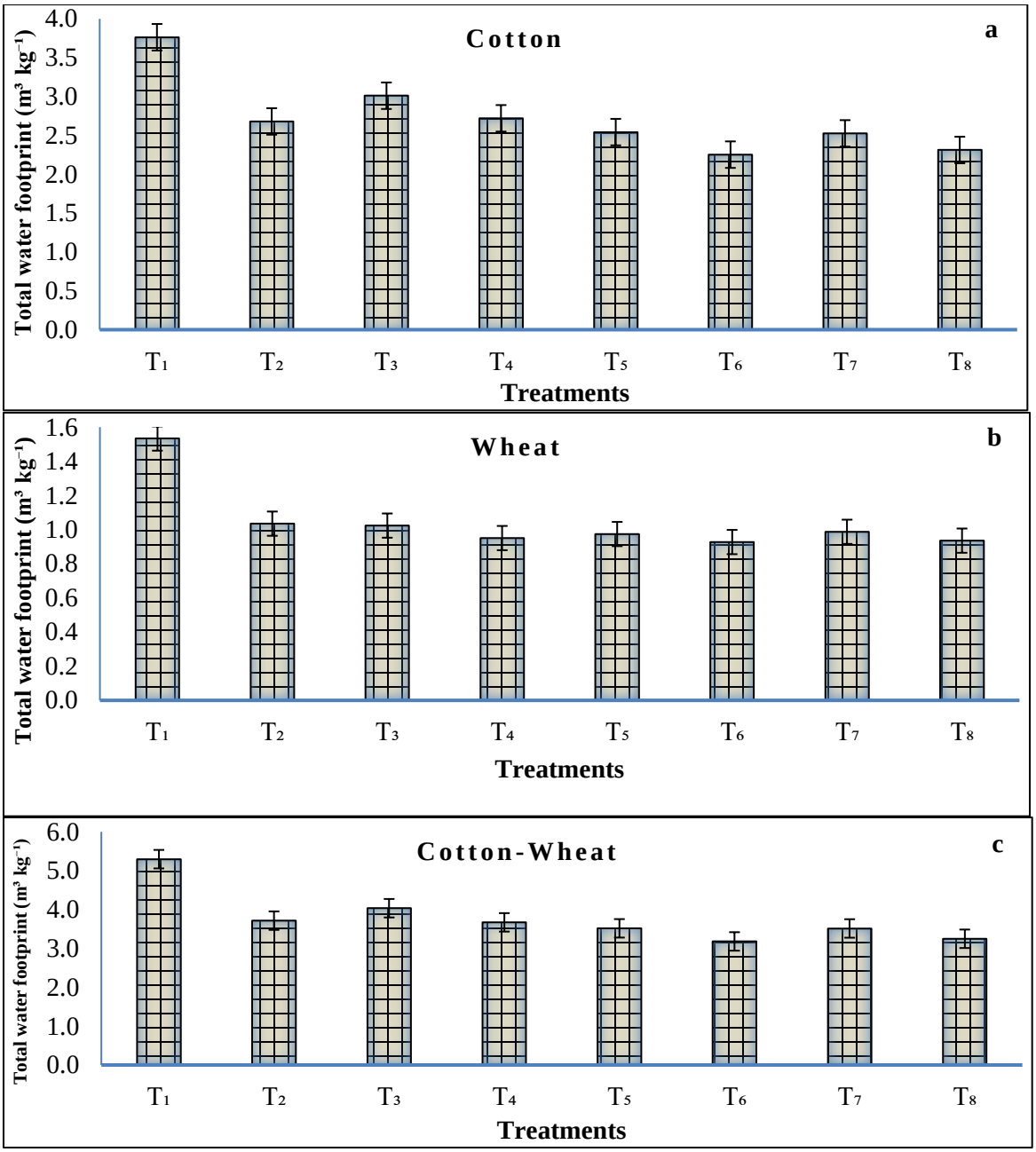


Fig.2: Total water footprint of cotton (a), wheat (b) and cotton-wheat cropping system (c)

T_1: SF with 100 % RDN; T_2: SD at 80% ETc with 100% RDN; T_3: SSDF at 60% ETc and 75% RDN; T_4: SSDF at 60% ETc and 100% RDN; T_5: SSDF at 80% ETc and 75% RDN; T_6: SSDF at 80% ETc and 100% RDN; T_7: SSDF at 100% ETc and 75% RDN; T_8: SSDF at 100% ETc and 100% RDN.

Note- SD: surface flood; SDF: surface drip fertigation; SSDF: subsurface drip fertigation; ETc: Crop evapotranspiration: RDN: recommended dose of nitrogen

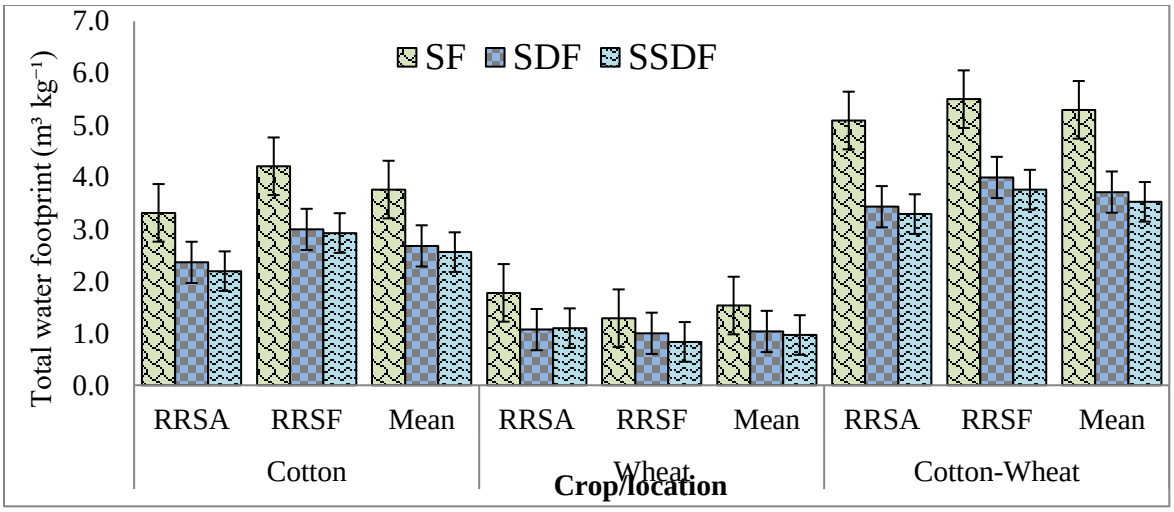


Fig.3: Spatial variations in total water footprint of CWCS under different irrigation/fertigation levels SF: surface flood irrigation; SDF: surface drip fertigation; SSDF: subsurface drip fertigation; RRSA: Regional Research Station, Abohar; RRSF: Regional Research Station, Faridkot

Case Study 2:

Water footprint of wheat under different irrigation practices at Faridkot, Punjab

(Choudhary et al. 2025)

Journal of Agrometeorology

ISSN : 0972-1665 (print), 2583-2980 (online)
Vol. No. 27 (1) : 22-26 (March - 2025)
https://doi.org/10.54386/jam.v27i1.2844
https://journal.agrimetassociation.org/index.php/jam

Research Paper

Water footprint of wheat under different irrigation practices at Faridkot, Punjab

SOURAV CHOUDHARY[1], SUDHIR KUMAR MISHRA[1]*, KULVIR SINGH[1], R. K. PAL[2] and PRABHJOT-KAUR[2]

[1]Punjab Agricultural University, Regional Research Station, Faridkot, Punjab
[2]Department of Climate Change and Agricultural meteorology, Punjab Agricultural University, Ludhiana
**Corresponding author Email: sudhirmet@pau.edu*

ABSTRACT

Field experiments were conducted during *Rabi* seasons at Punjab Agricultural University, Regional Research Station, Faridkot, Punjab for 13 years (2010-11 to 2022-23) to assess the water footprint (WF) of wheat crop irrigated through different methods such as conventional surface flood (SF) during 2010-11 to 2018-19, surface drip (SD) during 2019-20 to 2020-21, and subsurface drip (SSD) during 2021-22 to 2022-23. Results elucidated that quantity of the irrigation water applied to the wheat crop ranged between 209 and 375 mm in different years. Whereas, wheat yield ranged from 3450 kg ha^{-1} (2017-18) to 5471 kg ha^{-1} (2021-22). Wheat crop under SF irrigation recorded higher WF_{total} 0.98 to 1.57 m^3 kg^{-1}. The maximum rainfall 250.3 mm received in 2014-15 resulted highest WF_{green} (0.46 m^3 kg^{-1}) and lowest WF_{blue} (0.45 m^3 kg^{-1}). The wheat cultivation under SD and SSD reduced the WF_{grey} up to 35 % and WF_{blue} up to 35.0 – 42.8 % over SF. The higher crop yield and/or fewer water consumption both are associated with the lower WF. Therefore, for hydrological resource conservation and to ensure environmental sustainability, irrigation through SSD and SD should be promoted over the traditional SF method among the farming community.

Keywords: Irrigation, Surface drip, Sub-surface drip, Water footprint, Wheat.

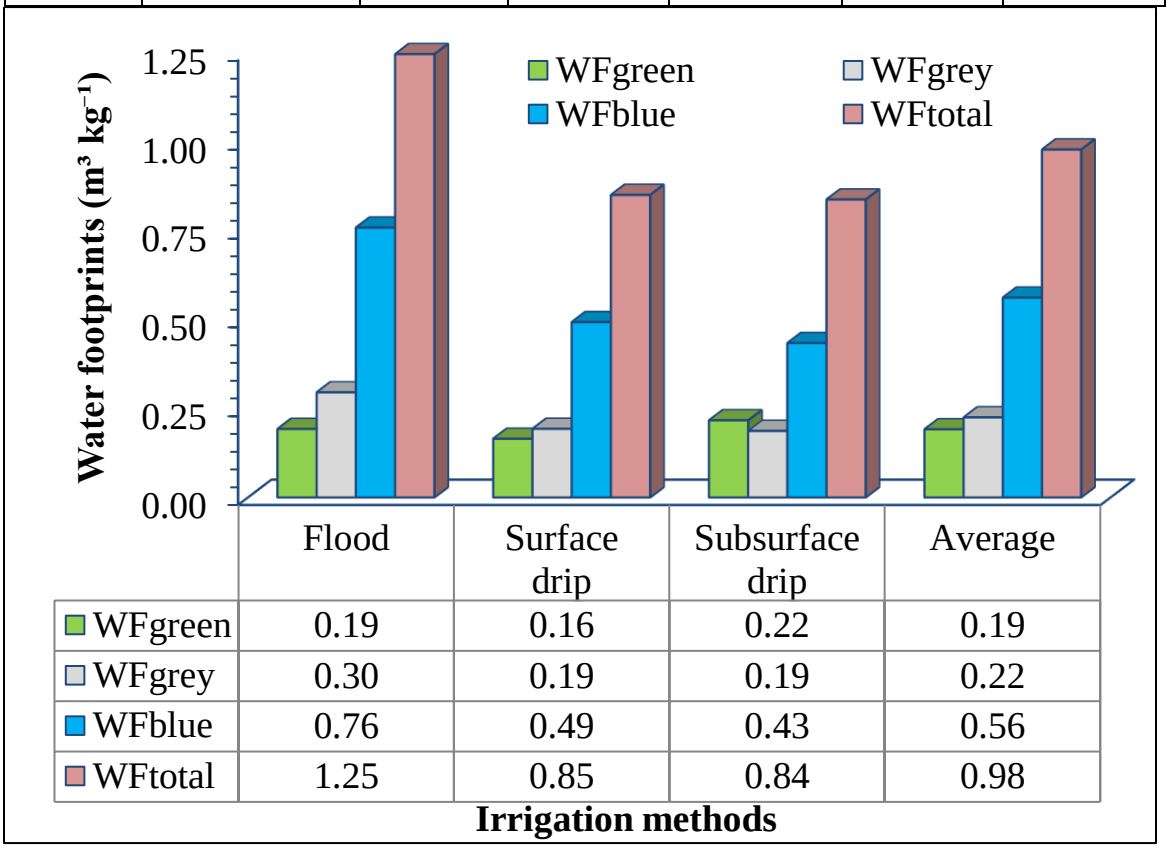
Table 4: Weather conditions and effective rainfall during wheat growing season at Faridkot

Year	Temperature (°C)		Relative humidity (%)		Rainfall (mm)	
	Maximum	Minimum	Maximum	Minimum	Total	Effective
Conventional flood irrigation						
2010-11	24.9	11.5	79	45	57.9	57.2
2011-12	24.2	10.8	74	49	15.0	14.8
2012-13	25.2	11.8	73	39	113.0	110.7
2013-14	24.2	10.9	80	53	104.9	103.0
2014-15	24.3	11.0	84	51	250.3	230.2
2015-16	25.8	11.5	84	43	104.0	100.7
2016-17	26.3	11.3	85	41	56.3	55.2
2017-18	25.8	10.9	86	42	41.6	41.2
2018-19	24.6	9.9	86	43	65.3	64.6
Mean	**25.0**	**11.1**	**81**	**45**	**89.8**	**86.4**
Surface drip irrigation						
2019-20	22.8	10.4	88	51	146.1	143.6
2020-21	25.2	10.5	84	42	27.4	27.2
Mean	**24.0**	**10.5**	**86**	**47**	**86.8**	**85.4**
Subsurface drip irrigation						
2021-22	26.2	10.9	84	43	125.0	118.3
2022-23	25.0	10.9	86	44	118.6	113.6
Mean	**25.6**	**10.9**	**85**	**44**	**121.8**	**115.9**

Table 5: Amount of irrigation, grain yield and water footprints of wheat during 2010-11 to 2022-23

	Amount of irrigation (mm)	Yield (kg ha^{-1})	WF_{green} (m^3 kg^{-1})	WF_{blue} (m^3 kg^{-1})	WF_{grey} (m^3 kg^{-1})	WF_{total} (m^3 kg^{-1})
	Conventional flood irrigation					
2010-11	375	3619	0.16	1.04	0.35	1.54
2011-12	375	4279	0.03	0.88	0.29	1.20
2012-13	225	4678	0.24	0.48	0.27	0.98
2013-14	300	5039	0.20	0.60	0.25	1.05
2014-15	225	5046	0.46	0.45	0.25	1.15
2015-16	300	4517	0.22	0.66	0.28	1.16
2016-17	375	4797	0.12	0.78	0.26	1.16
2017-18	375	3450	0.12	1.09	0.36	1.57
2018-19	300	3496	0.18	0.86	0.36	1.40
	Surface drip irrigation					
2019-20	212	5173	0.28	0.41	0.19	0.88

2020-21	299	5205	0.05	0.57	0.19	0.82
	Subsurface drip irrigation					
2021-22	256	5471	0.22	0.47	0.18	0.87
2022-23	209	5234	0.22	0.40	0.19	0.81



	Flood	Surface drip	Subsurface drip	Average
WFgreen	0.19	0.16	0.22	0.19
WFgrey	0.30	0.19	0.19	0.22
WFblue	0.76	0.49	0.43	0.56
WFtotal	1.25	0.85	0.84	0.98

Fig. 4: Water footprint of wheat under different irrigation systems

References

Allan J. A. (1997). Virtual Water: A Long-Term Solution for Water Short Middle Eastern Economies? (Vol. 5145). London, UK: School of Oriental and African Studies, University of London.

Allan J. A. (2003). Virtual water-the water, food, and trade nexus. Useful concept or misleading metaphor? *Water Int*., 28(1): 106-113.

Chapagain A. K., Hoekstra A. Y., Savenije H. H. and Gautam R. (2006). The water footprint of cotton consumption: An assessment of the impact of worldwide consumption of cotton products on the water resources in the cotton producing countries. *Ecol Econ*, 60(1), 186-203. doi:10.1016/j.ecolecon.2005.11.027.

Chouchane H., Hoekstra A.Y., Krol M.S. and Mekonnen M.M. (2015). The water footprint of Tunisia from an economic perspective. *Ecol Indic*, 52: 311-319, doi:10.1016/j.ecolind.2014.12.015

Choudhary S., Mishra S. K., Singh K., Pal R. K. and Prabhjot-Kaur (2025). Water footprint of wheat under different irrigation practices at Faridkot, Punjab. *J Agrometeorol*, 27(1): 22–26. doi:10.54386/jam.v27i1.2844

Davis K. F., Rulli M. C., Seveso A. and D'Odorico P. (2017). Increased food production and reduced water use through optimized crop distribution. *Nature Geosci*, 10(12): 919-924. doi:10.1038/s41561-017-0004-5

Fader M., Gerten D., Thammer M., Heinke J., Lotze-Campen H., Lucht W. and Cramer W. (2011). Internal and external green-blue agricultural water footprints of nations, and related water and land savings through trade. *Hydrol Earth Syst Sci*, 15(5): 1641-1660. doi:10.5194/hess-15-1641-2011

Falkenmark M. and Rockström J. (2006). Green Water for Food Security. *IFPRI Water Policy Brief*.

Falkenmark M. and Molden D. (2008). Wake Up to Realities of River Basin Closure. *International J Water Res Dev*, 24: 201–215. doi:10.1080/07900620701723570

Hoekstra A. Y. and Chapagain A. K., (2008). Globalization of Water: *Sharing the Planet's Freshwater Resources*, first ed. Wiley. https://doi.org/10.1002/9780470696224.

Hoekstra A. Y. and Chapagain A. K., Aldaya M. M. and Mekonnen M. M. (2011). The Water Footprint Assessment Manual: Setting the Global Standard (Earthscan, London).

Hoekstra A. Y. and Hung P. Q. (2002). Virtual water trade: A quantification of virtual water flows between nations in relation to international crop trade. *Value of Water Research Report Series No*. 11, UNESCO-IHE.

Hoff H., Falkenmark M., Gerten D., Gordon L., Karlberg L. and Rockström J. (2010). Greening the global water system. *J Hydrol*, 384(3-4): 177-186. doi:10.1016/j.jhydrol.2009.06.026

Jalilov S. M., Amer S. A. and Ward F. A. (2018). Managing the water-energy-food nexus: Opportunities in Central Asia. *J Hydrol*, 557: 407-425. doi:10.1016/j.jhydrol.2017.12.040

Karandish F. and Šimůnek J. (2020). A comparison of the HYDRUS (2D/3D) and SWAP models for simulating water balance and nitrate leaching in a sandy loam soil. *Agricultural Water Management*, 227, 105843.

Mekonnen M. M. and Hoekstra A. Y. (2011). The Green, Blue, and Grey Water Footprint of Crops and Derived Crop Products. *Hydrol Earth Syst Sci*, 15: 1577–1600. doi:10.5194/hess-15-1577-2011

Molden D., Oweis T., Steduto P., Bindraban P., Hanjra M. A. and Kijne J. (2010). Improving Agricultural Water Productivity: Between Optimism and Caution. *Agril Water Manag*, 97: 528–535. doi:10.1016/j.agwat.2009.03.023

Singh K., Mishra S. K., Singh M., Singh K. and Brar A. S. (2022). Water footprint assessment of surface and subsurface drip fertigated cotton-wheat cropping system–A case study under semi-arid environments of Indian Punjab. *J Clean Prod*, *365*: 132735.

Vanham D., Alfieri L. and Feyen L. (2022). National water shortage for low to high environmental flow protection. *Sci. Rep*. 12: 3037

Wada Y. and Bierkens M. F. P. (2014). Global Groundwater Depletion. *Science*, 344: 1241–1244.

Zeng Z., Liu J. and Savenije H. H. G. (2020). A global assessment of agriculture water sustainability. *One Earth*, 2(1): 56–65. https://doi.org/10.1016/j.oneear.2020.02.010

Meta-Analysis: Applications in Climate Resilient Agriculture

Santosha Rathod[1], Gayathri Chitikela[2], R Mahender Kumar[3], CH Padmavathi[3], B Nirmala[4] and S Arun Kumar[3]

[1]ICAR-National Institute of Abiotic Stress Management, Baramati-413115

[2]PJTSAU- College of Agriculture, Jagitiyal

[3]ICAR-IIRR, Hyderabad

[4]ICAR-NAARM, Hyderabad

Email: Santosha.Rathod@icar.org.in

Introduction:

Meta-analysis is a statistical method used to combine the findings from various studies concerning a specific variable. More precisely, it can be characterised as a quantitative or statistical technique for systematically merging results from prior research to draw conclusions about the overall body of work. The term "meta-analysis" was first introduced by Gene V Glass in 1976, who defined it as "the statistical analysis of a large collection of results from individual studies aimed at integrating the findings." He referred to it as "an analysis of analyses." The prefix "meta," derived from Greek, means after or beyond, indicating that meta-analysis extends beyond individual studies. Huque (1988) described it as "A statistical analysis that combines or integrates the results of several independent clinical trials that the analyst deems combinable." The historical progression of meta-analysis started with Karl Pearson, who was likely the earliest medical researcher to apply formal methods for merging data from different studies (1904) in assessing the effectiveness of typhoid vaccination. In 1952, Hans J. Eysenck concluded that psychotherapy had no positive effects, igniting a heated debate that 25 years and numerous studies could not resolve. Then in 1978, Gene V. Glass statistically compiled the results of 375 studies on psychotherapy outcomes, demonstrating that Hans J. Eysenck was wrong, and he coined the term "meta-analysis." In the context of climate-resilient agriculture, meta-analysis is an essential tool for assessing the effectiveness of various adaptation and mitigation strategies, including drought-resistant crop varieties, soil moisture conservation practices, and climate-smart resource management. By synthesizing empirical evidence from diverse agroecological zones and socio-economic situations, meta-analysis offers a thorough understanding of which interventions are most effective, the conditions under which they work best, and for which types of farming systems. This evidence-based

synthesis is vital for informing policy choices, directing research funding, and developing customized solutions to enhance agricultural resilience amid increasing climate variability and extreme weather conditions. Now a days, meta-analysis is also being employed in agricultural research to determine effect sizes using statistical estimation methods. At times, meta-analysis is mistaken for aggregation or systematic reviews, which typically present qualitative representative results from a collection of studies.

Purpose of Meta-Analysis:

- It is utilized when researchers have several studies that assess the same or similar hypotheses.
- It proves beneficial when researchers encounter numerous conflicting studies and intricate literature.
- To pinpoint heterogeneity in effects across different studies.
- To enhance statistical power and precision in detecting a sample effect size.
- To create, fine-tune, and evaluate suitable hypotheses.
- To minimize the subjectivity in study comparisons through robust statistical methods.
- To examine gaps in the existing knowledge base and propose directions for future research.

Initially, meta-analysis was mainly prevalent in medical and psychological studies, but it later extended to all fields requiring combined or aggregated results. A basic aggregation of effects from various studies or experiments tends to yield inaccurate and poor results, as it overlooks randomness, heterogeneity, and the significance of the studies in comparison criteria. In contrast, meta-analysis provides robust results, presenting a representative effect size with a confidence interval based on statistical methods.

Steps in Meta-analysis:

1. Formulate a research question based on a theoretical framework.
2. Conduct a literature search using PubMed, Google Scholar, and other relevant sources.
3. Review the titles and abstracts of the selected research papers.

4. Extract key information from the chosen final articles.
5. Evaluate the quality of the information in these articles by assessing their internal validity and applying the GRADE criteria.
6. Assess the degree of heterogeneity among the articles.
7. Calculate the overall effect size represented as an Odds Ratio using both fixed and random effects models, and create a forest plot.
8. Investigate the presence of publication bias among the articles and perform a funnel plot along with sensitivity analysis.
9. Carry out subgroup analyses and meta-regression to identify if certain subsets of research reflect the summary effects.

Effect size

The effect size in meta-analysis is a statistical measure that indicates the strength and direction of the relationship among variables. The definition of effect size varies and can differ across fields. In medical research, effect size is often referred to as the application effect and is quantified using the odds ratio, risk ratio or risk difference. In the social sciences, the term 'effect size' is commonly used, though it may sometimes be represented as the standardized mean difference or various relationships. The preferred method for calculating effect size will vary based on the study's objectives, design and data format. Generally, the most frequently used effect sizes fall into three categories: proportions, averages and correlational effects.

Effect sizes based on means:

The calculation of effect sizes is based on the raw differences between sample observations taken from the same measurement scale. Typically, for samples derived from continuous variables, raw mean differences are utilised. Consider a study that presents the means for two groups (treatment and control); if we wish to compare the means of these two groups, let μ_1 and μ_2 represent the true population means of the two groups. The population mean difference is expressed as

$$\Delta = \mu_1 - \mu_2$$

Let $\bar{x}_1$ and $\bar{x}_2$ denote the sample means from the two independent groups. The sample estimate of D is simply the difference between the means of the two treatment groups, given by

$$D = \bar{x}_1 - \bar{x}_2$$

Let S_1 and S_2 represent the sample standard deviations of the two groups, with n_1 and n_2 indicating the sample sizes of these groups. If we assume that the two population standard deviations are equal (which is commonly assumed in most parametric data analysis methods), then we have $\sigma_1 = \sigma_2 = \sigma$ and the variance of D can be defined as

$$V_D = \frac{n_1+n_2}{n_1 n_2} s^2_{pooled}$$

where, $s^2_{pooled} = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$

If we assume that the population standard deviations differ, then the variance of D is defined as

$$V_D = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

In both situations, the standard error of D is calculated as the square root of V,

$$SE_D = \sqrt{V_D}$$

Computing D from using paired or matched groups or pre-post scores

In matched group designs, the value of D can be estimated using the following equations: the sample estimate of Δ is simply the difference in sample means, denoted as D. When we have the difference score for every pair, we can find the mean difference ($\bar{x}_{diff}$) and the standard deviation of these differences (S_{diff}). Thus, we have:

$$D = \bar{x}_{diff}$$

$$V_D = \frac{s^2_{diff}}{n}$$

Where n represents the number of pairs, and $SE_D = \sqrt{V_D}$. However, to compute the standard deviation of difference scores, it is necessary to derive it from the standard deviations of each sibling's scores in matched pairs:

$$S_{diff} = \sqrt{s_1^2 + s_2^2 - 2 \times r \times s_1 x s_2}$$

Where r denotes the correlation between 'siblings' in matched pairs. If $S_1 = S_2$ then S_{diff} simplifies to

$$S_{diff} = \sqrt{2 \times s_{pooled}^2 (1 - r)}$$

As r approaches 1.0, the standard error of the paired difference will diminish, and when r equals 0, the standard error of the difference reverts to that of two independent groups, each consisting of sample size n.

Standardized mean difference

The raw mean difference serves as a valuable measure when the scales of measurement are identical; however, if the studies being examined use different measurement scales, it becomes necessary to standardize the differences. This approach aligns with Cohen's recommendations (1969, 1987) for characterizing the magnitude of effects in statistical analysis.

The standard mean difference can be viewed as comparable across studies through either of two arguments (Hedges and Olkin, 1985).

Let μ_1 and σ_1 represent the true (population) mean and standard deviation for the first group, while μ_2 and σ_2 represent the true (population) mean and standard deviation for the second group. When the standard deviations of both populations are equal, such that $\sigma_1 = \sigma_2 = \sigma$, then the parameter for the standard mean difference, or the population standardized mean difference, is expressed as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

Computing d from studies that used independent groups

The standardized mean difference (δ) can be estimated from studies that utilized two independent groups as $d = \frac{\bar{x}_1 - \bar{x}_2}{s_{within}}$

In this formula, $\bar{x}_1 and\ \bar{x}_2$ represent the sample means for the respective groups, while S_{within} denotes the pooled within-groups standard deviation.

$$s_{within} = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

Where n_1 and n_2 indicate the sample sizes of the two groups, and S_1 and S_2 are the standard deviations of the two groups. In research synthesis, the sample estimate of the standardized mean difference is commonly referred to as Cohen's d, with δ symbolizing the effect size parameter and d representing the sample estimate of that parameter.

The variance of d can be approximated as follows:

$$V_d = \frac{n_1+n_2}{n_1 n_2} + \frac{\mathrm{d}^2}{2(n_1+n_2}$$

The standard error of d is calculated as the square root of V_d

$$SE_d = \sqrt{V_d}$$

Computing d from pre and post scores and matched scores

We can calculate the standardized mean difference (δ) from studies utilizing matched groups or pre-post scores within a single group, where the sample estimate of d is represented as

$$d = \frac{\bar{y}_{diff}}{s_{within}} = \frac{\overline{y_1} - \overline{y_2}}{s_{within}}$$

For matched studies, the standard deviation within groups can be derived from the standard deviation of the difference by using

$$s_{within} = \frac{s_{diff}}{\sqrt{2(1-r)}}$$

where r represents the correlation between the paired observations (for instance, the correlation between pretest and posttest scores). The variance of d is expressed as

$$V_d = \frac{1}{n} + \frac{d^2}{2n} 2(1\text{-}r)$$

Where n represents the number of pairs. The standard error of d is calculated as the square root of V_d,

$$SE_d = \sqrt{V_d}$$

Estimation of D from dichotomous experiments

The calculation of effect size for dichotomous results depends on whether a certain phenomenon was detected or not. The most commonly used measures of effect size are the odds ratio (OR), the risk ratio (RR), and the risk difference (RD). The odds ratio represents the comparison of the likelihood of an event occurring (Littel *et al.*, 2008). In other words, the effect size is derived from the ratio of two potential outcomes (Borenstein *et al.*, 2009). The risk ratio, akin to the odds ratio, focuses on risk and is the ratio of one risk to another. The risk difference, on the other hand, is calculated as the disparity between two risks. The effect size determined by the odds ratio or the risk ratio is obtained by transforming the data into logarithmic values, while the risk difference employs raw data for its calculation.

Odds ratio:

The odds ratio for the study Ψ_k, k; refers to the ratio of the odds of an event occurring in the experimental group compared to the control group.

$$\Psi_k = \frac{\left(\frac{P_{ek}}{1-P_{ek}}\right)}{\left(\frac{P_{Ck}}{1-P_{ck}}\right)} = \frac{P_{ek}(1-P_{Ck})}{P_{ck}(1-P_{ck})}$$

If either of the two estimated event probabilities is zero, the log odds ratio, denoted as log Ψ_k, results in either -∞ or +∞ . If both probabilities are zero, the log odds ratio cannot be defined. The estimated odds ratio for study k is given by

$$\hat{\psi}_k = \frac{a_k d_k}{b_k c_k}$$

Due to the skewed distribution of this estimator with typical sample sizes, effect estimates, standard errors and confidence intervals are generally computed using the natural logarithm of $\hat{\psi}_k$ which we denote as log ($\hat{\psi}_k$). The final outcomes are then back-transformed to the original scale for presentation purposes. For typical sample sizes, the variance of the natural logarithm of the odds ratio can be accurately estimated by

$$\widehat{Var}(log\ \hat{\psi}_k) = \frac{1}{a_k} + \frac{1}{b_k} + \frac{1}{c_k} + \frac{1}{d_k}$$

where the accuracy of approximation increases as n_{ek} and n_{ck} increase. The estimated variance for this is expressed as;

$$\exp\ (\log \hat{\psi}_k \pm z_{1-\frac{\alpha}{2}}\ \text{S.E}\ (\log\hat{\psi}_k))$$

where the standard error S.E. (log $\hat{\psi}_k$) = $\sqrt{\widehat{Var}(log\ \hat{\psi}_k)}$ and $z_{1-\frac{\alpha}{2}}$ represents the 1- α/2 quantile from the standard normal distribution.

Risk Ratio

The risk ratio ϕ_k, commonly referred to as relative risk, is determined by the ratio of the probabilities of two events,

$$\phi_k = \left(\frac{P_{ek}}{P_{ck}}\right)$$

If either of the two estimated event probabilities is zero, the log risk ratio, log ϕ_k, results in either -∞ or +∞ . In cases where both probabilities are zero, the log risk ratio remains undefined. The risk ratio is estimated using the formula

$$\hat{\phi}_k = \frac{\left(\frac{a_k}{a_k+b_k}\right)}{\left(\frac{c_k}{c_k+d_k}\right)}$$

The variance of this estimate is approximated by

$$\widehat{Var}(log\ \hat{\phi}_k) = \frac{1}{a_k} + \frac{1}{c_k} - \frac{1}{a_k + c_k} - \frac{1}{c_k + d_k}$$

As with the log odds ratio, this approximation holds well for typical trials and improves as the number of patients, n_{ek} and n_{ck}, increases. A confidence interval for $\hat{\phi}_k$ can be expressed as

$$\exp(\log(\hat{\phi}_k) \pm z_{1-\frac{\alpha}{2}} \text{S.E}(\log\hat{\phi}_k))$$

Risk difference:

The risk difference, denoted as η_k, is calculated as the difference between the probabilities of two events, represented as

$$\eta_k = P_{ek} - P_{Ck}$$

The risk difference is consistently well-defined and ranges from -1 to 1. The natural estimate of the risk difference is

$$\hat{\eta}_k = \frac{a_k}{a_k + b_k} - \frac{c_k}{c_k + d_k}$$

with the corresponding variance estimate given by

$$\widehat{Var}(\hat{\eta}_k) = \frac{a_k b_k}{(a_k + b_k)3} + \frac{c_k d_k}{(c_k + d_k)3}$$

Consequently, a two-sided approximate (1-α) confidence interval for the risk difference can be expressed as $\hat{\eta}_k = \pm z_{1-\frac{\alpha}{2}} S.E(\hat{\eta}_k)$

With standard error $\text{S.E}(\hat{\eta}_k) = \sqrt{\widehat{Var}(\hat{\eta}_k)}$ and $z_{1-\frac{\alpha}{2}}$ representing the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution.

Models in meta-analysis

Fixed effect model

The fixed effect model assumes that the true effect size is identical across all studies, and any variation in effect size among studies is due solely to random error. Consequently, when assigning weights to different studies, we can primarily disregard information from smaller studies, as we have more reliable information about the effect size from larger

studies. Let k = 1,…,K denote the index for each study, $\hat{\theta}_k$ represent the intervention effect estimate for study k, and θ represents the intervention effect in the population that we aim to estimate. Let $\hat{\sigma}_k^2$ denote the sample estimate of Var ($\hat{\theta}_k$). The fixed effect model can be expressed as

$$\hat{\theta}_k = \theta + \sigma_k \varepsilon_k \ , \ \varepsilon_k \sim N(0,1)$$

Now consider the fixed effect estimate of θ, represented as $\hat{\theta}_F$. Given estimates ($\hat{\theta}_k, \hat{\theta}_F$), K-1,2,…, K , the maximum-likelihood estimate under this model is given by

$$\hat{\theta}_F = \frac{\sum_{k=1}^{k} \frac{\hat{\theta}_k}{\hat{\sigma} k^2}}{\sum_{k=1}^{k} 1/\hat{\sigma}_k^2} = \frac{\sum_{k=1}^{k} W_k \hat{\theta}_k}{\sum_{k=1}^{k} W_k}$$

Accordingly, $\hat{\theta}_F$ represents a weighted average of the individual effect estimates $\hat{\theta}_k$ where the weights are defined as $w_k = 1/\hat{\sigma}_k^2$. As a result, this approach is referred to as the inverse variance method. The variance of $\hat{\theta}_F$ can be approximated by

$$\widehat{Var}(\hat{\theta}_F) = \frac{1}{\sum_{k=1}^{k} W_k}$$

A (1-α) confidence interval for $\hat{\theta}_F$ can be determined using $\hat{\theta}_F = \pm z_{1-\frac{\alpha}{2}} S.E(\hat{\theta}_F)$, with standard error being S.E $(\hat{\theta}_F) = \sqrt{\widehat{Var}(\hat{\theta}_F)}$ and $z_{1-\frac{\alpha}{2}}$ denoting the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution. A corresponding test for an overall treatment effect can be formulated by utilising $\hat{\theta}_F /$ S. E $(\hat{\theta}_F)$ as test statistic.

Random effect model:

The random effects model assumes that different studies originate from populations that may vary in ways that could influence the treatment effect. The first type of variation is random error within studies, as seen in the fixed effect model. The second type is true variation in effect size across different studies. The random effects model aims to

account for the fact that the study effect estimates $\hat{\theta}_K$ are often more variable than assumed in the fixed effect model.

Under the random effects model,

$$\hat{\theta}_k = \theta + u_k + \sigma_k \varepsilon_k,\ \varepsilon_k \sim N(0,1);\ u_k \sim N(0,1)$$

where the u's and ε's are independent. The primary distinction from the fixed effect model is the question of whether the random effects model is appropriate. Several authors have suggested that, since smaller studies are more prone to bias, the fixed effect estimate is generally more desirable. The MLE function is

$$Q = \sum_{k=1}^{k} W_k \left(\hat{\theta}_k - \hat{\theta}_F\right)^2$$

Wk = 1/ $\hat{\sigma}_k^2$ represents the weighted sum of squares regarding the estimates of fixed effects. This is usually referred to as either the homogeneity test statistic or the heterogeneity statistic. Next define

$$S = \sum_{k=1}^{k} W_k - \frac{\sum_{k=1}^{k} W_k^2}{\sum_{k=1}^{k} W_k}$$

If Q < (K-1), then τ^2 is set to 0 and the random effects estimate $\hat{\theta}_R$ is made equal to the fixed effect estimate $\hat{\theta}_F$. Otherwise, the DerSimonian-Laird estimator of the between-study variance is given by

$$\hat{\tau}^2 = \frac{Q - (k-1)}{S}$$

And the random effects estimate and its variance can be calculated as

$$\hat{\theta}_R = \frac{\sum_{k=1}^{k} W_k^* \hat{\theta}_R}{\sum_{k=1}^{k} W_k^*}$$

$$\widehat{Var}\left(\hat{\theta}_R\right) = \frac{1}{\sum_{k=1}^{k} W_k^*}$$

With weights W_k^*= 1/ ($\hat{\sigma}_k^2 + \hat{\tau}^2$). The random effects estimator $\hat{\theta}_R$ is a weighted average of the individual effect estimates $\hat{\theta}_K$ with weights 1/ ($\hat{\sigma}_k^2 + \hat{\tau}^2$). Consequently, this approach is frequently referred to as the inverse variance method. A (1-α) confidence interval for $\hat{\theta}_R$ can be computed by using

$$\hat{\theta}_R = \pm z_{1-\frac{\alpha}{2}} S.E\left(\hat{\theta}_R\right)$$

With standard error S.E $(\hat{\theta}_R) = \sqrt{\widehat{Var}\left(\hat{\theta}_R\right)}$ and $z_{1-\frac{\alpha}{2}}$ denoting the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution. A corresponding test for an overall treatment effect can be constructed using $\hat{\theta}_R /$ S. E $(\hat{\theta}_R)$ as test statistic.

Tests and measures of heterogeneity:

The degree of heterogeneity is quantified in terms of Q, expressed as $Q = \sum_{k=1}^{k} W_k\left(\hat{\theta}_k - \hat{\theta}_F\right)^2$, which represents the weighted sum of squares about the fixed effect estimate $\hat{\theta}_F$. Higher values of Q suggest increased heterogeneity among the individual studies in a meta-analysis, as well as higher values of the between-study heterogeneity τ^2. Under the null hypothesis that τ^2=0,

$$Q \sim \chi^2_{k-1}$$

The test statistic against the null hypothesis is calculated as

$$H^2 = \frac{Q}{k-1}$$

$$I^2 = \begin{cases} (\mathrm{H2} - 1)/\ \mathrm{H2\ if\ Q} > (\mathrm{K} - 1) \\ 0, otherwise \end{cases}$$

Under the null hypothesis that τ^2= 0; Q has mean K -1; so H^2 has mean 1; again, large values of H^2 indicate greater heterogeneity. I^2 is a scaled version of H^2; lying between 0 and 1 (or 0 % and 100 %).

Funnel plot

The funnel plot is used to depict the publication bias in the meta-analysis, which shows the estimated treatment effects on the x-axis against a measure of their precision, on the y-axis, typically represented by the standard error, with the standard error positioned at the top,

i.e. an inverted axis. For the odds ratio the coordinates x_k and y_k of the funnel plot are defined as

$$x_k = log\,\hat{\psi}_K$$

$$Y_k = \text{S.E}\,(log\,\hat{\psi}_K)$$

Occasionally, the y-axis may represent the inverse standard error, inverse variance, or some function of sample size rather than the standard error.

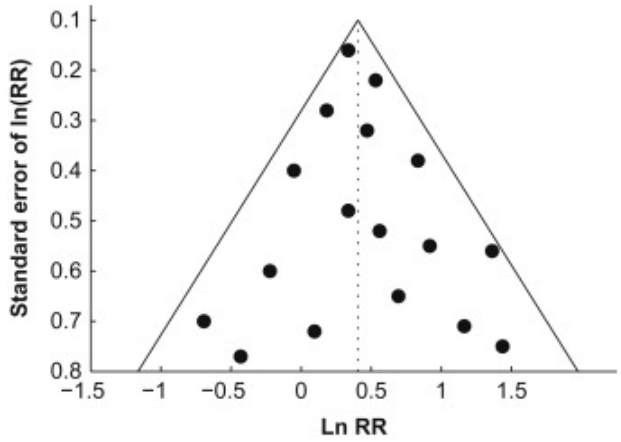


Fig 1: Ideal funnel plot

In the absence of small-study effects, treatment effects of large and small studies are dispersed around a common average treatment effect. If no excessive between-study heterogeneity exists, smaller studies (with larger standard errors) would scatter more than larger studies. That is, the funnel plot would resemble a symmetric triangle regarding the average treatment effect, with wide variability for small, imprecise studies at the bottom of the plot and minimal dispersion for large, precise studies at the top. An asymmetric funnel plot indicates the influence of small-study effects. The ideal funnel plot with 25 studies is shown in figure 1.

Sub group analysis:

Subgroup analyses are conducted to understand the treatment effect of the other factors or sub group treatment effects that influence the effect size of other treatments. The factor that defines the subgroups is said to be an effect moderator and it is essential to calculate

the treatment-subgroup interaction for the subgroup analysis. i.e. whether the treatment effect is modified, or moderated, by subgroup membership. The subgroup analysis will also help in identifying heterogeneity between studies.

Meta Regression:

Tests for differences among subgroups are based on a single covariate with a limited number of values, specifically a binary or categorical covariate. Meta-regression expands on this by allowing for either multiple binary/categorical covariates or a continuous covariate. In the case of meta regression, a subgroup analysis could result in subgroups of size one, i.e. each covariate value generates a subgroup. Let us consider the following meta-regression model

$$\hat{\theta}_k = \theta + \beta_1 x_{1k} + \cdots + \beta_p x_{pk} + u_k + \sigma_k \varepsilon_k, \varepsilon_k \sim N(0,1); u_k \sim N(0, \tau^2)$$

with k = 1, …, K and independent error terms u and ε. As this model has both fixed effect (βs) and random effects terms (u_k with variance τ^2), it is classified as called a mixed effects model. The fixed effect meta-regression represents a special case of the mixed effects model when the between-study variance $\tau^2=0$. Subgroup analyses involving more than two subgroups are essentially a meta-regression with a categorical predictor. For meta-regression, these subgroups are then dummy-coded.

Forest Plots:

The forest plot helps in presenting summary effect visualized with a point estimate bounded by its confidence interval. It indicates whether the overall effect is based on many studies or only a few. Additionally, it indicates whether these studies are precise or imprecise and reveals whether the effects reported across all studies are consistent or significantly different from one another. The forest plot is useful for verifying the proper interpretation of an entire meta-analysis and for highlighting anomalies, such as outliers, that need to be addressed. It represents a compilation of meta-analysis findings in a format that is easy to understand, even for those who do not understand the complex statistics involved in meta-analysis.An example of a forest plot is seen in figure 2.

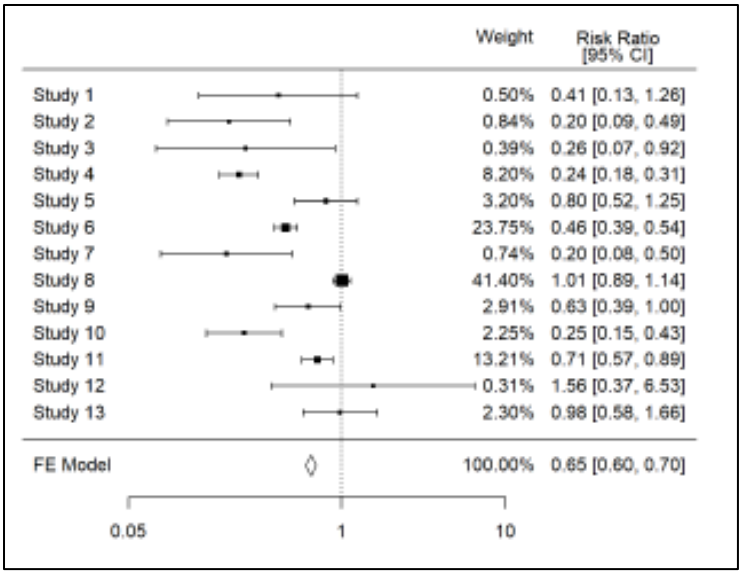


Fig.2: Forest plot

Galbraith plot:

The Galbraith plot, also referred to as Galbraith's radial plot or simply a radial plot, is a tool used to display several estimates of the same quantity that have different standard errors. It helps assess heterogeneity in a meta-analysis and can serve as either an alternative or a complement to a forest plot. To create a Galbraith plot, standardised estimates or z-statistics are calculated by dividing each estimate by its standard error (SE). This results in a scatter plot with z-statistics on the vertical axis and the inverse of the standard error on the horizontal axis. A larger radius in the Galbraith plot indicates smaller variance. Figure 3 provides an example of a radial plot.

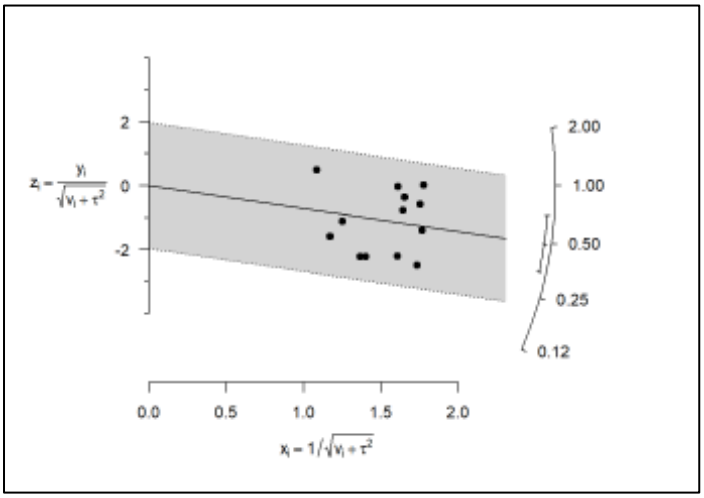


Fig.3: Radial plot

Conclusion:

In various research fields, a wide range of datasets has been produced, often utilising the same variables and similar study methods. This makes integrative analytical approaches, such as comparative analysis, feasible and potentially beneficial. However, they can also result in duplicative inferences, which may lead to confusion when generalising the concepts. Meta-analysis improves upon comparative analyses by providing a rigorous methodology for synthesising quantitative research. Although meta-analysis has drawbacks, including the substantial effort and expertise required and the risks of errors and biases from combining studies, these challenges can generally be addressed through careful research, planning and interpretation. Effect sizes are the most commonly used summary statistic in meta-analyses. However, Robinson and Levin (1997) suggest reporting effect sizes only after achieving statistical significance, but this could result in overestimating those effect sizes. In summary, meta-analysis is a scientific method for aggregating research findings and provides confirmatory results from small and large studies aimed at specific objectives.

R codes to perform meta-analysis:

Source: http://www.metafor-project.org/doku.php/metafor

Code source:

http://www.deeplytrivial.com/2018/04/v-is-for-meta-analysis-variance.html
http://www.deeplytrivial.com/2018/04/w-is-for-meta-analysis-weights.html
http://www.deeplytrivial.com/2018/04/e-is-for-effect-sizes.html

```
##install.packages("metafor")
library(metafor)
smd_meta<-data.frame(
  id = c("005","005","029","031","038","041","041","058","058","067","067"),
  study = c(1,2,3,1,1,1,2,1,2,1,2),
  author_year = c("Ruva 2007","Ruva 2007","Chrzanowski 2006","Studebaker 2000",
                  "Ruva 2008","Bradshaw 2007","Bradshaw 2007","Wilson 1998",
                  "Wilson 1998","Locatelli 2011","Locatelli 2011"),
  n1 = c(138,140,144,21,54,78,92,31,29,90,181),
  n2 = c(138,142,234,21,52,20,18,15,13,29,53),
  m1 = c(5.29,5.05,1.97,5.95,5.07,6.22,5.47,6.13,5.69,4.81,4.83),
  m2 = c(4.08,3.89,2.45,3.67,3.96,5.75,4.89,3.80,3.61,4.61,4.51),
  sd1 = c(1.65,1.50,1.08,1.02,1.65,2.53,2.31,2.51,2.51,1.20,1.19),
  sd2 = c(1.67,1.61,1.22,1.20,1.76,2.17,2.59,2.68,2.78,1.39,1.34)
##install.packages("metafor")
```

```
library(metafor)
smd_meta<-data.frame(
  id = c("005","005","029","031","038","041","041","058","058","067","067"),
  study = c(1,2,3,1,1,1,2,1,2,1,2),
  author_year = c("Ruva 2007","Ruva 2007","Chrzanowski 2006","Studebaker 2000",
                  "Ruva 2008","Bradshaw 2007","Bradshaw 2007","Wilson 1998",
                  "Wilson 1998","Locatelli 2011","Locatelli 2011"),
  n1 = c(138,140,144,21,54,78,92,31,29,90,181),
  n2 = c(138,142,234,21,52,20,18,15,13,29,53),
  m1 = c(5.29,5.05,1.97,5.95,5.07,6.22,5.47,6.13,5.69,4.81,4.83),
  m2 = c(4.08,3.89,2.45,3.67,3.96,5.75,4.89,3.80,3.61,4.61,4.51),
  sd1 = c(1.65,1.50,1.08,1.02,1.65,2.53,2.31,2.51,2.51,1.20,1.19),
  sd2 = c(1.67,1.61,1.22,1.20,1.76,2.17,2.59,2.68,2.78,1.39,1.34)
)
smd_meta <- escalc(measure="SMD", m1i=m1, m2i=m2, sd1i=sd1, sd2i=sd2, n1i=n1, n2i=n2,
                   data=smd_meta) # effect size using smd
or_meta<-data.frame(
  id = c("001","003","005","005","011","016","025","025","035","039","045","064","064"),
  study = c(1,5,1,2,1,1,1,2,1,1,1,1,2),
  author_year = c("Bruschke 1999","Finkelstein 1995","Ruva 2007","Ruva 2007",
                  "Freedman 1996","Keelen 1979","Davis 1986","Davis 1986",
                  "Padawer-Singer 1974","Eimermann 1971","Jacquin 2001",
                  "Ruva 2006","Ruva 2006"),
  tg = c(58,26,67,90,36,37,17,17,47,15,133,68,53),
  cg = c(49,39,22,50,12,33,19,17,33,11,207,29,44),
  tn = c(72,60,138,140,99,120,60,55,60,40,136,87,74),
  cn = c(62,90,138,142,54,120,52,57,60,44,228,83,73)
)
tibble(or_meta)
or_meta <- escalc(measure="OR", ai=tg, bi=(tn-tg), ci=cg, di=(cn-cg), data=or_meta)
smd_meta <- escalc(measure="SMD", m1i=m1, m2i=m2, sd1i=sd1, sd2i=sd2, n1i=n1,
                   n2i=n2,data=smd_meta)
or_meta <- escalc(measure="OR", ai=tg, bi=(tn-tg), ci=cg, di=(cn-cg),
                  data=or_meta)
smd.rma<-rma(yi,vi,method="FE",data=smd_meta)
summary(smd.rma)
smd.rma$zval
smd.rma$pval
or.rma<-rma(yi,vi,method="FE",data=or_meta)
summary(or.rma)
or.rma$zval
exp(or.rma$beta)
forest(smd.rma) ## Forest plot
smd.rma
funnel(smd.rma)  ## Funnel plot
radial(smd.rma)  ## Radial Plot
smd.rma.reml<-rma(yi,vi,method="REML",data=smd_meta)
summary(smd.rma.reml)
funnel(smd.rma.reml)
```

Suggested readings:

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2009). Introduction to meta-analysis. Chichester, U.K: John Wiley & Sons.

Greenland, S.: Invited commentary: a critical look at some popular meta-analytic methods. Am. J. Epidemiol. 140, 290–296 (1994)

Poole, C., Greenland, S.: Random-effects meta-analysis are not always conservative. Am. J. Epidemiol. 150, 469–75 (1999)

DerSimonian, R., Laird, N.: Meta-analysis in clinical trials. Control Clin. Trials 7, 177–188 (1986)

Hedges, L.V., Olkin, I.: Statistical Methods for Meta-Analysis. Academic, San Diego (1985)

Berkey, C.S., et. al. (1995). A random-effects regression model for meta-analysis. *Statistics in Medicine*, 14, 395–411.

Berlin, J.A., Santanna, J., Schmid, C.H., Szczech, L.A., Feldman, H.I.: Individual patientversus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. Stat. Med. 21, 371–387 (2002

Higgins, J.P.T., Thompson, S.G.: Controlling the risk of spurious findings from metaregression. Stat. Med. 23, 1663–1682 (2004)

Thompson, S.G., Higgins, J.P.T.: How should meta-regression analyses be undertaken and interpreted? Stat. Med. 21, 1559–1573 (2002)

Viechtbauer, W.: Conducting meta-analyses in R with the metafor package. J. Stat. Softw. 36(3), 1–48 (2010).

Construction of PCA-based Composite Indices

Ponnaganti Navyasree, Nobin Chandra Paul, Santosha Rathod, K. Ravi Kumar and Prabhat Kumar
ICAR-National Institute of Abiotic Stress Management, Baramati-413115
Email: navyaponnaganti97 @gmail.com

Introduction: What is a composite index?

A **composite index** is defined as a **quantitative or qualitative measure** that combines different dimensions and indicators to provide a single, comprehensive representation of a particular phenomenon or concept. It is fundamentally constructed by **aggregating a set of individual indicators into one summary measure**. The primary purpose of developing composite indices is to **capture complex, multi-dimensional concepts** that cannot be adequately represented by a single indicator. These indices are invaluable tools for understanding and measuring phenomena such as development, sustainability, or performance across various locations or over time.

2. Key Characteristics and Uses of Composite Indices

Composite indices are distinguished by several key characteristics and practical uses:

- **Multi-dimensionality**: They are designed to encompass multiple facets of a complex concept, providing a holistic view rather than focusing on a single aspect.
- **Aggregation**: They involve a structured process of combining diverse indicators, often with different units and scales, into a single, normalised score.
- **Comparability**: A significant advantage of composite indices is their utility in **comparing the performance of different entities**—such as districts, states, or countries—or for **tracking changes over time**.
- **Policy Formulation and Resource Allocation**: They serve as critical tools for policy formulation, enabling targeted interventions and efficient resource allocation to specific areas or issues.
- **Varying Complexity**: Composite indices can range from very simple constructions based on a few indicators to highly intricate ones incorporating a large number of variables.

3. Examples of Composite Indices

The lecture highlighted several widely recognised composite indices:

- **Human Development Index (HDI)**
- **Global Hunger Index (GHI)**

- **Corruption Perception Index (CPI)**
- **Global Innovation Index**
- **Technology Development Index**
- **Sustainable Livestock Production Index (SLPI)**: Developed at both district and state levels in India.
- **Sustainable Farming Index (SFI)**: Developed at the district level in India.

4. Broad Steps in Constructing a Composite Index

The construction of a robust composite index typically follows a systematic five-step process:

1. **Defining the Conceptual Framework and Purpose**: Clearly outlining what the index aims to measure and why.
2. **Identifying and Screening Indicators**: Selecting appropriate indicators that are relevant, measurable, and available.
3. **Normalisation**: Transforming indicators to a common scale to ensure comparability.
4. **Weighting Indicators and Dimensions**: Assigning relative importance to different components.
5. **Aggregation**: Combining the normalised and weighted indicators into the final composite index.
6. **Visualization of the results**

5. Detailed Explanation of Construction Steps

5.1. Step 1: Defining the Conceptual Framework and Purpose

This initial step is paramount. It involves **carefully defining the concept** that the index aims to understand what is being measured, who are the clients of the index, whether it is static or dynamic, and its specific purpose. It provides the basis for grouping of variables into a meaningful composite indicator under a fitness-for-purpose principle. This definition relies heavily on insights drawn from **existing literature and expert opinion**. A clear understanding of the purpose helps in delineating the multi-dimensional nature of the concept, which is crucial for defining its various sub-components or dimensions. For instance, a sustainability index might be conceptualised with economic, social, and ecological dimensions, each requiring specific indicators. Without a clear conceptual framework, the index risks being arbitrary or misleading.

5.2. Step 2: Identifying and Screening Indicators

Once the conceptual framework is established, the next step involves selecting the specific indicators that will be used to measure the different dimensions. Use proxy wherever the data is not available. It should have a balance between input and output; demand and supply; efficiency and inclusiveness; state, pressure and response. This process is not arbitrary and involves rigorous screening based on several criteria:

- **Relevance**: Indicators must be directly relevant to the concept and dimensions being measured.
- **Data Availability**: Practicality dictates that data for chosen indicators must be available or collectable.
- **Complexity and Cost-Effectiveness**: While comprehensive, the selection should avoid excessive complexity that might make data collection or analysis prohibitively expensive or time-consuming.
- **Geographical coverage**:
- **Policy Relevance**: Indicators should ideally offer insights that are actionable and useful for policy formulation.
- **Measurability**: The indicators must be quantifiable or categorizable in a meaningful way.
- **Relationship to the Theme**: Indicators should ideally have a clear and understandable relationship to the overall theme of measurement.
- **Balance**: It's important to ensure a balanced representation of indicators across all dimensions.

Statistical methods such as Correlation, PCA can assist in screening by identifying redundancy or grouping indicators, but the ultimate judgment should always be guided by the index's purpose.

5.3. Step 3: Normalisation

Indicators often come with different units of measurement (e.g., rupees, percentages, kilograms) and varying scales, which makes direct comparison or aggregation impossible. **Normalisation** is the process of transforming these diverse indicators into a common, unit-less scale. This ensures that each indicator contributes appropriately to the overall index without being disproportionately influenced by its original unit or range. Selection of the normalization depends on the purpose, nature of the data, variance in data, importance of extreme values, etc.

In general, there are five normalisation methods:

a) **Min-Max Normalisation**: This is a common and straightforward method, especially useful when comparing performances of districts or states relative to the minimum and maximum observed values within the dataset.
Formula for Positive Relationship (higher value = better performance):

$$\mathbf{I_{ijk}} = \frac{X_{ijk} - Min\ X_{ijk}}{Max\ X_{ijk} - Min\ X_{ijk}}$$

Formula for Negative Relationship (higher value = worse performance):

$$\mathbf{I_{ijk}} = \frac{Max\ X_{ijk} - X_{ijk}}{Max\ X_{ijk} - Min\ X_{ijk}}$$

X_{ijk} = value of the i^{th} variable representing j^{th} component of the k^{th} region.
It is also known as feature scaling or unitary method. No information loss as it retains the variability of the data and widen the range of indicators lying with a small interval. This method scales all values between 0 and 1. However, it can be sensitive to outliers, where extreme values can distort the normalisation range.

b) **Z-score Normalisation**: This method transforms data into standard deviations from the mean.

$$Z = \frac{X - Mean}{Standard\ Deviation}$$

This method is considered more statistically robust as it is less affected by outliers than Min-Max normalisation. The transformation standardizes all indicators to a common scale with a mean of zero and a standard deviation of one, ensuring consistency and minimizing distortions during aggregation, especially when indicators have varying original means. This approach is particularly suitable when it is desirable to reward or penalize exceptional values, as indicators with extreme deviations will exert a stronger influence on the resulting composite index.

c) **Ranking**: A simple method where entities (states, districts, countries) are ranked based on their indicator values. This method is simple to understand and implement, not sensitive to the outliers. It is useful in the case of lack of precise data, ordinal data. One of the disadvantages is loss of information about the magnitude of differences between ranks (e.g., the difference between rank 1 and 2 might be very small, while between 2 and 3 it might be large, but ranking treats them equally).

d) **Categorical scaling**: This method is almost similar to ranking which involves assigning numerical scores based on class or percentiles.

e) **Distance to Reference Point**: This approach measures the performance of an entity relative to one or more reference points. Reference points can be internal (e.g., the best performer in the dataset) or external (e.g., a global benchmark). Defining the reference point is crucial and results may be misleading if not defined properly.

$$I_{qc}^{t} = \frac{X_{qc}^{t}}{X_{qc=\bar{c}}^{t_0}}$$

$X_{qc}^{t} =$ is value of the indicator, q for country c, at time t, $\bar{c}$ is the reference country.

f) **Threshold-based Normalisation (as applied in SLPI/SFI)**: This method defines lower and upper thresholds for each indicator.

For **indicators positively related to sustainability** (higher value is better):

- If Actual Value <= Lower Threshold (LT), the score is **0**.
- If Actual Value >= Upper Threshold (UT), the score is **1**.
- If LT < Actual Value < UT, the score is calculated linearly: (Actual Value - LT) / (UT - LT).

For **indicators negatively related to sustainability** (lower value is better, e.g., Cost of Milk Production or Fertiliser Use):

- If Actual Value <= Lower Threshold (LT), the score is **1**. (This implies that very low cost/use is highly sustainable).
- If Actual Value >= Upper Threshold (UT), the score is **0**. (This implies that very high cost/use is unsustainable).
- If LT < Actual Value < UT, the score is calculated linearly: (UT - Actual Value) / (UT - LT). This method ensures that values beyond acceptable thresholds are capped at 0 or 1, preventing extreme outliers from disproportionately affecting the index, and focuses on performance within a defined 'sustainable' range.

5.4. Step 4: Weighting Indicators and Dimensions

After normalisation, the next step is to assign weights to indicators and/or dimensions. Weights reflect the relative importance or contribution of each component to the overall composite index.

Methods for assigning weights include:

a) **Equal Weighting (Simple Aggregation)**: This is the simplest method where all indicators or dimensions are given the same weight. It implies that all components are equally important. While easy to implement and understand, it often lacks a statistical or theoretical basis for choosing a different scheme or sufficient knowledge of casual

relationships or consensus on alternative solutions and may not reflect the true significance of different components.

b) **Different weights**

i. Weighing based on statistical tools.

- **Principal Component Analysis (PCA):** This method is useful when the indicators are linearly correlated. It uses statistical techniques to derive weights from the data itself, often based on the variance explained by each component. Weights are assigned by addressing the information of two or more correlated variables. If spurious correlations are present, this method reduces the influence of such indicators by assigning them to lower weights. However, the resulting weights may not necessarily align with the theoretical relevance or anticipated significance of the indicators.
- Data envelopment analysis

ii. Subjective/Expert-based Weighting: This involves

- Eliciting opinions from experts or stakeholders.
- Budget allocation process: Experts are given a "budget" of points (e.g., N =100 points) to distribute among indicators or dimensions based on their experience and subjective judgement or perceived importance. These methods allow for incorporating domain-specific knowledge but can be subjective and potentially influenced by biases.
- **Analytic Hierarchy Process (AHP)**: This method incorporates expert judgment in the weighting process and employs an analytical approach to break down a complex problem into a structured hierarchy of simpler, more manageable components or groups. A structured technique for organizing and analyzing complex decisions, involving pairwise comparisons of elements. Objectively derived from data can identify underlying structures and reduce redundancy among indicators. Finally, weights are computed using a eigenvector technique.

Choice of weight depends on objective of constructing index, reliability and pattern of data, and ease of weighting methodology. If the indicators are highly correlated, it is advised to use weight based on PCA. Theoretical factors and policy priority should be kept in mind in choosing weighting scheme Most of the composite indicators rely on equal weights.

5.5. Step 5: Aggregation

Aggregation is the last step where the individual, normalised, and weighted indicators are combined to form the composite index. The choice of aggregation method depends on the conceptual understanding of how the indicators interact and whether inferior performance in one area can be compensated by superior performance in another.

Three broad types of aggregation methodologies were discussed:

a) **Additive Aggregation**: This typically uses a **simple or weighted arithmetic mean** by summing up the rank based normalized scores of each indicator or aggregating the weighted normalized indicators using arithmetic mean**.** This approach implies **perfect substitutability** between indicators. This means that poor performance in one indicator can be fully compensated by excellent performance in another. For example, a low score in "access to education" could be offset by a high score in "health infrastructure" if both contribute to an overall "human development" index additively. **It is suitable when all the indicators have the same measurement unit.** This method rewards the indicators based on their relative weights. It should not be used when different goals are equally legitimate and vital.

$$CI_c = \sum_{q=1}^{Q} w_q I_{qc}$$

b) **Geometric Aggregation**: This uses a **geometric mean**. This method implies **imperfect substitutability** (partial compensability). While good performance in one area can still compensate for bad, it does so to a lesser extent than additive aggregation, and it penalises extreme imbalances. It cannot be applied if any indicator value is zero, as the product would become zero.

$$CI_c = \prod_{q=1}^{Q} X_{q.c}^{w_q}$$

c) **Non-Compensatory Aggregation**: Both linear and multiplicative aggregations are unsuitable if the modeller decides that the loss of one dimension cannot be compensated by increasing the performance of another dimension/indicator. This approach explicitly limits the ability of one indicator to compensate for another. It is used when certain minimum thresholds are critical for all dimensions or indicators. More relevant when extremely different dimensions are composited. Ranking is done based on multi-criterion approach (MCA)

6. Principal Component Analysis and Composite Index Generation

Principal Component Analysis (PCA) is a statistical method used to reduce the number of variables in a dataset while retaining as much of the original variability as possible. It does this by converting a group of correlated variables into a smaller number of uncorrelated variables known as principal components (PCs). These components are linear combinations of the original variables and are ranked based on the proportion of variance they capture from the data.

This document outlines the PCA process and the generation of a composite index, a single metric summarizing multidimensional data, such as the environmental performance of products or the impact of training on startups.

Principal Component Analysis (PCA): PCA involves several key steps to derive principal components from a dataset with n observations and p variables.

a) Standardization

Ensure variables with different scales contribute equally, they are standardized to have a mean of 0 and a standard deviation of 1:

$$X'_{ij} = \frac{X_{ij} - \bar{X}_j}{\sigma_j}, i = 1 \ldots . n ; j = 1, \ldots, p$$

Where X_{ij} is the original data value, $\bar{X}_j$ is the mean, and σ_j is the standard deviation of the j^{th} variable.

b) Correlation Matrix

When variables are standardized, the covariance matrix becomes the correlation matrix, with diagonal elements equal to 1. The sum of the eigenvalues of the correlation matrix equals the number of variables (ρ), which is the trace of the matrix.

c) Eigenvalues and Eigenvectors

The eigenvalues ($\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_p \geq 0$) and corresponding eigenvectors ($l_1, l_2, \ldots . . l_p$) of the correlation matrix are computed. Each principal component is defined as:

$$PC_1 = \boldsymbol{\ell_1}'\boldsymbol{X} = \ell_{11}X_1 + \ell_{12}X_2 + \ldots + \ell_{1p}X_p$$

$$PC_2 = \boldsymbol{\ell_2}'\boldsymbol{X} = \ell_{21}X_1 + \ell_{22}X_2 + \ldots + \ell_{2p}X_p$$

and so on till $PC_i = \boldsymbol{\ell_i}'\boldsymbol{X} = \ell_{i1}X_1 + \ell_{i2}X_2 + \ldots + \ell_{ip}X_p$

where $\boldsymbol{\ell_i} = (\ell_{i1}, \ell_{i2}, \ldots . \ell_{ip})$ is the i^{th} eigenvector, and $X = (X_1, X_2, \ldots \ldots X_p)$ is the vector of standardized variables.

d) Variance Explained

The variance of the i^{th} principal component is equal to its eigenvalue:

$$Var(PC_i) = l'_i \sum l_i = \lambda_i$$

Principal components are uncorrelated:

$$\text{Cov}(PC_i, PC_k) = l_i' \sum l_k = 0,\ i \neq k$$

The proportion of total variance explained by the r^{th} principal component is:

$$\frac{\lambda_r}{\lambda_1 + \lambda_2 + \lambda_3 + \cdots + \lambda_p}$$

Since the correlation matrix is used, the total variance equals p.

e) Selecting Principal Components

Typically, the first $r \leq p$ principal components are selected based on a cumulative variance threshold (e.g., 70–90%). These components capture the most significant patterns in the data.

f) Loadings

Loadings are the elements of the eigenvectors, representing the contribution of each original variable to a principal component. For use in a composite index, the absolute values of loadings may be normalized to sum to 1 for each principal component.

g) Composite Index Generation

A composite index (CI) aggregates multiple variables into a single metric to summarize complex phenomena. Using PCA, the composite index for the i^{th} observation (e.g., a startup or district) is:

$$CI_i = \frac{\lambda_1 PC_{1i} + \lambda_2 PC_{2i} + \lambda_Q PC_{Qi}}{\lambda_1 + \lambda_2 + \lambda_3 + \cdots + \lambda_Q}$$

CI_i: Composite index for the *i*-th observation.
λ_j:Eigenvalue of the *j*-th principal component.
PC_{ij} Value of the *j*-th principal component for the *i*-th observation.
Q: Number of selected principal components ($Q \leq p$).
The denominator normalizes the index by the sum of the eigenvalues of the selected components.

Normalization

To make the composite index comparable, it is normalized to a range of 0 to 1:

$$\boldsymbol{CI_{Ni}} = \frac{CI_i - min(CI)}{max(CI) - min(CI)}$$

where CI_{Ni} is the normalized composite index, and $min(CI)$ and $max(CI)$ are the minimum and maximum composite index values across all observations.

Interpretation

• **Principal Components**: Each PC_i is a linear combination of standardized variables, with weights from the eigenvector l_i. The first principal component explains the largest proportion of variance, followed by subsequent components.

• **Composite Index**: The composite index summarizes the selected principal components into a single score, weighted by their eigenvalues to their explanatory power.

• **Application**: Higher composite index values indicate better performance (e.g., positive impact of training on startups), while lower values suggest poorer performance.

Case Study: Evaluating Agricultural Performance Using PCA-Based Composite Index

In this case study, we evaluate the agricultural performance of twelve farms using multivariate data on key indicators:

i. **Crop Yield (x):** Measured in kilograms per hectare (kg/ha), this represents the total crop output from each farm and is a direct indicator of productivity.
ii. **Rainfall (y):** Recorded in millimetres (mm), this denotes the total precipitation received during the cropping season and significantly influences crop growth.
iii. **Fertilizer Use (z):** Represented in kilograms per hectare (kg/ha), this refers to the total quantity of fertilizer applied, affecting crop health and yield potential.
iv. **Pesticide Use (w):** Also measured in kilograms per hectare (kg/ha), this indicates the intensity of pest management practices employed by the farm.

Objective of the Study:

The goal is to integrate these four variables into a single composite index using Principal Component Analysis (PCA), which can be used to:

- ✓ Rank farms based on overall agricultural performance.
- ✓ Identify which variables contribute most to performance variation.
- ✓ Enable targeted interventions and policy recommendations.

Why Use PCA?

Agricultural performance is influenced by multiple interrelated factors. PCA helps in:

- ✓ Reducing data complexity by converting correlated variables into uncorrelated principal components.
- ✓ Prioritizing farms without subjectively assigning weights.
- ✓ Objectively combining variables into one metric that reflects multivariate performance.

By conducting unweighted PCA (equal importance to all farms), we ensure that the index reflects inherent performance rather than external survey biases.

Step 1: Prepare the Data

The dataset includes 12 farms with the following variables:

- Crop Yield (x) in kg/ha
- Rainfall (y) in mm
- Fertilizer Use (z) in kg/ha
- Pesticide Use (w) in kg/ha

The `Serial_No` is an identifier. The variables x, y, z, and w are extracted into a matrix for analysis.

Step 2: Compute the Covariance Matrix

The covariance matrix is calculated to capture the relationships among the variables, assuming equal weight for each observation. It provides the basis for PCA, showing how much the variables vary together.

Step 3: Perform Eigen Decomposition

Eigenvalues and eigenvectors are extracted from the covariance matrix. Eigenvalues represent the variance explained by each principal component (PC), while eigenvectors (loadings) indicate the contribution of each variable to each PC.

Step 4: Calculate Principal Component Scores

PC scores are computed by projecting the original data onto the principal components. These scores represent the transformed coordinates in the PCA space, capturing the underlying data structure.

Step 5: Compute the Composite Index

The composite index is calculated as a weighted sum of PC scores, where the weights are derived from the eigenvalues. This gives more importance to components explaining greater variance.

Step 6: Normalize the Composite Index

Min-max normalization is applied to rescale the composite index between 0 and 1. This allows for intuitive comparison across farms.

Step 7: Rank the Farms

Farms are ranked based on their normalized composite index, from highest to lowest. This ranking reflects their relative agricultural performance.

Case Study Interpretation: PCA-Based Composite Index of Agricultural Performance

This case study applies **Principal Component Analysis (PCA)** to develop a composite index that integrates four farm-level indicators:

- **x**: Crop yield (kg/ha)
- **y**: Rainfall (mm)
- **z**: Fertilizer use (kg/ha)

- **w**: Pesticide use (kg/ha)

The PCA composite index enables ranking the farms based on multivariate agricultural performance in a single, objective score normalized between **0 (worst)** and **1 (best)**.

R Code for PCA-based Composite Index

The following R code implements the steps described above, using a dataset with crop yield, rainfall, fertilizer, and pesticide use for 12 farms:

```
# Sample Data
Serial_No<- 1:12
x <- c(840,204,944,1009,745,811,883,593,254,215,172,169)  # Crop yield
y <- c(150,104,293,331,113,52,164,201,124,108,95,90)      # Rainfall
z <- c(350,214,693,731,413,352,364,301,214,208,186,130)   # Fertilizer
w <- c(50,20,70,90,45,55,62,49,23,21,19,18)               # Pesticide use (new)

# Combine into a data frame
data <- data.frame(Serial_No, x, y, z, w)

# Step 1: Extract variables matrix
var <- as.matrix(data[,-1])  # Remove Serial_No
n <- nrow(var)

# Step 2: Covariance matrix
cov_farm <- cov(var)

# Step 3: Eigen decomposition
eigenvalues <- eigen(cov_farm)$values
loadings<- eigen(cov_farm)$vectors

# Step 4: Principal component scores
scores<- var %*% loadings

# Step 5: Composite score using eigenvalue weights
score <- (scores %*% eigenvalues) / sum(eigenvalues)

# Step 6: Normalize composite index to [0,1]
a <- max(score)
b <- min(score)
comp_pcindex <- (score - b) / (a - b)

# Step 7: Rank the index
ranked_index <- cbind(data[order(-comp_pcindex), "Serial_No"], round(comp_pcindex
[order(-comp_pcindex)], 4), 1:n)
colnames(ranked_index) <- c("Sample_id", "Composite_index", "Rank")

# View the ranked result
print(ranked_index)
```

Composite Index Results & Rankings

Rank	Farm (Sample_id)	Composite Index
1	4	1.0000
2	3	0.9227
3	7	0.7269
4	1	0.6815
5	6	0.6453
6	5	0.6177
7	8	0.4524
8	9	0.1111
9	10	0.0726
10	2	0.0647
11	11	0.0248
12	12	0.0000

Key Interpretations

Top Performing Farms (Rank 1–3)

a) **Farm 4 (Index: 1.0000)**

- Best performing farm across all four variables.
- Extremely high crop yield (1009 kg/ha), highest rainfall (331 mm), high fertilizer (731 kg/ha), and high pesticide use (90 kg/ha).
- Dominates **PC1**, which captures the largest share of variance (~91% in your earlier setup).
- Sets the benchmark for 100% performance in this dataset.

b) **Farm 3 (Index: 0.9227)**

- Strong performer, especially in crop yield (944 kg/ha), rainfall (293 mm), and fertilizer (693 kg/ha).
- Slightly lower pesticide use than Farm 4, but still very good overall.
- Reflects a well-balanced input-output profile.

c) **Farm 7 (Index: 0.7269)**

- High crop yield (883 kg/ha), moderate rainfall and input use.
- Balanced input strategy with efficient use of fertilizer and pesticide.
- Performs well but not as input-intensive as Farm 3 or 4.

Moderate Performing Farms (Rank 4–7)

These farms show **moderate productivity** and **reasonable input use**, but may be lagging in one or two areas.

a) **Farm 1 (0.6815)** and **Farm 6 (0.6453)**: Strong yield and fertilizer use, but rainfall and/or pesticide inputs are relatively lower.
b) **Farm 5 (0.6177)**: Slightly lower yield and rainfall, possibly indicating more efficient use of resources.
c) **Farm 8 (0.4524)**: Shows moderate levels of performance, possibly lower input use or environmental constraints.

Low Performing Farms (Rank 8–12)

These farms are significantly lagging in most performance metrics.

a) Farm 12 (Index: 0.0000)

- Poorest performance across all variables.
- Lowest crop yield (169 kg/ha), lowest rainfall (90 mm), and very low fertilizer (130 kg/ha) and pesticide use (18 kg/ha).
- May face multiple constraints: poor soil, low investment, or drought stress.

b) Farm 11 (0.0248) and Farm 2 (0.0647)

- Very low yields and inputs.
- May be candidates for intervention — soil fertility, irrigation access, or capacity-building.

c) Farm 9 (0.1111) and Farm 10 (0.0726)

- Low yield with modest inputs.
- Likely operating under sub-optimal environmental or economic conditions.

Insights for Decision-Making

For Policymakers and Planners

- **Target support** to lowest-ranked farms (Farm 12, 11, 2, 10, 9) with interventions like irrigation, fertilizer subsidies, or soil health cards.
- **Best practices** from top-ranked farms (4, 3, 7) could be disseminated as model cases.

For Researchers

- Variables like crop yield and fertilizer use contributed significantly to **PC1**, the most vital component.
- PCA helped identify latent patterns — e.g., farms with high pesticide but low rainfall didn't perform well, indicating diminishing returns.

For Farmers

- Understanding the composite score helps benchmark performance.
- Encourages balanced use of inputs and highlights the importance of managing multiple factors (not just fertilizer or yield).

Conclusion

The importance of composite indices is essential for policy applications, data-driven decision-making, and understanding complex realities. It emphasised the need for transparency, careful selection of indicators and methodologies, and involving experts to ensure the reliability and validity of the constructed index. The choice of normalisation, weighting, and aggregation methods should always align with the overall purpose and conceptual framework of the index. In this case study, the PCA-based composite index simplifies multidimensional agricultural performance into a single, comparable metric. It provides: **Ranked performance** of each farm, **Variable contribution insights** via PCA and **Targeted policy suggestions** based on data

References

Manoj Kumar, M. K., Tauqueer Ahmad, T. A., Anil Rai, A. R., & Sahoo, P. M. (2013). Methodology for construction of composite index.

Mazziotta, M., & Pareto, A. (2024). Principal component analysis for constructing socio-economic composite indicators: theoretical and empirical considerations. *SN Social Sciences*, *4*(6), 114.

Johnson, R. A., & Wichern, D. W. (2002). Applied multivariate statistical analysis.

Difference-in-Difference (DiD) analysis using R

Ponnaganti Navyasree, Nobin Chandra Paul, Santosha Rathod, K. Ravi Kumar and Prabhat Kumar
ICAR-National Institute of Abiotic Stress Management, Baramati-413115
Email: navyaponnaganti97 @gmail.com

Introduction

The Difference-in-Differences (DiD) method is a cornerstone of impact evaluation, particularly in the field of econometrics, where it is often described as a "controlled before-and-after study." This powerful and intuitive approach is designed to estimate the causal effects of a policy, program, or intervention by leveraging a combination of temporal (before-and-after) and group-based (treatment and control) comparisons. Its widespread use in economics and related disciplines stems from its ability to provide credible estimates of treatment effects in settings where randomized controlled trials may not be feasible.

At its core, DiD seeks to answer a fundamental question: How has a specific intervention-such as a policy change, program implementation, or external treatment-impacted the individuals or entities exposed to it? This question is deceptively complex because, in reality, we cannot observe the counterfactual scenario-that is, what would have happened to those individuals had the intervention never occurred. For example, if a new labour policy is introduced in one region, we cannot directly observe how workers in that region would have fared without the policy, as the policy was indeed implemented. To address this challenge, DiD employs a clever strategy: it uses a control group—individuals or entities not exposed to the intervention—as a proxy for the counterfactual. The control group is assumed to experience similar trends to the treatment group in the absence of the intervention. By comparing the changes in outcomes (*e.g.*, employment rates, wages, or health metrics) between the treatment and control groups before and after the intervention, DiD isolates the effect of the intervention itself. Specifically, it calculates the treatment effect as the difference in the average post-intervention outcomes between the treatment and control groups, adjusted for any pre-existing differences in trends. The method relies on a key assumption known as the "parallel trends assumption," which posits that, in the

absence of the intervention, the treatment and control groups would have followed similar trajectories over time. While this assumption cannot be directly tested, researchers often examine pre-intervention trends to assess its plausibility. When applied correctly, DiD provides a robust way to disentangle the impact of an intervention from other confounding factors, making it a preferred method in fields like economics, public policy, and social sciences for evaluating real-world interventions.

The Difference-in-Differences (DiD) model is a widely used method to evaluate the impact of a specific change, such as a policy, program, or intervention, by comparing two groups over time.

Components of the Difference-in-Differences (DiD) Model:

The DiD model relies on four key elements that work together to estimate the effect of a change. These components allow us to isolate the impact of the change by comparing what happens to a group affected by it (the treatment group) to a similar group that isn't affected (the control group). Here's a detailed explanation of each component:

- A sudden **exogenous source of variation (Treatment)**: The "treatment" is the specific change, policy, or intervention you want to study. It's called "exogenous" because it comes from outside the system and isn't influenced by the people or things being studied. In other words, the treatment is something that happens suddenly and isn't caused by the behaviour of the treatment or control groups. The treatment is the event or action you're trying to evaluate. For DiD to work, the treatment should affect only the treatment group and not the control group, and it should happen at a clear point in time.
- A quantifiable and measurable **outcome:** The outcome is the specific thing you measure to see if the treatment had an effect. It's something you can put a number on, like income, test scores, or health outcomes. The outcome can either be the direct target of the treatment (e.g., the program aims to increase employment, so you measure employment rates) or an indirect proxy (e.g., the program aims to improve confidence, but you measure hours worked as a related indicator). A measurable outcome is crucial

because DiD relies on comparing numbers before and after the treatment for both groups. Without a clear, quantifiable outcome, you can't calculate the effect of the treatment.

- A **treatment group:** The treatment group consists of the people, organizations, or things that experience the treatment (the policy or change). These are the subjects directly affected by the intervention you're studying. By comparing their outcomes before and after the treatment, you can start to see the effect, but you'll need the control group to make sense of it.
- A **control group:** The control group is a group of people, organizations, or things that are similar to the treatment group in key characteristics (like age, location, or economic status) but do not experience the treatment. They act as a "comparison group" to show what would have happened without the change.

Method

The DiD method calculates the impact of the policy or program by comparing how the outcome (e.g., wages, test scores, or employment rates) changes over time for the treatment group versus the control group. The key idea is to isolate the effect of the policy by removing changes that would have happened anyway, which we see in the control group.

The DiD estimate is calculated using a simple formula that compares the average outcomes for the two groups (treatment and control) across the two time periods (before and after). Here's the formula, explained in plain language:

DiD Estimate = (Change in Treatment Group's Average Outcome) - (Change in Control Group's Average Outcome)

In mathematical terms, it is written as:

$$DiD = (\bar{y}_s = \text{Treatment, t=After} - \bar{y}_s = \text{Treatment, t=Before}) (\bar{y}_s = \text{Control, t=After} - \bar{y}_s = \text{Control, t=Before})$$

where y is the outcome variable like wages, test scores, or employment status.

$\bar{y}$ represents the average outcome for a group at a specific time, calculated by adding up the outcomes for all individuals in that group and dividing by the number of individuals.

s: This indexes the group, either "Treatment" (the group that gets the policy) or "Control" (the group that doesn't). The letter "s" is often used because many studies involve policies applied at the state level, but it can represent any group.

t: This indexes the time period, either "Before" (before the policy) or "After" (after the policy).

Individuals (i): The data is typically at the individual level, so the averages ($\bar{y}$) are calculated over many individuals, each indexed by "i" (e.g., each worker or student in the sample). The formula calculates the difference in the average outcome for the treatment group (after minus before) and subtracts the difference in the average outcome for the control group (after minus before). This "double difference" gives you the estimated effect of the policy.

How the DiD Method Analyses Data

The DiD method calculates the impact of the treatment by comparing how the outcome changes over time for the treatment group versus the control group. The key idea is to isolate the effect of the treatment by subtracting out changes that would have happened anyway, which we see in the control group. Here's how the analysis works in simple steps:

1. **Calculate the Change (or "Gain") for Each Group**:
    - For the **treatment group**, find the difference in the average outcome between the "after" period and the "before" period. This shows how much the outcome changed after the treatment was introduced.

$$\textbf{Change in Treatment Group} = \textbf{Average Outcome}_{\textbf{After, Treatment}} - \textbf{Average Outcome}_{\textbf{Before, Treatment}}$$

- For the **control group**, do the same: find the difference in the average outcome between the "after" period and the "before" period. This shows how much the outcome changed without the treatment.

Change in Control Group = Average Outcome$_{\text{After, Control}}$ - Average Outcome$_{\text{Before, Control}}$

Subtract the Change in the Control Group from the Change in the Treatment Group:

Determine the change over time within the treatment group and subtract the corresponding change observed in the control group. This difference-in-differences approach isolates the estimated impact of the treatment by accounting for trends that would have occurred in the absence of the intervention.

DiD Estimate = (Average Outcome$_{\text{After, Treatment}}$ - Average Outcome$_{\text{Before, Treatment}}$) - (Average Outcome$_{\text{After, Control}}$ - Average Outcome$_{\text{Before, Control}}$)

In mathematical notation, this is often written as:

$$\text{DiD} = (\bar{y}_{s=\text{Treatment}, t=\text{After}} - \bar{y}_{s=\text{Treatment}, t=\text{Before}}) - (\bar{y}_{s=\text{Control}, t=\text{After}} - \bar{y}_{s=\text{Control}, t=\text{Before}})$$

where:

- $\bar{y}$ is the average outcome (e.g., average test score or wage).
- s indicates the group (Treatment or Control).
- t indicates the time period (Before or After).

The Difference-in-Differences (DiD) method is a powerful tool used to figure out whether a specific change, like a new law, policy, or program, actually made a difference. It's called a quasi-experimental design because it mimics a scientific experiment without randomly assigning people to groups, which isn't always possible in real-world situations. Instead, DiD uses longitudinal data—information collected over time—to compare two groups: one that experiences the change (the treatment group) and one that doesn't (the control group). By looking at how outcomes change over time for both groups, DiD helps us estimate the

causal effect of the change, meaning the true impact it had, separate from other factors. DiD is especially useful for studying the impact of specific interventions or treatments, such as passage of a law, enactment of a policy, or large-scale program implementation: A nationwide job training program rolled out in some regions but not others.

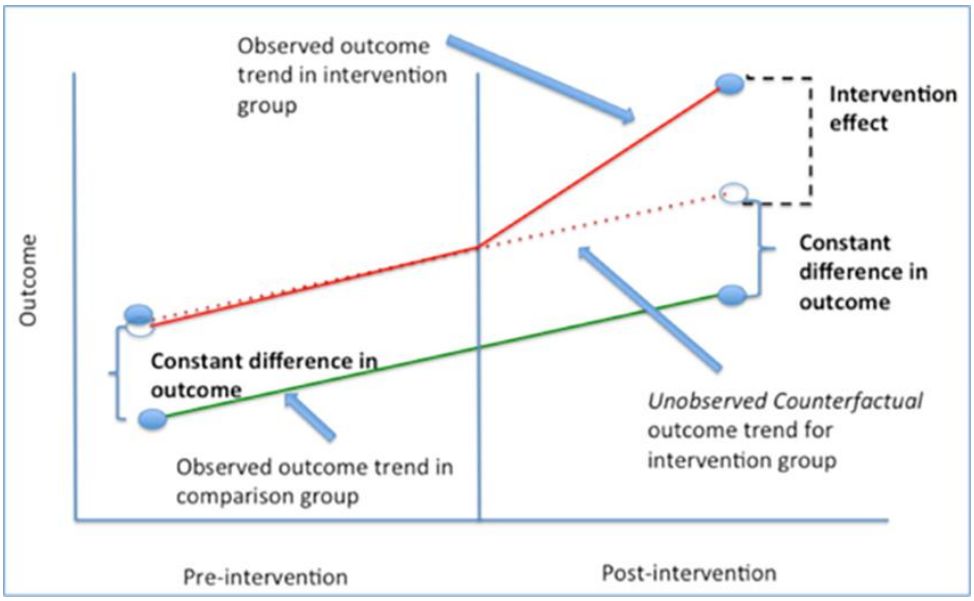


Figure 1. Graphical explanation of DiD estimation ***(Source: Conrad et al. 2021)***

Difference-in-Differences (DiD) method is a widely used analytical tool for evaluating the impact of a policy, program, or intervention (referred to as the "treatment") in observational settings, where researchers cannot randomly assign individuals to treatment or control groups. Unlike randomized controlled trials, where groups are made comparable through random assignment, DiD works with real-world data where the treatment is applied naturally—for example, a new law or a job training program implemented in one region but not another. By comparing changes in outcomes over time between a treatment group (those affected by the intervention) and a control group (those unaffected), DiD estimates the causal effect-the true impact of the intervention—while accounting for differences and trends that could skew results.

DiD in Observational Settings: Why Randomization Isn't Needed

In an ideal experiment, researchers would randomly assign people to a treatment group (e.g., receiving a new drug) or a control group (e.g., receiving a placebo) to ensure the groups are nearly identical in all ways except the treatment. This exchangeability means any difference in outcomes can be confidently attributed to the treatment. However, in observational settings, like studying the effect of a new minimum wage law in one state compared to another, random assignment isn't possible. States or individuals differ in ways that might affect outcomes (e.g., economic conditions, education levels, or local policies), so full exchangeability isn't realistic.

The Difference-in-Differences (DiD) approach addresses this issue by using a less stringent assumption called the **parallel trends assumption**. This assumption posits that, in the absence of the treatment, the unobserved differences between the treatment and control groups—such as factors like motivation, social norms, or community attitudes—would remain stable over time. By assuming these unmeasured characteristics evolve similarly in both groups, DiD helps isolate the true effect of the treatment. In other words, without the intervention, the outcome (e.g., employment rates or test scores) for both groups would have followed similar trends. By focusing on changes in outcomes over time rather than their absolute levels, DiD accounts for these differences and isolates the treatment's effect, making it a practical tool for real-world studies where randomization isn't feasible.

Data Requirements for DiD

To apply DiD, you need **longitudinal data**—information collected over time—to track outcomes before and after the intervention. This data must include:

- **Pre-Intervention Period**: Data from before the treatment is introduced, providing a baseline.
- **Post-Intervention Period**: Data from after (or during) the treatment, showing the outcome after the change.

The data can be collected in two main ways:

- **Cohort or Panel Data**: This involves tracking the *same individuals or entities* over time. For example, you might follow the same group of workers before and after a job training program, measuring their employment status at both points. Panel data is powerful because it reduces the variation by focusing on the same people.

- **Repeated Cross-Sectional Data**: This involves collecting data from *different individuals or groups* at each time period, ideally using random samples to represent the population. For example, you might survey different workers in a state before and after a policy change. This is useful when tracking the same individuals isn't possible.

The data is typically at the individual level (e.g., each worker's wages or each student's test scores) but can also be at the group level (e.g., average employment rates for a city or state). The key is to measure the same outcome for both the treatment and control groups in both time periods to ensure a fair comparison.

Causal Effects

The Difference-in-Differences (DiD) method is a useful for determining the causal effect of a policy, program, or intervention (the "treatment")—that is, the true change in an outcome, such as employment rates or test scores, directly caused by the intervention. DiD is most commonly used to estimate the treatment effect on the treated, which measures the impact on the group that receives the intervention. With stricter assumptions, it can also estimate the Average Treatment Effect (ATE), which reflects the effect if the intervention were applied to the entire population, including those not treated. DiD's strength lies in its simple yet effective approach to handling differences between groups, making it ideal for observational studies where groups may not be identical. Let's explore this in a clear and straightforward way.

Types of Causal Effects

DiD typically focuses on the treatment effect on the treated—the impact of the intervention on the group that experiences it. For instance, if a city launches a job training program, DiD can reveal how it affected the employment rates of the workers who participated. This is the causal effect for those "exposed" to the treatment.

In some cases, researchers aim to estimate the Average Treatment Effect (ATE), which is the expected impact if the intervention were applied to everyone, including the control group (those not receiving the treatment). Estimating the ATE requires stronger assumptions, such as the treatment having the same effect on the control group as on the treatment group if they were treated. This can be challenging in observational studies where groups differ, so DiD is most often used for the treatment effect on the treated, with ATE estimation reserved for studies with robust design and assumptions.

Key Assumptions for the Difference-in-Differences (DiD) Method:

1. **Stable Group Composition:** The DiD method assumes that the composition of the treatment and control groups remains consistent over the study period (before and after the intervention). This means the characteristics of the individuals or entities in each group—such as age, income, education, or other relevant factors—should not change significantly during the time frame of the study. Stable composition ensures that any changes in the outcome (e.g., employment rates or test scores) are due to the treatment, not shifts in who makes up the groups.
2. **No Spillover Effects:** The DiD method assumes that the treatment only affects the treatment group and has no spillover effects on the control group. Spillover effects occur when the intervention indirectly influences the control group's outcomes, contaminating their role as a counterfactual (what would have happened without the treatment).
3. **Exogenous Treatment (Treatment Not Determined by Outcome):** The DiD method assumes that the amount or assignment of the treatment is not influenced by the outcome being studied. In other words, the intervention is exogenous—it comes from outside the system and isn't driven by factors related to the outcome.

For example, a policy shouldn't be implemented because of changes in the outcome it aims to affect.

4. **Parallel Trends Assumption:** The DiD method assumes that, in the absence of the treatment, the treatment and control groups would have followed similar trends in the outcome over time. This is called the parallel trends assumption. It means the difference in the average outcome between the groups (e.g., a 3% higher employment rate in the treatment group) would remain constant over time if neither group received the treatment. In other words, both groups would experience similar changes in the outcome due to external factors, like economic trends.

Strengths and Weaknesses of Difference in Differences

Strengths of the DiD Method

1. Intuitive and Easy to Understand

DiD is straightforward and intuitive, making it accessible to researchers, policymakers, and stakeholders. It works by comparing the change in an outcome (e.g., employment rates or test scores) for the treatment group to the change in the same outcome for the control group. The difference between these changes—the "difference in differences"—isolates the treatment's effect. This simple logic (before vs. after, treatment vs. control) makes DiD easy to explain and apply, even for those new to statistical methods.

2. Flexible for Observational Data

DiD is highly flexible because it can estimate causal effects using observational data, where randomization (like in a clinical trial) isn't feasible. It's ideal for studying real-world interventions, such as new laws, public health campaigns, or educational programs, where groups are naturally assigned to treatment or control conditions (e.g., different states or cities). As long as the key assumptions are met (see below), DiD can provide reliable causal estimates without requiring perfectly identical groups.

3. Handles Different Starting Levels

Unlike methods that require groups to have similar baseline outcomes, DiD focuses on **changes** in outcomes rather than absolute levels. This means the treatment and control groups can start with different outcome levels (e.g., different employment rates) without biasing the results, as long as the differences are constant over time (the parallel trends assumption). By differencing out these fixed differences, DiD isolates the treatment's effect, making it versatile for diverse settings.

4. Accounts for External Factors

A major strength of DiD is its ability to control for changes in outcomes caused by factors other than the treatment, such as economic trends, seasonal effects, or societal shifts. These external factors often affect both the treatment and control groups similarly. By subtracting the control group's change (which captures these external influences) from the treatment group's change, DiD isolates the treatment's true effect.

Weaknesses of the DiD Method

1. Reliance on Strict Assumptions

DiD's effectiveness depends on several critical assumptions, and violating any of them can lead to biased or unreliable results. These assumptions include stable group composition, no spillover effects, exogenous treatment, and parallel trends (detailed below). If these conditions aren't met, the DiD estimate may not accurately reflect the treatment's causal effect, limiting its applicability in some cases.

2. Requirement for Baseline and Control Group Data

DiD requires data from both **before and after** the intervention (longitudinal data) for both the treatment and control groups. This can be a limitation in settings where baseline (pre-intervention) data is unavailable or where finding a suitable control group is challenging. Without a control group to serve as a counterfactual, or without pre-intervention data to establish trends, DiD cannot be applied.

3. Limitations Due to Assumption Violations

DiD should not be used if certain conditions are not met, as they undermine the method's ability to isolate the treatment's effect:

- **Treatment Determined by Baseline Outcome**: If the treatment's assignment or intensity is influenced by the outcome being studied (endogenous treatment), DiD results will be biased. For example, if a job training program is offered only to workers with low employment rates, the treatment is tied to the outcome, making it hard to separate the program's effect from pre-existing conditions.

- **Different Outcome Trends (Non-Parallel Trends)**: DiD assumes that, without the treatment, the treatment and control groups would have followed similar outcome trends (parallel trends assumption). If the groups have different trends before the intervention (e.g., one group's employment rates are rising faster due to local factors), the DiD estimate may attribute these trends to the treatment, leading to incorrect conclusions.

- **Unstable Group Composition**: DiD assumes the treatment and control groups' characteristics (e.g., age, skills, or demographics) remain stable over the study period. If the groups' composition changes significantly (e.g., new people join or leave), these changes could affect the outcome and be mistaken for the treatment's effect.

Conclusion:

The DiD method offers a straightforward and effective approach to estimating the causal effects of policies, programs, or interventions (referred to as the "treatment"). By comparing changes in outcomes over time between a treatment group (those affected by the intervention) and a control group (those unaffected), DiD isolates the true impact of the treatment, making it an intuitive and accessible tool for researchers and policymakers. Its simplicity lies in its focus on changes rather than absolute values, allowing it to work with real-world observational data where groups may differ at the outset. In recent years, the

DiD method has gained significant traction in management studies, where it is increasingly applied to evaluate the impact of organizational changes, workplace policies, or strategic interventions. Its flexibility makes it a viable research design across various subfields of management, such as human resources, organizational behavior, and strategic management. For example, DiD can assess how a new employee training program affects productivity or how a shift in corporate policy impacts firm performance. The choice of research method depends on the context, data availability, and study goals, but DiD stands out as a particularly well-suited approach for estimating the causal effects of policy changes or management practices. It excels in observational settings where controlled experiments or natural experiments are not feasible, providing policymakers and managers with critical insights into the effectiveness of their initiatives. By leveraging longitudinal data and a control group, DiD accounts for external factors and group differences, delivering reliable estimates of a treatment's impact. Globally, DiD has been widely adopted to study the effects of diverse policies and reforms, from labor market interventions to public health campaigns and educational initiatives. Its ability to provide actionable evidence without requiring randomization makes it a valuable tool for decision-makers seeking to understand the real-world impact of their actions. As management studies continue to embrace data-driven approaches, DiD's simplicity, flexibility, and robustness position it as a go-to method for evaluating change and informing evidence-based policy and practice.

Implementation of DiD Test Using R

Government of India introduced the **Pradhan Mantri Fasal Bima Yojana (PMFBY)** in 2016 to provide financial protection to farmers through crop insurance. While the scheme has been widely implemented, the actual impact of crop insurance on **farmers' income** remains a critical policy question. This study uses a **DiD** framework to estimate the causal impact of crop insurance on farm income by comparing income changes over time between districts that adopted the scheme at scale (**treated group**) and those with low or no implementation (**control group**).

Dataset Overview

- **Total districts analyzed:** 50
- **Treated districts:** 25- These districts represent areas where crop insurance was widely adopted between 2015 and 2019.
- **Control districts:** 25- These are districts with minimal or no adoption of crop insurance in the same period.
- **Time period covered:**
  - **Pre-intervention year:** 2015
  - **Post-intervention year:** 2019
- **Observations:** 100 (2 years × 50 districts)

Each district has two observations (one pre- and one post-policy), capturing their average farm income over time. The DiD model estimates how income changed in the treated districts compared to the control group, isolating the effect of crop insurance by controlling for common time trends and fixed district characteristics.

Synthetic Dataset (Excerpt)

Below is a snapshot of the synthetic dataset used for the DiD analysis:

District	Treated	Post	Year	Income (₹)
District_1	1	0	2015	73534.25
District_1	1	1	2019	75864.57
District_2	1	0	2015	62519.13
District_2	1	1	2019	72297.64
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
District_48	0	0	2015	63372.43
District_48	0	1	2019	63470.14
District_49	0	0	2015	58083.71
District_49	0	1	2019	60984.03
District_50	0	0	2015	60981.38
District_50	0	1	2019	63780.79

R code:

```
##... Implementation of DiD Test Using R …##

# Load required libraries
library(tidyverse)
library(broom)
library(readxl)
library(ggplot2)

# Load data
# setwd("your_directory_path") # Optional
data <- read_xlsx("DiD_Crop_Insurance_Data.xlsx")

# Create interaction term for DiD
data$did <- data$treated * data$post

# Run the DiD regression
model <- lm(income ~ treated + post + did, data = data)
summary(model)

# Create synthetic conceptual data
df_concept <- data.frame(
  time = rep(c(0, 1), 2),
  outcome = c(5, 6, 4, 4.5),  # treated and control
  group = rep(c("Treated", "Control"), each = 2)
)

# Add counterfactual point
counterfactual <- data.frame(time = 1, outcome = 5.5, group = "Counterfactual")
```

```
# Plot
ggplot(df_concept, aes(x = time, y = outcome, color = group)) +
  geom_line(aes(group = group), size = 1.2) +
  geom_point(size = 3) +
  geom_point(data = counterfactual, aes(x = time, y = outcome),
             shape = 1, size = 4, color = "black", stroke = 1.2) +
  annotate("segment", x = 1, xend = 1, y = 5.5, yend = 6,
           linetype = "dashed", color = "black", size = 1) +

  # Adjusted Annotations
  annotate("text", x = 1.05, y = 5.75, label = expression(bold("Intervention")),
           hjust = 0, size = 3.5) +
  annotate("text", x = 0.25, y = 6.05, label = expression(bold("Treated Group")),
           color = "#F8766D", size = 4) +
  annotate("text", x = 0.25, y = 4.55, label = expression(bold("Control Group")),
           color = "#00BFC4", size = 4) +
  annotate("text", x = 1.07, y = 5.35, label = expression(bold("Counterfactual")),
           color = "black", size = 3.5) +

  # Axis settings
  scale_x_continuous(breaks = c(0, 1), labels = c("Pre", "Post")) +
  scale_y_continuous(limits = c(3.5, 6.2)) +

  # Labels and theme
  labs(
    title = "Difference-in-Differences Plot",
    x = "Time",
    y = "Outcome",
    color = "Group"
  ) +

  coord_cartesian(clip = "off") +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.position = "none",
    plot.margin = margin(10, 50, 10, 50)  # top, right, bottom, left
  )
```

Suggested Readings

Fredriksson, A., & Oliveira, G. M. D. (2019). Impact evaluation using Difference-in-Differences. *RAUSP Management Journal*, *54*, 519-532.

Cao, Zhun et al. Difference-in-Difference and Instrumental Variabels Approaches. An alternative and complement to propensity score matching in estimating treatment effects.CER Issue Brief: 2011.

Columbia University Population Health Methods: DiD Estimation. https://www.mailman.columbia.edu/research/population-health-methods/difference-difference-estimation.

Conrad, D. A., Milgrom, P., Du, Y., Cunha-Cruz, J., Ludwig, S., & Shirtcliff, R. M. (2021). Impacts of innovation in dental care delivery and payment in Medicaid managed care for children and adolescents. *BMC Health Services Research*, *21*(1), 565

Donald, S. G., & Lang, K. (2007). Inference with difference-in-differences and other panel data. *The review of Economics and Statistics*, *89*(2), 221-233.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of econometrics*, *225*(2), 254-277.

Applications for Propensity Score Matching in Agricultural Economics Research.

Ponnaganti Navyasree[1], Rakesh N[2], Nobin Chandra Paul[1], Santosha Rathod[1], K. Ravi Kumar[1] and Prabhat Kumar[1]

ICAR-National Institute of Abiotic Stress Management, Baramati-413115

ICAR- Mahatma Gandhi Integrated Farming Research Institute, Motihari, Bihar-845429

Email: navyaponnaganti97 @gmail.com

1. Introduction:

Randomized Controlled Trials (RCTs) are considered the gold standard for evaluating the causal impact of treatments on outcomes. In RCTs, random assignment ensures that both observed and unobserved baseline characteristics do not systematically differ between treated and control groups, thus eliminating confounding. This allows for a reliable comparison of outcomes to estimate the treatment effect (Austin, 2011). However, in **observational studies**, where treatment assignment is not under the control of the researcher (Porta, 2008), individuals often self-select into treatment based on their characteristics. This selection process introduces confounding, as the treated and control groups may differ in ways that affect the outcome. To address this issue, **Propensity Score Matching (PSM)** is employed. PSM is a statistical method designed to reduce selection bias by pairing individuals in the treatment and control groups who have similar observable characteristics. Its goal is to replicate the balance achieved in randomized experiments by controlling for differences in pre-treatment variables (X), thereby providing a more accurate estimate of the treatment's causal effect. PSM has gained wide application across disciplines such as labor economics, pharmacoepidemiology, education, and the social sciences. In the **agriculture sector**, recent applications include evaluating the economic benefits of Integrated Pest Management (IPM) and Farmer Field Schools (FFS) (Sanglestsawai et al., 2015), assessing the effects of on-farm climate adaptation strategies (Gorst et al., 2015), examining the influence of irrigation technologies on rural poverty (Solomon and Ketema, 2015), and analyzing the impact of agricultural extension services on social capital (Rijn, 2015).

2. The Propensity Score:

Propensity scores offer a valuable alternative for estimating treatment effects when random assignment is not possible. A propensity score represents the probability that an individual receives a specific treatment based on their observed characteristics (covariates). This score is usually estimated using statistical models such as logistic regression, which predicts treatment likelihood using variables like age, income, or, in agricultural contexts, factors such as farm size and soil quality. By summarizing multiple covariates into a single value, the propensity score simplifies the task of achieving balance between treatment and control groups. For instance, in a study comparing organic and conventional farming practices, the propensity score reflects the likelihood that a farmer chooses organic methods, given their farm's attributes. Matching farmers with similar propensity scores

allows researchers to create equivalent groups, thereby reducing selection bias and producing more reliable estimates of treatment effects-such as differences in crop yield. The key advantage of the propensity score lies in its ability to adjust for numerous confounding factors simultaneously, without requiring exact matching on each individual variable. **Propensity Score Matching (PSM)** involves pairing individuals from treatment and control groups who have similar propensity scores (and potentially similar covariates), while discarding unmatched units to improve comparability (Inacio *et al.*, 2015). PSM is widely used to compare groups in observational studies where randomization is not practical.

Statistical Definition:

The estimated propensity score, for subject *i, (i = 1,..., N*) is the conditional probability of being assigned to a particular treatment given a vector of observed covariates x_i (Rosenbaum and Rubin, 1983):

$$e\ (X_i) = \Pr(Z_i = 1/X_i)$$

where:

Z_i=1, for treatment
Z_i=0, for control and
X_i, the vector of observed covariates for the i^{th} subject
Since the propensity score is probability, it ranges in value from 0 to 1. In randomized studies, covariates are variables that are not affected by the allocation of treatments to subjects.

3. History of Propensity score matching

The concept of Propensity Score Matching (PSM) was first introduced by **Rosenbaum and Rubin (1983)** in their seminal work titled *"The Central Role of the Propensity Score in Observational Studies for Causal Effects."* They established the theoretical foundation for using propensity scores to reduce bias in observational studies by balancing covariates between treatment and control groups. **Heckman (1997)** also significantly contributed to the advancement of PSM methodologies, particularly by addressing **selection bias** in non-randomized settings. His work focused on improving causal inference when treatment assignment is influenced by observed or unobserved factors. Furthermore, Heckman later developed the **Difference-in-Differences (DiD)** approach, which complements PSM and is often used in combination to strengthen causal analysis in observational studies.

4. Key Assumptions for PSM Validity

Assumption 1: (Unconfoundedness / Conditional Independence Assumption or CIA)
After accounting for a set of observable covariates, the potential outcomes become independent of the treatment assignment. Let Y_1 represent the outcome that would occur under the treatment condition and Y_0 represent the outcome under the control condition.

$$(Y_1 , Y_0) \perp D \mid X$$

This implies that, once we control for the observed covariates (X), the assignment to treatment can be considered effectively random. This assumption, known as

unconfoundedness or **selection on observables**, is crucial for identifying the true effect of a program or intervention. It allows for an accurate estimation of selection bias by ensuring that differences between treated and control groups are attributable to observed factors. As a result, it becomes possible to construct a valid counterfactual for the treatment group using comparable individuals from the control group.

Assumption 2: (Overlap Condition or Common Support Condition):

For each value of X, there is a probability of being both treated and control.

$$0<P(Xi|D=1)<1$$

This means that, for every possible value of the covariates (X), the probability of receiving treatment falls strictly between 0 and 1, and likewise, the probability of not receiving treatment also lies within this range. In other words, both treated and control units must exist for all values of X. This condition ensures sufficient overlap in the characteristics of both groups, allowing for meaningful comparisons and adequate matching-a concept referred to as the **overlap condition** or **common support**. When both this overlap condition and the unconfoundedness assumption are met, the treatment assignment is considered **strongly ignorable** (Rosenbaum and Rubin, 1983, pp. 41-50). This foundation allows researchers to reliably estimate causal effects using methods like Propensity Score Matching. This relationship can be illustrated in **Figure 1**.

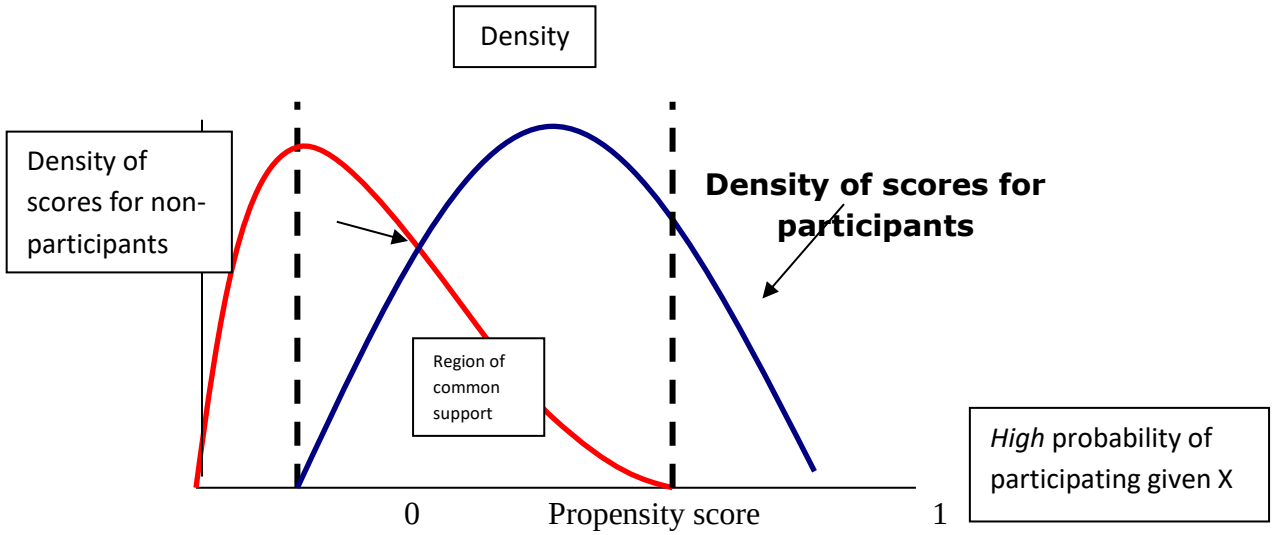
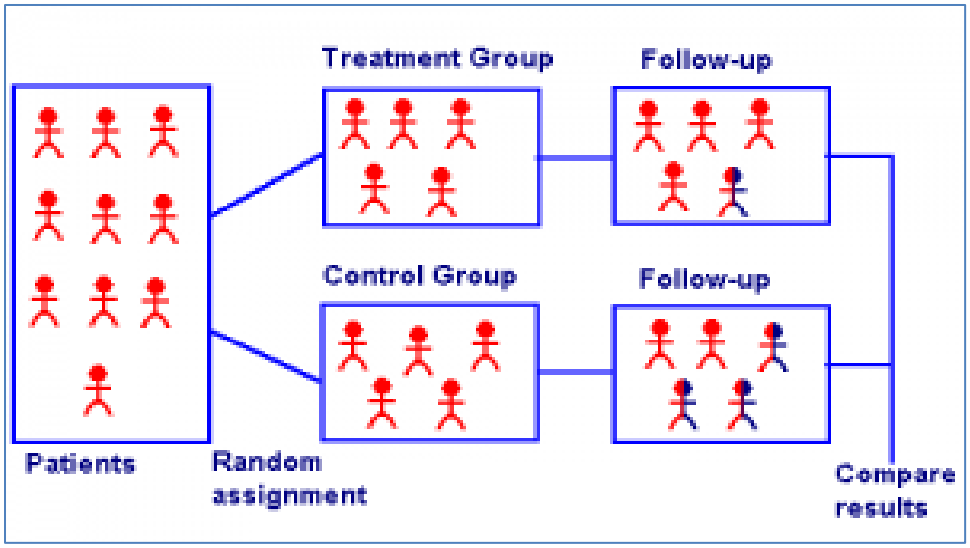


Fig 1: Propensity score density curves with common support region. (Source: Duy., 2012.)

5. Basic steps for propensity score matching

A **propensity score** represents the likelihood that a unit with specific observed characteristics will receive the treatment rather than being assigned to the control group. In observational studies, these scores are used to reduce or eliminate **selection bias** by ensuring that covariates (participant characteristics) are balanced across treatment and control groups. When this balance is achieved, it becomes significantly easier to match individuals based on multiple characteristics, thereby improving the validity of causal comparisons.



The primary objective is to replicate the conditions of a randomized experiment, thereby addressing many of the limitations associated with analyzing observational data. **Matching designs** can be classified as either **bipartite** or **non-bipartite**. In bipartite matching, each treated unit is paired with a unique control unit-similar to sampling without replacement-making it the more commonly used approach. In contrast, **non-bipartite matching** allows for the reuse of control units across multiple treated units, resembling sampling with replacement. This design is useful in situations where suitable matches are limited, and a single control unit may need to serve as a match for multiple treated individuals. Beyond matching, **propensity scores** can also be applied through other techniques to address confounding. These include **stratification** (dividing data into subgroups based on propensity score ranges), **regression adjustment** (incorporating the propensity score into regression models), and **inverse probability weighting**, which assigns weights to observations based on the inverse of their propensity scores to create a balanced pseudo-population.

Implementation Steps for PSM

Step 1: Model the Propensity Score

This initial step involves estimating each individual's probability of receiving treatment based on their observed covariates.

Model Choice:

a) For **binary treatment** cases (participation vs. non-participation), logistic regression (logit) or probit models are most used. These are preferred over linear probability models due to their ability to constrain predictions within the probability bounds. The choice between logit and probit is often not critical as they typically yield similar results.

$$\ln \frac{e\,(X_i)}{1 - e\,(X_i)} = \ln \frac{\Pr(Z_i = \frac{1}{X_i})}{1 - \Pr(Z_i = \frac{1}{X_i})} = a + \beta^T X_i$$

where:

$$e\,(X_i) = \Pr(Z_i = \frac{1}{X_i})$$

$$e\,(X_i) = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_i X_i$$

b_0 is the intercept
b_i is the regression coefficient
X_i , the treatment variables and covariates (random variables)
x_i , observed value of variables
In logistic regression, the dependent variable is binary, Z_i=1 is for the treatment and Z_i =0 for the control.

b) For **multiple treatment** cases (e.g., choosing among several programmes), the choice becomes more important. While multinomial probit is generally preferred over multinomial logit (due to the latter's "independence from irrelevant alternatives" assumption), it can be computationally burdensome. A practical alternative is to estimate a series of binomial models (e.g., comparing each treatment group to a reference group or to other treatment groups pairwise).

- **Variable Choice (Covariates)**:

a) **Careful selection of covariates (X)** is essential, as they must plausibly meet the **unconfoundedness assumption**. Only those variables that simultaneously affect both the likelihood of participating in the treatment and the outcome of interest should be included. The selection process should be informed by **economic theory**, **prior empirical studies**, and a solid understanding of the **institutional context**. Additionally, covariates must be measured **before treatment occurs** and should not be influenced by participation or the anticipation of it. Ideally, data for both treated and untreated individuals should be drawn from the **same data sources** to ensure consistency and comparability.

b) **Concerns with Exclusion/Inclusion**: Omitting important variables can significantly increase bias. However, including too many extraneous variables can exacerbate the common support problem or increase the variance of estimates, especially in smaller samples. This highlights a trade-off between ensuring the plausibility of the CIA and managing the variance of the estimates.

c) **Statistical Strategies for Selection**: While theory is paramount, formal statistical tests can assist:

- **"Hit or Miss" Method (Prediction Rate Metric)**: Variables are chosen to maximise the within-sample correct prediction rates of treatment assignment.
- **Statistical Significance**: Start with a parsimonious model and iteratively add variables that are statistically significant.

- **Leave-One-Out Cross-Validation**: Choose variables by comparing mean squared errors (MSEs) as different blocks of variables are added.

d) **Overweighting Key Variables**: If some variables are considered particularly important, researchers can "overweight" them by carrying out matching on subpopulations (e.g., matching men with men, women with women) or insisting on a perfect match for those specific variables prior to PSM.

Step 2: Choose a Matching Algorithm

Once propensity scores are estimated, individuals are matched. Different algorithms determine how a treated individual's counterfactual outcome is constructed from the comparison group. An appropriate matching technique is implemented with the estimated propensity score. Below are seven of the primary types of propensity score matching:

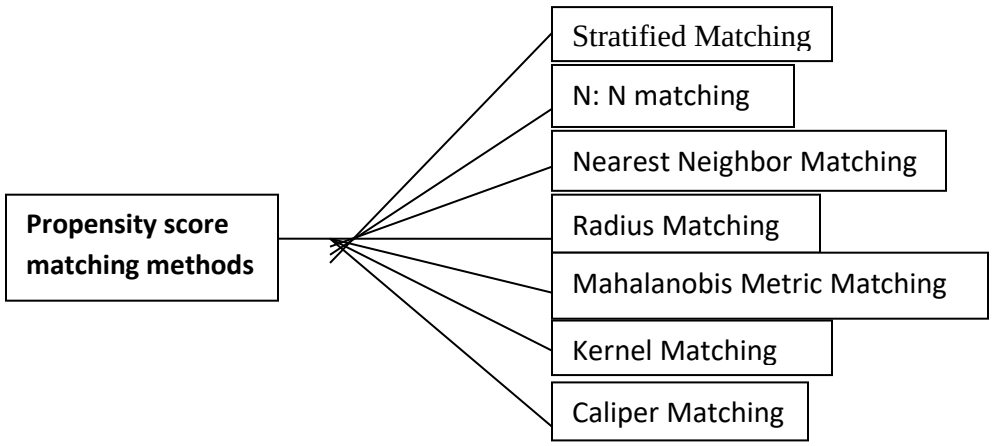


Fig 3: Propensity score matching methods.

(Source:http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_sciences/departmental_units/mchp/protocol/media/propensity_score_matching.pdf)

Stratified Matching:

Propensity scores are often grouped into intervals or strata based on their range of values. Within each interval, or **stratum**, the treatment and control units are expected to have similar propensity scores on average, allowing for meaningful comparisons. To estimate the **average treatment effect (ATE)**, the differences in outcomes between treated and control units are calculated within each stratum. These differences are then averaged, with each stratum weighted according to the proportion of treated units it contains. Dividing the sample into **five strata** (i.e., quintiles of the propensity score distribution) has been shown to remove **over 90–95% of the bias** arising from differences in covariates between groups (Cochran, 1968). This stratification method is a practical and effective way to reduce selection bias in observational studies.

N: N Matching:

In this approach, treatment and control subjects are randomly ordered, and each of the first *n* treated individuals is matched to *n* control individuals with the closest propensity scores. The most commonly used matching ratios include **1:1** (one treated matched to one control), **1:N** (one treated matched to multiple controls), and **N:1** (multiple treated units matched to a single control). These matching strategies help ensure comparability between groups by minimizing differences in propensity scores.

Nearest Neighbor Matching:

The goal is to minimize the **absolute difference** between the estimated propensity scores of treated and control subjects. After randomly ordering the treatment and control groups, each treated individual is matched with a control subject whose propensity score is **closest in value**, forming matched pairs or sets. This process ensures that matched units are as similar as possible in terms of observed covariates, thereby reducing selection bias and improving the validity of causal inference.

$$C(P_i) = \min |P_i - P_j|$$

Where:

C (P_i) indicates the group of control subjects *j* matched to treated subjects *i (*on the estimated propensity score)
The estimated propensity score for the treated subjects *i* denoted by P_i
The estimated propensity score for the control subjects *j* denoted by P_j

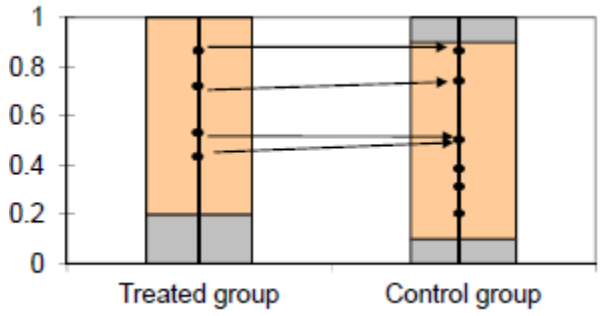


Fig 4: Nearest Neighbor Matching

Source: https://sites.google.com/site/econometricsacademy/econometrics-models/propensity-score-matching.

Radius Matching:

In this approach, each treated individual is paired with a control subject whose propensity score falls within a **specified range** of the treated subject's score. This ensures that matches are made only when the control subject is sufficiently similar to the treated one in terms of observed characteristics. However, because matches are restricted to a defined interval, not all treated subjects may find suitable control counterparts, limiting the number of possible comparisons. This method enhances the quality of matches by avoiding those with large differences in propensity scores.

$$| P_i - P_j | < r$$

Mahalanobis Metric Matching:

- The distance between the treated and control subjects are calculated after randomly ordering the subjects. The distance is:

$$D_{ij} = \sqrt{(x_i - y_j)^T S^{-1} (x_i - y_j)}$$

Where:

- S^{-1} indicates the sample covariance matrix of matching variables from the control subjects.
- x_i and y_i are the matching variable values including the propensity score where *i* represents the treated subjects and *j* the control subjects

Treatment and control units are matched based on the **smallest Mahalanobis distance**, which accounts for the correlations among covariates to identify the closest matches. The process continues until each treated subject is paired with a control subject, and any unmatched controls are excluded from further analysis. If a treated unit does not have a suitable control match, it is also excluded, as drawing causal inferences without a comparable counterpart would require extrapolation, which may compromise validity. This matching method is conceptually similar to **blocking** in randomized experiments, where subjects with similar characteristics are grouped together to control for confounding.

Kernel Matching:

In this method, each treated subject is matched with a weighted average of control subjects. The weights are assigned such that they are inversely proportional to the distance between the propensity scores of the treated and control subjects.

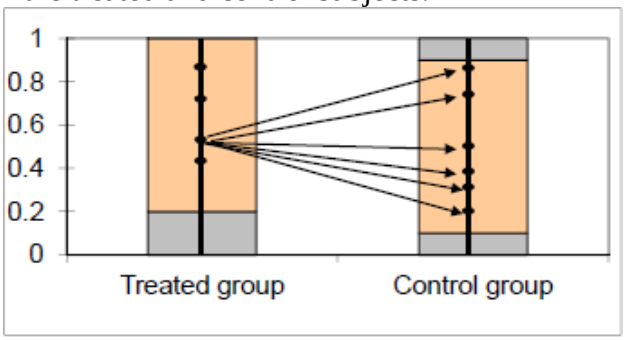


Fig 5: Kernel matching.

Source: https://sites.google.com/site/econometricsacademy/econometrics-models/propensity-score-matching.

$$W\ (i,j) = \frac{K\ (\frac{p_j - p_i}{h})}{\sum_{j=1}^{n_0} K(\frac{p_j - p_i}{h})}$$

Caliper Matching:

As per Sianesi (2002), this method uses a predefined range, typically set within 0.25 of the standard error of the estimated propensity score. Observations falling outside this range are excluded from the analysis.

The range is: $|P_i - P_j| < e$

Where:

Estimated propensity score for the treated subjects is denoted as P_i
Estimated propensity score for the control subjects is denoted as P_j
Pre-determined range of values is determined as *e*.

Step 3: Check Overlap and Common Support

This method is suitable when there is considerable overlap between treated and control groups. It involves checking the difference in mean propensity scores, the variance ratio, and the ratio of covariate residuals between groups. Mean scores should be close, and ratios should be near one. To estimate the treatment effect, the regression model includes both the treatment indicator and the propensity score as covariates, along with any additional relevant variables.

Limitations:

This method requires sufficient overlap between the treatment and control groups. If covariate distributions differ significantly, regression adjustment may be ineffective, as it adjusts toward the mean of the dependent variable, potentially misrepresenting individual group effects. The difference in mean propensity scores between groups should be small (around 0.5s), unless sample sizes are similar and distributions are symmetric with equal variance. Bias correction may be unreliable if the variance ratio of propensity scores is far from one. Likewise, the variance ratio of covariate residuals between groups should be close to one after adjusting for the propensity score.

Step 4: Assess Matching Quality and Estimate Treatment Effect

After matching, it is essential to check if the procedure has successfully balanced the distribution of the relevant covariates between the treated and matched control groups. This is critical because PSM conditions on the propensity score, not directly on each covariate.

Estimating Treatment Effects:

The estimated treatment effect is,

$$\hat{t} = (\bar{Y}_t - \bar{Y}_c) - b\,(\bar{X}_t - \bar{X}_c)$$

It is generally preferable to measure effects from the beginning of the programme to avoid endogeneity issues and align with policy decision-making. However, one must be mindful of potential "locking-in effects" where participants might initially show negative effects (e.g., reduced job search intensity while in a programme) before positive impacts emerge.

Step 4: Propensity Score Estimation: The most common method for estimating propensity scores is logistic regression.

$$\log\frac{e(x)}{1-e(x)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k$$

Where,

X_i are covariates (e.g., age, income)

β_i are coefficients estimated from the data

Step 5: Sensitivity Analysis

Testing the sensitivity of estimated treatment effects is an increasingly important topic in applied evaluation literature. It helps to assess the robustness of results to departures from key assumptions.

Estimating the Propensity Score with an example

$$e\,(X_i) = \Pr(T_i = 1 |\, x_{1,} x_{2,} \ldots, x_{ki})$$

$$\hat{e}(X_i) = \frac{1}{1 + e^{-(a+Sb_k x_k)}}$$

Where

Ti= TRT
T_i =0, without treatment
Ti =1, with treatment
x_{1i}= $SEX_{(1,2)}$
1= female and 2= male
x_{2i}= $AGE_{(CONTINUOUS)}$
x_{3i}= $HYP_{(0,1)}$
HYP=0, without hypertension
HYP=1, with hypertension
So the model becomes,

$$\hat{e}(X_i) = \frac{1}{1 + e^{-(a+b_1 SEX_i + b_2\ AGE_i + b_3 HYP_i)}}$$

If we estimate following parameters using maximum likelihood (ML) techniques:

a= -3.9

b_1= 0.63

b_2= 0.025

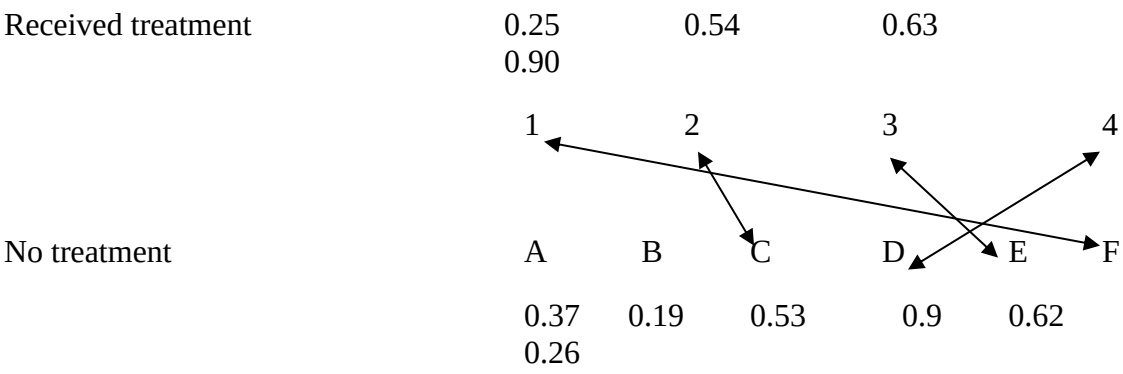
b_3= 0.343

$$\hat{e}(X_i) = \frac{1}{1+e^{-(-3.9+0.63SEX_i+0.025AGE_i+0.343HYP_i)}}$$

Suppose the subject is female (SEX=1), 75 (AGE =75) and has hypertension (HYP = 1).

$$e^{X_i} = \frac{1}{1+e^{-(-3.9+0.63(1)+0.025\ (75)+0.343\ (1))}} = \frac{1}{1+e^{-(-1.05)}} = 0.25$$

After calculating the estimated propensity scores, the treated subjects are matched with the control subjects that have the same/similar propensity score. The unmatched subjects are discarded from the analysis. This example follows a 1-to-1 match:



6. Step-by-step explanation of PSM and implementation in R

The case study evaluates the impact of organic farming (treatment) versus conventional farming (control) on crop yield, using a sample dataset with covariates: **farm size, farmer experience, soil quality, and irrigation access**. a sample dataset of 1000 farmers, with the following variables:

- **farm_size:** Farm size in acres (numeric)
- **experience:** Years of farming experience
- **soil_quality:** Soil quality index (0 to 1)
- **irrigation:** Access to irrigation (binary: 0 = No, 1 = Yes)
- **treatment:** Whether the farmer adopted organic farming (binary: 1 = Yes, 0 = No)
- **yield:** Crop yield (tons/ha), with 0.5 tons/ha added if the farmer adopted organic farming

```
> data
   farm_size experience soil_quality irrigation treatment yield
1       58.2         24        0.087          0         1 3.882
2       49.4         17        0.413          0         0 5.379
3       74.5         27        0.098          1         0 6.902
4       23.3         15        0.916          1         1 6.102
5       75.1         24        0.065          0         1 7.232
6       82.7         26        0.390          1         1 5.353
7       10.9         26        0.805          0         1 5.227
8       28.3         19        0.384          1         0 6.147
9        3.1         23        0.112          1         0 5.505
10      60.0          9        0.160          0         0 5.801
11      98.6         22        0.732          0         1 5.365
12      84.7         19        0.277          0         0 4.952
13      59.6         23        0.990          0         1 5.750
14      35.0         21        0.870          0         0 5.044
15      65.0         26        0.422          1         0 5.791
16       4.7         19        0.966          1         0 4.632
17      12.5         24        0.427          0         1 4.580
18      52.1         19        0.947          1         1 6.980
19      87.3         23        0.799          0         1 4.762
20      11.2         14        0.247          1         1 5.788
```

Figure 1. Sample dataset of 20 farmers.

Step 1: Model the Propensity Score

The first step in PSM is to estimate the propensity score, which is the probability of receiving the treatment given a set of covariates. We use logistic regression to model this probability, where the treatment (e.g., organic farming = 1, conventional = 0) is the dependent variable, and covariates like farm size, experience, soil quality, and irrigation access are predictors. The logistic regression model is:

$$logit\big(P(T = 1 \mid X)\big) = \beta_0 + \beta_1 * farm\ size + \beta_2 * experience + \beta_3 * soil\ quality + \beta_4 * irrigation$$

In the **R code**, we load the necessary packages ("**MatchIt"** for PSM and "**dplyr"** for data manipulation). The dataset includes farm_size (acres), experience (years), soil_quality (index), irrigation (binary), treatment (organic = 1, conventional = 0), and yield (tons/ha), with a 0.5 tons/ha treatment effect added for organic farming. Here, **glm** fits the **logistic regression** model, and **predict** calculates the **propensity scores**, stored in **data$pscore**.

The code for this step is:

```
## Simulated dataset of farmers ##
library(MatchIt)
library(dplyr)
set.seed(123)
n <- 1000
data <- data.frame(
  farm_size = runif(n, 1, 100),
  experience = rnorm(n, 20, 5),
```

```
  soil_quality = runif(n, 0, 1),
  irrigation = rbinom(n, 1, 0.4),
  treatment = rbinom(n, 1, 0.5),
  yield = rnorm(n, 5, 1)
)
data$yield <- data$yield + 0.5 * data$treatment
ps_model <- glm(treatment ~ farm_size + experience + soil_quality + irrigation,
                family = binomial(), data = data)
data$pscore <- predict(ps_model, type = "response")
```

Explanation: We use a logistic regression model to estimate the probability that a farmer adopts organic farming based on their characteristics. This is called the propensity score.

In our example, covariates include farm size, experience, soil quality, and irrigation access. Logistic regression is suitable because the treatment is binary (organic vs. conventional). Alternatives like probit regression or machine learning methods (*e.g.*, random forests) can be used, but logistic regression is simple and widely accepted. In the R code, the glm function with family ***= binomial()*** fits this model, and predict with type = "response" converts log-odds to probabilities (propensity scores).

Step 2: Match Individuals

After estimating propensity scores, we match treated and untreated individuals with similar scores to create comparable groups. The R code uses nearest-neighbor matching with a caliper of 0.1, meaning each treated farmer is paired with a control farmer whose propensity score is within 0.1 units. The caliper ensures high-quality matches but may exclude some treated units if no suitable control is found. The code is:

```
# Step 2: Perform matching using MatchIt
# Nearest neighbor matching with a caliper of 0.1
match_obj <- matchit(treatment ~ farm_size + experience + soil_quality + irrigation,
                     data = data,
                     method = "nearest",
                     caliper = 0.1,
                     ratio = 1)
matched_data <- match.data(match_obj)
```

Explanation: We match each treated (organic) farmer to a control (conventional) farmer with a similar propensity score (within a caliper of 0.1). This forms a balanced dataset.

The **matchit** function from the "**MatchIt" R package** performs the matching, specifying the treatment and covariates. The method = "nearest" indicates nearest-neighbor matching, caliper = 0.1 sets the maximum allowable difference in propensity scores, and ratio = 1 ensures one control per treated unit. The **match.data** function extracts the matched dataset, which includes only the paired individuals. This matching reduces bias by ensuring treated and control groups have similar covariate distributions, mimicking randomization.

Step 3: Check Balance

To ensure the matching was successful, we check whether the covariates are balanced between treated and control groups. A common metric is the Standardized Mean Difference (SMD), calculated as:

$$SMD = \frac{|\bar{X}_T - \bar{X}_C|}{\sqrt{(s_T^2 + s_C^2)/2}}$$

where $\bar{X}_T$ and $\bar{X}_C$ are the means of a covariate for the treated and control groups, and s_T^2 and s_C^2 are their variances. An SMD < 0.1 indicates good balance. We also use visual tools like ***jitter plots*** and ***histograms*** to assess propensity score overlap. The code is:

```
summary(match_obj)
par(mfrow = c(1, 2))
plot(match_obj, type = "jitter", main = "Propensity Score Distribution")
plot(match_obj, type = "histogram", main = "Propensity Score Histogram")
```

Explanation: *We examine whether matching has balanced the covariates using:* ***SMD****: A value < 0.1 indicates good balance and* ***Plots****: Show score overlap between groups.*

The summary(match_obj) output shows SMDs before and after matching. We should look for SMDs below 0.1 post-matching, indicating that covariates like farm size and soil quality are balanced. The jitter plot shows individual propensity scores, with overlap suggesting good matching, while the histogram compares score distributions. It is to be noted that the poor balance may require adjusting the matching method or adding more covariates to the propensity score model.

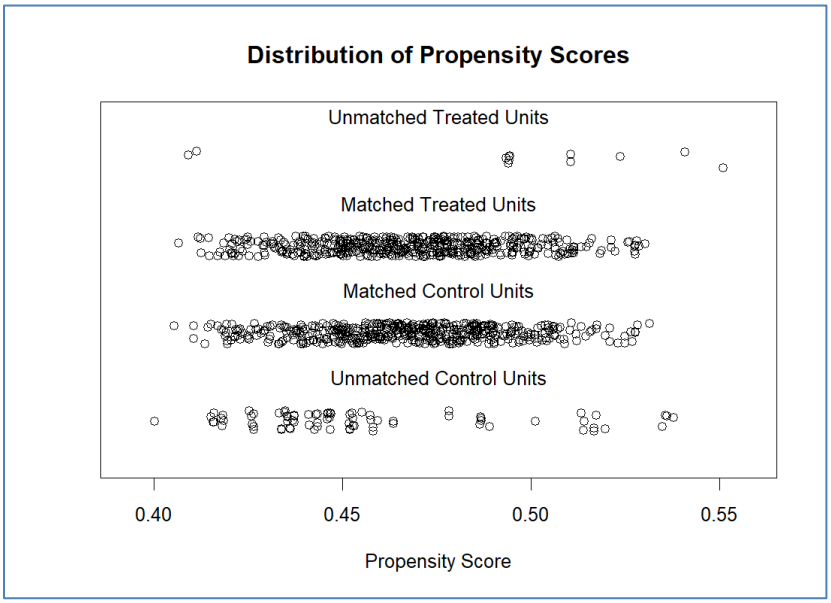


Figure 2. Jitter plot of propensity scores.

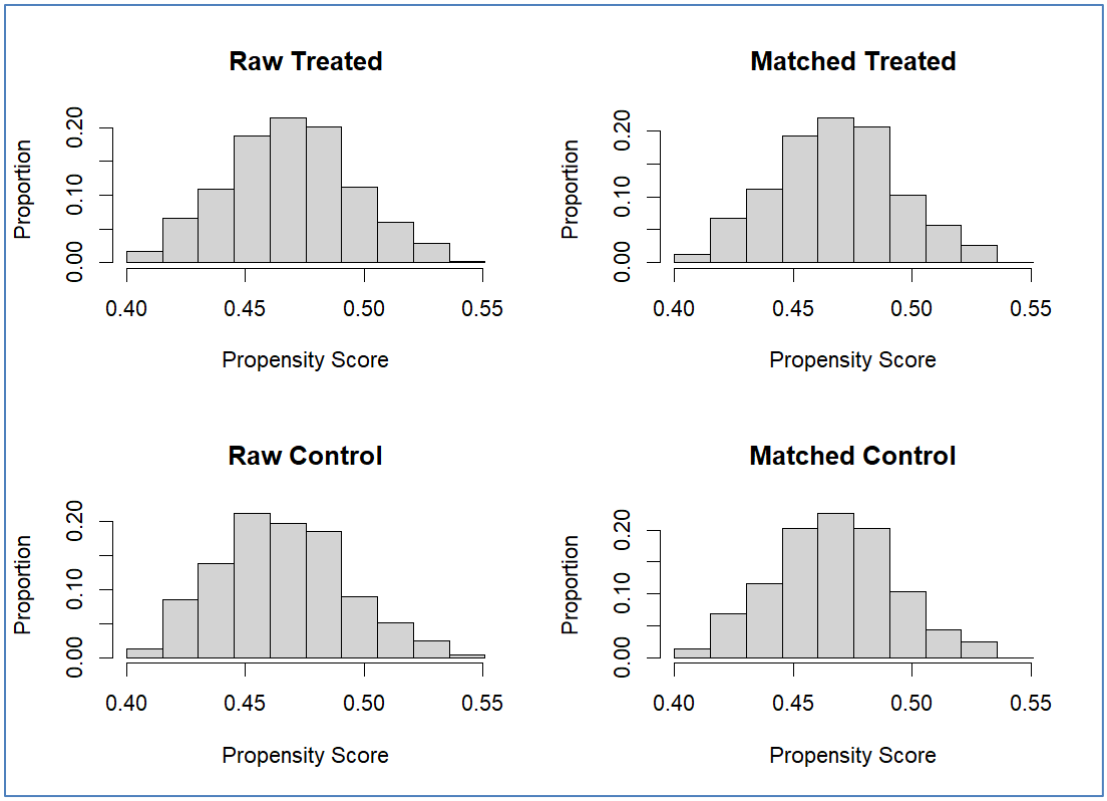


Figure 3. Histogram of propensity scores.

Step 4: Estimate Treatment Effect

After confirming balance, we estimate the treatment effect by comparing the outcome (yield) between matched groups. The code calculates the mean yield for each group and performs a t-test to check significance:

```
# Step 4: Estimate treatment effect
# Compare mean yield between treated and control groups in matched data
effect <- matched_data %>%
  group_by(treatment) %>%
  summarise(mean_yield = mean(yield))
print(effect)
t_test <- t.test(yield ~ treatment, data = matched_data)
print(t_test)
```

Explanation: *We calculate the* ***average treatment effect on the treated (ATT)*** *by comparing the mean yield between organic and conventional groups using a* ***t-test****.*

The code groups the matched data by treatment and computes the mean yield. In this example, organic farming might yield 5.5 tons/ha, while conventional yields 5.0 tons/ha, suggesting a 0.5 tons/ha increase. The t-test checks if this difference is statistically significant (p-value < 0.05).

Additionally, we use regression adjustment to account for any residual imbalances:

```
# Perform a t-test to check if the difference is significant
model <- lm(yield ~ treatment + farm_size + experience + soil_quality + irrigation,
            data = matched_data)
```

```
summary(model)
```

Explanation: *We can further refine the treatment effect estimate using regression on the matched sample, controlling for covariates.*

The linear model estimates the treatment effect while controlling covariates. The coefficient for treatment (e.g., 0.48) approximates the average treatment effect on the treated (ATT). This step quantifies the intervention's impact, and regression adjustment can improve precision by addressing minor imbalances.

Interpretation of the PSM results

In this example, we aimed to estimate the causal effect of a treatment (*e.g.*, agricultural intervention) on crop yield using Propensity Score Matching (PSM). The logistic regression model estimated each farmer's propensity to receive the treatment based on covariates including farm size, farming experience, soil quality, and irrigation availability. The output dataset shows propensity scores ranging from approximately 0.41 to 0.55, indicating the predicted probability of each farmer adopting organic farming. For example, the first farmer (farm_size = 58.2, experience = 24, soil_quality = 0.087, irrigation = 0, treatment = 1) has a propensity score of 0.491, suggesting a 49.1% chance of adopting organic farming based on their characteristics. Using nearest neighbor matching with a caliper of 0.1, 455 treated units were successfully matched with 455 control units, ensuring strong overlap in propensity scores. Balance statistics showed significant improvement post-matching, with standardized mean differences across covariates nearing zero, indicating good covariate balance. Visual checks through jitter plots and histograms confirmed overlapping propensity score distributions, reflecting adequate common support. The treatment effect was then estimated by comparing crop yields between the matched groups. The result shows that the mean yield for conventional farmers (treatment = 0) is **5.05** tons/ha, while for organic farmers (treatment = 1), it is **5.49** tons/ha, suggesting a 0.44 tons/ha increase for organic farming. A Welch two-sample t-test confirmed that this difference was statistically significant ($t = -6.55, p < 0.05$), with a 95% confidence interval ranging from -0.56 to -0.30. This suggests that the treatment led to an average yield increase of approximately 0.43 units. Hence, the intervention had a **positive and significant impact on yield**, after accounting for selection bias through matching.

7. Limitations of PSM

- This method requires a **large sample size** to be effective.
- Since **randomization is absent** and propensity scores are based on observational data, **bias from unobserved covariates** may persist, as matching only accounts for observed variables.
- To support strong causal inference, there must be **sufficient overlap** in propensity scores between treated and control groups. The method becomes ineffective when subjects with high propensity scores (treated) are compared to those with low scores (untreated).
- There are **no universally accepted guidelines** for selecting variables in the propensity score model. Some researchers include variables predicting treatment, others include those related to the outcome, and some include only variables associated with both treatment and outcome.

Characteristics of a Good PSM

- Variables are measured reliably and accurately.
- Substantial overlap between the treatment and control groups on the propensity scores.
- Covariates of the treated and untreated subjects are balanced adequately in the model.
- Minimizes group differences of many variables and adjusts for selection bias**.**
- Variables are chosen and adjusted for propensity scores based on logic, theory and empirical evidence

8. Conclusion

Propensity Score Matching (PSM) is a valuable and flexible statistical approach that allows agricultural researchers to estimate causal effects using observational data, particularly when randomized controlled trials are impractical. By balancing observed covariates between treated and control groups, PSM offers a solid foundation for assessing the impacts of interventions such as organic farming, irrigation practices, or the adoption of improved crop varieties on key outcomes like yield, income, and soil quality.
The method involves a series of steps-estimating propensity scores, matching individuals, evaluating covariate balance, and estimating treatment effects-all of which can be efficiently carried out in R using tools like the **MatchIt** package, as shown in the example. This approach supports researchers in making evidence-based inferences, aiding sustainable agricultural practices and policy formulation. However, the validity of PSM depends on careful **selection of covariates**, thorough **assessment of balance**, and an awareness of limitations, particularly **unobserved confounding**, which may still bias results despite matching.

Reference:

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3), 399-424.

Gorst, Ashley, Ben G., And Ali D., 2015, Crop productivity and adaptation to climate change in Pakistan. Centre for Climate Change Economics and Policy, *Working Paper No. 214.*

Inacio, M., C., S., Yuexin, C., Elizabeth W., Paxton, Robert S., Namba, Steven, M., Kurtz, And Guy C., 2015, Statistics in Brief: An Introduction to the Use of Propensity Scores. *Clinical Orthopaedics and Related Research*, 1-5.

Rijn, V., Fédes, Ephraim N., And Adewale A., 2015, The impact of agricultural extension services on social capital: an application to the Sub-Saharan African Challenge Program in Lake Kivu region. *Agriculture and Human Values*: 1-19.

Rosenbaum, P., & Rubin, D. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–50

SWOT Analysis: Tool for Strategic Planning in an Organization

K. Ravi Kumar, Ponnaganti Navyasree, Nobin Chandra Paul, Santosha Rathod, and Prabhat Kumar
ICAR-National Institute of Abiotic Stress Management, Baramati-413115
Email: kure.ravi@gmail.com

Introduction

In today's highly competitive environment, **micro-level planning, implementation, monitoring, and evaluation** are crucial for effective utilization of organizational resources. **Strategic planning** facilitates the systematic allocation of these resources across various operational and management levels. This process typically begins with defining the **vision, mission, goals, objectives**, and identifying appropriate methods to achieve desired outcomes. One widely used strategic tool is the **SWOT analysis**, which examines both internal (strengths and weaknesses) and external (opportunities and threats) factors influencing an organization or enterprise. The **origin of SWOT** can be traced back to the early 1950s, during a period of intense competition between the U.S. government and large corporate entities. Between 1948 and 1950, U.S. federal agencies—including the Justice Department, the Federal Trade Commission, and Congress—began scrutinizing large multinational corporations (Feinstein, p. 161). In this context, Stewart proposed a framework where organizations are composed of independent systems, structures, and procedures, allowing multiple internal and external stakeholders to contribute to strategic reasoning (Steiner & Cannon, 1966, p. 302). The technique gained popularity in the 1960s and 1970s, largely due to the efforts of **Albert Humphrey**, a management consultant at Stanford Research Institute (SRI), who initially developed the **SOFT analysis**-where 'S' stood for satisfactory, 'O' for opportunity, 'F' for fault, and 'T' for threat. This eventually evolved into the modern SWOT framework. **SWOT analysis** serves as a valuable tool for both **strategic planning and formative evaluation**. It provides insights into critical organizational aspects such as planning, staffing, coordination, control, and budgeting. By considering the organization as an interactive system comprising multiple dimensions and subsystems, SWOT helps in understanding the broader economic, social, political, and environmental context. It supports comprehensive decision-making and must be grounded in evidence from **case studies, market trends, forecasts**, and financial data. The primary advantage of using SWOT lies in its ability to create actionable strategies that leverage an organization's strengths-such as brand recognition, technology adoption, and efficient resource management-while also highlighting weaknesses like high operational costs, unfavorable locations, or underperforming research units, particularly in profit-driven enterprises. The insights gained from SWOT are instrumental in guiding decisions related

to new business strategies, product launches, market expansion, navigating competition, adapting to regulatory changes, and embracing technological innovation.

The external analysis is necessary to identify the probable threats and opportunities in competitive environment. The history and trend of competitive markets and how is has evolved over a period of time and a special preference towards economic crisis.

The process of SWOT analysis involves four components i.e., Strengths, Weakness, Opportunities and Threats. Which can be grouped into two dimensions Strengths and Weakness are considered as internal factors which mostly depends on organization overall structure and functioning, whereas opportunities and threats as external factors which is influenced by environment. For summary purpose it is drawn in four quadrant boxes.

Components of SWOT Analysis

Organization Strength	Characteristics that are relatively better than others in the global / national / regional market.
Organization Weakness	Characteristics that are negative which are disadvantage to the organization.
Organization Opportunities	External factors that will give benefit to the organization growth and development.
Organization Threats	External factors that could harm the organization growth and development.

Framework of SWOT

Organizational Strengths and Weaknesses

Advertisement	Information technology	Resources
Popular Brand	Leadership	Research and development
Competency of staff	Location	Reengineering
Distribution network	Management	Services
Decentralization	Organization structure	Technologies
Efficient system	Physical facilities	
Costs	Product quality	
Forecasting	Quality management	

Organizational Opportunities and Threats

Competition	Political factors	Technological factors
Judiciary system	Religious factors	
Market fluctuations	Social values	

Organizational Strengths:

Strength refers to any attribute or capability that enhances an organization's performance and efficiency, giving it a competitive advantage over others. These are favorable internal characteristics that contribute to innovation, effectiveness, and overall success. The effectiveness and efficiency of an organization largely depend on how well it utilizes its strengths. An organization can be considered strong, average, or weak in comparison to its competitors based on several criteria, including relative market position, financial structure, production and technical capacity, research and development potential, and human resource and managerial effectiveness (Dinçer, 2007: 145). Strength is essentially a distinctive competence that provides the organization with a comparative edge in the marketplace. As noted by Pearce and Robinson (1991: 182), strengths may relate to various areas such as financial resources, brand image, market leadership, and relationships with buyers or suppliers.

Organizational Weaknesses:

Weakness refers to deficiencies in competencies and structure of the organizations. Weaknesses is something relatively disadvantageous and negative characteristic which limits the potential functioning of operation. It can be skills, resources, facilities, budget and some times and brand image can be sources of weaknesses" (Pearce and Robinson, 1991: 182).

Both strengths and weakness analysis is needed to understand the efficient and inefficient quality as well as negative and positive characters of the organisations, so that authority could able to decide upon building efficient system.

Environmental Opportunities:

Opportunity refers to a situation or condition that is favorable for initiating or advancing an activity. It serves as an advantage and acts as a motivating factor, carrying a positive and beneficial nature. Opportunities arise from external environmental conditions and, when identified through analysis, can lead to favorable outcomes for an organization (Emet Gürel and Merba Tat, 2017: 998). To effectively leverage these opportunities, organizations must align them with their strategic objectives and determine the most appropriate course of action to enhance their competitive position.

Environmental Threats:

Threat refers to a situation or condition that poses a risk to the successful execution of an activity. It represents a disadvantage or external challenge that can hinder organizational performance or strategic goals. Due to its harmful and unfavorable nature, a threat is considered a negative factor that organizations should seek to anticipate, mitigate, or avoid (Emet Gürel and Merba Tat, 2017: 998).

Advantages of SWOT analysis:

The process of making plans or decisions is essential for both managers and employees within any organization. **SWOT Analysis** is a widely used and trusted method in strategic management and marketing, valued for its simplicity and effectiveness (Emet Gürel and Merba Tat, 2017: 998). It serves as a fundamental tool for analyzing internal and external factors that influence organizational success. However, the **categorization of variables** into the four SWOT quadrants-**Strengths, Weaknesses, Opportunities, and Threats**-can be complex. A single factor may fit into more than one category; for example, it can be viewed as both a strength and a weakness depending on context. Similarly, a current strength may become a weakness if not sustained, and opportunities that are ignored but seized by competitors can evolve into threats. The classification also varies depending on the **specific objective** of the analysis, making it a subjective and context-dependent exercise (Emet Gürel and Merba Tat, 2017: 998).

Conclusion:

In any business planning or entrepreneurial process, the assessments of social, economic, environmental factors is considered to be crucial exercise done mostly in initial phase of business development. SWOT analysis facilitates in strategic planning and decision making in development of plans short term and long term, allocation of resources, Human resource development and specific intervention needed for business development strategies. Although there are few limitations such as overlapping of factors based on situations, over generalization of characteristics, critics on categorization of factors into four quadrants which is mostly challenging and since it's a dynamic process it can interchange based on the situation hence flexibility needed for changing according to market situations. This tool is been used from decades which contributed significant for development of organizations. Identifying shortcomings and developing a actionable plan or strategy is the critical phase in any process. It helps in analyzing the present situation and facilitate in developing future action plans. SWOT analysis lays down foundation for strategic planning ensuring business success positioning efficiently itself in the market while overcoming internal and external factors that impact the organizational growth. Additionally, it is crucial for aligning organizational strategies with overall mission and vision which is vital for achieving long term goals.

References

DİNÇER, Ö. (2007). Stratejik Yönetim ve İşletme Politikası, (8. Baskı), İstanbul: Alfa Basım Yayım.

Fligstein, N. (1993). *The transformation of corporate control*. Harvard University Press.

GÜREL, E. and TAT, M. (2017). SWOT ANALYSIS: A THEORETICAL REVIEW. The Journal of International Social Research. **10.** 994-1006.

https://bschool.pepperdine.edu/personal-growth/article/best-practices-for-successful-swot-analysis.htm, Retrieved on 10 July, 2025.

https://quantive.com/resources/articles/swot-analysis, Retrieved on 10 July, 2025.

PEARCE, J. A. & ROBINSON, R. B. (1991). Strategic Management, (4th Edition), USA: Irwin, Inc

Puyt, Richard & Lie, Finn & Wilderom, Celeste. (2023). The origins of SWOT analysis. Long Range Planning. 102304. 10.1016/j.lrp.2023.102304.

Steiner, G. A., & Cannon, W. M. (1966). Multinational corporate planning.

Scientometric Analysis in Agriculture and Allied Sectors

Sivakumar S and Ramasubramanian V
ICAR-National Academy of Agricultural Research Management
Hyderabad-500030
Email: ram.vaidhyanathan@gmail.com

Introduction

Scientometrics is a broader field that specifically focuses on the quantitative study of science, including the production, dissemination, and use of scientific knowledge across various disciplines. It aims to understand the dynamics of scientific research and innovation, including how knowledge is created, shared, and applied. It encompasses not only publications but also research funding, patents, collaborations, and institutional activities. Scientometric applications used in science policy, research funding decisions, and understanding the growth and impact of scientific disciplines. The term scientometrics was coined in the 1960s by Nikolay Vasilyevich Pukinel in the Soviet Union, though the field gained significant traction after the 1970s. Boris S. Zuckerman and Nikolay S. Smirnov were important early contributors to the field, establishing scientometrics as a discipline that extended beyond bibliometrics, incorporating the analysis of research policies, the dynamics of scientific collaboration, and the economics of science.

Bibliometrics is the statistical analysis of written publications, such as books, articles, and other forms of literature. It is used to assess the impact of publications, track trends in research, and evaluate the performance of individual researchers, institutions, or journals. Bibliometric includes citation counts, h-index, impact factor, and publication frequency to evaluate the performance of journals, authors, or institutions. Bibliometrics often looks at a narrower view, focusing mainly on publications and their citations within the field of literature and information science. The term bibliométrie was indeed first introduced by Paul Otlet in 1934. Otlet, a Belgian bibliographer and information scientist, defined bibliométrie as the "measurement of all aspects related to the publication and reading of books and documents." The term bibliometrics was first introduced by Alan Pritchard in 1969 in his seminal paper titled "Statistical Bibliography or Bibliometrics?". In this paper, Pritchard proposed the use of statistical methods to analyse written publications, laying the foundation for what we now know as bibliometrics. His use of the term bridged the gap

between the earlier theoretical ideas of Otlet and the more systematic, empirical approach that characterizes modern bibliometric analysis.

Scientometric analysis refers to the quantitative evaluation of scientific literature to understand research trends, productivity, collaboration, and influence within a specific field. Originating from the broader domain of bibliometrics, scientometrics particularly focuses on measuring and analyzing scientific publications using tools such as citation analysis, co-authorship mapping, and keyword co-occurrence networks (Hood & Wilson, 2001). In recent decades, scientometric methods have emerged as crucial tools for understanding research dynamics in agriculture and allied sectors such as horticulture, animal husbandry, fisheries, forestry and agricultural engineering.

In the field of agriculture, scientometric analysis provides valuable insights into the evolution of research topics, institutional productivity, collaborative patterns, and the impact of research outputs. It enables stakeholders, including policymakers, funding bodies, and researchers, to evaluate the effectiveness of national and international research efforts. Studies by Garg and Kumar (2014) and Chaurasia et al. (2018) have revealed that agriculture research in India has experienced significant growth in recent decades, especially in the domains of crop science, biotechnology, and sustainable agriculture. These findings underscore the strategic role of scientometric studies in identifying research gaps, prioritizing funding, and promoting effective dissemination of agricultural innovations.

Theoretical Framework and Metrics

Foundational Scientometric Laws

1. **Bradford's Law (Journal Dispersion):** Bradford's Law describes how articles on a specific subject are distributed across scientific journals. It states that a core group of journals will produce the majority of articles on a topic, while a larger group will produce fewer. This law is useful in identifying core journals in a research domain.

Example: In agricultural sciences, journals like *Field Crops Research* or *Agricultural Systems* may appear in the core zone.

2. **Lotka's Law (Author Productivity):** Lotka's Law states that the number of authors publishing *n* papers is inversely proportional to n^2. That means, for every 100 authors publishing one paper, about 25 will publish two, 11 will publish three, and so on. It helps assess author contribution inequality.

Application: Identifies the small percentage of highly productive researchers in agriculture.

3. **Price's Law (Growth of Scientific Knowledge):** Price's Law suggests that scientific output grows exponentially, but only a small number of researchers contribute to half of the publications in a field. This law provides a framework for understanding the evolutionary nature of research growth.

Relevance: Highlights the concentration of agricultural research in elite institutions or countries.

Key Scientometric Metrics

1. **h-index:**

The h-index measures both productivity and citation impact of a researcher. A scholar has an h-index of *h* if *h* of their *N* papers have at least *h* citations each.

Example: An author with an h-index of 15 has 15 papers each cited at least 15 times.

2. **g-index:**

The g-index improves upon the h-index by giving more weight to highly cited papers. It is defined such that the top *g* articles together have at least g^2 citations.

Example: A researcher's top 10 articles have a total of 100 citations → g-index = 10.

3. **M index**:

The m-index is a metric used to assess the research productivity and impact of an academic researcher, particularly over time. It is related to the h-index, but while the h-index measures both productivity (the number of publications) and citation impact (the number of citations those publications have received), the m-index normalizes the h-index over time to account for the researcher's career length. The m-index is calculated as the h-index divided by the number of years since the researcher's first publication. The formula can be expressed as:

m index= h index / years since first publication

If a researcher has an h-index of 20 and has been published for 10 years, their m-index would be: 20/10=2.

4. **i10-index:**

This metric is specific to Google Scholar and counts the number of publications with at least 10 citations.

Example: A scientist with 25 papers cited at least 10 times each has an i10-index of 25.

5. **Total Citations:** The total number of times a researcher's or institution's work has been cited.

Used for: General impact assessment.

6. **Average Citations per Document:** The average number of citations received per publication. This metric reflects the per-paper impact.

Example: 200 citations across 50 publications = average 4 citations per paper.

7. **Degree of Collaboration:** It is the proportion of co-authored publications compared to total publications. It measures collaborative behaviour in research.

Formula: DC = (Number of co-authored papers) / (Total papers).

8. **Co-authorship Index:** Indicates the average number of authors per publication. It reveals collaboration intensity within the field.

Example: High co-authorship is common in multi-institutional agriculture research project

Relational and Network-Based Metrics

1. Co-citation: Co-citation occurs when two documents are cited together by a third document. If this happens frequently, the two are likely related in topic or conceptually linked.

Used for: Discovering research clusters or scientific schools of thought.

2. Bibliographic Coupling: When two documents cite the same third source, they are said to be bibliographically coupled. This implies a similarity in their knowledge base.

Used for: Detecting research communities working on similar problems.

3. Thematic Evolution: This involves tracking how keywords or topics evolve over time. It shows the emergence, growth, decline, or merging of research themes.

Application: Tracing how topics like "climate-smart agriculture" have grown over the years.

4. **Burst Detection:** This technique identifies sudden increases in the frequency of terms, citations, or topics — signalling emerging trends.

Example: A burst in the keyword "AI in agriculture" might indicate a new research front.

Popular Tools Used in Scientometric Analysis

1. **VOS viewer:** A free tool used for creating maps based on network data, particularly for co-authorship, co-citation, and keyword co-occurrence.

Strength: Excellent for visualizing large bibliometric networks.

2.**Cite Space:** Java-based software for visualizing and detecting emerging trends and research fronts over time.

Special Feature: Temporal bursts and cluster labelling.

3. **SciMAT** (Science Mapping Analysis Tool): Focuses on longitudinal analysis of themes and helps in understanding thematic evolution.

Application: Used to track how a field like "organic farming" has changed over 30 years.

4. **Bibliometrix (R Package):** An R-based software for comprehensive bibliometric analysis. It also offers a web interface called *Biblioshiny*.

Strength: Fully customizable, supports advanced statistics.

5. **Gephi:**

An open-source network analysis tool, especially used for large and complex networks. Commonly used for collaboration networks.

Application: Visualizing institutional or country-level research collaborations.

Scopus and Web of Sciences

Scopus and Web of Science are two of the largest and most widely used academic databases that provide access to scholarly articles, journals, conference papers, patents, and other academic content. Both databases are essential tools for researchers, institutions, and academics to search for literature, track citations, and measure research impact.

Scopus is owned by Elsevier and covers a wide range of disciplines, including science, technology, medicine, social sciences, and arts & humanities. It includes over 23,000 journals, conference proceedings, and other content types. Scopus provides citation analysis, author profiles, and metrics like the h-index. It also offers tools for tracking citation trends and measuring research performance. Scopus is known for its

comprehensive indexing of conference proceedings, which makes it especially valuable for those in fast-evolving fields where conference presentations are important.

Web of Science is owned by Clarivate Analytics (formerly Thomson Reuters). The web of Science includes journals across a variety of disciplines, with a strong emphasis on science, social sciences, arts & humanities, and indexes over 21,000 journals, books, and patents.

Features are similar to Scopus, Web of Science provides citation analysis, author and institutional profiles, and impact metrics. It includes a powerful citation search tool, which allows users to track the number of articles that have been cited and explore citation relationships. The Web of Science is particularly well-regarded for its rigorous selection process for journal inclusion, meaning the journals indices are often considered to be of high scholarly quality.

Benefits of Scientometric Analysis

Scientometric analysis helps identify areas where further research is needed, enabling targeted investments in agricultural research and development. By analyzing citation patterns and research output, scientometric analysis can assess the impact of research in agriculture and allied sectors. Scientometric analysis can provide insights into emerging trends and areas of research focus, informing policy decisions and resource allocation. Data analytics helps farmers identify optimal conditions for each crop variety and field location, leading to improved yields and reduced waste. Technologies like drones, sensors, and satellite imagery enable farmers to make data-driven decisions, optimizing resource use and maximizing yields.

Analytics helps farmers reduce environmental impact through targeted interventions, such as precision irrigation and fertilizer application.

Scientometric analysis in agriculture and allied sectors involves evaluating research activity and productivity in these fields. Several scientometric studies have used databases such as Scopus, Web of Science, CAB Abstracts, and AGRIS to analyse agricultural research output. For example, Singh et al. (2022) conducted a comprehensive analysis of Indian agricultural research using Scopus data and found that institutions of Indian Council of Agricultural Research (ICAR) like Indian Agricultural Research Institute (IARI), and State Agricultural Universities like Tamil Nadu Agricultural University (TNAU) were among the most prolific contributors. The study also indicated an increase in international

collaboration, particularly with the United States, China, and European countries, driven by global issues like food security and climate change.

Methodologically, scientometric studies in agriculture involve several analytical techniques, including citation analysis (Garfield, 1979), co-authorship networks (Newman, 2001), keyword co-occurrence analysis (Callon et al., 1983), and productivity analysis. Tools like VOS viewer (Van Eck & Waltman, 2010), CiteSpace (Chen, 2006), and Bibliometrix in R are widely used to visualize and interpret the data. These tools help identify core authors, influential journals, thematic clusters, and emerging research fronts. For instance, Das and Mishra (2020) used VOS viewer to analyse the growth of literature in climate-smart agriculture, highlighting the increasing attention to sustainable farming practices and adaptation strategies.

The scope of scientometric studies in agriculture has expanded to cover diverse subfields. In crop science, researchers analyse the publication trends in precision farming, plant breeding, and genetic engineering. In horticulture, analysis often focuses on high-value crops such as mango, banana, and tomato. Animal sciences have seen increasing publications in areas like dairy management, animal health, and feed technology (Kumar & Garg, 2021). Similarly, research in soil science, forestry, and fisheries is being evaluated to understand their contributions to ecological sustainability and economic development.

India has demonstrated significant research output in areas such as rice and wheat breeding, soil nutrient management, and agrometeorology. However, several studies (e.g., Raj, 2019; Singh et al., 2022) point out under-researched areas like post-harvest management, value chain development, and farmer-centric technological adoption. There is also a need for better integration of socio-economic factors in agricultural research, especially in domains like agricultural extension, market intelligence, and rural livelihood resilience.

Challenges in conducting scientometric analysis in agriculture include language bias (most databases prioritize English publications), name variations among authors and institutions, limited access to full-text articles, and underrepresentation of regional or non-indexed journals (Moed, 2005). Moreover, research outputs from low-income countries are often excluded from global scientometric reviews, thereby limiting the comprehensiveness of such studies.

Despite these limitations, scientometric analysis offers significant opportunities for enhancing agricultural research and development. The integration of altmetrics which includes mentions in news, social media, and policy documents provides an additional layer of understanding beyond citations. The use of machine learning and natural language processing (NLP) is also gaining popularity in processing large volumes of agricultural texts to extract themes and sentiments (Zhang et al., 2021). With the growing emphasis on interdisciplinary research, scientometric approaches can help bridge gaps between agricultural sciences and other domains like environmental science, economics, and public policy.

Limitations and Ethical Considerations

While scientometric analysis provides critical insights, several limitations exist:

- Underrepresentation of non-English and regional-language research
- Over-dependence on citation counts, which may not reflect true impact in applied fields
- Algorithmic bias in databases and tools
- Ethical issues in data scraping or misrepresentation of co-authorship networks

Researchers are encouraged to triangulate findings with qualitative assessments and expert judgment.

Conclusion

Scientometric analysis in agriculture and allied sectors is a powerful tool for assessing research productivity, detecting trends, and informing strategic decisions. The increasing complexity of agricultural challenges ranging from food insecurity and climate change to resource degradation requires a robust and evidence-based research ecosystem. Scientometric studies, supported by reliable databases and advanced analytical tools, can play a pivotal role in shaping this ecosystem. As agriculture moves towards digitization and data-driven innovation, scientometric insights will become indispensable for optimizing research investments, fostering international collaborations, and guiding national research agendas in a sustainable and inclusive manner.

References

Hood, W. W., & Wilson, C. S. (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, 52(2), 291–314.

Garfield, E. (1979). Citation indexing: Its theory and application in science, technology, and humanities. *John Wiley & Sons*.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *PNAS*, 98(2), 404–409.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research. *Scientometrics*, 22(1), 155–205.

Singh, A. K., Kumari, R., & Meena, H. R. (2022). Scientometric analysis of Indian agricultural research: A Scopus-based study. *Journal of Scientometric Research*, 11(3), 148–157.

Raj, N. (2019). Bibliometric mapping of Indian agricultural research journals. *Library Philosophy and Practice (e-journal)*, 2642.

Das, A., & Mishra, R. (2020). Growth of climate smart agriculture literature: A scientometric analysis. *Journal of Agricultural Sciences*, 12(4), 54–63.

Kumar, S., & Garg, K. C. (2021). Trends in global agricultural research outputs: A bibliometric study. *Current Science*, 121(7), 951–957.

Moed, H. F. (2005). Citation analysis in research evaluation. *Springer*.

Zhang, L., Chen, C., & Xu, Y. (2021). Artificial intelligence and agriculture: A scientometric review. *Computers and Electronics in Agriculture*, 189, 106381.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 359–377.

Time Series Forecasting using Grey Model

Mrinmoy Ray
AKMU, ICAR-IARI, New Delhi-110012
Email: mrinmoy@iari.res.in

1. Introduction:

Forecasting in agricultural systems plays a crucial role in effective planning. In the field of time series analysis, various statistical approaches have been developed for modeling and forecasting in different domains. However, each modeling approach has its own set of advantages and limitations. One common limitation shared by many time series models is the requirement of a substantial number of observations, typically 50 or more, to construct an accurate model. Unfortunately, there are situations where we are only able to gather a limited number of observations due to societal constraints, rapid technological advancements, or evolving policies. This poses a challenge when applying conventional time series models. To address this issue, the grey model, introduced by Deng in 1982, offers a non-traditional forecasting technique based on scarce and fuzzy information. The grey model utilizes differential equations to describe future tendencies within a time series. One notable advantage of the grey model is its ability to generate reliable forecasts even when only a few observations, as few as four, are available for the prediction process. This flexibility makes the grey model a valuable tool in situations where data scarcity or limited historical records hinder the application of traditional time series models. The grey model provides an innovative approach to time series forecasting, particularly in cases where conventional models face limitations due to a small number of observations. By leveraging scarce and fuzzy information, the grey model offers a viable solution for forecasting in agricultural systems and other domains with limited data availability.

The grey model finds applications in various domains, showcasing its versatility and usefulness. Numerous studies have demonstrated the effectiveness of the grey model in different fields. For instance, it has been employed for high precision forecasting, as shown by Lin et al. in 2001. The model has also been utilized in estimating vehicle fatality risk, as demonstrated by Mao and Chirwa in 2006. Additionally, the grey model has proven valuable in time series prediction, as highlighted by Kayachan et al. in 2010. Moreover,

the grey model has been successfully employed in diverse areas such as predicting the growth trend of internet users, revenue, and the online game industry, as explored by Chang et al. in 2013. It has also been applied to forecast the output of the integrated circuit industry, as researched by Wang and Hsu in 2007.Beyond its applications in specific domains, the grey model has found utility in other research fields, including linear planning, financial forecasting, and tourism demand analysis. Notably, one of the strengths of the grey model lies in its simplicity of computation. Conventionally, the Ordinary Least Square (OLS) technique is employed to estimate the parameters of the grey model, further emphasizing its simplicity and ease of implementation. In summary, the grey model's wide range of applications in various domains, coupled with its straightforward computation procedure using techniques like OLS, make it a valuable tool for forecasting and analysis in fields such as linear planning, finance, tourism, and beyond.

2. Grey Model (GM)

Following Mao and Chirwa (2006), Grey model with One variable in First order can be represented as GM (1,1). Estimation of GM (1,1) by employing conventional approach is as follows:

Step 1: Let the actual time series observations be:

$x^{(0)} = \{x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(k)\}$ where k is the number of observations (1)

Step 2: A new time series can be obtained by using the first-order Accumulated Generating Operation (1-AGO) on $x^{(0)}$, i.e.

$x^{(1)} = \{x^{(1)}(1),\ x^{(2)}(2), \dots,\ x^{(1)}(k)\}$ (2)

with $x^{(1)}(1) = x^{(0)}(1)$ and

$x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i), k \geq 2$

Step 3: The background value $z^{(1)}$ is computed employing the method of generations based on average:

$z^{(1)}(k) = 0.5 * [x^{(1)}(k) + x^{(1)}(k-1)], k \geq 2$ (3)

Step 4: Considering $x^{(1)}$ as an exponential variation, the Grey differential equation employed in the GM(1,1) estimation is $\frac{dx^{(1)}(k)}{dk} + ax^{(1)}(k) = b$ and its difference equation is expressed by

$$x^{(0)}(k)+a\, z^{(1)}(k)=b \quad (4)$$

where a and b are the parameters of the model known as development coefficient and Grey control variable respectively.

Step 5: The parameters are estimated using OLS approach as

$$\begin{bmatrix} a \\ b \end{bmatrix} = (B^T B)^{-1} B^T Y \quad (5)$$

where $Y = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \cdot \\ \cdot \\ x^{(0)}(n) \end{bmatrix}, B = \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ -z^{(1)}(n) & 1 \end{bmatrix}$

Step 6: The formula of the Grey differential equation under the initial condition $x^{(1)}(1)= x^{(0)}(1)$, is given by

$$\hat{x}^{(1)}(k) = \left(x^{(0)}(1) - \frac{\hat{b}}{\hat{a}}\right) e^{-\hat{a}(k-1)} + \frac{\hat{b}}{\hat{a}}, k \geq 2 \quad (6)$$

Since the Grey forecasting model is constructed using the 1st-order Accumulated Generating Operation (1-AGO) data rather than the original dataset, it is necessary to convert the predicted values back to their original scale. This conversion is performed using the **first-order Inverse AGO (1-IAGO)** transformation, which is represented by:

$$\hat{x}^{(1)}(k) = \left(x^{(0)}(1) - \frac{\hat{b}}{\hat{a}}\right) e^{-\hat{a}(k-1)}(1 - e^{\hat{a}}), k \geq 2 \quad (7)$$

2. Illustration

In a research study conducted by Ray et al. (2022), the grey model was applied to analyze the annual yield of Bt cotton in India from the years 2002-03 to 2016-17. The data used for this analysis were obtained from the Cotton Advisory Board of India. Out of the fifteen available observations, the first eleven data points were used to fit the grey model, while the remaining four data points were reserved for forecasting purposes. By utilizing the grey model, Ray et al. aimed to examine its effectiveness in predicting future yields of Bt cotton based on historical data patterns. The grey model, known for its ability to handle limited

data and fuzzy information, provided a suitable framework for forecasting the yield of Bt cotton in this study. Through the process of model fitting, the grey model was calibrated to capture the underlying trends and patterns present in the yearly Bt cotton yield data. Once the model was fitted using the initial eleven observations, the researchers tested its forecasting capabilities by predicting the yield for the remaining four years. This illustration from Ray et al.'s research serves as a practical example of applying the grey model in the agricultural domain, specifically in analyzing and forecasting Bt cotton yield in India. By partitioning the data into a fitting subset and a forecasting subset, the researchers were able to assess the grey model's performance in capturing and predicting the variations in cotton yield over time. The study showcases how the grey model can be a valuable tool for agricultural forecasting, enabling researchers, policymakers, and stakeholders to make informed decisions based on historical data patterns. It demonstrates the potential of the grey model to provide insights and aid in planning and managing agricultural systems, even with limited available data.

The estimated parameters $a = -0.018$ and $b = 450.03$ are obtained using the **Ordinary Least Squares (OLS)** method applied to the difference equation described in **Step 5 of Section 2**. Based on these estimates, the conventional forecast equation for the **GM(1,1)** model (refer to Equation 7) is given by:

$$\hat{x}_0^{(0)}(k) = [x^{(0)}(1) - \frac{450.033}{0.018}](1 - e^{-0.018})e^{0.018(k-1)}$$

3. Conclusion

In conclusion, the grey model proves to be a valuable forecasting technique, particularly in situations where conventional time series models face limitations due to data scarcity or limited observations. Its ability to handle sparse and fuzzy information sets it apart from traditional approaches. The grey model has demonstrated successful applications in various domains, ranging from high precision forecasting to vehicle fatality risk estimation, time series prediction, and industry output forecasting. It has shown promise in fields such as agriculture, linear planning, financial forecasting, and tourism demand analysis. One of the significant advantages of the grey model is its simplicity of computation. The use of

techniques like Ordinary Least Square (OLS) to estimate model parameters contributes to its ease of implementation and wide accessibility. Moreover, the grey model has proven effective even when working with a small number of observations. It offers a viable solution for forecasting purposes, allowing reliable predictions to be made even with limited data points. This characteristic makes it particularly useful in rapidly evolving environments or situations where a small number of observations are available due to societal or technological factors. Overall, the grey model serves as a powerful alternative for time series forecasting, enabling researchers, practitioners, and decision-makers to make informed decisions and plan effectively. Its versatility, simplicity, and ability to handle scarce data make it a valuable tool for analyzing and predicting trends in various domains, contributing to improved forecasting accuracy and decision-making processes.

R code:

```
v1  <-c(640,724,813,1145,1509,2122,1883,2413,2834,4235,7144,5269)
#testing
GM_test<-function(data){
 kk=length(data)-1
 a1<-vector("numeric")
 for (i in 1:kk) {
  a1[[i]]=(data[i]/data[i+1])}
 yf<-vector("numeric")
 for (i in 1:kk){
  if(a1[i] > 7.389){
   yf[[i]]=1
  } else {
   yf[[i]]=0
  }}
 yf1<-vector("numeric")
 for (i in 1:kk){
  if(a1[i] < 0.1345){
   yf1[[i]]=1
  } else {
   yf1[[i]]=0
  }}
 yf2=yf+yf1
 yf3=sum(yf2)
 if (yf3>0){
  print(" Data is not suitable for Grey modelling")
 } else {
```

```
   print("Data is suitable for Grey modelling")
  }}
GM_test(data=v1)


x <- cumsum(v1)

#average
t <-length(v1)-1

k <-vector("numeric")
y <-0

for ( i in 1:t) {

  y[i] <- (x[i] + x[i+1])/2

  print(y[i])
  k[i] <-y[i]

}
```

References:

Mao, M. and Chirwa, E. C. (2006).Application of grey model GM (1, 1) to vehicle fatality risk estimation. *Technol. Forecast. Soc. Change*, **73**, 588-605.

Ou, S. (2012). Forecasting agricultural output with an improved grey forecasting model based on the genetic algorithm. *Comp. Electro. Agri.*, **85**, 33-39.

Ray, M., Rai, A., Singh, K. N. and V., Ramasubramanian. (2017).Modeling and forecasting of hybrid rice yield using a grey model improved by the genetic algorithm. *Int. J. Agri. Stat. Sci.*, **13 (2),** 563-566.

Markov Chain Analysis

Prity Kumari
Department of Basic Science, College of Horticulture, Anand Agricultural University, Gujarat
Email: psingh2506@aau.in

Introduction:

Many real-life systems evolve over time due to a mix of randomness and dependence on their current state. To model and anticipate behaviour of such systems, stochastic processes are employed, mathematical frameworks that account for both unpredictability and dynamic progression. A Markov chain is a special type of stochastic process that describes a system where probability of moving to next state depends only on current state, not on the path taken to reach it. This characteristic is known as *Markov property*.

For example, consider a farmer who monitors daily moisture level of his field. The soil can be in one of two conditions: *well irrigated (W)* or *under irrigated (U)*. Each day, the state of soil can either remain same or change, influenced by weather or irrigation. To estimate tomorrow's condition, farmer doesn't need to recall moisture levels from past several days; just today's condition is sufficient. This makes it a typical example of a Markov chain, where system's *states* are W & U and *transition probabilities* might state that if today soil is well irrigated then there is 80% chance it remains well irrigated tomorrow.

Markov chains are valuable tools for modelling such time-dependent transitions in various agricultural scenarios, such as soil moisture status, crop growth stages, pest or disease progression, market shift, price prediction model etc. They are particularly useful when focus is on how process evolves over time and when future predictions rely solely on present state.

Terminologies used in Markov chain:

I. **State ($s_1, s_2, \ldots, s_n$):**

A state refers to a specific condition of agricultural system at a given point in time. *Example*: Soil moisture levels can be classified into three states: s_1= Dry, s_2 = Moist and s_3 = Saturated At any day t, soil will be in one of these states.

II. **Transition Probability (P_{ij}):**

The probability of moving from one state to another or probability of system moves from state i to state j in one time step. If soil is Moist today (s_2), the chance it becomes Dry tomorrow (s_1) is: $P_{21} = P(X_{t+1} = s_1 \mid X_t = s_2)$.

III. **Transition Probability Matrix (TPM), P:**

A square matrix P that shows all one-step transition probabilities between states.

Each row represents a current state and each column shows probability of going to a future state. *Example*, 3 soil states:

$$\mathrm{P} = \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix} \text{Where, } \mathrm{P}_{12} = \mathrm{P}(\text{Dry} \rightarrow \text{Moist})$$

$$\sum_{j=1}^{n} P_{ij} = 1$$

One-step transition probability:

$p_{ij} = P(X_{t+1} = j \mid X_t = i)$

$$\mathrm{P} = \begin{bmatrix} P_{11} & P_{12} & \ldots\ldots & P_{1n} \\ P_{21} & P_{22} & \ldots\ldots & P_{2n} \\ . & . & . & . \\ . & . & . & . \\ P_{n1} & P_{n2} & \ldots\ldots & P_{nn} \end{bmatrix}$$

N-Step Transition Probability:

$p_{ij}^{(n)} = P(X_{t+n} = j \mid X_t = i)$

$$P^n = \begin{bmatrix} {P_{11}}^{(n)} & {P_{12}}^{(n)} & \ldots\ldots & {P_{1n}}^{(n)} \\ {P_{21}}^{(n)} & {P_{22}}^{(n)} & \ldots\ldots & {P_{2n}}^{(n)} \\ . & . & . & . \\ . & . & . & . \\ {P_{n1}}^{(n)} & {P_{n2}}^{(n)} & \ldots\ldots & {P_{nn}}^{(n)} \end{bmatrix}$$

IV. Initial Probability Distribution (π):

This is a row vector representing the probability of being in each state at start (t = 0), where $\pi = [\pi_1, \pi_2, \ldots, \pi_n]$. If soil moisture starts with 20% Dry, 50% Moist and 30% Saturated, then $\pi = [0.2, 0.5, 0.3]$

V. Steady-State Distribution (π^*):

This is a long-term probability distribution where the probabilities no longer change after many steps.

$$\pi^* = \pi^* P \quad \text{and} \quad \sum_{i=1}^{n} \pi_i^* = 1$$

After simulating many days of irrigation and weather patterns, soil stabilizes to a state where $\pi^* = (0.1, 0.6, 0.3)$ $i.e.,$ in long run, soil will be Moist 60% of time.

VI. Markov Property:

The future state depends only on current state and not on past history.

$$P(X_{t+1} = j \mid X_t = i, X_{t-1}, \ldots, X_0) = P(X_{t+1} = j \mid X_t = i)$$

If today soil is Moist, chance of it being Wet tomorrow depends only on the fact that it's Moist today, not on whether it was Dry or Wet in the past.

Classification of States:

To better understand Markov chains, concept different states should be familiar. Following are the types of state:

I. **Accessibility (i → j):**
State j is **accessible** from state i if $P_{ij} > 0$. From "Moist" soil (state i), "Dry" soil (state j) can be reached in some number of steps.

II. **Communication (i ↔ j):**
States i and j communicate if both are accessible from each other like i → j and j → i ⇒ i ↔ j.

III. **Recurrent State (Persistent):**
A state is recurrent if the system is guaranteed to return to it eventually. Example: Soil will eventually become "Moist" again due to rainfall/irrigation cycle.

IV. **Transient state:**
A state is considered **transient** if, once the process enters this state, there is a **nonzero probability that it may never return** to it. More formally, **state *i*** is transient if there exists at least one **state *j*** that is **reachable from state *i***, but **state *i* is not reachable from state *j***. In other words, the process can move from state *i* to state *j*, but there is no path back from *j* to *i*.

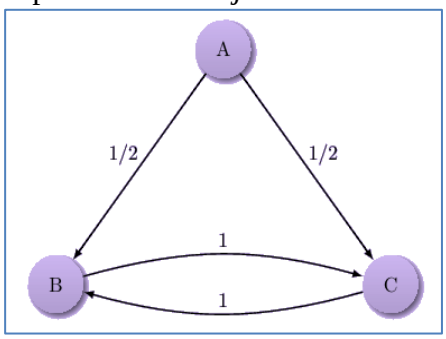


Fig 1: A Markov chain with one transient state and two recurrent states (Source: Brilliant.org)

V. **Absorbing State:**
A state i is absorbing if once entered, the process stays there forever. If a plant reaches "Dead" state due to drought, it cannot return to a "Recovering" stage.

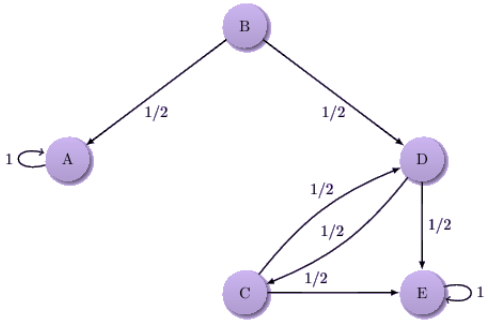


Fig 2: Absorbing states A & E (Source: Brilliant.org)

VI. **Periodic & Aperiodic:**

A state is periodic if it can only be returned to after fixed intervals. A system always goes from Seed → Flower → Fruit → Seed in exact order, like a fixed cycle. A state is aperiodic if it can be returned to at any time, not just fixed intervals.

VII. **Ergodic State:**

A state is ergodic if it is recurrent and aperiodic.

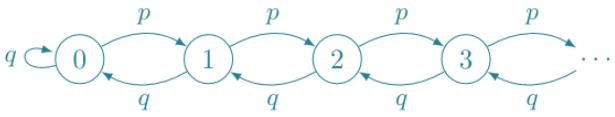


Fig 3: **Ergodic States** (Source: Brilliant.org)

Characteristics of Markov Chains:

A. Ergodic Chain:

A chain is ergodic if all its states are ergodic. Only ergodic chains have well defined steady-state distributions. Soil transitioning daily between all three moisture states with no fixed cycle.

B. Irreducible Chain:

A chain is irreducible if all states communicate with each other. The system can move from any state to any other (possibly in multiple steps). Soil moisture transitioning between Dry, Moist, Wet.

C. Reducible Chain:

A chain is reducible if at least one state cannot be reached from another. After moving from "Seedling" to "Harvest", the cycle never returns to "Seedling".

D. Reversibility:

A Markov chain is said to be reversible if the likelihood of moving from one state to another is the same as the likelihood of returning to the original state from the second state.

Types of Markov Chain

There are two main types of Markov chains: **Discrete-Time Markov Chain (DTMC)** and **Continuous-Time Markov Chain (CTMC)**. As their names suggest, **DTMC** involves transitions occurring at fixed, discrete time intervals, whereas **CTMC** allows transitions to happen at any continuous point in time (Ross, 2014; Lawler, 2018). These two types of Markov chains are summarized in **Table 1**.

Table 1: Types of Markov chain

Type	Full Form	State	Time
DTMC	Discrete Time Markov Chain	Discrete	Discrete
CTMC	Continuous Time Markov Chain	Discrete	Continuous

A. Discrete-time Markov chains (DTMC): In DTMC, transitions happen at fixed time intervals. This is where Hidden Markov Models (HMMs) fit in.

✓ **Hidden Markov Model (HMM):**

A Hidden Markov Model (HMM) is a statistical model in which the system being modelled is assumed to follow a Markov process with unobservable (hidden) states (Rabiner, 1989). The actual sequence of states is not directly visible, but each state generates an observation that is visible. Hence, model consists of:

a. *Hidden States*: $S = \{s_1, s_2, \ldots, s_N\}$

b. *Observations*: $O = \{o_1, o_2, \ldots, o_T\}$

c. *Initial State Probability*: Initial distribution over states,

$$\pi = \{\pi_i\} \quad \text{where} \quad \pi_i = P(q_1 = s_i)$$

d. *Transition Probability Matrix*:

Probability of transitioning between states:

$$A = [a_{ij}] \quad \text{where} \quad a_{ij} = P(q_{t+1} = s_j \mid q_t = s_i)$$

e. *Emission Probability Matrix*:

Describes the probability of observing a symbol given a state:

$$B = [b_j(k)] \quad \text{where} \quad b_j(k) = P(o_t = v_k \mid q_t = s_j)$$

Here, q_t denotes the hidden state at time t while o_t denotes observed output at time t and v_k represents a specific possible observation of output series.

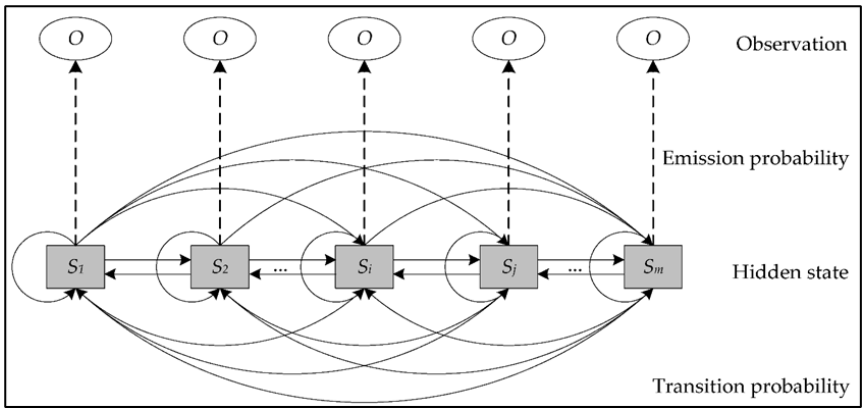


Fig 4: Hidden Markov Model (HMM) (Wu, 2020)

✓ **Algorithms for solving HMMs**

Hidden Markov Models (HMMs) utilize various algorithms, each designed for specific tasks, such as Forward Algorithm, Backward Algorithm and Viterbi Algorithm. However, this chapter focuses exclusively on the Viterbi Algorithm.

The Viterbi Algorithm is specifically used to address what is known as "*decoding problem"* in Hidden Markov Models (HMMs). Given a known HMM; comprising defined **a.** *Hidden states* **b.** *Transition probabilities* **c.** *Emission*

probabilities ***d.*** *Initial state distributions* **e.** *Sequence of observed events*, the objective is to determine the most likely sequence of hidden states responsible for generating those observations. In other words algorithm decodes most probable path through hidden layer of the model, offering a powerful tool for inference in sequential data (Viterbi, 1967).

While the decoding problem in HMMs may seem straightforward, "identifying most likely sequence of hidden states behind a given observation", it quickly becomes computationally intense as sequence length grows. For example, with just three hidden states and ten observations, there are over 59,000 (3^{10} = $(\text{State})^{\text{observations}}$ = 59,049) possible paths to consider. Exhaustive search is impractical for longer sequences.

The Viterbi Algorithm (Forney, 1973) solves this challenge efficiently using dynamic programming. Rather than evaluating all possible paths, it incrementally builds the most probable path using two key structures as mentioned in table 2. This step-by-step optimization makes decoding tractable even for long sequences.

Table 2: Two main structures of Viterbi Algorithm

Table	Purpose
Viterbi Table (V)	Highest probability of any path reaching each state at time t
Backpointer Table	Which previous state led to that highest probability

✓ **Steps to be followed in Viterbi Algorithm:**

1. **Initialization:** At time t = 0, calculate:

$$V_1(s) = \pi(s) \cdot P(O_0 \mid s)$$

Where, $\pi(s)$ is probability of starting in state S_1, $P(O_0 \mid s)$ is probability that state S_1 emits first observation O_0. Also at this time, backpointer will be none (this is the start).

2. **Recursion:** (for t = 1 to t-1) and for each state S_j

$$V_t(s) = \max_{s'}[V_{t-1}(s') \cdot P(s \mid s')] \cdot P(O_t \mid s)$$

$P(s \mid s')$ *is probability* of going from state s' to s while $P(O_t \mid s)$ *is emission* probability of observation O_t from state s. Save index of state s' that gave the max value in the backpointer.

3. **Termination:** At final time t, find:

$$P^* = \text{Bestpathprob} = \max_s V_T(s)$$

This gives the highest probability of any complete state path. Note the state q_t^* where this highest probability ends.

4. **Backtracking:** From t−1 to 0, follow:

$$q_t^* = backpointer(q_{t+1}^*)$$

Most likely full sequence of hidden states:

$$Q^* = [\, q_0^*,\ q_1^*,\ q_2^*, \ldots\ldots \,]$$

B. Continuous-time Markov chains (CTMC):

CTMC is a stochastic process where system transitions between discrete states continuously over time. Unlike DTMCs, CTMCs allow transitions at any real-valued time point and are governed by *transition rates* rather than probabilities. Components of CTMC are given table 3.

Table 3: Core component of CTMC

Components	Description
States (S)	A finite set of states
Holding Time (T)	The time spent in a state before transitioning, modelled as exponential distribution: $T \sim Exp(\lambda_i)$, where λ_i is the transition rate from state i.
Transition Rate (λ)	Inverse of average holding time: $\lambda = 1/\tau$
Generator (Q) Matrix	Diagonal entries are negative (rate of leaving a state); rows sum to 0.

In Continuous-Time Markov Chain (CTMC), these transitions are governed by *exponential random variables*, meaning the time a system spends in any given state, called the *holding time* and it follows an exponential distribution.

To illustrate, consider **life cycle of pest**, which progresses through four stages: **Egg → Larva → Pupa → Adult.**

At any moment, a pest can transition from its current stage to the next. However, the exact timing of this transition is random and varies for each stage. In such systems, once the timer for the current stage "expires", the pest transitions to the next stage. This unidirectional progression, typical in biological development, does not involve multiple competing paths, so the "competing exponentials" mechanism commonly seen in network systems or general CTMCs with multiple transition options per state is not applicable here. Instead, the pest always moves to a fixed next state, making the transition deterministic in direction, though stochastic in timing.

Despite the fixed path, the CTMC framework still applies. The system can be described using a Q-matrix (generator matrix), which encodes the transition rates. In this matrix:

- Off-diagonal entry q_{ij} represents the **rate of transition from state i to state j**.
- Diagonal entries q_{ii} are negative and equal to **negative sum of all outgoing rates** from state i, ensuring each row sums to zero.

This behavior forms what is known as a jump process. While timing of transitions is continuous and governed by exponential distributions, the sequence of transitions forms a discrete-time Markov chain, called the embedded or jump chain. A defining feature of CTMCs is the Markov property, often described as "memorylessness".

✓ **Steps to be followed in CTMC:**

1. Compute mean duration (τ) for leaving the stage
2. Calculate transition rate (λ) for each stage:

$$\lambda = {}^{1}/_{mean\ duration} = {}^{1}/_{\tau}$$

3. Q matrix

$$Q = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & 0 \\ 0 & -\lambda_2 & \lambda_2 & 0 \\ 0 & 0 & -\lambda_3 & \lambda_3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

4. Simulate Holding Times T_i

$$T_i \sim Exp(\lambda_i) = Waiting\ time = -\frac{ln(U)}{\lambda};\ \ U \sim Uniform(0,1)$$

Here U is random number between 0 to 1.

5. Add cumulative time at each stage transition

$$T_{total} = T_1 + T_2 + T_3 + \ldots..$$

Practical examples of Markov Chain:

Category: Discrete-Time Markov Chain (DTMC)

Q1. Given 15 days of daily rainfall data (in mm), forecast the probability of rainfall (Wet state) over the next 10 days using the DTMC approach.

Days	Rainfall (mm)	Days	Rainfall (mm)
1	0.0	9	12.1
2	5.2	10	15.3
3	0.0	11	0.0

4	0.0	12	0.0
5	7.3	13	2.0
6	0.0	14	0.0
7	0.0	15	0.0
8	0.0		

Solution:

Convert Rainfall into discreate states

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14	Day 15
Rainfall	**0.0**	**5.1**	**0.0**	**0.0**	**7.3**	**0.0**	**0.0**	**0.0**	**12.1**	**15.3**	**0.0**	**0.0**	**2.0**	**0.0**	**0.0**
State(0/1)	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**0**	**0**

Count Transitions

From \ To	Dry (0)	Wet (1)	Total
Dry (0)	5	4	**9**
Wet (1)	4	1	**5**

Transition Probability Matrix (TPM)

$$TPM=\begin{pmatrix} 5/9 & 4/9 \\ 4/5 & 1/5 \end{pmatrix} = \begin{pmatrix} 0.5556 & 0.4444 \\ 0.8 & 0.2 \end{pmatrix}$$

Initial Probability Distribution (π_0): (Dry = 10; Wet = 5; Total = 15)

$$\pi_0 = \begin{pmatrix} 10/15 & 5/15 \end{pmatrix} = (0.6667 \quad 0.3333)$$

$$\pi_1 = \pi_0 * TPM = (0.6667\ 0.3333) * \begin{pmatrix} 0.5556 & 0.4444 \\ 0.8 & 0.2 \end{pmatrix}$$

$$=[0.6667 \times 0.5556 + 0.3333 \times 0.8, \quad 0.6667 \times 0.4444 + 0.3333 \times 0.2]$$

$$=[0.3704 + 0.2667,\ 0.2963 + 0.0667]=[0.6371,\ 0.3630]$$

$$\pi_2 = \pi_1 * TPM = (0.6371\ 0.3630) * \begin{pmatrix} 0.5556 & 0.4444 \\ 0.8 & 0.2 \end{pmatrix}=[\ 0.5941\ \ 0.3431\ ]$$

$$\pi_3 = \pi_2 * TPM = (0.5941\ \ 0.3431) * \begin{pmatrix} 0.5556 & 0.4444 \\ 0.8 & 0.2 \end{pmatrix}=[\ 0.6045\ \ 0.3326\ ]$$

$$\pi_4 = \pi_3 * TPM = (0.6045\ \ 0.3326) * \begin{pmatrix} 0.5556 & 0.4444 \\ 0.8 & 0.2 \end{pmatrix}= [0.6019 \quad 0.3351]$$

$$\pi_5 = \pi_4 * TPM = (0.6019 \quad 0.3351) * \begin{pmatrix} 0.5556 & 0.4444 \\ 0.8 & 0.2 \end{pmatrix}=[0.6024\ \ 0.3345\ ]$$

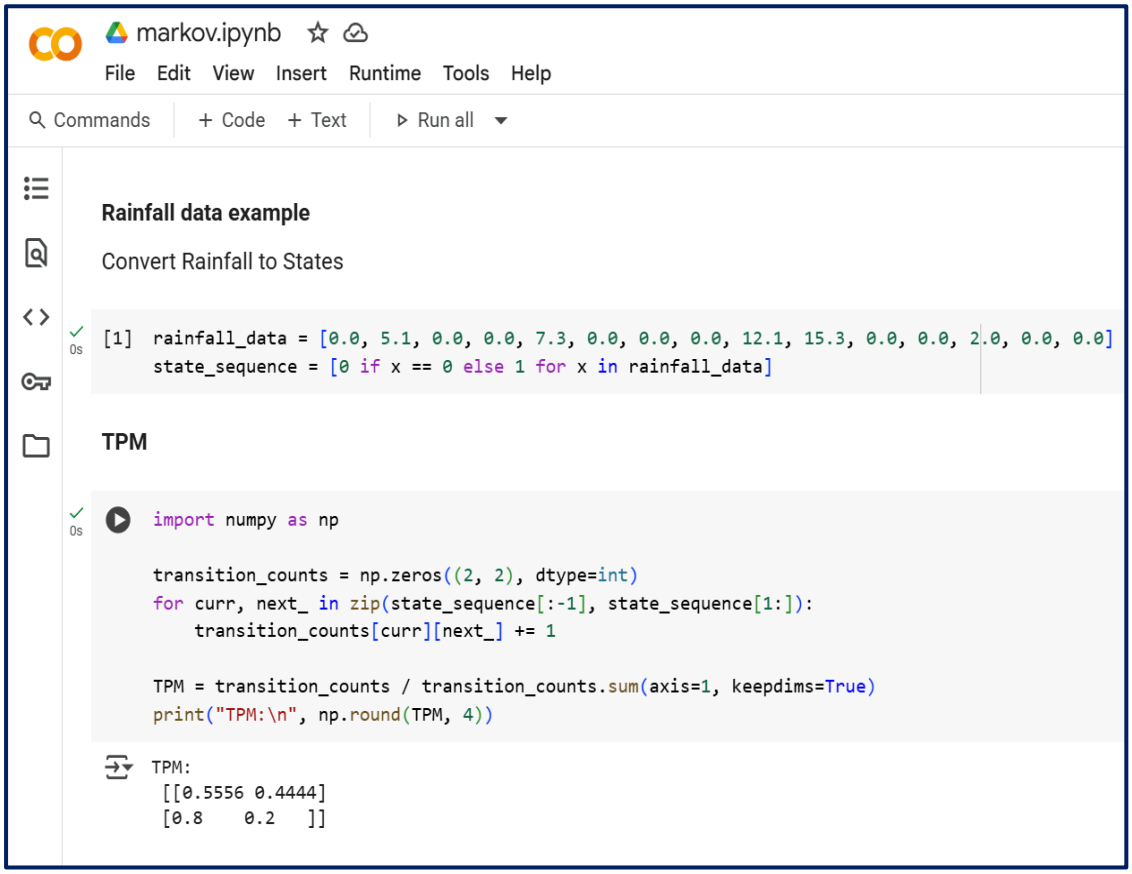
$$\pi_6 = \pi_5 * \text{TPM} = (0.6024 \quad 0.3345) * \begin{pmatrix} 0.5556 & 0.4444 \\ 0.8 & 0.2 \end{pmatrix} = [\ 0.6022 \quad 0.3345\ ]$$

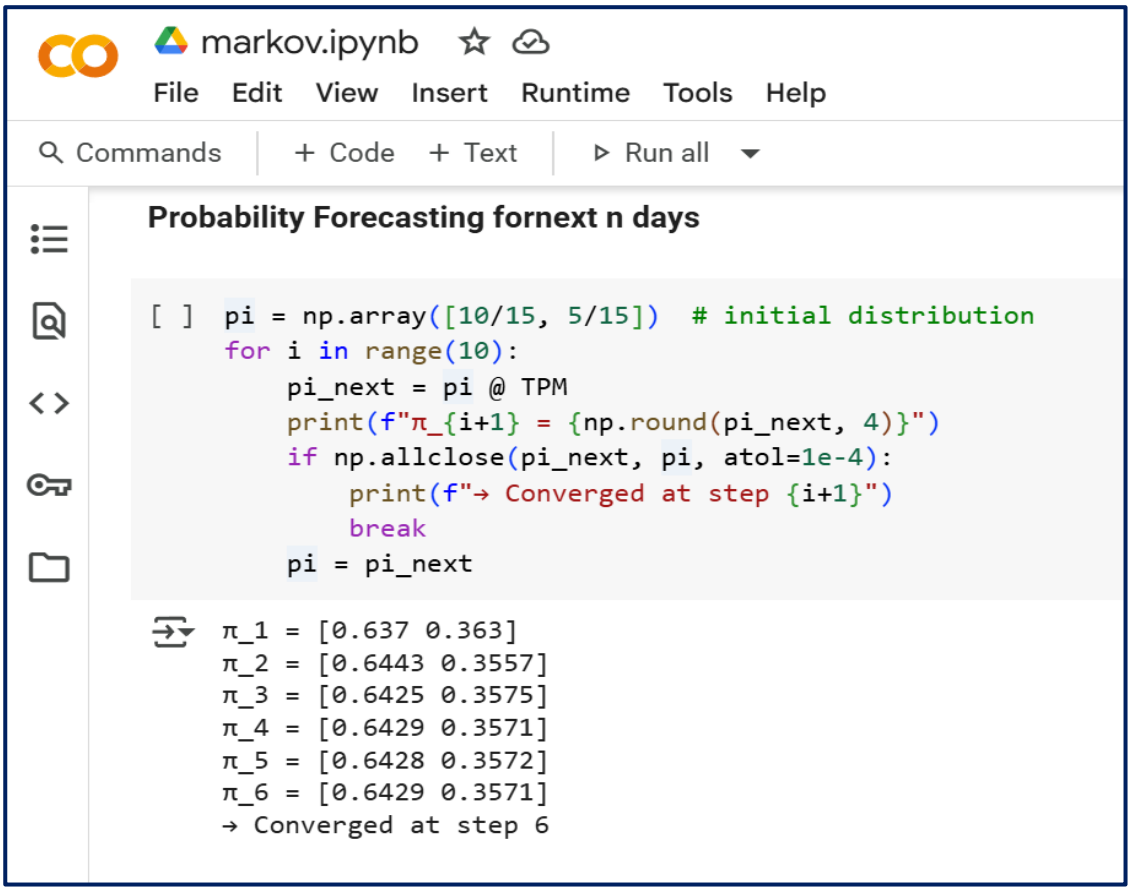
$$\pi_7 = \pi_6 * \text{TPM} = (0.6022 \quad 0.3345) * \begin{pmatrix} 0.5556 & 0.4444 \\ 0.8 & 0.2 \end{pmatrix} = [\ 0.6021 \quad 0.3346\ ]$$

……. keep multiplying until π_{10}

Probabilty of next 10 days = π_0 * $(\text{TPM})^{10}$

Python Code Implementation Rainfall Forecast

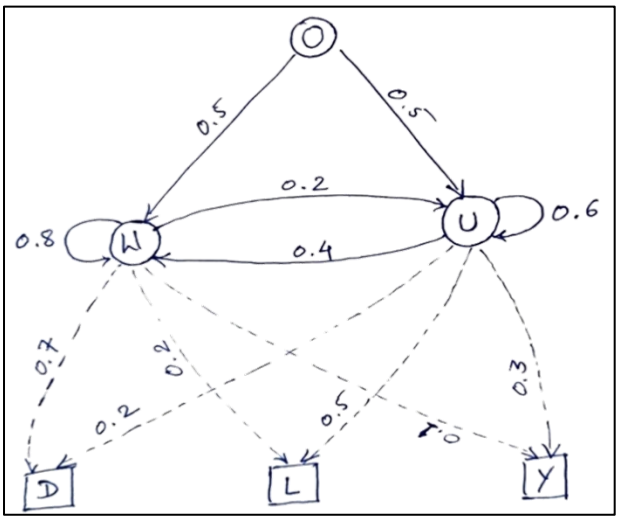




Q2. A farmer, in absence of soil sensors, relies on leaf color observations from drone images recorded over 3 consecutive days to estimate soil moisture status. The probabilities are given in diagram below.

I. Determine most probable sequence of soil moisture states over 3 days, given observed leaf color sequence as L → D → Y.

Observed Leaf Colours:	Hidden States (Soil Moisture Condition):
Day 1 → Light Green (L) **Day 2** → Dark Green (D) **Day 3** → Yellowish (Y)	**W**: Well Irrigated **U**: Under Irrigated



Solution:

Enlist all matrices first which is required to solve the problem using Viterbi algorithm:

π (Initial state) = $\{0.5 \quad 0.5\}$

A (State Transition Matrix) = $\begin{Bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{Bmatrix}$

B (Emission Matrix) = $\begin{Bmatrix} 0.7 & 0.2 & 0.1 \\ 0.2 & 0.5 & 0.3 \end{Bmatrix}$

Viterbi algorithm steps:

i. **Initialization:**

$$V_1(s) = \pi(s) \cdot P(O_1 \mid s)$$

ii. **Recursion (for t ≥ 2):**

$$V_t(s) = \max_{s'}\left[V_{t-1}(s') \cdot P(s \mid s')\right] \cdot P(O_t \mid s)$$

iii. **Last step:**

$$\textbf{Best path prob} = \max_s V_T(s)$$

i. Initialization (Day 1 → L)

$$V_1(W) = \pi(W) \cdot P(L \mid W) = 0.5 * 0.2 = 0.1$$

$$V_1(U) = \pi(U) \cdot P(L \mid U) = 0.5 * 0.5 = 0.25$$

ii. Recursion (Day 2 → D)

$$V_2(W) = max[V_1(W) * P(W \mid W), V_1(U) * P(W \mid U)] * P(D \mid W)$$

$$= \max[0.1 * 0.8,\ 0.25 * 0.4] *\ 0.7 = \max[0.08,\ 0.10] * 0.7$$

$$= 0.10 * 0.7 = 0.07$$

$$V_2(U) = \max[V_1(W) * P(U \mid W), V_1(U) * P(U \mid U)] * P(D \mid U)$$

$$= \max[0.1 * 0.2,\ 0.25 * 0.6] * P(D \mid U) = \max[0.02,\ 0.15] * 0.2$$

$$= 0.15 * 0.2 = 0.03$$

iii. Recursion (Day 3 → Y)

$$V_3(W) = max[V_2(W) * P(W \mid W), V_2(U) * P(W \mid U)] * P(Y \mid W)$$

$$= max[0.07 \cdot 0.8,\ 0.03 \cdot 0.4] \cdot 0.1 = max[0.056,\ 0.012] \cdot 0.1$$

$$= 0.056 \cdot 0.1 = 0.0056$$

$$V_3(U) = \max[V_2(W) * P(U \mid W), V_2(U) * P(U \mid U)] * P(Y \mid U)$$

$$= \max[0.07 \cdot 0.2,\ 0.03 \cdot 0.6] \cdot P(Y \mid U) = \max[0.014,\ 0.018] \cdot 0.3$$

$$= 0.018 \cdot 0.3 = 0.0054$$

So the most probable path ends in W and the maximum path probability is 0.0056

Then trace back:

Day 3: W (← came from W)

Day 2: W (← came from U)

Day 1: U

So most probable hidden state sequence for 3 consecutive days is U → W → W

Python Code Implementation for Soil Moisture Estimation Using Leaf Color

markov.ipynb

File Edit View Insert Runtime Tools Help

Commands + Code + Text Run all

Install package for hidden markov model

hmmlearn

```
!pip install hmmlearn
```

```
Collecting hmmlearn
  Downloading hmmlearn-0.3.3-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadat
Requirement already satisfied: numpy>=1.10 in /usr/local/lib/python3.11/dist-packages (from hmm
Requirement already satisfied: scikit-learn!=0.22.0,>=0.16 in /usr/local/lib/python3.11/dist-pa
Requirement already satisfied: scipy>=0.19 in /usr/local/lib/python3.11/dist-packages (from hmm
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from s
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages
Downloading hmmlearn-0.3.3-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (165 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 165.9/165.9 kB 3.1 MB/s eta 0:00:00
Installing collected packages: hmmlearn
Successfully installed hmmlearn-0.3.3
```

Define HMM model and its componnents

```
[10] model = CategoricalHMM(n_components=2, init_params="")

     # initial probabilities (π), transition matrix (A), emission matrix (B)
     model.startprob_ = np.array([0.5, 0.5])  # π: P(W), P(U)
     model.transmat_ = np.array([[0.8, 0.2],  # From W
                                 [0.4, 0.6]])  # From U
     model.emissionprob_ = np.array([[0.2, 0.7, 0.1],  # W: P(L), P(D), P(Y)
                                     [0.5, 0.2, 0.3]])  # U
```

```
#Decode by Viterbi algorithm
logprob, hidden_states = model.decode(obs_seq, algorithm="viterbi")

#Convert statenumbers to names
state_map = {0: 'W', 1: 'U'}
decoded_path = [state_map[state] for state in hidden_states]
max_prob = np.exp(logprob)

#results
print("Most probable state sequence:", decoded_path)
print("Maximum path probability:", max_prob)
```

```
Most probable state sequence: ['U', 'W', 'W']
Maximum path probability: 0.005600000000000025
```

Category: Continuous-Time Markov Chain (CTMC)

Q3. Lifecycle data of 10 eggs of a single pest species recorded under uniform conditions. Each pest transitions through four stages: Egg → Larva → Pupa → Adult..

I. Calculate average time in each state and corresponding transition rates

II. Construct CTMC Q matrix for 4-stage lifecycle

III. Simulate adult emergence of 100 pests and prepare a day-wise emergence table

IV. How many reaching Adult stage by Day 5, 6, 7 and determine peak infestation day

Pest ID	Egg Laid (Day 0)	Larva Emerged (Day)	Pupa Formed (Day)	Adult Emerged (Day)
P1	0.0	2.1	5.4	9.3
P2	0.0	1.8	4.7	8.8
P3	0.0	2.2	5.2	9.5
P4	0.0	2.0	5.0	9.1
P5	0.0	1.9	4.8	8.9
P6	0.0	2.3	5.3	9.6
P7	0.0	2.1	5.1	9.4
P8	0.0	2.0	5.2	9.2
P9	0.0	2.2	5.4	9.5
P10	0.0	2.1	5.3	9.3

Solution:

A. *Estimation of Stage Durations*

- For each pest, compute the time spent in each stage:
  - Egg duration = Larva Emerged − Day 0
  - Larva duration = Pupa Formed − Larva Emerged
  - Pupa duration = Adult Emerged − Pupa Formed
- Compute the mean duration for Egg, Larva and Pupa stages across all 10 pest ID manually.

B. *Estimation of Transition rate & Q matrix*

- Calculate the transition rate (λ) for each stage :

$$\lambda = {}^{1}/_{mean\ duration} = {}^{1}/_{\tau}$$

- Q matrix

$$Q = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & 0 \\ 0 & -\lambda_2 & \lambda_2 & 0 \\ 0 & 0 & -\lambda_3 & \lambda_3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

C. Simulation for 100 pests via using python

Pest ID	Egg Laid (Day 0)	Larva Emerged (Day)	Pupa Formed (Day)	Adult Emerged (Day)		Duration		
						Egg (τ_1)	Larva(τ_2)	Pupa(τ_3)
P1	0	2.1	5.4	9.3		2.1	3.3	3.9
P2	0	1.8	4.7	8.8		1.8	2.9	4.1
P3	0	2.2	5.2	9.5		2.2	3	4.3
P4	0	2	5	9.1		2	3	4.1
P5	0	1.9	4.8	8.9		1.9	2.9	4.1
P6	0	2.3	5.3	9.6		2.3	3	4.3
P7	0	2.1	5.1	9.4		2.1	3	4.3
P8	0	2	5.2	9.2		2	3.2	4
P9	0	2.2	5.4	9.5		2.2	3.2	4.1
P10	0	2.1	5.3	9.3		2.1	3.2	4
					Mean	**2.07**	**3.07**	**4.12**
					Lambda = 1/mean	0.4831	0.3257	0.2427

- $\boldsymbol{\tau}$ $(mean)$ and $\boldsymbol{\lambda}$ (lamda=transition rate) are calculated in the above table. Additionally, **Q** matrix is formulated with help of λ.

$$\boldsymbol{Q} = \begin{pmatrix} \mathbf{-0.4831} & \mathbf{0.4831} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{-0.3257} & \mathbf{0.3257} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{-0.2427} & \mathbf{0.2427} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}$$

- Compute holding time **T**

$$\boldsymbol{T_i \sim Exp(\lambda_i) = Waiting\ time} = -\frac{\ln(U)}{\lambda};\ \boldsymbol{U \sim Uniform(0,1)}$$

  - Randomly choose U values for each pest and each stage (from [0.1–0.9]):

- Pest 1
  - U_1 (Egg) = 0.7
  - U_2 (Larva) = 0.4
  - U_3 (Pupa) = 0.3

$$T_{U_1} = -\frac{ln(0.7)}{0.4831} \approx 0.7384 \text{ days}$$

$$T_{U_2} = -\frac{ln(0.4)}{0.3257} \approx 2.8134 \text{ days}$$

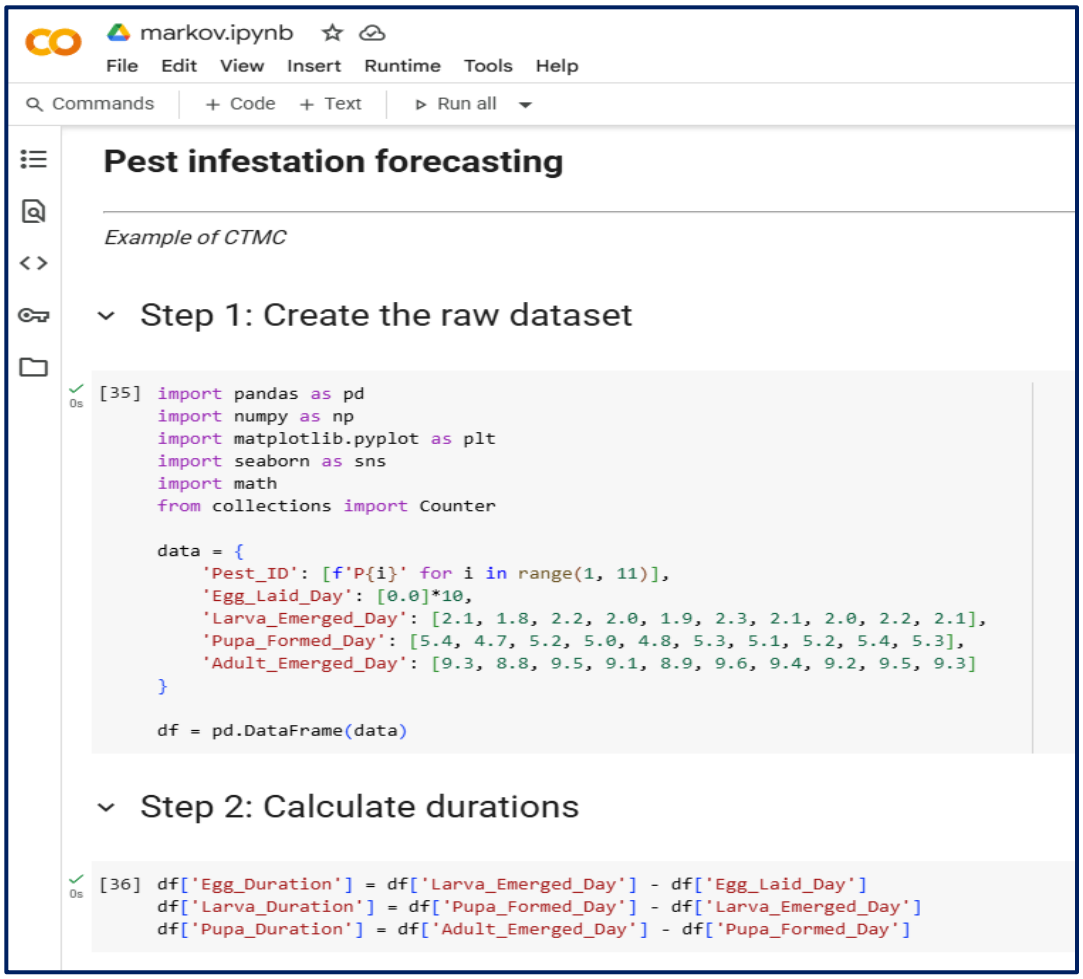
$$T_{U3} = -\frac{ln(0.3)}{0.2427} \approx 4.9636 \text{ days}$$

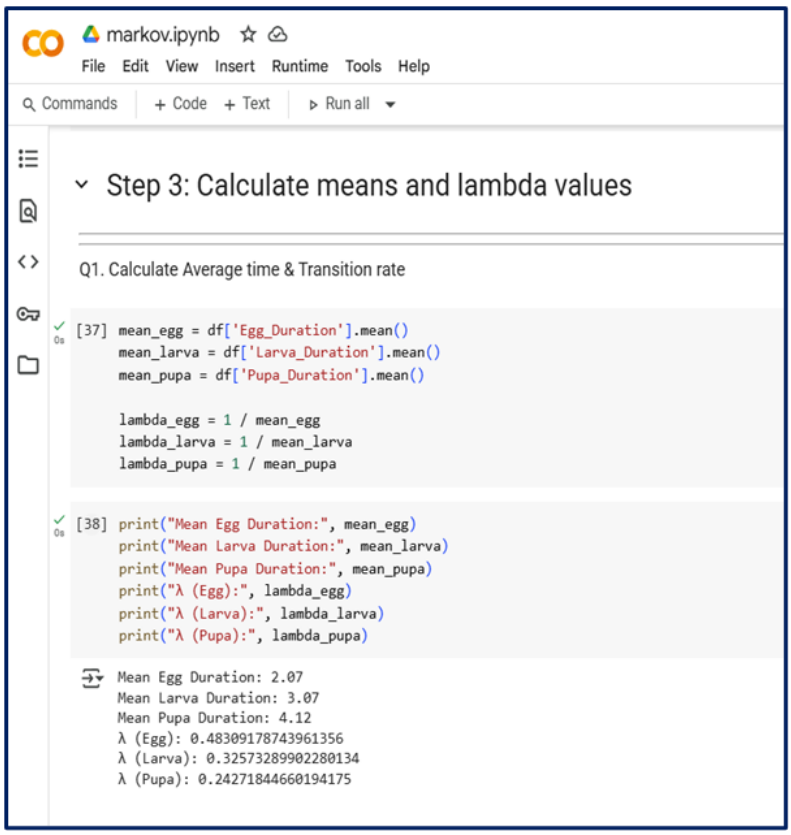
- **Total time** to adult = 0.7384 + 2.8134 + 4.9636 = 8.5154 days
- Similarly it is **simulated** for all 10 eggs

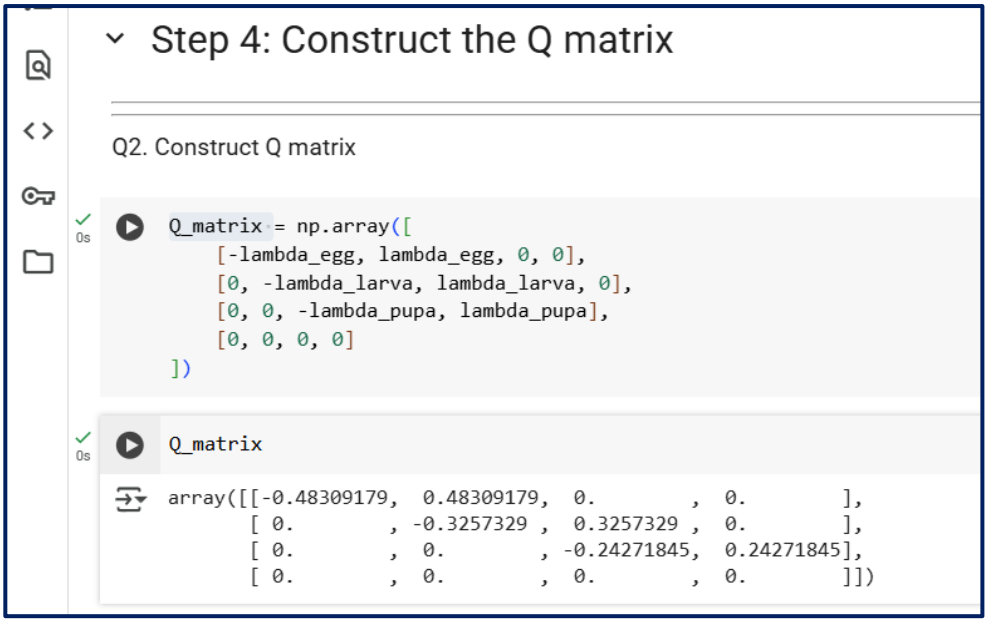
Cohort Forecasting (Simulation for 10 eggs)

	U_1 (Egg)	U_2 (Larva)	U_3 (Pupa)	T_Egg	T_Larva	T_Pupa	Total Time
P1	0.29	0.90	0.65	2.56	0.33	1.66	4.6
P2	0.34	0.71	0.66	2.23	1.07	1.60	4.9
P3	0.89	0.33	0.50	0.24	3.47	2.68	6.4
P4	0.71	0.30	0.51	0.71	3.77	2.60	7.1
P5	0.43	0.22	0.76	1.75	4.74	1.06	7.5
P6	0.40	0.37	0.33	1.90	3.11	4.28	9.3
P7	0.60	0.55	0.55	1.06	1.87	2.31	5.2
P8	0.20	0.21	0.39	3.33	4.88	3.63	11.9
P9	0.50	0.39	0.25	1.43	2.95	5.35	9.7
P10	0.54	0.66	0.65	1.28	1.30	1.66	4.2

Python Code Implementation for Pest Infestation Forecasting (CTMC)




markov.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text Run all
Step 3: Calculate means and lambda values
Q1. Calculate Average time & Transition rate
[37] mean_egg = df['Egg_Duration'].mean()
mean_larva = df['Larva_Duration'].mean()
mean_pupa = df['Pupa_Duration'].mean()
lambda_egg = 1 / mean_egg
lambda_larva = 1 / mean_larva
lambda_pupa = 1 / mean_pupa
[38] print("Mean Egg Duration:", mean_egg)
print("Mean Larva Duration:", mean_larva)
print("Mean Pupa Duration:", mean_pupa)
print("λ (Egg):", lambda_egg)
print("λ (Larva):", lambda_larva)
print("λ (Pupa):", lambda_pupa)
Mean Egg Duration: 2.07
Mean Larva Duration: 3.07
Mean Pupa Duration: 4.12
λ (Egg): 0.48309178743961356
λ (Larva): 0.32573289902280134
λ (Pupa): 0.24271844660194175


Step 4: Construct the Q matrix
Q2. Construct Q matrix
Q_matrix = np.array([
[-lambda_egg, lambda_egg, 0, 0],
[0, -lambda_larva, lambda_larva, 0],
[0, 0, -lambda_pupa, lambda_pupa],
[0, 0, 0, 0]
])
Q_matrix
array([[-0.48309179, 0.48309179, 0. , 0.],
[0. , -0.3257329 , 0.3257329 , 0.],
[0. , 0. , -0.24271845, 0.24271845],
[0. , 0. , 0. , 0.]])

Step 5: Simulate 100 pests

Q3(a). Simulate adult emergence of 100 pests

```
[50] simulated_pests = []
     for _ in range(100):
         t_egg = -np.log(np.random.uniform()) / lambda_egg
         t_larva = -np.log(np.random.uniform()) / lambda_larva
         t_pupa = -np.log(np.random.uniform()) / lambda_pupa
         total_time = t_egg + t_larva + t_pupa
         simulated_pests.append(round(total_time, 1))
```

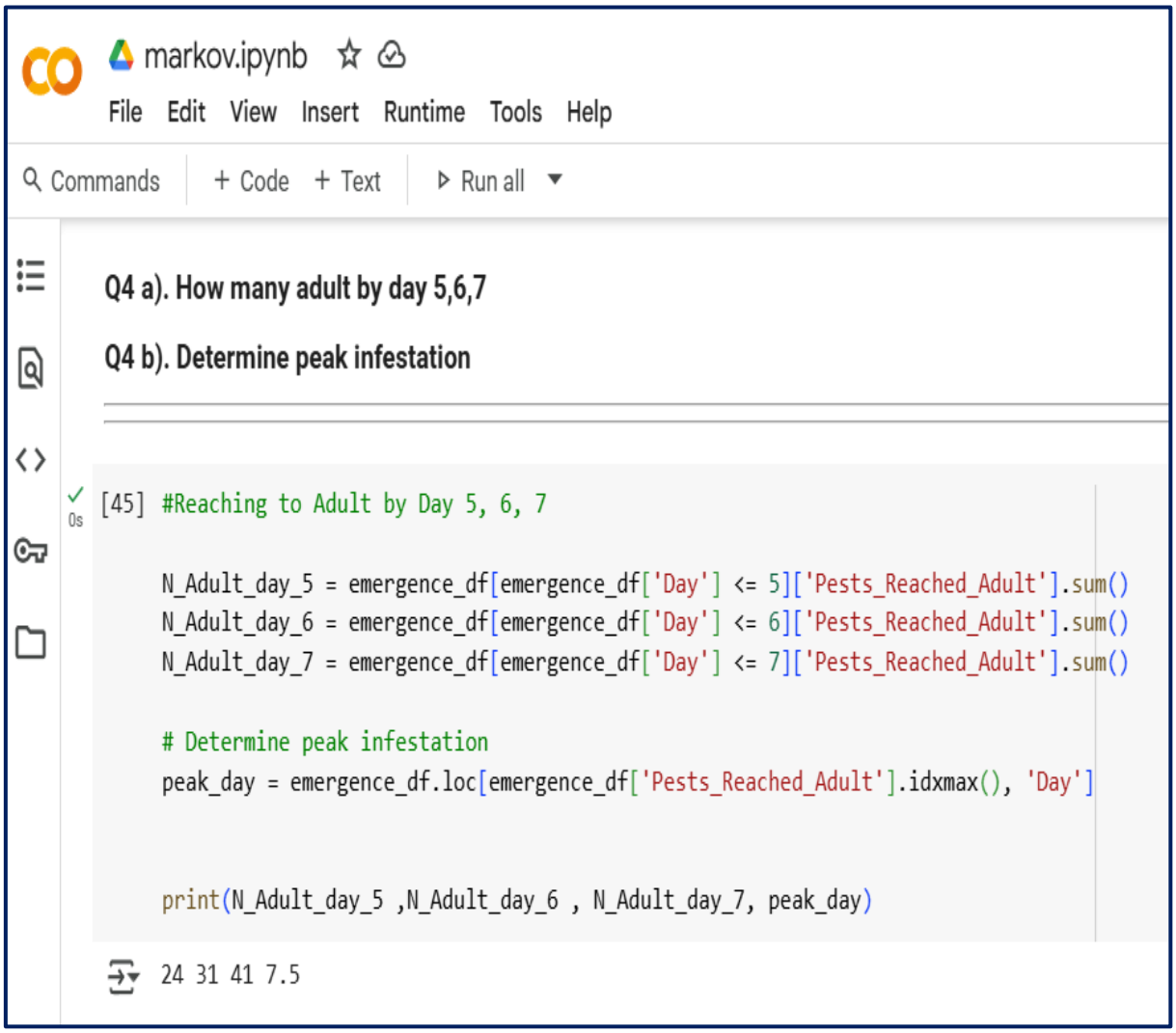
```
import pandas as pd
df_simulated = pd.DataFrame(simulated_pests, columns=["Total Time (Days)"])
df_simulated.index = [f"P{i+1}" for i in range(len(df_simulated))]
df_simulated.index.name = "Pest ID"
df_simulated.head(10)
```

Pest ID	Total Time (Days)
P1	6.3
P2	4.7
P3	8.9
P4	6.2
P5	8.1

Q3(b). Prepare day wise emergence table

```
import pandas as pd
from collections import Counter

emergence_counter = Counter(simulated_pests)        #Count number of pests emerging at each day
emergence_df = pd.DataFrame(sorted(emergence_counter.items()), columns=['Day', 'Pests_Reached_Adult'])
emergence_df['Cumulative_Adults'] = emergence_df['Pests_Reached_Adult'].cumsum()       #cumulative number of adult

emergence_df
```

	Day	Pests_Reached_Adult	Cumulative_Adults
0	1.5	1	1
1	1.7	1	2
2	1.8	1	3
3	2.2	1	4
4	2.3	1	5
...	...	...	...
71	20.3	1	96
72	20.8	1	97



References:

Brilliant.org. "*Absorbing Markov Chains*." Brilliant Math & Science Wiki. Accessed June 28, 2025. https://brilliant.org/wiki/absorbing-markov-chains/.

Ross, Sheldon M. Introduction to probability models. Academic press, 2014.

Lawler, Gregory F. Introduction to stochastic processes. Chapman and Hall/CRC, 2018.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286. https://ieeexplore.ieee.org/document/18626

Wu, J.; Jin, Y.; Hu, S.; Fei, J.; Zhang, Y. Approach to Risk Performance Reasoning with Hidden Markov Model for Bauxite Shipping Process Safety by Handy Carriers. *Appl. Sci.* **2020**, *10*, 1269. https://doi.org/10.3390/app10041269

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory, 13(2), 260-269. https://ieeexplore.ieee.org/document/1054010

Forney, G. D. (1973). The Viterbi algorithm. Proceedings of the IEEE, 61(3), 268-278. https://ieeexplore.ieee.org/document/1455960

Analytical Hierarchy Process (AHP) – case study in India

Dr. Amrender Kumar
AKMU, ICAR-Indian Agricultural Research Institute, Pusa, New Delhi -110012
Email: akjha@iari.res.in

Abstract

The Analytic Hierarchy Process (AHP), developed by Thomas Saaty in the 1970s, is a structured and widely used methodology for decision-making, particularly in complex and multi-criteria environments. Its first real-world application was in 1973, and it was formally published in 1980. Since then, AHP has become one of the most applied techniques in decision analysis. The method is based on three core principles: decomposition, comparative judgment, and synthesis of priorities. The decomposition principle involves breaking down a complex decision problem into a hierarchical structure composed of different levels. The top level represents the overall objective, the intermediate levels represent criteria and sub-criteria, and the bottom level includes the decision alternatives. Each level is independent of the others, allowing a clear and organized approach to problem-solving from general goals to specific elements. The second principle, comparative judgment, entails constructing pairwise comparison matrices in which elements within a level are compared with respect to their importance relative to a criterion in the level above. These comparisons are used to populate a reciprocal matrix, from which the principal eigenvector is derived to determine the relative priority of each element.

Finally, the synthesis principle involves aggregating the local priorities downward through the hierarchy to produce global priorities for the decision alternatives. This is done by multiplying the local priorities of each element by the priority of its parent criterion in the upper level and summing the results. AHP is particularly valuable for incorporating subjective expert judgments into a structured framework and enables consistent evaluation of alternatives. The method's strengths lie in its ability to handle both qualitative and quantitative factors, its hierarchical organization of complex problems, and its systematic approach to deriving and aggregating priorities for informed decision-making.

According to Saaty (1980), priorities in the Analytic Hierarchy Process (AHP) are synthesized from the second level downward by multiplying each element's local priority by the priority of its corresponding criterion in the level above. These products are then summed for each element in a level, based on the criteria they relate to. This process yields the composite or global priority for that element, which is subsequently used to weight the local priorities of the elements in the level below when making comparisons based on that criterion, and the procedure continues down to the bottom level of the hierarchy. As with any methodological framework, AHP is grounded in a set of foundational axioms. These axioms serve as the base assumptions that support the structure and logic of the process. The **reciprocal axiom** requires that if a paired comparison between elements A and B with respect to a parent element C (denoted PC(A, B)) indicates how many times more A possesses a certain property compared to B, then the inverse comparison PC(B, A) must equal the reciprocal value, i.e., 1 / PC(A, B). The **homogeneity axiom** asserts that elements being compared should not vary too widely in the property being assessed. To avoid large judgmental errors, the linguistic scales used for comparisons should be limited to within one order of magnitude. The **synthesis axiom** implies that judgments regarding the priorities of elements in a hierarchy should remain unaffected by elements in lower levels. This ensures that the relative importance of higher-level objectives is not influenced by the priorities of lower-level components. While it doesn't require complete independence of all factors, it ensures that the hierarchical composition remains valid for additive aggregation of priorities. Lastly, the **expectation axiom** maintains that individuals should ensure their beliefs and reasoning are adequately captured within the decision structure. This means the final output priorities should not drastically deviate from any well-informed expectations or prior knowledge the decision-maker holds. Together, these axioms uphold the consistency, rationality, and reliability of the AHP framework.

The Analytical Process

The AHP is essentially an **additive weighted aggregation** method, where **priority scores** are computed by combining weights assigned to various criteria. These weights are derived from **subjective pairwise comparisons** of the lowest-level factors or criteria within the decision hierarchy. Through this structured comparison, AHP quantifies the relative importance of each factor and aggregates them to produce an overall priority score for decision alternatives.

Hierarchical Decomposition of the Decision

The decision-making process using the AHP begins by **decomposing the problem into a hierarchical structure** that includes all the essential and sufficient elements required for the decision. This hierarchy typically starts with the overall objective at the top, followed by higher-level factors or categories, and then further broken down into more specific sub-factors or criteria at the lower levels. In the illustrative structure shown in Figure 1, element **A** represents the overall priority of an alternative, determined by considering all relevant factors. The higher-level elements **B, C, and D** might represent broad decision categories such as **Lifecycle Cost, Benefits,** and **Risks**, while the lower-level elements (depicted as ellipses in Figure 1) represent **sub-criteria** within each of these main categories. To clarify the methodology and computational steps of AHP, we reference a military application related to enhancing **airborne surveillance capabilities**. In this case, five potential alternatives have been shortlisted for evaluation, forming the basis for comparison and prioritization using the AHP framework. This approach allows decision-makers to systematically assess each alternative by considering its performance against multiple structured criteria.

Alt1: Airborne Warning & Control System, Alt2: Strategic Surveillance System

Alt3: Tactical Surveillance System Alt4: Electromagnetic Intelligence

Alt5: Battlefield Surveillance System

The top level categories are:

B = Acquisition Feasibility, C = Sustainability, D = Military Gain

The next level factors are:

B1 = Availability, B2 = Cost

C1 = Import Content, C2 = Technology Absorption, C3 = Indigenous Production

D1 = Non-existing Capability, D2 = Enhancement of Exis ting Capability,

D3 = Reduction in Enemy Capability, D4 = Morale Booster for Own Force.

And the lowest level factors are:

B21 = Set-up Cost, B22 = Running Cost, B23 = Annual Equivalent Cost

Pairwise Comparisons

In the AHP, two types of pairwise comparisons are conducted. The first involves comparing pairs of factors within the same level of the hierarchy, where analysts provide input on their relative importance. These comparisons yield computed values known as factor weights, which are crucial for the final merit aggregation across the hierarchy. The weights are derived from a top-down comparison process and are normalized so that their total under each parent node equals one. The second type of comparison is between pairs of alternatives, evaluating their relative performance with respect to each leaf or terminal node in the hierarchy. To carry out all pairwise comparisons, a graded rating scale is used to express the strength of preference between the compared elements.

Selection of Comparison Scale Type

Since the interpretation of a numerical rating depends on the nature of the scale it is derived from-such as ordinal, interval, or ratio-it is essential to define the scale type being used. According to Saaty [27], the AHP operates using a **ratio scale**, allowing for meaningful comparisons of relative magnitudes between elements.

Selection of Comparison Scale Units

When applying a ratio scale for mutual comparisons, the assigned numbers indicate the **relative magnitude** of the attribute or property held by the two factors under comparison. However, in the Analytic Hierarchy Process (AHP), decision-makers typically express their judgments using **verbal or linguistic terms** to convey relative importance. These qualitative assessments are then translated into **numerical values** to populate the pairwise comparison matrix. A commonly used scale, as proposed by Saaty, assigns values such as

1 for equal importance, 3 for moderate importance, 5 for strong importance, 7 for very strong importance, and 9 for extreme importance, with even numbers (2, 4, 6, 8) serving as intermediate values for finer discrimination.

	Equal Importance	Mildly Stronger	Stronger	Much	Stronger	Extremely Stronger
	1	3	5	7		9

	ratings of relative importance
Equally preferred	1
Equally to moderately preferred	2
Moderately preferred	3
Moderately to strongly preferred	4
Strongly preferred	5
Strongly to very strongly preferred	6
Very strongly preferred	7
Very to extremely strongly preferred	8
Extremely preferred	9

The reciprocals of the numbers represent the inverse of the original comparisons, aligning with the reciprocity axiom. The corresponding set of values for the 5 main grades mentioned above thus includes their respective reciprocals:

{ 0.11, 0.14, 0.20, 0.33, 1, 3, 5, 7, 9 }

The AHP calculates factor weights and alternative priorities from square matrices of pairwise comparison ratings using matrix algebra, specifically through eigenvalue-based methods.

For any square matrix A, λ is an eigenvalue associated with a vector Ψ such that $A\Psi = \lambda\Psi$, where Ψ is called the corresponding eigenvector.

Step 1: Determine the largest eigenvalue (λ) that satisfies the characteristic polynomial of the comparison matrix.

Step 2: Compute the principal eigenvector corresponding to this maximum eigenvalue.
Step 3: Normalize the principal eigenvector so that the sum of its elements equals one.

The normalized elements indicate the relative weights of factors or the relative priorities of alternatives with respect to a given criterion. To assess the consistency of rankings within any options matrix (or criterion priority matrix), the **Consistency Index (CI)** is first computed using the formula provided below. For a case with 4 options, $\boldsymbol{E_{max}}$ represents the estimated maximum eigenvalue:

$$CI = (\lambda_{max} - n) / (n-1)$$

where λ_{max} is the largest eigenvalue and n is the number of criteria. The **Consistency Ratio (CR)** is then obtained by dividing the **CI** by the **Random Consistency Index (RCI)**, which varies based on the number of options. The RCI values, as provided by Saaty (1980), are listed in the table below for different values of n.

RANDOM CONSISTENCY INDEX TABLE	
Number of Options	**RCI**
1	0
2	0
3	0.58
4	0.90
5	1.12
6	1.24

7	1.32
8	1.41
9	1.45

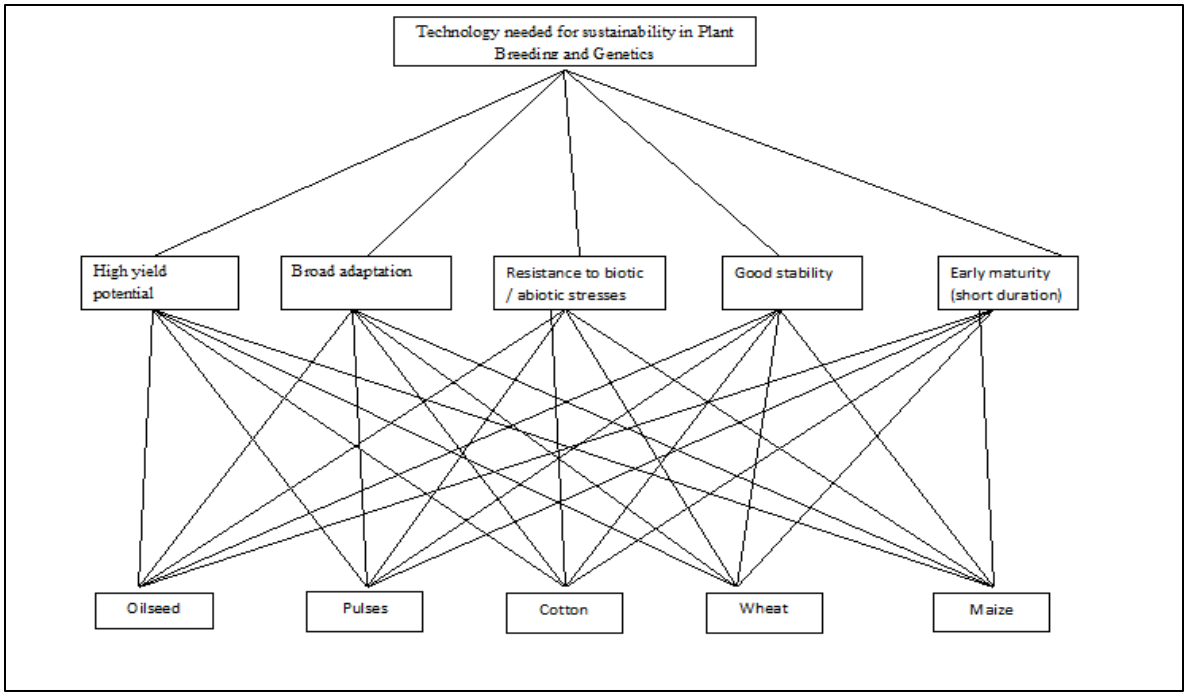
The consistency ratio (CR) is given by:

$$CR = CI / RCI$$

A CR value greater than 0.10 (or 10%) indicates the need to re-examine and revise the pairwise comparisons in the matrix. The ideal of the AHP pairwise comparison model offers a strong and justifiable framework for identifying the most suitable (reasonable and feasible) options from a wide set of location alternatives.

Case study for AHP

The research aims to determine the most suitable technology for achieving sustainability in Plant Breeding and Genetics, focusing on five major crops: oilseeds, pulses, cotton, wheat, and maize. The evaluation is based on five key criteria: high yield potential, broad adaptation, resistance to biotic/abiotic stresses, good stability, and early maturity. The initial step in the AHP involves creating a graphical representation of the decision problem, structured as a hierarchy. Figure 1 illustrates this hierarchy, with the top level representing the overall goal-selecting the appropriate technology for sustainability. The second level outlines the five evaluation criteria that contribute to this goal. The third level includes the five crop alternatives, which serve as the options to be evaluated and compared based on the defined criteria.



To apply the AHP, the decision-maker must express judgments regarding the relative importance of each criterion in contributing to the overall goal. This is typically done through pairwise comparisons. For example, the decision-maker may be asked, "Between High yield potential and Broad adaptation, which is more important for setting research priorities across food crops?" Similar questions are posed for all possible pairs of criteria to construct a pairwise comparison matrix, following the method outlined by Saaty (1986). A sample priority matrix for the criteria (for demonstration purposes) is presented in Table 1.

Table 1: The relative importance of various criterion.

	High yield potential	Broad adaptation	Resistance to biotic / abiotic stresses	Good stability	Early maturity
High yield potential	1	3	1/2	1/2	5

Broad adaptation	1/3	1	5	4	4
Resistance to biotic / abiotic stresses	2	1/5	1	1/5	4
Good stability	2	1/4	5	1	1/7
Early maturity	1/5	1/7	1/4	7	1

From this preference matrix a corresponding set of weights (the eigenvector w) are determined by the AHP. These are

Citeria	Weight
High yield potential	0.250
Broad adaptation	0.259
Resistance to biotic / abiotic stresses	0.157
Good stability	0.185
Early maturity	0.146

The next step is to make pairwise comparisons for each crop alternative with respect to each of the criteria which is given in table 2 to 6.

Table 2: The relative importance of factors Broad adaptation, Resistance to biotic / abiotic stresses, Good stability, Early maturity with respect to High yield potential.

	Oilseed	Pulses	Cotton	Wheat	Maize
Oilseed	1	1/2	1/4	7	3
Pulses	2	1	1/3	1/7	1/4
Cotton	4	3	1	1/9	1/5
Wheat	1/7	7	9	1	3
Maize	1/3	3	5	1/3	1

Table 3: The relative importance of factors High yield potential, Resistance to biotic / abiotic stresses, Good stability, Early maturity with respect to Broad adaptation.

	Oilseed	Pulses	Cotton	Wheat	Maize
Oilseed	1	1/3	1/5	1/7	2
Pulses	3	1	4	1/5	1/2
Cotton	5	1/4	1	1/2	7
Wheat	7	5	2	1	1/4
Maize	1/4	2	1/7	4	1

Table 4: The relative importance of factors High yield potential, Broad adaptation, Good stability, Early maturity with respect to Resistance to biotic / abiotic stresses.

	Oilseed	Pulses	Cotton	Wheat	Maize
Oilseed	1	1/3	1/2	1/5	2
Pulses	3	1	1/6	1/7	1/4
Cotton	2	6	1	1/5	1/2
Wheat	5	7	5	1	1/5
Maize	1/2	4	2	5	1

Table 5: The relative importance of factors High yield potential, Broad adaptation, Resistance to biotic / abiotic stresses and Early maturity with respect to Good stability.

	Oilseed	Pulses	Cotton	Wheat	Maize
Oilseed	1	1/4	3	1/4	1/3
Pulses	4	1	6	1/5	1/4
Cotton	1/3	1/6	1	1/4	1/2
Wheat	4	5	4	1	1/5
Maize	3	4	2	5	1

Table 6: The relative importance of factors High yield potential, Broad adaptation, Resistance to biotic / abiotic stresses and Good stability, with respect to Early maturity.

	Oilseed	Pulses	Cotton	Wheat	Maize
Oilseed	1	1/5	1/3	1/7	1/5

Pulses	5	1	5	6	8
Cotton	3	1/5	1	1/3	1/2
Wheat	7	1/6	3	1	1/7
Maize	5	1/8	2	7	1

	Oilseed	*Pulses*	*Cotton*	*Wheat*	*Maize*
High yield potential	0.280	0.081	0.169	0.317	0.149
Broad adoption	0.067	0.185	0.241	0.296	0.209
Stress res	0.139	0.083	0.154	0.319	0.301
Good stability	0.095	0.186	0.072	0.258	0.387
Early maturity	0.045	0.499	0.084	0.156	0.241

High yield potential *Broad adoption* *Stress res* *Good stability* *Early maturity*

$$\begin{matrix} Oilseed \\ Pulses \\ Cotton \\ Wheat \\ Maize \end{matrix} = [0.250]\begin{matrix} 0.280 \\ 0.081 \\ 0.169 \\ 0.317 \\ 0.149 \end{matrix} + [0.259]\begin{matrix} 0.067 \\ 0.185 \\ 0.241 \\ 0.296 \\ 0.209 \end{matrix} + [0.157]\begin{matrix} 0.139 \\ 0.083 \\ 0.154 \\ 0.319 \\ 0.301 \end{matrix} + [0.185]\begin{matrix} 0.095 \\ 0.186 \\ 0.072 \\ 0.258 \\ 0.387 \end{matrix} + [0.146]\begin{matrix} 0.045 \\ 0.499 \\ 0.084 \\ 0.156 \\ 0.214 \end{matrix}$$

	Priority
High yield potential	0.134
Broad adoption	0.189
Stress res	0.155
Good stability	0.278
Early maturity	0.242

The composite score indicates that stability has highest research priority followed by early maturity, broad adoption, stress resistance and high yield potential.

References

Bhatt, R., Macwan, J. E. M., Bhatt, D., & Patel, V. (2010). Analytic hierarchy process approach for criteria ranking of sustainable building assessment: A case study. World Applied Sciences Journal, 8(7), 881-888.

Ghosh, A., & Kar, S. K. (2018). Application of analytical hierarchy process (AHP) for flood risk assessment: a case study in Malda district of West Bengal, India. *Natural Hazards*, *94*, 349-368.

Saaty, T. L. (2001). Fundamentals of the analytic hierarchy process. The analytic hierarchy process in natural resource and environmental decision making, 15-35.

Saaty, T. L. (2013). Analytic hierarchy process. In Encyclopedia of operations research and management science (pp. 52-64). Springer, Boston, MA.

Saaty, T. L., & Vargas, L. G. (2012). Models, methods, concepts & applications of the analytic hierarchy process (Vol. 175). Springer Science & Business Media.

National Animal Disease Referral Expert System (NADRES V2.0) - Risk Prediction System for Livestock Diseases in India

Dr. K. P. Suresh
ICAR-NIVEDI, Bengaluru-560 119, Karnataka, India
Email: suresh.kp@icar.org.in

I. Introduction to NADRES v2

The geographic and seasonal distribution patterns of numerous infectious livestock diseases are closely associated with climatic factors, positioning seasonal climate forecasts as crucial components of disease early warning systems (EWS). Recognizing this relationship, ICAR-NIVEDI conceptualized and developed the National Animal Disease Referral Expert System (NADRES), a web-based platform, to enhance proactive surveillance and early detection of major livestock diseases in India. Initiated during the early 2000s and supported by the National Agricultural Technology Project (NATP), NADRES was officially launched in 2005. Initially built on an Oracle database and later migrated to a MySQL architecture, the system enabled access to forewarning information on 16 prioritized livestock diseases at the district level with a two-month lead time. With advancements in remote sensing, satellite data, and AI/ML technologies, NADRES evolved into a more dynamic and scalable system NADRES V2.0. This upgraded version integrates machine learning algorithms, high-resolution climatic and disease datasets, and a robust processing pipeline to support national-level livestock disease forecasting and decision-making.

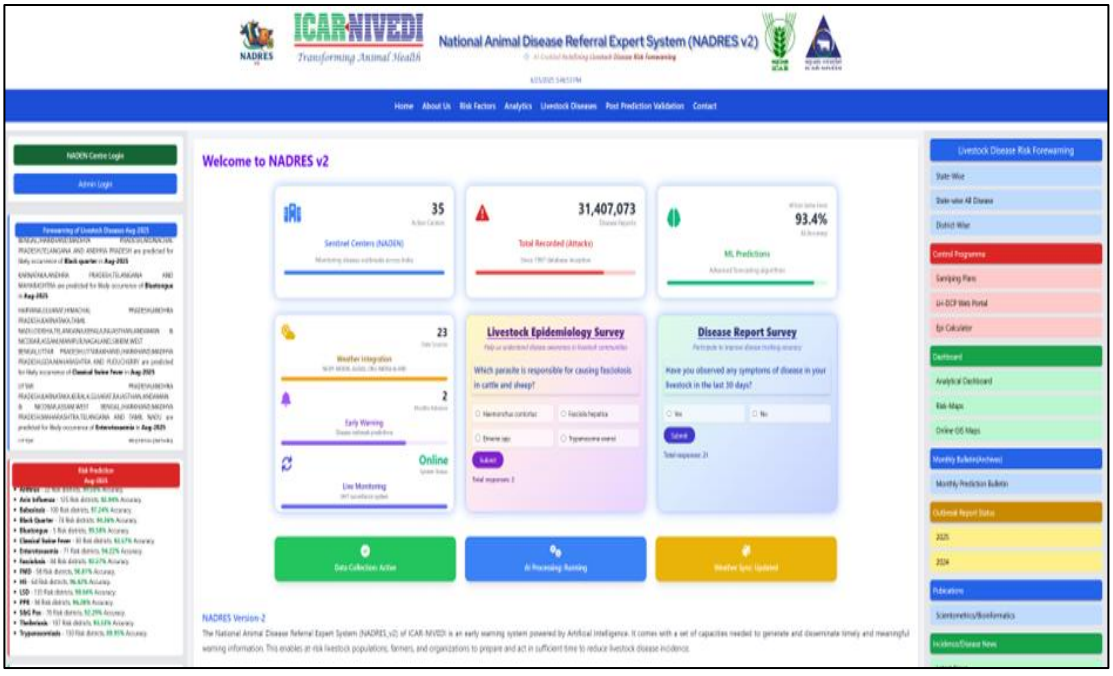


Fig 1. NADRES V_2 Home page

II. Forewarning Methodology

Preamble

NADRES V2.0 functions as a comprehensive early warning system powered by AI and ML, providing timely and actionable information to safeguard livestock health. Its architecture is designed to support preparedness among farmers, veterinarians, and institutions through advanced risk prediction and disease control.

Objectives

- Development of forecasting models for major livestock diseases with a lead time of two months.
- Design and implementation of state-of-the-art communication models for effective dissemination of disease risk information.

1. Materials and Data Acquisition

- **Livestock Disease Data**: A decade's worth of historical livestock disease outbreak data is retrieved from the NADRES database and linked with relevant risk factors.
- **Livestock Population Data**: Population data for cattle, buffalo, sheep, goat, and pig at the village level were obtained from the 20th Livestock Census (2019), DAHD, GoI.

2. Meteorological and Remote Sensing Data

The disease risk forecasting framework integrates livestock census data with multiple climatic and ecological parameters. These include:

- **Remote Sensing Indicators**: LST, NDVI, EVI, PET, LAI
- **Meteorological Variables**: Air temperature, rainfall, wind speed, soil moisture, surface pressure, specific humidity, etc.
- **Data Sources**: NASA's GLDAS, MODAPS, and CRU datasets from the University of East Anglia.

3. Delta Weather Parameters

- **Static Set**: Long-term deltas (2001–2021) to capture climatic trends.
- **Dynamic Set**: Recent averages (2018–2023) for short-term forecasting.
- A total of 46 weather parameters (23 each for static and dynamic sets) are now used for robust modelling.

4. Ecological Parameters

Additional data layers include soil pH, elevation (min/max/mean), and datasets on carbon emissions and water bodies (under process).

5. Risk Modelling Workflow

- Data collection and cleaning
- Feature scaling and extraction
- ARIMA-based weather forecasting (validated by IMD)
- Dimensionality reduction using PCA
- Application of multiple machine learning algorithms
- Model validation using ROC, Cohen's Kappa, and other metrics
- Risk classification into six levels for intervention planning

6. Seven-Step Risk Prediction Methodology

1. Spatial Endemicity
2. Temporal Endemicity
3. Autocorrelation
4. Space-Time Clustering
5. Linear Discriminant Analysis
6. Risk Modelling & Mapping
7. Secondary Infection Analysis (R0)

II. NADRES v2 Data Flow and Data Processing Diagram

A) **Data Flow Diagram:**

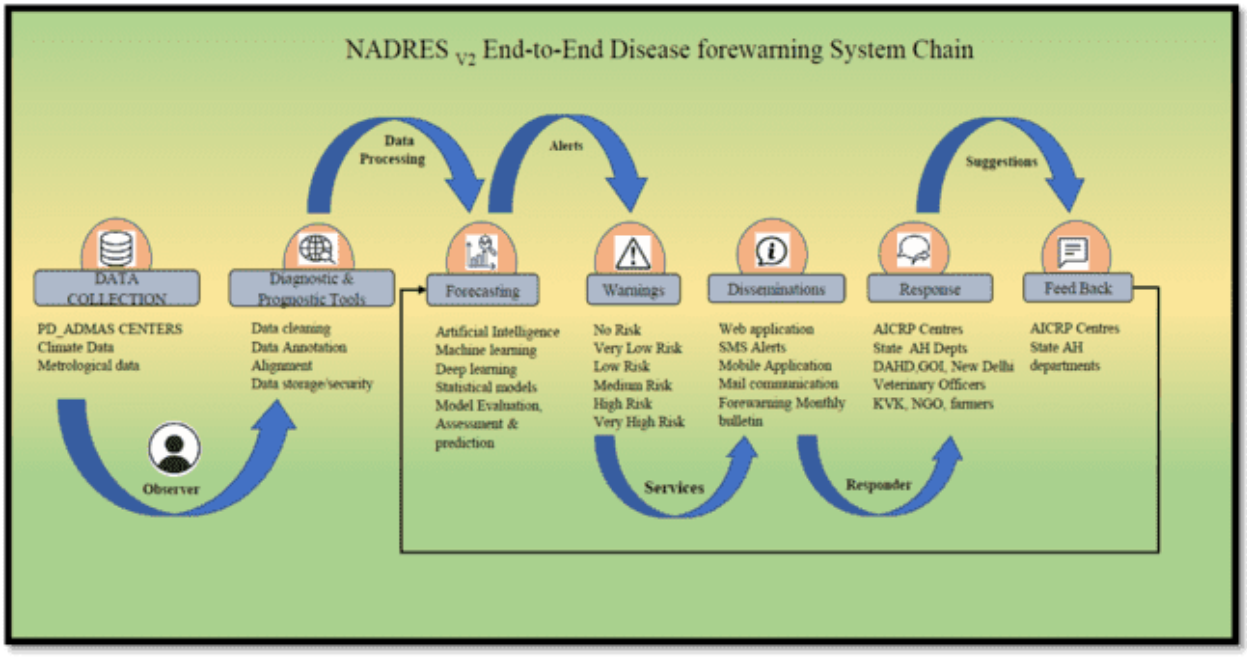


Fig 2. NADRES $_{V2}$ Data Flow Diagram

B) **Artificial Intelligence enabled Data Capturing and Forewarning System:**

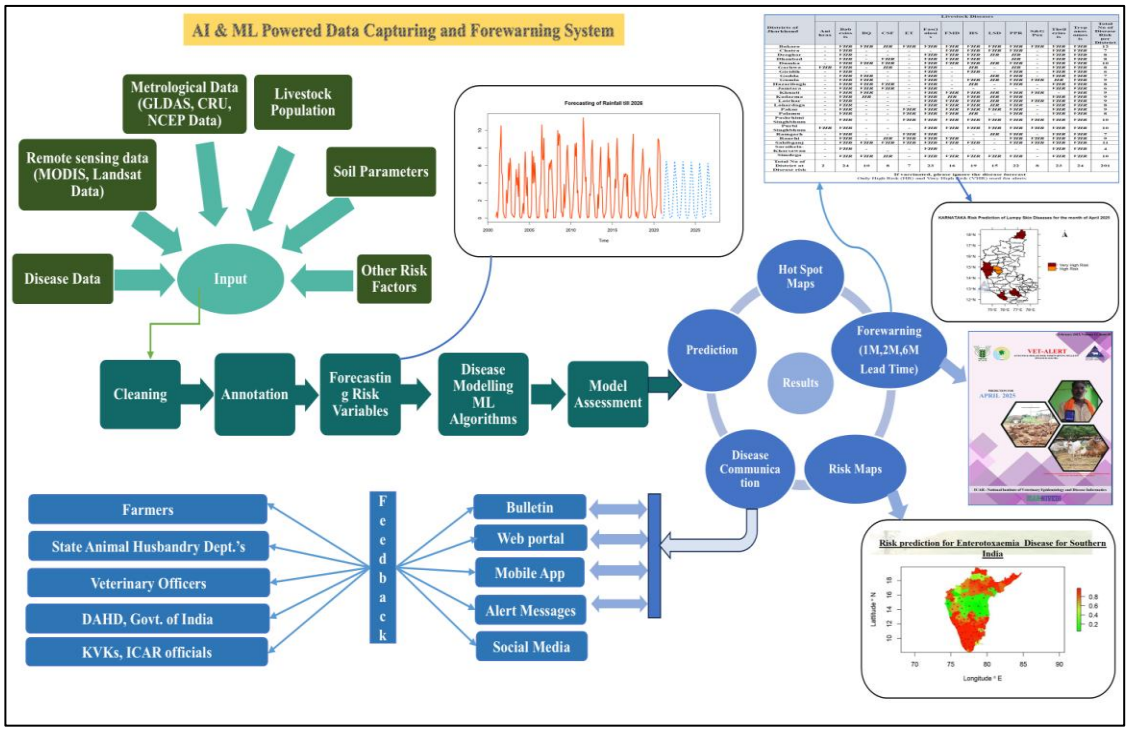


Fig 3. Data Capturing and Forewarning system

III. Machine Learning Models for Disease Prediction

Disease occurrence data is geospatially aligned and linked with climatic risk variables. The following algorithms have been employed:

- **Regression Models**: GLM, GAM
- **Machine Learning Models**: RF, BRT, ANN, MARS, FDA, CTA, XGM, SVM, Tree_prob, LASSO, GP, NN, Multinom_probs, KSVM, Ridge, Elastic Net, CRF

Model performance is assessed using metrics such as ROC, Kappa, TSS, Accuracy, Precision, Sensitivity, Specificity, F1 Score, MAE, RMSE, and Gini Coefficient. Ensemble results are combined using raster stacking. All models are tested for overfitting and optimized for generalization.

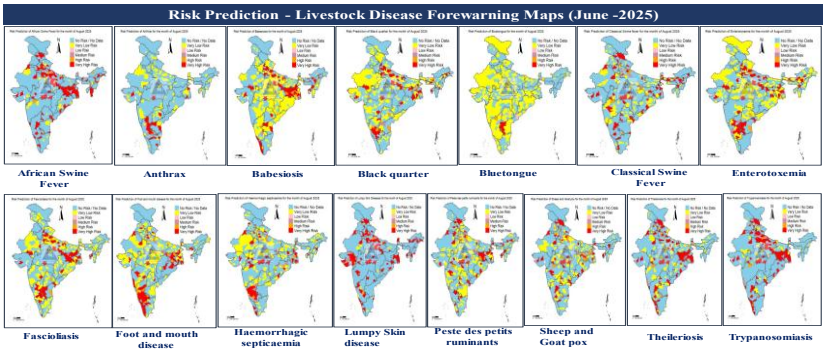


Fig. 6. Risk Prediction – Livestock Disease Forewarning Maps (June 2025)

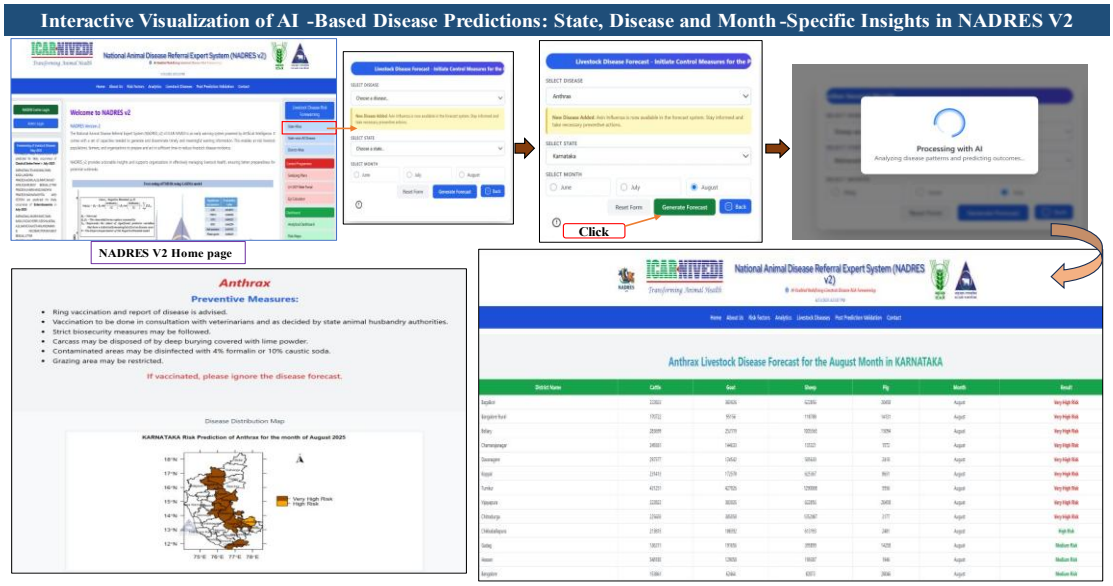
Figure 7. Livestock disease risk forewarning Bulletin Cover pages

IV. Disease Risk Communication Strategy

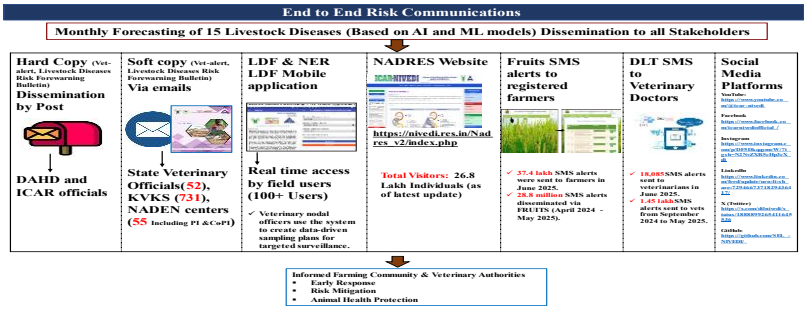
A structured, multi-modal risk communication strategy ensures timely and inclusive dissemination of disease forecasts:

- **Bulletins**: Hard copies sent to ICAR and DAHD officials; digital versions emailed to state departments, KVKs, and NADEN centers.

- **LDF Mobile App**: Offers real-time, location-specific risk data for veterinary officers; supports surveillance and sampling.
- **Web Portal**: Public access to risk maps and forecasts; over 26.8 lakh users recorded.



- **SMS Alerts via FRUITS**: Integrated with the Government of Karnataka's FRUITS platform; 1.22 crore SMS alerts issued between April 2023–March 2024.
- **DLT Alerts**: Direct communication with field veterinarians for rapid response.
- **Social media**: Outreach via YouTube, LinkedIn, Twitter promotes awareness and participatory surveillance.



LDF App Download Links:

- APK: http://www.nivedi.res.in/android_nadres/LDF.apk
- Google Play: https://play.google.com/store/apps/details?id=info.androidhive.ldf

References

Suresh, K. P., Bylaiah, S., Patil, S., Kumar, M., Indrabalan, U. B., Panduranga, B. A., Srinivas, P. T., Shivamallu, C., Kollur, S. P., Cull, C. A. and Amachawadi, R. G., 2022a. A New Methodology to Comprehend the Effect of El Niño and La Niña Oscillation in Early Warning of Anthrax Epidemic Among Livestock. Zoonotic Dis., 2(4):267-290.

Suresh, K. P., Sengupta, P. P., Jacob, S. S., Sathyanarayana, M. K. G., Patil, S. S., Swarnkar, C. P. and Singh, D., 2022b. Exploration of machine learning models to predict the environmental and remote sensing risk factors of haemonchosis in sheep flocks of Rajasthan, India. Acta Trop., 233:106542. doi: 10.1016/j.actatropica.2022.106542.

Suresh, K.P., Hemadri, D., Kurli, R., Dheeraj, R. and Roy, P. 2019. Application of artificial intelligence for livestock disease prediction. Indian Farm, 69: 60-62.

Suresh, K. P., Patil, S., Gowda, R. S., Mohan, G. S. K. D. and Hemadri, D., 2020. Short message service (SMS) alert of occurrence of 13 economically important livestock diseases two months' in advance to veterinary officers at Taluka/Region level of Karnataka state: A Methodology. Clinical Research in Animal Science, 1(2): CRAS.000507.2020.

Suresh K P (2024). NADRES V2-Complete R programming Codes for Livestock Disease Risk Prediction, ICAR- NIVEDI, Bengaluru, 1-105. https://www.nivedi.res.in/Nadres_v2/pdf/manual/Rcodes.pdf

https://www.nivedi.res.in/Nadres_v2/bulletin.php

Social Network Analysis using R for Climatic Data

Prabhat Kumar, K. Ravi Kumar, Ponnaganti Navyasree, Nobin Chandra Paul and Santosha Rathod,
ICAR-National Institute of Abiotic Stress Management, Baramati-413115
Email: kure.ravi@gmail.com

Introduction

Social Network Analysis (SNA) is a methodological approach used to study the relationships and interactions among individuals, groups, systems, or entities. Unlike traditional data analysis that focuses on individual characteristics, SNA emphasizes the connections and patterns of interactions within a network. These networks consist of nodes (also called vertices) representing entities, and edges (or links) representing the relationships or interactions between them. Originally rooted in sociology, SNA has grown into a multidisciplinary tool applied in areas such as epidemiology, communication studies, economics, and increasingly, environmental and agricultural sciences. In recent years, SNA has proven valuable for analyzing complex systems, such as those governing climatic interactions and abiotic stress patterns across regions.

What is SNA?

Social Network Analysis (SNA) is a methodological approach used to study the relationships and interactions among a set of entities referred to as "nodes" that are connected by "links" or "edges". These nodes can represent individuals, organizations, locations, concepts, or any elements capable of forming relationships, while edges capture the nature and strength of their interactions.

SNA allows researchers to move beyond individual attributes and focus on the structure and pattern of relationships within a system. It is widely used across disciplines including sociology, biology, economics, information science, and systems research.

At its core, SNA involves:

- **Nodes**: The entities or actors in the network (e.g., people, organizations, systems).

- **Edges**: The relationships or interactions between the nodes (e.g., communication, collaboration, similarity).
- **Weights**: The strength or intensity of connections (e.g., frequency, importance, or numerical correlation).

How Does Social Network Analysis Work?

SNA operates by:

- **Modeling the system** as a graph, where entities are represented as nodes and their relationships as edges.
- **Analyzing the structure** of the graph using network metrics such as:
    - **Degree centrality** (how many connections a node has),
    - **Betweenness centrality** (how often a node acts as a bridge),
    - **Closeness** and **eigenvector centrality** (how influential a node is).
- **Visualizing the network** to detect hidden patterns, relationships, or clusters.
- **Identifying communities or subgroups**, which are tightly connected portions of the network, using community detection algorithms.

SNA offers valuable insights into system dynamics, interdependencies, and influence pathways that are not easily captured through conventional linear analysis. It is particularly useful for understanding complex systems, mapping influence flows, and optimizing network-based interventions.

Data Requirements for Social Network Analysis in Climatic Studies

To perform Social Network Analysis (SNA) effectively in the domain of climate science and abiotic stress research, diverse and high-quality data are essential. The construction of a meaningful and accurate network depends on the availability of structured information that captures the relationships between different climate-related entities; such as locations, time periods, or variables.

In the context of climatic studies, the following types of data are commonly required:

- **Temporal Climatic Data**: Time series data of temperature, rainfall, humidity, wind speed, solar radiation, etc.
- **Spatial Data**: Geographic coordinates or locations of weather stations or regions under study.
- **Derived Similarity Metrics**: Correlation or distance matrices that quantify the similarity or dissimilarity between stations or variables.
- **Remote Sensing and Satellite Data**: Large-scale environmental observations that can be spatially linked across regions.
- **Sensor-Based IoT Data**: Real-time climatic data from automated sensors deployed in agricultural fields.
- **Anomaly or Event Data**: Information about extreme events (e.g., droughts, heatwaves) and their spatial-temporal occurrence.

Why and When to Use SNA in Climatic Studies?

SNA is particularly useful in climate science for:

- Exploring spatial and temporal interactions in climate variables.
- Identifying zones or stations most affected by or contributing to climatic extremes.
- Revealing early warning signals by monitoring central or bridge nodes.
- Understanding regional connectivity in the spread of abiotic stress.

Use SNA when:

- Climate variables show interdependency across space or time.
- The goal is to analyze system-level behavior (not just individual variables).
- Traditional statistical approaches are insufficient to model complex, dynamic systems.

Analysis and Visualization in R

Purpose of the Analysis

To analyze correlation patterns in rainfall across multiple stations using Social Network Analysis. This helps identify spatial dependencies, similar climatic behavior, and zones with shared rainfall patterns—vital for abiotic stress monitoring (like droughts/floods).

Simulated Rainfall Dataset

We simulate simulated dataset of rainfall (monthly rainfall in mm) for 10 hypothetical weather stations over 12 months, followed by R code to perform Social Network Analysis using correlation-based edges.

Simulate rainfall data

```
set.seed(123)
rainfall_data <- matrix(round(runif(120, 10, 300), 1), nrow = 10, ncol = 12)
rownames(rainfall_data) <- paste0("Station_", 1:10)
colnames(rainfall_data) <- paste0("Month_", 1:12)
rainfall_df <- as.data.frame(rainfall_data)
# Correlation between stations
cor_matrix <- cor(t(rainfall_df))  # transpose: station-wise correlation
# Melt correlation matrix
library(reshape2)
cor_df <- melt(cor_matrix)
# Filter: remove self-loops and keep strong correlations (>0.5)
cor_df <- subset(cor_df, Var1 != Var2 & value > 0.5)
# Build network
library(igraph)
g <- graph_from_data_frame(cor_df, directed = FALSE)
# Plot
plot(g,
     vertex.label.cex = 0.8,
     vertex.color = "skyblue",
     edge.width = cor_df$value * 2,
     main = "Rainfall Correlation Network")
library(tidygraph)
library(ggraph)
g_tbl <- as_tbl_graph(g)
ggraph(g_tbl, layout = "fr") +
  geom_edge_link(aes(width = value), color = "darkgray") +
```

```
  geom_node_point(color = "skyblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1.5, size = 4) +
  theme_void() +
  ggtitle("Rainfall Correlation Network (SNA)")
```

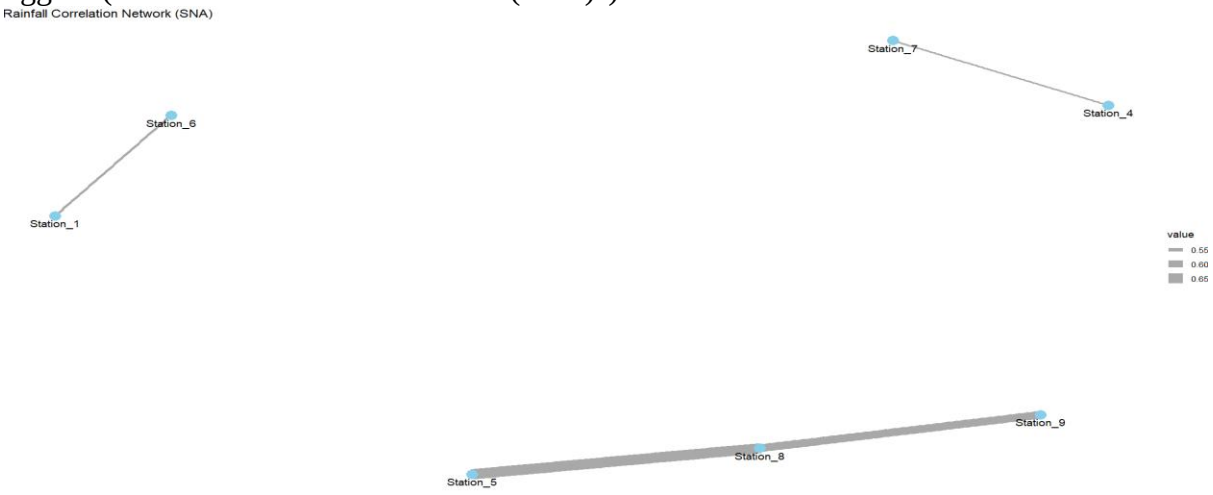


Fig. Network Plot

Interpreting the Network Plot

- **Nodes:** Weather stations
- **Edges:** Significant correlations (r > 0.6)
- **Thicker Edges:** Stronger similarity in rainfall patterns
- **Isolated Nodes:** No significant link with others

E.g., if Station_1, Station_5, and Station_9 are tightly connected, they likely share similar rainfall patterns → zonal behavior → important for climate zoning or irrigation planning.

Conclusion

Social Network Analysis (SNA) provides a robust framework to uncover hidden patterns, relationships, and zones of similarity within complex climatic data. By transforming rainfall correlations among weather stations into a network structure, researchers can visualize and quantify how different regions are interlinked through similar climatic behavior. This approach goes beyond conventional pairwise analysis by emphasizing system-level connectivity, helping identify core zones, bridge regions, and clusters that

might share or propagate climatic stresses like droughts or floods. Integrating SNA into climatic studies supports better-informed decisions for climate-smart agriculture, resource allocation, early warning systems, and regional planning. The example using simulated rainfall data demonstrates how R can effectively model, analyse, and visualize such networks, providing a foundation for further real-world applications.

Suggested Readings

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2018). *Analyzing Social Networks* (2nd ed.). SAGE Publications.

Knoke, D., & Yang, S. (2019). *Social network analysis*. SAGE publications.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications.

Kolaczyk, E. D., & Csárdi, G. (2014). *Statistical analysis of network data with R* (Vol. 65). New York: Springer.

Csardi, G., & Nepusz, T. (2006). The igraph software. *Complex syst*, *1695*, 1-9.

Team, R. C. (2020). RA language and environment for statistical computing, R Foundation for Statistical. *Computing*.

Survival Analysis

Santosha Rathod, Ponnaganti Navyasree, Nobin Chandra Paul, K. Ravi Kumar and Prabhat Kumar
ICAR-National Institute of Abiotic Stress Management, Baramati-413115
Email: Santosha.Rathod@icar.org.in

Introduction:

Survival analysis is a branch of statistics which deals with the expected duration of time until one or more events occurs, such as death in biological organisms and failure in mechanical systems. Survival Analysis is called as reliability theory/ analysis in engineering, duration analysis or duration modelling in economics, and event history analysis in sociology. Survival analysis is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. The event of interest can be can be death, occurrence of a disease, marriage, divorce, unemployment after education, etc. The time to event or survival time can be measured in days, weeks, years, etc. For example, if the event of interest is heart attack, then the survival time can be the time in years until a person develops a heart attack.

In logistic regression, we were interested in studying how independent variables were associated with the occurrence of categorical dependent event/ disease. Sometimes, though, we are interested in how an independent variable affects the dependent event to occur and data underlying process is assumed to be normal. But, in case of time to event occurred data, the underlying process may not follow the normal distribution it may falls under other than normal distributions. In these cases, logistic regression is not appropriate, then survival analysis come into the picture. In survival analysis, subjects are usually followed over a specified time period and the focus is on the time at which the event of interest occurs. Linear regression cannot model the survival time as a function of predictor variables because, times are typically positive numbers; ordinary linear regression may not be the best choice Secondly, ordinary linear regression cannot effectively handle the censoring of observations. The response in survival analysis is often referred to as a failure time, survival time, or event time. The Survival data can be derived from laboratory studies

of animals or from clinical trials, epidemiologic studies and secondary data on crop insurance, time to disease or pest incidence, secondary data on unemployment etc.

Commonly used terminologies:

Survival Time: Survival time refers to a variable which measures the time from a particular starting time (e.g., time initiated the treatment) to a particular endpoint of interest (e.g., attaining certain functional abilities). It is important to note that for some subjects in the study a complete survival time may not be available due to censoring.

Survival Rate: Survival rate is defined as the percent of observations/ peoples who survive a disease such as cancer for a specified amount of time. For example, if the five-year survival rate for a particular cancer is 34 percent, this means that 34 out of 100 people initially diagnosed with that cancer would be alive after 5 years.

Time-to-event: The main variable of interest in survival analysis is time-to-event. It is the time until the event occurs. Time-to-event is a positive random variable. For example, Times to death of patients with certain disease, Remission duration of certain disease in clinical trials, Incubation times of certain disease, such as AIDS, SARS etc., Failure times of certain manufactured products, Life times of elderly in particular social programs.

Censoring: In longitudinal studies exact survival time is only known for those individuals who show the event of interest during the follow-up period. For others (those who are disease free at the end of the observation period or those that were lost) all we can say is that they did not show the event of interest during the follow-up or observational or study period. These individuals are called censored observations. An attractive feature of survival analysis is that we are able to include the data contributed by censored observations right up until they are removed from the risk set. The following terms are used in relation to censoring:

Right censoring: A subject is right censored if it is known that failure occurs sometime after the recorded study period.

Left censoring: A subject is left censored if it is known that the failure occurs some time before the recorded follow-up period. For example, you conduct a study investigating factors influencing days to first oestrus in dairy cattle. You start observing your population at 40 days after calving but find that several cows in the group have already had an oestrus event. These cows are said to be left censored at day 40.

Interval censoring: A subject is interval censored if it is known that the event occurs between two times, but the exact time of failure is not known. For example, an event occurred between date A and date B, but exact date is not known.

Hazard: The instantaneous rate at which a randomly-selected individual known to be alive at time (t - 1) will die at time t is called the conditional failure rate or instantaneous hazard.

Life table: Also called a mortality table or actuarial table, shows, for each age, what the probability is that a person of that age will die before his or her next birthday ("probability of death"). In other words, it represents the survivorship of objects from a certain population, also called as Kaplan–Meier survival life table.

The Hazard and Survival Functions:

Let *T* be a non-negative random variable representing the waiting time until the occurrence of an event. Say, the event of interest as `death' and to the waiting time as 'survival' time, but the techniques to be studied have much wider applicability. They can be used, for example, to study age at marriage, the duration of marriage, the intervals between successive births to a woman, the duration of stay in a city (or in a job), and the length of life.

The Survival Function:

A function describing the proportion of individuals surviving to or beyond a given time. Let T denote the survival time, then the survival function S(t) is defined as follows;

$$\hat{S}(t) = \frac{No.of\ observations\ survives\ longer\ than\ t}{Total\ number\ of\ obseravtions}$$

S(t) = P(surviving longer than time t)

= P(T > t)

The function S(t) is also known as the cumulative survival function, which lies between 0 to 1.

Example: Four animal's survival time are 10, 20, 35 and 40 months. Estimate the survival function.

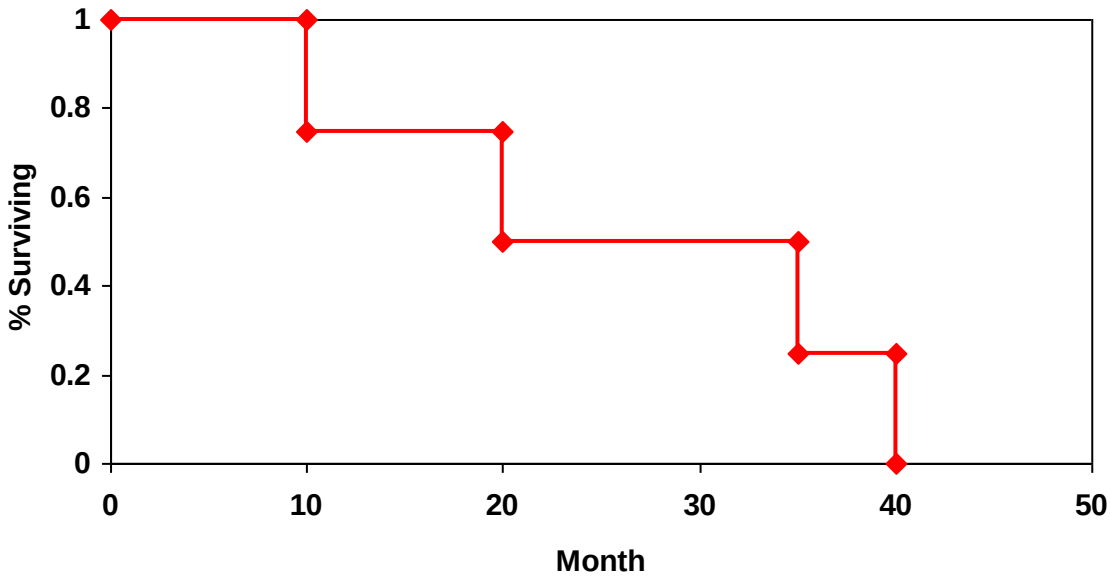


Fig.1: Survival Curve

Let us assume *T* is a continuous random variable with probability density function (p.d.f) *f(t)* and cumulative distribution function (c.d.f) $F(t) = \Pr\{T < t\}$, giving the probability that the event has occurred by duration *t.* It will often be convenient to work with the complement of the c.d.f, *i.e.* the survival function.

$$S(t) = \Pr\{T \geq t\}$$

$$S(t) = 1 - F(t)$$

$$S(t) = \int_t^{\infty} f(x)dx$$

which gives the probability of being alive just before duration *t*, or more generally, the probability that the event of interest has not occurred by duration *t*.

The Hazard Function:

An alternative characterization of the distribution of *T* is given by the hazard function, or instantaneous rate of occurrence of the event, `defined as follows;

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt | T \geq t\}}{dt}$$

The numerator of this expression is the conditional probability that the event will occur in the interval [t, t+dt) given that it has not occurred before, and denominator is the width of the interval.

In other words, the hazards function is defined as the ratio of rate of occurrence of the event at duration *t* equals the density of events at *t*, to the probability of surviving to that duration without experiencing the event.

$$\lambda(t) = \frac{f(t)}{S(t)}$$

The cumulative hazard function is

$$H(t) = \int_0^t h(u)du$$

$$S(t) = e^{-H(t)}$$

It means, the higher the hazard, the lower the survival.

Estimation of Survival Function:

One objective of the analysis of time-to-event data is to estimate and plot the survival function. A very widely used method of estimating survival function is by using Kaplan–Meier method and the resulting plot is called as **Kaplan–Meier curve**. This is a non-parametric method of estimating the survival function. Non-parametric methods are simple methods which lacks the distributional assumptions. Non-parametric methods are very

useful for summarizing survival data and making simple comparisons but cannot so easily deal with more complex situations.

Let $t_1 < t_2 < \ldots < t_2$ be the observed event times and $n = n_0$ the sample size. Let d_j be the number of individuals who have an event at time t_j, where $j = 1, \ldots, k$, and m_j the number of individuals censored in the interval $[t_j, t_j + 1)$. Then $n_j = (m_{j+} d_j) + \ldots + (m_k + d_k)$ is the number of individuals at risk just prior to t_j. The Kaplan–Meier (or product-limit) estimator is a non-parametric estimator of the survival function is

$$\hat{S}(t) = \prod_{j:t_{j \le t}} \frac{n_j - d_j}{n_j}$$

Standard errors can be calculated using Greenwood's formula, which approximates the variance as

$$\widehat{Var}\{\hat{S}(t)\} = \{\hat{S}(t)\}^2 \prod_{j:t_{j \le t}} \frac{d_j}{n_j(n_j - d_j)}$$

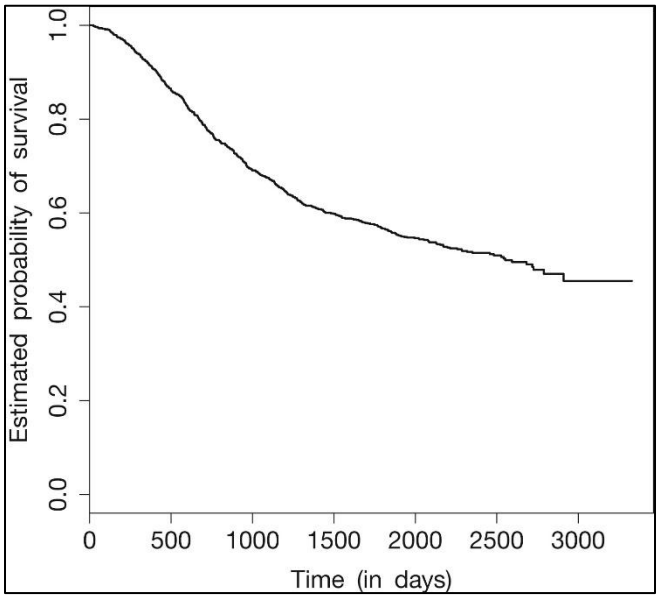


Fig.2: Kaplan-Meier Survivor curve

Figure 2 is an example of a Kaplan–Meier curve, Confidence intervals can be plotted around the curve. An alternative, less commonly used but very similar, non-parametric estimate of the survival function is the life table estimator, based on dividing the time scale

into cells. Kaplan–Meier curves can be used in simple analyses of which the aim is to compare survival times of two or more generally a small number of groups.

To estimate and plot the cumulative hazard function, the Nelson–Aalen estimator can be used. The Nelson–Aalen estimator is a non-parametric estimator of the cumulative hazard function, is as follows;

$$\widehat{H}(t) = \sum_{:t_{j \le t}} \frac{d_j}{n_j} = \sum_{:t_{j \le t}} \hat{h}$$

Where, d_j is the number of individuals who have an event at time t_j, where j=1,…,*k*, and n_j is the number of individuals at risk just prior to t_j.

Comparison of Survival Curves:

When we have more than two groups or survival functions, then we go for comparison of survival times of two or more groups. A simple test of statistical significance is the **log-rank or Mantel–Haenzel test**. It can be used to test whether the survival of individuals in two or more groups is significantly different and it is similar to the $\chi 2$ (chi-squared) test for association.

The Log-Rank Test:

The method calculates at each event time, for each group, the number of events one would expect since the previous event if there were no difference between the groups. These values are then summed over all event times to give the total expected number of events in each group, say E_i for group *i*. The logrank test compares observed number of events, say O_i for treatment group *i*, to the expected number by calculating the test statistic.

Null hypothesis ($H_{0)}$: No difference between (true) survival curves

$$\chi^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2, g - 1$$

This value is compared to a $\chi 2$ distribution with $(g-1)$ degrees of freedom, where g is the number of groups. In this manner, a P-value may be computed to calculate the statistical significance of the differences between the complete survival curves.

Parametric Models:

An alternative basis for estimation and testing in survival analysis is the use of parametric models. Parametric methods are methods in which we make assumptions about the patterns of survival times. The distribution of survival times can be represented using continuous parametric survival models. This can be most easily thought of as assuming that the hazard, as a function of time, has a particular type of shape, with its exact shape being determined by one or more parameters which are estimated using the observed data. Some commonly used distributions in survival analysis are the exponential, the Weibull and the log-logistic distribution. The exponential distribution is the simplest with a single parameter to be estimated and a special case of the Weibull distribution. The choice of parametric family to be used depends on the shape of the distribution. Fitting an exponential distribution to a set of data assumes that the underlying hazard function is constant in time, that is, it assumes that the occurrence of events in time is totally random. A Weibull distribution allows a monotonic (either continuously increasing or decreasing hazard) and a log-logistic distribution allows either a monotonic or a unimodal hazard function.

When using parametric models, we make assumptions the plausibility of which should be investigated. For example, if the event of interest is death of any cause and the time origin is an individual's birth (that is, our time scale is age), then using an exponential model is not a valid option, as the instantaneous all cause death probability (i.e. the hazard) is unlikely to be constant with age. Such a model can be fitted to a set of survival data in order to summarize the features of the data. It may also facilitate the comparison of two or more sets of data. Parametric models can be used in regression analyses of survival data when the effects of other variables on survival are to be investigated. Estimation of the parameters can be done using maximum likelihood. The parameter estimates are found by differentiation the log likelihood with respect to the unknown parameters, setting the derivatives to zero and solving the resulting equations with respect to the parameters

Regression models:

One of the objectives of the analysis of survival data might be to examine whether survival times are related to other features. Regression models can be used to assess the effect of covariates on the outcome. These are similar to regression analyses for other types of outcomes, such as linear regression for a continuous numeric outcome or logistic regression for binary outcomes.

Two common types of regression models for survival data, classified by the way in which covariates are assumed to affect the survival times are the Cox proportional hazards model and the accelerated failure time (or accelerated life) model.

Cox proportional hazards model:

A Cox proportional hazards model is represented as follows;

$$h(t;x) = h_0(t)exp^{\beta x}$$

Where, $h_0(t)$ is the baseline hazard, x is a covariate and b is a parameter to be estimated, representing the effect of the covariate on the outcome. The baseline hazard is the hazard when, in the case of a single covariate, the covariate is equal to zero. The main assumption implied is the proportional hazards assumption, is that the hazard ratio, that is the ratio of the hazard function to the baseline hazard, is constant over time. The use of the exponential function ensures that the hazard is positive. The unknown parameters b in a Cox proportional hazards model can be estimated using the partial likelihood.

The Weibull model (and thus the exponential, being a special case of the Weibull) can be used in regression as a proportional hazards model. If we assume that the explanatory variable acts multiplicatively on the hazard, starting from the survival function of a Weibull model and replacing the baseline hazard with the hazard that includes the explanatory variable effect, we end up with a survival function which has the form of a Weibull distribution but with different parameters. Therefore, the Weibull model belongs

to the family of proportional hazards models. It can be shown that the Weibull model can also be written as an accelerated life model.

Accelerated life models:

An accelerated life model (or accelerated failure time model) is a regression model in which the survival function is assumed to have the same shape for all individuals and explanatory variables are assumed to affect survival by changing the speed with which individuals move on the curve. That is, some individuals move across it more slowly or more quickly than others. So instead of having the hazard multiplied by a quantity as in a proportional hazards model, here the survival time is multiplied by a quantity. This can be represented as;

$$T = \psi T_0 \; or \; T = T_0(t) exp^{-\beta}$$

The resulting survival function is

$$S(t; x) = S_0(exp^{\beta x})$$

Where, $S_0(.)$ is the is the 'baseline' survival of an individual with explanatory variable taking the value zero. The factor e^b is called as the acceleration factor, which, represents how faster or slower an individual would move on the survival curve for a unit increase in the explanatory variable x. If the acceleration factor is greater than 1 then individuals with higher values of x will tend to have earlier event times, whereas if it is less than 1 then individuals with higher values of x will tend to have later event times, that is their survival times will be longer. The hazard and density function can be deduced from the survival function. Unlike proportional hazards models, accelerated life models are usually fully parametric. The log-logistic model is an example of an accelerated life model.

Illustration:

Data from a clinical trial on colon cancer adjuvant therapy1 are used as an illustration. A group of colon cancer patients are followed up from diagnosis to death. That is, the time scale has origin the time of diagnosis of colon cancer and endpoint the time of death from colon cancer. The dataset is freely available in R Software R (dataset 'colon' in package 'survival'), contains 929 observations on colon cancer. Following are the first 20 observations on a subset of the colon data.

Table 1: Colon Cancer data

Id	Study	Rx	Sex	Age	Obstruct	Perfor	Adhere	Nodes	Status	Differ	Extent	Surg	Node4	Time	etype
1	1	Lev+5FU	1	43	0	0	0	5	1	2	3	0	1	1521	2
2	1	Lev+5FU	1	63	0	0	0	1	0	2	3	0	0	3087	2
3	1	Obs	0	71	0	0	1	7	1	2	2	0	1	963	2
4	1	Lev+5FU	0	66	1	0	0	6	1	2	3	1	1	293	2
5	1	Obs	1	69	0	0	0	22	1	2	3	1	1	659	2
6	1	Lev+5FU	0	57	0	0	0	9	1	2	3	0	1	1767	2
7	1	Lev	1	77	0	0	0	5	1	2	3	1	1	420	2

8	1	Obs	1	54	0	0	0	1	0	2	3	0	0	3192	2
9	1	Lev	1	46	0	0	1	2	0	2	3	0	0	3173	2
10	1	Lev+5FU	0	68	0	0	0	1	0	2	3	1	0	3308	2
11	1	Lev	0	47	0	0	1	1	0	2	3	0	0	2908	2
12	1	Lev+5FU	1	52	0	0	0	2	0	3	3	1	0	3309	2
13	1	Obs	1	64	0	0	0	1	1	2	3	0	0	2085	2
14	1	Lev	1	68	1	0	0	3	1	2	3	0	0	2910	2
15	1	Obs	1	46	1	0	0	4	0	2	3	0	0	2754	2
16	1	Obs	1	68	0	0	0	1	0	2	3	1	0	3214	2
17	1	Lev	1	62	1	0	1	6	1	2	3	0	1	406	2
18	1	Lev	1	79	1	0	0	1	1	2	3	1	0	522	2
19	1	Lev+5FU	0	34	1	0	0	1	1	2	2	0	0	887	2
20	1	Lev	0	50	0	0	1	1	0	2	3	0	0	3329	2

Table 2: Data description

Id:	Patient I.D
Study:	1 for all patients
Rx:	Treatment - Obs(ervation), Lev(amisole), Lev(amisole)+5-FU
Sex:	1=male
Age:	in years
Obstruct:	obstruction of colon by tumour
Perfor:	perforation of colon
Adhere:	adherence to nearby organs
Nodes:	number of lymph nodes with detectable cancer
Time:	days until event or censoring
Status:	censoring status
Differ:	differentiation of tumour (1=well, 2=moderate, 3=poor)
Extent:	Extent of local spread (1=submucosa, 2=muscle, 3=serosa, 4=contiguous structures)
Surg:	time from surgery to registration (0=short, 1=long)
Node4:	more than 4 positive lymph nodes
etype:	event type: 1=recurrence,2=death

The variable ‘status’ indicates whether a patient has died or alive, taking the value 1 if a patient has died and 0 otherwise, and ‘time’ is the survival time since diagnosis in days. ‘age’ is the patients’ age at the time of entry into the study, ‘nodes’ is the number of lymph nodes with detectable cancer and ‘node4’ is a binary variable taking the value 1 if the patient has more

than four lymph nodes with cancer and 0 if the patient has fewer than or equal to four positive lymph nodes.

The survival function: A very widely used method of doing that is calculating and plotting a Kaplan–Meier curve. In figure 3 and 4 Kaplan–Meier survival curve, is depicted calculated from the colon data by considering sex and rx variables.

Cox proportional hazards regression model: As part of the analysis Cox proportional hazards regression model was also fitted and results are depicted in Table 3.

Accelerated Failure Time Model: Results of Accelerated Failure Time Model are depicted in Table 4.

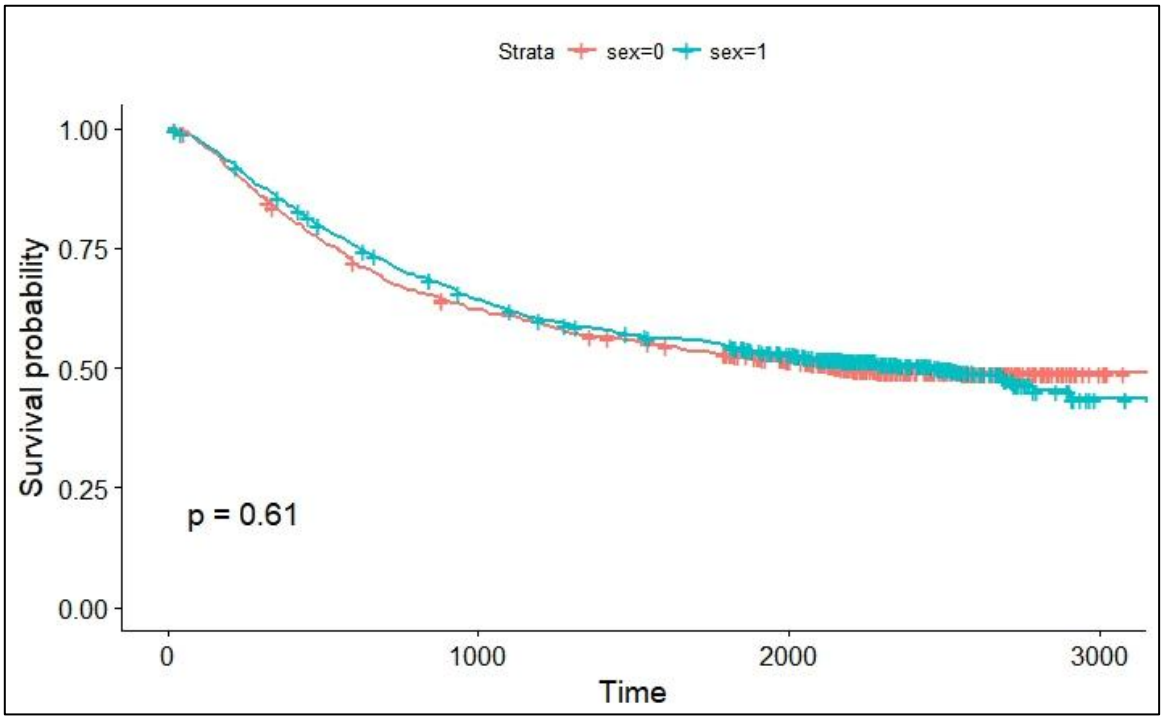


Fig.:3 Kaplan–Meier survival curve for variable sex

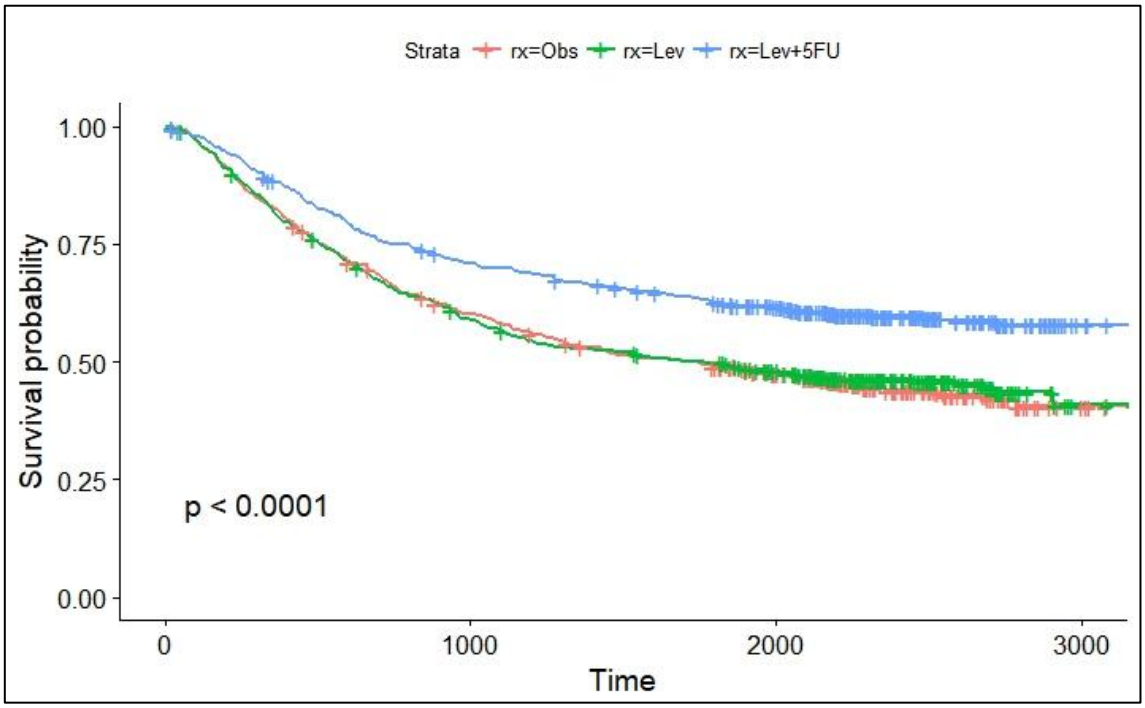


Fig.:4 Kaplan–Meier survival curve for variable rx

Table 3: Fitting of Cox proportional hazards model

	coef e	xp(coef)	se(coef)	z	p
sex	-0.06607	0.93606	0.06828	-0.97	0.33323
age	0.00213	1.00213	0.00291	0.73	0.46476
rxLev	-0.03789	0.96282	0.07957	-0.48	0.63397
rxLev+5FU	-0.43677	0.64612	0.08622	-5.07	4.10E-07
obstruct	0.22062	1.24685	0.08471	2.6	0.00921
perfor	0.12898	1.13767	0.18578	0.69	0.48752
adhere	0.16846	1.18348	0.09254	1.82	0.06871
nodes	0.04075	1.04159	0.01078	3.78	0.00016
differ	0.13872	1.14881	0.07015	1.98	0.04797

extent	0.45389	1.57442	0.08401	5.4	6.60E-08
surg	0.24301	1.27508	0.07432	3.27	0.00108
node4	0.62764	1.87319	0.10059	6.24	4.40E-10
etype	-0.26914	0.76404	0.06778	-3.97	7.20E-05

Table 4: Fitting of Accelerated Failure Time Model

	Value	Std. Error	z	p
(Intercept)	9.69056	0.40934	23.6737	6.73E-124
sex	0.09091	0.08304	1.0948	2.74E-01
age	-0.00455	0.00348	-1.3065	1.91E-01
rxLev	0.00637	0.09868	0.0645	9.49E-01
rxLev+5FU	0.46576	0.10267	4.5367	5.71E-06
obstruct	-0.39551	0.10353	-3.8203	1.33E-04
perfor	-0.04704	0.23454	-0.2006	8.41E-01
adhere	-0.2058	0.11633	-1.7691	7.69E-02
nodes	-0.05982	0.01693	-3.5333	4.10E-04
differ	-0.20853	0.08205	-2.5414	1.10E-02
extent	-0.5706	0.09607	-5.9396	2.86E-09
surg	-0.29579	0.0921	-3.2116	1.32E-03
node4	-0.77004	0.1386	-5.5557	2.76E-08
etype	0.49455	0.08237	6.0039	1.93E-09
Log(scale)	0.43598	0.02657	16.4107	1.60E-60

R codes:

Install the Package “Survival”

Create the response variable:

```
>s <- Surv(colon$time, colon$status)
> class(s)
> head(colon, 20)
> s1fit <- survfit(Surv(time, status)~sex, data=colon)
> plot(s1fit)
> s1fit
> install.packages("survminer")
> ggsurvplot(s1fit, pval=TRUE)
#For rx
> s2fit <- survfit(Surv(time, status)~rx, data=colon)
> ggsurvplot(s2fit, PVal=TRUE)
> summary(s1fit)
> fit1 <- coxph(Surv(time, status)~sex+age+rx+obstruct+perfor+adhere+nodes+differ+extent
+surg+node4+etype,data=colon)
> fit1
>ggsurvplot(survfit(Surv(time, status)~rx, data=colon), pval=TRUE)
#Accelerated Failure Time Model
> artfit <- survreg(Surv(time, status)~sex+age+rx+obstruct+perfor+adhere+nodes+differ+ext
ent+surg+node4+etype,dist="lognormal",data=colon)
> artfit
> summary(artfit)
```

Suggested Readings:

Cox DR. Regression models and life-tables. J R Stat Soc B 1972; 34: 187-220.

Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. 2nd edn. New York: John Wiley & Sons, 2002.

Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958; 58: 457-481.

Kartsonaki C, Cox DR. Some matched comparisons of two distributions of survival time. Biometrika 2016; 103: 219-224.

Kleinbaum, David G., and Mitchel Klein. Survival Analysis: A Self-Learning Text. 3rd ed. ed. Springer, 2012. Statistics for Biology and Health.

Moeschberger ML, Klein JP. Survival analysis: techniques for censored and truncated data. 2nd edn. New York: Springer Science and Business Media, Inc, 2003.

Therneau T (2015). *A Package for Survival Analysis in S*. version .38,

https://CRAN.R-project.org/package=survival.

Van Houwelingen H, Putter H. Dynamic prediction in clinical survival analysis. CRC Press, 2011.

Statistical Methods for Appraisal of Soil Quality Index: Applications in Field and Regional Scale Studies

Rajagopal Vadivel and KS Reddy
ICAR-National Institute of Abiotic Stress Management, Baramati-413115
Email: rgpssac@gmail.com

Abstract

The Soil Quality Index (SQI) has emerged as a comprehensive tool for assessing soil health by integrating multiple soil indicators across physical, chemical, and biological domains. This chapter explores the statistical methods essential for developing SQI and their application in both field-level experiments and regional-scale studies. It traces the conceptual evolution of soil quality from productivity-centered views to multifunctional ecosystem perspectives endorsed by global agencies. Emphasis is placed on the importance of SQI in sustainable land management, climate resilience, and food security. The chapter outlines key components of SQI development, including indicator selection, data collection methods, preprocessing, normalization, dimensionality reduction (e.g., PCA and regression), scoring function transformation, and index construction techniques. It also discusses advanced tools such as fuzzy logic, geostatistics, and machine learning for enhancing spatial and predictive capabilities. Analytical frameworks such as ANOVA, cluster analysis, and geospatial interpolation are reviewed to interpret SQI outcomes. Challenges related to scale dependency, temporal variability, data gaps, and the need for integrating traditional knowledge are critically examined. Finally, the chapter highlights future directions involving remote sensing, open-access digital platforms, and artificial intelligence to improve SQI applicability and decision-making in soil and land resource management.

Keywords: Soil Quality index, Indicator selections; Scoring functions; Data reductions; Artificial Intelligence

1. Introduction

The concept of soil quality has evolved considerably over the past few decades, reflecting the growing recognition of soils as living systems that not only support plant growth but also perform critical ecosystem functions. Initially, soil quality was equated with soil fertility and productivity. For example, during the Green Revolution era, the United States Department of Agriculture (USDA) and related agronomic institutions defined soil quality primarily as the capacity of a soil to support high crop yields. This productivity-centric perspective dominated early discourse, with an emphasis on chemical inputs and crop response. However, a more holistic understanding began to emerge in the

1990s. A widely accepted definition by Doran and Parkin (1994) described soil quality as "the capacity of a specific kind of soil to function within natural or managed ecosystem boundaries, to sustain plant and animal productivity, maintain or enhance water and air quality, and support human health and habitation." This marked a significant conceptual shift, placing equal emphasis on environmental quality and ecosystem functioning alongside productivity. It recognized soils as dynamic systems involved in nutrient cycling, water filtration, and biological activity (Doran & Parkin, 1994; Karlen et al., 1997).

Following this, the USDA's Natural Resources Conservation Service (NRCS) further refined the definition by identifying five key functions of soil: sustaining biological productivity, regulating water flow, filtering and buffering pollutants, cycling nutrients, and supporting biodiversity and root growth. Around the same time, international agencies such as the Food and Agriculture Organization (FAO) began emphasizing soil quality in terms of sustainability. The FAO defined soil quality as the ability of the soil to perform its functions and provide ecosystem services such as biomass production, carbon sequestration, nutrient cycling, and water regulation (FAO, 2020). Their approach highlighted soil's role in supporting ecosystem resilience, biodiversity, and climate regulation. In more recent years, the concept has continued to expand, particularly with the emergence of global initiatives such as the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) and the European Union Soil Strategy for 2030. These frameworks emphasize the contribution of healthy soils to nature-positive solutions, climate change mitigation, and the achievement of Sustainable Development Goals (SDGs) (Bünemann et al., 2018).

As the concept of soil quality became broader and more integrative, the Soil Quality Index (SQI) emerged as a practical tool to quantify and assess soil function. The SQI condenses complex information from multiple soil indicators spanning physical, chemical, and biological domains into a single score that can be used to compare management practices, monitor degradation or improvement, and guide land use planning (Andrews et al., 2004). While there is no universally accepted definition of SQI, it is generally described as a composite index derived from a minimum data set of soil indicators using statistical or rule-based methods to evaluate the soil’s capacity to perform key functions.

Over time, the conceptual development of soil quality and SQI can be traced through key milestones. In the 1960s and 1970s, the focus was on soil fertility and yield maximization. By the 1990s, researchers began considering multifunctionality and environmental impacts. From the 2000s onward, soil quality has been increasingly linked to ecosystem services, policy frameworks, and resilience. In the 2020s, emphasis is being placed on soil biodiversity, climate change mitigation, and the integration of soil health indicators into national and global reporting systems. In summary, the definition of soil quality has

evolved from a narrow focus on productivity to a comprehensive view of soils as foundational to ecosystem health, climate regulation, and human sustainability. Different agencies and frameworks have contributed to this evolution by emphasizing varied but complementary aspects such as productivity, environmental quality, ecosystem services, and policy relevance. This progression has laid a strong foundation for the development and application of statistical methods for SQI appraisal at both field and regional scales.

Importance of SQI in sustainable land management, climate resilience, and food security

The SQI plays a pivotal role in advancing sustainable land management (SLM) by offering a science-based, integrative approach to evaluating the functional capacity of soils. As soil health underpins agricultural productivity, water regulation, biodiversity support, and nutrient cycling, SQI serves as a critical diagnostic and decision-support tool for land users and policymakers (Lal, 2015). In the context of climate resilience, soils act as both sources and sinks of greenhouse gases. An SQI framework can capture soil organic carbon levels, biological activity, and structural stability, which are key to enhancing soil carbon sequestration and reducing emissions. Moreover, SQI enables the monitoring of soil degradation and restoration progress under changing climatic conditions. From a food security perspective, SQI ensures that soil's capacity to support crop growth is assessed not only by yield potential but also by long-term soil health indicators such as nutrient balance, microbial activity, and water-holding capacity. In this way, SQI aligns with global goals for sustainable intensification and resilience in agri-food systems (Pretty & Bharucha, 2014).

Need for robust statistical methods in SQI development

The development of an effective Soil Quality Index depends heavily on the use of robust statistical methods to reduce data complexity, ensure indicator relevance, and maintain scientific rigor. Soils are inherently variable across space and time, and integrating multiple indicators—physical, chemical, and biological—requires analytical techniques that can handle multivariate datasets. Statistical tools such as Principal Component Analysis (PCA), correlation analysis, and regression models are commonly used to identify a minimum data set (MDS), eliminate redundancy, and determine weighting schemes for indicators (Shukla et al., 2006; Ghaemi et al., 2014). Without such techniques, SQI construction can be arbitrary or biased, potentially misrepresenting soil functionality. Moreover, the reliability and comparability of SQI across studies and scales depend on consistent, transparent statistical protocols. Advanced methods like fuzzy logic, geostatistics, and machine learning further enhance the sensitivity and predictive power of SQI models, making them more responsive to complex land management scenarios and environmental changes (Qi et al., 2009).

Difference in Approach Between Field-Level Experiments and Regional Assessments

The approach to SQI development and application differs considerably between field-level experiments and regional-scale assessments, primarily due to differences in scale, objectives, and data availability. In field experiments, SQI is typically used to compare the effects of management practices (e.g., tillage, fertilization, crop rotation) on soil quality under controlled conditions. These studies involve detailed measurements of multiple soil properties with replication and randomized experimental design, enabling rigorous statistical testing such as ANOVA or regression (Andrews et al., 2004). In contrast, regional-scale SQI assessments are designed to evaluate soil health patterns across landscapes, agroecological zones, or administrative boundaries. These assessments often rely on spatial sampling schemes, secondary data, and geographic information systems (GIS) to cover broad areas. Here, techniques like stratified sampling, geostatistics, fuzzy logic modeling, and remote sensing are used to interpolate and map SQI spatially (Qi et al., 2009; Ghaemi et al., 2014). Additionally, while field studies may prioritize short-term functional changes due to specific treatments, regional assessments aim to support land use planning, policy interventions, and long-term sustainability goals. The choice of indicators, scoring methods, and interpretation frameworks must therefore be adapted to suit the scale and purpose of the SQI analysis.

2. Indicator Selection and Data Acquisition

2.1. Indicator Selection

Selecting appropriate soil quality indicators is a foundational step in developing a reliable Soil Quality Index (SQI). The effectiveness and relevance of the SQI depend on the careful selection of a set of indicators that are sensitive to management practices, measurable using standard procedures, and functionally relevant to soil processes (Andrews et al., 2004; Bünemann et al., 2018). The main criteria for indicator selection include sensitivity, measurability, and functional relevance. Sensitivity refers to an indicator's ability to respond to land use changes, agricultural practices, and environmental stressors within a reasonable time frame. Measurability emphasizes the practicality and reproducibility of the indicator across sites and laboratories, ensuring that data can be consistently gathered and interpreted. Functional relevance links each indicator to specific soil functions such as nutrient cycling, water retention, aeration, organic matter decomposition, and microbial activity. Indicators that reflect multiple soil functions or integrate across physical, chemical, and biological processes are especially valuable for inclusion in an SQI framework (Bünemann et al., 2018).

Soil indicators can be broadly grouped into three major categories: physical, chemical, and biological.

Physical indicators describe the structural and hydrological properties of soil, which directly influence root growth, water infiltration, and soil aeration. Commonly used physical indicators include bulk density, which provides insight into compaction and root penetration; porosity, which determines the soil's ability to store air and water; and aggregate stability, which reflects resistance to erosion and structural degradation. High-quality soils generally exhibit low bulk density, high porosity, and well-formed aggregates.

Chemical indicators assess the soil's nutrient content, buffering capacity, and chemical environment. Parameters such as soil pH and electrical conductivity (EC) reflect the chemical environment influencing nutrient availability and microbial activity. Organic carbon (OC) content is a central chemical indicator as it serves as a source of energy and nutrients for soil organisms and plants. The availability of macro-nutrients like nitrogen (N), phosphorus (P), and potassium (K) is directly related to plant productivity, while cation exchange capacity (CEC) serves as an indicator of the soil's ability to retain and supply nutrients.

Biological indicators are increasingly recognized as critical components of soil quality assessments because they reflect dynamic processes such as decomposition, nutrient cycling, and microbial interactions (Bünemann et al., 2018). Key biological indicators include microbial biomass carbon (MBC), which measures the living portion of soil organic matter and serves as a proxy for microbial activity and soil fertility. Soil enzyme activities (e.g., dehydrogenase, phosphatase, urease) are also widely used as they provide insights into the biochemical functioning of the soil and its capacity for nutrient transformations. Additionally, microbial diversity assessed using molecular tools or functional assays offers valuable information about the resilience and adaptability of the soil biological community. In summary, the selection of soil quality indicators must be context-specific, balancing scientific relevance with practical feasibility. While some indicators may be universally important (e.g., organic carbon, pH, bulk density), others may be chosen based on land use, agroecosystem type, or specific research objectives. A well-designed indicator set ensures that the SQI reflects not just the current state of the soil but also its capacity to sustain productivity and ecosystem services over time.

2.2. Data Collection Methods

The reliability and interpretability of the Soil Quality Index (SQI) are fundamentally influenced by the quality and representativeness of the data collected. Data collection methods must be tailored to the scale of assessment—whether field-level experiments or regional-scale studies—as each demand's different designs and sampling strategies to capture variability and ensure statistical robustness (Karlen et al., 1997).

In field experiments, where the primary objective is to evaluate the effect of specific treatments or land management practices on soil quality, data collection is typically guided by experimental designs that allow for replication, control, and statistical testing. One of the most widely used designs is the Randomized Block Design (RBD), where treatments are randomly assigned within blocks that account for known variability such as slope or soil type. This design minimizes experimental error and enables clear comparison of treatment effects on soil indicators. The split-plot design is another common approach, especially when dealing with two factors such as tillage and fertilization. In this method, one factor is applied to main plots and another to sub-plots, allowing researchers to test interactions between factors efficiently. Strip trials are often used in on-farm research or farmer participatory studies where treatments are applied in large, continuous strips across a field, mimicking real-world conditions while capturing spatial variability. Soil samples in field experiments are usually collected at multiple depths and time intervals to capture both spatial and temporal dynamics of soil quality indicators (Andrews et al., 2004).

In contrast, regional-scale soil quality assessments are aimed at understanding patterns across landscapes, watersheds, or administrative zones. These studies require sampling methods that ensure representation of the heterogeneity found across diverse soil types, land uses, and topographies. Stratified random sampling is a commonly used method in which the region is divided into homogeneous strata—such as by land use, slope position, or soil type—and random samples are collected within each stratum. This approach increases sampling efficiency and improves statistical representativeness. Another widely used method is transect-based sampling, where samples are collected systematically along linear paths that traverse key landscape features. This method is particularly useful in hilly or undulating terrains where slope and aspect influence soil properties. Soil survey grid sampling is often employed for spatial mapping of SQI using geostatistical tools. In this approach, samples are collected at regular intervals (e.g., 1 km × 1 km grid) across the study region, and data are interpolated using techniques such as kriging to generate spatially continuous SQI maps (Qi et al., 2009).

Each method has its own advantages and limitations. While experimental designs offer controlled environments for testing hypotheses and comparing practices, regional sampling approaches are essential for generating large-scale insights that inform land-use planning, policy-making, and resource management. The choice of data collection method should be aligned with the study objectives, scale, variability of the study area, and available resources to ensure that the resulting SQI is both meaningful and actionable.

3. Data Preprocessing and Normalization

Preprocessing soil quality data is a critical step before applying statistical analysis and constructing the Soil Quality Index (SQI). Raw soil data, especially when collected

from diverse sites or treatments, often contain inconsistencies, outliers, missing values, and varying scales of measurement. To ensure valid comparisons and robust results, data must be cleaned, normalized, and transformed appropriately. This step enhances the reliability of statistical outputs and prevents the dominance of indicators with larger numerical ranges over those with smaller but equally important values (Nielsen & Winding, 2002; Sun et al., 2013).

One of the first steps in data preprocessing is outlier detection, as extreme values can skew statistical results and misrepresent the true state of soil properties. Visual tools like boxplots are commonly used for preliminary detection of outliers by identifying data points lying outside the interquartile range (IQR). For more formal testing, the Grubbs test is frequently employed to determine whether a specific value in a univariate dataset is significantly different from the rest (Mowlick et al., 2014). Identified outliers can either be removed or winsorized, depending on their cause and potential influence.

After cleaning, normalization of soil indicators is essential because the variables used in SQI calculations often differ in units and ranges. For instance, soil pH typically ranges from 4 to 9, whereas microbial biomass carbon may range from a few mg/kg to several hundred. To enable meaningful integration of such indicators, various normalization techniques are applied. Min–Max scaling is a widely used method where each value is transformed to a range between 0 and 1, using the formula: (value – min) / (max – min). This method preserves the relative position of data points and is suitable for indicators with known boundaries. Another commonly used method is Z-score normalization, which transforms data based on the mean and standard deviation, effectively centering the data around zero with a standard deviation of one. This approach is particularly useful for datasets that are normally distributed and when detecting deviations from the mean is important. Percentile ranking is also used in some SQI models to assign scores based on the position of a value relative to the rest of the dataset. This non-parametric approach is less sensitive to outliers and is effective in large-scale or regional studies (Ghaemi et al., 2014).

In addition to normalization, handling missing data is a crucial component of data preprocessing. Missing values may arise due to sampling errors, instrument failure, or laboratory issues. Ignoring missing data or handling it improperly can lead to biased or incomplete analysis. Common strategies include imputation, where missing values are estimated based on statistical relationships using methods such as mean substitution, regression imputation, or nearest neighbor estimation. Alternatively, if the proportion of missing data is small and randomly distributed, deletion rules such as listwise or pairwise deletion may be applied (Qi et al., 2009). The choice of method depends on the nature of

the missing data, the size of the dataset, and the statistical techniques to be used in subsequent steps.

In summary, data preprocessing including outlier detection, normalization, and handling of missing values ensures that the dataset used for SQI construction is clean, consistent, and comparable across indicators. These preparatory steps form the backbone of reliable SQI development, especially when integrating diverse indicators collected across field plots or large-scale landscapes.

4. Indicator Reduction Techniques

Developing a robust and meaningful Soil Quality Index (SQI) often involves the challenge of handling a large number of soil indicators, many of which may be correlated or redundant. Including too many variables can dilute the interpretability of the index and introduce statistical noise. Therefore, indicator reduction techniques are applied to identify a core set of meaningful and non-redundant variables, often referred to as the Minimum Data Set (MDS). The MDS includes the most informative indicators necessary to capture the variability in soil quality without compromising the functional representation of soil health (Andrews et al., 2002; Shukla et al., 2006).

4.1 Principal Component Analysis

One of the most widely used techniques for reducing dimensionality in soil datasets is Principal Component Analysis (PCA). PCA transforms the original correlated variables into a smaller number of uncorrelated variables known as principal components, which capture the majority of variance in the dataset. Each principal component is a linear combination of the original variables, weighted by their contribution to total variability (Ghaemi et al., 2014).

The process begins with standardizing the dataset to ensure comparability among variables with different units. The selection of principal components for MDS is usually guided by the eigenvalue >1 rule, where components with eigenvalues greater than one are retained. This criterion ensures that only those components explaining more variance than a single original variable are considered significant. A scree plot, which displays the eigenvalues in descending order, is used to visually identify the point where additional components contribute minimally to explaining variance (the “elbow” point). To improve interpretability, Varimax rotation is often applied to the factor loadings, allowing the original variables to load more strongly onto specific components, making it easier to assign functional meaning to each component (Qi et al., 2009).

For example, in a field trial evaluating the impact of different nutrient management systems on soil quality, 15 soil indicators were initially measured including pH, EC, bulk density,

organic carbon, available nitrogen, phosphorus, potassium, microbial biomass carbon, and enzyme activities. PCA reduced these 15 variables into five principal components that explained over 80% of the total variance. From these, a Minimum Data Set was derived by selecting one indicator from each principal component typically the one with the highest absolute factor loading and functional relevance (e.g., organic carbon from PC1, available phosphorus from PC2). This streamlined set of indicators was then used to construct the SQI with reduced redundancy and increased clarity.

4.2 Correlation Analysis and Stepwise Regression

Another useful approach for indicator reduction is the combination of correlation analysis and stepwise regression. Correlation analysis helps identify highly correlated variables, which may provide overlapping information about soil processes. When two variables show a high correlation (e.g., $r > 0.85$), one of them may be dropped to avoid redundancy. However, care must be taken to retain indicators that represent distinct soil functions, even if they are statistically correlated (Sharma et al., 2005).

Stepwise regression, either forward or backward, can be used to determine which variables best explain variation in a dependent variable of interest—such as crop yield, carbon sequestration, or a composite soil function score. This method involves iteratively adding or removing variables based on statistical criteria such as the Akaike Information Criterion (AIC), p-values, or adjusted R^2, ultimately retaining only those predictors that make a significant contribution to explaining the outcome. When used in conjunction with PCA or as a standalone method, stepwise regression is particularly valuable in studies with a clear response variable and when modeling relationships between soil indicators and ecosystem services (Velasquez et al., 2007).

Together, PCA and regression-based methods offer powerful tools for simplifying complex soil datasets. While PCA is more appropriate for exploratory, unsupervised dimensionality reduction, correlation and stepwise regression are suitable for confirmatory analyses aimed at identifying functionally independent and predictive indicators. The choice of method depends on the study design, data structure, and research objectives, but both approaches enhance the efficiency, interpretability, and accuracy of SQI models.

5. Scoring Functions for Indicator Transformation

Once relevant soil quality indicators have been selected and pre-processed, the next step in the construction of a Soil Quality Index (SQI) is to transform raw indicator values into standardized scores. This transformation enables integration of diverse indicators—often measured in different units and scales—into a common scoring system (typically ranging from 0 to 1 or 0 to 100). The choice of scoring function is crucial because it directly

affects how soil conditions are interpreted and ranked. Several types of scoring functions are commonly used in SQI development, including linear, non-linear, threshold-based, and expert- or rule-based scoring methods (Andrews et al., 2004; Masto et al., 2008).

Linear scoring functions are among the simplest and most widely used. They assume a direct, proportional relationship between the raw value of an indicator and soil quality. For example, if higher values of organic carbon or microbial biomass are considered beneficial, a "more is better" linear scoring function is applied. Conversely, for parameters like bulk density or electrical conductivity, where lower values are preferred, a "less is better" function is used. Linear functions are easy to implement and interpret but may not accurately reflect complex, non-linear soil processes (Karlen et al., 2003).

To address such complexities, non-linear scoring functions such as sigmoid or exponential models are often used. These functions capture diminishing returns or threshold effects, where improvements in soil quality indicators may level off beyond a certain point. For example, the biological activity may increase rapidly with organic carbon up to a critical threshold but show little improvement thereafter. Sigmoid curves (S-shaped) are particularly useful in modelling such responses, while exponential functions can capture rapid changes at low or high ends of an indicator range. These models are more realistic for biological and ecological processes but require careful parameterization and calibration (Velasquez et al., 2007). Threshold-based scoring involves assigning scores based on whether an indicator falls within predefined “optimum,” “acceptable,” or “poor” ranges. This approach is especially useful for indicators with well-established agronomic or ecological thresholds, such as pH (e.g., optimum range 6.5–7.5) or salinity levels (e.g., EC < 2 dS/m for non-saline conditions). Scores are typically assigned categorically or on a step-wise basis. Though simple and intuitive, threshold methods may not capture the continuous nature of soil variability and can introduce abrupt changes in scoring (Qi et al., 2009).

Finally, expert opinion or rule-based scoring is applied when dealing with qualitative indicators or when empirical scoring functions are unavailable. This method relies on domain knowledge, field experience, or consensus among experts to assign scores. Examples include assessing soil structure quality based on visual ratings or evaluating land use intensity through management histories. While subjective, this approach adds value in contexts where measurable data are limited or not fully descriptive of soil function (Karlen et al., 1997). The choice of scoring function should align with the nature of the indicator, available data, and the purpose of the SQI. In many cases, a combination of scoring methods is used within the same SQI framework to account for the unique behaviour of different soil indicators.

Table 1. Scoring Functions and Their Suitability for Soil Quality Indicators

Indicator	Indicator Type	Preferred Value	Scoring Method	Remarks
Bulk Density	Physical	Lower	Linear (less is better)	Compaction increases with value; linear decrease in soil quality.
Soil Organic Carbon (SOC)	Chemical	Higher	Linear or Sigmoid	Biological response plateaus at high SOC; sigmoid may be more realistic.
Soil pH	Chemical	Optimum range	Threshold-based	Best between 6.5–7.5; sharp decline outside this range.
Electrical Conductivity (EC)	Chemical	Lower	Exponential decay	High EC adversely affects plant growth; exponential function captures this.
Available Phosphorus	Chemical	Moderate to higher	Sigmoid or threshold	Low P limits growth, excess P can cause runoff; non-linear preferred.
Microbial Biomass Carbon	Biological	Higher	Sigmoid	Increases microbial activity up to a saturation point.
Dehydrogenase activity	Biological	Higher	Linear or sigmoid	Reflects microbial oxidation activity; more is better up to a point.
Soil Structure (visual)	Physical (qual.)	Good structure	Rule-based	Scored based on expert observation or classification.

6. Index Construction Methods

After scoring individual soil quality indicators, the next step in SQI development is to aggregate these scores into a single composite index. This process, known as index

construction, allows diverse indicators spanning physical, chemical, and biological dimensions to be integrated into a unified value that represents the soil's overall health or functionality. The choice of aggregation method can significantly influence the final SQI outcome and its interpretation, especially across different scales and land-use systems. Common methods include weighted additive models, aggregation of scoring functions using mathematical means, and advanced decision-making tools such as fuzzy logic and Analytic Hierarchy Process (AHP).

6.1 Weighted Additive Method

The weighted additive method is the most commonly used approach for SQI construction, particularly in field-scale studies. In this method, each scored indicator is multiplied by a weight representing its relative importance, and the weighted scores are summed to derive the final index. Weight assignment is a critical step and can follow either an equal weightage approach or a PCA-based weightage approach.

In the equal weightage approach, all indicators are treated as equally important, and each is given the same weight (e.g., if five indicators are used, each receives a weight of 0.2). This approach is simple, transparent, and often used when no prior knowledge exists regarding the relative contribution of each indicator to soil function. However, it may oversimplify complex soil processes and ignore the differential influence of certain parameters.

In contrast, PCA-based weighting assigns weights based on the contribution of each indicator to the total variability explained in principal components. Indicators with higher factor loadings in the principal components receive greater weights, thereby incorporating the statistical significance and data structure into the index construction. This method reduces subjectivity and better reflects the variability observed in experimental data, making it more suitable for data-driven field studies.

6.2 Scoring Functions Aggregation

Another set of methods involves aggregating the individual scores using mathematical functions such as the additive, multiplicative, or geometric mean models. In the additive model, the final SQI is simply the sum (or average) of all normalized scores. This method assumes that poor performance in one indicator can be compensated by high performance in others. While this makes the additive model tolerant to variability, it may also mask critical limitations if a key soil function is underperforming.

The multiplicative model (product of scores) assumes that each indicator contributes interactively to soil function. A low score in one indicator can significantly lower the overall index, thereby emphasizing the weakest link. Similarly, the geometric mean model

balances between the additive and multiplicative models, preserving sensitivity to both good and poor values without exaggerating extremes. These models are often selected based on the theoretical relationships among indicators and the desired level of compensability.

A widely used standardized system that incorporates aggregation rules is the Soil Management Assessment Framework (SMAF). SMAF uses non-linear scoring curves, assigns context-specific weights, and applies mathematical rules for combining scores to derive a functional SQI. It is particularly useful in long-term agricultural trials and has been adapted globally for various cropping systems.

6.3 Fuzzy Logic and Analytic Hierarchy Process (AHP)

For regional or watershed-level soil quality assessments, where indicators are often heterogeneous and relationships are non-linear or uncertain, more advanced decision-support methods such as fuzzy logic and Analytic Hierarchy Process (AHP) are preferred.

Fuzzy logic allows for the modeling of uncertainty and gradation in soil indicators by converting crisp input values into fuzzy membership functions. Each indicator is assigned a degree of membership to qualitative categories such as “low,” “medium,” or “high.” This method is particularly useful when indicator thresholds are not rigid or when expert knowledge needs to be incorporated. Fuzzy models enable nuanced interpretation of soil quality, especially in spatially variable and data-scarce regions.

The Analytic Hierarchy Process (AHP) is a multi-criteria decision-making tool that relies on pairwise comparisons of indicators to derive relative weights. These weights reflect expert judgments about the importance of each indicator in relation to soil functions. AHP is highly applicable in regional assessments where local stakeholder input and policy relevance are critical. It provides a transparent framework for integrating qualitative and quantitative information and supports hierarchical evaluation of soil quality at landscape scales.

In conclusion, index construction methods vary in complexity and application depending on the scale, data type, and decision-making context. While additive and PCA-based models are suitable for controlled field studies, fuzzy logic and AHP offer flexibility and interpretive strength for regional planning and natural resource management. The choice of method should align with the study's objectives, the nature of the data, and the end-users of the SQI.

7. Statistical Analysis and Interpretation

Once the Soil Quality Index (SQI) has been constructed, robust statistical analyses are essential to interpret the results, identify significant differences across treatments or landscapes, and validate the classification and spatial distribution of soil quality. A variety of statistical tools from classical inferential techniques to advanced spatial and multivariate methods are used to draw meaningful inferences from SQI data.

To assess the influence of different treatments (e.g., nutrient management systems, tillage practices) or site conditions (e.g., slope positions, land use types) on SQI, Analysis of Variance (ANOVA) or Multivariate Analysis of Variance (MANOVA) is commonly employed. ANOVA allows researchers to determine whether mean SQI values differ significantly among predefined groups. MANOVA, an extension of ANOVA, is used when multiple dependent variables (e.g., individual indicator scores along with SQI) are analyzed simultaneously. These tests are especially useful in field experiments with replicated designs, helping identify which management practices or environmental variables significantly affect soil quality.

Cluster Analysis is a powerful exploratory tool used to group similar soil profiles, land units, or regions based on their SQI values or indicator profiles. Hierarchical or non-hierarchical clustering methods classify sites into homogeneous groups without prior knowledge of group membership. This helps in identifying soil quality zones or management domains, especially in regional-scale studies. For example, cluster analysis can be used to distinguish high-performing, moderately degraded, and severely degraded soil clusters, which can guide targeted soil conservation interventions.

To evaluate how well a set of indicators or derived SQI scores differentiate among known groups (e.g., land use systems, management zones), Discriminant Function Analysis (DFA) is used. DFA develops functions based on weighted linear combinations of variables that best separate the groups. It also serves as a validation tool by checking the accuracy of classification and identifying which indicators contribute most to group discrimination. In SQI research, DFA helps verify whether the selected Minimum Data Set (MDS) and scoring models effectively differentiate between soil quality categories.

For spatial assessments, geostatistical techniques such as Kriging and Inverse Distance Weighting (IDW) are used to interpolate SQI values between sampling points. Kriging is a more advanced interpolation method that considers both the distance and the spatial autocorrelation between points to provide unbiased and minimum-variance estimates. In contrast, IDW estimates values at unsampled locations based on weighted averages of nearby observations, with weights decreasing with distance. Both methods allow the generation of continuous SQI surface maps that reveal spatial patterns of soil quality across regions or watersheds.

To make these spatial patterns accessible and visually interpretable, Geographic Information Systems (GIS) are employed. GIS enables the overlay of SQI maps with other spatial datasets such as land use, topography, rainfall, or administrative boundaries allowing integrated spatial analyses. Visualization tools within GIS platforms (e.g., QGIS, ArcGIS) can be used to classify SQI into categories (e.g., low, moderate, high), highlight critical zones, and support evidence-based land use planning and resource allocation. In summary, statistical analysis and interpretation not only validate the SQI but also enhance its utility in decision-making. From hypothesis testing at the plot level to regional mapping and classification, these analytical methods form the backbone of quantitative soil quality assessment and contribute to the scientific and practical credibility of SQI-based evaluations.

8. Challenges and Limitations

While the Soil Quality Index has emerged as a valuable tool for assessing and managing soil health, its development and application are not without challenges. These limitations arise from methodological, practical, and contextual factors that can influence the reliability, comparability, and interpretability of SQI results especially across different agroecological zones and land use systems.

One major limitation is the scale dependency of indicators and methods. Indicators that are effective at the field level, such as microbial enzyme activities or short-term nutrient availability, may not retain their significance or measurability at the regional or watershed scale, where soil heterogeneity, land use diversity, and climatic variability come into play. Similarly, statistical techniques and scoring models developed for controlled experiments may not be appropriate for large-scale assessments with sparse or inconsistent data. This scale mismatch can lead to either oversimplification or overcomplication, affecting the utility of the SQI for practical land management decisions. Another critical issue is the influence of temporal variability. Soil quality indicators are not static; they respond dynamically to seasonal changes, management interventions, and climatic fluctuations. For instance, microbial biomass and enzyme activity may vary significantly between wet and dry seasons, while soil moisture and nutrient availability can change within days due to rainfall or irrigation events. A single-point sampling approach, although practical, may not adequately capture these temporal dynamics, leading to misleading conclusions about long-term soil quality. Repeated or seasonal sampling adds logistical complexity and cost, but may be necessary for accurate temporal interpretation. In regional assessments, the availability and quality of data pose significant constraints. Soil databases, particularly in developing regions, may be outdated, fragmented, or collected using inconsistent methodologies. This can hinder the application of geostatistical tools or the development of spatially explicit SQI models. Moreover, indicators like biological properties or land use

histories are often missing from national databases, reducing the comprehensiveness of regional SQI evaluations. Limited access to high-resolution remote sensing data, lack of ground-truthing, and institutional constraints further exacerbate these issues.

Finally, a frequently overlooked limitation is the lack of integration of traditional knowledge and farmer perceptions in SQI frameworks. Farmers often possess nuanced, experiential insights about soil behavior, crop response, and long-term land changes, which are not captured through conventional scientific indicators. For example, indigenous soil classifications based on color, texture, or vegetation response may correlate well with scientific indicators but remain underutilized. Including participatory approaches and farmer-derived indicators can enhance the relevance, acceptance, and sustainability of SQI-based recommendations, especially in community-managed or smallholder landscapes. In summary, while SQI is a powerful concept, its application must be context-sensitive, flexible, and inclusive. Addressing challenges related to scale, temporal variability, data quality, and stakeholder engagement is essential to improve the robustness and impact of soil quality assessments. Future work should focus on harmonizing methods, enhancing monitoring systems, and bridging scientific and local knowledge systems to make SQI a truly integrative tool for sustainable land management.

9. Future Perspectives

The advancement of soil quality assessment is entering a transformative era, driven by innovations in data science, earth observation technologies, and participatory digital tools. As the complexity of soil–plant–climate interactions becomes increasingly evident, there is a pressing need for more predictive, scalable, and accessible approaches to Soil Quality Index (SQI) modeling. These emerging perspectives aim to enhance the precision, timeliness, and usability of SQI for researchers, planners, and farmers alike.

A promising frontier lies in the application of machine learning algorithms for predictive SQI modeling. Unlike traditional statistical methods, which often assume linearity and require predefined relationships among variables, machine learning can handle non-linear interactions, large datasets, and complex patterns. Algorithms such as Random Forests (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) have shown significant potential in classifying soil quality levels, identifying key predictors, and generating high-accuracy spatial models. For example, Random Forests can rank the importance of physical, chemical, and biological indicators in determining SQI, while SVM and ANN can predict soil quality classes based on diverse input variables, including topography, climate, and land use. These models improve the objectivity and generalizability of SQI frameworks, especially in heterogeneous and data-rich environments.

Another major advancement is the integration of SQI modeling with remote sensing and geospatial technologies. Satellite data from platforms such as Sentinel-2, Landsat 8, or hyperspectral sensors can provide continuous, near-real-time observations of land surface characteristics like vegetation indices, surface moisture, soil brightness, and land cover. These variables can be correlated with ground-based soil indicators, allowing for the remote prediction and monitoring of SQI across large areas and over time. The coupling of remote sensing data with machine learning models enables the creation of dynamic SQI maps, which can support early warning systems for land degradation, drought impact assessments, and precision land management interventions.

Equally important is the development of open-access, user-friendly digital platforms that democratize SQI knowledge. Web-based interfaces and mobile applications can allow farmers, extension workers, and planners to input local soil and crop information, access SQI scores, and receive tailored management recommendations. Integrating these platforms with national soil health databases, GPS tagging, and weather APIs can provide real-time, location-specific soil quality advisories. Such tools will empower stakeholders—particularly smallholder farmers with actionable insights for nutrient management, conservation practices, and climate-smart agriculture. Initiatives that combine open data standards, participatory design, and multilingual interfaces are crucial to maximize adoption and equity.

In conclusion, the future of soil quality assessment lies in building smart, scalable, and inclusive SQI systems. Machine learning and remote sensing offer powerful tools for predictive and spatially explicit modeling, while digital platforms bring soil information directly to the hands of users. The integration of these technologies promises a shift from static assessments to dynamic soil quality monitoring, aligned with the goals of sustainable land management, climate resilience, and food security.

References

- Andrews, S. S., Karlen, D. L., & Cambardella, C. A. (2004). The soil management assessment framework: A quantitative soil quality evaluation method. *Soil Science Society of America Journal, 68*(6), 1945–1962. https://doi.org/10.2136/sssaj2004.1945
- Bünemann, E. K., Bongiorno, G., Bai, Z., Creamer, R. E., De Deyn, G., de Goede, R., ... & Brussaard, L. (2018). Soil quality – A critical review. *Soil Biology and Biochemistry, 120*, 105–125. https://doi.org/10.1016/j.soilbio.2018.01.030
- Doran, J. W., & Parkin, T. B. (1994). Defining and assessing soil quality. In J. W. Doran, D. C. Coleman, D. F. Bezdicek, & B. A. Stewart (Eds.), Defining Soil

Quality for a Sustainable Environment (pp. 3–21). Soil Science Society of America. https://doi.org/10.2136/sssaspecpub35.c1

- FAO. (2020). Soil pollution: A hidden reality. Food and Agriculture Organization of the United Nations. https://www.fao.org/documents/card/en/c/ca0148en/
- Ghaemi, M., Astaraei, A. R., Emami, H., Nassiri Mahallati, M., & Sanaeinejad, S. H. (2014). Determining soil indicators for soil sustainability assessment using principal component analysis of Astan Quds Razavi farm in Iran. *Ecological Indicators, 36*, 102–110. https://doi.org/10.1016/j.ecolind.2013.07.017
- Karlen, D. L., Ditzler, C. A., & Andrews, S. S. (2003). Soil quality: Why and how? *Geoderma, 114*(3–4), 145–156. https://doi.org/10.1016/S0016-7061(03)00039-9
- Karlen, D. L., Mausbach, M. J., Doran, J. W., Cline, R. G., Harris, R. F., & Schuman, G. E. (1997). Soil quality: A concept, definition, and framework for evaluation. Soil Science Society of America Journal, 61(1), 4–10. https://doi.org/10.2136/sssaj1997.03615995006100010001x
- Lal, R. (2015). Restoring soil quality to mitigate soil degradation. Sustainability, 7(5), 5875–5895. https://doi.org/10.3390/su7055875
- Masto, R. E., Chhonkar, P. K., Singh, D., & Patra, A. K. (2008). Alternative soil quality indices for evaluating the effect of intensive cropping, fertilization and manuring for 31 years in the semi-arid soils of India. *Environmental Monitoring and Assessment, 136*(1–3), 419–435. https://doi.org/10.1007/s10661-007-9692-z
- Mowlick, S., Acharya, S. S., Layek, J., & Ghosh, B. C. (2014). Soil quality assessment using minimum data set under different land use systems in North Eastern Himalayas. *Indian Journal of Soil Conservation, 42*(2), 144–150.
- Nielsen, M. N., & Winding, A. (2002). Microorganisms as indicators of soil health. *National Environmental Research Institute Technical Report No. 388*, Denmark. https://inis.iaea.org/search/search.aspx?orig_q=RN:34011977
- Pretty, J., & Bharucha, Z. P. (2014). Sustainable intensification in agricultural systems. Annals of Botany, 114(8), 1571–1596. https://doi.org/10.1093/aob/mcu205
- Qi, Y., Darilek, J. L., Huang, B., Zhao, Y., Sun, W., & Gu, Z. (2009). Evaluating soil quality indices in an agricultural region of Jiangsu Province, China. *Geoderma, 149*(3–4), 325–334. https://doi.org/10.1016/j.geoderma.2008.12.019
- Sharma, K. L., Mandal, U. K., Srinivas, K., Vittal, K. P. R., Grace, J. K., & Ramesh, V. (2005). Soil quality and productivity improvement under rainfed conditions–Indian case studies. *Soil and Tillage Research, 80*(1–2), 1–14. https://doi.org/10.1016/j.still.2004.03.003

- Shukla, M. K., Lal, R., & Ebinger, M. (2006). Determining soil quality indicators by factor analysis. *Soil and Tillage Research*, *87*(2), 194–204. https://doi.org/10.1016/j.still.2005.03.011
- Sun, B., Zhang, X., & Chen, L. (2013). Evaluating soil quality using a soil quality index for black soil region of northeast China. *Soil and Tillage Research, 130*, 122–130. https://doi.org/10.1016/j.still.2013.02.003
- Velasquez, E., Lavelle, P., & Andrade, M. (2007). Unraveling the relationships between soil properties and macrofauna biodiversity and abundance in soils of Amazonia and the Andes. *Pedobiologia*, *51*(5–6), 289–299. https://doi.org/10.1016/j.pedobi.2007.06.003

भाकृअनुप ICAR

राअस्ट्रैप्रसं NIASM

Twenty-One Day Online Training Program on

Advanced Statistical and Machine Learning Techniques for Data Analysis Using Open-Source Software for Abiotic Stress Management in Agriculture

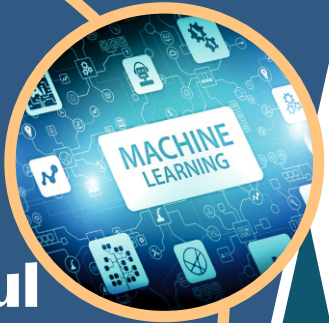
16 July – 5 August 2025

Chief Patron

Dr. K Sammi Reddy, Director, ICAR-NIASM

Patron

Dr. Nitin P Kurade, Head, SSSPS, ICAR-NIASM



Course Directors

Dr. Santosha Rathod
Dr. Nobin Chandra Paul
Ms. Ponnaganti Navyasree
Mr. K Ravi Kumar

Blue Sky SOFTWARE

Google Earth Engine

Organised by:

School of Social Science and Policy Support
ICAR-National Institute of Abiotic Stress Management Baramati, Maharashtra - 413115

About ICAR-NIASM

ICAR-NIASM is the premier institute of ICAR established in 2009 at Baramati. The institute aims at exploring the avenues for the management of abiotic stresses affecting the very sustainability of national food production systems. Besides focusing on developing climate resilient solutions through cutting-edge technologies for managing abiotic stresses, NIASM also aims to enhance scientific capacity through multidisciplinary research and capacity building programs.

About the Training Program

This 21-day online training program offers hands-on experience in advanced statistical, machine learning, and deep learning techniques for analyzing agricultural data. The training is not limited to abiotic stress management; it is applicable across all research disciplines where data analysis plays a critical role. Participants will work with large datasets using open-source tools such as R, Python, QGIS, VassarStats, and BlueSky Statistics through practical, application-oriented sessions.

Key Objectives

The training program aims to:

- Train the participants in multivariate statistics, AI- ML, and deep learning, agroecological modeling tools, remote sensing &GIS
- Provide hands-on experience with open-source software
- Enable independent application of these techniques in research work

Course Content

The program combines theoretical foundations with hands-on practical sessions, enabling participants to apply these techniques to their own datasets efficiently.

Module 1: Software Tools for Data Analysis

- Pre-training session on Installation guide to R/Python/other tools
- Introduction to R
- Introduction to Python
- Introduction to Bluesky Statistics & VassarStats
- Data Visualization in R
- R Shiny and R packages

Module 2: Regression & Multivariate Statistical Methods

- Regression Analysis
- Regression for Categorical Data
- Nonlinear Growth Models
- Regularization Techniques in Regression Models
- Panel Data Regression
- Non-Parametric Analysis
- Data Classificatory Techniques (CA, DA)
- Data Reduction Techniques (FA, PCA)

Module 3: Design of Experiments & Statistical Genetics

- Analysis of Complete and Incomplete Block Designs
- Analysis of Incomplete Block Designs
- Analysis of Groups of Experiments (GOE)
- Response Surface Methodology
- Generation Mean Analysis
- Mating Designs
- Path Analysis
- Stability Analysis
- QTL Analysis
- Transcriptomic Analysis
- Genome Wide Association Studies (GWAS)
- Genomic Selection
- Selection Index
- Meta-QTL Analysis
- Meta-Genomics

Module 4: Machine Learning & Deep Learning Techniques

- Introduction to Machine Learning
- k-nearest neighbor (KNN)
- Artificial Neural Network
- Support Vector Machine
- CART and Decision Tree
- Random Forest Regression
- Extreme Learning Machine
- XGBoost
- Deep Learning: RNN, GRU, CNN, LSTM, Transformer DL
- ML Optimization Techniques
- Yield Forecasting using AI

Module 5: Time Series & Forecasting Methods

- Trend Analysis
- Time Series Analysis
- ARCH Family of Models
- Bayesian Forecasting Models
- Count Time Series Models
- Spatiotemporal Time Series Modelling
- Hybrid Modelling
- Ensemble Modelling
- VAR and Cointegration Analysis

Module 6: Spatial & Environmental Data Analysis

- Introduction to RS & GIS
- Introduction to QGIS
- Introduction to Google Earth Engine
- Spatial Interpolation Techniques
- Introduction to Sampling & Spatial Sampling Strategy
- Application of ML in RS & GIS (ASIS portal)
- Application of UAVs in agricultural data modelling

Module 7: Agro-Ecological Modelling

- Biomass Modelling &Carbon Sequestration using Allometric Models
- CMIP6 GCM Models
- Crop Simulation Modelling (DSSAT and APSIM)
- High-throughput Plant Phenotyping
- Assessment of Extreme Weathers

Module 8: Emerging & Interdisciplinary Topics

- Importance of Data Science in Agricultural Research
- Meta Analysis
- AHP and Grey Model: Technology Forecasting
- Markov Chain Analysis
- Social Network Analysis
- Bibliometric Analysis
- Economic Index Development
- Impact Assessment Modelling, Trend Impact Analysis
- Statistical Modelling in Disease Epidemiology
- Fuzzy Regression Analysis

Expected Learning Outcomes

- By completing this program, participants will master in advanced statistical, machine learning, and deep learning techniques to analyze complex agricultural and environmental datasets.
- They will gain hands-on experience with open-source tools (R, Python, QGIS and others) for data analysis, image processing, and designing efficient workflows to address real-world abiotic stress challenges.

Who Can Apply?

- Researchers and scientists from agriculture, climate science, environmental studies, and allied fields
- Data analysts looking for transition from classical statistics to ML/DL approaches
- Academicians and students seeking proficiency in open-source statistical and geospatial tools

Registration Fee

- ₹ 1000/- for students and research scholars
- ₹ 2000/- for scientists, researchers, faculty members, and working professionals from public organizations
- ₹ 5000/- for participants from private industries

Bank Account Details

Account Holder Name: ICAR UNIT-NIASM, Baramati
Account Number: 30862846914
Name of the Bank: State Bank of India
Branch Address: Afzalpurkar Building, Bhigwan Road,
Baramati, Maharashtra-413102
IFSC: SBIN0000321
UPI Code : icarniasmbmt@sbi

ICAR UNIT NIASM BARAMATI

SCAN & PAY

UPI ID: icarniasmbmt@sbi

Important Dates

- Last Date for Receipt of Applications: 30th June 2025
- Information to Selected Candidate: 2nd July 2025

Registration Link: https://forms.gle/5cMmTxnS19DvWoc48

Contact for Registration Related Queries

Dr. Santosha Rathod
Senior Scientist (Agricultural Statistics)
School of Social Science and Policy Support
ICAR-NIASM, Baramati
Mob: 9900912188

Dr. Nobin Chandra Paul
Scientist (Agricultural Statistics)
School of Social Science and Policy Support
ICAR-NIASM, Baramati
Mob: 8851954194

Ms. Navyasree Ponnaganti
Scientist (ABM)
School of Social Science and Policy Support
ICAR-NIASM, Baramati
Mob: 8639110291

Mr. K. Ravi Kumar
Scientist (Agricultural Extension)
School of Social Science and Policy Support
ICAR-NIASM, Baramati
Mob: 9133120921

Contact Email Id:ssspsniasm@gmail.com

भाकृअनुप
ICAR

हर कदम, हर डगर
किसानों का हमसफर
भारतीय कृषि अनुसंधान परिषद

Agrisearch with a human touch

Application Form

Twenty-One Day Online Training Program

on

Advanced Statistical and Machine Learning Techniques for Data Analysis Using Open-Source Software for Abiotic Stress Management in Agriculture

16 July to5 August 2025

1.	Full Name (in BLOCK letters)				
2.	Highest degree with specialization				
3.	Present Institute Name				
4.	Address for Correspondence				
5.	E-mail address: *Telephone Number Mob/O/R:*				
6.	Date of Birth				
7.	Sex (Male/Female/other)				
8.	Education Qualification:				
	Degree	Subject	Year of passing	Class / Division / Equivalent	University / Institute
	Bachelors Masters Ph.D. Any Other				
9.	Level of Knowledge in Statistics			Beginner / Intermediate / Expert	
10.	Level of Knowledge in R/ Python/ other software			Beginner/Expert	
11.	Area of ongoing research work				
13	Expectations from the training				

**Candidate must fill in all the details*

Signature of the Applicant with date

CERTIFICATE

It is certified that information furnished above is correct.

Signature of the Recommending Authority
/ Head of the Department/ Institute along with Seal