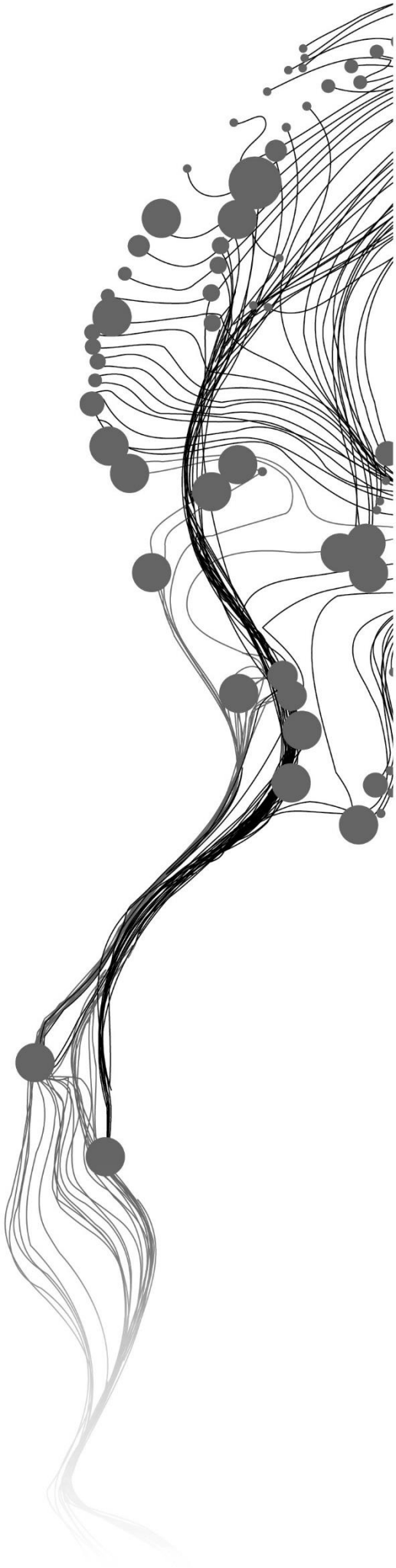
# Integrating Earth Observation Data into Area Frame Sampling Approach to Improve Crop Production Estimates

SARDAR SALAR SAEED DOGAR

June, 2022

SUPERVISORS:
Dr. ir. C.A.J.M. de Bie (Kees)
V. Venus (Valentijn)



# Integrating Earth Observation Data into Area Frame Sampling Approach to Improve Crop Production Estimates

SARDAR SALAR SAEED DOGAR
Enschede, The Netherlands, June, 2022

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Natural Resources Management

SUPERVISORS:
Dr. ir. C.A.J.M. de Bie (Kees)
V. Venus (Valentijn)

THESIS ASSESSMENT BOARD:
Prof. Dr. A.D. Nelson (Chair)
Dr. IR. L.G.J. Boerboom (External Examiner, Faculty of Geo-Information Science and Earth Observation (ITC))

# ABSTRACT

Accurate and reliable agricultural statistics are crucial for understanding the current crop dynamics and improving food security, especially in developing countries. Yield gap analysis provides insights into crop dynamics across the agricultural landscape. It lays the foundation to identify yield constraint factors within fields and improve practices to close the yield gap. However, in many developing countries, current agricultural surveys are established using administrative boundaries and do not reflect the country's agricultural landscape. The most common survey approach is Area Frame Sampling which typically uses Admin areas as primary stratification and does not incorporate AEZ. This research adopts a hybrid approach to identify site-specific crop yield variability and extrapolate it to area-specific crop production estimates by combining statistical and open-source earth observation data. To identify yield constraint factors and quantify crop production function, 503 site-specific wheat yield samples from Punjab, Pakistan, were analysed using Comparative performance analysis (CPA). Long-term NDVI climatology of 20 years is used to capture the agro-climatic conditions over a complex and fragmented agricultural landscape. ISODATA unsupervised classification is used to identify crop phenological cycles and produce Crop Production System Zones (CPS zones). Regression analysis is used to assess the relationship of site-specific measured yield with CPS zones and admin areas. Results revealed that site-specific field parameters explained 41.2 percent of the yield variability. CPS zones based on Earth observation approach explained 23.3 percent of yield variability. A combination model was developed and evaluated to determine the combined impact of site-specific factors and CPS zones. The final model derived through stepwise multiple linear regression included two CPS zones (one from irrigated zone and second from a rainfed zone). This final model explained 43.2 percent of the yield variability. The main findings of this research were as follows: i) UREA fertilizer, broadcasting sowing pattern and seed treatment are identified to be an important field parameters (i.e., explained 28, 19.5 and 14.4 percent of the deviance), ii) Long-term NDVI identified clear crop phenological cycle exists in the study area, iii) CPS zones could differentiate between different rainfed and irrigated croplands. Overall, the study's findings and comparisons support the premise that hyper-temporal earth observation can effectively capture climatological changes in a fragmented agricultural landscape to identify crop yield variability and improve crop production estimates. The method can be applied by government departments and researchers for further studies and aid in decision-making related to closing yield-gap, cropping and food security goals.

**Keywords:** CPS Zones, NDVI, Yield-gap, ISODATA, ProbaV, SPOT, Temporal, Punjab, Pakistan, Yield variability, Area Frame Sampling, Crop Production

# ACKNOWLEDGEMENTS

***"Without education, it is complete darkness, and with education, it is light."***
***-Muhammad Ali Jinnah***

To my primary supervisor Dr. Kees de Bie, I would like to extend sincere gratitude for his continuous support, patience, guidance, and care in every aspect of the research, especially when I have questions about the statistical processes. He not only helped me in finding the way but also boosted my confidence throughout this whole period. I feel no shame in admitting that I could not have completed this research without his guidance and encouragement. We also explored different research angles and had fruitful insights during the research work. He has helped me in develop critical thinking and learn 'seeing, believing, and interpreting.' He always answered all my questions with patience and supported me in my challenging times, for which I will always be grateful to him, and I hope we shall share more in the future.
Furthermore, I would like to thank my secondary supervisor, Mr. Valentijn Venus, for his valuable comments and suggestions in making my result more explainable. I also want to thank Dr. Micheal Marshall for his time and critical comments on my proposal document. His comments pushed me to take control of my thesis topic and complete it correctly.

To my mentor and NRM course coordinator, Drs. R.G. Nijmeijer for his guidance and support during my MSc. I want to show my sincere gratitude to my International and Pakistani friends at ITC; I appreciate all the discussions and moments that we had since our first days at ITC. I would like to specially mention a few names with whom I share the best time in ITC: Sonam, Wondi, Devanshi, Gadisa, Ruhi, Ashfak, Mahnoor, Enting, Akshay, and Mahek.

To senior colleagues Dr. Hammad Gilani (Researcher- Remote Sensing and GIS) from IMWI- Pakistan and Dr. Ahmad Khan (Post-Doctoral Associate) from the Department of Geographical Sciences, the University of Maryland, USA, for their valuable time, comments, and suggestions on my thesis.

To Dr. Abdul Qayyum (Director General) for providing the field data to pursue this research and Mr. Zulfiqar Ali Mayo (Deputy Director Statistics) from Crop Reporting Service, Wing of Agriculture Department of Punjab, Pakistan, for his support and assistance regarding questions of field data.

To the Netherlands Government for providing me with the opportunity to pursue my MSc study in the Netherlands. Without the scholarship, I would not have been here.

Last but not least, to my entire family, including my parents, my beloved wife Amrozia, and my dear siblings Maryam and Uzair for constantly cheering me up. I would like to dedicate my thesis to my dear Mother; it was only through your prayers that I was able to reach this point.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ACRONYMS

| | |
|---|---|
| AEZ | Agro-Ecological Zone |
| AFS | Area Frame Sampling |
| CPA | Comparative Performance Analysis |
| CPSZ | Crop Production System Zone |
| CRS | Crop Reporting Service |
| ERDAS | Earth Resources Data Analysis System |
| ESA | European Space Agency |
| FAO | Food and Agriculture Organization |
| GOP | Government of Pakistan |
| ISODATA | Iterative Self-Organizing Data Analysis Techniques |
| MLR | Multiple Linear Regression |
| MODIS | Moderate Resolution Imaging Spectroradiometer |
| NDVI | Normalized Difference Vegetation Index |
| PROBA-V | Project for On-Board Autonomy- Vegetation |
| RMSE | Root Mean Square Error |
| SDG | Sustainable Development Goal |
| SMLR | Stepwise Multiple Linear Regression |
| SPOT | Satellite for Observation of Earth |
| SPSS | Statistical Package for Social Sciences |

# 1. INTRODUCTION

## 1.1. Background

Food security is a global concern, with around 690 million people suffering from food insecurity worldwide (McCarthy et al., 2018). Most of the world's population lacks access to adequate food; emerging countries in Asia and Africa have the most affected undernourished people; roughly 381 million lives in Asia and 250 million lives in Africa are in danger (United Nations, 2021). The growing global population puts a strain on food demand; on the other hand, climate variabilities such as changes in rainfall pattern, season duration, temperature and crop diseases impact agricultural production, resulting in food insecurity (Becker-Reshef et al., 2020; Zhao et al., 2017). The agriculture sector requires a tremendous deal of care and attention in order to meet this food demand through increased production. Food security is at the heart of the United Nation's Sustainable Development Goals (SDGs) due to its global importance. The SDG's second target is "End hunger, achieve food security, improve nutrition, and promote sustainable agriculture." This goal's targets 2.2 and 2.3 are focused on increasing agricultural productivity and incomes of small farms (Target 2.2) and ensuring sustainable food production systems (Target 2.3) (FAO, 2017; Lobell et al., 2020).

Agricultural statistics are essential for achieving these goals. Crop production, crop acreage, and farmer practices such as field management, land, and crop genetics are all covered in agricultural statistics. The most common agricultural field survey method is Area frame sampling (AFS) (Qayyum et al., 2019; Pan et al., 2010). In the AFS method, randomly selected segments (a piece of land/field) in the country's agricultural landscape are used as sampling units to collect crop analytics (FAO, 2015). For data collection in the field, the Area Frame Sampling (AFS) method employs various sampling strategies. These sampling strategies differ by country due to geography, administrative boundaries, and agricultural activities. Survey methods are developed in accordance with the agricultural profile of the country and the total area under agricultural activities (Pan et al., 2010).

Agricultural surveys must be designed in such a way that they collect complete crop analytics while also explaining agricultural dynamics, particularly in developing countries where, due to the large population and fragmented landscape, the majority of farmers hold small farm holdings (<2ha). Smallholder farming systems are of foremost importance due to their contribution to global crop yield (Jin et al., 2019). 75-80% of the world's agricultural land among 500 million farms is considered as smallholdings (Lowder et al., 2016; Jin et al., 2019). Small farms mostly grow cereal crops (wheat, maize, rice, barley, oat, and sorghum) and produce 30-34% of global crop yield (Ricciardi et al., 2018). Heterogeneity in the fields raises the need for accurate agricultural surveys to capture crop performance, which leads to the identification of yield constraint factors. Yield at the field level is influenced by various on and above-ground factors/ parameters such as temperature, precipitation, soil conditions, crop genetics, management practices (e.g., timing of sowing, application of fertilizer, irrigation), pests, weeds, and crop diseases (Bairagi and Hassan, 2002). All these yield constraint factors affect crop productivity and cause yield gaps. The yield gap is the difference between potential yield and actual yield achieved at the farm level (Lobell et al., 2009a). Equation 1 shows the formula for estimating the yield gap, where Yp is potential yield in a controlled situation and Ya represents actual yield produced by the farmer with limited resources. Potential yield is rarely achieved when all the above-mentioned factors are

considered. Reducing yield gaps within fields can help improve food security and farmer livelihood; as a result, it aids in achieving SDG goals (Beza, 2017).

$$Yield\ gap = Yp - Ya \quad \text{..... (Equation 1)}$$

As the demand for food production rises, so does the pressure on the agricultural system. As a result, it is critical for researchers, agronomists, and farmers to understand the crop dynamics during the cropping season to identify specific factors limiting crop yield within fields (Lobell et al., 2009b). This is also important to estimate yield variability within fields to minimize the yield gap and increase productivity (Dehkordi et al., 2020). Crop yield variability varies between fields, where the size of agriculture fields is small and diverse farming practices are being used (Lambert et al., 2017). The traditional method to collect agricultural statistics is based on agricultural surveys designed using administrative boundaries. Surveys based on administrative boundaries as input for collecting crop analytics do not provide an accurate picture of a country's agricultural system (Kang and Özdoğan, 2019). Agro-ecological zonation[1] (AEZ) is the appropriate and reliable input for designing agriculture surveys to collect crop statistics. AEZ briefly explains the country's agricultural landscape(Kayad et al., 2019). Many developed countries use AEZ as an integral part of their agricultural surveys.

In contrast, it is challenging to plan and execute broader surveys in developing countries at first, and if such survey systems exist, they are based solely on administrative boundaries, which do not indicate the actual places where specific crops are grown (M.R.Khan et al., 2010; Mohammed, 2019). The concept of AEZs has been comprehensively introduced to strengthen the agricultural sector and improve the agricultural statistics. AEZs are created based on the country's geography, climate, soil characteristics, crops, and cropping seasons(Mohammed et al., 2020; FAO, 1978). Agro-ecological zones can improve the quality of field surveys and minimize the extensiveness and bias if they are adequately designed and consider the parameters listed above (Ali et al., 2012; de Bie and Nelson, 2021).

Agricultural field surveys are time-consuming, labour-intensive, and expensive; as a result, the various field- survey methods do not generate accurate agricultural statistics (Jain et al., 2016). Field surveys are undertaken by agriculture extension officers and field enumerators, and the possibility of human error exists, resulting in discrepancies in the field survey, mainly when the agricultural landscape of the country is diverse and fragmented, with mixed cropping patterns. This may impact the country's agriculture policy framework, import and export, and farmers' current farming practices. Many countries, particularly in Asia and Africa, are concerned about the accuracy and reliability of agricultural statistics. In developing countries, the agricultural sector plays a vital role in transforming a country's economic growth and ensuring food security for the growing population (Skakun et al., 2021). To overcome the agricultural survey challenges and improve agricultural statistics, earth observation has proven effective for agricultural mapping. Earth observation provides consistent, efficient, cost-effective, and reliable data for large-scale agricultural areas (Hunt et al., 2019). Integrating earth observation in the area frame sampling approach can improve the ground surveys and aid in producing accurate agricultural statistics of the country.

### 1.2. Earth Observation and Agriculture

Earth observation enables the acquisition of valuable and efficient data for agricultural mapping, which is required for precision agriculture (Jin et al., 2019). With the advancement of satellite technology, remote sensing (RS) technology has been used for agricultural mapping since the 1970s. Remote

[1] Agro- ecological zonation (AEZ): A zonation or stratification defined by climate, topography, and soils.

sensing has proven to be a promising method of obtaining agricultural information from space (Carfagna and Gallego, 2005). The remote sensing satellite industry has evolved over-time and it is classified according to spatial resolution as coarse resolution (>250m), moderate resolution (10-30m), and high resolution (<5m) (Mohammed et al., 2020; Trivedi, 2020). However, globally freely available Sentinel-1 and Sentinel-2 have 10-20m spatial resolution and 5-6 days temporal frequency classified as high-spatial resolution satellites (Chen et al., 2021).

Higher-temporal satellite imagery is distinguished by a very high revisit frequency, typically between one and two days. This high revisit frequency helps in the capture of seasonal and interannual variation between crop fields over both short and prolonged periods (Mohammed et al., 2020; Trivedi, 2020). Freely available satellite imagery of MODIS, SPOT-V & PROBA-V provides 16-days and 10-days of composite NDVI imagery that can be used to analyse vegetation's climatic behaviour over the years. The NDVI offers information about the health of vegetation. Healthy vegetation has higher reflectance in the near-infrared (NIR) wavelength and a lower reflectance in the red wavelength, resulting in high NDVI values (ÜNAL and KEES, 2017). This aids in evaluating the climatic behaviour of vegetation over the years. Due to frequent and cloud-free availability, NDVI products can be used to create a homogeneous landcover stratification (Trivedi, 2020). The long-term NDVI climatology temporal profiles can be used not only for stratification of landcover types but also be used to stratify agricultural landscape into Crop Production System Zones (CPS zones) or AEZs (Ali et al., 2012). It provides insights into cropping seasonality, calendar & management practices (de Bie, C. A. J. M., & Nelson, A. D. 2021). NDVI temporal profiles help in creating CPS zones based on agricultural productivity at a country level to estimate crop yield. (Ali et al., 2012; Kees de Bie et al., 2011; M. R. Khan et al., 2010; Mohammed et al., 2020; Trivedi, 2020) used earth observation satellite imagery with agricultural ground datasets for the stratification of agricultural landscapes in Mekong delta (Vietnam), Nizamabad district (India), Andalucía (Spain), Oromia region (Ethiopia), Eastern region (Ghana). 10-days SPOT VGT and PROBA-V NDVI composites at 1km and 16-days MODIS Terra + Aqua composites at 250m spatial resolution used to stratification agricultural landscape into CPS zones. ISODATA unsupervised classification technique applied for creating CPS zones through stratification. This approach captured intercrops and differentiate between crops/non-crops land.

Very high spatial resolution sensors (IKONOS, QuickBird, SPOT5, RapidEye) capture detailed spatial variation between crop fields, which is helpful for mapping yield variability between fields (Hunt et al., 2019); (Skakun et al., 2021). Due to very high spatial resolution, the acquisition of images during the whole season becomes a hurdle to capture the yield, and the image size is usually large and thus requires a long time for data processing that is technically unsuitable when the area size is large. On the other hand, open-source high spatial resolution satellites like Sentinel-2 (A+B) product by the European Space Agency (ESA) capture ground images with 10m spatial resolution and five days revisit frequency from 2017 to the present and are very useful for crop monitoring and yield estimates (Hunt et al., 2019; Kayad et al., 2019). Compared to Landsat spectral bands, Sentinel-2 is enriched with three additional red-edge bands centred at 705, 740, and 783nm, which are sensitive to capturing crop growth during the growing season (Delegido et al., 2011). (Kayad et al., 2019) used Sentinel-2 spectral bands and vegetation indices to map corn grain yield spatial variability within the field scale in 22 ha of an agriculture field in North Italy. The performance of several vegetation indices was assessed with field data and found that the Green Normalized Difference Vegetation Index (GNDVI) showed the highest $R^2$ of 0.48 for monitoring within-field yield variability. High-resolution satellites worldview-3, planet along with Sentinel-2, and Landsat-8 were used to assess within-field corn and soybean yield variability in 30 fields from Iowa State University. Results showed that moving to moderate from high resolution 10m, 20m, and 30m reduces the explained variability (Skakun et al., 2021). (Hunt et al.,

2019) combined vegetation indices and environmental data with Sentinel-2 to estimate within-field wheat yield in 39 wheat fields in the United Kingdom. They found that Sentinel-2 data with ecological factors improve estimates of within-field yield variability. Optical satellites sometimes counter severe problems for use in crop management because of climatic conditions such as clouds (Chen et al., 2021; Khabbazan et al., 2019). This problem limits the mapping and classification of crop fields (Abubakar et al., 2020; Holtgrave et al., 2020).

Microwave or (SAR) satellites can monitor the Earth from space in all weather conditions, day and night. Researchers and scientists have studied and investigated SAR data for agricultural purposes (Fieuzal et al., 2017, 2013; Larranaga et al., 2013). The most significant limitations of using SAR satellites are data availability, understanding and interpretation of data, and noise interference compared to optical earth observation images (Holtgrave et al., 2020). The recent launch of Sentinel-1 (SAR) opened a new gateway of research horizons in the agricultural domain. Sentinel-1 is the combination of two sensors, Sentinel-1A and Sentinel-1B. The constellation of both provides temporal resolution (revisiting frequency) of 12 days globally and six days specific part over the globe (Abubakar et al., 2020; Veloso et al., 2021). (Khabbazan et al., (2019)) explored Sentinel-1 for crop monitoring and detecting crop growing dates in the Netherlands. Both Sentinel-1 and Sentinel-2 can identify crops, differentiate biophysical characteristics among different crops, and monitor crop growth phases (Mateo-Sanchis et al., 2019).

In previous studies, earth observation proved to be an efficient tool for monitoring and mapping agricultural landscapes (e.g., agricultural zonation, crop type mapping, crop area estimation, yield prediction, and estimation) from small to large scale (Hunt et al., 2019; Kayad et al., 2019; Lambert et al., 2017; Skakun et al., 2021). However, the studies related to yield variability and production were performed on small-scale areas (e.g., field stations, small agricultural lands) mainly due to the unavailability of the field data. Thus, the integration of earth observation with existing survey approaches for identifying yield variability within fields and improving crop production estimates in large regions is still unexplored.

### 1.3. Problem Statement

The gap in various data sources and methods for collecting crop statistics, such as agricultural field surveys and earth observation, is the primary cause of the problem with agricultural statistics. Agricultural statistics are typically collected using the Area Frame Sampling method, which is based on administrative boundaries (M. R. Khan et al., 2010; Mohammed et al., 2020; Qayyum et al., 2019), and does not integrate agro-ecological zonation of agricultural landscape as previously stated. Furthermore, agricultural statistics data are inconsistent over time, limiting their utility for agricultural estimates. Data collection for agricultural statistics is costly, time-consuming, and labour-intensive. Earth observation can be helpful in delineating such agricultural landscapes into homogeneous strata where fragmented crop fields exist, and much heterogeneity between fields exists (Khan et al., 2010).

In contrast, earth observation has been utilized for agricultural mapping, including yield estimations (Rattalino Edreira et al., 2021). The varied satellite designs separated Earth observation into spatial and temporal resolution. Long-term NDVI climatology (Hyper-temporal Imagery) captures the seasonal and interannual variation in agricultural fields and can use to create stratified CPS zones of an agricultural landscape at a coarse resolution with similar soil, climate, and weather characteristics (Kees de Bie et al., 2011; M. R. Khan et al., 2010; Mohammed et al., 2020). It is crucial to explore the potential of integrating earth observation data into the existing area frame sampling approach for producing accurate crop yield estimates.

In this study, a hybrid method was developed and evaluated the feasibility of improving crop production estimates in a complex landscape with smallholder farms by combining earth observation data into an area frame sampling approach. The method employed i) the statistical approach Comparative Performance Analysis to identify yield constraint factors and develop crop production models, ii) coarse resolution earth observation for capturing seasonal and interannual climatological variation and producing CPS zones, iii) assessed yield variation between CPS zones and administrative area, and iv) used step-wise multiple linear regression (SMLR) to quantify the combined impact of field parameters and CPS zones on yield variability.

## 1.4. Research Objectives and Questions

This research aims to integrate earth observation data with existing Area Frame Sampling methods to improve crop production estimations to extrapolate from site-specific yield to area/region-specific crop production estimates. In addition to site-specific data, Crop Production System Zones (CPS zones) derived from long-term NDVI climatology will be used to capture spatial variability in crop performance within zones during cropping seasons.

### 1.4.1. Main Objective

The main objective of this study is to identify patterns of yield variability between surveyed fields that can lead to improved crop production estimates by integrating earth observation data into the existing area frame sampling approach. To achieve this objective, provided below are the sub-objectives and research questions.

1. To assess the causal relationships between site-specific data gathered by agricultural officers and the measured crop yields.
    a) Which field-specific factors (genetics-G, management-M, land-L) correlate significantly with crop yields, and to what extent?
2. To produce crop production system zones (CPSZs) in the agricultural landscape using long-term NDVI climatology from 1999-to 2020.
3. To assess the relationship of site-specific measured crop yields with CPS Zones vs. Tehsilwise Stratification (Admin Areas).
    a) How do yields of sampled sites vary within and between CPS Zones vs. Administrative Areas?
4. To merge all the above into one assessment to quantify their combined impact on crop yield variability, as required to extrapolate site-specific yield data to area-specific crop production estimates.
    a) To what extent can an integrated assessment of all the above-studied parameters explain variability in measured yields?

### 1.4.2. Research Hypothesis

Based on the research questions, the hypothesis adopted in this study are as follows:

H0 a: There is no significant relationship between crop yield and site-specific field parameters collected by Agricultural officers and reported by farmers.

$$Yield \neq f\ (site\text{-}specific\ field\ parameters\ (GxMxL))$$

H1 a: There is a significant relationship between crop yield and site-specific field parameters collected by Agricultural officers and reported by farmers.

H0 b: There are no significant relationship between NDVI- based crop production zones and site-specific measured wheat yield.

$$Yield \neq f\ (NDVI\text{-}CPS\ Zones)$$

H1 b: There is a significant relationship between NDVI- based crop production zones and site-specific measure wheat yield.

H0 c: There is no significant relationship between site-specific measured wheat yield and i) site-specific field parameters and ii) NDVI- CPS Zones

$$Yield \neq f\ (site\text{-}specific\ field\ parameters\ (GxMxL),\ NDVI\text{-}\ CPS\ Zones)$$

H0 c: There is a significant relationship between crop yield and i) site-specific field parameters and ii) NDVI- CPS Zones.

## 1.5. Conceptual Diagram

The study's conceptual diagram is shown in **Figure 1**. The conceptual diagram is divided into two parts: one depicts the current method for crop production surveys, and the other depicts the study's recommended method. Punjab, a Pakistani province, is the system's boundary. Punjab is divided into nine administrative divisions, having 36 districts and 147 tehsils. An area frame sampling approach with a 2-stage sampling method is used in this region to collect crop yield statistics (Qayyum et al., 2019). A village is the smallest administrative entity in this system. Villages represent Union Council, and the union council represents Tehsils. Yield data collected from villages are aggregated to the tehsil level, and then final estimates release. Villages are chosen in the first step from the union councils based on total cultivated land; the more cultivated land, the higher the possibility. Since the selection of villages depends on more croplands (bias of the sampling method), aggregated areas (averaged yield) estimates are likely affected by the sampling method. Following the selection of the village, a land piece of 150 acres is selected, and two 6x8ft plots per crop are sampled. A total of 5500 segments are chosen from the province, with two segments in each union council on average. During the cropping season, Crop Reporting Services (CRS) and Agriculture Department staff survey samples. Estimates are generalized to the districts after gathering samples from the entire tehsil. Agricultural data are calculated using substantial fieldwork undertaken during the growing season. The proposed method of this investigation is depicted in the second portion of this conceptual diagram. Villages will select randomly from the CPSZs that are assumed homogenous in cropping intensity and potential yields. The aggregation and averaging will thus not be biased. The weather may have created the variability within CPS zones; that is why this study also looks at the possibility of identifying these in-season biases too.

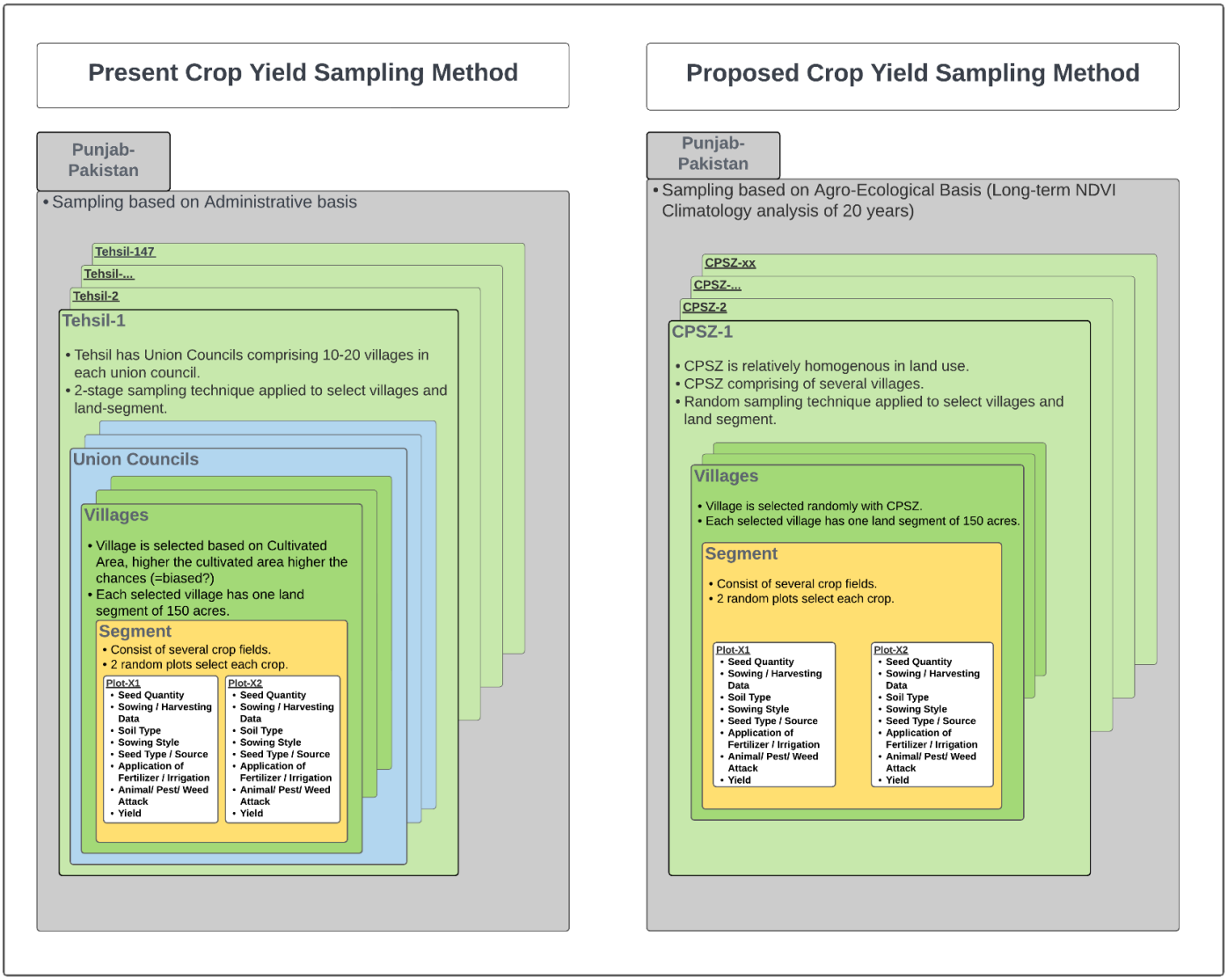



Figure 1: Conceptual Diagram.

# 2. STUDY AREA AND DATASETS

This chapter presents the study area and data that will use in this study. The first section presents the study area, and the second section presents the field data and earth observation data.

## 2.1. Study Area

Punjab is the most populated province of Pakistan, with 53% of the total population. It is the second largest in terms of area with 205,344 km$^2$ which is approximately 25.8% of the total landmass of Pakistan. The province has a significant role in the country's economy; about 42% population is engaged with agriculture sector. Punjab, with 12.5 million hectares (Mha) of cultivated land, is considered the breadbasket of Pakistan, and plays an important role in ensuring the food security of the country's total population. Pakistan is amongst the world's top ten producers of wheat and rice. According to FAOSTAT, Pakistan is ranked 8$^{th}$ in wheat production **Figure 2**.


Figure 2: Pakistan Ranking in the World (Crop Production).

Despite these high ranks and strong production figures, according to UN World Food Program (WFP), only 63.1% of the country's population is food secure, while 36.9% of the population faces food insecurity. According to the official statistics, there are approximately 5 million farms in the province, and majority of these farms are categorized as small farms (<2ha). The average size of the farms is 2ha (GOP, 2010). Yield variability between fields is large, especially in developing countries like Pakistan, where small farm holdings dominate the agricultural landscape. The average provincial total grain yield gap between potential and actual yield is approximately 4335 kg/ha (Khan et al., 2021). Closing the yield gap is possible by identifying yield constraint factors (Lobell et al., 2009b). Improvements in agricultural practices and land use management are fundamental to achieving high yield productivity and understanding yield constraints that cause yield gaps in crop productivity (Lobell et al., 2009b; Mohammed, 2019).

Punjab consists of a total of 36 Districts[2] with a total of 147 Tehsils[3]. In this study, based on the availability of field data, three districts, Sheikhupura- Gujrat, and Gujranwala of Punjab, were selected. The study area is situated between 33.03N to 31.36N Latitude and 73.59E to 74.69E Longitude. This region has a hot arid climate and is considered a warm and temperate climate zone. The total area of the study area is approximately 9827.01 Km$^2$ covers almost 4.79% of Punjab province. The altitude ranges from 153m to 416m above mean sea level. **Figure 3** shows the map of the study area in Punjab.

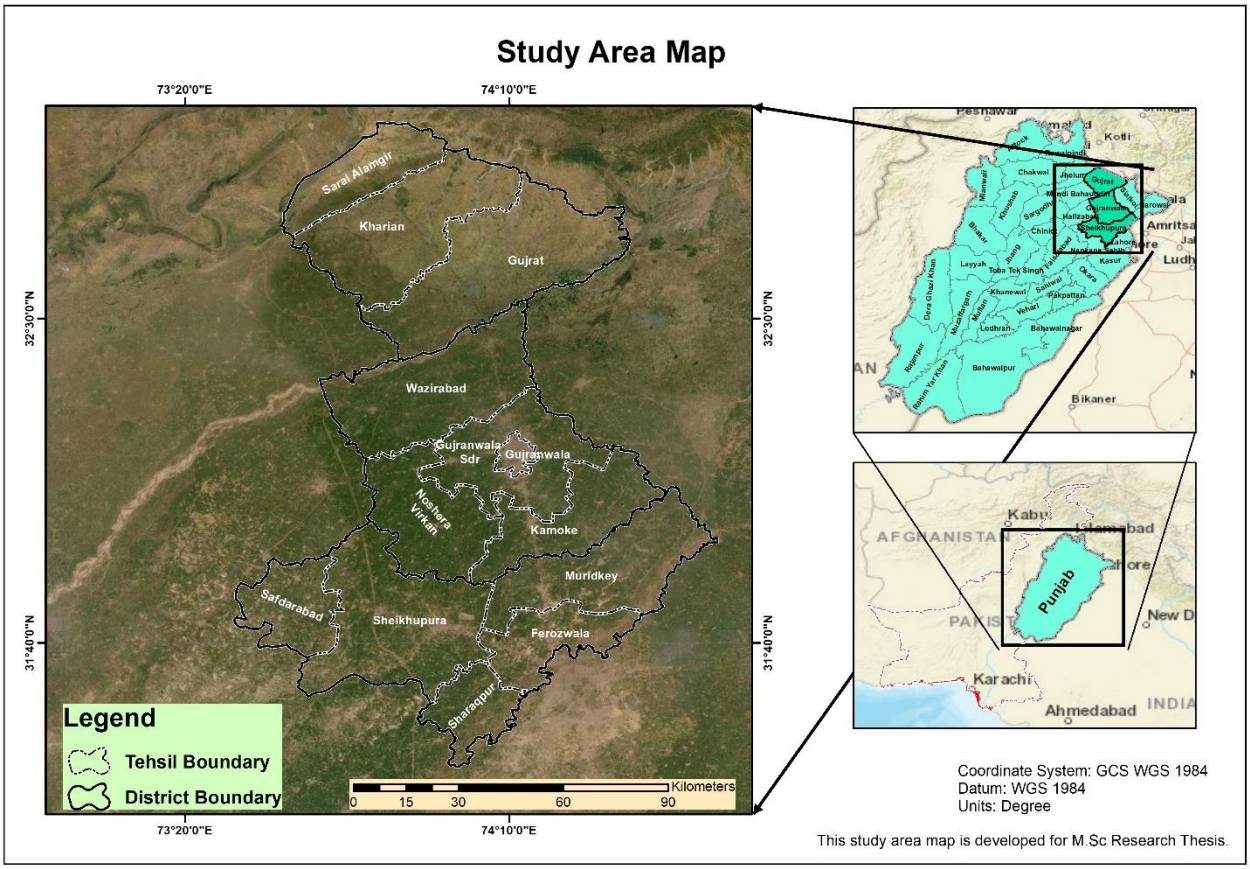



Figure 3: Study Area Map.

## 2.2. Datasets

In this study, data from various sources have been used to estimate crop yield variability at the field level. The data included: Site-specific wheat yield data 2019-2020, reported wheat production data 2019-2020, and hyper-temporal NDVI of Spot-ProbaV satellite 1999-2020.

### 2.2.1. Site-Specific Yield Data

In this study, site-specific yield data of 503 sites were obtained on request from Crop Reporting Services (CRS), Agriculture Department Punjab, Pakistan **Figure 4**. CRS is the official agency responsible for collecting agricultural statistics in Punjab (CRS, Punjab). CRS provided the site-specific yield data of wheat for three years (2017-18, 2018-19, 2019-20). The data was provided in SPSS tabular format. The tables contained the information on 46 different field parameters about genetics-G,

[2] District is the third-order administrative divisions of Pakistan, below provinces and divisions, but forming the first tier of local government (Wikipedia).
[3] Tehsil is an administrative sub-division of a District.

management-M, and land-L (G*M*L) **Table 1**. For this present study, wheat yield data for 2019-2020 was used. The reason for using only one season data was the availability of geo-coordinates of the sites. Geo-coordinates for the earlier two seasons were not available.

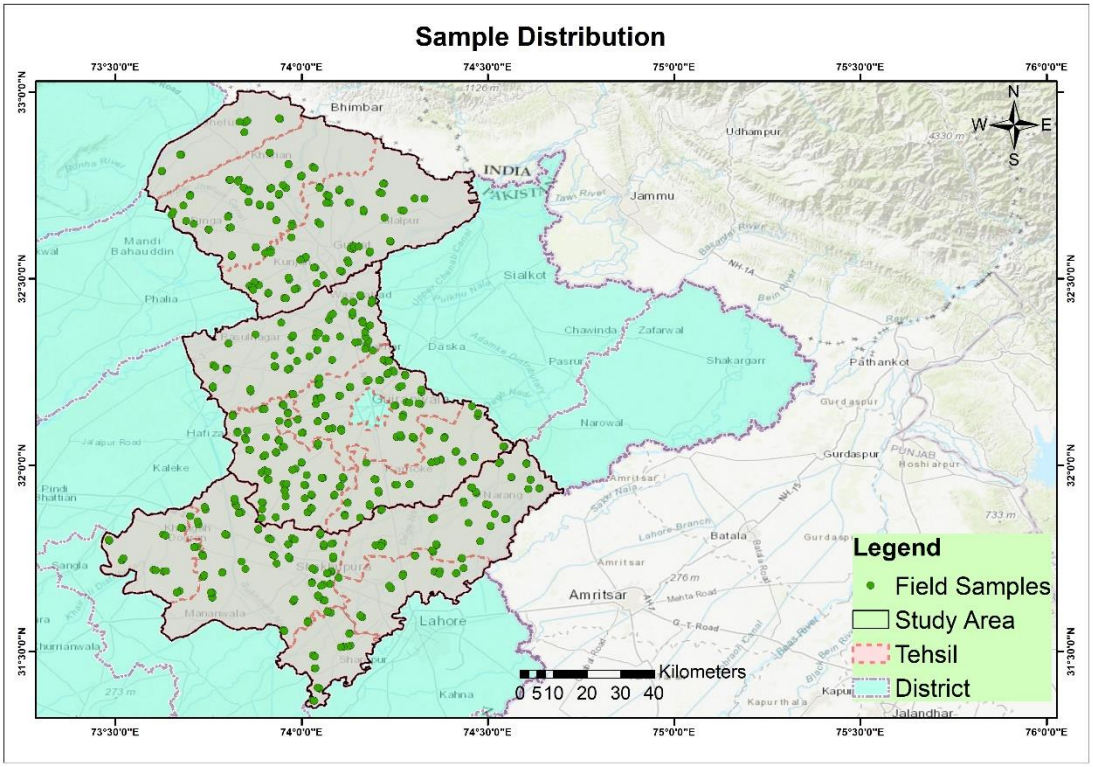



Figure 4: The distribution of surveyed sites in the study area.

Table 1: Field Parameters Collected by Agriculture Department.

| **Variables** | **Detail** |
|---|---|
| Location | District/ Tehsil/ Union Council/ Village/ Geo-coordinates |
| Farmer Detail | Name/ Phone/ Total Land |
| Yield | Achieved by Farmer |
| Wheat Variety | Various varieties (reported by farmer) |
| Planting/ Harvesting Time | Timing of Planting and Harvesting |
| Preparation/ Planting/ Harvesting Method | Different methods reported by farmers |
| Seed Type | Certified/ Uncertified |
| Seed Source | Domestic, Research Centre, Punjab Seed Corporation, Private Company |
| Seed Treatment | Treated/ Untreated |
| Application of Fertilizers | DAP- Urea- Farm-yard Manure |
| Water Management | Irrigated (Tube well or Canal)/ Unirrigated |
| Seed Quantity | Kg/ha |
| Sowing Pattern | Manual/ Machine (Broadcasting/ Line) |
| Soil Texture | Loam- Silt- Sandy |
| Pest/ Weed | Attack/ Infestation and Application |
| Previously land use | Last crop in the field |

### 2.2.2. Reported Wheat Production Data 2019-2020

Reported wheat production for the year 2019-2020 in kilogram (kg) per hectare was downloaded from CRS official website (http://www.crs.agripunjab.gov.pk/). The data contained the reported District average yield in kilograms per hectare. This data was used to develop a GIS map of reported crop yield tehsil wise and later, this map was used to compare with CPS zones.

### 2.2.3. SPOT- ProbaV NDVI

This study used the coarse SPOT- ProbaV NDVI series (Jan-1999 to June-2020). The SPOT- ProbaV sensor has a revisit time of one day (i.e., daily images) and spatial resolution of 1km. However, the NDVI product is a 10-day (dekad[4]) maximum value composite. SPOT-ProbaV NDVI was presented as DN values (0-255) that representing . The sensor's metadata provides data flags for omitted pixels (missing data, clouds, snow, water bodies, and backdrop), which aided in the data processing.

## 2.3. Software

This section reports the software used in this study, as shown in **Table 2**.

Table 2: List of Software used in this study.

| Software | Function |
|---|---|
| SPSS Statistics 27 | Descriptive Statistical Analysis (one by one relationship)<br>Identification of Significant variables through Stepwise Multiple Linear Regression. |
| MS Office Excel | Development of Crop Production Model |
| R Software | To prepare Violin, Box and whisker plots using libraries (ggplots2, dplyr, hrbrthemes, viridis, tidyverse). |
| ERDAS Imagine 2020 | To pre-processing, layer stacking, cleaning and unsupervised classification and produce CPS zones. |
| ENVI Classic 5.6 | To apply Savitzky- Golay filter for smoothening and produce percentiles 10 & 90. |
| ArcMap 10.8.1 | To prepare all maps used in this study. |
| Google Earth Pro | To analyse the study area vs. field parameters. |
| GDAL | To acquire SPOT- ProbaV 774 Images from the ITC archive. |

[4] Dekad is a period of 10 days.

# 3. METHOD

The method is outlined as improving crop production estimates by quantifying the site-specific field parameters to identify yield constraint factors and estimate yield gap; producing homogenous Crop Production System Zones (CPS zones) using long-term NDVI climatology of 20 years; assessing the relationship of site-specific measured yield with CPS zones vs. Tehsil wise stratification (administrative area), and integrating the above-identified field parameters and CPS zones into one model to assess their combined impact on crop yield variability **Figure 5**.

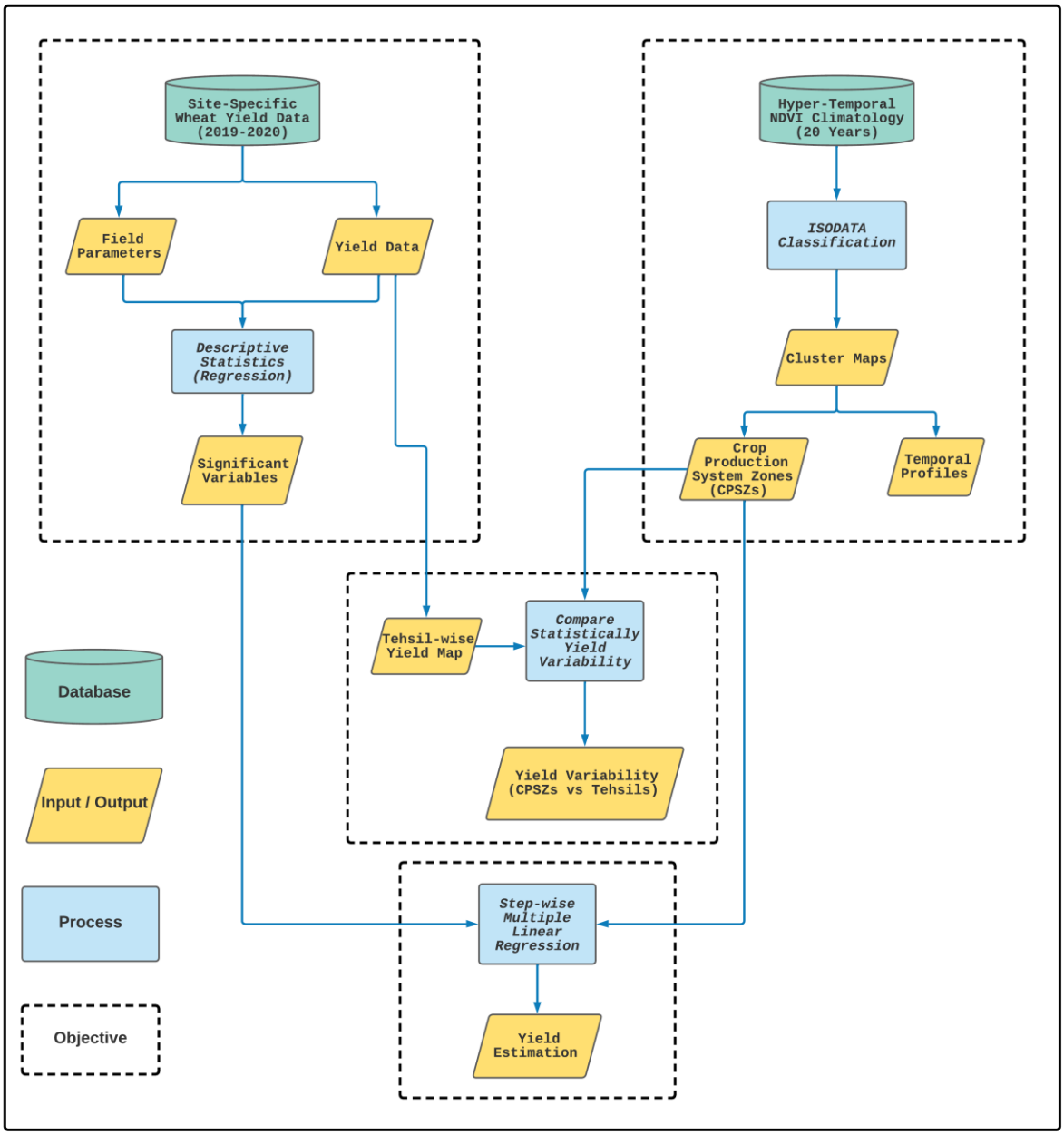



Figure 5: Flowchart of the research method.

## 3.1. Quantify Site-Specific Field Parameters (Comparative Performance Approach)

Site-specific field parameters collected during wheat cropping season 2019-2020 have consisted of various field parameters. Field parameters of genetics, land, management, and yield data were acquired through the field survey conducted by the CRS. Field parameters were categorized into three groups to analyse site-specific data, as shown in the **Table 3**. Each parameter was further divided into sub-parameters. In this study, all these field parameters were analysed individually to identify yield constraint parameters. After analysing individually, the model was developed using the specified significant parameters. This model was used to quantify their combined impact on wheat yield. Thus, the yield gap was also estimated. In order to identify and quantify, regression analysis was performed in SPSS software.

Table 3: Site-Specific Field Parameters Collected During Survey.

| Field Parameters | Category |
|---|---|
| Genetics Data | • Wheat Variety<br>• Seed Type (Certified/ Uncertified) |
| Management Data | • Seed Treatment (Treated/ Untreated)<br>• Field Preparation<br>• Seed Quantity<br>• Planting Time<br>• Sowing Method<br>• Fertilizer Application<br>• Pest Attack<br>• Weed Infestation<br>• Pest Application<br>• Weed Application<br>• Harvesting Time<br>• Harvesting Method<br>• Last Crop (Land use Pattern) |
| Land Data | • Soil Texture<br>• Water Management (Irrigated by Tube well or Canal/ Unirrigated) |

### 3.1.1. Identification of Significant Field Parameters Through Regression

At first, field parameters were analysed individually with the yield variable using a linear regression algorithm. Non-significant field parameters were eliminated during this process. The remaining parameters were combined to assess their impact on wheat yield. Stepwise multiple linear regression was applied to determine the most important parameters that influences the yield. Lastly, a production function and parameter statistics were derived and used to determine the mean and best values for each explanatory parameter, and the quantified impact by yield constraint and its contribution to the overall yield gap was estimated.

## 3.2. Estimation of Wheat Yield at Coarse Resolution (Producing Crop Production System Zones)

### 3.2.1. SPOT- ProbaV NDVI Pre-processing

NDVI climatology of SPOT- ProbaV is acquired from the ITC archive through GDAL. After downloading the 1km SPOT- ProbaV NDVI climatology (1999-2020), all the NDVI images were stacked using ERDAS Imagine software. After stacking, the temporal filtering method is used to clean

the data, and flagged pixels (DN values > 250) were replaced with zero values (i.e., "251 for missing (Bad radiometry), 252 for cloud or shadow,253 for sea, and 255 for background (missing input data)") (Gragn, 2021; Mohammed, 2019). 2nd temporal cleaning was then carried out through an iterative smoothing process by applying the Savitzky-Golay filter via the NRS tool in the ENVI Classic software (de Bie, 2020). The Savitzky-Golay (SAVGOL) filter uses a simplified least square procedure to fill gaps, and smooth data inconsistencies by suppressing disturbances and replacing each data value by a linear combination of nearby value in a time window (Beltran-Abaunza, 2009). Window size should be defined by the user, defining small window size can overfit the time-series while large size may over smooth. In this study, the upper- envelop filtering was carried out with window size 4, from both the left- hand and right- hand neighbours. After temporal filtering the clean stacked image is passed through layer stacking process again to stack each dekad of 20 years by excluding first nine 10-days images (Jan-March) to allow proper temporal coverage of 20 years. Median, 10th and 90th stack percentiles are retrieved for each dekad over all the years in order to speed up the classification runs by reducing the amount of data to processed. The tail of the distribution curve, where anomalies are located, is represented by the 10th and 90th percentiles (de Bie and Nelson, 2021; Oto, 2017). This resulted in 108 data layers by multiplying 36 dekads times 3. Finally, the cleaned image is classified in ERDAS Imagine through ISODATA unsupervised classification algorithm. Only pixels inside the research area have been processed. The image was cropped to the study area using a shapefile of three districts to ensure that the pixels belonged to the study region.

#### 3.2.2. NDVI Stratification

ISODATA algorithm is the most common, robust and well- understood statistical unsupervised classification approach and feasible when training data is not available for the study area (Abburu and Golla, 2015; Al-Ahmadi and Hames, 2009; Oto, 2017; Scarrott, 2022). In this study, ISODATA algorithm is used to stratify the fragmented agricultural landscape into homogenous zones (i.e., crop production system zones). Unsupervised categorization was used because the algorithm relies on minimizing user participation. 3 times ISODATA algorithm is carried out to produce 50-20 & 10 classes[5]. The number of iterations was set to 50 with a convergence criterion of 1.00 (Mohammed et al., 2020). The median value of each class was extracted and combined with the site-specific field data in SPSS software. The median value is least influenced than the mean.

#### 3.2.3. CPS zones vs. Site-Specific Field Parameters (Visualization)

CPS zones were produced using 20 years of NDVI imagery through an unsupervised classification algorithm that does not use field signatures to classify the satellite image. The land might be going through abrupt changes during this period. Therefore, before analysing the accuracy of CPS zones, a visual comparison was carried out between CPS zones and various important field parameters identified significantly through descriptive statistical analysis.

### 3.3. Assess the Relationship of Measured Crop Yield Between CPSZs vs. Tehsil Stratification (Administrative Areas)

#### 3.3.1. Box and whisker Plot

After producing the crop production system zones (CPS zones) based on 20 years of NDVI climatology, the next step was to explain the relationship of both approaches, i) CPS zones and ii) Admin areas with site-specific measured yield from the field data. To do this, a box and whisker plot

[5] In this study, the term class, cluster, zone, and stratum are used interchangeably to refer to crop production system zones (agro-ecological zones). In contrast, the term admin areas, tehsil wise are used to refer administrative boundaries.

was used to plot yield variation between CPS zones and Admin areas. This method helps understand the data flow so that variation can be seen between different scenarios. The primary purpose was to assess how measured yield varies in CPS zones compared to yield variation in Admin areas.

#### 3.3.2. Regression of Measured Yield variation between CPS zones vs. Admin area

Considering the results derived in the previous step that CPS zones may provide better results over the Tehsil stratification approach (admin areas), a multiple linear regression was carried out in the 2nd stage of analysis. The results of this analysis will aid in understanding the yield variation in CPS zones and Admin areas. The regression model for both approaches was carried out on the SPSS software environment.

##### 3.3.2.1. Multiple Regression Analysis of CPS Zones

A total of 10 zones were produced through NDVI stratification. Data distribution of site-specific samples showed that a total sample of 503 existed in 9 zones. **Table 4** shows the zones with no. of samples exist. Zone-2 & 3 were considered constant (reference) for the regression equation to perform the multiple regression analysis. Each site-specific field point's cluster value has been combined with the field data file. The site-specific yield from the yield survey data connected to the NDVI clusters is then estimated at 1km resolution using multiple regression. This method was developed because agricultural fields have distinct temporal NDVI profiles (i.e., crop phenology cycles) compared to other land cover types. The regression model in SPSS was used can be expressed by:

$$Y = B_1C_1 + B_2C_2 + B_3C_3 + \ldots\ldots + B_nC_n$$

Where Y is the crop yield (kg/ha) from yield survey data, B1, B2, B3…. Bn is the coefficients, and C1, C2, C3….. Cn is the NDVI clusters within Tehsils. The model was forced through origin because not every pixel contains agricultural fields. The regression model was created using the SPSS software package. Based on the occurrences of the distinct NDVI clusters within each district, the regression model distributed the predicted wheat yield (kg/ha) over the study region at a 1km resolution.

Table 4: Site-Specific Data Distribution in CPS Zones.

| Zone | No. of Samples |
|---|---|
| Zone-1 | 0 |
| Zone-2 | 2 |
| Zone-3 | 27 |
| Zone-4 | 62 |
| Zone-5 | 39 |
| Zone-6 | 39 |
| Zone-7 | 28 |
| Zone-8 | 80 |
| Zone-9 | 84 |
| Zone-10 | 142 |

#### 3.3.2.2. Multiple Regression Analysis of Admin areas

The study area consisted of 12 Tehsil in 3 Districts. **Table 5** shows the admin areas with no. of samples collected from each Tehsil. Tehsil Sarai Alamgir was used as a reference to develop a regression equation. The regression model in SPSS was used can be expressed by:

$$Y = B_1X_1 + B_2X_2 + B_3X_3 + \ldots\ldots + B_nX_n$$

Where Y is the crop yield (kg/ha) from yield survey data, $B_1$, $B_2$, $B_3$…. Bn is the coefficients, and $X_1$, $X_2$, $X_3$….. Xn is the Tehsils within Districts in the study area.

Table 5: Site-Specific Data Distribution in Admin Areas.

| Tehsil | District | No. of Samples |
|---|---|---|
| Gujrat | Gujrat | 65 |
| Kharian | | 49 |
| Sarai Alamgir | | 10 |
| Gujranwala Saddar | Gujranwala | 60 |
| Nowshera Virkan | | 66 |
| Wazirabad | | 64 |
| Kamoke | | 34 |
| Sheikhupura | Sheikhupura | 66 |
| Muridkey | | 41 |
| Ferozwala | | 15 |
| Sharaqpur | | 16 |
| Safderabad | | 17 |

### 3.4. Merge All Studied Parameters to Quantify Combined Impact on Yield Variability

After deriving important field-specific site parameters, producing crop production system zones, and assessing the relationship of CPS zones and Admin areas with site-specific measured yield, the last step of this study was to integrate all those parameters into one model and to find out at what extent these all-combined parameters explain yield variability. Field parameters identified through one on one descriptive statistics explained the importance of various parameters of management & land and their impact on the overall productivity of crops during the cropping season. CPS zones created using NDVI climatology of 20 years explained the study area's long-term agricultural activities. In order to evaluate the combined impact on yield variability all the studied parameters along with CPSZs and Tehsil wise area frame sampling approach were merged together in stepwise multiple linear regression.

### 3.4.1. Stepwise Linear Regression Analysis

To do this process, outputs of sub-objective (i) and (ii) were merged to form a final model to compare and quantify the performance of EO based approach on yield variability with the existing sampling survey approach. **Table 6** shows the variables combined in the model. Stepwise regression was applied and carried out in SPSS software. Stepwise regression is an automated process that helps determine which factors/ parameters are essential and remove uncorrelated or redundant data. This process helps in making the model robust.

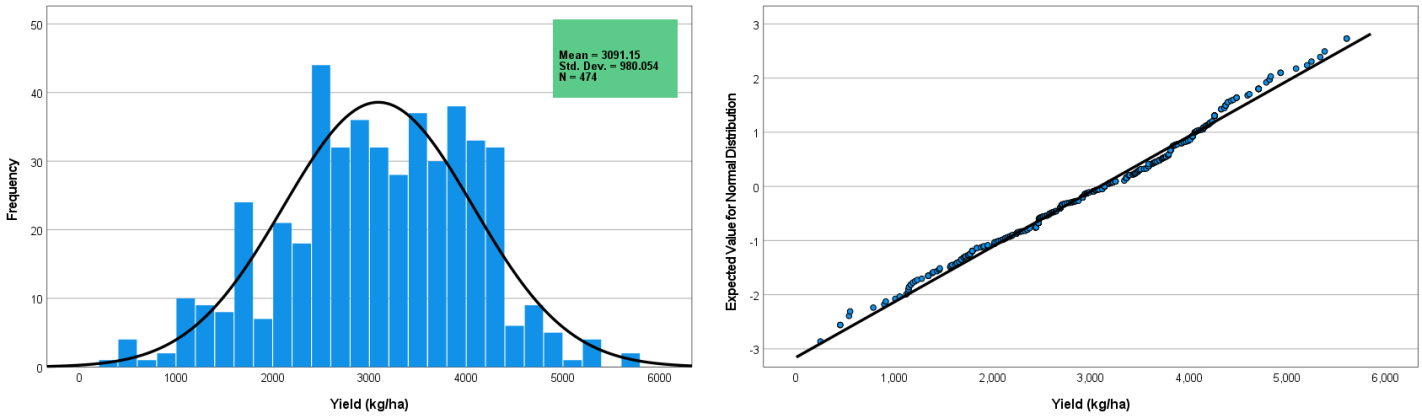
Table 6: Final Model of Integrating EO with Site-Specific Field Parameters.

| **Model (EO Based)** |
|---|
| Urea |
| Irrigated by Tube well |
| Faisalabad Treated Seed |
| Sowing pattern: Machine broadcasting |
| Spray Pesticides |
| Attack by Pest |
| CPS zone 4 |
| CPS zone 5 |
| CPS zone 6 |
| CPS zone 7 |
| CPS zone 8 |
| CPS zone 9 |
| CPS zone 10 |

# 4. RESULTS

## 4.1. Quantifying Site-Specific Field Parameters Using Statistical Analysis (Comparative Performance Analysis)

**Figure 6a** shows the distribution of wheat field data collected during the cropping season of 2019-2020. The distribution curve shows the normal distribution of field data. To confirm the normality, Kolmogorov- Smirnov test was applied, and the result showed that data is completely significant and normally distributed. The z- score plot in **Figure 6b** shows how well the data was distributed along the reference line. The closer the points are to the reference line follows a normal distribution.




a b

Figure 6: Data Distribution and Z-Score Plot.

### 4.1.1. Genetics Data

#### 4.1.1.1. Wheat Variety

Various varieties were used by farmers in the study area. Farmers chose these varieties based on location, climate, land type, and season suggested by the agriculture departments. In this study, eight varieties were found grow by farmers. In order to keep the data distribution normal for analysis, all those varieties with a frequency of fewer than ten reports were merged into other categories. After combining the varieties, a total of 5 classes were analysed statistically. **Figure 7** shows the variation of yield by different wheat varieties grown. The majority of the farmers 405x reported the use of the Faisalabad variety. 50x grown Galaxy variety and achieved higher yield as compared to all other varieties, 11x reported using Sehar variety, 14x used Inqlab-91, and the remaining 23x used other varieties. The impact of various varieties on yield variability could be quantified;

*Yield (kg/ ha) = 3043 – 469 × (if variety used; Inqlab-91) + 626 × (if variety used; Galaxy)** + 194 × (if variety used; Sehar) – 77 × (if wheat variety used; Faisalabad)*
[n= 498; Adj- $R^2$ = 4.3%; * sign. At 10% ** sign. at 5%; *** sign. at 1%]


Figure 7: Variation of Yield by Wheat Varieties Grown.

#### 4.1.1.2. Seed Type

The detail about seed type (certified or un-certified) was collected during the field survey. Often farmers purchase certified seeds and then save them for two or more consecutive years before being purchased again. This increases the risks for growers; although the cost of certified seeds is high, it always pays for itself through increases in yield compared to the other seed saved by farmers. In the data collected from the farmers, 38x reported using certified seeds, and the remaining 465x used un-certified seeds. **Table 7** contains the information about seed type w.r.t various wheat varieties farmers grown. Wheat varieties were categorized into certified and uncertified. Based on the data provided by the farmers, all those varieties with less than 10 counts were merged into others respectively, others certified and others uncertified; the remaining varieties were categorized accordingly Error! Reference s ource not found.. Certified seeds provided better yield as compared to uncertified seeds, except one variety Galaxy uncertified which performed relatively better than remaining uncertified seeds. The reason of outperformed other certified varieties could be that Galaxy variety is high yielding, high tolerant variety suitable for irrigated areas except rice zones developed in 2013 and either farmer purchased it directly from agriculture department or utilised the seed save from last crop (J et al., 2019) The summary of seed type could be quantified;

*Yield (kg/ha) = 2906 + 298 × (if Faisalabad certified used) - 296 × (if Inqlab-91 uncertified used) + 776 × (if Galaxy uncertified used)*** + 296 × (if Sehar uncertified used) + 40 × (if Faisalabad uncertified used) + 516 × (if others certified used)*

[n= 496; Adj- $R^2$ = 4.3%; * sign. At 10% ** sign. at 5%; *** sign. at 1%]

Table 7. Crosstabulation of Seed Type with Wheat Varieties.

| Wheat Variety | Seed Type | | Total |
|---|---|---|---|
| | Certified | Uncertified | |
| Inqlab-91 | 2 | 12 | 14 |
| Galaxy | 4 | 46 | 50 |
| Sehar | 1 | 10 | 11 |
| Lasani | 0 | 1 | 1 |
| Waten | 1 | 2 | 3 |
| Faisalabad | 28 | 377 | 405 |
| Others | 2 | 17 | 19 |
| Total | 38 | 465 | 503 |

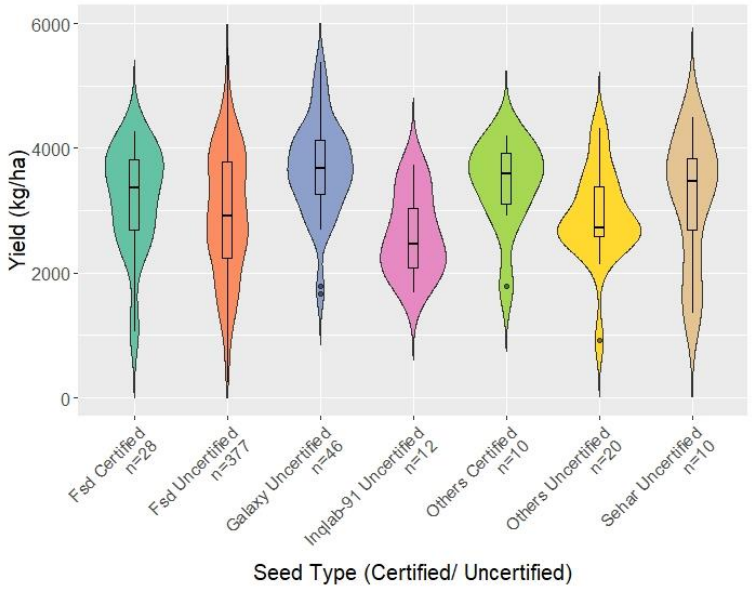



Figure 8: Variation of Yield by Seed Type.

### 4.1.2. Management Data

#### 4.1.2.1. Seed Treatment

Seed treatment is a process in which seeds are treated with physical, chemical (fungicides or pesticides), and biological agents to protect them from the diseases caused by seed, soil, and insects. Data collected from farmers about seed treatment were crossed tabulated with wheat varieties. **Table 8** shows the detail of wheat varieties and seed treatment. Based on the data provided by the farmers, all those varieties with less than ten counts were merged into others, respectively treated and untreated, and the remaining varieties were categorized accordingly. Error! Reference source not found. shows that t reated seeds are associated with higher yields.

The impact of treated seeds on yield could be quantified;

*Yield (kg/ ha) = 2790 + 983 × (if Faisalabad Treated)*** + 879 × (if Galaxy Untreated)*** + 483 × (if Others Treated)* + 18 × (if Faisalabad Untreated)*

[n= 498; Adj- $R^2$ = 14.5%; * sign. At 10% ** sign. at 5%; *** sign. at 1%]

Table 8. Seed Treatment vs. Wheat Varieties.

| Wheat Variety | Seed Treatment | | Total |
|---|---|---|---|
| | Yes | No | |
| Inqlab-91 | 4 | 10 | 14 |
| Galaxy | 0 | 52 | 52 |
| Sehar | 9 | 6 | 15 |
| Faisalabad Certified | 12 | 22 | 34 |
| Faisalabad Uncertified | 61 | 345 | 406 |
| Others | 4 | 22 | 26 |
| Total | 45 | 502 | 547 |

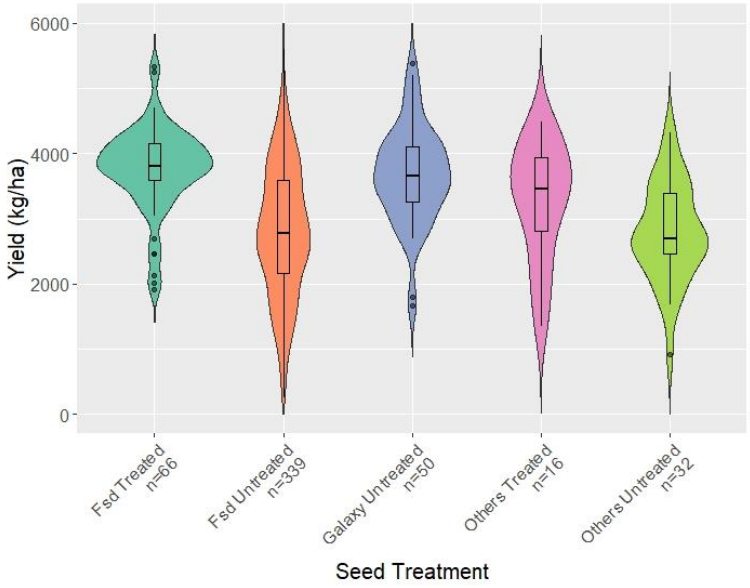



Figure 9: Variation of Yield by Seed Process (Treated vs. Untreated).

#### 4.1.2.2. Field Preparation

Field preparation before starting of new cropping season can be done with different mechanical and manual equipment. According to data received from farmers, land preparation was done using three distinct machines **Figure 10**. 417x reported use of disc plough **Figure 10a**, 52x reported using rotavator for land preparation **Figure 10b**, while the remaining 34x farmers reported using chisel plough **Figure 10c**.

a b c

Figure 10: Different Methods of Field Preparation.

**Figure 11** shows the variation of yield (kg/ha) in these three field preparation methods reported by farmers. The summary could be quantified;

*Yield (kg/ha) = 2995 + 479 × (if field is prepared; by Rotavator)*** -154 × (If field is prepared; by Chisel Plough)*
[n= 500; Adj- $R^2$ = 2.0%; * sign. At 10% ** sign. at 5%; *** sign. at 1%]

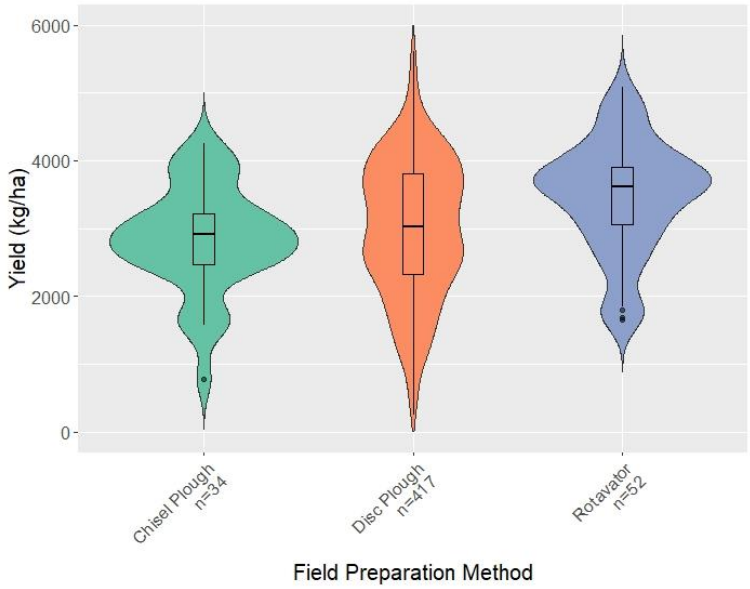



Figure 11: Land Preparation Method vs. Yield.

#### 4.1.2.3. Seed Quantity (kg/ha)

The proposed average wheat seed quantity by the agriculture department is 120kg/ha (Imran, 2019). The research program of PARC[6] suggested 120kg/ha for normal sowing and 150kg/ha for late sowing. Figure 12 shows details of farmers' reported seed quantities during the cropping season 2019-2020. 250x used 120 kg/ha, 128x used 110 kg/ha and 125x used 99kg/ha. The linear curve indicates that the quantity of seed used had a slightly positive impact on yield **Figure 12**. One kg of extra seed resulted in 17 kg of additional yield.
The impact of seed quantity could be quantified;

*Yield (kg/ha) = 1057 + 17 × (seed quantity used; kg/ha)****
[n= 501; Adj- $R^2$ = 3.1%; * sign. At 10% ** sign. at 5%; *** sign. at 1%]

[6] Pakistan Agricultural Research Council (PARC) is the apex national organization working in close collaboration with other federal and provincial institutions in the country to provide science-based solutions to agriculture of Pakistan through its statutory functions. (http://www.parc.gov.pk)

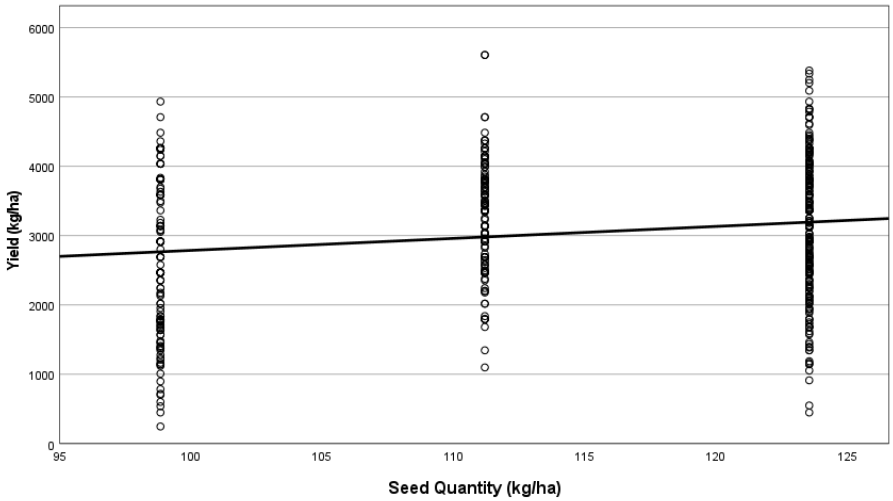



Figure 12: Seed Quantity (kg/ha) vs. Yield.

#### 4.1.2.4. Planting Time (day of year)

Provided sowing dates in the field data were already generalized into two weeks. The mean date is picked and converted into the day of the year for analysis in this study. The proposed time for planting wheat in irrigated areas starts from 1st November- 10th December (between 305- 344 day of the year), while for rainfed areas, the dates start around 20th October and goes up to 20th November (between 293- 324 day of the year). **Figure 13** shows that most of the farmers planted their crops during the 2nd part of November (between 321- 330), and this period was indeed the optimum planting period (median day of year 327). Data collected from farmers shows that 350x planted crops in the fourth week of November (± day nr. 327), and 109x planted around the second week of November (± day nr. 312), 7x planted early in the last week of October (± day nr. 297) and 37x farmers planted later in December (between day nr. 340 and 360). The selection of planting time depends on the temperature and post-harvest land condition due to the last crop grown on the land. The impact of sowing time could be quantified;

*Yield (kg/ha) = 2479 + [ 28 × (sowing; day after DOY 295)* – 0.298 × (sowing; day after DOY 295)2*]*
[n= 500; Adj- $R^2$ = 1.2%; * sign. At 10% ** sign. at 5%; *** sign. at 1%]

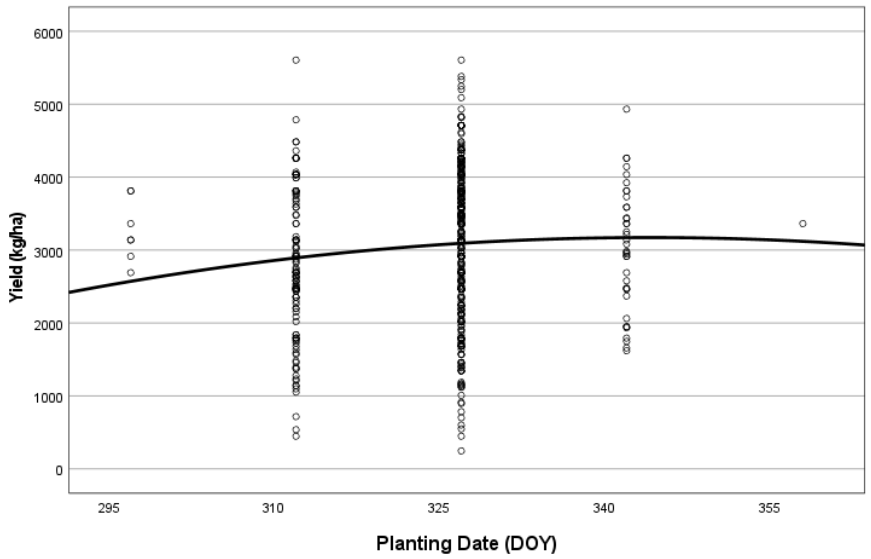



Figure 13: Planting date (DOY) vs. Yield.

#### 4.1.2.5. Sowing Method

Several sowing patterns exist like broadcasting, sowing in furrows (line), drilling, zero-till seed drill, furrow irrigated raised bed, etc. Data showed that 336x were planted manually, 131x were planted in a broadcasting pattern using a machine, and 36x were planted in a line pattern using a machine **Figure 14**. The yield associated with the machine line method is lower than the other method. **Figure 15** shows the spatial distribution of the reported sowing methods used by farmers. The relationship between sowing pattern and yield could be quantified;

Yield (kg/ha) = 2902 + 739 × (if sowing pattern: Machine Broadcasting)*** - 843 × (if sowing pattern: Machine Line)***

[n= 500; Adj- $R^2$ = 17.2%; * sign. At 10%; ** sign. at 5%; *** sign. at 1%]

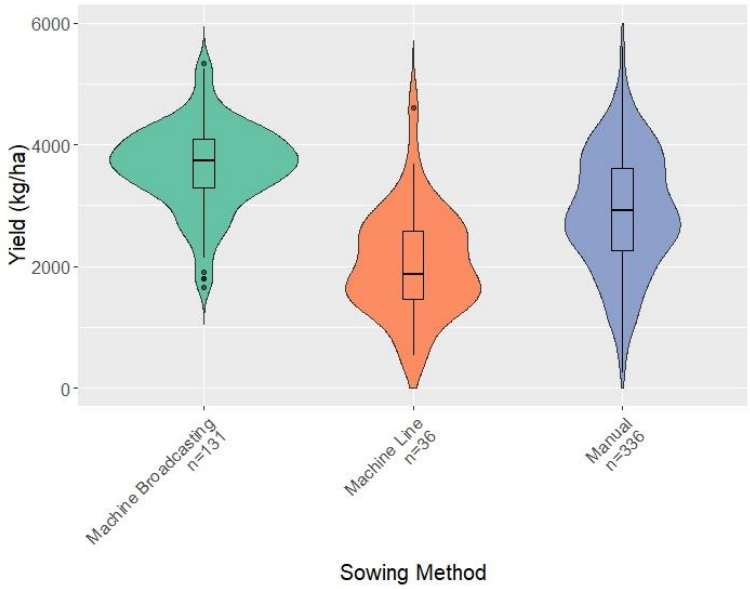



Figure 14. Sowing Method vs. Yield.

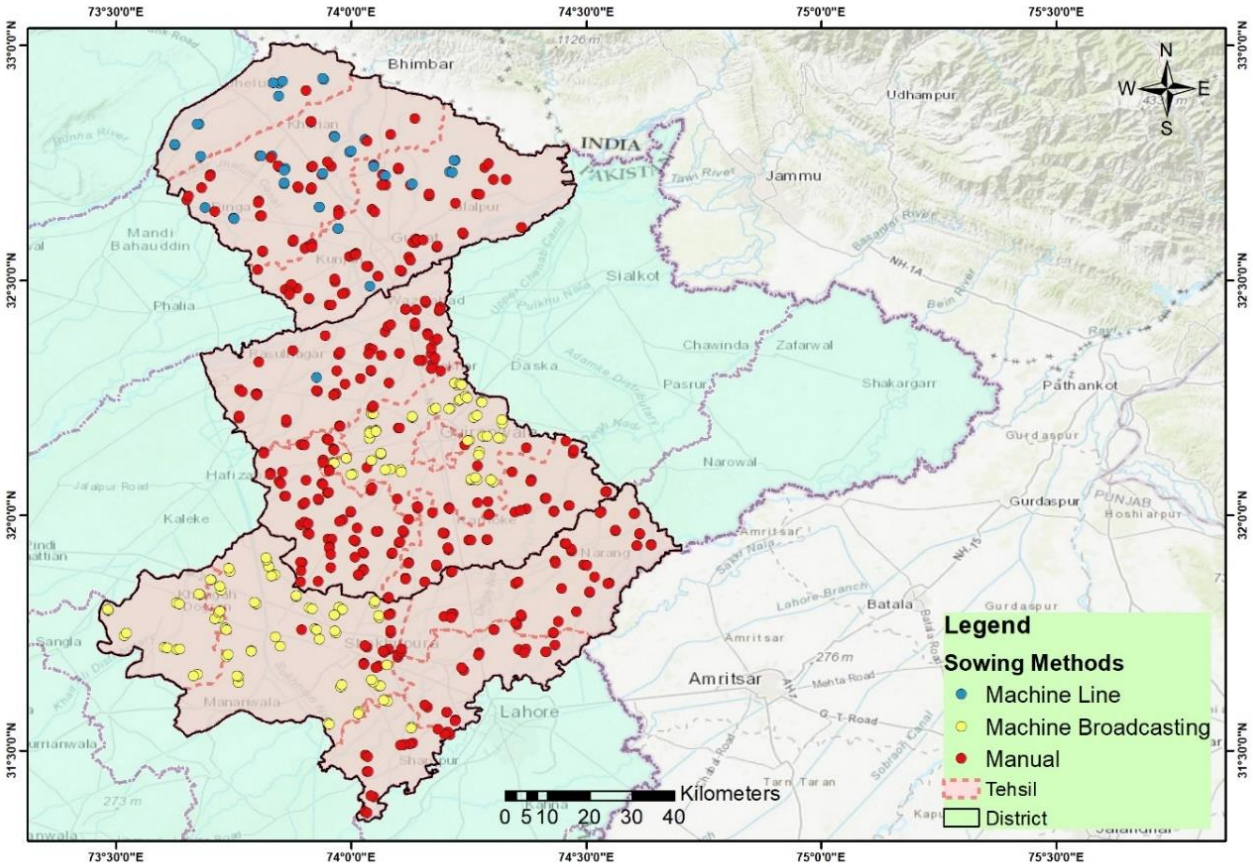



Figure 15: Spatial Distribution of Sowing Method Samples.

#### 4.1.2.6. Fertilizer Application

In Pakistan farmers predominately use Urea (*46:0:0*) and Di-ammonium Phosphate (DAP) (*18:46:0*). Fertilizers are used to add nutrition or change the properties of the soil. The three primary nutrients most commonly used by farmers are N-P-K ($N$-$P_2O_5$-$K_2O$). These three letters refer to the ratio of these nutrients in a bag of different fertilizers like N and DAP. In Pakistan, DAP with *18:46:0*, which is the NPK content in Di-Ammonium Phosphate, this ratio explains 18% N, 46% $P_2O_5$, and 0% $K_2O$. Similarly, for UREA, the ratio is 46:0:0. Data about fertilizers, including Urea, DAP, and Farm-yard manure (FYM), were collected, and analysed in this study.

##### 4.1.2.6.1. Urea (kg/ha)

Urea (*46:0:0*) is the most common use by farmers. The majority of the farmers' data collected during the field survey applied between 100-250 kg/ha. 209x farmers applied 240 kg/ha, 103x applied 185 kg/ha, 159x applied 124 kg/ha. 10x applied more than 300 kg/ha, 12x applied less than 100 kg/ha, and 10x did not apply Urea. Application of Urea significantly improves in wheat productivity. The quadratic curve indicates that increase in the amount of Urea kg/ha at a certain limit, increases productivity **Figure 16**. The relationship between Urea quantity and yield could be quantified;

$$Yield\ (kg/ha) = 690 + [20 \times (Urea;\ kg/ha)^{***} - 0.035 \times (Urea;\ kg/ha)2^{***}]$$

[n=500; Adj- $R^2$ = 28.1%; * sign. At 10%; ** sign. at 5%; *** sign. at 1%]

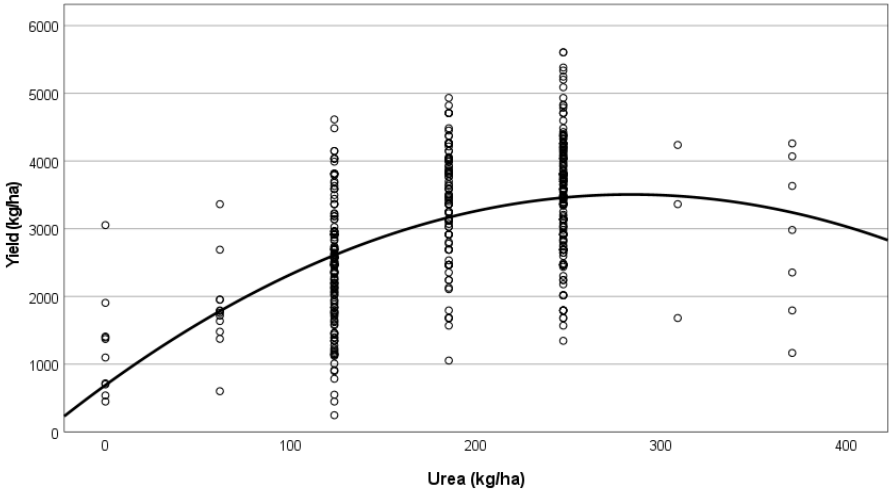



Figure 16: Urea Fertilizer vs. Yield.

#### 4.1.2.6.2. DAP (kg/ha)

Farmers apply fertilizer usually twice during the cropping season, one at the time of plantation while the second time with irrigation. 428x farmers applied 120 kg/ha, 24x applied 185 kg/ha, 7x applied more than 200kg/ha, 18x applied 61 kg/ha, and 26x farmers did not apply DAP. Relating through regression, the quadratic curvet indicates that adding DAP in the field increases the yield **Figure 17**. The impact of DAP on yield could be quantified;

*Yield (kg/ha) = 1708 + [15 × (DAP; kg/ha)*** – 0.033 × (DAP; kg/ha)2***]*
[n= 500; Adj- $R^2$ = 11.1%; * sign. At 10% ** sign. at 5%; *** sign. at 1%]

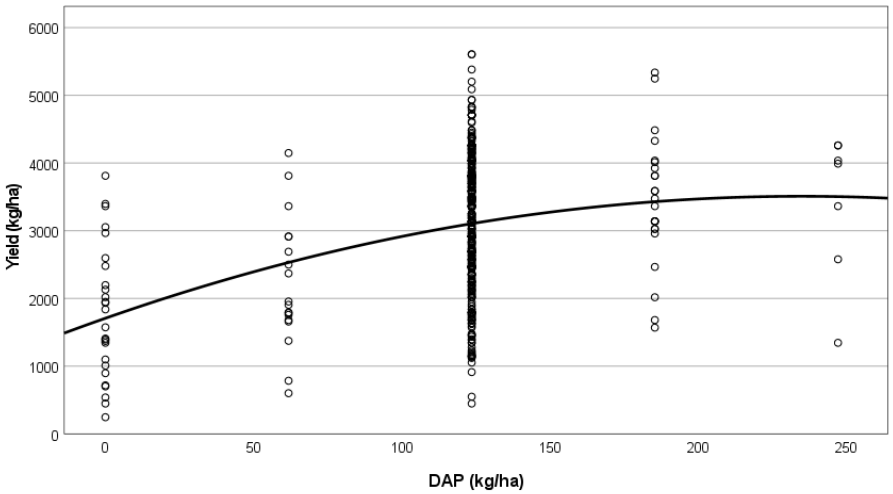



Figure 17: DAP Fertilizer vs. Yield.

#### 4.1.2.6.3. Farm-yard Manure (Ton/ha)

Farmers applied farm-yard manure up to 1-2 ton/ha during the land preparation to increase the soil nutrition and improve the soil biodiversity. In the available field data majority of farmers did not apply farmyard manure. Summary shows that 497x did not apply FYM, only 6x applied and result was not that much satisfactory. Therefore, no further analysis was required to perform.

#### 4.1.2.7. Weeds and Pests Attack

In agriculture, weed & pest attacks produce significant losses in crop yield with the increase of agricultural inputs such as new seed varieties, irrigation, pesticides, and fertilizers. Therefore, it is important to measure the effect of weeds on yield loss. Studies show that yield loss may go up to 20 to 30% due to weed (Oad et al., 2007). In this study, we only have data about weed and pest attacks on the crop but do not have details about the type of weeds and pests and how many times farmers weeded their crops. Field data shows that 347x farmers reported weed infestation on their crops while 156x farmers reported no infestation Figure 18. Similarly, 203x reported pest attacks, and 300x reported no pest attacks Figure 19. Farmers reported no pest attack, and no infestation of weeds achieved better yield compared to those who faced attacks. The impact of both pests attack and weed infestation could be quantified;

*Yield (kg/ha) = 3227 – 280 × (if weed infested: yes/no)****
[n= 501; Adj- $R^2$ = 1.5%; * sign. At 10% ** sign. at 5%; *** sign. at 1%]

*Yield (kg/ha) = 3167 – 331 × (if pest attacked: yes/no)****
[n= 501; Adj- $R^2$ = 2.4%; * sign. At 10% ** sign. at 5%; *** sign. at 1%]

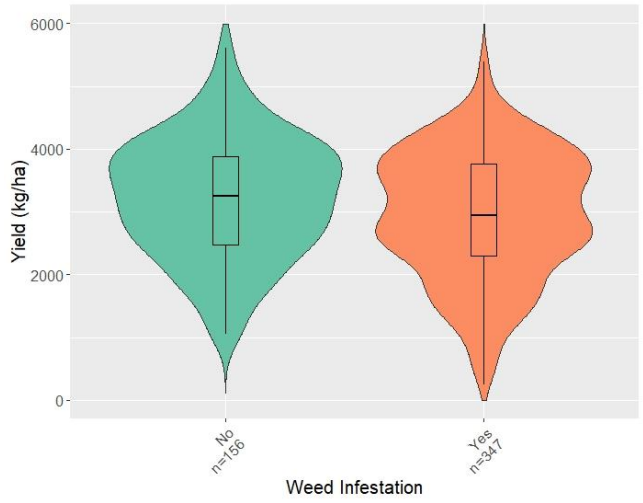



Figure 18: Weed Infestation vs. Yield.

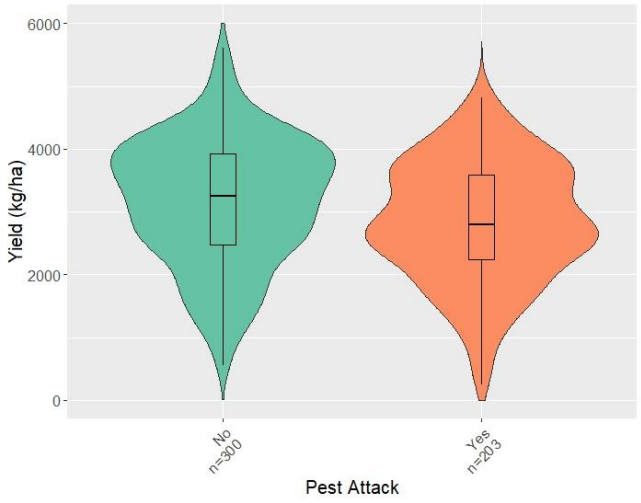



Figure 19: Pest Attack vs. Yield.

#### 4.1.2.8. Pest Application

Farmers reported using different pesticides to prevent their crops from damaging during the crop growing season. Data regarding pest application was collected during the field survey. 295x farmers did not apply any pesticide during the crop season, while 208x farmers used pesticides. Farmers who applied pesticides to their crops to prevent from pest attacks were seen to achieve better yields **Figure 20**. The impact of pest application on yield could be quantified;

*Yield (kg/ha) = 2805 + 554 × (if pest spray applied)****
[n= 501; Adj- $R^2$ = 7.3%; * sign. At 10% ** sign. at 5%; *** sign. at 1%]

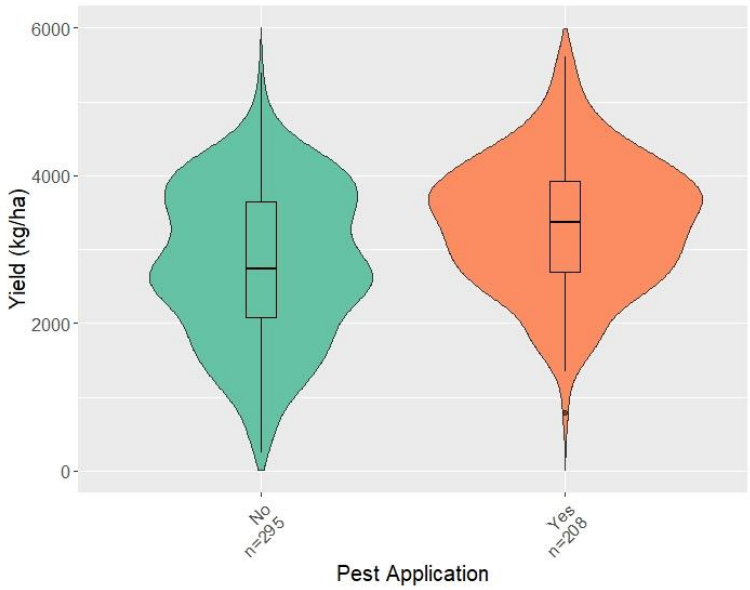



Figure 20: Pest Application vs. Yield.

#### 4.1.2.9. Weed Application

Farmers reported using different weedicides to prevent their crops from being damaged during the crop growing season. Data about applied weedicides was collected during the field survey. 78x farmers did not apply any spray, and 425x applied the weedicide sprays on their crops. Farmers that used weedicides achieved higher yields compared to others **Figure 21**. The impact of weed application on yield could be quantified;

*Yield (kg/ ha) = 2365 + 791 × (if weed spray applied)****
[n= 501; Adj- $R^2$ = 8.0%; * sign. At 10% ** sign. at 5%; *** sign. at 1%]

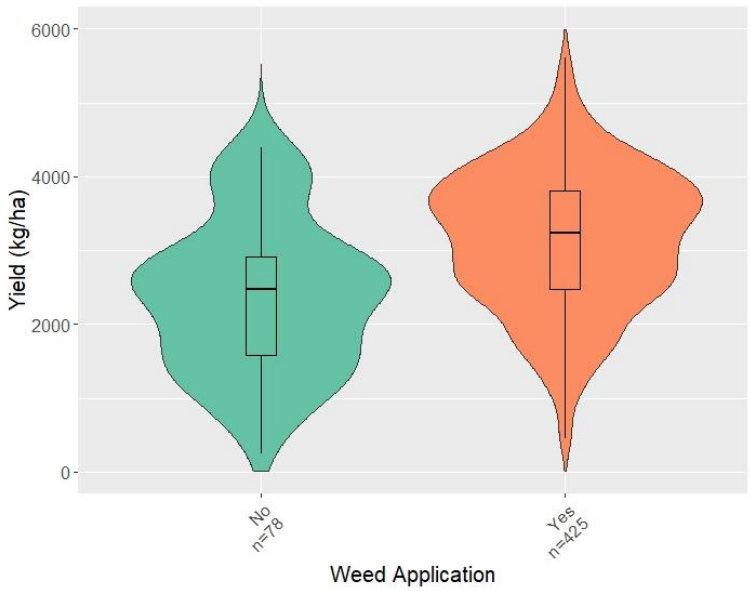



Figure 21: Weed Application vs. Yield.

#### 4.1.2.10. Harvesting Time (day of year)

Harvesting took place around the first week of April and goes until the end of May. The average length of the growing cycle was 140-160 days which starts from October and ends in April-May. 59x farmers harvested wheat in the second week of April (day nr. 106), 440x harvested around the end of the fourth week of April (day nr. 116), while only 4x waited long before harvesting their crops and achieved lower yield. The length of the growing period had a significant (positive) effect on yield, but this may be more related to the timing of planting than the timing of harvesting. The quadratic curve indicates that harvesting took place around day nr. 116 achieved higher yield, and delay in harvesting once crop reach to maturity level can cause a reduction in the productivity **Figure 22**. This could be quantified:

*Yield (kg/ha) = 2013 + [133 × (harvesting date; days after DOY 105)*** – 4.11 × (harvesting date; day after DOY 105)2**]*
[n= 500; Adj- $R^2$ = 2.2%; * sign. At 10% ** sign. at 5%; *** sign. at 1%]

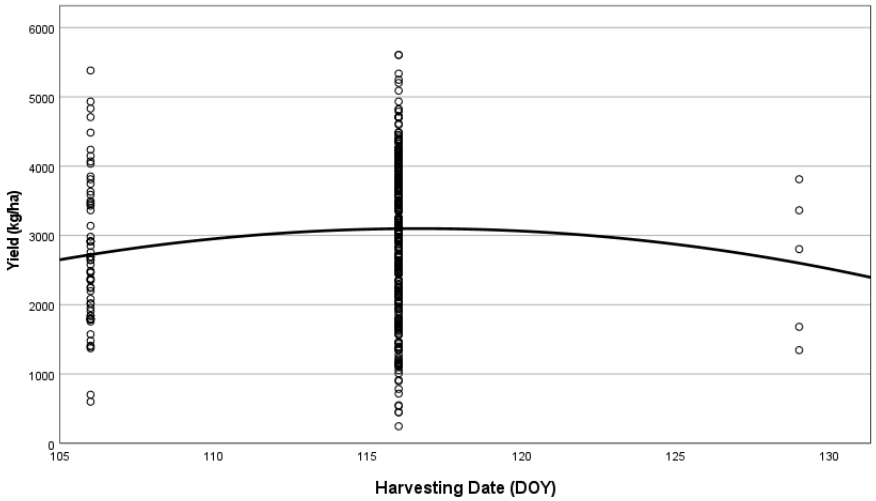



Figure 22: Harvesting Time (DOY) vs. Yield.

#### 4.1.2.11. Harvesting Method

Farmers were also asked about their crop harvesting methods during the field survey. **Figure 23** depicts the most prevalent harvesting method in Pakistan. Farmers described their wheat harvesting methods. Harvesting by hand was recorded by 110 people, while harvesting by combine harvester was reported by 436 people. Shattering wheat grains during harvest not only reduces overproduction, but also cause additional expense for picking residuals (Payne, 2002). The collected data has been included in the study due to the importance of this phase for yield. **Figure 24** shows that those farmers which used manual method for harvesting had lower yields as compared to the those who harvested using combine harvester. The relationship could be quantified:

*Yield (kg/ha) = 2422 + 779 × (if harvested by Combine Harvester)****
[n= 501; Adj- $R^2$ = 10.00%; * sign. At 10% ** sign. at 5%; *** sign. at 1%]

a: Manual Harvesting.


b: Mechanical Harvesting.

Figure 23: Harvesting method of Crop.

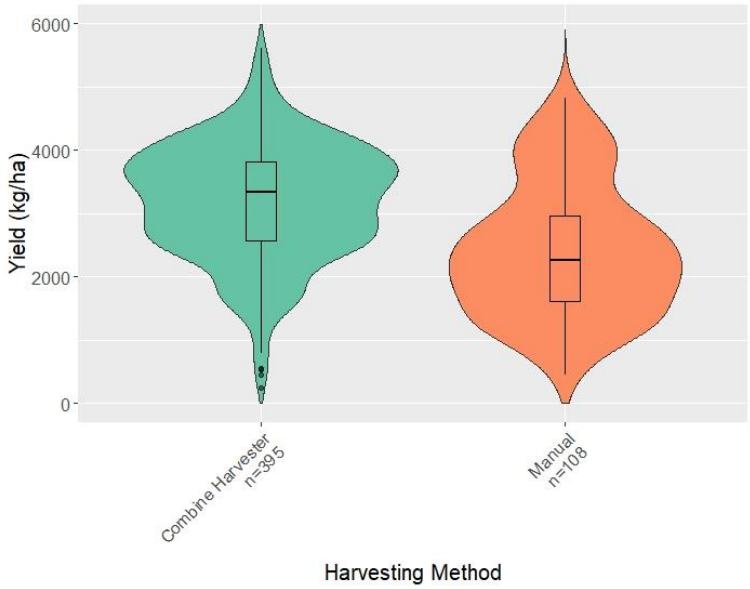



Figure 24: Harvesting Method vs. Yield.

#### 4.1.2.12. Last Crop (Crop Patten)

In Pakistan, two cropping seasons exist; the Rabi and Kharif. Rabi season starts from October/November and ends around March/April, while Kharif season is from June/July to October. The fertility of land also depends upon how the land is being used. Data about the earlier use of land was also collected from the farmers. 453x previously sown rice in their field before sowing wheat, 37x sown fodder in their fields, while 56x left their unplanted. **Figure 25** shows the wheat yield variation in the different state of the field before wheat plantation. The impact of last crop could be quantified;

*Yield (kg/ ha) = 2206 + 999 × (if previously land used for rice)*** + 145 × (if previously land was fodder)*
[n= 500; Adj- $R^2$ = 13.0%; * sign. At 10% ** sign. at 5%; *** sign. at 1%]

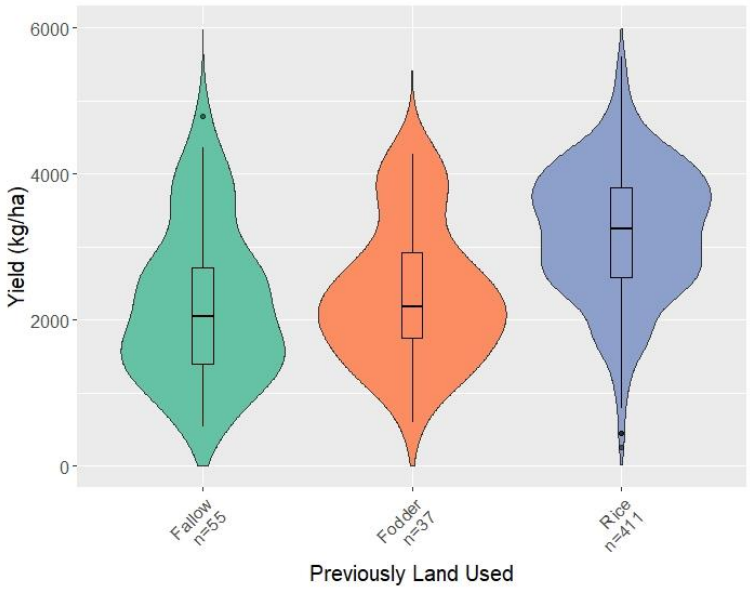



Figure 25: Previously land used vs. Yield.

### 4.1.3. Land Data

#### 4.1.3.1. Soil Texture

Soil texture data collected during the field survey was classified into three types. The uptake of water, oxygen, and nutrients by plants is influenced by soil texture, this ultimately have effects on crop growth. Farmers' opinions are used to collect data. Based on the data gathered from farmers, it is clear that crops cultivated in loam and silt soil textures produced greater results and yields than those planted in sandy soil **Figure 26**. To visualise the distribution of soil texture, it is spatially mapped. Farmers' opinions are used to construct a soil texture map in the study area **Figure 27**. The impact of soil texture on yield could be quantified;

*Yield (kg/ha) = 2567 + 571 × (if soil texture: Silt)*** + 490 × (if soil texture: Loam)****
[n= 500; Adj- $R^2$ = 1.7%; * sign. At 10% ** sign. at 5%; *** sign. at 1%]

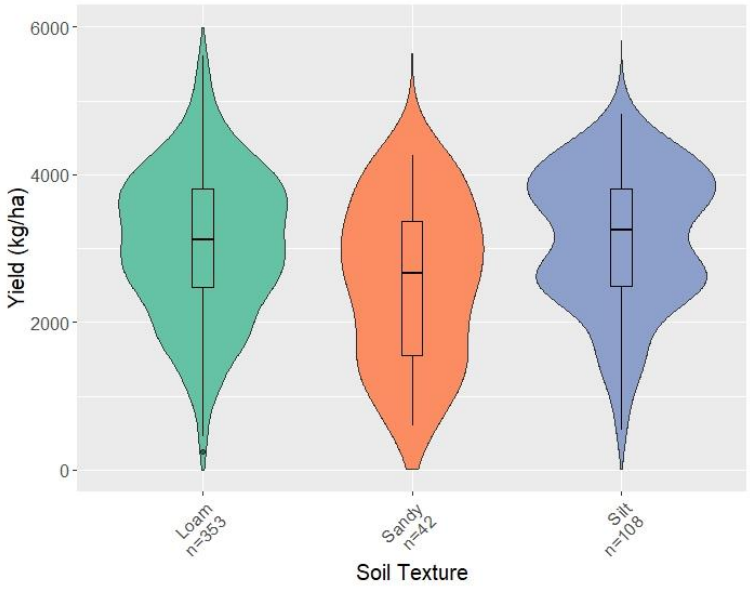



Figure 26: Soil Predominant Texture (Reported by Farmers).

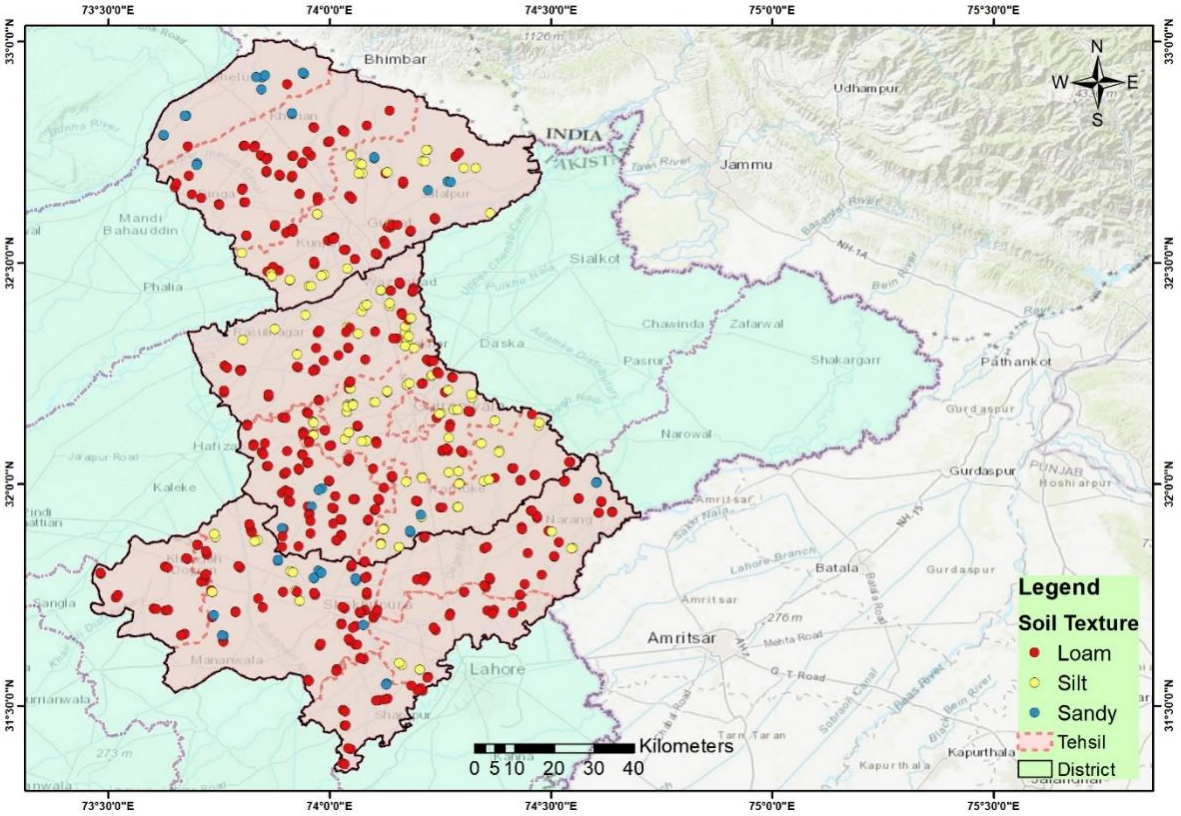



Figure 27: Spatial Distribution of Soil Texture Samples (Reported by Farmers).

#### 4.1.3.2. Water Management (Land type)

Water availability on agricultural land is an important factor for achieving higher yields. In this study, Land type was classified into two types (i.e., Irrigated and Un-irrigated). The definition of agricultural irrigated land refers to the agricultural area purposely provided with water, including land irrigated by controlled flooding. Field data showed that 422x reported using tube well, 10x used canal water for irrigation during the cropping season, and 71x sites reported as unirrigated and possibly relied on rain. Unirrigated agricultural land is generally the land with no supply of water; also considered to be dependent on rainfall to fulfil the water requirement for crops. **Figure 28** shows the variation of yield w.r.t the land type. Farmers reported that unirrigated agricultural land achieved a low yield compared to the yield produced in irrigated lands. **Figure 29** shows the water management map of the study area developed based on farmers' opinions. The spatial distribution of samples shows that the northern part of the study area District Gujrat is considered unirrigated land. Farmers reported that unirrigated land is from District Gujrat. According to the agriculture report 2019-2020 published by CRS, Punjab, Gujrat has 108275 ha, Gujranwala 174 ha, and Sheikhupura reported with the null area of unirrigated agricultural land.

The impact of water management on yield could be quantified;

*Yield (kg/ ha) = 1994 + 1212 × (if irrigated by tube well)*** + 1153 × (if irrigated by canal)****

[n=500; Adj- $R^2$ = 17.4%; * sign. At 10% ** sign. at 5%; *** sign. at 1%]

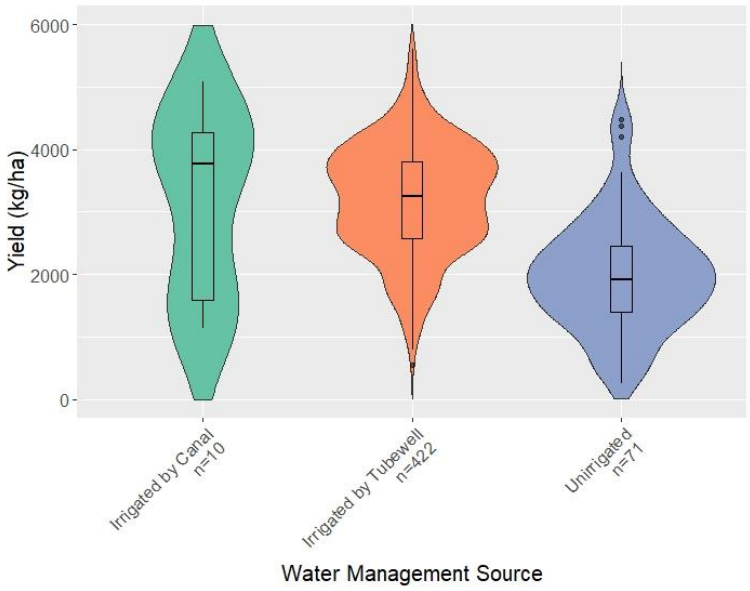



Figure 28: Water Management vs Yield.

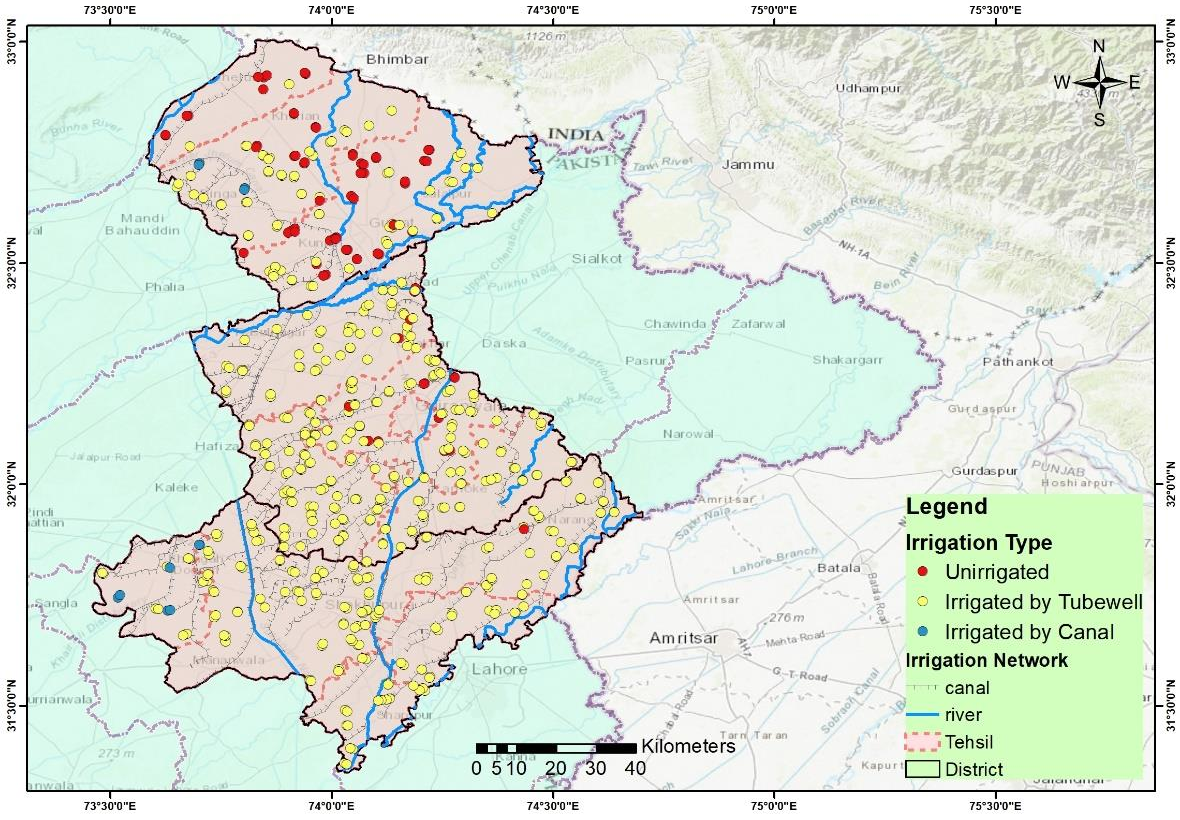



Figure 29: Spatial Distribution of Land Type samples (Water Management) reported by farmers.

#### 4.1.4. Crop Production Model

Based on the results of descriptive statistics in **Table 9**, a list of independent variables was selected for consideration in an approximate production function obtained through stepwise multiple linear regression. The resulting function is presented in Table 10. The model explained yield variability with 41.2% (adjusted- $R^2$) within fields.

The stepwise multiple linear regression analysis suggests the following notable changes in the deductions from descriptive statistics:

- DAP found important parameters that positively impact crop yield but could no longer be detected as significant in the final model.
- Treatment of Seed before sowing was found to be significant; therefore, only one wheat variety that was treated became an important variable.
- Planting time is crucial for proper crop germination, but results showed no such significance of planting time with yield.

Table 9: Summary of results obtained through descriptive statistics.

| Yield (kg/ha) | is: | each explanatory variable is tested individually | and explained % |
|---|---|---|---|
| 28 | x | Sowing date (day of year) after 295 | 1.20 |
| -0.298 | x | Sowing date$^2$ | |
| -280 | if | weed infested (yes/no) | 1.50 |
| 571 | if | Soil texture "silt" (yes/no) | 1.70 |
| 490 | if | Soil texture "loam" | |
| 133 | x | Harvesting date (day of year) after 105 | 1.80 |
| -4.11 | x | Harvesting date$^2$ | |
| 479 | if | Field prepared by Rotavator (yes/no) | 2.00 |
| -331 | if | Pest attacked (yes/no) | 2.40 |
| 17 | x | Seed quantity (kg/ha) | 2.90 |
| 626 | if | Variety used "Galaxy" | 4.30 |
| 776 | if | Galaxy uncertified seed used | |
| 554 | if | Pest spray applied (yes/no) | 7.30 |
| 791 | if | Weed spray applied (yes/no) | 8.00 |
| 779 | if | Harvested by combine harvester | 10.00 |
| 15 | x | DAP (kg/ha) | 11.10 |
| -0.033 | x | DAP (kg/ha)$^2$ | |
| 999 | if | Previously land used for "Rice" | 13.00 |
| 474 | if | Others treated seeds used | 14.50 |
| 933 | if | Faisalabad treated seed used | |
| 817 | if | Galaxy untreated seed used | |
| 739 | if | Sowing pattern "Machine Broadcasting" | 17.20 |
| -843 | if | Sowing pattern "Machine Line" | |
| 1212 | if | Irrigated by Tube well | 17.40 |
| 1153 | if | Irrigated by Canal | |
| 20 | x | Urea (kg/ha) | 28.10 |
| -0.035 | x | Urea (kg/ha)$^2$ | |

Table 10: Established overall production function.

| **Yield (kg/ha) =** | **925** | **Explained %** |
|---|---|---|
| 14 | x Urea (kg/ha) | 91.7 |
| -0.026 | x Urea (kg/ha)$^2$ | -63.7 |
| 427 | if Faisalabad treated seed used | 14.4 |
| 406 | if Irrigated by tube well | 14.9 |
| 444 | if Sowing pattern (machine broadcasting) | 19.5 |
| 352 | if pest spray applied | 17.4 |
| -279 | if pest attacked | 13.7 |
| N= 495, Adjusted $R^2$ = 41.2% | | |

#### 4.1.5. Estimation of Yield Gap by Yield Constraint

The 'mean' and 'best' values for each explanatory parameter determined by stepwise multiple linear regression were evaluated using the production function and comparative performance analysis (CPA). The combined impact of these explanatory variables was assessed, and their contribution to the overall yield gap was estimated in **Table 11, Figure 30**. Estimates of the respective contribution were based on comparisons of the average yield with the best possible yield value reported from the 503 surveyed sites. **Figure 31** shows linear regression between measured yield and predicted yield whereas few outliers are also found in the data which either represesting a different land change or errors in the site-specific field data. Sites with yield around 3000-4000 kg/ha found close to the regression line. Predicted yield model with six identified field parameters showed a positive relation with measured yield model.

Table 11: Impact by Field Parameters and its estimated contribution to the overall yield gaps.

| **Model** | **Coefficients** | **Descriptive Statistics** | | | | **Best Value** | **x coefficient** | | **Yield Gap** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **Min** | **Max** | **Mean** | **St. Dev** | | **Mean** | **Best** | **Diff** | **Perc** |
| Constant | 925 | | | | | | 925 | 925 | | |
| Urea (kg/ha) | 14 | 0 | 371 | 188 | 66 | 247 | 1681 | 1828 | 148 | 8 |
| Urea-Squared | -0.026 | - | - | 35400 | - | 61009 | - | - | | |
| Faisalabad Treated Seed | 427 | 0 | 1 | 0.13 | 0.34 | 1 | 56 | 427 | 372 | 21 |
| Irrigated by Tube well | 406 | 0 | 1 | 0.84 | 0.37 | 1 | 341 | 406 | 65 | 4 |
| Machine Broadcasting | 444 | 0 | 1 | 0.26 | 0.44 | 1 | 115 | 444 | 329 | 18 |
| Spray Pesticides | 352 | 0 | 1 | 0.41 | 0.49 | 1 | 144 | 352 | 208 | 12 |
| Attack by Pest | -279 | 0 | 1 | 0.40 | 0.49 | 1 | -112 | 0 | 112 | 6 |
| Estimated | | | | | | | 3150 | 4382 | 1232 | |
| Actual | | | | | | | 3034 | 5606 | 2572 | |

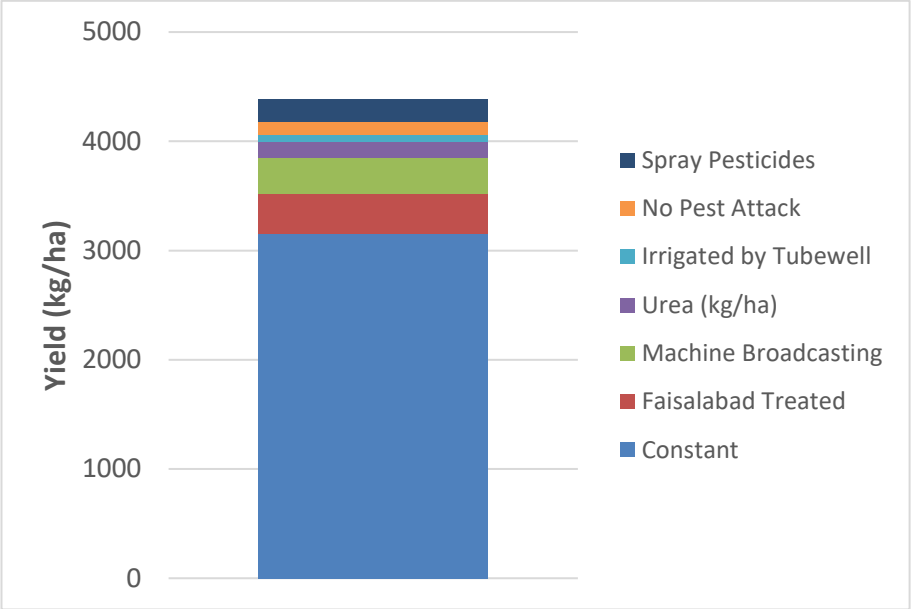



Figure 30: Contribution of significant parameters to the yield gaps.

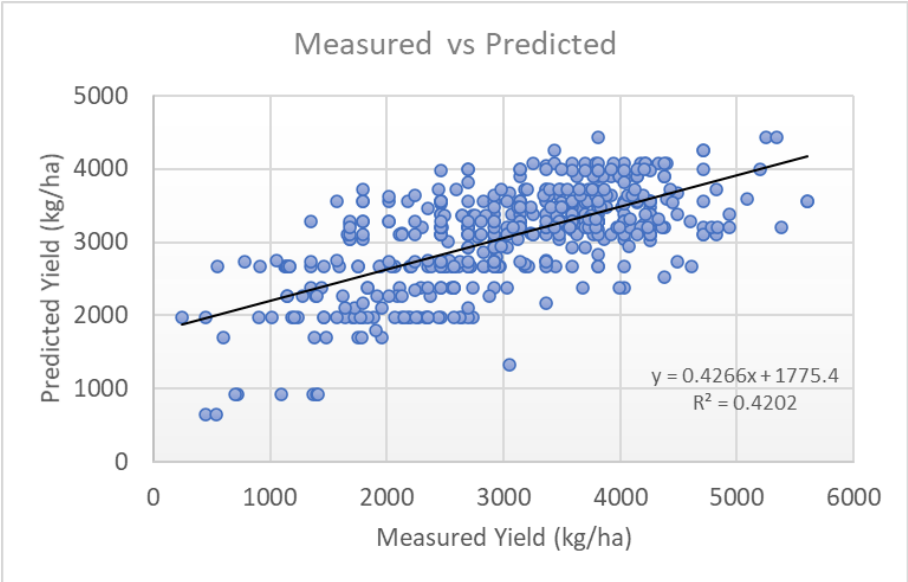



Figure 31: The relationship between measured yield (kg/ha) and predicted yield (kg/ha) after descriptive statistics.

## 4.2. Producing NDVI- based Crop Production System Zones

After the pre-processing and filtering process of NDVI climatology of 20 years, a stack of the whole series from 1999-to 2020 was grouped into ten classes using ISODATA unsupervised classification algorithm. An optimum number of clusters to produce zonation are unknown. Therefore 50-20-10 classes were produced to find the most suitable and optimum number keeping in mind the spatial coverage of the study area. As a result, ten classes were produced. **Figure 32** shows the spatial distribution of the produced CPS zones.

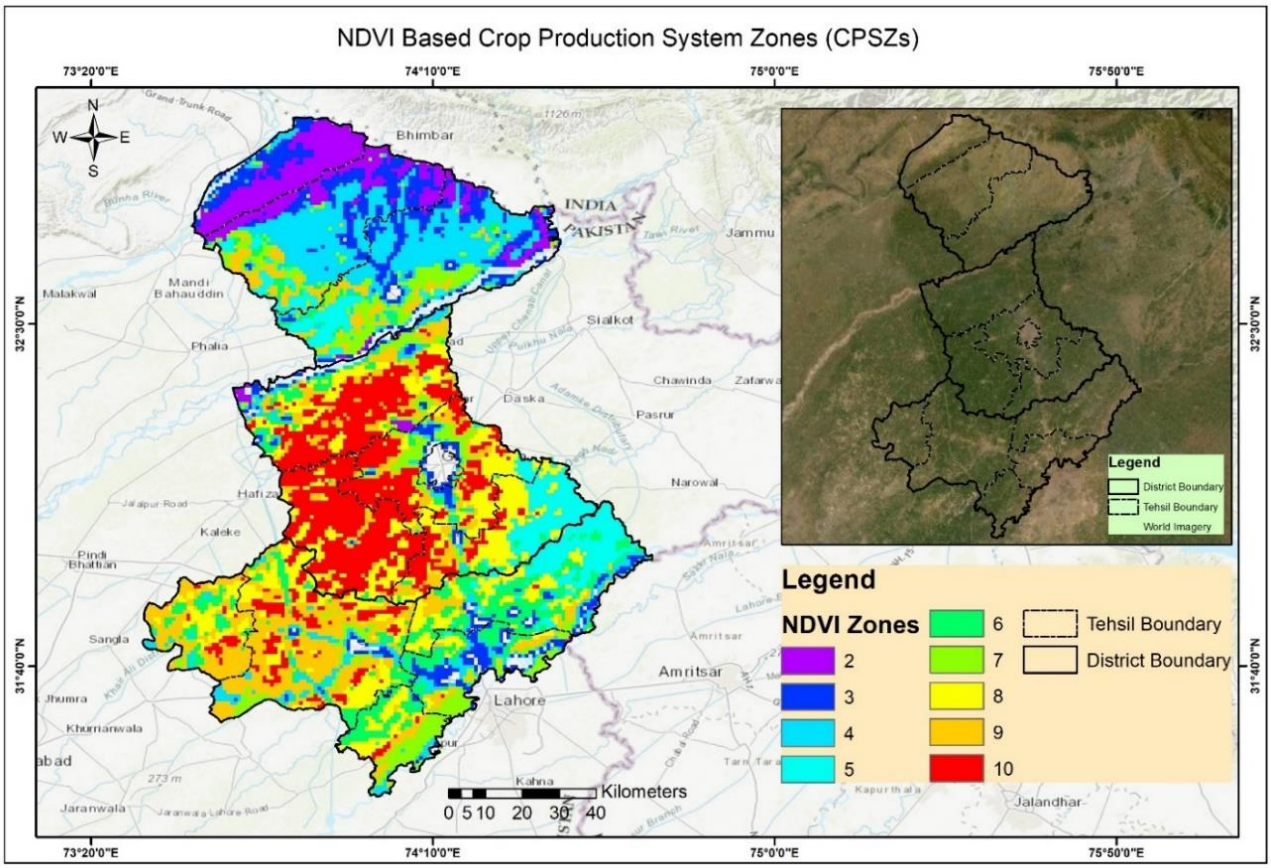



Figure 32: NDVI-based Clustering of CPS Zones.

To elaborate in more detail about the meaning of relatively similar classes, the temporal behaviour of NDVI climatology median among 10 classes is shown in **Figure 33**. Temporal behaviour based on NDVI climatology can indicate different land cover types, including the long-term productivity of agricultural patterns. As discussed in the **subsection 3.2.1**, the temporal behaviour of one year is categorized into 10-50-90 percentiles. Dekad 1-36 showed the NDVI- profiles at the 10th percentile, dekad 37 to 72 at the median (50th) percentile, and dekad from 73 to 108 at the 90$^{th}$ percentile. The 10$^{th}$ and 90$^{th}$ percentile help in understanding the lowest and highest possibilities of NDVI values in each land-cover pattern. Temporal profiles indicated the existence of two cropping seasons in the study area (Rabi and Kharif). Rabi season starts between dekad 30 and 33 (Oct-Nov) and ends around dekad 12 (end of April). The major crops in the study area during the Rabi seasons 2019-2020 were Wheat, Barley, Maize, Gram, and Potato (Autumn). In this study, the focus was on median percentile and the specific wheat season as mentioned in the Figure 33.

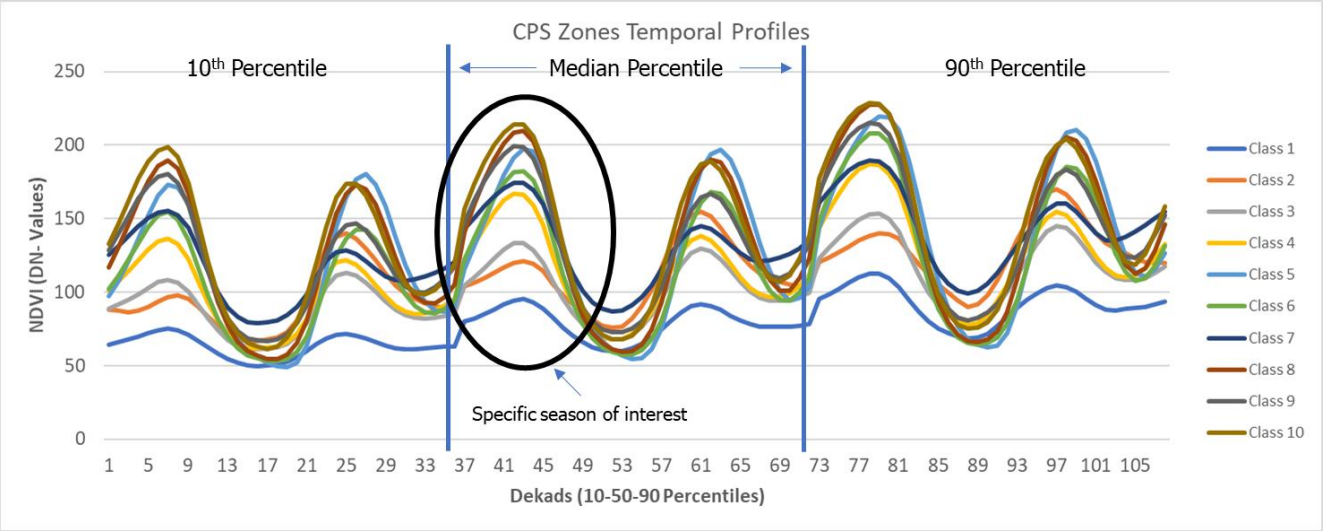



Figure 33: Temporal Behaviour of One Year with (10-50-90 Percentile).

Further, temporal profiles of all these 10 CPS zones were analysed individually to identify land-cover types such as crop type (i.e., rainfed or irrigated) and non-agriculture land types. **Figure 34** shows the temporal behaviour of each zone produced using NDVI climatology. Figure **34-a** indicates a temporal behaviour of the non-agriculture type; the NDVI value is relatively low and stable throughout the year with no abrupt changes. Thus, this indicates the bare land, including urban areas, and water dominance within the zone, including rivers and canal networks. Figure **34-b** and **34-c** indicate landscape including bare land near urban blocks, waterbodies, grass, and shrublands in the north part of the study area with high altitude. Figure 34-d and 34-g indicate agricultural land mixed with other land cover types with two growing peaks in a year. The first peak in both figures around dekad 6-7 (march) shows a slightly high NDVI value, indicating cropland (i.e., rainfed wheat or potato). In contrast, the second peak with a short span of time is not that high and shows the presence of greenness which indicates that either there was any crop grown for a short period to make the land fertile for the next season or animal fodder possibly. **Figures 34-(e,f,h,i,j)** indicate agricultural land with two high NDVI value peaks. These high peaks indicate towards few points about crop and land; the wheat is grown in this area is better than in the other zones, and the land is irrigated.

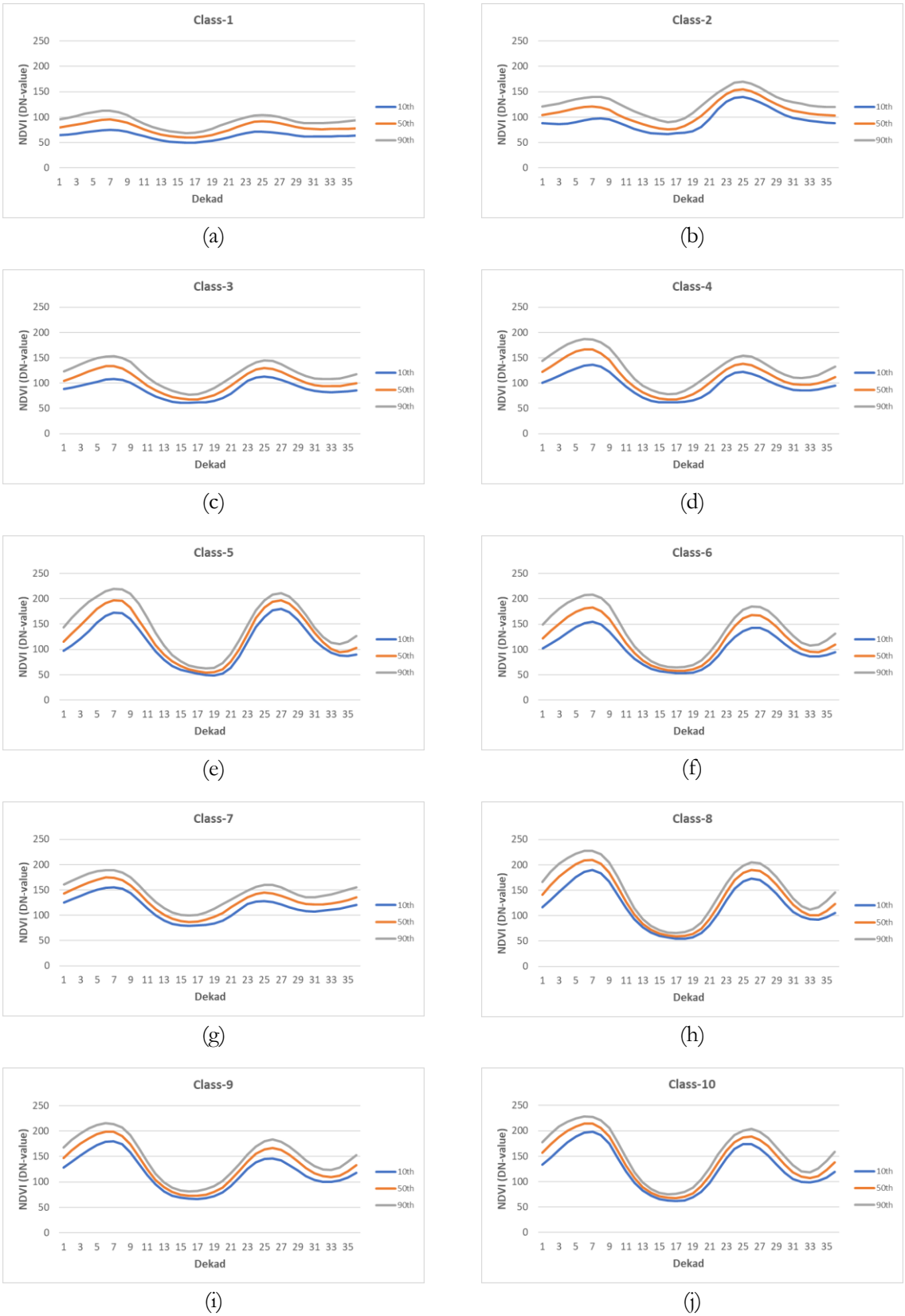



Figure 34: Classes Produced Through NDVI Climatology.

### 4.2.1. Comparing CPS Zones with Site Parameters

As mentioned earlier, wheat crop yield data for one year (2019-2020) was used, whereas the CPS zones produced using SPOT- ProbaV sensor are based on 20 years of NDVI time-series. There is a possibility that land-use changes occurred in the landscape during this period as people left farming or converted their lands into housing properties (Mukhtar et al., 2018). Therefore, a visual comparison was made to observe the similarities and differences between remotely sensed produced CPS zones and site-specific field parameters. **Figure 35** shows a comparison of CPS zones with various field parameters to assess the NDVI bases stratification of the study area. **Figure 35-a** shows ten classes of CPS zones. **Figure 35-b** shows the soil texture of surveyed sites in the study area. From the previous **section 4.1**, it was found that fields with silt texture achieved higher yields, whereas sandy soils showed lower yields. In the map, the distribution showed the north part, and a few sites in the south part reported sandy texture, whereas the central part reported mostly loam and silt texture. Wheat produced in Silt and Loam soil texture was reported to achieve a higher yield than sandy soil (Mojid et al., 2020). **Figure 35-c** shows the land type distribution by water resources (i.e., Irrigation by Tube well or Canal/ Unirrigated). The farmers from the north part of the study area reported unirrigated land, which refers to the rainfed wheat crop. Based on NDVI stratification, a clear difference can be seen in the study area, differentiating the study area into irrigated and unirrigated land (Irrigated wheat/ Rainfed wheat). **Figure 35-d** shows the land-use pattern in the study area. Land use patterns also impact crop yield; data showed three types, i.e., Fallow, Fodder, and Rice. The majority of the farmers grow two crops in a year, and this can be seen in the map that except north part of the study area where farmers kept their field lands fallow in the last season or grown fodder for animal feeds remaining in all study area rice was grown by farmers. **Figure 35-e** shows the sowing pattern distribution in the study area, i.e., Manual, Broadcasting, and Line pattern. As discussed in **section 4.1**, farmers with sowing pattern broadcasting achieved higher yields compared to other patterns; this pattern can be seen in the long-term NDVI-based CPS zones map. In the broadcasting method, wheat grain yield increases as compared to other methods and, resulting in high NDVI values (Abbas et al., 2009).

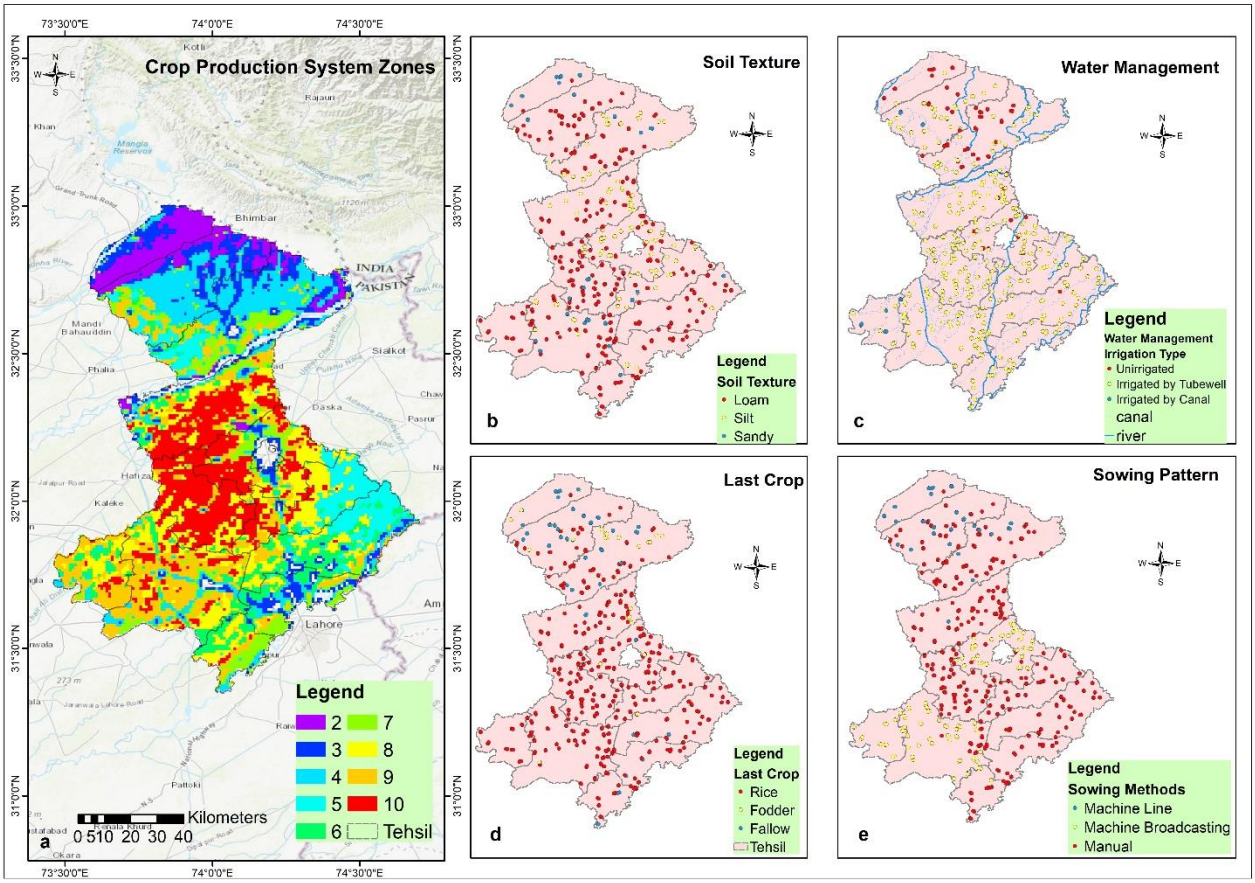



Figure 35: Comparison of CPSZs with Field Parameters (Reported by Farmers).

## 4.3. Assessing Relationship of Site-specific Measured Crop Yield between CPSZs vs. Admin Areas

After producing CPSZs from long-term NDVI climatology of 21 years, the following process assesses the relationship of measured yield with CPSZs vs. the existing sampling method approach. Regression models for both approaches were developed to find how much both approaches could explain yield variability. Later, the yield variation between CPS zones and Admin area was plotted using box and whisker plots.

### 4.3.1. Developing Regression Model for CPZ zones vs. Admin areas with Yield

The relationship of wheat yield between CPS zones and Admin areas was investigated through multiple linear regression. In the first model, the predictors were the CPS zones within each district, and the response variable was the yield (kg/ha). Whereas in the second model, the predictors were the Tehsil (Admin areas) within each district, and the response variable was yield (kg/ha). The result of the regression model is shown in **Table 12**; the coefficient column in the table represents the yield variability explained within each cluster. Class 4 was identified as less significant than the rest of the classes. Two NDVI classes explained more than 50% yield variability within fields in each cluster, four classes were able to explain yield variability between 30% and 10%, and one class could explain less than 10%. This regression model could explain 23.3% (adjusted- $R^2$) yield variability between CPS zones with RMSE 876.64 kg/ha. **Table 13** showed the regression model for admin areas and yield explained 40.5% (adjusted- $R^2$) yield variability. The result showed that the admin areas approach performed relatively better than the CPSZs approach.

Table 12: Regression model result between measured yield and NDVI zones.

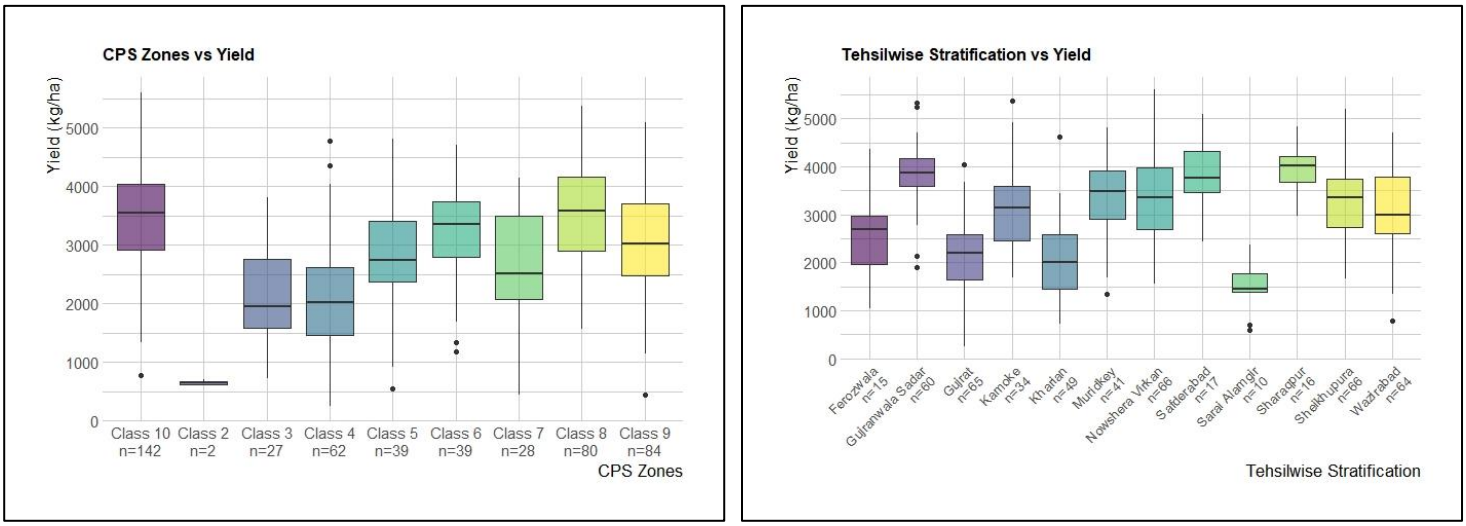
| NDVI-Zone | Coefficient | Sig. |
|---|---|---|
| Class-4 | 0.01 | 0.83 |
| Class-5 | 0.19 | 0.00 |
| Class-6 | 0.30 | 0.00 |
| Class-7 | 0.13 | 0.02 |
| Class-8 | 0.51 | 0.00 |
| Class-9 | 0.36 | 0.00 |
| Class-10 | 0.60 | 0.00 |
| (n= 495; adjusted-$R^2$ = 23.3%) | | |

Table 13: Regression model result between measured yield and Admin areas.

| Tehsils | Coefficients | Sig. |
|---|---|---|
| Gujrat | 19.3 | 0.03 |
| Kharian | 18.2 | 0.02 |
| Gujranwala Saddar | 76.6 | 0.00 |
| Nowshera Virkan | 63.3 | 0.00 |
| Wazirabad | 53.2 | 0.00 |
| Kamoke | 41.7 | 0.00 |
| Sheikhupura | 58.5 | 0.00 |
| Muridkey | 52.7 | 0.00 |
| Ferozwala | 18.5 | 0.00 |
| Sharaqpur | 43.4 | 0.00 |
| Safderabad | 41.6 | 0.00 |
| (n= 491; adjusted- $R^2$ = 40.5%) | | |

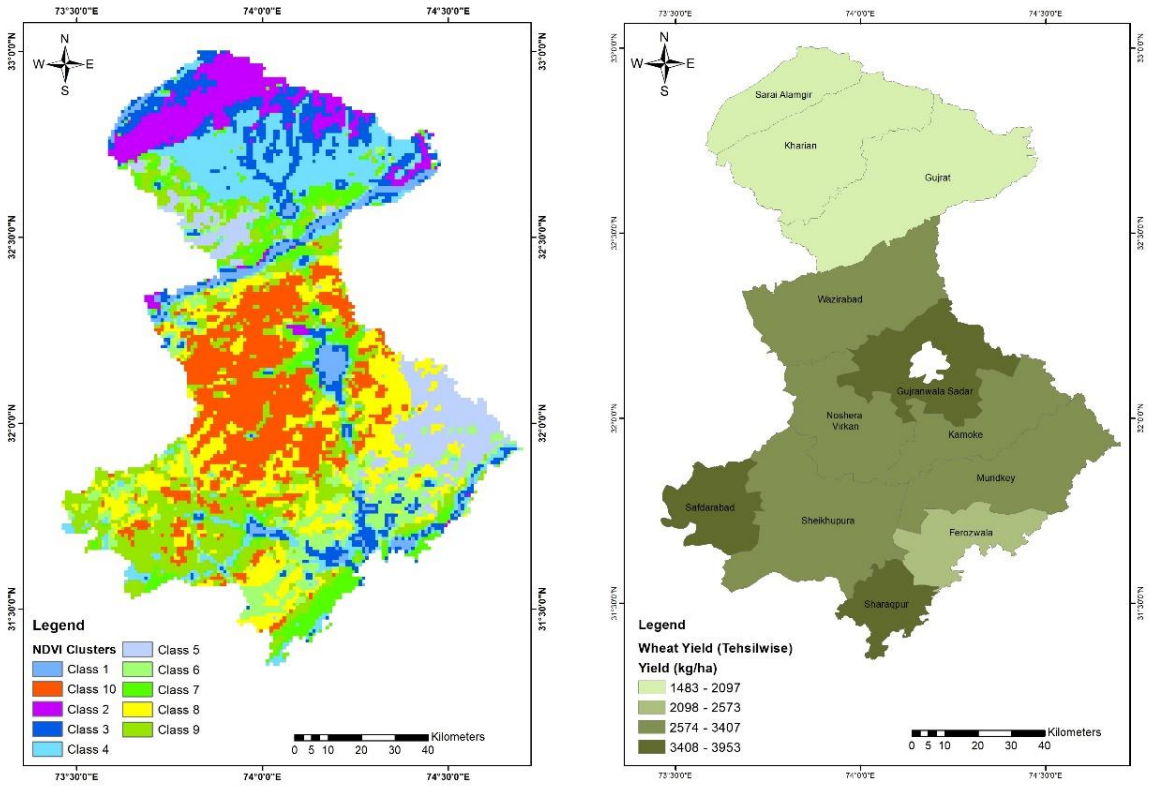
#### 4.3.2. Plotting Yield Variation Between CPS zones vs. Admin areas

Box and whisker plots were plotted to observe how yield varies in both scenarios and to what extent these two approaches explain. Figure 36 shows the box and whisker plots of two approaches (CPS zones vs. Admin areas) with yield. Figure 36-a shows the variation of yield in 9 CPS zones. From the analysis output, this would be easy to differentiate between zones with low & high yield performances. This also indicates the type of land, i.e., irrigated and rainfed land. Figure 36-b showed the variation of yield in tehsil-wise stratification (sampling approach followed by CRS- Punjab). Sampling size tehsil wise was not uninformed, as can be seen in the figure. As mentioned in the Conceptual Diagram, villages within Tehsils are selected based on the village's total cultivated land. That was the reason the number of samples did not look uniform. Yield variation in tehsils also explained the difference between land types, as mentioned earlier. Tehsil Sarai Alamgir (10x) was identified with the lowest yield. Similarly, Tehsil Kharian (49x) and Gujrat (65x) were identified with relatively high yields from Sarai Alamgir but lower than other tehsils. Tehsil Sharaqpur (16x), Gujranwala (60x), and Safderabad (17x) were identified with high yield among these 12 tehsils. Figure 37 shows a comparison made between CPS zones and the Tehsil wise wheat yield map produced using published wheat production for the year 2019-2020. In the map, a clear difference between low and yield areas can be seen in both situations. Rainfed areas in the north part of the study area are identified with lower yield in Figure 37-a, whereas in CPS zone map Figure 37-b, the north part is classified in zone 2-3 & 4 with lower NDVI temporal profiles.




a b

Figure 36: Yield variation between CPS zones vs. Admin areas.




a b

Figure 37: Visualization of Both CPSZs vs. Admin Areas.

### 4.4. Combining Studied Parameters of Sub-objective (i) and (ii) into One Integrated Model

To assess the combined impact of significant field parameters (G*M*L) and cps zones on yield variability, a final model was developed by combining all these identified parameters in sub-objective (i) and (ii). SMLR was applied to quantify their impact. 14 independent variables were taken into consideration to develop the final model. The model is explained with an adjusted-$R^2$ of 43.2% for yield variability in the reported crop yield estimations. The result of the regression model is shown in **Table 14**. Results showed that among field parameters (G*M*L), the model identified two CPS zones as important variables for explaining crop yield estimation Class 4 and Class 8. As mentioned earlier in the section, Class 4 was identified with relatively low NDVI value and rainfed area with two annual peaks; the first peak around March-April and 2nd peak around Sep-Oct. Conversely, Class 8 was identified as having high NDVI values with two annual peaks. The regression model also explained the similar situation that farmers within Class 4 achieved -388.80 kg/ha less than the average yield. In contrast, the farmers within Class 8 achieved +225 kg/ha more than the average. The combined impact of these 8 variables could be quantified:

*Yield (kg/ha) = 1044.44 + 14.07 × (Urea (kg/ha) – 0.028 × (Urea (kg/ha)² ± 384 × (If Faisalabad Treated Seed (yes/no) ± 361 × (If Irrigated by Tube well (yes/no) – 388.80 × (Field in Class-4) ± 392 × (If Sowing pattern; Broadcasting by Machine (yes/no) – 277 × (If Pest Attacked) + 288 × (If Pest Spray Applied) + 225 × (Fields in Class-8).*

Table 14: Established Combined Impact of Significant Variables and CPS Zones.

| Parameters | Coefficients (B) | Coefficients (Importance) | Sig. | VIF |
|---|---|---|---|---|
| (Constant) | 1044.44 | - | 0.00 | - |
| Urea (kg/ha) | 14 | 93.3 | 0.00 | 18.52 |
| Urea-Squared | -0.028 | -63.2 | 0.00 | 17.13 |
| Treated seed (Faisalabad) | 385 | 13.0 | 0.00 | 1.28 |
| Irrigated by Tube well | 361 | 13.3 | 0.00 | 1.36 |
| CPS-4 | -389 | 12.8 | 0.00 | 1.19 |
| Sowing pattern (Machine Broadcasting) | 392 | 17.2 | 0.00 | 1.38 |
| Attacked by Pest | -278 | 13.6 | 0.00 | 1.23 |
| Spray Pesticides | 288 | 14.2 | 0.00 | 1.43 |
| CPS-8 | 225 | 8.2 | 0.01 | 1.06 |
| (n= 493; Adjusted-$R^2$ = 43.2%) | | | | |

# 5. DISCUSSION

## 5.1. On Quantifying Site-Specific Field Parameters (Descriptive Statistical Analysis)

The results confirm that site-specific field parameters could help understand the crop dynamics during the growing season. Each site-specific parameter collected in the survey was analysed individually to estimate its impact on wheat yield. Wheat, considered the most important staple food crop in Pakistan and Punjab, is of particular importance as it contributes 70-75% to the total annual wheat production of the country (Mudasser et al., 2001). A recent report on Food Crises 2020 published by The Global Network Against Food Crises alarmed about the food security severe threat to the country. In such challenging and demanding situations, it is essential to understand the crop dynamics and find solutions for improvement. This study used an easy and attractive statistical approach called comparative performance analysis to analyse site-specific field parameters (Kees de Bie, 2002). This approach was found to be applied to identify the significant field parameters and quantify their impact on wheat yield by estimating the yield -gap.

This study started with the descriptive statistics of the site-specific parameters. Field data of 600 wheat yield samples (2019-2020) were provided by CRS, Punjab. In this study data cleaning process was conducted twice at two different stages. First data cleaning was performed at the beginning of the analysis in **section 3.1** (sub-objective 1), where all those points were outside the study area, or incomplete data entries were removed, and a total of 543 samples remained in the database for analysis. The second round of data cleaning was performed before **section 3.2** (sub-objective 2). After this total of 503 field samples were taken into analysis. Field data about 46 different parameters were categorized into three data types (genetics-G, management-M, land-L). All parameters were taken into consideration to identify the most significant variables. In individual descriptive statistics, 23 parameters showed significance with wheat yield. In the second analysis, a further deduction was made to eliminate redundant, less important variables and make the model more efficient (goodness of fit). Stepwise multiple linear regression analysis was carried out to eliminate the least contributive parameters and develop a model to estimate the yield gap. The accuracy of the outputs depends on how farmers report to the enumerator during the survey. The final regression model could explain 41.2% (adjusted R2) yield variability within fields. Six variables were identified as the most significant among all site-specific field parameters.

Urea (46:0:0) is a low-cost nitrogen fertilizer type utilized as an essential input to supply crop plants with green leafy growth and improve crop photosynthesis. Urea was identified as one of the total six important field parameters by the regression model. Urea was the most important predictor, with deviance explaining 28% yield variability. There are certain recommendations about the timing of the application of Urea, which were not collected in the survey. The important finding of the Urea application was that it helps increase the yield at a certain limit; after that, results showed a decline in yield. This cause might have occurred due to application timing or soil fertility. Next to Urea was the Machine Broadcasting sowing pattern that explained 19.5% yield variability. Machine broadcasting outperformed the machine line (drill) method as studies showed that the drill method is considered the best way to plant wheat. Still, the result showed that farmers with machine broadcasting methods achieved higher yields than the line method, whereas manual broadcasting also showed better results. Another important parameter identified significant was irrigation by tube well. Punjab is known as the land of water, but in recent years the irrigation system faced severe problems due to climate change. Thus, farming systems shifted to tube well systems (electrical or diesel-based). Results also showed a

similar picture that most farmers reported Irrigated by Tube well. Interestingly only 10x farmers reported Canal water while 422x reported Tube well; this also highlights the severity of the Canal irrigation system in the country. The seed treatment process was found to be very important before sowing. Treated seeds are able to protect against many diseases from a fungicide, whereas some seed treatment products also provide additional protection against season insects and thus produce a high yield. In this study, farmers who used treated seeds achieved approximately 400 kg/ha additional yield as compared to untreated seeds. By comparing the calculated average yield with the best yield, the overall yield gap was estimated to be around 1802 kg/ha. In comparison, the yield gap between the reported average and best yield was approximately 2575 kg/ha.

The accuracy of the model could be higher than the achieved, as various field parameters and information were missed, for example, i) the timing of sowing and harvesting was generalized, whereas the actual dates could provide extra information about crop dynamics in a specific region, ii) timing of fertilizer application was missed, there is usually a recommendation for fertilizer application such as in Punjab, Pakistan the recommendations for irrigated and rainfed zones are different. In the irrigated zone, 5 bags of DAP and 2.5 bags of Urea and Potash per hectare are recommended at sowing time and 1 bag of Urea with 1$^{st}$ and 2$^{nd}$ irrigation. For the rainfed zone, 3.5 bags of DAP and 2.5 bags of Urea and Potash are recommended at sowing time (Wheat Program, PARC). This information about timing was missed. Similarly, irrigation times, soil fertility, and climate data may provide additional information about crop yield. These parameters could improve the model accuracy for explaining yield variability. For further studies, the quality of site-specific field surveys can be improved by incorporating the missing parameters in the survey.

### 5.2. On producing NDVI- based Crop Production System Zones

Various inputs such as topography, soil, climate, and land use data are required to produce Agro-ecological zonation (AEZ) (Ahmad et al., 2019; de Bie and Nelson, 2021; de Bie, 2020; Mohammed, 2019). The surprising fact is that AEZ produced with such an amount of data inputs does not update frequently and is thus used for many years. In Pakistan, the AEZ map that has been used until 2019 was produced in 1980, and recently, FAO launched a new AEZ of Punjab in 2019. Long-term NDVI stratification was found to be an appropriate alternative solution in identifying the crop productivity pattern based on the fact that the greenness of NDVI differs (de Bie, 2020).

Unsupervised classification technique ISODATA algorithm was used to produce CPS zones through NDVI climatology for 20 years from 1999 to 2020. Results showed that long-term NDVI of SPOT-ProbaV could be used to stratify CPS zones. As discussed in **section 3.2.2**, the ISODATA algorithm minimizes the human interference in the classification process and merges classes with similar response patterns. Areas with a similar pattern were merged into specific zones based on the long-term crop productivity pattern. The crop production system zones clearly explained crop phenology (two cropping seasons) in the study area, Rabi and Kharif seasons. Two major crops in the study area are Wheat in the Rabi season and Rice in the Kharif season. The wheat crop in the northern part of the study area relies on rainfall, and the central and southern part is considered to be irrigated land, which can be seen in the Crop Production System Zones. NDVI temporal profiles with stable low values indicate bare land and non-agriculture lands, including shrub lands and grasslands. The stable NDVI temporal profile with high values indicates crop phenology of two seasons in the study area, including wheat crops in both rainfed and irrigated regions. The study area was stratified into 10 clusters based on pixel responses.

Based on the spatial extent of the study location, 10 clusters were found suitable for the study area's stratification. The number of clusters was determined based on the spatial extent of the study area. 50-20 and 10 clusters were produced to explain the agricultural activities in the study area. Ten clusters were considered the optimal number to reduce the similar clusters. This can be considered a simple and time-efficient approach for optimizing the image processing and producing crop production system zones. For further studies, separability analysis (i.e., minimum, and average separability) can be a better alternative to determine the optimum number of clusters, especially when the spatial extent is complex and consist of a fragmented agricultural landscape (Ali et al., 2012; de Bie, 2020). Existing Agro-ecological zones of Punjab published in 2019 classified the study area into two zones (Rice-Wheat and Rice) based on many inputs. **Figure 38** shows the different inputs used to produce AEZs of Punjab (taken from FAO report "Agro-Ecological Zones of Punjab-Pakistan 2019). This zonation does not depict the farm-level variability in the country's complex agricultural landscape of the country especially when the majority of the farmers hold small lands. On the other side, the earth observation-based stratification helps in understanding the crop dynamics and captures the yield variability at 1km resolution. To justify the advocacy of NDVI-based stratification, **Figure 39** shows the map of Pakistan developed by the faculty of ITC in 2012 and explains detailed agro-climatic variability (Kees de Bie, 2012). The NDVI temporal profiles represent unique responses for land cover types, which helps classify data into clusters sharing similar responses.

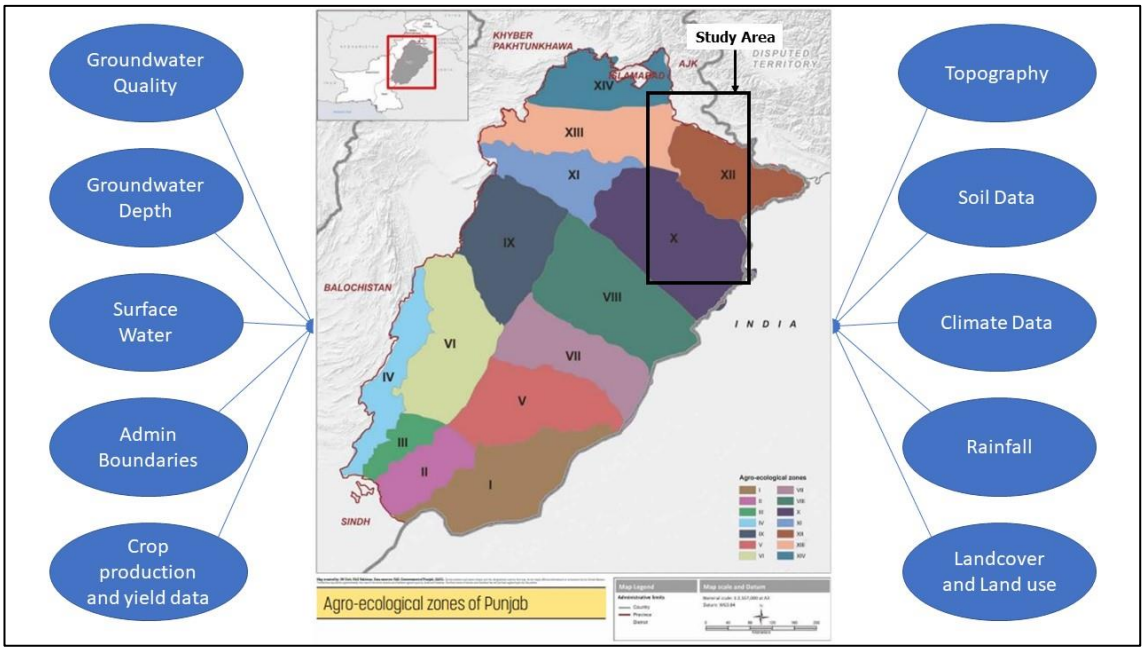



Figure 38: AEZ Map of Punjab (Ahmad et al., 2019)

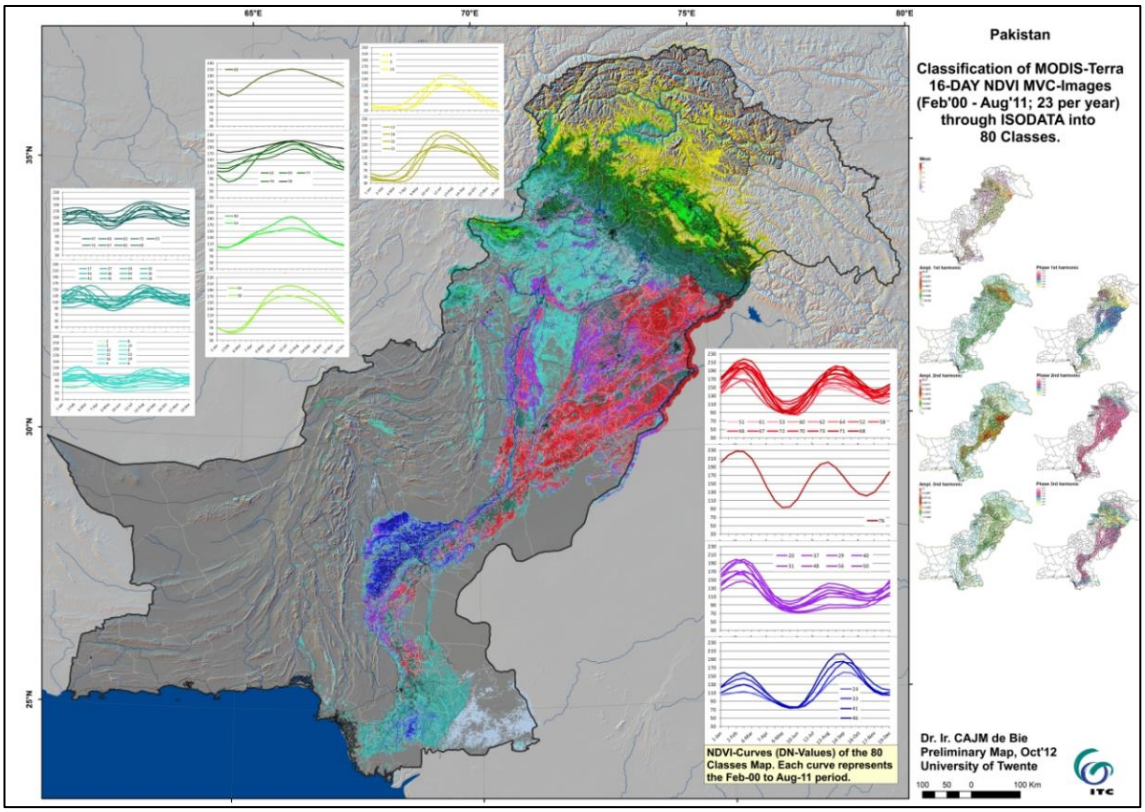



Figure 39: NDVI Based Stratification of Pakistan (Produced by Kees de Bie in 2012).

### 5.3. On Assessing Site-specific Measured Yield between CPSZs vs. Admin Areas

After producing CPS zones, the next step was to develop a regression model and assess how much CPS zones explain yield variability compared with the administrative areas sampling approach. The NDVI-based crop production system zones explained 23.3% of the wheat yield variability in the study area, while the existing approach in the study area explained 40.5% of yield variability. Both models

explained yield variability with low adjusted $R^2$; however, this does not mean that stratification of CPS zones in the study area does not represent it well. A comparison was made between the average reported yield Tehsil wise map and the NDVI-based CPS zones map to visualize. CPS zones not only explained the on-ground reality but also highlighted a few significant points as CPS zones could differ between rainfed and irrigated fields, soil texture, and fertilizer applications.

In this study, the field data inaccuracies were found to be the key factor behind the low performance of CPS zones. This can raise questions about the published statistics. After performing the 2-stage data cleaning step, a third series of error found the GPS locations of data entries. **Figure 40** shows the one error found in the detailed monitoring of field entries on Google Earth Pro; **Figure 40-a** shows the site-specific entries found inside the urban area; when further zoomed-in, four site-specific entries from different villages were found at the exact point **Figure 40-b**. These inaccuracies in the field data are referred to as human errors; these human errors might not affect the existing AFS approach in the study area but can affect the earth observation-based approach with low adjusted-$R^2$.

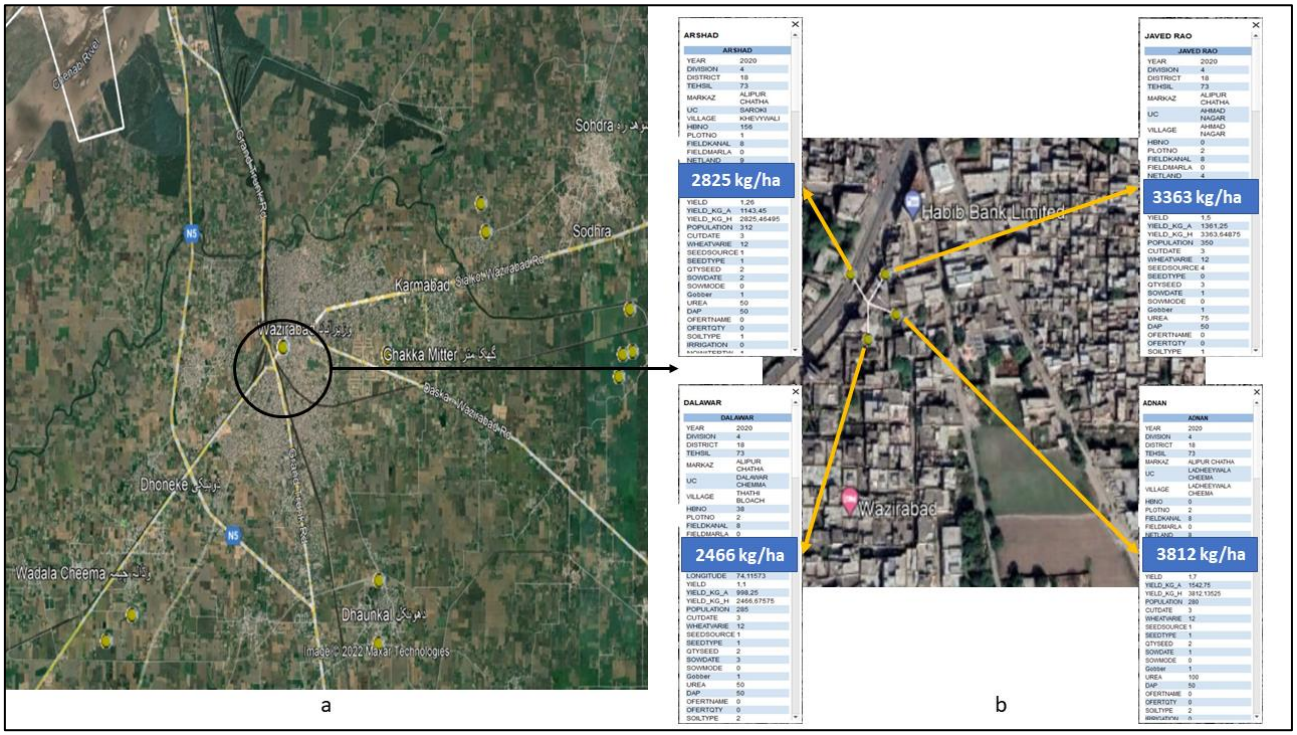



Figure 40: Field Data Errors.

Further studies can overcome this low model performance by improving the field data accuracy. According to CRS officials, this was the first-time spatial information collected with the survey yield data. In further field surveys, proper training for enumerators about maintaining GPS accuracy can help get better field samples. Here are some field pictures in **Figure 41** from a similar study in India, where rice yield data was collected. After facing the same issue, the innovative approach was adopted with a mobile application called NOTECAM. **Figure 41-a** represents the sample plot for a survey; in the small box, detail of the sample point can be seen with latitude, longitude, elevation, accuracy, time, and note. **Figure 41-d** shows the weighing of yield collected from the surveyed plot with the same information. This four-stage picture approach (i., sample demarcation ii. crop cutting iii. manual harvesting iv. weight of surveyed sample yield) can improve survey quality and data quality and indeed improve the site-specific yield estimates, which will lead to extrapolating the site-specific estimates to area-specific productions.


Figure 41: Examples of Collecting Field Data maintaining accuracy problems.

Earlier studies used a similar approach for crop area estimation, while this approach was used to estimate site-specific crop yield and improve crop production estimates. In a study by Muhammad et al. (2019) in Ethiopia, the author achieved an adjusted $R^2$ of 91.4% for estimating field fractions in the reported crop areas. Similarly, in another study by Khan et al. (2010) in southern Spain, the authors achieved adjusted $R^2$ of 98.8%, 97.5%, and 76.5%, respectively, for rainfed wheat, rainfed sunflower, and barley. Thus, the value of adjusted $R^2$ is not comparable with these studies as the main objective was clearly different. To elaborate, the relatively lower adjusted $R^2$ achieved in this current study is perhaps due to the questionable quality of site-specific data collected by agricultural officers.

### 5.4. On Assessing Combined Impact of Studied Parameters on Yield Variability

The last but not the least part of this study was to combine the identified significant field parameters and CPS zones into one model to assess their combined performance to explain variability in measure yield. 6 parameters from site-specific field data and 7 CPZ zones were integrated together in the final model to find out their combined impact makes any specific difference on the model performance. The final model explained 43.2% of the yield variability. The regression model kept two CPS zones in the final model with field parameters. As discussed earlier, CPS zones performed relatively low but rejected the null hypothesis that there was no significant relationship between NDVI-based CPS zones and measured site-specific yield. The result shows a medium positive relationship between measured yield and CPS zones. In further studies, besides enhancing the field quality and producing CPS zones with separability divergence analysis, high-resolution earth observation sensors like Landsat-8 and Sentinel-2 can further integrate with this approach to capture the in-season yield variability within fields and zones.

# 6. CONCLUSION AND RECOMMENDATION

This research developed and evaluated a method for quantifying crop yield variability at the farm level and improving crop production estimations in a fragmented agricultural landscape by integrating earth observation data (at 1-km spatial resolution) into the existing area frame sampling survey approach. Site-specific field parameters help in understanding the crop dynamics during the growing season, explaining 41.2% yield variability within fields. Urea fertilizer was founded to be an essential parameter for achieving high yield, with 28% importance.

CPS zones were produced using long-term NDVI climatology of 20 years (1999-2021). The regression model explained 23.3% of yield variability; this is somewhat low compared to earlier studies, but the stratification technique clearly distinguishes between irrigated and rainfed croplands, soil texture, and crop productivity. The performance of the earth observation approach can be improved by enhancing the field data quality. After combining the significant field parameters and CPS zones, the overall model explained 43.2% yield variability. The regression analysis added CPS zone 4 and 8 in the final model, which explains the importance of CPS zones.

This research provides significantly satisfactory results for the stratification of the agricultural landscape. It explains yield variability between zones with SPOT- ProbaV, which shows an important element in long-term NDVI climatology. Nonetheless, further study could focus on 1) e-training to field enumerators on how to capture field data accurately from the field, 2) estimating season-specific crop yield variability by producing CPS zones, 3) explore the possibility to incorporate Sentinel-3 NDVI product, 4) within zones yield variability captures using high-resolution sensors like Landsat-8 and Sentinel-2, 5) developing an effective method that allows integrating of multi-sensor data into existing sampling approach for quantifying crop production function and extrapolate from site-specific to an area/region-specific crop production. This method is easy and can be effectively integrated with the existing agricultural systems in countries with complex and fragmented landscapes.

# 7. SCIENTIFIC AND SOCIETAL IMPACT

This research investigated the possibilities of incorporating earth observation data into an existing area frame sampling approach for determining yield variability within fields and extrapolating site-specific yield to area-specific crop production estimates. In a complex and fragmented agricultural landscape, the coarse resolution satellite Spot- ProbaV 1km generated significant results for creating homogeneous stratification and producing crop production system zones. These zonation maps can assist government departments (such as agriculture and statistics) in improving their survey approach to obtain more precise and dependable crop output estimates. Agronomists can employ yield-gap analysis reports to assist farmers in closing yield-gaps within fields, allowing them to increase crop productivity. Stakeholders and decision-makers must close yield gaps within fields and improve crop output projections in order to achieve food security and SDG targets.

# 8. ETHICAL CONSIDERATION

The earth observation data of SPOT- ProbaV used in this research, are available under EU law, granting users access. It also allows users for reproducibility and distribution of data. While downloading the earth observation data the rules and regulations were strictly adhered by the student. The R-code used in this research was either written by the student or accessed from open source. The used code will be appropriately cited. The ground data of the three districts (Gujrat- Gujranwala & Sheikhupura) of Punjab, Pakistan, has been obtained on special request from Crop Reporting Services (CRS), Agriculture Department, Punjab. All stakeholders of this data are properly informed and agreed to use for this research.

# LIST OF REFERENCES

Abbas, G., Ali, M.A., Azam, M., Hussain, I., 2009. IMPACT OF PLANTING METHODS ON WHEAT GRAIN YIELD AND YIELD CONTRIBUTING PARAMETERS. The Journal of Animal & Plant Sciences 19, 30–33.

Abburu, S., Golla, S.B., 2015. Satellite Image Classification Methods and Techniques: A Review, International Journal of Computer Applications.

Abubakar, G.A., Wang, K., Shahtahamssebi, A., Xue, X., Belete, M., Gudo, A.J.A., Shuka, K.A.M., Gan, M., 2020. Mapping Maize Fields by Using Multi-Temporal Sentinel-1A and Sentinel-2A Images in Makarfi, Northern Nigeria, Africa. Sustainability (Switzerland) 12, 1–18. https://doi.org/10.3390/su12062539

Ahmad, A., Khan, M.R., Shah, S.H.H., Kamran, M.A., Wajid, S.A., M. Amin, A.K., Arshad, M.N., Cheema, M.J.M., Saqib, Z.A., Ullah, R., Ziaf, K., Huq, A. ul, Ahmad, S., Fahad, M., Waqas, M.M., Abbas, A., Iqbal, A., 2019. Agro-Ecological Zones of Punjab- Pakistan. FAO.

Al-Ahmadi, F.S., Hames, A.S., 2009. Comparison of four classification methods to extract land use and land cover from raw satellite images for some remote arid areas, Kingdom of Saudi Arabia. Journal of King Abdulaziz University, Earth Sciences 20, 167–191. https://doi.org/10.4197/Ear.20-1.9

Ali, A., de Bie, C.A.J.M., Scarrott, R.G., Ha, N.T.T., Skidmore, A.K., 2012. COMPARATIVE PERFORMANCE ANALYSIS of A HYPER-TEMPORAL NDVI ANALYSIS APPROACH and A LANDSCAPE-ECOLOGICAL MAPPING APPROACH. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 1, 105–110. https://doi.org/10.5194/isprsannals-I-7-105-2012

Bairagi, G.D., Hassan, Z.U., 2002. Wheat crop production estimation using satellite data. Journal of the Indian Society of Remote Sensing 30, 213–219. https://doi.org/10.1007/BF03000364

Becker-Reshef, I., Justice, C., Barker, B., Humber, M., Rembold, F., Bonifacio, R., Zappacosta, M., Budde, M., Magadzire, T., Shitote, C., Pound, J., Constantino, A., Nakalembe, C., Mwangi, K., Sobue, S., Newby, T., Whitcraft, A., Jarvis, I., Verdin, J., 2020. Strengthening agricultural decisions in countries at risk of food insecurity: The GEOGLAM Crop Monitor for Early Warning. Remote Sensing of Environment 237, 111553. https://doi.org/10.1016/j.rse.2019.111553

Beltran-Abaunza, J., 2009. Method development to process hyper-temporal remote sensing (RS) images for change mapping. 2009 MSc theses GEM 53.

Beza, E.A., 2017. Citizen science and remote sensing for crop yield gap analysis.

Carfagna, E., Gallego, F.J., 2005. Using remote sensing for agricultural statistics. International Statistical Review. https://doi.org/10.1111/j.1751-5823.2005.tb00155.x

Chen, Y., Hou, J., Huang, C., Zhang, Y., Li, X., 2021. Mapping maize area in heterogeneous agricultural landscape with multi-temporal sentinel-1 and sentinel-2 images based on random forest. Remote Sensing 13, 1–22. https://doi.org/10.3390/rs13152988

Crop Reporting Service, Punjab [WWW Document], n.d. URL http://www.crs.agripunjab.gov.pk/ (accessed 6.16.22).

de Bie, C.A.J.M., 2002. Yield gap studies through comparative performance evaluation. Commission VII, Working Group VII/2.1 on Sustainable Agriculture, Pre-Symposium Tutorial 11.

de Bie, C.A.J.M., Khan, M.R., Smakhtin, V.U., Venus, V., Weir, M.J.C., Smaling, E.M.A., 2011. Analysis of multi-temporal SPOT NDVI images for small-scale land-use mapping. International Journal of Remote Sensing 32, 6673–6693. https://doi.org/10.1080/01431161.2010.512939

de Bie, C.A.J.M., Nelson, A.D., 2021. Next generation crop production analytics: Dynamic area sampling frames for improved crop analytics. Enabling Crop Analytics at Scale.

de Bie, K., 2020. W2 - Procedures and Tools to Process NDVI-Images: Spatio-temp. Analysis RS for food&water (2020-2B) [WWW Document]. URL https://canvas.utwente.nl/courses/7448/pages/w2-procedures-and-tools-to-process-ndvi-images?module_item_id=219955 (accessed 6.17.22).

de Bie, K., n.d. Hyper-temporal RS: EO for Natural Resources Management (2020-2A) [WWW Document]. URL https://canvas.utwente.nl/courses/7432/pages/06b-hyper-temporal-rs-day-1am-the-intro-lecture-ppt?module_item_id=212549 (accessed 6.13.22).

De Bie, K., Venus, V., Skidmore, A.K., 2011. Improved mapping and monitoring with hyper-temporal imagery. Proceedings of the internetional workshop on advanced use of satellite - and geo - information for agricultural and environmental intelligence 122–133.

Dehkordi, P.A., Nehbandani, A., Hassanpour-bourkheili, S., Kamkar, B., 2020. Yield Gap Analysis Using Remote Sensing and Modelling Approaches: Wheat in the Northwest of Iran. International Journal of Plant Production 14, 443–452. https://doi.org/10.1007/s42106-020-00095-4

Delegido, J., Verrelst, J., Alonso, L., Moreno, J., 2011. Evaluation of sentinel-2 red-edge bands for empirical estimation of green LAI and chlorophyll content. Sensors. https://doi.org/10.3390/s110707063

FAO, 2017. The future of food and agriculture: trends and challenges. Rome.

FAO, 2015. HANDBOOK ON Master Sampling Frames for Agricultural Statistics Frame Development, Sample Design, and Estimation.

FAO, 1978. Report on the agro-ecological zones project. FAO, Rome.

Fieuzal, R., Baup, F., Marais-Sicre, C., 2013. Monitoring Wheat and Rapeseed by Using Synchronous Optical and Radar Satellite Data—From Temporal Signatures to Crop Parameters Estimation. Advances in Remote Sensing 02, 162–180. https://doi.org/10.4236/ars.2013.22020

Fieuzal, R., Marais Sicre, C., Baup, F., 2017. Estimation of Sunflower Yield Using a Simplified Agrometeorological Model Controlled by Optical and SAR Satellite Data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10, 5412–5422. https://doi.org/10.1109/JSTARS.2017.2737656

GOP, 2010. Agriculture Census of Punjab 2010. 1103.

Gragn, Y.G., 2021. Mapping Arable Field Fractions with Muultisensor Remote Sensing Data-Driven Gradient Boosted and Classical GAMs Mapping Arable Field Fractions with Multisensor Remote Sensing Data-Driven Gradient Boosted and Classical GAMs.

Holtgrave, A.K., Röder, N., Ackermann, A., Erasmi, S., Kleinschmit, B., 2020. Comparing Sentinel-1 and -2 data and indices for agricultural land use monitoring. Remote Sensing 12. https://doi.org/10.3390/RS12182919

Hunt, M.L., Blackburn, G.A., Carrasco, L., Redhead, J.W., Rowland, C.S., 2019. High resolution wheat yield mapping using Sentinel-2. Remote Sensing of Environment 233, 111410. https://doi.org/10.1016/j.rse.2019.111410

Imran, A., 2019. Wheat Crop Development in Central Punjab.

J, Anwar, MH, T., AW, S., J, Ahmad, M, S., 2019. A new high yielding durable rust resistant variety named Galaxy-2013 for the irrigated areas of Punjab, Pakistan. Academia Journal of Agricultural Research.

Jain, M., Srivastava, A.K., Balwinder-Singh, Joon, R.K., McDonald, A., Royal, K., Lisaius, M.C., Lobell, D.B., 2016. Mapping smallholder wheat yields and sowing dates using micro-satellite data. Remote Sensing 8, 1–18. https://doi.org/10.3390/rs8100860

Jin, Z., Azzari, G., You, C., Di Tommaso, S., Aston, S., Burke, M., Lobell, D.B., 2019. Smallholder maize area and yield mapping at national scales with Google Earth Engine. Remote Sensing of Environment 228, 115–128. https://doi.org/10.1016/j.rse.2019.04.016

Kang, Y., Özdoğan, M., 2019. Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. Remote Sensing of Environment 228, 144–163. https://doi.org/10.1016/j.rse.2019.04.005

Kayad, A., Sozzi, M., Gatto, S., Marinello, F., Pirotti, F., 2019. Monitoring within-field variability of corn yield using sentinel-2 and machine learning techniques. Remote Sensing 11. https://doi.org/10.3390/rs11232873

Khabbazan, S., Vermunt, P., Steele-Dunne, S., Arntz, L.R., Marinetti, C., van der Valk, D., Iannini, L., Molijn, R., Westerdijk, K., van der Sande, C., 2019. Crop monitoring using Sentinel-1 data: A case study from The Netherlands. Remote Sensing 11, 1–24. https://doi.org/10.3390/rs11161887

Khan, I., Lei, H., Khan, Ahmad, Muhammad, I., Javeed, T., Khan, Asif, Huo, X., 2021. Yield gap analysis of major food crops in Pakistan: prospects for food security. Environmental Science and Pollution Research. https://doi.org/10.1007/s11356-020-11166-4

Khan, M.R., de Bie, C.A.J.M., van Keulen, H., Smaling, E.M.A., Real, R., 2010. Disaggregating and mapping crop statistics using hypertemporal remote sensing. International Journal of Applied Earth Observation and Geoinformation 12, 36–46. https://doi.org/10.1016/j.jag.2009.09.010

Lambert, M.J., Blaes, X., Traore, P.S., Defourny, P., 2017. Estimate yield at parcel level from S2 time serie in sub-Saharan smallholder farming systems. 2017 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images, MultiTemp 2017. https://doi.org/10.1109/Multi-Temp.2017.8035204

Larranaga, A., Alvarez-Mozos, J., Albizua, L., Peters, J., 2013. Backscattering behavior of rain-fed crops along the growing season. IEEE Geoscience and Remote Sensing Letters 10, 386–390. https://doi.org/10.1109/LGRS.2012.2205660

Lobell, D.B., Azzari, G., Burke, M., Gourlay, S., Jin, Z., Kilic, T., Murray, S., 2020. Eyes in the Sky, Boots on the Ground: Assessing Satellite- and Ground-Based Approaches to Crop Yield Measurement and Analysis. American Journal of Agricultural Economics. https://doi.org/10.1093/ajae/aaz051

Lobell, D.B., Cassman, K.G., Field, C.B., 2009a. Crop Yield Gaps: Their Importance, Magnitudes, and Causes. http://dx.doi.org/10.1146/annurev.environ.041008.093740 34, 179–204. https://doi.org/10.1146/ANNUREV.ENVIRON.041008.093740

Lobell, D.B., Cassman, K.G., Field, C.B., 2009b. Crop yield gaps: Their importance, magnitudes, and causes. Annual Review of Environment and Resources 34, 179–204. https://doi.org/10.1146/annurev.environ.041008.093740

Lowder, S.K., Skoet, J., Raney, T., 2016. The Number, Size, and Distribution of Farms, Smallholder Farms, and Family Farms Worldwide. World Development 87, 16–29. https://doi.org/10.1016/j.worlddev.2015.10.041

Mateo-Sanchis, A., Piles, M., Muñoz-Marí, J., Adsuara, J.E., Pérez-Suay, A., Camps-Valls, G., 2019. Synergistic integration of optical and microwave satellite data for crop yield estimation. Remote Sensing of Environment 234, 111460. https://doi.org/10.1016/j.rse.2019.111460

Mc Carthy, U., Uysal, I., Badia-Melis, R., Mercier, S., O'Donnell, C., Ktenioudaki, A., 2018. Global food security – Issues, challenges and technological solutions. Trends in Food Science and Technology 77, 11–20. https://doi.org/10.1016/j.tifs.2018.05.002

Mohammed, I., 2019. Mapping Crop Field Probabilities Using Hyper Temporal and Multi Spatial Remote Sensing in a Fragmented Landscape of 1–63.

Mohammed, I., Marshall, M., de Bie, K., Estes, L., Nelson, A., 2020. A blended census and multiscale remote sensing approach to probabilistic cropland mapping in complex landscapes. ISPRS Journal of Photogrammetry and Remote Sensing 161, 233–245. https://doi.org/10.1016/j.isprsjprs.2020.01.024

Mojid, M.A., Mousumi, K.A., Ahmed, T., 2020. Performance of Wheat in Five Soils of Different Textures under Freshwater and Wastewater Irrigation. Agricultural Science 2, p89. https://doi.org/10.30560/AS.V2N2P89

Mudasser, M., Hussain, I., Aslam, M., 2001. CONSTRAINTS TO LAND-AND WATER PRODUCTIVITY OF WHEAT IN INDIA AND PAKISTAN: A COMPARATIVE ANALYSIS International Water Management Institute (IWMI).

Mukhtar, U., Zhangbao, Z., Beihai, T., Naseer, M.A.U.R., Razzaq, A., Hina, T., 2018. Implications of Decreasing Farm Size on Urbanization: A Case Study of Punjab Pakistan. Journal of Social Science Studies 5, 71. https://doi.org/10.5296/JSSS.V5I2.12746

Next generation crop production analytics: Dynamic area sampling frames for improved crop analytics — University of Twente Research Information [WWW Document], n.d. URL https://research.utwente.nl/en/publications/next-generation-crop-production-analytics-dynamic-area-sampling-f (accessed 6.15.22).

Oad, F.C., Siddiqui, M.H., Buriro, U.A., 2007. Growth and yield losses in wheat due to different weed densities. Asian Journal of Plant Sciences 6, 173–176. https://doi.org/10.3923/AJPS.2007.173.176

Oto, L.H.D.E., 2017. Exploring an alternative approach for deriving NDVI-based forage scarcity in the framework of index-based livestock insurance in East Africa Exploring an alternative approach for deriving NDVI-based forage scarcity in the framework of index-based livestoc.

Pan, Y., Wang, M., Wei, G., Wei, F., Shi, K., Li, L., Sun, G., 2010. Application of Area-frame sampling for agricultural statistics in China, in: Proceedings of the Fifth International Conference on Agiculture Statistics (ICAS-V). FAO, Rome.

Payne, T., n.d. Harvest and storage management of wheat [WWW Document]. URL https://www.fao.org/3/Y4011E/y4011e0u.htm (accessed 2.24.22).

Qayyum, A., Jamil Shera, H.M.M., 2019. Method of Area Frame Sampling Using Probability Proportional to Size Sampling Technique for Crops' Surveys: A Case Study in Pakistan. Journal of Experimental Agriculture International 41, 1–10. https://doi.org/10.9734/jeai/2019/v41i230395

Rattalino Edreira, J.I., Andrade, J.F., Cassman, K.G., van Ittersum, M.K., van Loon, M.P., Grassini, P., 2021. Spatial frameworks for robust estimation of yield gaps. Nature Food 2, 773–779. https://doi.org/10.1038/s43016-021-00365-y

Ricciardi, V., Ramankutty, N., Mehrabi, Z., Jarvis, L., Chookolingo, B., 2018. How much of the world's food do smallholders produce? Global Food Security 17, 64–72. https://doi.org/10.1016/j.gfs.2018.05.002

Scarrott, R.G., 2022. OCEAN-SURFACE HETEROGENEITY MAPPING: EXPLOITING HYPERTEMPORAL DATASETS IN SUPPORT OF SEASCAPE ECOLOGY RESEARCH. University College Cork, Cork, Ireland.

Skakun, S., Kalecinski, N.I., Brown, M.G.L., Johnson, D.M., Vermote, E.F., Roger, J.C., Franch, B., 2021. Assessing within-field corn and soybean yield variability from worldview-3, planet, sentinel-2, and landsat 8 satellite imagery. Remote Sensing 13, 1–18. https://doi.org/10.3390/rs13050872

Trivedi, M.B., 2020. MAPPING PROBABILITIES OF ARABLE FIELDS USING MODIS , SENTINEL-1 AND SENTINEL-2 BASED IMAGE FEATURES IN GHANA SIS ] MAPPING PROBABILITIES OF ARABLE FIELDS USING MODIS , SENTINEL-1 AND SENTINEL-2 BASED IMAGE FEATURES IN GHANA SIS ].

ÜNAL, E., KEES, D.B., 2017. Zaman Serisi NDVI Verileri ve Resmi Tarım İstatistikleri Kullanarak Türkiye Buğday Alanlarının Haritalandırılması. Tarla Bitkileri Merkez Araştırma Enstitüsü Dergisi 26, 11–11. https://doi.org/10.21566/tarbitderg.323560

United Nations, 2021. The Sustainable Development Goals Report 2021, United Nations publication issued by the Department of Economic and Social Affairs.

Veloso, A., Mermoz, S., Bouvet, A., Toan, T. le, Dejoux, J., Ceschia, E., Veloso, A., Mermoz, S., Bouvet, A., Toan, T. le, Planells, M., 2021. Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications To cite this version : HAL Id : hal-03272371 Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for ag.

Wheat Program [WWW Document], n.d. URL http://www.parc.gov.pk/index.php/en/faq-s/60-faqs/270-faqs-wheat (accessed 6.17.22).

Wikipedia [WWW Document], n.d. URL https://en.wikipedia.org/wiki/Districts_of_Pakistan (accessed 6.10.22).

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D.B., Huang, Y., Huang, M., Yao, Y., Bassu, S., Ciais, P., Durand, J.L., Elliott, J., Ewert, F., Janssens, I.A., Li, T., Lin, E., Liu, Q., Martre, P., Müller, C., Peng, S., Peñuelas, J., Ruane, A.C., Wallach, D., Wang, T., Wu, D., Liu, Z., Zhu, Y., Zhu, Z., Asseng, S., 2017. Temperature increase reduces global yields of major crops in four independent estimates. Proc Natl Acad Sci U S A 114, 9326–9331. https://doi.org/10.1073/pnas.1701762114

# ANNEXES

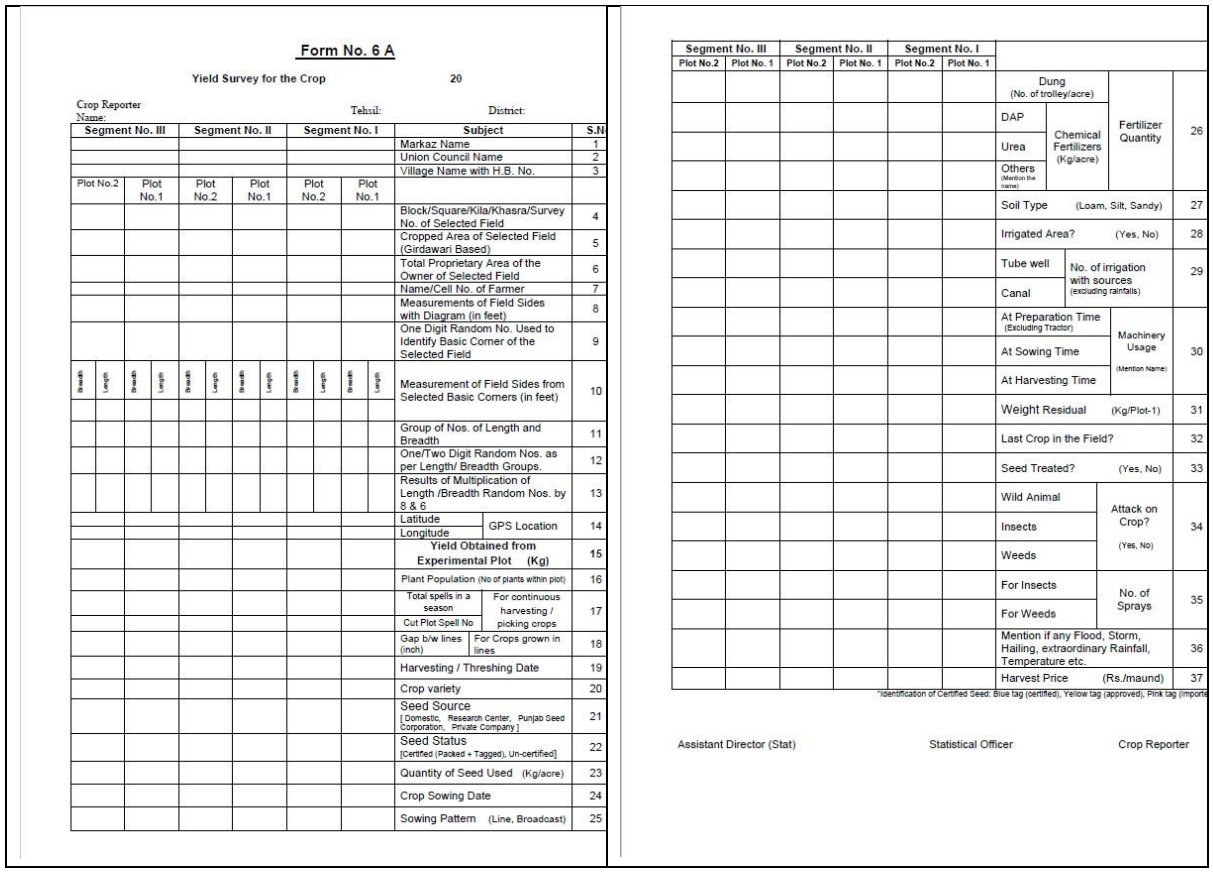
Annex 1: R-code used to produce Violin Plots.

```
#R-code to produce Violin Plots
# Libraries
library(ggplot2)
library(dplyr)
library(hrbrthemes)
library(viridis)

#Set working Directory
setwd("D:\\ITC_Study\\M.Sc Plans\\Data\\Current_Files")
d<-read.csv("Wheat Yield (6x8 ft Plot) Punjab 2020_V3_copy1.csv",header=TRUE, sep=",")

# create a dataset
d <- data.frame(
  Legend= d$(Var Name on X axis),
  value= d$(Var Name on y axis)
)


# sample size
sample_size = d %>% group_by(Legend) %>% summarize(num=n())

# Plot
d %>%
  left_join(sample_size) %>%
  mutate(Legend = paste0(Legend, "\n", "n=", num)) %>%
  ggplot( aes(x=Legend, y=value, fill= Legend)) +
  geom_violin(trim = FALSE) +
  geom_boxplot(width=0.1, color="black", alpha=0.5) +
  scale_fill_viridis(discrete = TRUE) +
  scale_fill_brewer(palette="Set2") +
  theme(
    legend.position = "none",
    axis.title = element_text(size=15),
    axis.text.x = element_text(size= 12, angle = 45, vjust = 1, hjust = 1),
    axis.text.y = element_text(size = 12),
  ) +

  #ggtitle("Title Name") +
  xlab("Label Name") + ylab("Label Name")+ ylim(0,6000)
```

Annex 2: Code used to produce Boxplots.

```
#Libraraies
library(tidyverse)
library(hrbrthemes)
library(viridis)

#Set working directory and call the csv file
setwd("D:\\ITC_Study\\M.Sc Plans\\Data\\Current_Files")
d<-read.csv("Wheat Yield (6x8 ft Plot) Punjab 2020_V3_copy1.csv",header=TRUE, sep=",")

# create a dataset
data <- data.frame(
  name=d$(Var name on x-axis),
  value=d$(Var name on y-axis)
)

# sample size
sample_size = data %>% group_by(name) %>% summarize(num=n())

# Plot
data %>%
  left_join(sample_size) %>%
  mutate(name = paste0(name, "\n", "n=", num)) %>%
  ggplot( aes(x=name, y=value, fill=name)) +
  geom_boxplot() +
  scale_fill_viridis(discrete = TRUE, alpha=0.6) +
  theme_ipsum() +
  theme(
    legend.position="none",
    plot.title = element_text(size=12),
    axis.title.x = element_text(size=13),
    axis.title.y = element_text(size=13),
    axis.text.x = element_text(size =10, angle = 45, vjust = 1, hjust = 1),
  ) +
  ggtitle("Title name") +
  xlab("Var name") + ylab("Var name")
```

# Annex 3: Yield Survey Form



**Form No. 6 A**

**Yield Survey for the Crop** 20

Crop Reporter Name: Tehsil: District:

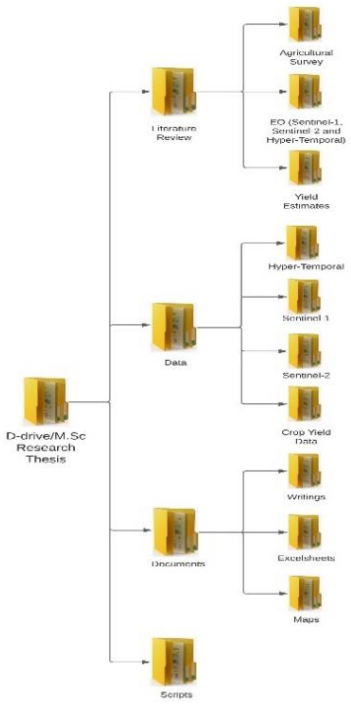
| Segment No. III | | | | Segment No. II | | | | Segment No. I | | | | Subject | S.No |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | Markaz Name | 1 |
| | | | | | | | | | | | | Union Council Name | 2 |
| | | | | | | | | | | | | Village Name with H.B. No. | 3 |
| Plot No.2 | | Plot No.1 | | Plot No.2 | | Plot No.1 | | Plot No.2 | | Plot No.1 | | | |
| | | | | | | | | | | | | Block/Square/Kila/Khasra/Survey No. of Selected Field | 4 |
| | | | | | | | | | | | | Cropped Area of Selected Field (Girdawari Based) | 5 |
| | | | | | | | | | | | | Total Proprietary Area of the Owner of Selected Field | 6 |
| | | | | | | | | | | | | Name/Cell No. of Farmer | 7 |
| | | | | | | | | | | | | Measurements of Field Sides with Diagram (in feet) | 8 |
| | | | | | | | | | | | | One Digit Random No. Used to Identify Basic Corner of the Selected Field | 9 |
| Breadth | Length | Breadth | Length | Breadth | Length | Breadth | Length | Breadth | Length | Breadth | Length | Measurement of Field Sides from Selected Basic Corners (in feet) | 10 |
| | | | | | | | | | | | | Group of Nos. of Length and Breadth | 11 |
| | | | | | | | | | | | | One/Two Digit Random Nos. as per Length/ Breadth Groups. | 12 |
| | | | | | | | | | | | | Results of Multiplication of Length /Breadth Random Nos. by 8 & 6 | 13 |
| | | | | | | | | | | | | Latitude / Longitude — GPS Location | 14 |
| | | | | | | | | | | | | **Yield Obtained from Experimental Plot (Kg)** | 15 |
| | | | | | | | | | | | | Plant Population (No of plants within plot) | 16 |
| | | | | | | | | | | | | Total spells in a season / Cut Plot Spell No — For continuous harvesting / picking crops | 17 |
| | | | | | | | | | | | | Gap b/w lines (inch) — For Crops grown in lines | 18 |
| | | | | | | | | | | | | Harvesting / Threshing Date | 19 |
| | | | | | | | | | | | | Crop variety | 20 |
| | | | | | | | | | | | | Seed Source [ Domestic, Research Center, Punjab Seed Corporation, Private Company ] | 21 |
| | | | | | | | | | | | | Seed Status [Certified (Packed + Tagged), Un-certified] | 22 |
| | | | | | | | | | | | | Quantity of Seed Used (Kg/acre) | 23 |
| | | | | | | | | | | | | Crop Sowing Date | 24 |
| | | | | | | | | | | | | Sowing Pattern (Line, Broadcast) | 25 |

| Segment No. III | | Segment No. II | | Segment No. I | | | | |
|---|---|---|---|---|---|---|---|---|
| Plot No.2 | Plot No. 1 | Plot No.2 | Plot No. 1 | Plot No.2 | Plot No. 1 | | | |
| | | | | | | Dung (No. of trolley/acre) | Fertilizer Quantity | 26 |
| | | | | | | DAP / Urea / Others (Mention the name) — Chemical Fertilizers (Kg/acre) | | |
| | | | | | | Soil Type (Loam, Silt, Sandy) | | 27 |
| | | | | | | Irrigated Area? (Yes, No) | | 28 |
| | | | | | | Tube well / Canal — No. of irrigation with sources (excluding rainfalls) | | 29 |
| | | | | | | At Preparation Time (Excluding Tractor) / At Sowing Time / At Harvesting Time — Machinery Usage (Mention Name) | | 30 |
| | | | | | | Weight Residual (Kg/Plot-1) | | 31 |
| | | | | | | Last Crop in the Field? | | 32 |
| | | | | | | Seed Treated? (Yes, No) | | 33 |
| | | | | | | Wild Animal / Insects / Weeds — Attack on Crop? (Yes, No) | | 34 |
| | | | | | | For Insects / For Weeds — No. of Sprays | | 35 |
| | | | | | | Mention if any Flood, Storm, Hailing, extraordinary Rainfall, Temperature etc. | | 36 |
| | | | | | | Harvest Price (Rs./maund) | | 37 |

*Identification of Certified Seed: Blue tag (certified), Yellow tag (approved), Pink tag (imported

Assistant Director (Stat) Statistical Officer Crop Reporter

Annex 4: Data Management Plan

| Data Management Plan | |
|---|---|
| Student | Sardar Salar Saeed Dogar |
| Student ID | S2268922 |
| Course | M-Geo, NRM, 2020-2022 |
| Research Theme | FORAGES |
| Supervisors | 1) Dr. IR. C.A.J.M. De Bie<br>2) Valentijn Venus |
| Research Title | Integrating Earth Observation Data into Area Frame Sampling Approach to Improve Crop Production Estimates |
| Research Description | This research aims is to capture yield variability and produce accurate crop yield estimates through integration of multi-sensor Earth Observation data into existing Area Frame Sampling Approach. |
| Research Duration | November 2021 to June 2022 |
| Research Data Manager | Salar Saeed with supervision of Associate Prof. Dr. IR. C.A.J.M. De Bie and Lecturer Valentijn Venus |
| Date of This Plan | 20/01/2022 |
| **Research Data** | |
| File Naming (Standards) | All the files and folders under this research will have a meaningful name, followed by the date and the version. For example: typeofdata_yymmdd_v# |
| Data Privacy | The data collected from Government Department will solely use for this research only. |
| Backup Plan | For the safety of the data under this research, copy of the data will store on the office cloud (Microsoft OneDrive) |
| Data Storage and Organization |  |