scientific reports

Check for updates

OPEN

TPTC: topic-wise problems' trend clusters for smart agricultural insights extraction and forecasting of farmer's information demand

Samarth Godara[1], Shbana Begam[2]✉, Ram Swaroop Bana[3]✉, Jatin Bedi[4], Rajni Jain[5], Md. Ashraful Haque[1], Rajender Parsad[1], Sudeep Marwaha[1], Madhu Patial[3], Saber Shirzad[3,6]✉ & Ravi Nirmal[3]

To meet the challenges of increasing food production demand globally, extracting insights regarding the persistent agriculture-related problems on a nationwide scale is the need of the hour. Policymakers now have limited possibilities for acquiring a comprehensive knowledge of the difficulties that farmers face on a national level. In this direction, the presented work proposes a new artificial intelligence-based pipeline to gain insights at country level regarding the farmers' demand for assistance in India. The presented study uses the data from the Kisan Call Centres, a nationwide network of farmer's helplines, including 28.6 million call-log records, made available by the Ministry of Agriculture & Farmers' Welfare, Government of India. Additionally, the extracted insights are presented in the form of "Topic-wise Problems' Trend Clusters" (TPTC), which can be used by policymakers in both the government and private sectors to aid decision-making. The article also introduces a pipeline for designing forecasting models to estimate the monthly frequency of farmer inquiries (in terms of the number of query calls). The seven statistical forecasting models were examined in the study with the TBATP1 (Trigonometric seasonal components with Box-Cox transformation incorporating ARIMA errors and Trend including the Seasonal components) model attaining the lowest error rates in terms of Root Mean Square Error (0.034) and Mean Absolute Error (0.107). The study also explores numerous applications of the derived insights in the real world as well as the future scope of the presented work.

Keywords Agricultural problems analytics, Artificial intelligence in agriculture, Forecasting, Helpline centre data, Machine learning.

Agriculture has always played a key role in India's socio-economic growth that generates around one-third of the nations' GDP[1]. Nonetheless, currently, the yields are not only lower than predicted but also unstable, owing to the climate change and technological transfer gaps[2]. Policymakers need to have broad perspectives on farmers' issues in order to reduce the national yield gap. Nevertheless, the Indian government periodically conducts surveys to better understand the state of agriculture and farmers in preparation for the implementation of new policies. These surveys are prolonged and expensive to provide the information that too conducted (not more than) once a year. Furthermore, most of the large-scale surveys conducted by the government are production-oriented rather than being agricultural problems-oriented (Comprehensive Scheme for Studying the Cost of Cultivation of Principal Crops (CoC), Directorate of Economics and Statistics (DEC), National Sample Survey Organization data (NSSO), etc.). In this direction, a study has been designed using Artificial Intelligence (AI) techniques with the following two objectives:

1. Obtaining novel insights to deliver information regarding the most frequent increasing and decreasing agricultural problems in India over the previous years, and.

[1]ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India. [2]ICAR-National Institute for Plant Biotechnology, New Delhi, India. [3]ICAR- Indian Agricultural Research Institute, New Delhi, India. [4]Thapar Institute of Engineering And Technology, Patiala, Punjab, India. [5]ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi, India. [6]JACK, Kabul, Afghanistan. ✉email: shbana.begam@icar.gov.in; rsbana@gmail.com; sabershirzad100@gmail.com

2. Designing forecasting models to predict the monthly farmers' need for assistance in the target area on the selected agricultural issue.

The proposed approach solves the existing issues by clustering agricultural problems based on their time-series trends, enabling policymakers to identify and address the most persistent and emerging issues across regions and crops. In order to achieve the defined objectives, in this work, we use data from the Kisan Call Centres (KCC), a program initiated by the Department of Agriculture & Cooperation (DAC), Ministry of Agriculture, Government of India, on January 21, 2004, across the country to provide extension services to farmers. The data generated through the KCCs includes over 28.6 million call-log records corresponding to the query calls made by the Indian farmers over the past eight years since March 2013. These knowledge centers operate across India, assisting and guiding farmers in their local and regional languages in order to solve problems over phone calls. This facilitates the rapid diffusion of technology among farmers, identifying location-specific problems and developing location-specific solutions by allowing access to the timeline of farmer questions. KCCs generate a large amount of factual information directly from farmers, that is already been utilized to extract valuable insights and understand the agriculture-related issues in India so that timely management can be done on both the national and local levels[3,4].

In the present work, first, the query call-log data is fetched from the KCC servers and processed using a number of critical data mining pre-processing. In the next phase, for extracting the insights, we introduce the concept of Topic-wise Problems' Trend-based Clusters (TPTC), where, the yearly time-series corresponding to the most common agricultural problems are extracted initially. Later, using the linear regression integrated K-modes clustering algorithms, clusters are extracted and results are visualized. Here, linear regression is used to identify whether the agricultural problem trend is increasing or decreasing over time, while K-modes clustering groups hundreds of agricultural problems based on these trends. The TPTCs are extracted from a total of 11,836 agricultural problems i.e., a combination of all the inquired crops from all the Indian states regarding top four major topics including seeds/varieties, weed management, fertilizer usage, and plant protection. For developing the forecasting models, first, the monthly query-call time series are extracted. Later, seven statistical models, including ARIMA, Prophet, TATS, TBAT, TBATS1, TBATP1, and TBATS2, are used to make predictions. In order to compare the forecasting performances of the models, two metrics, including RMSE and MAE are considered. Furthermore, to compare the performance of all these forecasting models, total 100 time series (a combination of top 5 crops of top 5 states with 4 topics) considered in the experiment.

The outputs of the proposed pipelines reveal promising results that might be leveraged to offer new perspectives to decision-makers. Aside from that, the framework offers a new way to obtain explicit information regarding the activities taking place across the entire agriculture sector. Further, the models are helpful in developing intelligent systems such as prediction models, early warning, recommender systems, market predictions and many more.

The KCC scheme of Govt of India, is rapidly becoming an important tool for technology transmission in agriculture and allied sectors, and it is a reliable and simple source of agriculture-related information. In this scheme, a toll-free number '1800–180–1551' is provided for 24×7 support to the farmers for their agriculture-related problems. The calls information made available by the helpline centers can be used to obtain many important insights regarding the problems faced by Indian farmers. In recent years, several attempts were made on the KCC data using various computational techniques. Chouhan et al.[5] conducted a study on the KCC dataset for the Bhopal district of Madhya Pradesh to get the monthly frequency of discipline-wise query calls. They also studied and analyzed the constraints while making answers for the queries. Viswanath et al.[3] used Hadoop-based MapReduce algorithms to analyze three years of KCC data (2015–17) to extract intriguing insights such as the crops that have been questioned the most by farmers and the hour when the most calls are made. Aside from that, the authors used Natural Language Processing (NLP) to group similar queries in order to figure out which one farmers commonly ask.

Kavitha and Anandaraja[6] examined the KCC data patterns by district, sector, crop, and topic-wise. They reported that the highest number of calls were received from the Warangal and Mahaboobnagar districts of Maharashtra. The objective of the study was to assist the Agriculture Extension Centers (AECs) in facilitating and improving the Transfer of Technology process. Godara and Toshniwal[7] proposed a new approach that uses association rule mining integrated with a multi-criteria decision-making technique (TOPSIS) to extract only the most relevant patterns from the KCC dataset. In 2022, Godara and Toshniwal[7] presented several machine learning and deep learning-based models to forecast the futuristic query-call counts from the KCC datasets.

Some studies based on the KCC dataset have been observed to advance the process of KCC. Mohapatra and Upadhyay[8,9] developed a model to generate query responses in text format using NLP on the KCC dataset. The authors achieved this goal by incorporating Latent Dirichlet al.location (LDA) and Latent Semantic Indexing (LSI) into the TF-IDF model's pipeline. They expanded on their work by training a model to extract query information based on the similarity of query sentences and then finding the best possible answer based on the similarity. To detect similar searches, the term-frequency-inverse document frequency (TF-IDF) method was utilized. Although existing research does extract some insights from the KCC helpline data, these studies are mostly focused on improving the present KCC model and do not provide policymakers with useful information.

Arora et al.[10] developed a Long Short Term Memory (LSTM) technique-based natural language generative chatbot namely "Agribot" for farmers to provide electronic message service in regional language. Ajawan et al.[11] developed "Smart Sampark" an automatic responsive model for KCC. In this study, experiments were conducted with over 100 questions from the KCC dataset. For each test query, the five most related responses based on the cosine similarity were considered. Momaya et al.[12] developed a farmers chatbot "Krushi" using the KCC dataset. This is an end-to-end trainable learning model that can be used to build a conversational system with minimal error and respond to inquiries regarding current conditions.

Furthermore, various studies have been conducted to examine the impact of KCCs on farmers[6,13–15] and farmers' attitude towards it[16,17]. KCC queries related to animal husbandry[18] were analyzed to assess the information needed by farmers and livestock owners to develop specific information services for them. Despite these useful insights, most of the existing studies were conducted at the district, state or regional level, whereas country-level insights and information are required to formulate relevant policies. The current work focuses on national level insights including extracting common crop problems over several states, common problems over several crops in a single state, and many more. The following are the major research contributions of the presented study:

1. **National-Level Insights**: The current study offers a comprehensive analysis of agricultural problems faced by farmers at a national scale, considering a broader range of crops and regions, which can support policy-making efforts.
2. **Use of KCC Data for Problem-Oriented Analysis**: Unlike previous studies, which focused on production, this research emphasizes the identification of common agricultural problems through KCC data, helping policymakers to prioritize issues affecting farmers.
3. **Topic-wise Problem Trend Clusters (TPTC)**: The study introduces the novel concept of TPTC, which clusters agricultural issues based on their time-series trends, providing insights into the most frequent and emerging problems.
4. **Multi-Criteria Approach for Problem Identification**: The study uses a combination of top crops and topics (e.g., seeds/varieties, fertilizer usage) across multiple states to extract meaningful trends, offering a multi-dimensional understanding of farmers' issues.
5. **Comparison of Diverse Models**: By testing various forecasting models, the study not only provides insights into the best-performing models but also offers a systematic comparison, advancing the understanding of model suitability for agricultural problem forecasting.

Results

In this section, we discuss the obtained experimental insights and results from eight years of query call data recorded under the KCC scheme from March 2013 to November 2021 utilizing the proposed framework. The computations of the proposed methodology are executed with python 3.0 script on the Google Colab platform with dual Intel(R) Xeon(R) CPU @ 2.20 GHz microprocessor, 13GB RAM and 108GB disk space. Moreover, the outputs corresponding to each module are as follows:

Topic-wise problems' trend clustering

The first step in the extraction of the TPTCs is to obtain multiple yearly time series. In order to achieve this, we first chose the 294 crops (all the crops in the dataset) under the 32 Indian states and union territories while taking into account the top 4 query types (seeds and varieties, weed management, fertilizer use, and plant protection) in order to obtain yearly time-series data points for extracting the target insights. After the extraction, the empty time series are removed from the. Later, the remaining time-series are linearly regressed and the coefficient of determination is calculated. Figure 1 displays examples of a few time-series along with linear model examples and coefficient of determination.

Figure 1(a), (b) and (c) represent the linear relationship of weed management-related queries for paddy crop in West Bengal, fertilizer usage-related queries for wheat crop in Gujarat state, and fertilizer usage-related queries for cotton crop in Telangana state, respectively. From the figures, it is observed that the demand for assistance for these particular topics are increasing over the period of time. As shown in Fig. 1(a) and (b), since the data points are closely packed to the regression line, the coefficient of determination is also high, i.e. >0.8. Whereas, in Fig. 1(c), due to non-linear behavior of the data point, the coefficient of determination is comparatively less, i.e., 0.569.

Furthermore, Fig. 1(d), (e) and (f) represent the decreasing trends of fertilizer usage-related queries in Tamil Nadu for paddy crop, seeds and varieties-related queries in Haryana state for paddy crop, and fertilizer usage-related queries in Odisha state for chili crop, respectively. Since the coefficient of determination in Fig. 1(d) and (e) is higher than the cutoff value (>0.7), these problems are not discarded in the filtration process. Whereas, the problem mentioned in Fig. 1(f) is filtered out.

Upon investigating the obtained slopes of the extracted 11,836 agricultural problems, it was noted that the number of increasing problems in India (with positive slopes) in the past years is approximately the same as the decreasing problems (with depleting slopes) as shown in Fig. 2.

Furthermore, Fig. 3 represents the coefficient of determination values of all the considered problems. It is to be noted that since the coefficient value is set to be 0.7, all the agricultural problems below this value are discarded. Moreover, in this process, a total of 26,364 problems are discarded for not showing a strong linear trend (either in increasing or decreasing manner).

With the filtered agricultural problems, first the 'slope' attribute associated with each problem is converted into categorical values (Table 1, 'Slope' column). Later, the dataset is clustered using the K-modes algorithm discussed in the previous section. In the present study, the K-modes algorithm is executed with two inputs, i.e., the state, query-type, and slope are passed to the clustering algorithm as the properties of the data points in order to acquire the state-wise insights. Second, the crop, query-type, and slope are taken into account as the attributes to identify the clusters in order to gain the crop-wise insights.

A sample of the obtained clusters is given in Tables 1 and 2, meanwhile, the complete output is given with the supplementary information. The table has five columns, including.

- Cluster: defines the cluster number of the particular problem,

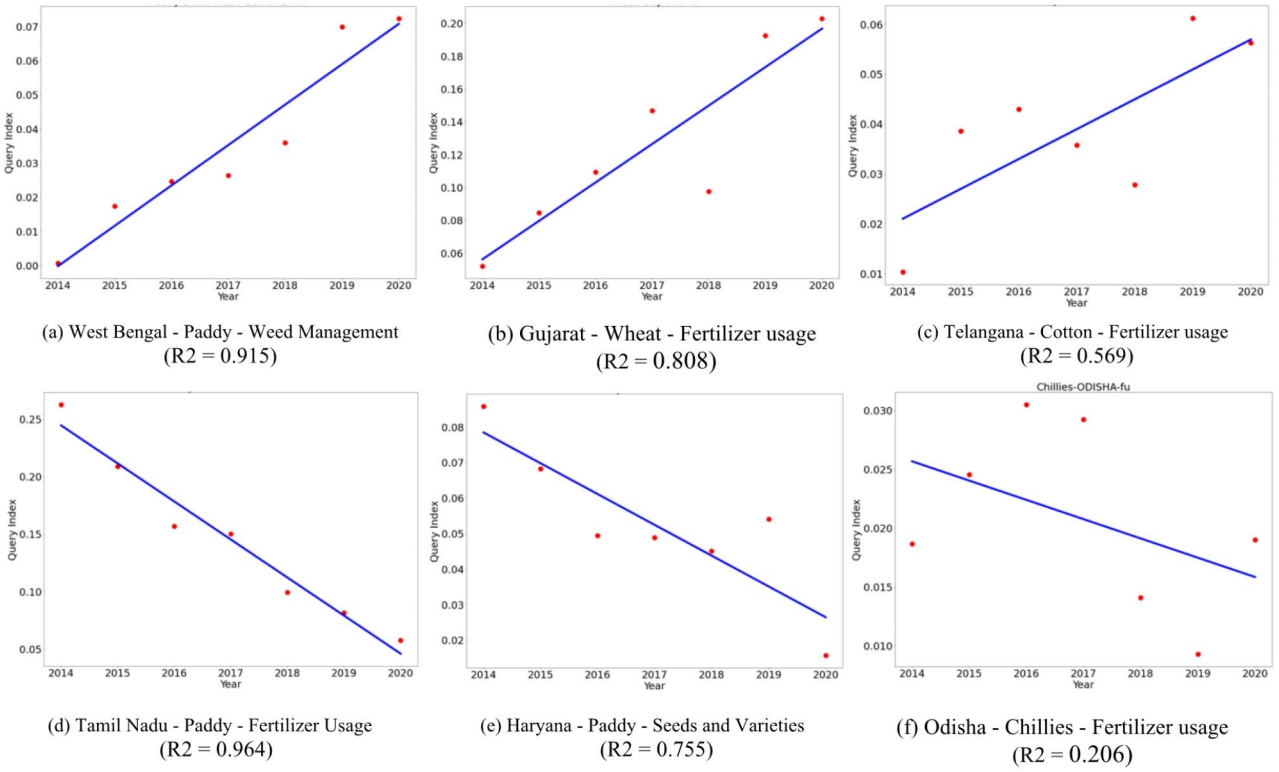


Fig. 1. Extracted yearly time series corresponding to various agricultural problems along with the coefficient of determination (R^2) of the linear regression model.

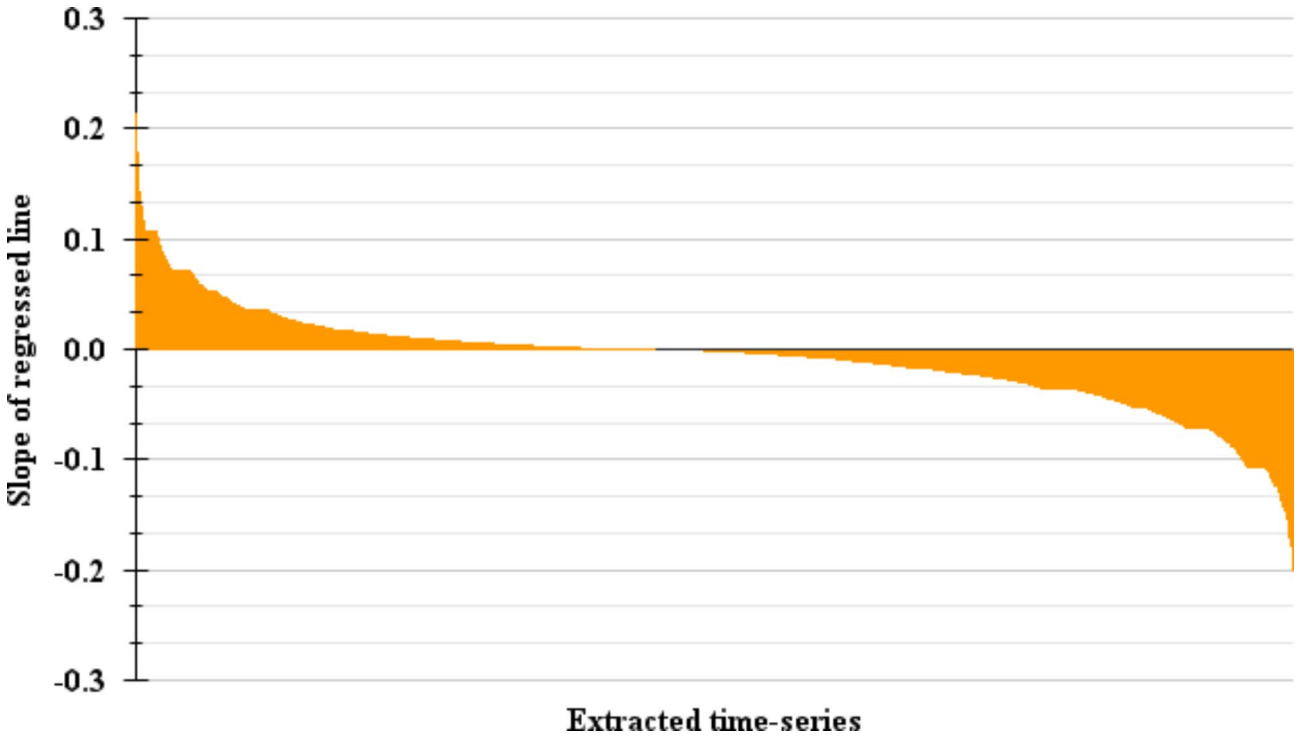


Fig. 2. Slope values corresponding to the extracted yearly time series. X-axis represents the slope values, Y-axis represents the corresponding yearly time series.

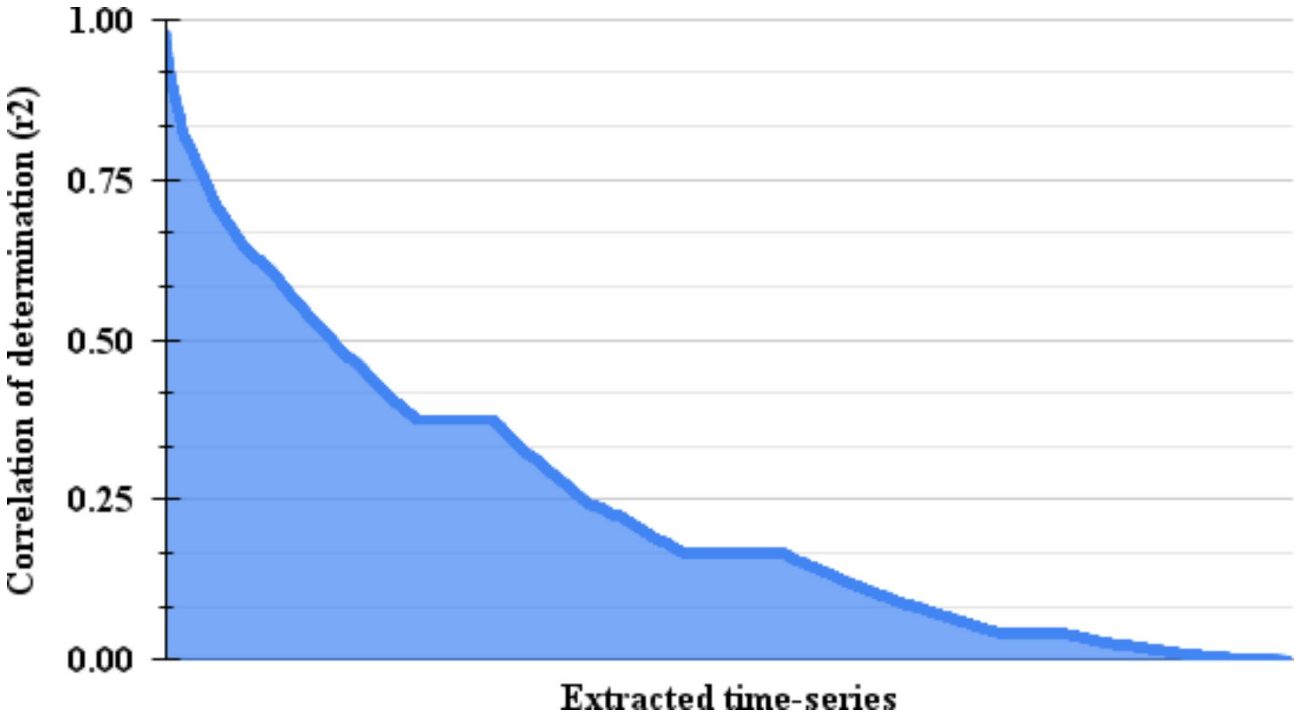


Fig. 3. Extracted agricultural problems against their respective correlation coefficients with the regressed line.

Cluster	Crop	State	Query type	Slope
2	Chili, Black Gram, Banana, Coconut, Bitter Gourd, Papaya, Pumpkin, Orange, Acid Lime, and Tuberose	West Bengal	Plant Protection	Decreasing Queries
0	Onion, Sugarcane, Tomato, Green Gram, Mango, Pigeon pea, Citrus, Coriander, Ginger, Berseem, Aonla, Melon	Uttar Pradesh	Fertilizer Usage	Increasing Queries

Table 1. TPTC state-wise insights output sample.

Cluster	State	Crop	Query type	Slope
2	Uttar Pradesh, Rajasthan, Madhya Pradesh, Uttarakhand	Tulsi	Plant Protection	Decreasing Queries
9	Uttar Pradesh, Jharkhand, Gujarat, Uttarakhand	Tomato	Fertilizer Usage	Increasing Queries
10	Punjab, Haryana, Rajasthan, Madhya Pradesh, Chhattisgarh, Delhi, Jammu And Kashmir	Wheat	Weed Management	Increasing Queries
21	Rajasthan, Gujarat, Maharashtra, Odisha, Madhya Pradesh, Chhattisgarh	Groundnut	Weed Management	Increasing Queries

Table 2. TPTC crop-wise insights output sample.

- Slope: defines if the problem is increasing or decreasing over past years,
- Query Type: type of the problem,
- Crop: the problem associated with which crop, and.
- State: the particular problem is observed in which Indian state.

As observed from Table 1, one of the two clusters contains the agricultural problems with decreasing trend (cluster 2), whereas the other one represents an increasing-trend among the problems (cluster 0). It is also noted that cluster 2 represents the problem of plant protection in the West Bengal state for the following 10 crops, i.e., Chili, Black Gram, Banana, Coconut, Bitter Gourd, Papaya, Pumpkin, Orange, Acid Lime, and Tuberose. Furthermore, cluster 0 shows that farmers from the Uttar Pradesh state are increasingly demanding for help regarding fertilizer usage in the following 12 crops, i.e., Onion, Sugarcane, Tomato, Green Gram, Mango, Pigeon pea, Citrus, Coriander, Ginger, Berseem, Aonla, and Melon. Moreover, Fig. 4 demonstrates a pictorial representation of the obtained insights.

Table 2 depicts a few cluster-sample outputs of the TPTC module with crop-wise insights. From the table it is noted that farmers from the four Indian states including Uttar Pradesh, Rajasthan, Madhya Pradesh, and Uttarakhand have been asking Plant protection-related questions in the Tulsi (Basil) crop in decreasing manner over the past few yea16rs (Fig. 5a). Moreover, cluster no. 9 in the table shows that farmers from Uttar Pradesh, Jharkhand, Gujarat, and Uttarakhand have been asking questions about Fertilizer usage in the crop of

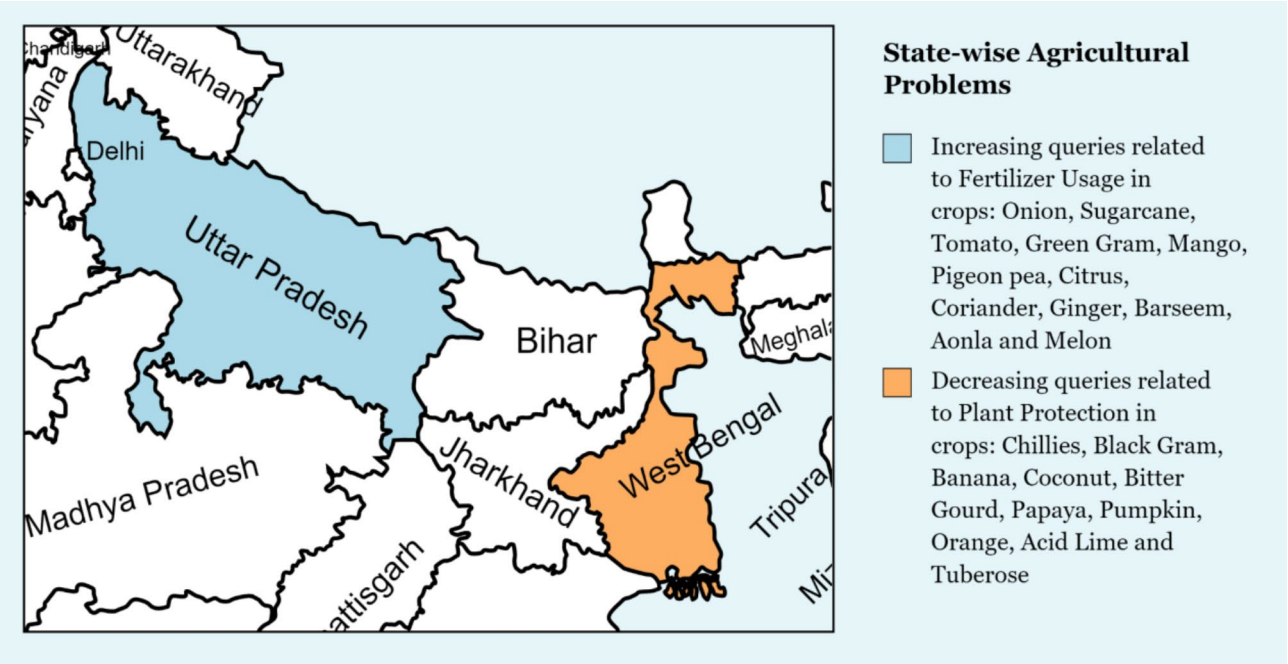
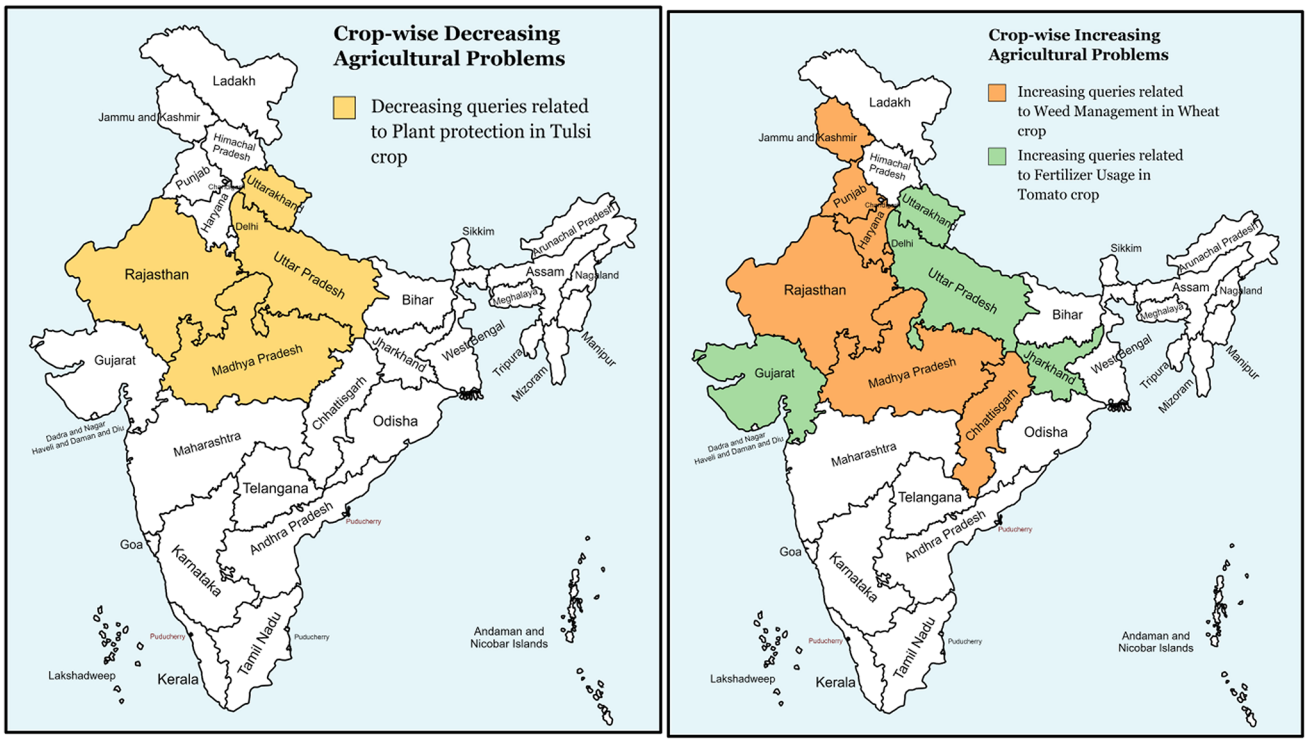


Fig. 4. TPTC state-wise insights - visualized results (map created by: https://www.mapchart.net/india.html).



(a) Crop-wise decreasing problems *(b) Crop-wise increasing problems*

Fig. 5. Visual representation of the crop-wise agricultural problems extracted by the TPTC module (map created by: https://www.mapchart.net/india.html).

tomato increasingly since 2013 (Fig. 5b). Furthermore, a similar pattern is noted in the queries related to Weed management in the wheat crop in the states of Punjab, Haryana, Rajasthan, Madhya Pradesh, Chhattisgarh, Delhi, and Jammu and Kashmir. Besides, from the table, it is observed that the farmers from the states of Rajasthan, Gujarat, Maharashtra, Odisha, Madhya Pradesh, and Chhattisgarh also ask questions related to Weed management in the groundnut crop increasingly.

Figure 5 gives an example of how the output of the TPTC module can be represented visually on a geographical map. Figure 5(a) illustrates the states where farmers have been asking Plant protection-related queries in the Tulsi crop decreasingly over the past few years. Furthermore, Fig. 5(b) gives information regarding the states where farmers have been asking questions related to weed management in wheat crop and fertilizer usage in tomato crop in increasing order.

Forecasting of monthly topic-wise query calls

In this study, seven statistical time-series forecasting models viz. 'ARIMA', 'Prophet', 'TATS', 'TBAT', 'TBATS1', 'TBATP1' and 'TBATS2' were used to forecast the query counts and their comparative prediction performances are presented in this section. In order to perform the forecasting of monthly topic-wise calls of 100 time-series including 84 data points each has been used with 7 different forecasting models, therefore, in the study, a total of 700 models have been developed and evaluated. In the study, the data was split 75% for training and 25% for testing, with careful selection to avoid data leakage. Cross-validation was not used due to the small dataset size (84 points per series). While the models are adaptable to other domains, they were specifically optimized for agricultural data in this study. The sample output of the developed forecasting models for four different agricultural problems is presented in Fig. 6. From the figure, it is observed that most of the models successfully capture the seasonal patterns of the farmers' query calls.

Furthermore, to evaluate the performance of the forecasting models two metrics are taken into account, i.e., RMSE and MAE. The box plots of the models' RMSE and MAE on all the considered time series are presented in Fig. 7(a and b). In addition, the average RMSE and average MAE of all models are given in Fig. 7(c) and Table 3. From the results, it is observed that the TBATP1 model performed better in terms of RMSE and MAE as compared to the other models with the RMSE and MAE values of 0.034 and 0.107, respectively. Whereas, the performance of the TBAT-based model is noted to be the lowest in terms of both RMSE and MAE, with values 0.089 and 0.191, respectively. The above results suggest that, in comparison to the other models, on average, the TBATP1-based model reflected the times-series data query calls more accurately.

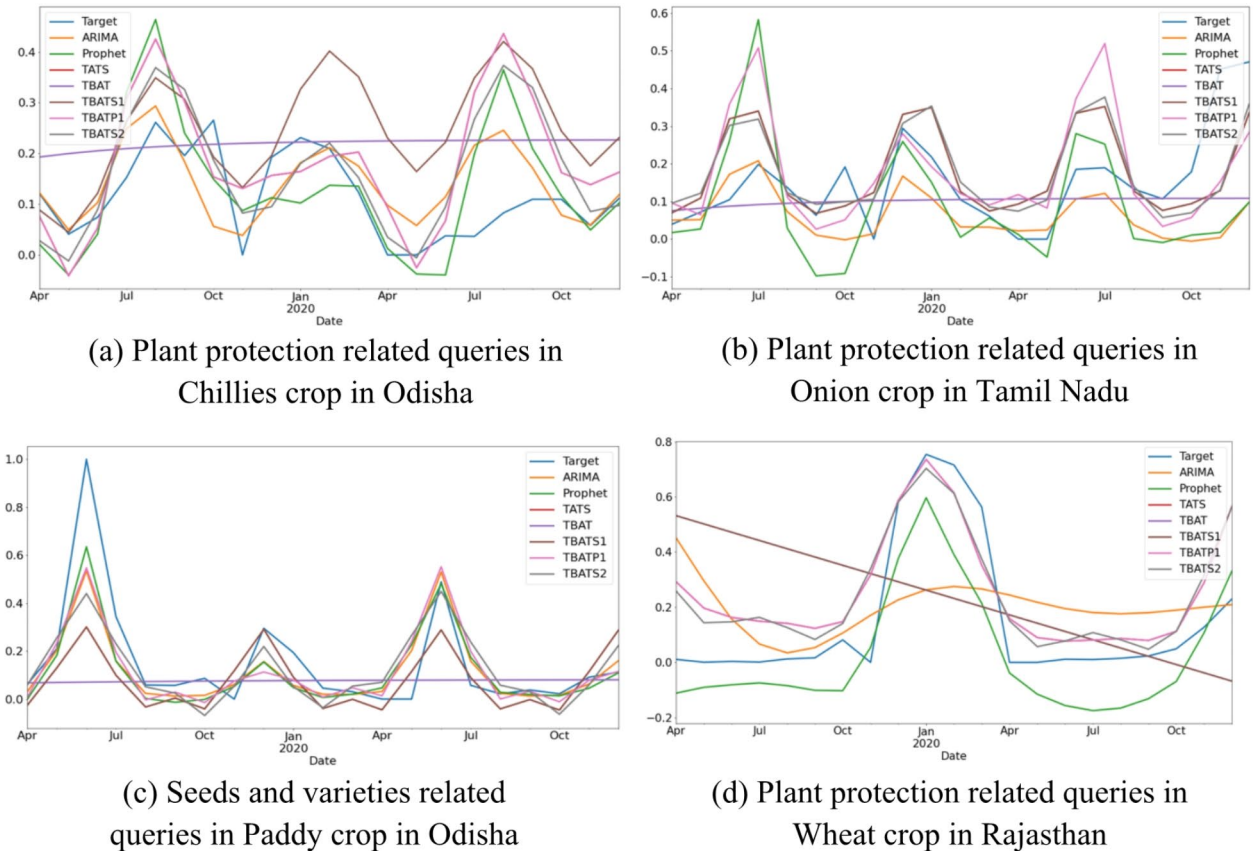
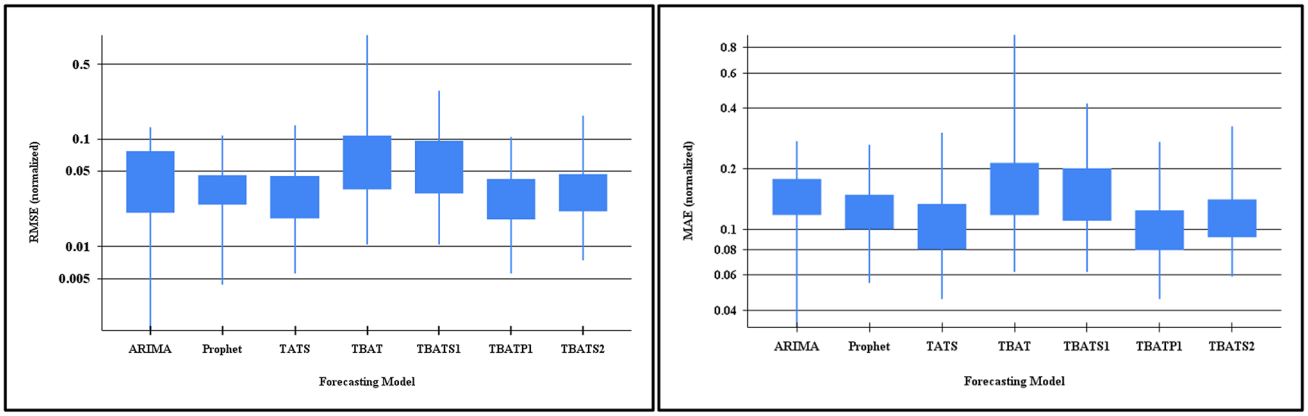
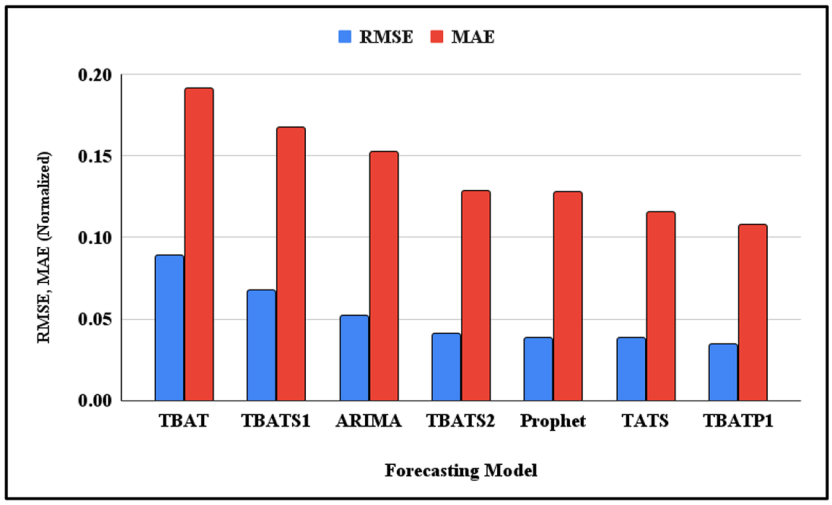


Fig. 6. Sample output of the developed forecasting models.



(a) Models' RMSE on multiple time series (b) Models' MAE on multiple time series



(c) Models' RMSE and MAE comparison

Fig. 7. RMSE and MAE comparisons of the forecasting models.

S. No.	Model	RMSE	MAE
1.	TBAT	0.089	0.191
2.	TBATS1	0.067	0.167
3.	ARIMA	0.051	0.152
4.	TBATS2	0.040	0.128
5.	Prophet	0.038	0.127
6.	TATS	0.038	0.115
7.	TBATP1	0.034	0.107

Table 3. RMSE and MAE comparison of forecasting models.

Discussion

Topic-wise problems' Trend clusters

In the Discussion section, we focus on the insights from the output of the TPTC module mentioned in Tables 1 and 2, including a total of six clusters of problems (two state-wise and four crop-wise clusters). Nonetheless, the complete output of the TPTC module is given in supplementary information. The obtained insights are beneficial for the nationwide and micro-level decision-makers to predict the futuristic market demand, greenhouse gas emissions and to introduce non-chemical based farming practices. The insights are also useful for designing agricultural extension policies, agricultural research activities and marketing strategies for optimizing and matching spatio-temporal demands of herbicide usage. Another major use of TPTC insights is the impact

assessment of the introduced governmental programs. Furthermore, other utilities of the TPTC-based insights include obtaining the specific topic-level queries that farmers ask. This consequently helps in multiple scenarios including focusing the attention of the authorities on particular problems of the regions, designing research studies, agricultural products and many more.

State-wise clusters with decreasing and increasing agricultural problems

It is noted that farmers from the Uttar Pradesh state have been increasingly asking queries regarding Fertilizer usage in various crops including Onion, Sugarcane, Tomato, Green Gram, Mango, Pigeon pea, Citrus, Coriander, Ginger, Berseem, Aonla, and Melon. These observations are supported by the increased fertilizer usage in the Uttar Pradesh state by the Directorate of Economics and Statistics, Department of Agriculture and Farmers' welfare[20,21].

The findings suggest that in order to reduce chemical use for plant nutrition and protection, AI, IoTs and precision agriculture can help in knowing precise requirement of these chemicals and avoid over dosage. However, it may not be possible for smallholders in the country to adopt such technologies on their own. Hence, facilitating policies are the need of the hour.

Crop-wise clusters with increasing agricultural problems

From Table 2, it is observed that farmers from the Indian states including Uttar Pradesh, Rajasthan, Madhya Pradesh, and Uttarakhand have been asking questions regarding Plant protection topic in the Tulsi crop in decreasing number over the past few years (Fig. 5a). The primary cause behind this observation includes the increased awareness regarding seed treatment, sowing time adjustment and other agronomic practices including inter-cropping among farmers[22].

Moreover, cluster no. 9 in the table shows that farmers from Uttar Pradesh, Jharkhand, Gujarat, and Uttarakhand have been asking questions about Fertilizer usage in the crop of tomato increasingly (Fig. 5b). This seems to be the consequence of increased area of cultivation of tomato crop in the states[23].

In addition, it is also noted that, the queries related to Weed management in the wheat crop from the states of Punjab, Haryana, Rajasthan, Madhya Pradesh, Chhattisgarh, Delhi, and Jammu and Kashmir are increasing. This seems to be the consequence of the development of herbicide resistance in different weed species in wheat-growing season. Farmers are interested to know alternative herbicides and herbicidal rotation options[24], and the enhancement of area under zero-tillage wheat in the Indo-gangetic plains which demands herbicidal weed management[25].

Furthermore, from the table, it is also observed that the farmers from the states of Rajasthan, Gujarat, Maharashtra, Odisha, Madhya Pradesh, and Chhattisgarh ask questions related to Weed management in the groundnut crop increasingly. The possible reason behind it is that area of cultivation have increased for the crop in the past few years[26] and, also the increased farmers' awareness programs toward oil-seed crops by the government[27]. Furthermore, the development of new post emergence herbicide molecules in the recent past for the legume crops[28] also supports the conclusions.

Monthly topic-wise query-calls forecasting

In the present study, 100 time series corresponding to the monthly query calls of the top 5 crops for the top 5 states and 4 most frequently asked topics were taken into account for the assessment of the forecasting performances. From the comparison results of the performances of the forecasting models trained over 100 time series, it is noted that the TBATP1-based models achieved the best performance on average (RMSE = 0.034, MAE = 0.107). Moreover, the performance of TATS, Prophet and TBATS2-based models was found to be comparable to that of the best model (RMSE = 0.038–0.04, MAE = 0.115–0.128). Whereas, the ARIMA, TBATS1 and TBAT-based models were noted to have the highest error rate (RMSE = 0.051–0.089, MAE = 0.152–0.191). Monthly, predictions from the forecasting models are valuable in generating the farmer advisories in advance. This consequently can be useful in a number of intelligent systems including early warning systems, recommender systems, and market price forecasting systems.

The performance of the forecasting models in the study can be attributed to several factors related to their underlying methodologies and how well they fit the data characteristics of the monthly query calls for crops. The TBATP1 model likely combines various components that account for seasonal patterns, trends, and any irregularities present in the time series data. This complexity allows it to adapt to different data characteristics, improving its accuracy. Furthermore, the TBATP1 models may have benefited from effective hyperparameter tuning, optimizing their configurations to better fit the training data and improve generalization on the test data.

In contrast, the ARIMA models assume linear relationships in the data and may struggle with complex non-linear patterns typical in agricultural time series data, leading to poorer performance. Besides, the TBATS1 and TBAT models, while capable of handling seasonality, may not capture the specific seasonal patterns present in the query calls as effectively as the TBATP1 model.

The presented study faces several limitations, including challenges in generalizing the results to other datasets or regions due to potential differences in agricultural practices and climatic conditions. Additionally, there may be issues related to data quality from the Kisan Call Centres, such as inconsistencies or inaccuracies in the recorded query calls, which could affect the reliability of the forecasting models.

Likewise, to enhance the TPTC framework and forecasting models, future research could explore the integration of more complex machine learning algorithms, such as ensemble methods or deep learning techniques, which may improve predictive accuracy. Additionally, testing the models on diverse agricultural datasets, including varying crop types or climatic conditions, could provide insights into their adaptability and robustness. Incorporating external variables, such as weather patterns or economic factors, could also enhance model performance by capturing broader influences on agricultural queries.

Conclusion

The present study is designed to obtain novel insights regarding the nationwide agricultural problems and forecast the demand for help using the farmers' helpline data. The study offers the concept of TPTC that delivers insights regarding nationwide common problems related to the agriculture sector along with their increasing or decreasing trend over the past few years. The paper also outlines the stages of the forecasting models development that will be used to predict the monthly number of query calls from the farmers of the target states corresponding to the particular agricultural problems. The proposed methodology uses data mining integrated with advanced statistical and machine learning techniques to extract insights from the helpline data. The article also elaborates on various practical agricultural problems pointed out by the proposed study along with the possible reasons behind them. Additionally, the comparison of the forecasting models' performance shows that TBATP1 is the most suitable model for predicting the purpose of such a time series. The reason is that the model integrates multiple components that capture seasonal patterns, trends, and irregularities within the time series data. The extracted insights and the developed models in the study are useful for agriculture-related decision-making, and the development of systems including recommender systems, early warning systems and also smart-market analytics systems. As for the future scope of the present study, the authors tend to use Natural Language Processing-based models to extract insights based on the question asked by the farmers and use Deep Learning-based forecasting models in the subsequent studies.

Materials and methods

The whole methodological aspect of the present study can be divided into three modules, i.e. Data acquisition and pre-processing module, Topic-wise problems' trend clustering module, and query-calls forecasting modules (Fig. 8). First, the raw data i.e. the call-log record files were downloaded from the Kisan Call Centre Servers and pre-processed to eliminate the inconsistencies, noise, and inaccuracies. Subsequently, yearly time series were extracted from the pre-processed dataset. The extracted crop-wise yearly time series were fed to the TPTC module to extract insights and trends of the agricultural problems. Later, the monthly time series were extracted from the pre-processed dataset and the crop-wise monthly time series were provided to the forecasting module to train and evaluate the forecasting models.

Data acquisition and pre-processing module

In general, this module deals with obtaining the dataset from the KCC servers and transferring it to our disc storages, as well as with preparing the raw data to eliminate noise and inconsistencies. The following subsections provide a thorough explanation of these steps:

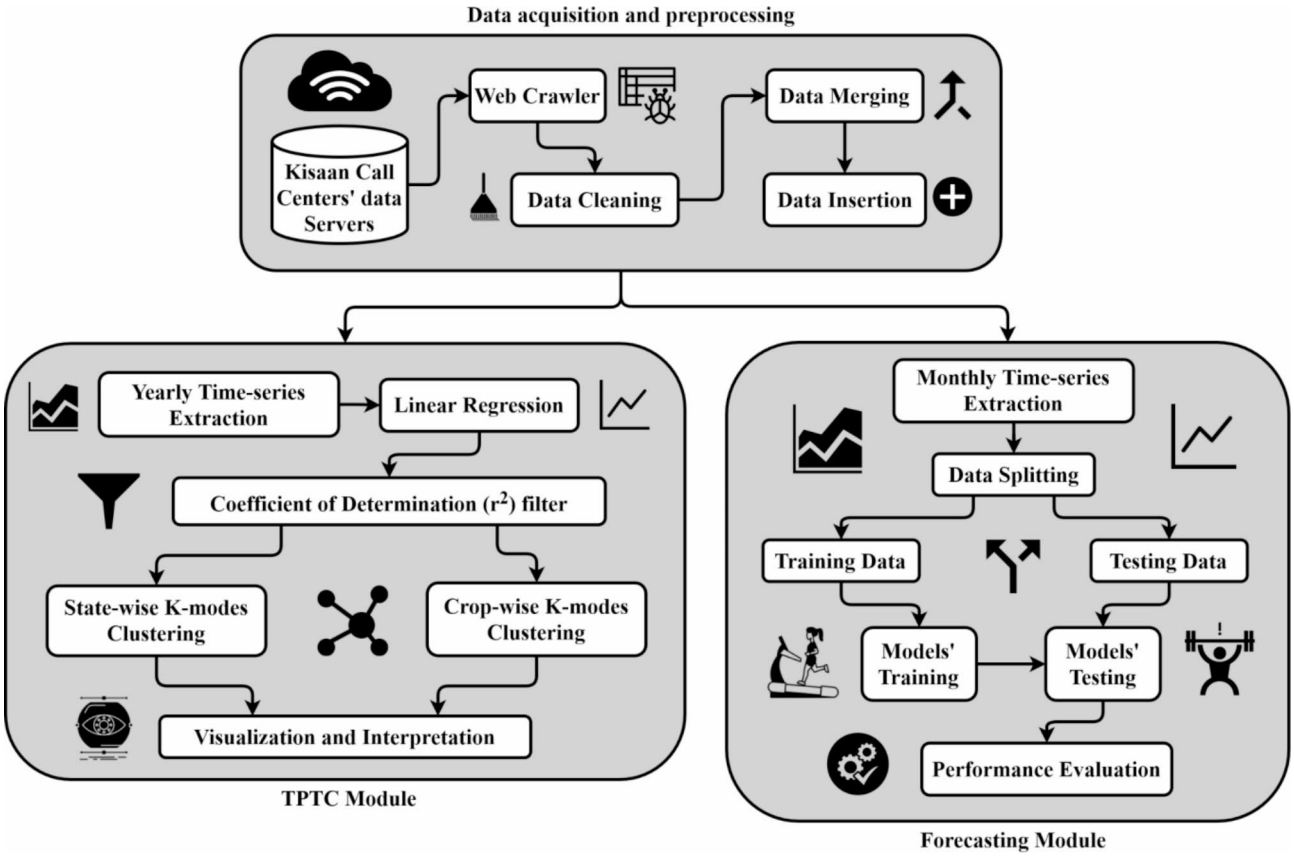


Fig. 8. Workflow of proposed methodology.

KCC dataset and its acquisition
The dataset used in this study was first gathered from the official Kisan Call Center (KCC) scheme of India website. KCC is an initiative by the Indian government that provides a help-line service for the queries of the farmers of the country. The acquired dataset is in .json format, and primarily in textual format. In this step, total 55,844 files downloaded from the helpline server. The dataset contains 26,874,198 queries of farmers from all over India from March 2013 to November 2021. Table 4 provides a thorough explanation of the call-log records' parameters.

Data pre-processing
Pre-processing helps to improve the quality of the primary data by removing inconsistencies, noise, inaccuracies, etc. The raw dataset for this study undergone the following data pre-processing techniques:

a) **Data Cleaning**: Since noisy data can induce unpredictable results, we used data cleaning to deduct noise from the raw data files. In this step, all characters from the records except the alphabets, digits, and a handful of special symbols including commas, space, hyphens, etc. excluded from the records.
b) **Data Merging**. Subsequently, all of the data files were merged into a single .csv file. A single-file-dataset makes it more comfortable to regulate and execute operations on the data records. The output of this step includes a single .csv file consisting of 26,874,198 queries is obtained by merging files from all states.
c) **Data Selection**. Next, the attributes irrelevant to the present study were removed from the dataset. The output file contains the following four attributes, i.e., CreatedOn, StateName, QueryType and Crop.
d) **Data Insertion**. The "CreatedOn" attribute includes details regarding the phone call-query's year, month, day, and time. In this phasetwo new attributes: "Year" and "Month", were added to the dataset by separating values from the "CreatedOn" attribute.

Topic-wise problem-trend clusters (TPTC) module

In this module, the TPTCs were obtained from the extracted time-series data points of query-calls count using linear regression in combination with the k-modes clustering technique. A detailed explanation of the whole process is as follows:

Yearly time-series extraction
First, the yearly query-calls count times series from the pre-processed dataset were extracted. The time series consists of the number of queries made by the farmers every year, denoted by the Eq. 1. Each time series corresponds to the combination of crop name, state name and topic. The extracted time series can be mathematically represented as:

$$T = (t_1,\ t_2,\ \ldots, t_N) \tag{1}$$

where, T represents the yearly time-series corresponding to the selected crop and t_i represents the number of query calls made by the farmers from the selected combination of state, crop and topic in the i^{th} year. The value of each data point t_i is extracted from the dataset using the relational algebraic Eq. 2.

$$t_i = \psi\ (\sigma_{\ \gamma}\ (KCC\_dataset)) \tag{2}$$

$$\gamma = ((Statename == S) \wedge\ (CropName == C) \wedge\ (QueryType == Q) \wedge\ (Year == Y)) \tag{3}$$

Here γ is the condition which is to be satisfied by the dataset records in order to get selected, σ is the selection function, and ψ represents the cardinality of the set of selected records[29]. In the present work, total 37,632 time series were extracted using this step, i.e., a sets of yearly time series corresponding to the 294 crops present in the dataset, from the 32 Indian states/union territories with 4 topics (seeds and varieties, fertilizer usage, weed

Attribute Name	Description	Sample values
BlockName	Block name of farmer	Dondilohara, Konta, A.Chowki(Td), A.Konduru, Aalampur Jafarabad, Aaminpur
tableDistrictName	District name of farmer	Ahmadabad, Ahmednagar, Aizawl, Ajmer, Akola
StateName	State name of farmer	Andhra Pradesh, Arunachal Pradesh, Assam, Bihar, Chhattisgarh, Gujarat, Haryana
CreatedOn	Date and time of the query	2013-03-04 13:20:00.000
Season	Season of year	Jayad, Kharif, Rabi
Category	Query Category	Animal, Avian, Beekeeping, Cereals, Condiments and Spices, Drug and Narcotics
Crop	Target crop of query	Acid Lime, African Sarson, Almond, Aloe Vera, Amaranthus/Grain Amaranthus, Amarphophallus
QueryType	Type of Query	Field Preparation, Plant Protection, Water Management, Agriculture Mechanization, Animal Breeding, Animal Nutrition, Animal Production
Sector	Target sector of query	Agriculture, Animal Husbandry, Fisheries, Horticulture
QueryText	Query in textual format	Control of tsetse fly in cashew, asked about reddening of leaves, asked about leaf webber, sucking pest
KccAns	Query Response	spraying of carbaryle 3 grms/lit water, spray micromax, suggest to spray chloripyriphos, carbaril 2 g/l

Table 4. Detailed description of the KCC dataset.

management, and plant protection) each. Moreover, not all the extracted time-series were useful in the analysis, as many combinations of the state, crop and topics do not produce any records, which is why such time-series are eliminated that contain no data points. After the removal of such time-series, linear regression is performed on all the 11,836 yearly time-series separately.

Linear regression on the obtained time-series
In order to extract the rate (or slope) and intercept of each of the obtained time series, a linearly regressed model is fitted represented by Eq. 4[30].

$$y = mx + c \tag{4}$$

Here, *y* is a linear representation of the dependent variable i.e., query-call counts, m represents the slope of the regression line, x represents the values of the independent variable i.e. year. The values of *y* and *x* are known to the system from the time series, whereas, the Eqs. 5 & 6 were used to calculate the values of m (slope) and c (intercept) for each time series:

$$m = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x^2\right)} \tag{5}$$

$$c = \frac{n\sum(xy) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \sum(x^2)} \tag{6}$$

Here, *n* represents the total number of observations in each time-series, i.e. a total number of years into consideration.

Coefficient of determination (R2) based time-series filtering
In the next step, the coefficient of determination between the observed (actual data points in the time series) and predicted values of *y* (using the linear regression model) were calculated using Eq. 7[31].

$$R^2(T) = \left(\frac{\sum\left(\hat{y}_i - \underline{\hat{y}}\right)\left(y_i - \underline{y}\right)}{\sqrt{\sum\left(\hat{y}_i - \underline{\hat{y}}\right)^2 \left(y_i - \underline{y}\right)^2}} \right)^2 \tag{7}$$

Here, $\widehat{y}$and y represents the predicted and actual values of time series *T*, respectively. Subsequently, the time series are filtered using the R^2 values, as shown in Eq. 8:

$$p = \{\, T \mid R^2(T) > 0.7 \,\} \tag{8}$$

Here *p* represents the set of problems (time series corresponding to a particular combination of state, crop and topic) which are left after the filtering procedure. Since 0.7 indicated a good fitness of the linear model, in this study we have used this value for filtration. Decision maker can tune the value according to the requirement of the situation. An R^2 value of 0.7 means that 70% of the variation in the data is explained by the model, which suggests a strong fit. We selected an R^2 threshold of 0.7 because it is often considered a good benchmark in agricultural data analysis for identifying reliable trends. This threshold ensures that the trend being detected, whether an increase or decrease in agricultural problems, is based on a model with sufficient predictive accuracy. While this is a commonly used standard, we specifically chose it to ensure robustness in the context of our dataset's complexity.

State-wise and crop-wise K-modes clustering
Until this step, the extracted time series are those which show a high coefficient of determination, i.e. the problems are showing a strong linear relationship (either increasing or decreasing) with time. This also serves in validating the obtained results as only the most promising facts are considered for clustering. Moreover, in this step, four attributes of the filtered time series (i.e. state name, crop name, topic name, and slope) are used for clustering the problems. Among the considered four attributes, only slope is a numerical valued attribute, others are categorical in nature. Therefore, the slope attribute is converted from numerical to categorical using Eq. 9.

$$c(m) = \text{“}increasing\text{”} \quad if\ m > 0,\ \text{“}decreasing\text{”} \quad otherwise \tag{9}$$

Here, *m* represents the slope value that is to be replaced by the corresponding categorical value, furthermore, if the input value is greater than 0, it is assigned the “increasing” category, and “decreasing” otherwise. Next, the k-modes clustering algorithm[32] is used to group the similar agricultural problems together based on four attributes (state, crop, topic, slope). K-modes is a widely used algorithm for grouping categorical data because it is simple to implement and effectively handles enormous quantities of data. It uses the distance metric of “mismatches” between the input data points (problems in our case). The lesser the dissimilarities (similar

problems) the closer the input data points are. Furthermore, it uses the 'mode' of the cluster data points, instead of the 'mean'. Following is the algorithm for the same:

K-modes clustering algorithm **Input** Data points Z (each point comprising a vector of four values), Number of clusters K to be generated.

Step 1: Randomly choose the initial K number of modes, C_j, such that $j = 1,2,\ldots,K$ from the data points.

Step 2: Calculate the dissimilarity between K initial cluster modes and each data point using Eqs. 10 & 11.

$$d\left(z_i,\ q_r\right) = \sum_{j=1}^{m} \delta\left(z_{ij}, q_{rj}\right) \tag{10}$$

$$\delta\ \left(z_{ij}, q_{rj}\right) = 1 \qquad if\ \ z_{ij} = q_{rj},\ 0 \qquad if\ \ z_{ij} \neq\ q_{rj} \tag{11}$$

Here, z_i represents the ith data point of the dataset, q_r represents the mode data point of cluster r, m is the total number of attributes that each data point contains.

Step 3: Assign the data points to the closest cluster modes.

Step 4: Revise the modes using the frequency-based approach on newly assembled clusters.

Step 5: Repeat step 2 and step 4 until the clusters have no modifications.

K-modes clustering is a machine-learning technique used for grouping data with categorical attributes. Unlike traditional K-means clustering, which works with numerical data, K-modes allow us to group agricultural problems into clusters based on similarity in categorical features.

In the present study, two types of TPTC insights are extracted, i.e. state-wise common problem all over the country, and nation-wide crop-wise common problems. In order to capture the state-wise problems showing a similar pattern, the clustering algorithm is given only the state, topic and slope as input. Next, to obtain the crop-wise problems (similar problems in different states), the clustering algorithm is given the crop, topic and slope as input. The clustering insights can help policymakers identify prevalent agricultural issues across regions, enabling targeted interventions. By grouping similar problems, decision-makers can prioritize resource allocation and develop tailored solutions for specific farming challenges.

Data visualization and interpretation

Next, the output of the k-modes clustering is visualized using geographical maps (Figs. 4 and 6) and in tabular form (Table 2). The geographical map is used to display the similar problems that are being asked by the farmers in increasing/decreasing trends. Furthermore, the tabular format is used to display the clusters with the actual values and helps in obtaining detailed information regarding the clustered data points.

Visualizations play a crucial role in enhancing the understanding of clusters and forecasting results in our study. By representing complex data in a graphical format, stakeholders can easily interpret patterns and relationships that may not be immediately apparent in raw data. For instance, visualizations of clustering results allow users to identify distinct groups of agricultural problems based on query patterns, helping them prioritize issues that require immediate attention.

Furthermore, visual outputs of forecasting results, such as time-series plots, provide clear insights into expected trends and fluctuations in query calls. This aids stakeholders, including policymakers and farmers, in making informed decisions, such as optimizing resource allocation and implementing timely interventions. By translating statistical findings into accessible visuals, stakeholders can engage more effectively with the data, fostering collaboration and improving outcomes in agricultural practices and policy design.

Forecasting monthly topic-wise query calls

The forecasting module developed in the study can be further divided into four basic steps, these are:

Monthly time-series extraction

First, the monthly query-calls count times series from the pre-processed dataset were extracted in the similar manner as discussed in subsection 3.2.1. The extracted time series consist of the number of queries made by the farmers every month, the time series can be denoted by the Eq. 1. Moreover, the value of each data point t_i is extracted from the dataset using the relational algebraic Eq. 2. In contrast to the previous extraction, the condition used for time-series extraction (Eq. 3) is substituted by Eq. 12.

$$\gamma = ((Statename == S) \wedge\ (CropName == C) \wedge\ (QueryType == Q) \wedge\ (Month == M)) \tag{12}$$

Data splitting

After obtaining the target time-series, each of the series consisting of the query-call counts was splitted into two parts, i.e. training data and testing data in the ratio of 75:25. There were a total of 100 such time series present in the dataset and each series consists of 84 data points (12 months × 7 years), 63 of which were used to train the models and the last 21 months of data are used for models' testing.

Model training

After splitting the time series, seven statistical forecasting models ('ARIMA', 'Prophet', 'TATS', 'TBAT', 'TBATS1', 'TBATP1' and 'TBATS2') were trained on the training data. In the study, we opted for statistical models rather than machine learning (ML) or deep learning (DL) models due to the limited number of data points available. Modern ML and DL techniques typically require large datasets for effective training, which was not feasible with

our time-series data (only 84 data points in each time-series). Therefore, statistical models were more suitable for this study. Details of the considered models are as follows:

- **Auto-Regressive Integrated Moving Average (ARIMA)**: ARIMA is a statistical time-series forecasting model that uses previous values of a time series to predict future values[3]. It consists of three components:
  - Auto-Regression (AR): A model that regresses a variable on its own past values.
  - Integrated (I): Differencing of raw observations to stabilize the time series.
  - Moving Average (MA): Incorporates the dependency between an observation and residual errors.
- **Prophet**: Prophet is an additive time-series forecasting model that fits non-linear trends with yearly, weekly, and daily seasonality[34]. It works well for time series with strong seasonal patterns and multiple seasonal cycles. The core equation of the model involves three components:
  - g(t): Describes a linear or logistic growth trend over time.
  - s(t): Describes the seasonal pattern.
  - h(t): Describes the effect of holidays or specific events.
  - ε(t): Denotes the error term.
- **TBATS**: TBATS is a time-series forecasting model designed for data with complex seasonal patterns[35]. It stands for:
  - T: Trigonometric seasonal components.
  - B: Box-Cox transformation.
  - A: ARIMA errors.
  - T: Trend.
  - S: Seasonal components.

The model uses exponential smoothing and handles multiple seasonalities. A Box-Cox transformation is applied to stabilize variance and normalize the data. Several variations of TBATS are used, such as:

- TATS: Trend and seasonal, no Box-Cox.
- TBATS1: Trend with one seasonal component and Box-Cox.
- TBATS2: Trend with two seasonal components and Box-Cox.
- TBATP1: TBATS1 with hard-coded seasonal periodicity.

Model testing
In the present study, the forecasting models were evaluated on the testing data which comprises of the last 21 data points of the time series. After training the models, the subsequent query-call counts of 21 months are predicted with each model on every time series comprising of overall 700 models (100 time series × 7 forecasting models) being tested.

Models' performance evaluation
The prediction performance of the trained forecasting models was evaluated using two performance metrics, i.e., root mean squared error (RMSE) and mean absolute error (MAE) denoted by the following Eqs. 20 & 21:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \widehat{y_i})}{N}} \tag{20}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y_i}| \tag{21}$$

where y_i is the ith observed time series, $\widehat{y_i}$ is ith the predicted time series and N is the total number of observations.

Data availability
The KCC data utilized in this study is available at https://kcc-chakshu.icar.gov.in/.

Received: 8 December 2023; Accepted: 19 November 2024
Published online: 26 November 2024

References
1. Pulicherla, K. K., Adapa, V., Ghosh, M. & Ingle, P. Current efforts on sustainable green growth in the manufacturing sector to complement make in India for making self-reliant India. *Environ. Res.* **206**, 112263 (2022).
2. Agrawal, S., Kumar, V., Kumar, S. & Shahi, S. K. Plant development and crop protection using phytonanotechnology: a new window for sustainable agriculture. *Chemosphere* **299**, 134465 (2022).
3. Viswanath, V. K., Chandu, G. V. M., Chamarthi, R. & Saravanan, S. & Manju V. Hadoop and natural language processing based analysis on kisan call center (kcc) data. *In International Conference on Advances in Computing, Communications and Informatics*. 1142–1151 (2018). (2018).

4. Godara, S. & Toshniwal, D. Sequential pattern mining combined multi-criteria decision-making for farmers' queries characterization. *Comput. Electron. Agric.* **173**, 105448 (2020).
5. Chouhan, R. S., Kumar, D. & Sharma, H. O. Performance of Kisan Call Centre: A case study of Kisan Call Centre of Indian Society of Agribusiness Professionals Bhopal (Madhya Pradesh). *Indian J. Agric. Econ.* **66**, 534 (2011).
6. Kavitha, S. & Nallusamy, A. A study on socioeconomics characteristics of Kisan call centre beneficiaries and non-beneficiaries in Mahbubnagar district of Telangana. *J. Pharmacogn Phytochem*. **8**, 4660–4663 (2019).
7. Godara, S. & Toshniwal, D. Deep learning-based query-count forecasting using farmers' helpline data. *Comput. Electron. Agric.* **196**, 106875 (2022).
8. Mohapatra, S. K. & Upadhyay, A. Using TF-IDF on the Kisan call centre dataset for obtaining query answers. In *International Conference on Communication, Computing and Internet of Things.* 479–482 (2018). (2018).
9. Mohapatra, S. K. & Upadhyay, A. Query Answering for Kisan Call Center with LDA/LSI. In *International Conference on Advances in Computing, Communication Control and Networking.* 711–716 (2018).
10. Arora, B. et al. Agribot: a natural language generative neural networks engine for agricultural applications. In *International Conference on Contemporary Computing and Applications (IC3A)*. 28–33 (2020). (2020).
11. Ajawan, P., Desai, P. & Desai, V. Smart Sampark-An approach towards building a responsive system for Kisan Call Center. In *2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC).* 1–5 (2020).
12. Momaya, M., Khanna, A., Sadavarte, J., Sankhe, M. & June). Krushi–The Farmer Chatbot. In *International Conference on Communication information and Computing Technology.* 1–6 (2021). (2021).
13. Jaisridhar, P. *Impact of Kisan call Centre on Technological Adoption among Dairy Farmers of Tamilnadu* (National Dairy Research Institute, 2013).
14. Chachra, K. et al. The impact of Kisan call centres on the farming sector, in Environmental and Agricultural Informatics: Concepts, Methodologies, Tools, and *IGI Global* 66–78 (2020).
15. Behera, B. S. et al. E-governance mediated agriculture for sustainable life in India. *Procedia Comput. Sci.* **48**, 623–629 (2015).
16. Gandhi, M., Pallae, A. & Srinivas, A. A scale to measure the attitude of farmers towards Kisan call centre. *Interaction* **31**, 56–59 (2013).
17. Sudharani, V., Pallae, A. & Srinivas, A. A scale to measure the attitude of farmers towards Kisan call centre-development and standardization. *Interaction* **31**, 77–80 (2013).
18. Rupasi, T., Sharma, M. C. & Singh, B. P. Animal health information needs of livestock owners: A Kisan Call Centre analysis. *Indian J. Anim. Sci.* **80**, 187–188 (2010).
19. Mukherjee, S. Agricultural diversification of West Bengal: Nature and Policy implications. *Indian J. Agric. Econ.* 76 (2021).
20. GOI. Chemical Fertilizers: NPK Consumption: Uttar Pradesh: Nitrogen. *NPK Consumption: Uttar Pradesh* (2021). https://www.ceicdata.com/en/india/chemical-fertilizers-nitrogen-phosphate-and-potash-npk-consumption-by-states/chemical-fertilizers-npk-consumption-uttar-pradesh-nitrogen
21. Pandey, C. & Diwan, H. Assessing fertilizer use behaviour for environmental management and sustainability: A quantitative study in agriculturally intensive regions of Uttar Pradesh, India. *Environ. Dev. Sustain.* **23**, 5822–5845 (2021).
22. Mishra, D. & Ghadei, K. Farmers' Knowledge about Safe Use of Plant Protection Measures in Eastern Uttar Pradesh, India (2021).
23. Gulati, A., Wardhan, H., Sharma, P. & Tomato Onion and Potato (TOP) value chains in *Agricultural Value Chains in India* 33–97Springer, (2022).
24. Choudhary, A. K. et al. Post-emergence herbicides for effective Weed Management, enhanced Wheat Productivity, profitability and quality in North-Western Himalayas: a 'Participatory-Mode' Technology Development and Dissemination. *Sustainability* **13**, 5425 (2021).
25. Kumar, V., Bana, R. S., Singh, T. & Louhar, G. Ecological weed management approaches for wheat under rice–wheat cropping system. *J. Environ. Sustain.* **4**, 51–61 (2021).
26. Kateshiya, G. B. Area under Groundnut Cultivation Crosses 19 Lakh Hectares for First Time in 16 Yrs. *Area under Groundnut Cultivation* (2020). https://indianexpress.com/article/business/commodities/gujarat-area-under-groundnut-cultivation-crosses–19-lakh-hectares-in–16-yrs–6507693/
27. Reddy, A. Policy implications of Minimum Support Price for Agriculture in India. *Acad. Lett. Article* 2406 (2021).
28. Nayak, A., Lokesha, H. & Gracy, C. P. Growth and instability analysis of Groundnut production in India and Karnataka. *Econ. Aff.* **66**, 61–69 (2021).
29. Gray, P. *Logic, Algebra and Databases* (E. Horwood, 1984).
30. Montgomery, D. C., Peck, E. A. & Vining, G. G. *Introduction to Linear Regression Analysis* (Wiley, 2021).
31. Asuero, A. G., Sayago, A. & González, A. G. The correlation coefficient: An overview. *Crit. Rev. Anal. Chem.* **36**, 41–59 (2006).
32. Chaturvedi, A., Green, P. E. & Caroll, J. D. K-modes clustering. *J. Classif.* **18**, 35–55 (2001).
33. Saboia, J. L. M. Autoregressive integrated moving average (ARIMA) models for birth forecasting. *J. Am. Stat. Assoc.* **72**, 264–270 (1977).
34. Taylor, S. J. & Letham, B. Forecasting at scale. *Am. Stat.* **72**, 37–45 (2018).
35. Livera, A. M., Hyndman, R. J. & Snyder, R. D. Forecasting time series with complex smoothing exponential using seasonal patterns. *J. Am. Stat. Assoc.* **106**, 1513–1527 (2011).

Acknowledgements

We would like to express our sincere gratitude to Dr. Y. S. Shivay from the Agronomy Division at ICAR-IARI, New Delhi, and Dr. P.S. Birthal from ICAR-NIAP, New Delhi, whose invaluable contributions and insightful discussions greatly enriched this paper.

Author contributions

SG and SB collected the data, developed the model code and executed them. JB, RJ, AH designed the experiments. RSB, RP, SM, MP, SS and RN interpreted the results and prepared the manuscript with contributions from all co-authors. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Funding

No funding source available now.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-80488-x.

Correspondence and requests for materials should be addressed to S.B., R.S.B. or S.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.