ANALYSIS AND PREDICTION OF SMART DATA USING MACHINE LEARNING

RADHIKA SREEDHARAN[1]

Ex- Faculty of CMR University (Management and Commerce), Bangalore, India

ABSTRACT

In the field of agriculture, Big data creation and analysis had been one of the most important technologies. The need aroused as the sensor technologies were proved to be advantageous in agricultural industry. Various sectors like food safety and breeding had its contribution, because agriculture got improvised by that. The data on agriculture were taken from Tamil Nadu data set. A comparison of consecutive years (2009-2013) was made in the production of crops among different seasons like Rabi, Kharif. The data available helped in the prediction of crop yield. Thereby, the analysis of this big data allowed farmers as well as companies for retrieving the value from certain data and also improved productivity. The Indian economy basically relied on the agricultural sector. Agriculture products needed a variety of protection like protection from insects, protection against rodents and many such undesired attacks in the field of agriculture. Growing status of crops was tracked by segregating, recognizing and measuring areas of different crops in Tamil Nadu and also estimated production early in the year. One of the biggest problems to be tackled is agricultural planning. According to study, crop yield rate, soil classification, weather prediction could be done using Machine Learning techniques. Crop selection was a major issue where cropping using available resources was a major concern.

KEYWORDS: Agriculture Big Data, IoT, Machine Learning

Protection of crops is the practice of managing plant diseases, weeds and other pests which damage forestry and agricultural crops. For producing higher quality crops with minimal wastage, proper crop protection is required. Pesticides protect plants from weeds, fungus, rodents, insects and diseases etc.

Farmers are helped to defend their crops against pests, protect natural resources like soil, water and controlling weeds etc. This is possible by innovations like glyphosate which is used for regulating plant growth.

Proper protection of crops is very essential for producing good and high quality crops with very less waste. If this productivity increases, water, labor and land which are required for crops become less. The philosophy of crop forecasting is based on various kinds of data collected from various sources: data based on weather conditions, data based on water holding capacity of soil, remotely sensed, statistics involved in agriculture. Lots of indices are derived based on climatic or agrarian data. Those indices are relevant variables to determine yield of crops, crop water satisfaction, surplus and excess moisture, etc. Prediction of crops is the practice of production and estimation of crop yields in tonnes or hectares. This happens few months in advance before actual harvest. Machine learning algorithms and statistical techniques are used to predict production and its variance.

LITERATURE REVIEW

[1]Crop Selection Method (CSM) was proposed to figure out problem of crop option and to maximize crops' net yield rate to achieve highest prudence growth of the country. For forecasting of crop yield, Support Vector Machine was utilized which was called Support Vector Regression. Production rate of crops relied on geography of a region, conditions of weather, composition and types of soil and methods of harvesting. However, the limitation of CSM method was it had to improve crops' net yield rate and those crops had to be planted. The accuracy and prediction of CSM method depends on predicted values of determined parameters like weather, type of soil and crop, water density), so a new prediction method needs to be adopted for more accuracy and better performance.[2]Researches were carried out in the field of automatic farming. With the help of given data set, A Valarmathi and Vinciya analyzed the categorization of inorganic, real-estate and organic data for crop prediction. In the research of extracting and predicting useful information, data mining technology was used. Multiple linear regressions were used for the selected region. Ashwani Kumar and her co-author used an algorithm calledas "Agro-Algorithm "for predicting yield of crop and to suggest good crops.

Sandeep and Snehal had discussed use of Artificial Neural Network (ANN) for crop yield prediction. The research was limited to fixed datasets. The future promised that when more data are added, it can be analyzed with

[1]Corresponding author

more approaches of machine learning for generating crop decisions with higher quality correctness.[3]A system of IoT had instruments which had ability to send actual time natural data to cloud repository and Algorithm of Machine Learning to forecast environmental conditions to for fungal identification and inhibition.[4] A machine learning algorithm was developed to process raw data and forecast day-to-day air temperature. This algorithm made use of Support Vector Machine Regression (SVMr).The Algorithm could predict relative air humidity. However, their accuracy needs to be improved because more empirical data need to be collected by these devices in future.[5]The exploratory data forecasting which were made by IoT system could assist crop field managers by facilitating better management of inhibition of fungal diseases. The IoT based devices were under control and they were supervised from remote locations. They were implemented in agricultural fields, grain stores for protection purposes.

METHODOLOGY

Regression coefficient: The quantity which represents a change in subordinate variable which results from change in predicted value, when all variables are unvarying is known as regression coefficient.

$$R=\frac{\mathbf{n\sum(xy)-(\sum x)(\sum y)}}{\sqrt{\mathbf{n}}\mathbf{\sum x2-(\sum x)2}\sqrt{\mathbf{n}}\mathbf{\sum y2-(\sum y)2}}$$

Regression models relate a response variable to one or many forecasters.

The number that lies between -1 and 1 and indicates the level to which the given variables Area and Production are linearly comparable is called as Pearson Correlation.

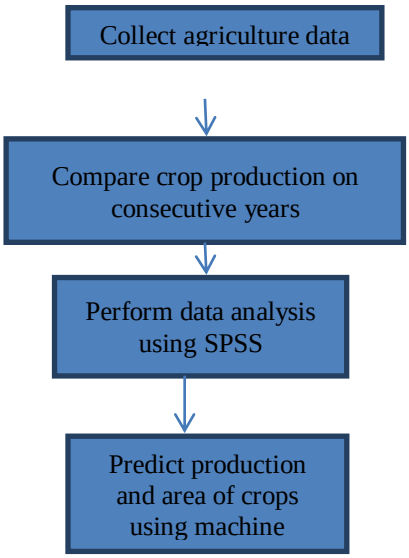
$$R=\frac{\mathbf{\sum xy}}{\mathbf{\sum x2\sum y2}}$$

Variance is the likelihood of the squared fluctuation of a random variable from its mean. Informally, it measures how far a set of (random) numbers are radiated out from their average value.

SPSS could take lots of data. However, while generating graph, it was limited to very finite data. If data are more, it created error while creating graphs. That limitation could be overcome by Machine Learning algorithm

One of the unsupervised learning algorithms is k-means clustering. In this type of algorithm, data can be clustered into k number of partitions.

One of the supervised learning algorithms is decision tree. In this case, it works for both area and production of crops. It is one of the very robust methods of prediction. By giving the value for area and production, based on the given data, it can forecast the type of crop to be produced.



IMPLEMENTATION

Python 3.7.0

Matplotlib

Numpy compact cpython 36

SPSS

RESULTS AND OUTCOME

Data Analysis for various districts of Tamil Nadu:

Ariyalur District

In 2009, area and production of rice was much more compared to small millets in Kharif season. Production of sugarcane was highest compared to other crops for the whole year in 2009. In 2010, production of rice was much higher compared to other crops in Kharif season. In the whole year, only sugarcane crop was produced and its production was much higher compared to

other crops. However, its production was reduced compared to previous year by approximately 8.76%.In 2011,production of cotton was more in Kharif season compared to Rabi season by 86.6%. In Rabi season, groundnut production was more compared to other crops. Its production is more compared to Rabi season by 97.7%. In Kharif season, production of rice was more compared to other crops.

Production of coconut was very high in the whole year. There was no production in 2010; however, there was drastic change in its production in 2011. Compared to year 2009, its production increased by almost 500 times in 2011. Production of sugarcane was increased by 1.19% from 2010 to 2011.The variance in production of crops decreased from 2011 to 2012, and then it again increased in 2012.In year 2012, only sugarcane was produced during whole year. However, its production was reduced by 19.6%.In Kharif season of same year, production of rice was maximum followed by maize, groundnut and cotton.

In 2013, production of rice was highest in Khariff season, followed by maize, cotton (lint) and groundnut. Production of crops like bajra, jowar, urad was very less. In Rabi season, production of groundnut was highest compared to other crops. Compared to 2011, its production was increased by 20.2% in Rabi season. In the whole year, production of coconut was much higher, followed by production of sugarcane.

However, compared to 2011, coconut's production was reduced by 34.61%. The production of sugarcane increased by 57.19% from 2012-2013

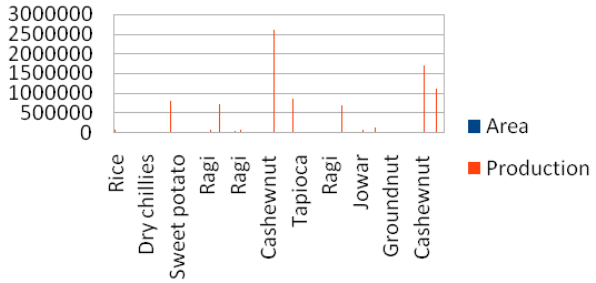
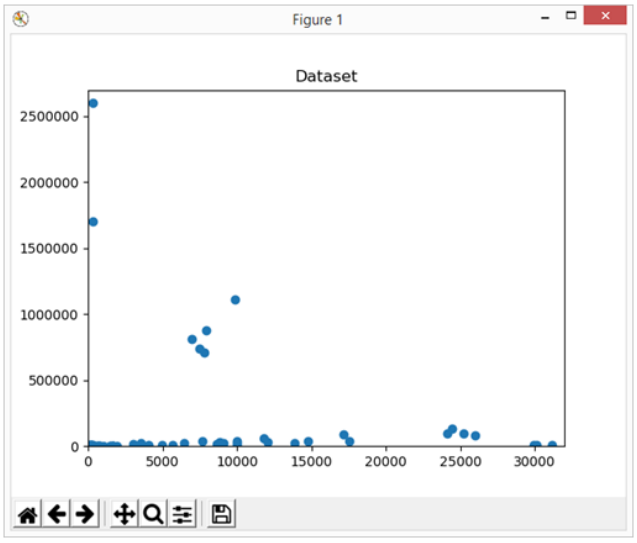
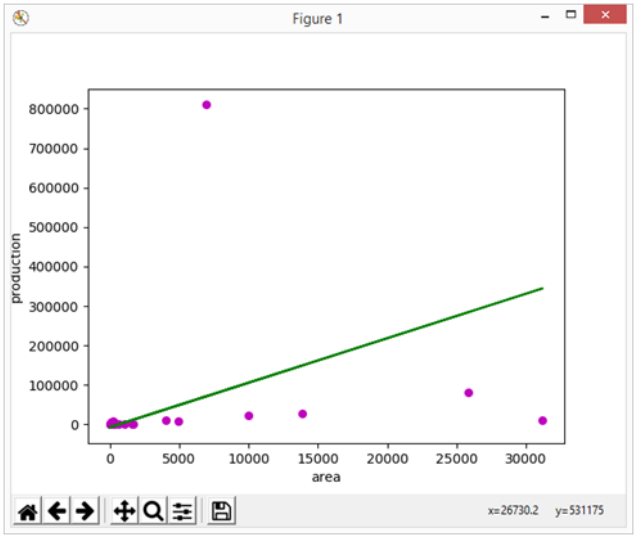


Figure 1: Overall production of crops in Ariyalur

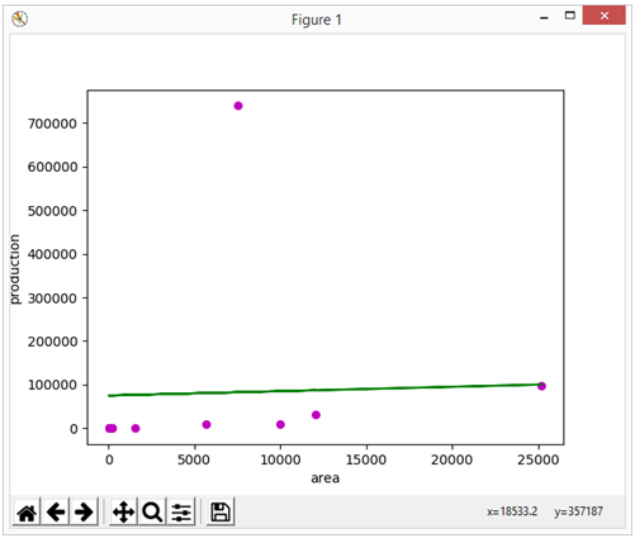
Data analysis using k-means



Data plot for 2009 using regression

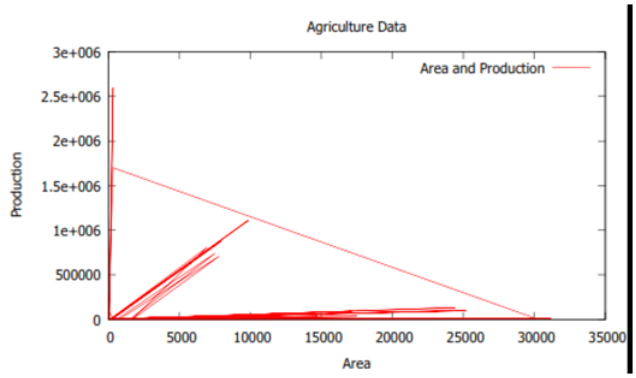


Data plot for 2010



Regression coefficient was very low in 2009. However, it increased drastically in 2010. Then, again it was reduced in 2011. Then, it was slightly increased in 2012, and then decreased in 2013.

Data analysis using gnu plot



Coimbatore District

In Kharif season of 2009, production of rice was much higher compared to small millets. Other crops were not produced during Kharif season. Production of coconut was highest followed by banana and sugarcane in the whole year.

In the whole year of 2010, sugarcane was produced. Its production compared to previous year increased by approximately 4.5%. In Kharif season of the same year, production of maize was maximum followed by production of groundnut and rice. However, its production is reduced by 20.9% compared to previous year 2009 as it was available for the whole year in 2009. The production of groundnut was improved by 3.4% in Kharif season of 2010, although it was produced whole year in 2009, however, its production was less compared to coconut, banana and sugarcane in 2009.

In Khariff season of 2011, production of maize was highest followed by groundnut, jowar, rice, cotton (lint). The production of ragi and jowar was very low. In Rabi season of 2011, production of maize was highest followed by jowar and groundnut. Production of other crops was very low.

The correlation coefficient was highest for Coimbatore compared to other 4 districts.

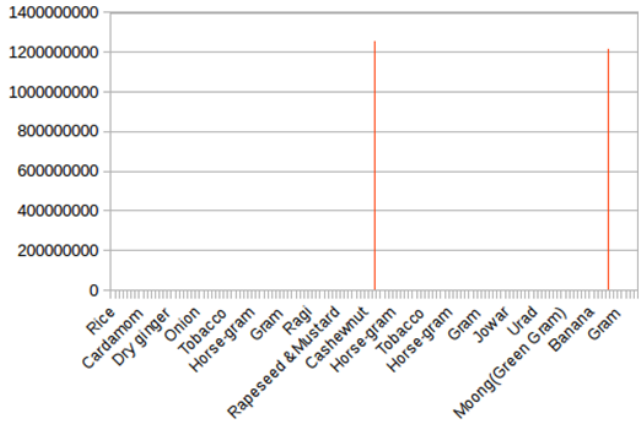
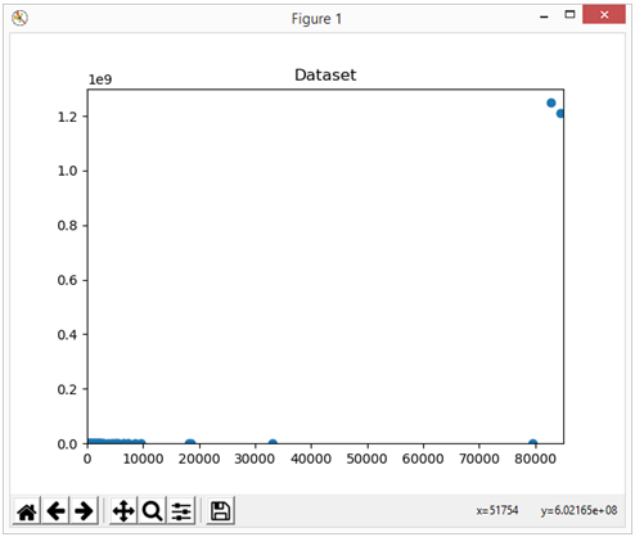


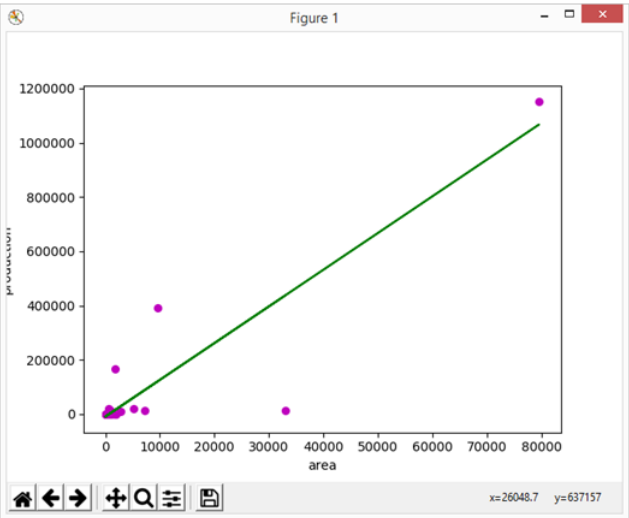
Figure 2: Overall production of crops in Coimbatore

Data analysis using k-means

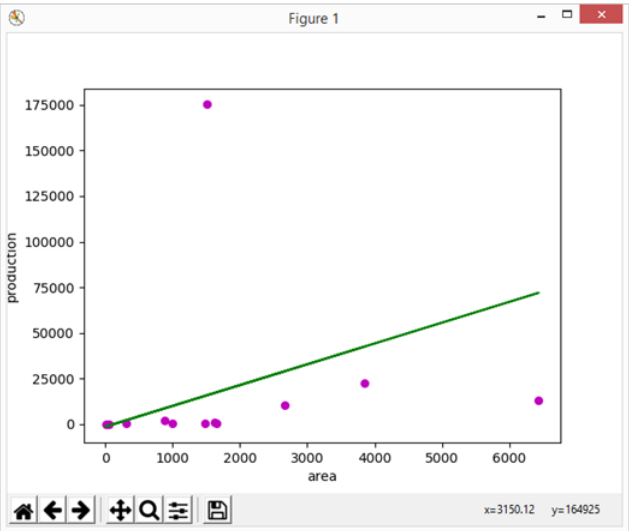


Data analysis using regression

For 2009



For 2010



Cuddalore District

In Kharif season of 2009, production of rice was very high compared to other crops. Production of small millets was one-sixth of rice production.

Production of sugarcane was very high compared to other crops in the whole year of 2009 and 2010. Its production was increased by 12.13% in 2010. However, its production was reduced by 11.08% in 2011. Production of rice was very high compared to other crops in Khariff season of 2010. However, its production decreased by 6.64% compared to 2009.

Production of rice was high compared to other crops in Kharif season of 2011. However, its production reduced by 17.66% compared to previous year. Production of groundnut was high compared to other crops in Rabi season of 2011. Production of coconut was very high in whole year of 2011. Production of coconut was very high in whole year of 2011.

In the whole year of 2012, sugarcane was produced and its production was very high compared to other crops in whole year of 2012. Its production was increased by 16.87% compared to 2011.Production of rice was very high compared to other crops in Kharif season of 2012. Its production was increased by 36.67% compared to 2011.

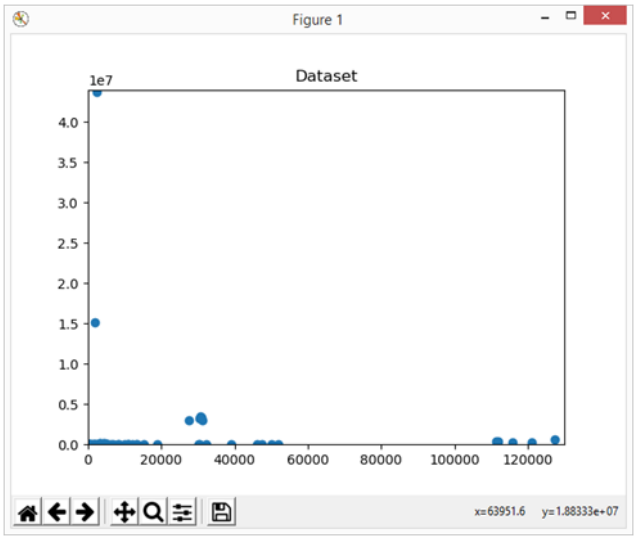
In 2013, rice production was very high compared to other crops in Kharif season. Its production was increased by 66.8%. Production of urad crop was highest followed by groundnut in Rabi season and production of coconut was maximum for whole year.

The correlation coefficient was negative for this district and it was least compared to remaining 4 districts.

The decision tree algorithm could predict that in 2009, production of urad crop was lower compared to banana.

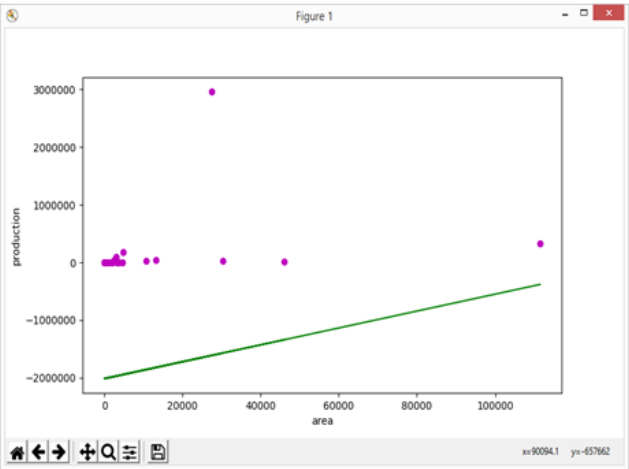
In 2010, it was predicted that maize production was much higher than sunflower.

Data analysis using k-means

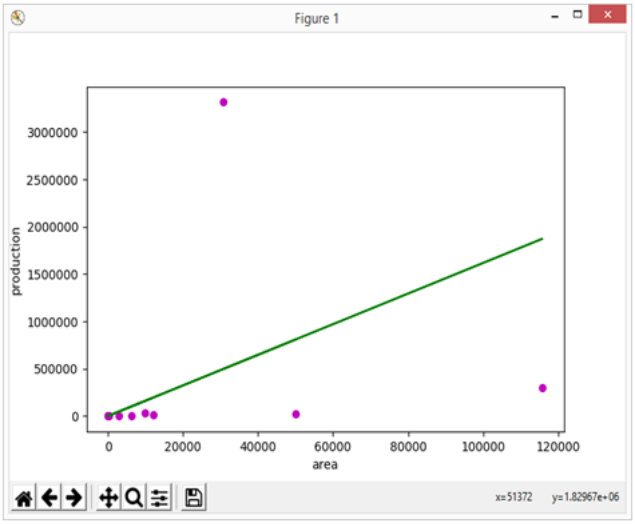


Data analysis using regression

2009



2010



There was drastical increase in regression coefficient from 2009 to 2010. However, in 2011, it was reduced drastically and again in 2012, it was increased drastically, then in 2013, it was decreased by one-eleventh.

Dharmapuri District

In Kharif season of 2009, production of rice was more than 8 times the productionof small millets. Production of sugarcane was very high compared to other crops in whole year of 2009.Gram was produced in Rabi season of 2010. Sugarcane was produced in whole year of 2010. Its production increased by approximately 18.7% from 2009-2010. Production of rice was much greater than other crops in Kharif season of 2010. Production of tur, maize, sunflower, urad, green gra, and horse-gram was very low. Production of rice was four times more than Ragi production.

In Khariff season of 2011, production of rice was highest, followed by groundnut, ragi, cotton (lint), maize and jowar. Rice production increased by 37.57% from 2010 to 2011.

In Rabi season of 2011, production of ragi was much greater, followed by groundnut and cotton (lint). Production of remaining crops was very less. Coconut production was very high in the whole year of 2011. Compared to 2009, its production increased by 830 times.

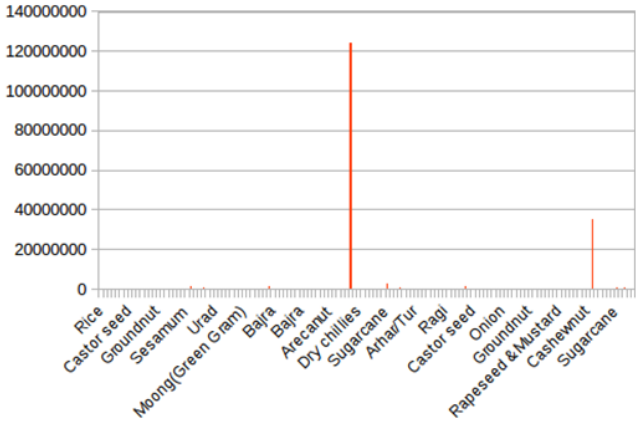


Figure 3: Overall production of crops in Dharmapuri

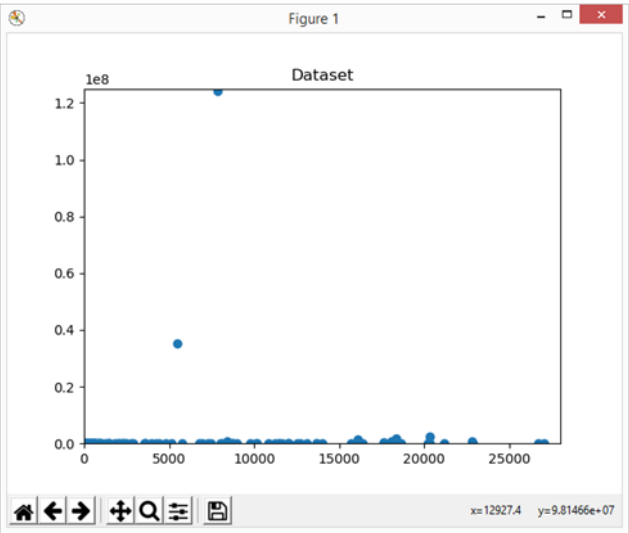
In whole year of 2012, sugarcane was produced. In Rabi season of 2012, gram was produced. In Kharif season of 2012, rice production was highest, followed by cotton (lint), groundnut, ragi, horse-gram and maize. Production of tur and urad was very less. Production of rice reduced by 58.14% compared to previous year. Production of groundnut was reduced by 38.26% compared to previous year Kharif season. Production of cotton (lint) was increased by 47.9% in Kharif season of 2012. Production of ragi was reduced by approximately 47.79% and production of maize was reduced by 56.8%.

In Kharif season of 2013, production of rice was highest followed by maize, ragi, etc. Its production was increased by approximately 2.47 times compared to Kharif season of 2012. Maize production was increased by 9.14 times. Ragi production was increased by almost 2.86 times. In Rabi season, production ofragi was highest, followed by

horse-gram, groundnut and maize. Production of coconut was very high in the whole year of 2013. However, compared to 2011, its production reduced by 71.4%.

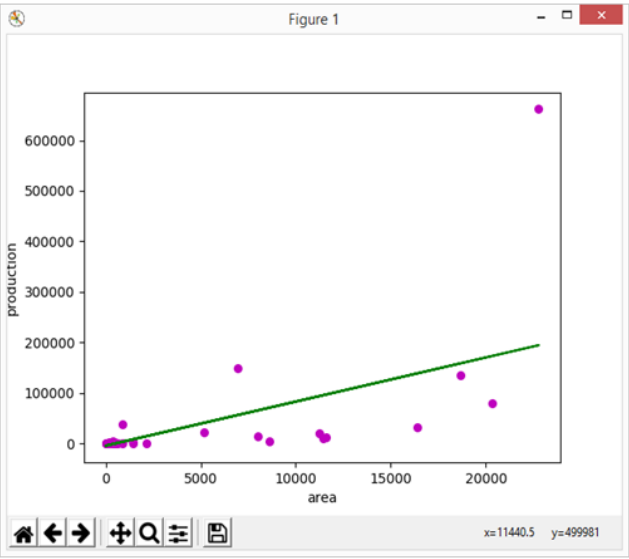
The decision tree algorithm could predict that production of coconut was higher than sugarcane in 2011. In 2009, rice production was higher compared to moong (green gram)

Data analysis using k-means

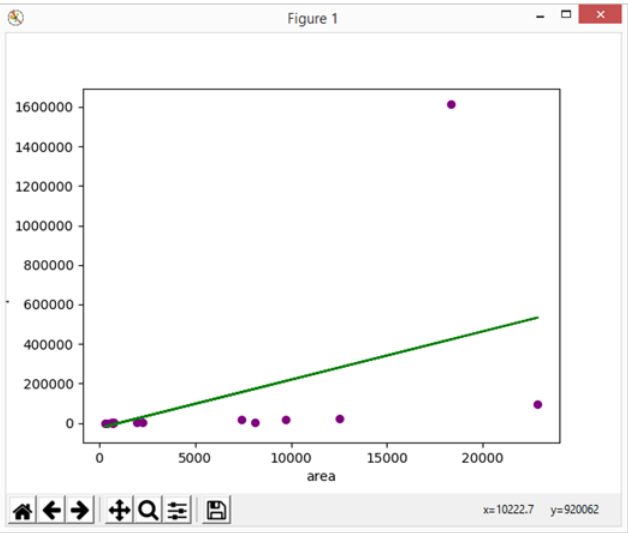


Data analysis using regression

2009



2010



Dindigul District

In Kharif season of 2009, production of rice was very high compared to small millets.

In the whole year of 2009, production of sugarcane was highest followed by coconut, maize and banana.

In Rabi season of 2010, Gram was produced. In the whole year of 2010, sugarcane was produced. In Kharif season of 2010, production of maize was highest followed by rice. Production of sugarcane was decreased by 6.1% in 2010 compared to 2009.

In Kharif season of 2011, production of rice was highest followed by maize and jowar. Production of rice was increased by 5.4% and production of maize was reduced by 57.1% .In Rabi season, production of maize was maximum compared to other crops. In the whole year, coconut production was very high.

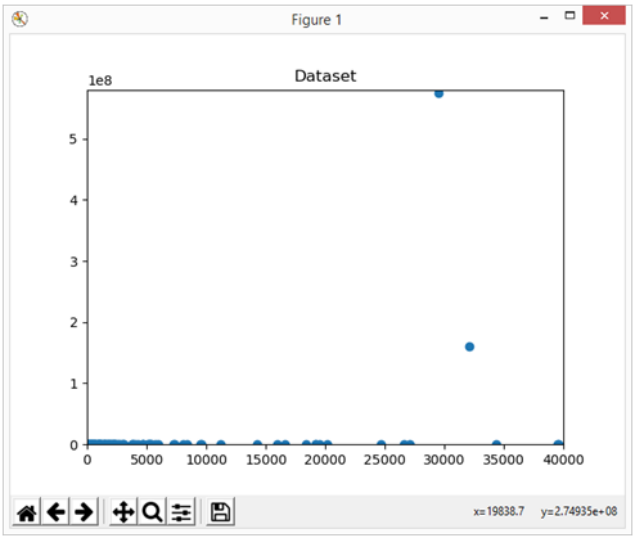
In Kharif season of 2012, maize production was very high compared to other crops. Its production increased by approximately 28%.In Rabi season, only gram was produced. In the whole year of 2012, sugarcane was produced. However, its production was reduced by 60%.

In Kharif season of 2013, production of maize was highest followed by rice, jowar and potato.

Production of rice was reduced by 16.38% and production of maize was reduced by 63.23%. In Rabi season, maize production was maximum followed by jowar.

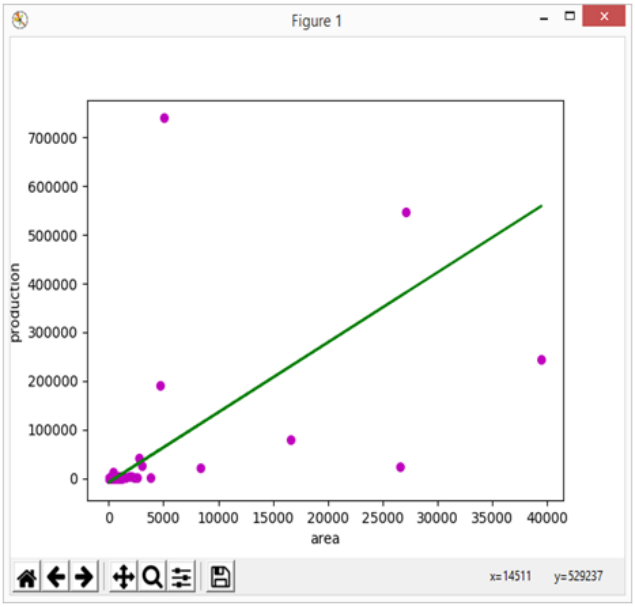
In the whole year, coconut production was very high. Production of sugarcane was increased by 3.87% in the whole year.

Data analysis using k-means

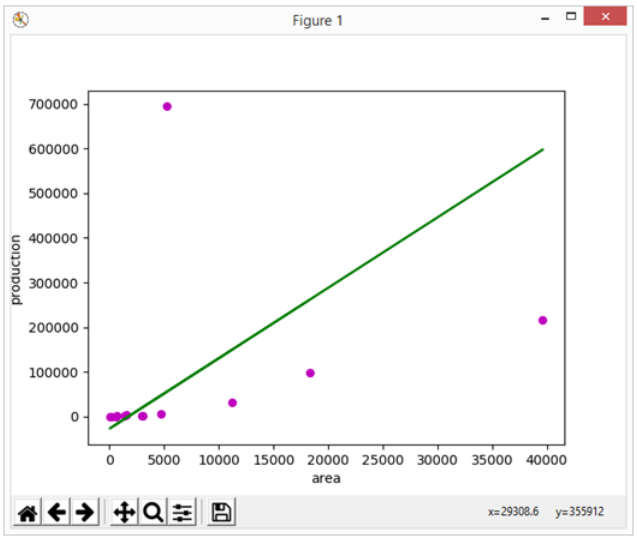


Data analysis using regression

2009



2010



Data analysis using SPSS (Statistical Package for Social Sciences)

Ariyalur

A neural network is a set of tools which are of type non-linear data modeling and they consist of input and output layers and also one or two hidden layers. A Neural network in IBM SPSS Wizard has a feature called Multi-layer Perceptron.

Multi-layer perceptron consists of too many perceptron which are organized into layers. They are used to produce Case Processing Summary, Network Information which gives information on number of latent layers, no. of units in latent layer etc. They are also used to give Model Summary.

Case HandlingCompilation			
		N	Percent
Sample	Training	73	68.9%
	Testing	33	31.1%
Valid		106	100.0%
Omitted		2	
Total		108	

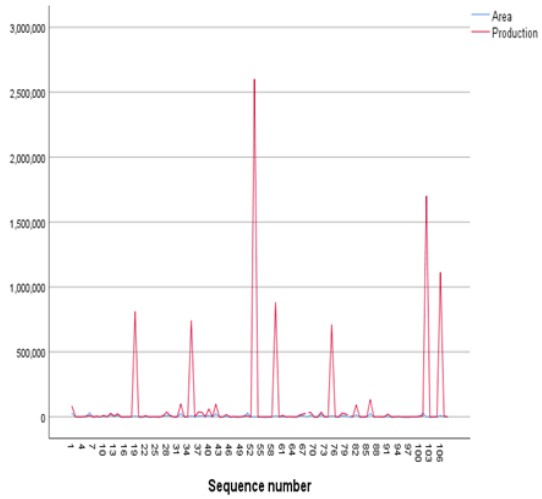
Network Information

Input Layer	Factors	1	Season
		2	Crop Year
	Number of Units		8
Latent Layer	Number of Latent Layers		1
	Number of Units in Latent Layer 1[a]		1
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	Area
		2	Production
		3	Crop
	Number of Units		27
	Rescaling Method for Scale Dependents		Standardized
	Activation Function		Identity
	Error Function		Sum of Squares
a. Excluding the bias unit			

Sequence Plot

Model Information

Name of Model		MOD_2
Series or Sequence	1	Crop_Year
	2	Area
	3	Production
Transformation		None
Non-recurrent Differencing		0
Recurrent Differencing		0
Length of Recurrent Period		No periodicity
Horizontal Axis Labels		Sequence numbers
Intervention Onsets		None
For Each Observation		Values not joined
Applying the model specifications from MOD_2		

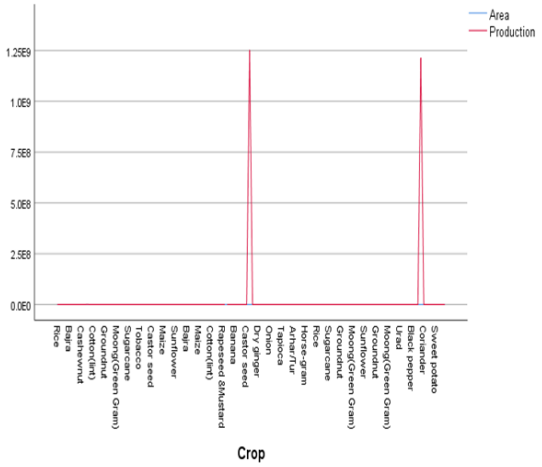


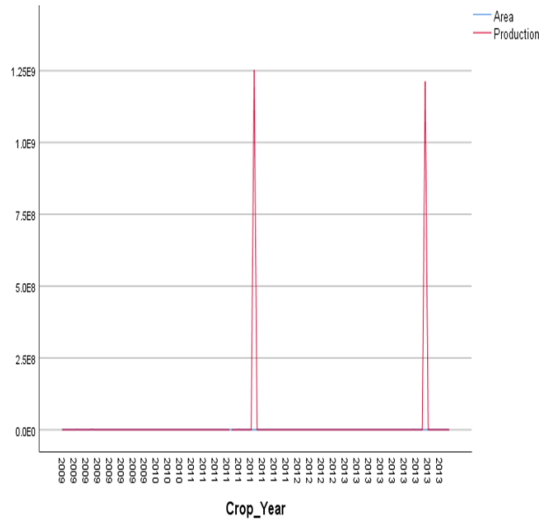
Ts plot of Area and Production

Coimbatore District

The Bayesian inference is made from a perspective that is approached by Posterior distribution characterization for one-sample mean. The marginal posterior distribution of the parameters can be investigated by integrating out the other nuisance parameters, and construct Bayesian confidence intervals to draw direct inference. This is the default setting. Here, it computes the Mode, Mean and Variance for Crop Year, area and production of crops.

Sequence Plot: Sequence Charts are used to plot time-series. Here, they have produced series or sequence length of crop year, Area and Production.





Cuddalore District

NPar Tests

In the procedure of Chi-Square test, variable is tabulated into categories and chi-square statistic is computed. This test is known as goodness-of-fit test. It compares the detected and awaited periodicities in each category. It is used to check whether all groupings contain same distribution of values.

Here, it computes the Residual by comparing Observed and Expected frequencies of each crop year, Area and Production.

Test Statistics

	Crop_Year	Area	Production
Chi-Square	19.880a	21.615b	26.581c
df	4	105	106
Asymp. Sig.	.001	1.000	1.000
a. 0 cells (0.0%) have awaited periodicities less than 5. The least awaited cell periodicity is 23.4.			
b. 106 cells (100.0%) have awaited periodicities less than 5. The least awaited cell periodicity is 1.1.			
c. 107 cells (100.0%) have awaited periodicities less than 5. The least awaited cell periodicity is 1.1.			

Dharmapuri District

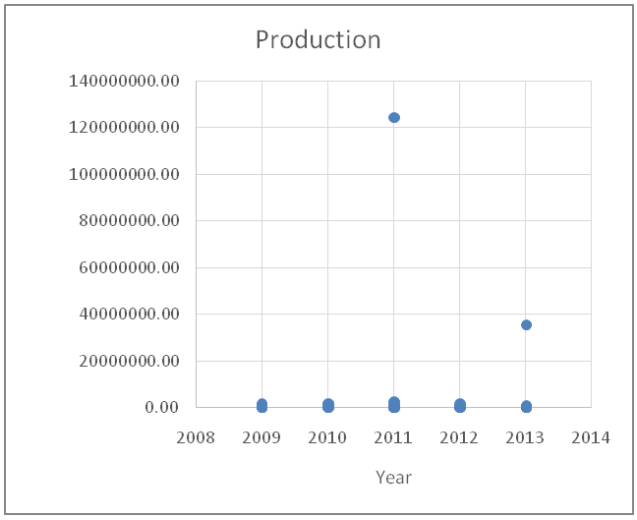
Test Statistics

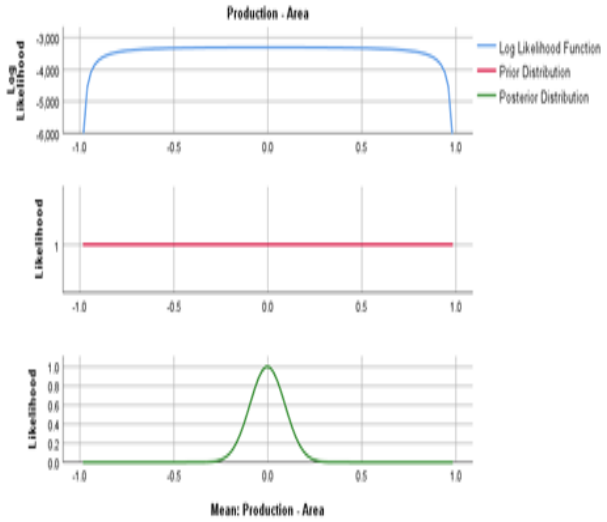
	Year	Area	Production
Chi-Square	24.916a	13.084b	2.857c
df	4	124	122
Asymp. Sig.	.000	1.000	1.000
a. 0 cells (0.0%) have awaited periodicities less than 5. The least awaited cell periodicity is 26.2.			
b. 125 cells (100.0%) have awaited periodicities less than 5. The least awaited cell periodicity is 1.0.			
c. 123 cells (100.0%) have awaited periodicities less than 5. The least awaited cell periodicity is 1.0.			

Model Description

Name of Model		MOD_3
Subordinate Variable	1	Area
	2	Production
Mathematical Statement	1	Linear Equation
Predictor Variable		Year
Constant		Incorporated
Variable Whose Values Label Observations in Plots		Crop Season

Variable Handling Compilation				
		Area	Production	Year
No.of +ve values		131	126	131
No.of 0s		0	0	0
No.of-ve values		0	0	0
No. of missing values	User-Missing	0	0	0
	System-Missing	0	5	0





Dindugal District

Bayesian Correlation

It is obtained by Analyze-> Bayesian Statistics-> Pearson Correlation in SPSS.

This measures the linear relation between two scale variables Production and Area jointly following a bivariate normal distribution.

Posterior Distribution Characterization for Pairwise Correlations

			Production	Area
Production	Posterior	Mode		-0.003
		Mean		-0.003
		Variance		0.008
	95% Credible Interval	Lower Bound		-0.180
		Upper Bound		0.177
	N		117	117
Area	Posterior	Mode	-0.003	
		Mean	-0.003	
		Variance	0.008	
	95% Credible Interval	Lower Bound	-0.180	
		Upper Bound	0.177	
	N		117	117

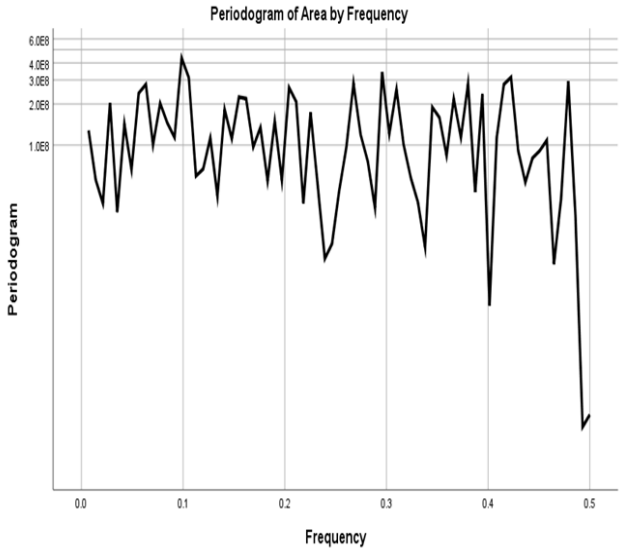
The analysis assumes reference priors c=0

Spectral Analysis

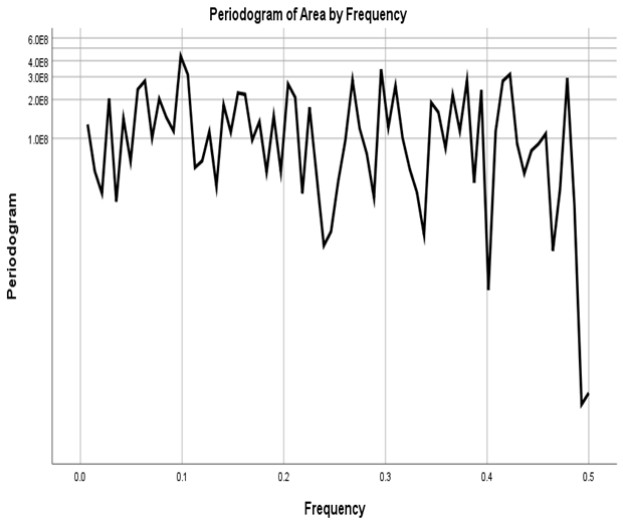
It is the forecasting tool in SPSS. It is used to calculate

Periodo-gram of time series.

Area



Production



CONCLUSION

With machine learning algorithms, the prediction of crops could be predicted. The Decision tree model is used to retrieve names of crops based on area and production. We could get overall production of crops through scatter plots with the help of regression and k-means algorithm. The algorithms generate graph no matter how big the data is. In future, a prediction model used to predict the area and production of particular crops based on given data set might be developed.

REFERENCES

KamilarisA., Karkatoullis A, Prenafeta-Boldi, A review on practice of big data analysis in Agriculture, Computer and Electronics in Agriculture, 143, 2017, 00.23-37

Masayuki Hirafuji ,'A strategy to create Agricultural Big Data',2014 Annual SRII Global Conference, pp. 249-250, 2014, DOI:10.1109/SR11.2014.43

Wang Juyun ,'Prediction of crop yield using Big Data', 2015 8th International Symposium on Computational Intelligence and Design, Vol. 1,pp 255-260, 2015. DOI: 10.1109/ISCID.2015.191

TanmayBaranwal,Nitika,Pushpendra Kumar Pateriya ,,Development of IoT based smart security and monitoring devices for agriculture' , 2016 6th International Conference-Cloud System and Big Data Engineering,pp 597-602, 2016. DOI: 10.1109/CONFLUENCE.2016.7508189

Rakesh Kumar, M P Singh, Prabhat Kumar, J P Singh,'Crop Selection method to maximum crop yield rate using Machine Learning Technique', 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, pp 138-145, 2015 DOI:10.1109/ICSTM.2015.7225403

Md. Tahmid Shakoor, Karishma Rahman, Sumaiya Nasrin Rayta, Amitabha Chakraborty, 'Agricultural Production Output Prediction using supervised Machine Learning Techniques', 2017 1st International Conference on Next Generation Computing Applications, pp: 182-187, 2017DOI: 10.1109/NEXTCOMP.2017.8016196

Thomas Truong, AnhDinh, Khan Wahid ,'An IoT Environmental Data Collection system for fungal detection in crop fields' , 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering, pp 1-4, 2017.