Agriculture Productivity Performance of North-Western Zone in Tamilnadu Using Data Mining Techniques

C. Arul Kumar [1], G. Manimannan[2*]

[1]Assistant Professor, Vivekanandha, College of Arts and Sciences for Women, Tiruchengode, Namakkal, Tamilnadu, INDIA.

[2]Assistant Professor, Madras Christian College, Tambaram, Chennai, Tamilnadu, INDIA.

[*]Corresponding Address:

manimannang@gmail.com

Research Article

Abstract: Agriculture sector is the mainstay and backbone of the Indian economy. Agriculture plays a vital role in the development of a country. It contributes nearly fifteen percent of Gross Domestic Product (GDP) of India. Seventy percent of the population depends on agriculture for their livelihood. Over the past decade the agriculture production in districts of Tamil Nadu had faced increased yield in all the crops. However, agriculture productivity differs from region to region, which needs a detailed investigation. The main objective of this paper is to analyze the agriculture productivity of Major Crops in North-western zones of Tamil Nadu using Data Mining and Geographical Information System (GIS) techniques. The data was collected from secondary source of Department of Economics and Statistics, Tamilnadu during the period of 2003 to 2012. In this study we examined yield variations, spatial patterns and classification of fifteen major crops. The results achieved three different classifications and are labeled High Yield Districts (HYD), Medium Yield Districts (MYD) and Low Yield Districts (LYD) based on their mean yield of various crops.

Keywords: Agriculture Productivity, Productivity Regions, Crop Yield, Factor Analysis, k- mean Clustering Techniques, Geographical Information System and Multivariate Discriminant Analysis

Introduction

Agriculture is the back bone of our India. Agriculture and its allied sectors contribute nearly 15 percent of Gross Domestic Product (GDP) of India. Over 70 percent of the population is significantly dependent on agriculture for their livelihood. This sector provides 50 percent of the employment at the national level and it helping for the growth of the economy. The agriculture development is a requirement not only providing the food security of the growing population but also a number of industrial raw materials. In nutshell, agriculture plays an essential role in the process of economic development of the country. Agriculture continues to the major predominant sector of the Tamilnadu economy. The state has a total area of 1.3 lakh sq. km with a gross cropped area around 63 lakh hectares and the contribution to Gross Domestic Product and state domestic product, has higher compared with other sector. Agriculture development of a country or region is dependent on the production of crops. Several efforts have been taken to increase the production and productivity level at different points of time. Agriculture productivity measures that the area are performing less or higher compared with other regions nearby. The concept of agriculture productivity has been extremely used to explain the spatial organization and pattern of agriculture. Productivity is generally is considered into two directions, i) productivity of land ii) Productivity of infrastructure, productivity of land is linked with infrastructure closely. Productivity of land is an important factor of agriculture since it is permanent fixed factor among the three categories of input land, labor and capital. Agriculture productivity of land explained by production of crops in terms of yield per unit. Spatial analysis of agriculture productivity is very important because it can highlight the structure and pattern of production. (Dharmasiri, 2009) Agriculture productivity is defined by several researchers with their views and discipline. Agriculture productivity may be defined as the 'ratio of index of agriculture output to index of total input used in farm production' (shafi, 1984) In geography agriculture productivity is defined on "output per unit of input" or "output per unit of land area". "Yield per unit" should be considered to indicate agriculture productivity (Singh and Dhilion, 2000).

Several studies have been made to measure the agriculture productivity and spatial pattern using different methods. Kendal used factor analytic approach and calculated latent roots to assign weight of individual crop production for the assessment of agricultural productivity patterns which emerged in England. Stamp (1958) suggested a method for measuring the agricultural productivity by converting the total agricultural productivity in calories. Shafi (1960) applied kendal ranking coefficient method for measuring the agriculture efficiency in the state of Uttar Pradesh. W. Tadesse and E

Bekele (2001) attempted factor analysis to determine the relationship between the yield components and morphological characters of several genotypes of grasspea.

Review of Literature

In Agriculture study so many researches done in the past decade. In this section we are collecting some important research papers and present their findings and techniques. Jahan mohan *et al.,* (2005) in his study found that area and production of cereal crops registered a negative trend in the entire agro climatic zone with the exception of western and southern zone that showed positive trend in the area of cotton. Dharmasiri (2009) had applied average productivity index to measure the agriculture productivity in Srilanka. Kalaivani *et al*. (2010) used Compound Growth Rate to measure the growth actions of area, production and yield for selected crops in Tamilnadu. In his study maize was recorded a positive trend in Tamilnadu. Sakthi mandal (2012) studied the spatial variation of agriculture productivity using Z score model and categorized the blocks as Very high, High, Medium and Low of agricultural productivity in south 24 PGS, Districts of West Bengal. Muthumurugan *et al*. (2012) suggested composite index analysis to study the agriculture development of Tamilnadu, he classified the districts based on the index value as highly developed, medium developed and low developed. Pajer Mansaram Pandit (2012), his study revealed that road, railway, agricultural labor, bank, good climate and modern technology are the important factors in the agriculture development of Nashik District. A S Rotaru *et al*. (2012) demonstrated factor analysis in agriculture by downsizing the data with various crops in the regions of Romania. Shafiquallah (2013) applied the technique Z-score model to study regional disparities and agricultural development in Uttar Pradesh.

Study Region

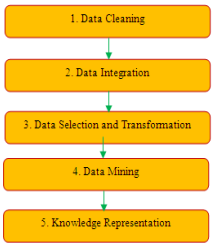
Tamilnadu has been classified into seven Agro Climatic Zones based on soil characteristics, rainfall distribution, irrigation pattern, cropping pattern and other social characteristics. The present study we are chosen North Western Agro climatic zone of Tamilnadu. This zone comprises the revenue districts of Dharmapuri, Krishnagiri, Salem and Namakkal. This zone is situated between 11and $12^{0}55^{'}$ north latitude and $77^{0}28^{'}$ and $78^{0}50^{'}$ east longitude; this zone covered an area of 16,150 sq km equivalent to 12.4 percent of the state area. Out of total geographical area of 17.31 lakh hactares, 8.01 lakh hectares are cultivated. The Average annual rainfall of the zone is 878 mm. This zone enjoys rainfall from both South West and North West monsoon seasons. The average temperature ranges maximum from 20° C to 42° C and minimum from 10°C to 31°C. Paddy, Maize, Ragi, Cumbu, Tapicaco and Sugarcane are the major crops cultivated in this zone. The main objective of this paper is to analyze the agriculture productivity, crops yield and spatial pattern and classification of North-western agro-climatic zone of Tamil Nadu.

Database and Methodology

The agriculture productivity data published by Department of Economics and Statistics, Chennai, Statistical Hand Book of Tamil Nadu was considered as database. The data consists of crops of each agro climatic zone for the time period of ten years (from 2003 to 2012). Among the listed crops, number of crops varied over the study period owing to removal of those crops for which the required data are not available. In this study, 15 crops are carefully chosen among the many that had been used in previous studies. These 15 mostly cultivated crops are chosen to asses agro productivity performance in the zone, they are Paddy, Cholam, Cumbu, Ragi, Maize, Bengal gram, Red gram, Green gram, Black gram, Horse gram, Tapioca, Ground nut, Gingely, Coconut and Sugarcane.

Data mining Techniques

Data Mining or Knowledge Discovery in Databases (KDD) is the process of discovering previously unknown and potentially useful information from the data in databases. In the present context data mining exhibits the patterns by applying few statistical data mining techniques namely, factor analysis, k-means clustering, Discriminant Analysis and Geographical Information System (GIS). As such KDD is an iterative process, which mainly consist of the following figure on the data collected;



Of these above iterative process 4 and 5 are most important. If appropriate techniques are applied in 5, it provides potentially useful information that explains the hidden structure. This structure discovers knowledge that is represented visually to the user, which is the final phase of data mining.

Factor Analysis

Factor Analysis is a generic term for a family of statistical techniques concerned with the reduction of a set of observed variables in terms of a small number of latent

factors. The primary purpose of Factor analysis is data reduction, reducing data complexity by reducing the number of variables being studied. It has been developed for analyzing relationship among a number of measureable entities. It describes the covariance relationship among many variables in terms of few underlying by unobservable, random quantities called factors. Factor model is motivated by the following arguments.

1. To identify those variables with a particular group that are highly correlated among them but have small correlations with variables in different groups.
2. To reduce a large no of variables to a small number of factors.

Factor Extraction

There are several criteria for the number of factors to be extracted. Among the different extraction methods, the most widely used method; the principal component analysis is applied in this study to assess the performance of yield in the specified region. The eigen value greater than one rule suggested by Kaiser has been used in this study to extract the factors.

Factor Rotation

After deciding the number of factors extracted, the next step is to determine the method of Rotation. The rotation is used to simplify the factor structure and to achieve a more meaningful and interpretable solution. In the present study, factor analysis is initiated to uncover the patterns underlying crops variables. In factor extraction method the number of factors is decided based on the proportion of sample variance explained. Orthogonal rotations such as Varimax and Quartimax rotations are used to measure the similarity of a variable with a factor by its factor loading (Everitt and Dunn, 2001; Hair, Black, Babin and Anderson, 2010).

Non- Hierarchical Clustering Techniques

Cluster analysis is a multivariate method which aims to classify a sample objects on the basis of a set of measured variables into a number of different groups such that similar subjects are placed in the same group. Cluster analysis encompassed a number of different algorithms and methods for grouping objects of similar kinds into respective categories. The goal of clustering algorithm is to group the objects into a set of meaningful subclasses. This algorithm is used to discover structure in data without providing an explanation and determine the grouping of data themselves. In the present study one of the popular clustering algorithms suggested by Mac Queen (1967) known as k-means is used to identify the k-classes in the data set. The k-means method follows the following steps.

Step 1: Specify the number of clusters and, arbitrarily or deliberately, the members of each cluster.
Step 2: Calculate each cluster's centroid, and the distances between each observation and centroid. If an observation is nearer the centroid of a cluster other than the one to which it currently belongs, re-assign it to the nearer cluster.
Step 3: Repeat Step 2 until all observations are nearest the centroid of the cluster to which they belong.
Step 4: If the number of clusters cannot be specified with confidence in advance, repeat Steps 1 to 3 with a different number of clusters and evaluate the results

Multivariate Discriminant Analysis

Multivariate Discriminant function analysis builds a predictive model for group membership. The model is composed of a discriminant function based on linear combinations of predictor variables. Those predictor variables provide the best discrimination between groups. A discriminant score can be calculated based on the weighted combination of the independent variables. The main purpose of the discriminant analysis, to maximally separated the groups, to determine the most economical way to separate the groups and finally, to discard the variables which are little related to group distinctions. It is similar to regression analysis. This is achieved by the statistical decision rule which maximizes the between group variance relative to the within group variances. The discriminant ananlysis derives the linear combination from an equation that takes the following form

$Z = W_1 X_1 + W_2 X_2 + \ldots\_ W_n X_n$, where, Z - Discriminant score, W_i - discriminant weights and X_i - Independent variables. In this research paper, we are using MDA for classifying the agriculture yield data in the North West Zone of Tamilnadu.

Geographical Information System

A Geographic Information System (GIS) is a system designed to capture, store, manipulate, analyze, manage, and present all types of geographical data. The first known use of the term "Geographic Information System" was by Roger Tomlinson in the year 1968 in his paper "A Geographic Information System for Regional Planning". GIS or spatial data mining is the application of data mining methods to analyze spatial data. Data mining, which is the partially automated search for hidden patterns in large databases, offers great potential benefits for applied GIS-based decision making. In the present study, GIS map is used to exhibit groups graphically and judge the nature of overall performance of the agriculture yield. A brief step-by-step algorithm to classify the crops during each of the study period based on their overall performances is described below:

Step 1: Factor analysis is initiated to find the structural pattern underlying the data set.

Step 2: k –means analysis is used to partition the data set into **k**-clusters using the original parameters in **Step 1** as input.
Step 3: Discriminant analysis is then performed with the original variables by considering the groups formed by the **k**-means algorithm.
Step 4: GIS districts map is drawn with the standardized canonical discriminant function and centroid values.

Results and Discussions

In this section we discuss our results of different techniques used in this study. Factor analysis is extended with the techniques of Varimax and Quartimax criterion for orthogonal rotation. Even though the results obtained by both the criterions were very similar, the varimax rotation provided relatively better pattern of crops. Consequently, only the results of varimax rotation are reported here. We have decided to retain 81 percent of total variation in the data, and thus accounted consistently four factors for each districts with eigen values little less than or equal to unity. Table 2 shows variance accounted for each factors and in each districts. The crops yield loaded in the factors are presented in *Table 3 to 6.* Only those crops with higher loadings are indicated by factor analysis. From *Tables 3 to 6* it is very clear that the clustering of crops is stable during the study period. We observed slight changes in factor loadings during the periods considered. The differences in factor loadings may be due to statistical variations in the original data. Formations of extracted factors are named as Seasonal crops, Utility value crops, Irrigated crops, Cash crops. After performing factor analysis, the next stage is to assign initial group labels to each district. Step 2 of the algorithm is explored with original parameters by Step 1, by conventional **k**-means clustering analysis. Formations of clusters are explored by considering 2-clusters, 3-clusters, 4-cluster and so on. Out of all the possible trials, 3-cluster exhibited meaningful interpretation than two, four and higher clusters. Having decided to consider only 3 clusters, it is possible to rate districts as High Yields Districts (**HYD)**, Moderate Yields Districts (**MYD)** or Low Yields Districts (**LYD)** depending on whether the districts belonged to Cluster 1, Cluster 2 or Cluster 3 respectively. Cluster 1 (**HYD**) is a group of districts that have high yields for the crops, indicating that these districts are performing well. The districts with lower yields for the crops are grouped into Cluster 3 (**LYD**). This suggests that Cluster 3 is a group of districts with low yields. Cluster 2 (**MYD**) are those districts which perform moderately well as compared to the Cluster 1 and Cluster 3. Incorporating the results for each district and in each year, only the summary statistics are reported in *Table 1.*

Table 1: No. of years in each group

District	Group 1	Group 2	Group 3
Dharmapuri	5	2	3
Krishnagiri	3	1	6
Namakkal	1	2	7
Salem	5	1	4

1 – HYD, 2 – MYD, 3 – LYD.

Table 2: Eigen Values and Percentage of Variance explained by factor

Factor	Dharmapuri		Krishnagiri	
	Eigen Value	Variance	Eigen Value	Variance
1	5.930	39.535	4.361	29.074
2	3.362	22.416	3.381	22.541
3	1.960	13.064	2.767	18.445
4	1.591	10.606	2.084	18.892
Total		**83.620**		**83.953**

Factor	Namakkal		Salem	
	Eigen Value	Variance	Eigen Value	Variance
1	5.011	33.406	6.304	50.141
2	2.700	18.002	2.299	15.326
3	2.403	16.017	1.850	12.335
4	2.174	14.454	1.767	11.778
Total		**81.920**		**89.583**

Table 3: Variable in Rotated Factors (Dharmapuri)

Factor No	1	2	3	4
Factor Name	**Utility Value Crops**	**Irrigated Crops**	**Cash Crops**	**Seasonal Crop**
	Horsegram	Bengalgram	Black	Greengram
	Maize	Paddy	gram	
	Sugarcane	Ragi	Tapioca	

	Redgram		Gingely	
	Cholam			
	Groundnut			
	Cumbu			
	Cocanut			

Table 4: Variable in Rotated Factors (Krishnagiri)

Factor No	**1**	**2**	**3**	**4**
Factor Name	**Utility Value Crops**	**Irrigated Crops**	**Cash Crops**	**Seasonal Crop**
	Horsegram	Ragi	Black	Redgram
	Sugarcane	Paddy	gram	Greengram
	Maize	Bengalgram	Tapioca	
	Cumbu		Gingely	
	Cocanut			
	Cholam			
	Groundnut			

Table 5: Variable in Rotated Factors (Namakkali)

Factor No	**1**	**2**	**3**	**4**
Factor Name	**Utility Value Crops**	**Irrigated Crops**	**Cash Crops**	**Seasonal Crop**
	Horse Gram	Bengalgram	Black	Greengram
	Maize	Paddy	gram	
	Sugarcane	Ragi	Tapioca	
	Redgram	Bengalgram	Gingely	
	Cholam			
	Groundnut			
	Cumbu			
	Cocanut			

Table 6: Variable in Rotated Factors (Salem)

Factor No	**1**	**2**	**3**	**4**
Factor Name	**Utility Value Crops**	**Irrigated Crops**	**Cash Crops**	**Seasonal Crop**
	Maize	Paddy	Cumbu	Tapioca
	Groundnut	Sugarcane	Blackgram	**Bengalgram**
	Ragi	**Redgram**	**Gingelly**	
	Greengram			
	Cocanut			
	Cholam			

Table 7: Wilk's Lambda

Districts	**Test of Functions**	**Wilks' Lambda**	**Chi-square**	**df**	**Sig.**
Dharmapuri	1 through 2	.004	22.09	14	.077
Krishnagiri	1 through 2	.144	7.753	14	.902
Namakkal	1 through 2	.016	16.51	14	.283
Salem	1 through 2	.002	24.68	14	.038.

Table 8: Classification Results*

Districts	**Group 1**	**Group 2**	**Group 3**
Dharmapuri	5	2	3
Krishnagiri	3	2	5
Namakkal	2	2	6
Salem	6	1	3

* 98.7 % original grouped classes correctly classified

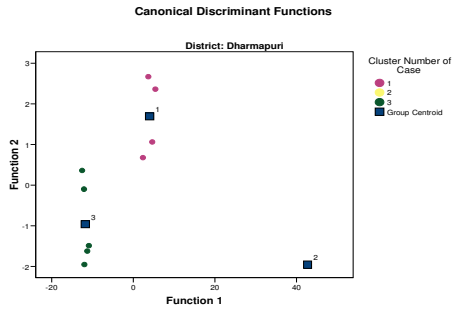


Figure 1: Classification Map of Dharmapuri

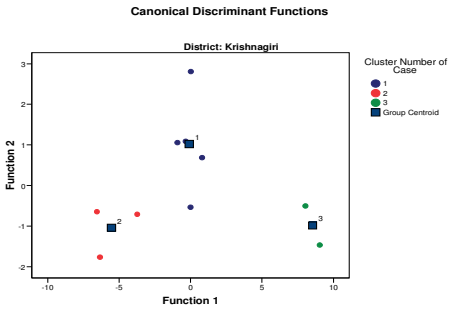


Figure 2: Classification Map of Krishnagiri

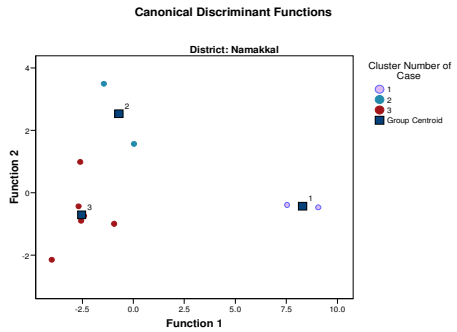


Figure 3: Classification Map of Namakkal

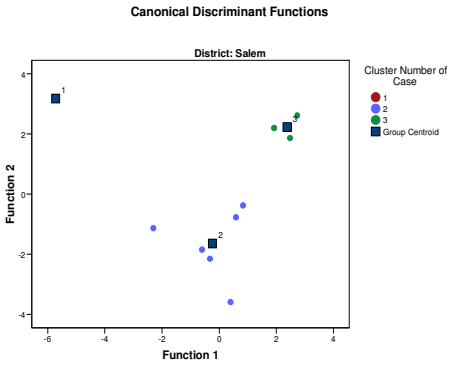


Figure 4: Classification Map of Salem

In *Table 1* provides the groupings done by cluster analysis. *Figure* 1 to 4 shows the groupings of crops into 3 clusters for each districts of the study period in Table 1, we rated the members in the first cluster as High Yield District (**HYD**), and the second as Moderate Yield District (**MYD**) and the third as Low Yield District (**LYD**). Crops belonging to **HYD** category are the ones that perform better than those of **MYD** and **LYD**. Similarly the crops belonging to **MYD** category are superior to those of **LYD**, indicating the members in the category **LYD** are at a low report in terms of the crops yield considered in the present analysis. Table 7 shows that Wilks' lambda is a measure of how well each function separates cases into groups. Smaller values of Wilks' lambda indicate greater discriminatory ability of the function. The associated chi-square statistic tests the hypothesis that the means of the functions listed are equal across groups. The small significance value indicates that the discriminant function does better than chance at separating the groups. In this study, Salem district is significantly different from other districts. The classification matrix is reported in table 8. The pictorial representation of GIS mapping is drawn using the standardized discriminant functions evaluated at the group centroids. From the GIS maps in *Figure 5*, it is evident that the three groups of rated crops are very well separated and represented in the GIS maps for four districts. The Spatial pattern of districts (*Figure 5*) shows that Salem district get High Yield, Krishnagiri and Namakkal get Moderate Yield and Dharmapuri gets Low yield performance during the study period. The low yield may be due to shortage of rainfall, Irrigated land usage and variations in the cropping pattern during the study period.

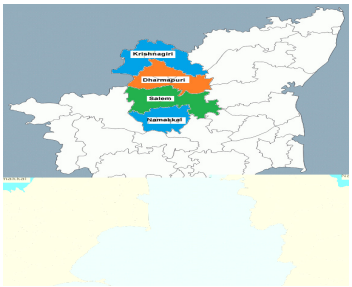


Figure 5: Spatial Patterns of Districts Map

MYD **LYD** **HYD**

Conclusion

In this paper we identified meaningful groups of districts that are classified as best with respect to their total yield in terms of crops using data mining and GIS techniques. An attempt is made to analyze the agriculture productivity yield data relating to major crops of North-western agro-climatic zone over a period of ten years from 2003 to 2012. The present analysis shows that only 3 groups could be meaningfully formed for each district.

This indicates that only 3 types of yields existed over a period of ten years. Further, the districts find themselves classified into *High Yield District* (HYD)*, Moderate Yield* District (MYD) and *Low Yield District* (LYD) categories depending on the crops. Agriculture Analysts can make use of these techniques of classifying, and the districts can project the yield on the basis of crops that has been considered in this study. Geographical Information System (GIS) may be used for crops yield, evaluation, and tracking changes in crops, among other uses. A generalization of the results is under investigation to obtain a set of 3 groups of yields for any given year.

References

1. Brain S. Everitt and Graham D (2001), Applied Multivariate Data Analysis, 2/e, Oxford University Press, Newyork.
2. Everitt, B.S., Landau, S. and Leese, M. (2001), Cluster Analysis, Fourth edition, Arnold.
3. Tadesse W and Bekele E (2001), Factor Analysis of components of yield in grasspea, lathyrus Lathyrism Newsletter 2 pp 91-93
4. Lal Mervin Dharmasiri (2010), Measuring Agricultural Productivity using the Average Productivity Index, Srilanka Journal of Adanced Social Studies vol.1 – No.2
5. Kalaivani M and A. Saravanadurai (2010), Growth Actions of Selected Cereal crops in Tamil Nadu state, International Journal of Applied Biology and Pharmaceutical Technology, Vol I; issue 3, pp 778-785.
6. Sakthi Mandal and Arijit Dhara (2012), Measurement of Agricultural productivity and levels of development in south 24 pargans district, West Bengal, International Journal of Agricultural Science and Research Vol.2 Issue 4 pp 91-98
7. Muthumurugan *et al*. (2012), Composite Index Analysis of Inter-Regional Variations in Agricultural Development of Tamil Nadu, International Journal of Social Sciences and Inter Disciplinary Research Vol.1 No.4 pp 58-62
8. Pager Mansaram Pandit (2012), Agricultural Development and Land Use Pattern in Nashik District of Maharastra, Mediterrronean Journal of Social Sciences, Vol.3 (16) pp. 151-161
9. Rotaru AS *et al*. (2012), Usefulness of Principal Component Analysis in agriculture, Bulletin UASVM Horticulture 69(2) pp 504-509
10. Safiquallah (2013), Impact of Regional disparities on Agricultural development in Uttar Pradesh – A Geographical Analysis, Global Journal of Human Social Science, Geography, Geo science, Environmental Management. Vol 2 issue 5 version1.0 pp 36-46
11. Stamp L.D (1958), The measurement of Land Resources, the Geographical Review Vol.48 No.1 pp.110-116
12. Kendal M.G (1939) The Geographical Distribution of Crop Production in England, Journal of Royal Statistical Society, Volume 102 pp. 21-26
13. Season and Crop Report, Department of Economics and Statisitcs, Chennai. (various issuses)
14. State of Indian Agriculture 2012-2013, Government of India, Ministry of Agriculture, Department of Agriculture and Cooperation, Directorate of Economics and Statistics, New Delhi
15. Statistical Hand Book of Tamilnadu 2012-2013, Department of Economics and Statisitcs, Chennai.
16. Shafi M (1960), Measurement of Agricultural efficiency in Uttar Pradesh, Economic Geography, Volune 36, No. 34 pp.296-305
17. Shafi M (1984), Agricultural productivity and Regional imbalances, New Delhi, Concept Publishing Company.
18. Singh J., *and* Dhillion, S.S, (2000), Agricultural Geography (2nd edition) New Delhi, Tata McGraw Hill.
19. Jahan Mohan K.R *et al*. (2005), Growth Performance of Agriculture in Agro-Climatic Zones of Tamil Nadu, Agriculture Situation in India, Vol. LXI, pp. 679-686.