# Cluster Analysis Evaluation of Price variation of Catla Fish in India

Raghavendra Prabhu
School of Information Sciences
Manipal University

H.G.Joshi
Department of Commerce
Manipal University

## ABSTRACT
Each and every sector in this modern world is undergoing rapid change due to the impact of IT field. Most of the social problems are analyzed by using various statistical and computational tools. In India, Fisheries is an unorganized sector and fishermen community are socially backward due to the exploitation of the middlemen. This paper, firstly evaluates the trends in the variation of price of Catla fish in different coastal regions of India from 2006 to 2013 and then analyze and predict the data for 17 markets of 5 states of India between the year 2006 to 2013 using clustering algorithm. At the end the result illustrates the cluster relation based on state, market, date of arrival and average price of Catla fish.

## General Terms
Sequence Clustering Algorithm

## Keywords
Clustering algorithm, Data Mining, Fish Markets,

## 1. INTRODUCTION
The problem in developing countries is not that markets are absent but that they are functioning badly. The prices of fish varies from district to district even though there is a good connectivity through road and train. Most developing countries have poorly functioning fish markets with an uneven distribution of information, which hinders negotiations and limits what can be contracted by the farmers. It is important to analyze the variation of price and demand of fish in various markets to understand the demand cycle and disparity of price in the region.

Catla, the second most important species after rohu is used as the surface feeder component in Indian major carp polyculture systems. Catla are marketed mostly in local markets, where they are sold fresh. Annual export of Catla fish from India has increased dramatically from year 2004 ( Figure 1). Post-harvest processing and value-addition of this species is almost non-existent at present in any of the producing countries.

Despite the wide variety of techniques available for grouping individuals into market segments on the basis of survey data, clustering remains the most popular and most widely applied method [1]. Clustering is the process of grouping or making sets of similar or nearly similar type of physical or abstract objects. The groups thus formed are known as clusters. It is the process of grouping the data into classes or clusters, so that the objects within the same cluster have higher degree of similarity in comparison to one another but are very much dissimilar to the objects in different clusters [2].

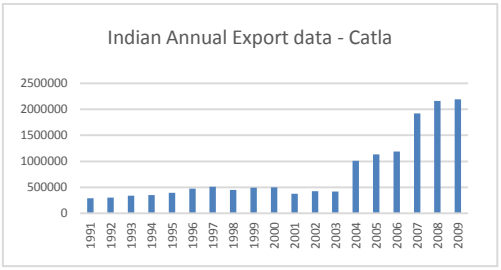


**Figure 1 (Data Source: FAO)**

The Sequence Clustering algorithm is used to group or cluster data based on a sequence of previous events. For example, web application users can browse the site through a variety of path. This algorithm can group end-users based on their sequence of pages through the site to help analyze users and determine if some paths are more profitable than others. This algorithm can also be used to predict, such as predicting the next page a user may visit.[3].

In this paper we use Sequence Clustering Algorithm of Business Intelligence tool of SQL Server Analysis Service 2005 to analyze and predict the price of Catla fish for 17 markets of 5 states of India during the period of 2006 to 2013.

## 2. METHODOLOGY
Cluster Analysis is a data mining technique used to group a set of objects in such a way that objects in the same group are more similar to each other compared to the other groups. It is a main task of exploratory data mining, and a common technique for statistical data analysis used in many fields, including pattern recognition, image retrieval, machine learning, information retrieval and bio-informatics. Cluster analysis itself is not one specific algorithm, but the general task to be solved. Clusters can be achieved by various algorithms, it differs by what constitutes an efficient cluster [3]. Popular clustering techniques include k-means clustering and expectation maximization (EM) clustering**.** k-means clustering is a method of vector quantization , originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cell.

Following steps are involved to build the clusters (Figure 2)

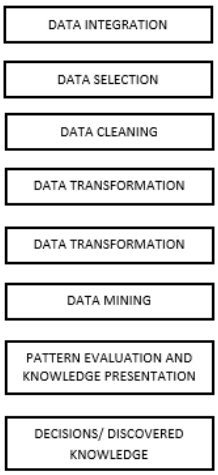


**Figure – 2 (Data source: [4] )**

1. Data Integration: From all the different sources data is collected and integrated.
2. Data Selection: Selection of data that are useful to build the clusters.
3. Data Cleaning: The data we have collected may contain errors, missing values, noisy or inconsistent data. Appling different techniques to get rid of such anomalies.
4. Data Transformation: Transforming of data into a form appropriate for mining. The techniques used to accomplish this are smoothing, aggregation, normalization etc.
5. Data Mining: Applying techniques on the data to discover the interesting patterns like clustering and association analysis.
6. Pattern Evaluation and Knowledge Presentation: This step involves visualization, transformation, removing redundant patterns etc from the patterns we generated.
7. Decisions / Use of Discovered Knowledge: This step helps user to make decision based on knowledge acquired.

In this paper, we make use of Microsoft SQL Server Analysis Services to analyze the data using clustering algorithm and sequence clustering algorithm. K-mean clustering technique is used to find the cluster with nearest mean and realize the scientific clustering of 21 fish market of India based on the average price and arrival date of Catla fish.

## 3. VARIABLES AND DATA

In India, fisheries is an unorganized sector. The prices are determined by the middlemen based on demand (export and local), quality and size of the fish. Price of selected fish is available for the selected markets across India. Data of most of the markets are unavailable due to control of middlemen in the supply chain of fish. n this paper, secondary data is collected from data.gov.in an open government data access portal. Even the data available here is random in nature. We one more drawback we faced was there unavailability of catch on a particular day. Data of 21 markets of 11 states are available in this portal. Due to poor quality of data in 4 markets of 4 states, 17 markets of 5 states are considered for this study. In this paper, data was collected from the year June 2006 - Dec 2013 for the fish variety Catla . Catla fish is largely farmed in Northern India. Market price of 17 regions of 5 states namely Delhi, Orissa, Tripura, Uttar Pradesh and West Bengal are considered. Data available from this portal on Catla fish are State, District, Market, Fish Variety (small/ Big), Arrival Date, Minimum Price, Maximum Price and Modal Price. So these 8 parameters are considered as variables for analyzing the data using cluster algorithm. Data cleaning is done and then stored into the database.

## 4. RESULTS

In this paper, raw data was normalized and stored into the database. Using the source database cube was created. State, District, Market, Arrival Date, MinValue, MaxValue and ModalValue are taken as dimesions. Where MinValue, MaxValue and ModalValue are the minimum, maximum and average price of Catla Fish on a particular arrival date. For Building the cluster following parameter are considered as input and predictable output as shown in Table 1.

**Table - 1**

| Dimension | Mode |
|---|---|
| ArrivalDate | Input |
| State | Input |
| District | Input |
| Market | Input |
| MaxValue | Input |
| MinValue | Input |
| ModalValue | Predict |

A total of 10 clusters where generated depending on k-mean clustering technique to group together similar objects as seen in Figure - 3.

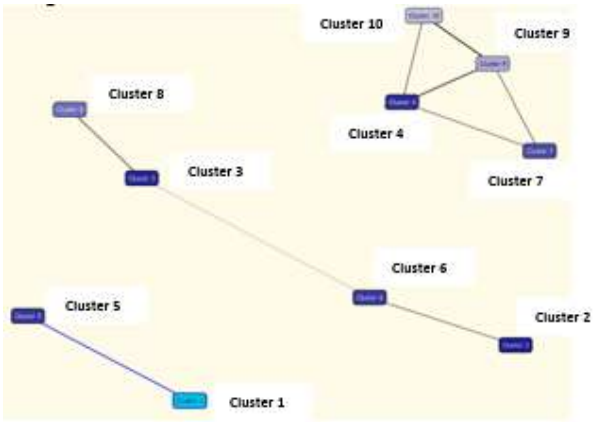


**Figure - 3**

Clusters linked together determines the discrimination between different class of clusters that can be characterized by differing covariance or spectral structures is of importance in applications occurring in the analysis of economic structure of Catla fish in these different regions. Stronger the line linked between the clusters determines the maximum disparity between the groups. But here the data was collected from 2006 to 2013, it discriminates the clusters on data arrival. So to understand the clusters with disparity of price at a given time need to analyze the cluster profile

Figure – 4 shows the cluster profiles based on different parameters used to build this cluster model. It clears illustrates the ratio of influence of each parameter on the given cluster.

This helps to analyze the price variation in each cluster based on arrival date, state, district, market and price. With this cluster profile we can determine the determine the disparity between different cluster groups.

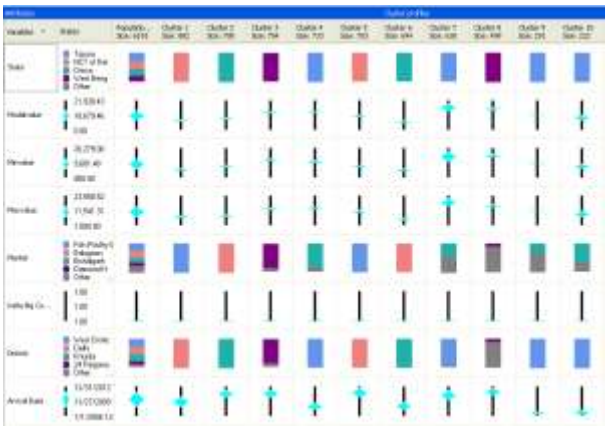

**Figure- 4**

**Cluster 9 and Cluster 10:** the size of both the cluster is around 250. Cluster 9 and cluster 10 represents Taliamura Market, West District of Tripura State and Bishalgarh Market, West District of Tripura State respectively.. In the year 2006, modal value of Catla fish in cluster 9 is Rs. 12000 and Cluster 10 is Rs. 9800. A difference of Rs. 2200 per 100kg of fish. Distance between to two market is 65 kms. We can find the discrimination of scores for cluster 9 and cluster 10 in Figure – 5

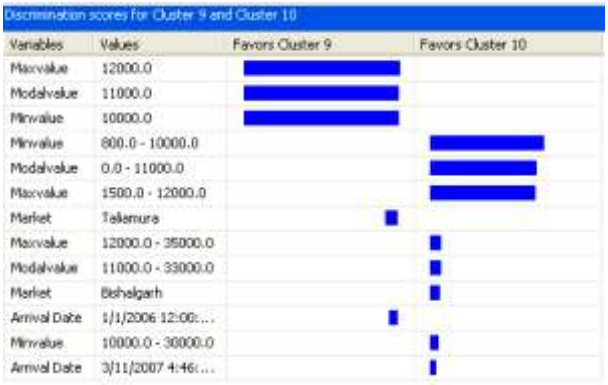


**Figure – 5**

**Cluster 2 and Cluster 7:** the size of Cluster 2 is 785 and Cluster 7 is 630. Cluster 2 and cluster 7 represents Balugaoun Market, Khurda District of Orissa State and Bishalgarh Market, West District of Tripura State respectively.. In the year 2011-12, modal value of Catla fish in cluster 2 is Rs. 8300 and Cluster 7 is Rs. 14500. A difference of Rs. 6700 per 100kg of fish. We can find the discrimination of scores for cluster 9 and cluster 10 in Figure – 6.

Similarly we can see the market opportunities in various market for the fish variety Catla for the benefit of fishing community. Due to non-availability of fish data in public forum and need to analyze the data on available resource has made the analysis bit difficult across all the regions. Along with Clustering algorithm various other analysis and statistical tool need to be investigated to find the opportunities in different regions of India.


**Figure-6**

## 5. CONCLUSION

Microsoft Analysis Services provides a new way of analyzing the data with little knowledge of Clustering analysis. Demand of Catla fish has increased drastically from the year 2009. With demand we can also see the increase in the price of various markets. This papers shows the disparity between the price of Catla fish in the regions of Tripura and Orissa. We find necessary infrastructure has to be developed to facilitate easy transportation of fish from one region to another. For all these, availability of real time data of fish price in various markets in very essential. With this, we can bridge the gap between demand and supply, so that consumer and fishing community both will get the benefit.

Along with Clustering algorithm various other analysis and statistical tool need to be investigated to find the market available of Catla fish in different regions of India.

## 6. REFERENCES

[1] Sara Dolnicar, *"Using cluster analysis for markets segmentation - typical misconceptions, established methodological weaknesses and some recommendations for improvement"*, Australasian Journal of Market Research, 2003, 11(2), 5-12.

[2] Mamta Tiwari, *"Application of Cluster Analysis in Agriculture – A Review Article*", International Journal of Computer Applications, Volume 36– No.4, December 2011

[3] http://technet.microsoft.com/en-us/library/ms345131 (v=sql.90 ).aspx

[4] http://dataminingwarehousing.blogspot.in/2008/10/data-mining-steps-of-data-mining.html