

A Dissertation Report

On

**Localized Pattern Discovery in Time Series Data from Urban Sources**

Submitted in partial fulfilment of the requirements for the award of degree

of

Master of Technology

in

Computer Science and Engineering

Submitted By

**Ayush Rathi**

**(16535006)**

Under the guidance of

**Dr. Durga Toshniwal**

Associate Professor, Dept. of Computer Science and Engineering



Department of Computer Science and Engineering

**INDIAN INSTITUTE OF TECHNOLOGY ROORKEE**

**Roorkee – 247667**

**May, 2018**

# CANDIDATE’S DECLARATION

I hereby declare that the work presented in this dissertation “**Localized Pattern Discovery in Time Series Data from Urban Sources**” towards the fulfilment of the requirements for award of the degree of Master of Technology in Computer Science, submitted to the Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India is an authentic record of my own work carried out during May, 2017 to May, 2018 under the guidance of Dr. Durga Toshniwal, Associate Professor, Department of Computer Science and Engineering, Indian Institute of Technology Roorkee. The content presented in this dissertation has not been submitted by me for the award of any other degree of this or any other institute.

Date

Place: Roorkee

Ayush Rathi

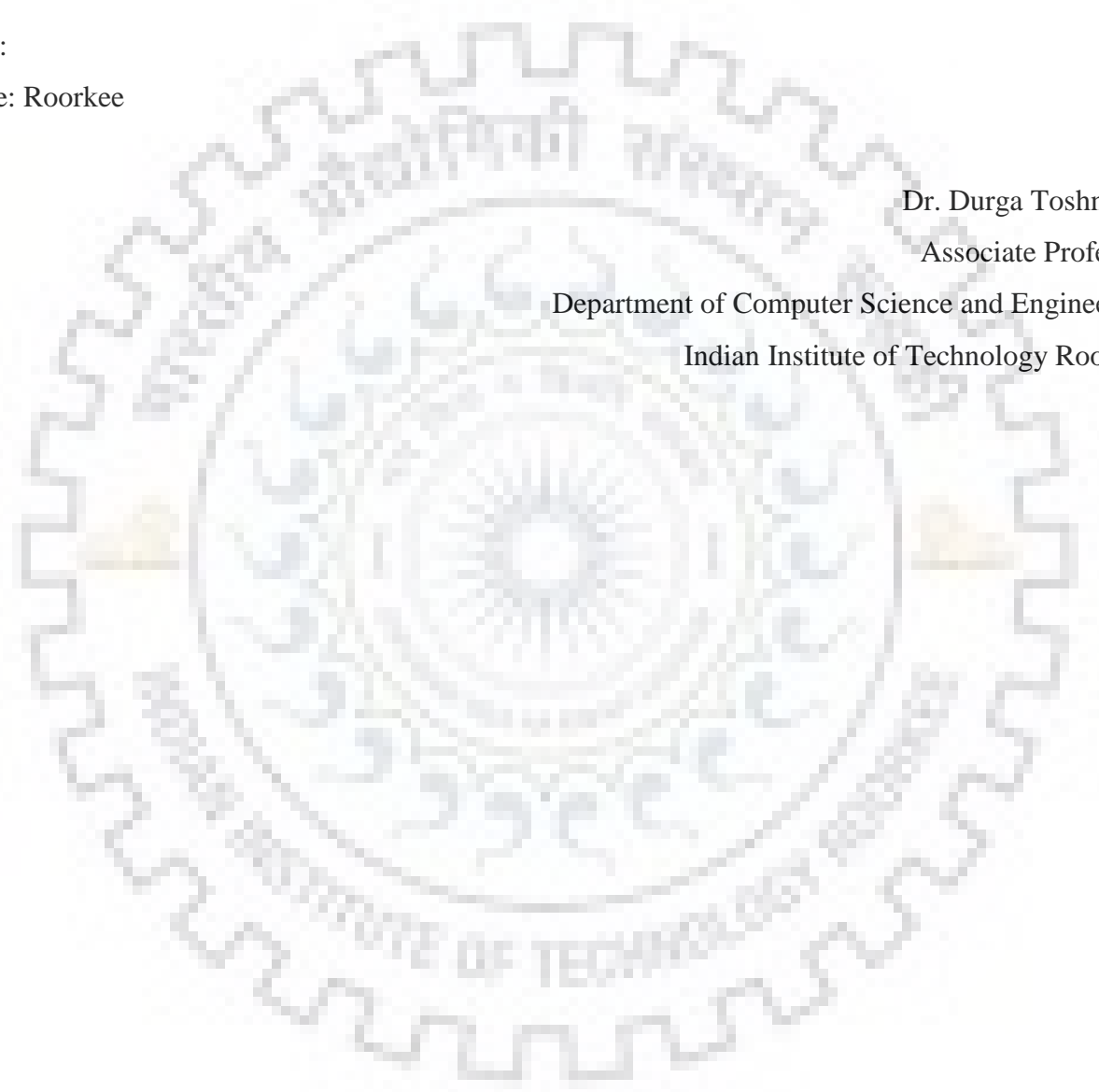


# CERTIFICATE

This is to certify that the statement made by the candidate in declaration is correct to the best of my knowledge and belief.

Date:

Place: Roorkee



Dr. Durga Toshniwal  
Associate Professor  
Department of Computer Science and Engineering  
Indian Institute of Technology Roorkee

# ACKNOWLEDGEMENTS

I would like to thank my guide Dr. Durga Toshniwal, Associate Professor, Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, for her encouragement and supervision during my research work. This work would not have been possible without her guidance.

I am thankful to all my lab mates and staff members of Department of Computer Science and Engineering for their support.

I would also like to thank my family and friends for helping me directly or indirectly and giving support and motivation for my work



## **ABSTRACT**

Time Series data are of crucial importance as they depict trends among various entities over time. Finding similarity in patterns in time series data can help in getting meaningful insights in it. This work presents a technique that finds the similarity amongst time series data by taking distance between angular measure rather than absolute differences. Further, local patterns with relative increase or decrease are also included in computing over all distances. These distances can then be used in clustering or classification algorithms to find which data exhibit similar local patterns. The proposed technique aims in mining similarity between various time series by considering local patterns over global trend.

This technique was mainly developed for mining urban data, such as municipal budgets. City planning is mainly governed by the budget declared by the municipal corporation. Budget data can be utilized for finding how the city is progressing, what are the factors critical for its growth, why is it lagging in some factors as compared to other cities etc. To be able to derive meaningful conclusions, various mathematical and data mining techniques need to be applied on widely available municipal budget data. The budgets of Urban Local Bodies of India follow of uniform structure and the systematic analysis of them can give interesting insights. Although analyses of municipal budget data have been done in past but most of them have static nature and there is need for some generalized framework. This report presents techniques for broad analysis of Municipal Budgets by considering historical budget allocations as time series data.

The technique proposed is also applied on other time series data sets which comprises of both urban and non-urban (generic) data. The experiments gave better results as compared to the existing techniques.

## Table of Contents

|  |           |
|--|-----------|
| Abstract   | v         |
| <b>1. INTRODUCTION</b>   | <b>1</b>  |
| 1.1 Introduction and Motivation.....   | 1         |
| 1.2 Problem Statement .....  | 1         |
| 1.3 Organization of the Report.....  | 1         |
| <b>2. LITERATURE REVIEW</b>  | <b>2</b>  |
| 2.1 Background and Related Work.....   | 2         |
| 2.2 Research Gaps Identified.....  | 4         |
| <b>3. PROPOSED WORK</b>  | <b>6</b>  |
| 3.1 An approach for mining similarities in localized pattern in Time Series Data                   | 6         |
| 3.1.1 Angular feature vector (AFV).....  | 6         |
| 3.1.2 Increase – Decrease Pattern as Feature Vector (IDP).....                                     | 7         |
| 3.1.3 Combining Angular and Increase – Decrease Pattern Feature<br>Vectors.....                    | 7         |
| 3.1.4 Similarity Search in time series with Angular Feature and Increase-<br>Decrease Pattern..... | 7         |
| <b>4. EXPERIMENTS AND DISCUSSION</b>   | <b>8</b>  |
| 4.1 Synthetic Time Series Data .....   | 8         |
| 4.2 Municipal Budget Dataset of Single City .....  | 11        |
| 4.3 Municipal Budget Dataset of Multiple Cities .....  | 17        |
| 4.4 Retail Store Dataset .....   | 21        |
| 4.5 Rainfall Dataset .....   | 24        |
| <b>5. CONCLUSION AND FUTURE SCOPE</b>  | <b>26</b> |
| References .....   | 27        |
| List of Publications.....  | 29        |

## List of Figures

| <b>Fig No</b>  | <b>Page No</b> |
|--|----------------|
| Fig 2.1 Use of Dynamic Time Warping for time series clustering                                   | 3              |
| Fig 4.1 Cluster formed for synthetic data set for weight = 0                                     | 9              |
| Fig 4.2 Cluster formed for synthetic data set for weight = 0.5                                   | 10             |
| Fig 4.3 Trend of Functional Groups   | 11             |
| Fig 4.4 Trend of “Other Functions”   | 12             |
| Fig 4.5 Cluster formed using Hierarchical Clustering based on Euclidean Distances                | 13             |
| Fig 4.6 Clusters formed using Hierarchical Clustering based on Normalized Euclidean Distances    | 14             |
| Fig 4.7 Clusters formed using String Based Clustering based on Edit Distances                    | 15             |
| Fig 4.8 Cluster formed after taking angles as feature vector                                     | 15             |
| Fig 4.9 Cluster formed after taking angles as feature vector                                     | 16             |
| Fig 4.10 One of the cluster formed after combining all account heads of Ahmedabad and Hyderabad. | 17             |
| Fig 4.11 Revenue expenditure estimates of various cities   | 18             |
| Fig 4.12 Sample Cluster formed for Department wise Aggregated Revenue Expenditure                | 19             |
| Fig 4.13 Sample Cluster formed for Department “Engineering - Public Works (Zonal)”               | 20             |
| Fig 4.14 Sample Cluster for Revenue Expenditure Aggregated as per “Minor Head”                   | 20             |
| Fig 4.15 Clustering of Various US Retail Stores on basis of sales.                               | 22-23          |
| Fig 4.16(a) Clusters formed by proposed method for rainfall across districts in India            | 25             |
| Fig 4.16(b) Monsoon arrival map of India   | 25             |

# 1. INTRODUCTION

## *1.1 Introduction and Motivation*

Time series have gain lot of attention in the field of data mining as they carry a lot of information which can be extracted if proper techniques are applied. Particularly mining local patterns in time series of various entities can depict similarity in variation among them.

The urban regions are tremendous source of time series data. Municipal Budgets are one such data source which provide insights about urban region's growth and development. Mining Municipal Budgets data can give interesting results which can help in better growth of urban regions and find out major issues being faced. Apart from that, Urban Regions have various other data sources such as Retail Sales data, analysis of which can provide insight about behaviour of people (like their shopping choices) in that region.

## *1.2 Problem Statement*

Urban regions generate tremendous amount of time series data. The aim of this work is to develop techniques to mine localized patterns in time series data.

The technique is applied primarily to the time series created by historic values of municipal budget data. The municipal budgets reflect the overall focus of the local government. This work presents approach for overall analysis of municipal budget data which can be understood by common person. The technique presented is applied for clustering of budget heads to predict relations between them and further to predict the trends across cities.

Since the technique proposed is generic and can be applied to various time series data, the work is further extended for various other urban and non-urban time series data.

## *1.3 Organization of the Report*

The section 2 of this report contains the literature review. Basics of time series mining are covered first in this section. Then the research gaps in time series mining is presented. Further as municipal budget analysis is done in this work, research gaps in analysis of municipal budget are also explained. Section 3 contains the proposed work in area of time series data. Section 4 contains the experiments and data sets used for analysis of municipal budget data. This includes exploratory analysis as well as application of proposed techniques for clustering the budget heads showing similar patterns. Further, the technique is applied to several other time series data sets including urban and generic data. This is followed by conclusion and future scope in Section 5.



## 2. LITERATURE REVIEW

This section presents the background and related work in the field of Mining Time series. A brief overview of work done in municipal budget analysis is also presented. Note that the special focus is given to municipal budget data since the method for mining similar patterns in time series data was developed as a result of finding similarities in Municipal Budget data set. However, the method proposed is applicable to other time series data as well. The other data sets used for performing experiments are self-explanatory and their description is deferred to section 4 (Experiments and Results) itself.

### 2.1 Background and Related Work

#### 2.1.1 Time Series Data Mining

Time series have gain lot of attention in the field of data mining as they carry a lot of information which can be extracted if proper techniques are applied. Various data mining tasks such as clustering, classification, anomaly detection etc can be applied on time series data to extract meaningful information. Esling and Agon [1] provide an overview of Time Series Data Mining Tasks.

Similarity Search in Time series has been a topic of research from long time and a lot of work has been done for the same. Similarity search in time series can be used for clustering the similar time series into groups or classifying them into existing classes. Clustering and classification essentially involves a measure for computing distance between data points and many methods for the same have been proposed. Clustering involves grouping the similar data such that objects within same group have high similarity. Classification on the other hand assigns classes to data based on the features exhibited by it. The data is assigned to that class, with which its features resembles the most. A detailed description of clustering techniques can be found in Xu and Tian [2]; Han et al. [3]. Study of classification techniques can be found in Han et al. [3].

Aghabozorgi et al. [4] have presented the comparison of various time series-based clustering techniques and explored all four components of time series clustering viz. Clustering algorithms, Prototypes (e.g. averaging etc.), Similarity measures (e.g. Euclidean, DTW, LCSS, etc.) and Dimensionality reduction (e.g. PAA, SAX, etc.).

Berndt and Clifford [5] have proposed the use of dynamic time warping to find patterns in time series. This technique wraps the time axis of the signals (or time series) to achieve better alignment between them. That is, the sequences are aligned to get best possible match.

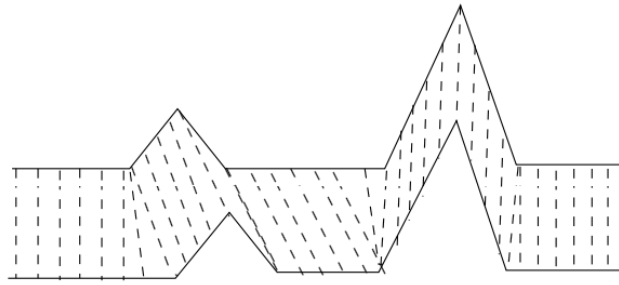


Fig 2.1<sup>1</sup> Use of Dynamic Time Warping for time series clustering

With DTW, for 2 sequences S1 and S2 having m and n points respectively, cumulative distance between ith and jth point,  $d(i,j)$  is given by

$$d(i,j)=\text{dist}(i,j)+\min[d(i-1,j),d(i,j-1),d(i-1,j-1)]$$

DTW is an efficient technique for time series clustering, but shape-based profiling can't be obtained when variation in amplitude of sequences is not regular.

Lin and Li[6] have proposed a bag-of-word technique (analogous to clustering of documents) to cluster time series. This approach focuses on global pattern instead of local shapes.

Paparrizos and Gravano [7] have proposed k-shape algorithm, for clustering time series based on shapes.

For mining similar patterns, approaches based on derivatives or slopes have also been proposed. One such approach is derivative dynamic time warping, Keogh and Pazzani [8]. This approach handles the singularity issues with classic DTW by taking slope as distance measure instead of actual Euclidean distances. Toshniwal and Joshi [9] proposed another approach for similarity search by taking cumulative slopes.

Sometimes, a symbolic representation of time series can have various advantages. For e.g., representing time series in form of symbols may provide computational advantage of String based algorithms on Time series data, Lin et al. [10]. Further, methods such as Piece wise aggregate approximation have been proposed to reduce the dimensionality of time series data and enhancing the computational speed, Keogh and Pazzani[11].

<sup>1</sup> Image Source: [https://upload.wikimedia.org/wikipedia/commons/a/ab/Dynamic\\_time\\_warping.png](https://upload.wikimedia.org/wikipedia/commons/a/ab/Dynamic_time_warping.png)

### **2.1.2 Analysis of Municipal Budgets**

City planning is mainly governed by the budget declared by the municipal corporation. Each year the budget is published by all municipal corporations of urban cities. Budget data can be utilized for finding how the city is progressing, what are the factors critical for its growth, why is it lagging in some factors as compared to other cities etc. Municipal budgets have not received much attention as state or union level budgets. Every year voluminous research is directed towards the analysis of India's national budget and to a certain extent towards the budgets of different states. The budgets at the local level, however, do not receive the kind of attention they should despite the growing importance of urban areas and the increasing focus on decentralisation and devolution of financial powers, Sekhar and Bidarkar [12]. This section presents few works done in the field of analysis of municipal budget data analysis in India as well as other countries.

Sekhar and Bidarkar [12], have presented analysis of budgets of 5 municipal corporations of India, for the period of 6 years. Key findings on revenues and expenditures have been presented.

Parkhimovich and Vlasov, [13] presented the need of open budget data at municipal level in Russia and how can this data be made open in digital form.

Benito et al. [14], provided analysis of Spanish municipal budget data taking into consideration deviations of forecasts. Mayper et al., [15] provides an analysis of municipal budget variances to find if they are systematically biased and to find factors affecting magnitude and direction of this bias.

[16] Presents the status of Municipal Finances in selected municipal corporations of Punjab in India. [17] Presents analysis of budgets of 15 municipalities of Maharashtra. [18] Gives an overview of trends in municipal finances across 35 municipal corporations in India for the period 1999-2000 till 2003-2004.

### **2.2 Research Gaps Identified**

A lot of work has been done in mining time series data. A lot of method exists for finding trends across time series, detecting anomalies and finding similarities. However, time series may also exhibit local patterns and methods which are capable of finding global trends may not work in finding these local patterns.

General methods such clustering on the basis Euclidean distances over the actual values are not robust from factors such as amplitude scaling or magnitude differences. Metrics such as Euclidean Distances, work solely on magnitude and hence cannot find relative similarity based on pattern, discarding amplitude differences. On the other hand, with normalized Euclidean distance, the impact of amplitude differences is totally discarded which is also not desirable. This is also the case if

symbolic approach is used where time series is represented in form of word (with alphabet of word representing pattern) and string distance-based metric such as Edit distance is used.

The technique presented in this work effectively finds similarities in local pattern. It uses angular measure (similar to that of slope-based measure as in Keogh and Pazzani [8] and Toshniwal and Joshi [9]). Apart from angular measure, a symbolic approach is also used to encode the increase decrease patterns. The mathematical formulation of technique is presented in detail in section 3.

As far as municipal budgets are concerned, while there is lot of work done in analysis of municipal budget data, but still it is very less as compared to Union or State Budget analyses. Further most of the ULB budget analyses have given the conclusions for the specific dataset only, and there is a need of some generalized framework through which key points of municipal budget can be extracted.

With advancement in data mining and analysis techniques, there lies a tremendous scope in analysis municipal budgets and results of the same can be used in profiling cities, finding trends, predicting the focus of newly elected local Government in a city based on its work in some other city, finding anomalies and much more.

Budget data of several years can be considered as a time series data. Clustering of time series can help in grouping various budget heads and find similarities in them. Multiple budget may have magnitude differences but mining local patterns can help in finding similarities. Similarly, budget of different cities may vary drastically because of differences in economy, population etc. But similarity in pattern can be used to get insights like areas where Government is focusing, what are priorities of different Governments etc.

The technique proposed in this work is applied over municipal budget data for Knowledge Discovery. Also, the technique is applied to various other data sets as explained in section 4.

### 3. PROPOSED WORK

The overall aim of this work is to present approach to mine similarities in time series data based on localized patterns. This approach is needed to cluster time series data based on similarity in shape. The proposed approach was applied to various time series data sets. First among them was Municipal Budget Data set. For municipal budget data analysis, firstly individual city is considered and exploratory analysis for the same is done. Then proposed approach is applied for clustering of data into logical groups so as to find out relations between various departments (and expenses made by those departments) for a city. The work is extended to country level and above analyses is done for major cities of the country. Again, the proposed technique for clustering is applied to get relations between estimates made by respective municipal corporations. These analyses may be useful for various purposes, such as giving recommendation for improvement of cities, finding priorities of local Government(s), Predicting focus of newly elected Government by using results of past analysis for the same etc.

Further, the proposed approach is applied to other time series data as well in which mining similarities in pattern can give meaningful insights.

The proposed approach is discussed next. The Experiments are covered in section 4.

#### ***3.1 An approach for mining similarities in localized pattern in Time Series Data***

This section describes approach for mining similar patterns in time series. This approach helps in formulating distances between time series considering similarity in local patterns, which can be then be used to cluster similar time series or classify to the most similar class.

The proposed method uses angles and well as increase-decrease patterns as features to find similarity in time series variations.

##### **3.1.1 Angular feature vector (AFV)**

Angular feature can be computed measuring clockwise angle between consecutive line joining the time series values, considering time series is plotted on a graph and. Once, time series points are converted into angular features, distances between AVF of two time series,  $\Delta_{Angular}$  can be found in similar way as Euclidean distances are computed, but by taking angles as feature instead of actual value.

### 3.1.2 Increase – Decrease Pattern as Feature Vector (IDP)

The increase decrease patterns can simply be captured by representing time series through a word, where alphabet of word encodes the delta change (i.e. Increase, Decrease or Constant). Also, impact of differences in increase decrease pattern is more when increase or decrease is relatively high. Hence, to find the difference corresponding to IDP feature in two time series, a linearity factor should be also considered along with distance in corresponding letters of words. Formally, for two time sequences, if the string-based difference for the  $k^{\text{th}}$  pair of consecutive letter is  $\partial_k$ , and corresponding linearity factor is  $LF_k$ , then the effective difference on basis of IDP between them is given by  $\partial_k * LF_k$ . The overall difference on basis of IDP,  $\Delta_{IDP}$  can be calculated by adding differences for all such pairs of consecutive letters.

### 3.1.3 Combining Angular and Increase – Decrease Pattern Feature Vectors

The similarity search based on solely increase decrease pattern will have issues of symbolic approaches explained earlier, where amplitude factor gets fully discarded. There may be scenarios where pattern is not same, but angles are same and hence these cannot be covered by angular features. However, if both the features are combined, all such scenarios can be covered. Hence, distance between the two-time series can be calculated by taking effectively the contributions of both equations (1) and (2). To add these two distances, some weights should be multiplied so that linear combination of both the factors is obtained. So, the effective distance therefore becomes:

$$\Delta_{effective} = \Delta_{Angular} + weight * \Delta_{IDP} \quad (2)$$

Where, all the symbols have their usual meanings as described earlier. The value of *weight* depends on dataset features like domain and characteristics of data, user requirements like clustering on basis of certain factors or classification with high accuracy etc.

### 3.1.4 Similarity Search in time series with Angular Feature and Increase-Decrease Pattern

Suppose there are ‘M’ time series in data set. The distance between each  $\binom{M}{2}$  pair of time series can be founded by applying method described in 3.1.3. Finally, a distance matrix can be created and tasks such as clustering and classification can be done on that to cluster or classify time series following similar pattern.



## 4. EXPERIMENTS AND DISCUSSION

This section covers the discussion of results obtained after method described in 3.1 was applied to variety of data sets.

### 4.1 Synthetic Time Series Data

The synthetic data set was designed to explain the working of approach. It consisted of 10 sequences (considered as time series) with each sequence comprising of 7 data points (hence 4 angles would be formed for each sequence). The value of data points was taken randomly in range [0,100]. The distance between each pair of them was calculated using the discussed approach. Finally, Hierarchical agglomerative clustering was applied with complete linkage to cluster them into 4 groups.

To plot the above sequences as time series on x-y graph with as 10:20 as aspect ratio of y:x, the value of  $\Delta x$  (that is the distance of two consecutive time instances on x axis) is taken to be 33.33. As an example, one of the time series in data set had values [26,16,71,96,65,44]. For above setting, the angular representation of this time series would be [104.51915853035277, 201.91169958980961, 259.7927228806812, 169.28810248883659].

Similarly, each time series was converted into string form with each letter depicting either of increase, decrease or constant.

These word and angular presentations were then combined by method explained in section 3.1 to find the similarities in time series and finally they were clustered into 4 groups. The experiments were done both with zero value of *weight* and with non-zero value of *weight*.

#### Results with zero value of ‘weight’

Fig 4.1 shows the clusters formed when ‘*weight*’ value is zero (i.e. only angular features are considered). The legend in the plot depicts the id number of time series. X axis shows the data instance (or time instance) and Y axis represents corresponding values. As can be seen from the graphs, time series showing similar patterns were grouped in same cluster.

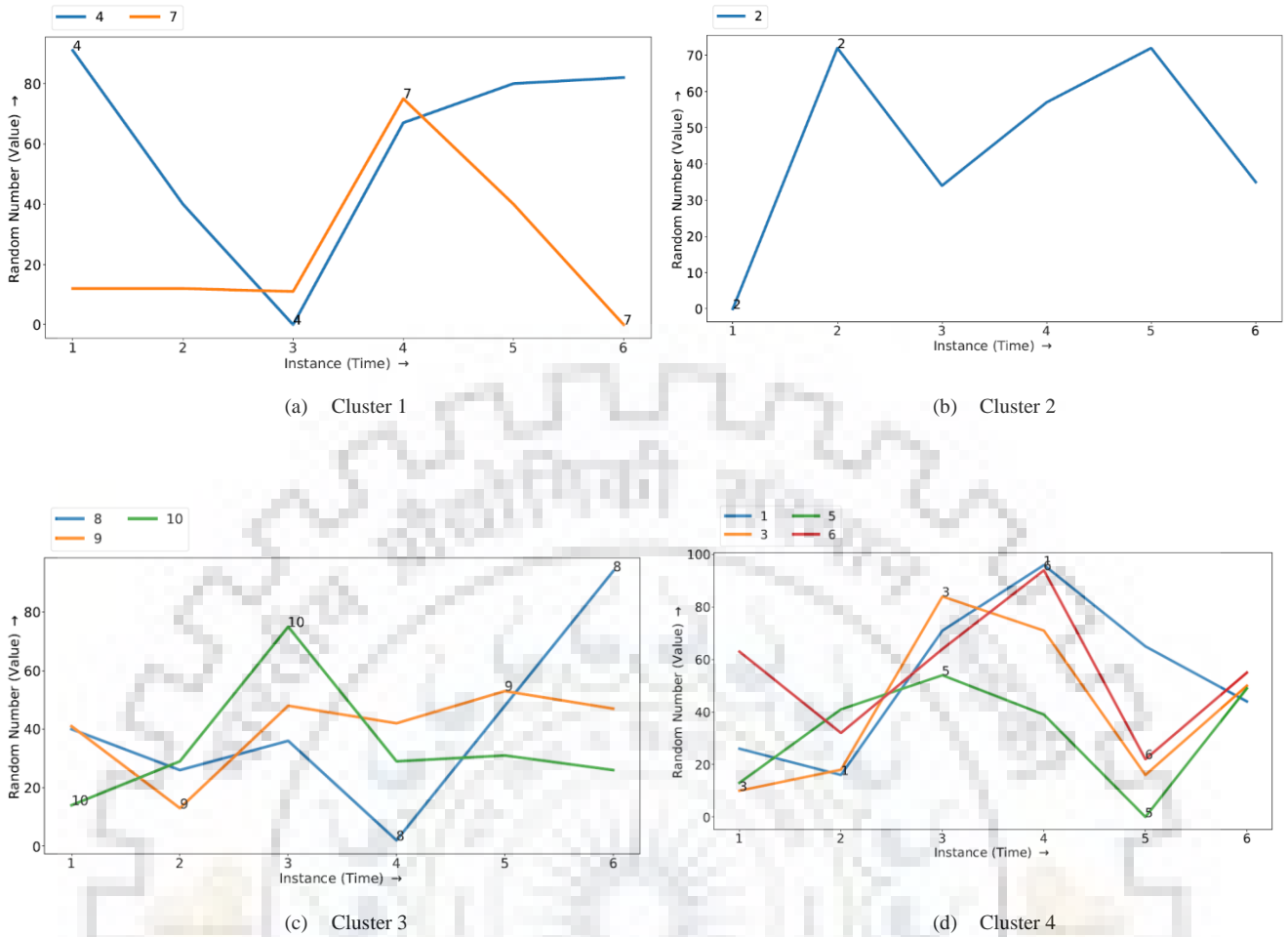
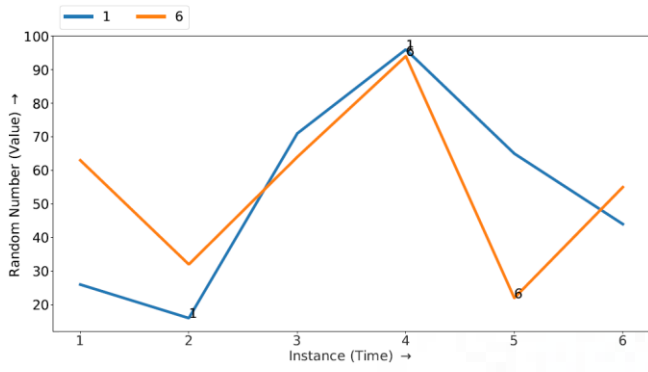


Fig 4.1 Cluster formed for synthetic data set for weight = 0

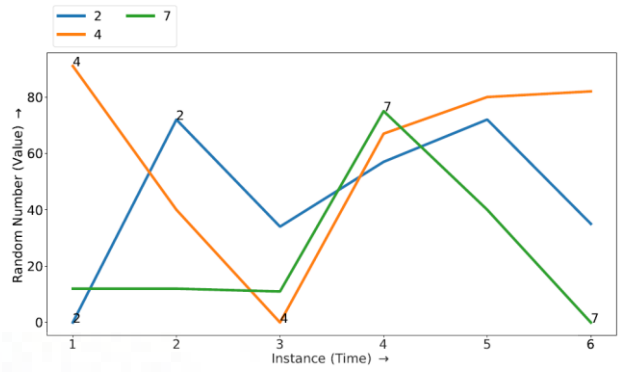
### Results with non-zero value of ‘weight’

In Fig 4.1, in fourth cluster, time series ‘3’ and ‘5’ have opposite pattern then ‘1’ and ‘6’ for first three data points. However, the cumulative sum of their angular differences was such that they came in same cluster. Non-zero value of ‘weight’ tackled this difference. After observing the angular and string-based distances ‘weight’ was set to 0.5, as this value made both magnitude of differences for both these features roughly of the same order. As can be seen in Fig 4.2, ‘3’ and ‘5’ formed a new cluster. But ‘2’, ‘4’ and ‘7’ were also moved to same cluster to accommodate a new cluster. Changes in ‘weight’ value can give better results depending on data characteristics. Further, optimal number of clusters can be found by applying various techniques measuring goodness of clusters.

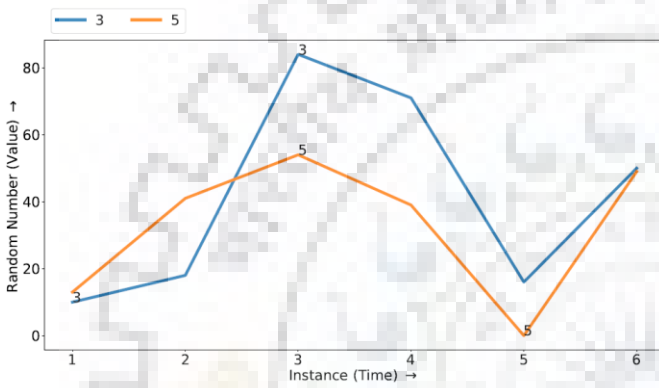




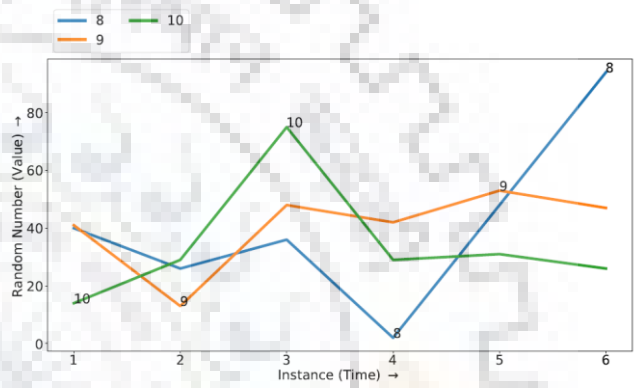
(a) Cluster 1



(b) Cluster 2



(c) Cluster 3



(d) Cluster 4

Fig 4.2 Cluster formed for synthetic data set for weight = 0.5

## 4.2 Municipal Budget Dataset of Single City

This section covers exploratory analysis of municipal budget of particular a ULB. Further correlations between departments of that ULB is found using clustering techniques. Traditional techniques based on measures such as Euclidean distances, edit distances have not given significant results, as compared to the method proposed in section 3.1.

### 4.2.1 Dataset Used

The dataset used for exploratory analysis is budget data of Indore Municipal Corporation. This data simply contains estimates of functionalities over various account heads for a period of 9 years. The data contains 10 function groups divided into 27 function descriptions which are further divided into 59 departments.

1493 unique department - budget head combinations form the total records with each record having estimates for 9 consecutive years from 2009-10 till 2017-18

### 4.2.2 Exploratory Data Analysis

The aggregation of data into logical groups gave some insights about city, few of which are summarized below:

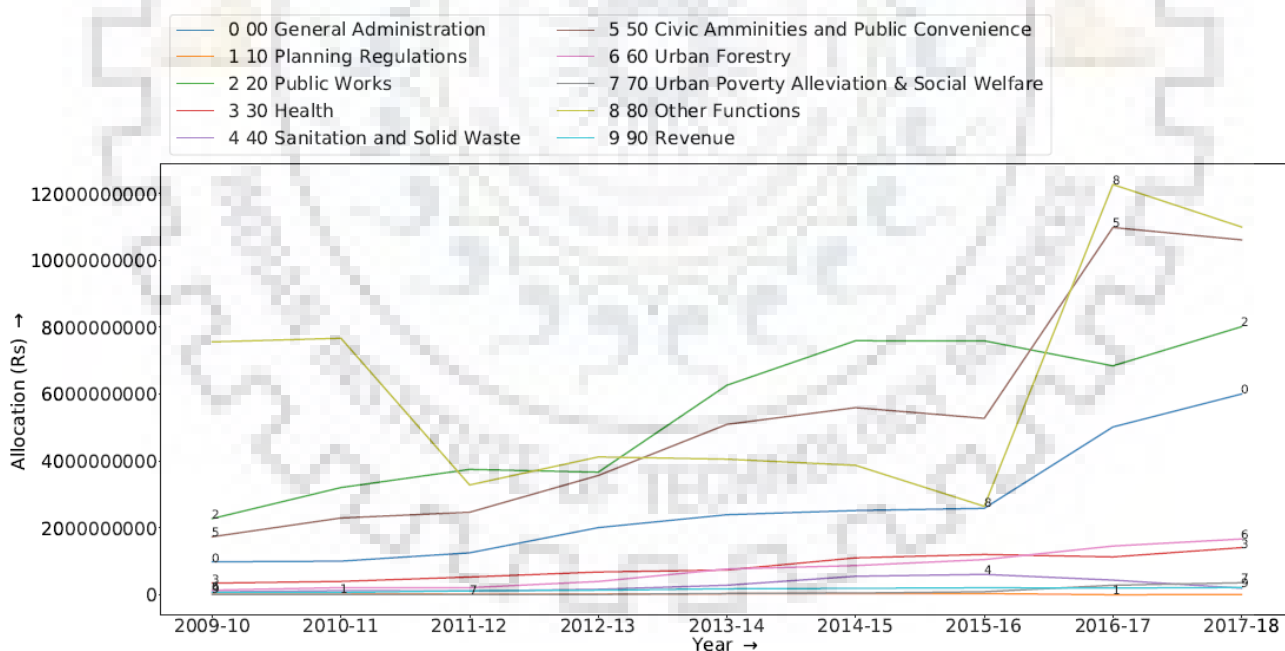


Fig 4.3 Trend of Functional Groups

As can be seen, Government increased the focus on “other functions” in 2016-17 after 2010-11. Civic Amenities and public Convince was also of prime focus in 2016 -2017. The focus on public works was less in 2016-17

Other groups have followed a gradual increasing pattern

Further, “other function” followed below trend

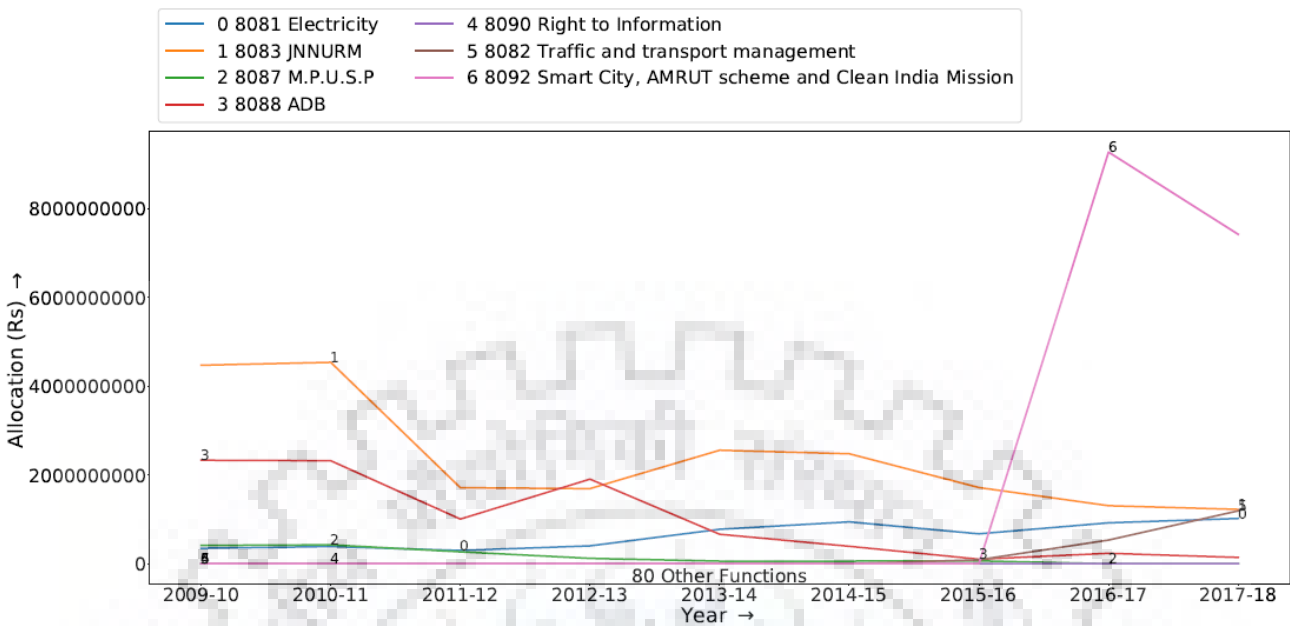


Fig 4.4 Trend of “Other Functions”

Other function was influenced by JNNURM initially in 2010-11. In 2016-17, increase in other work was because of Smart city, AMRUT scheme and Clean India Mission which also depicts government’s current focus.

Similarly, trends for each and every department can be analysed.

### 4.2.3 Clustering Techniques Applied

The motive of clustering was to group together departments showing similar trend. If budget allocation for a particular head in given department is increasing/decreasing over the years in a manner similar to budget allocation in some other head (within department or with some other department), then this can be grouped together for analysing the relation (if any) between them.

Various clustering methods have been tried, but no method give better quality clusters then the one which was specifically modelled for this purpose. All the methods and their performance are covered below, along with the modified one which is detailed after all these methods.

- **Hierarchal Clustering Based on Euclidean Distances**

This is the most general method used for clustering. The data is considered in form of matrix where each row represents budget allocation for a department and the columns are the values in lacks for 9 years. Euclidean distances for each pair are taken and the ones having relatively less distance are grouped together.

However, as explained in section 2.2 issue with this technique is that if 2 allocations are following similar pattern of increase/decrease but one is having numerically less values then other, than their distance will be high and hence will not be grouped in same cluster.

Fig 4.5 illustrates a sample cluster formed by this technique. The Fig shows plots of various budget heads groped under a cluster. As can be seen, this technique requires similarity in quantitative terms, while for getting trend only relative increase/decrease is required.

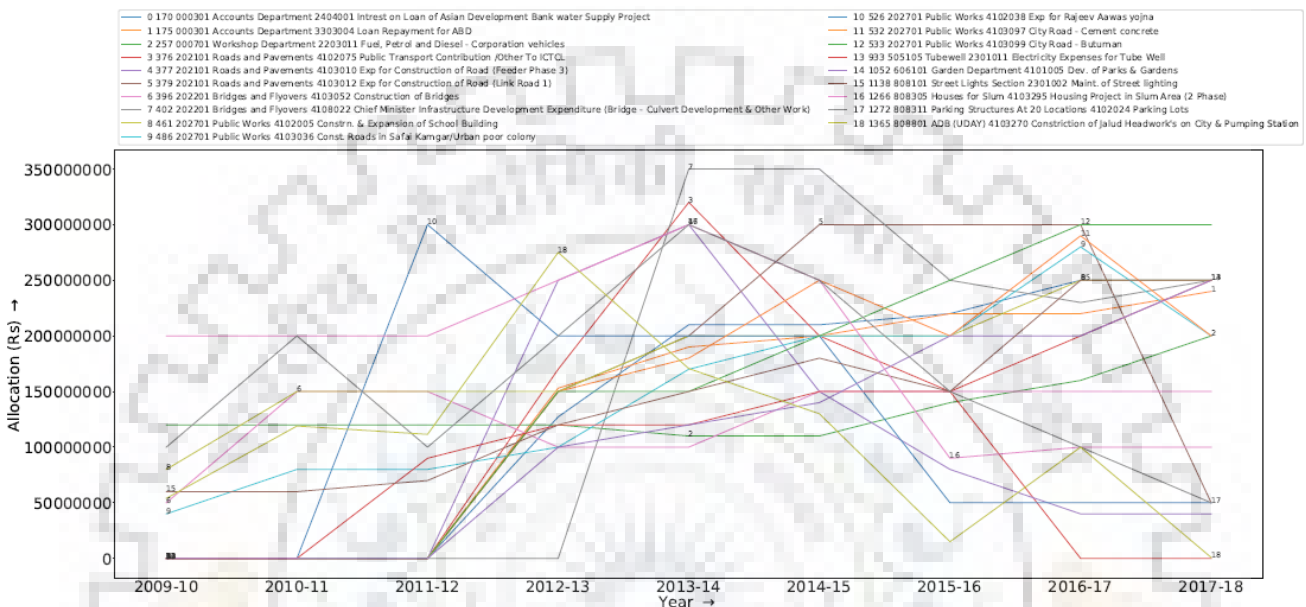


Fig 4.5 Cluster formed using Hierarchical Clustering based on Euclidean Distances

- **Hierarchal Clustering Based on Normalized Euclidean Distances**

The clusters created by previous technique depend heavily on the actual allocation values rather than relative increase/decrease. To overcome this, the budget allocation for given head were normalized in way that every allocation comes in the range (0,1). This significantly improved quality of clusters and the heads in which similar trend occurred were clubbed in the same cluster. But this technique also has an issue as explained in section 2.2. Consider one head in which there is a short increase in allocation for some year, while in other there is short decrease. Both will have chances of getting clubbed together if their values are relatively close, despite that they have shown opposite trends.

Fig 4.6 shows two sample clusters obtained by this technique. The first one is the good quality cluster obtained by this technique. The second one shows the issues stated in section 2.2 for this technique.

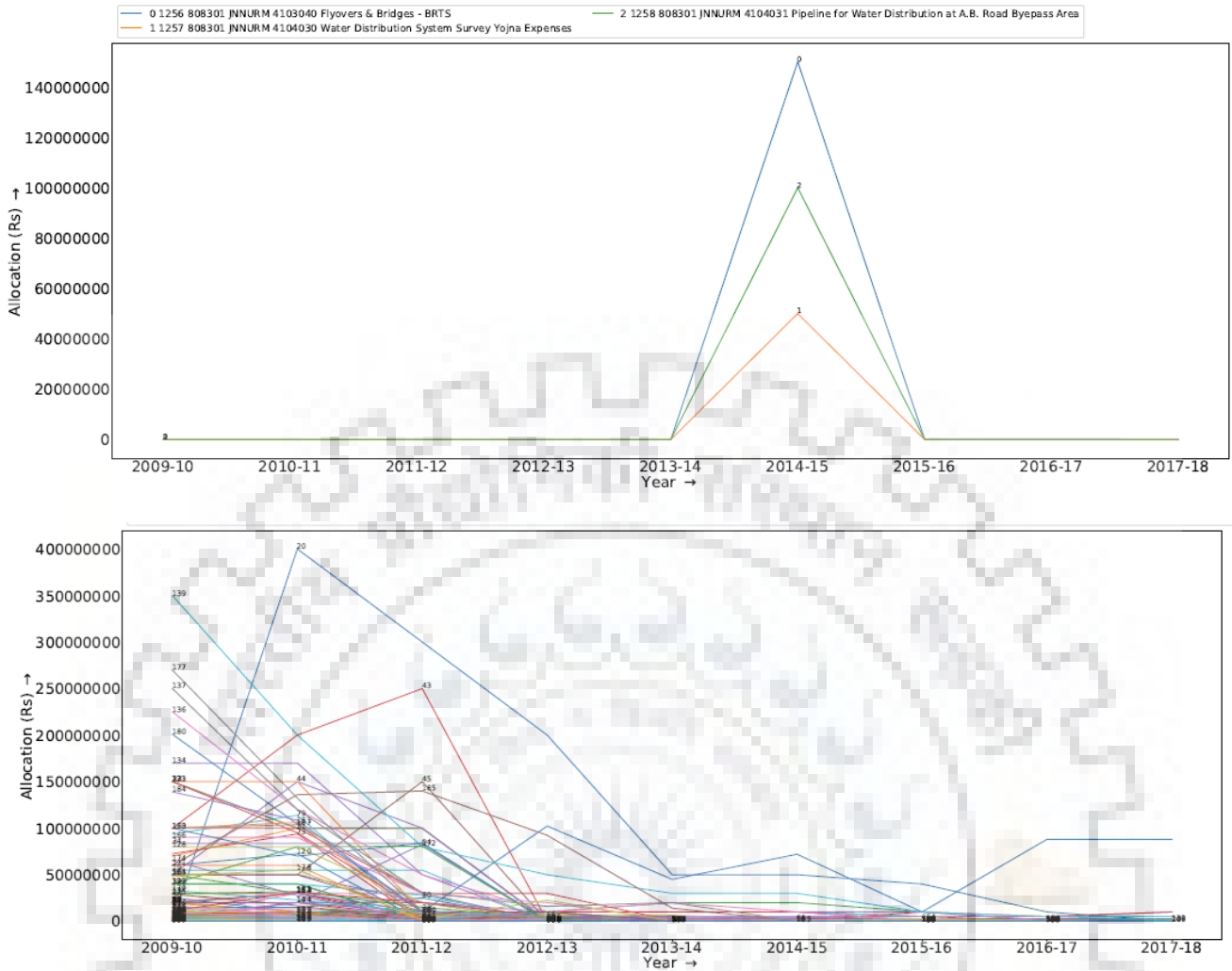


Fig 4.6 Clusters formed using Hierarchical Clustering based on Normalized Euclidean Distances

- **String based clustering**

To group together the heads showing similar patterns over the years in terms of increase/decrease in allocation, each head was represented by a feature vector of string of length (n-1) where n is number of years taken into account. Each character of depicts whether the allocation has increased/decreased or remained constant as compared to previous year.

For about 1500 records, 400 unique strings of such kinds were obtained leading to order of 400 clusters if exact matching is taken into account. So, edit distances between each pair of strings were taken and then normal hierarchical clustering was applied. But this lead to poor quality clusters as well since even edit distance of 1 can correspond to opposite trend (e.g one is increasing and other is decreasing). Fig 4.7 shows an example of cluster formed using string matching technique. Instead of edit distance of 1, hierarchical clustering was done, and 20 clusters were made

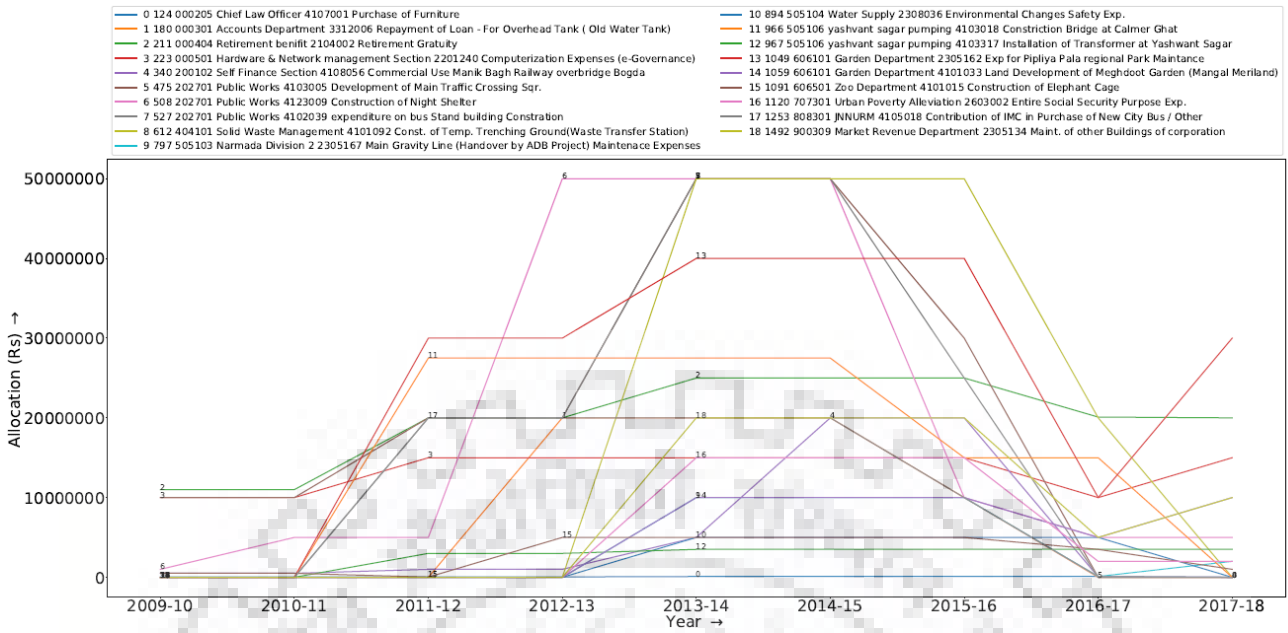


Fig 4.7 Clusters formed using String Based Clustering based on Edit Distances

- **Hierarchal Clustering Based on angular distances**

As discussed earlier, instead of taking the normalized budget allocation as feature vector, angles between the curves showing budget trend is taken.

The angles for similar trend will be closer than to the ones following some other trend. Hence Euclidean distance between angles for all pairs of heads is taken and hierarchical clustering is applied on the same. The results were better as compared to the case where feature vector was normalized value of allocation amount.

Figure 4.8 shows sample of cluster obtained by this technique.

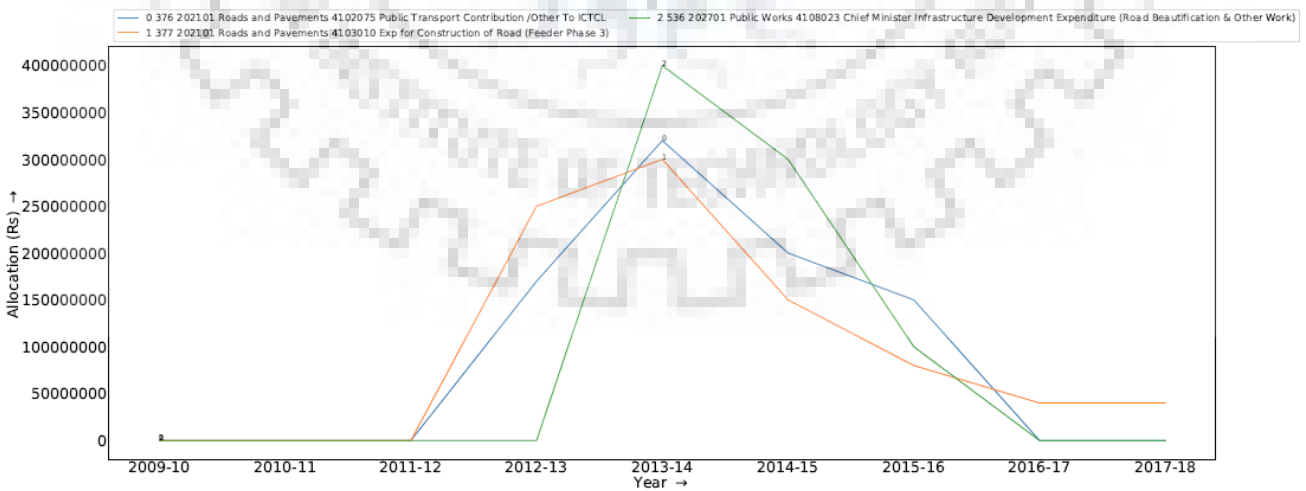


Fig 4.8 Cluster formed after taking angles as feature vector



Fig 4.9 shows another cluster formed with this technique.

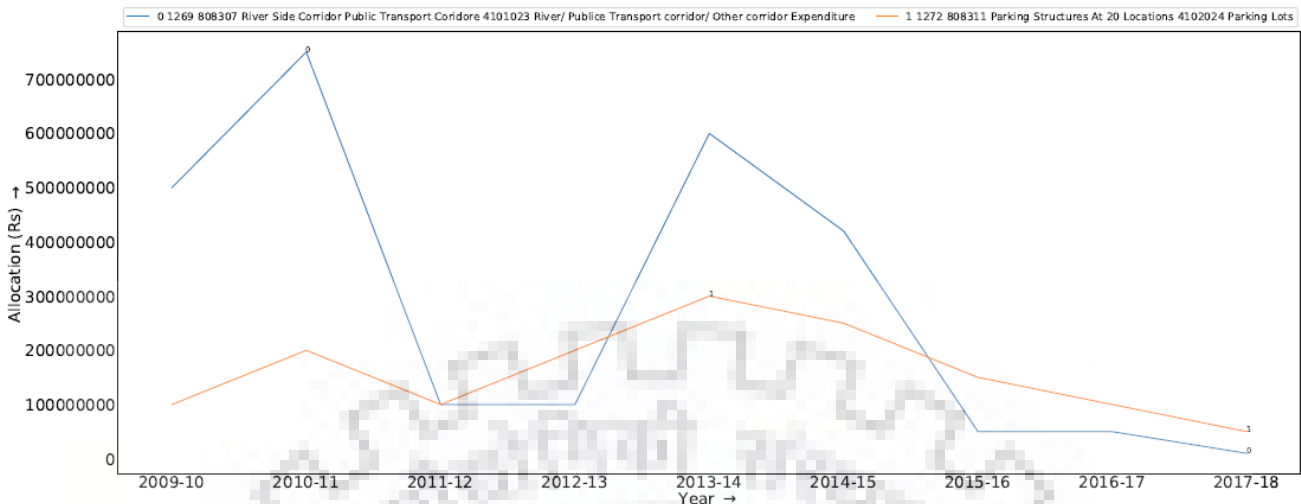


Fig 4.9 Cluster formed after taking angles as feature vector

### Deciding Number of Clusters-

While technique of clustering is an important aspect, equally important is deciding on number of clusters. Various metrics which measure goodness of cluster can be used to decide the optimal number of clusters.

The effectiveness of cluster can also be measured by determining the overall shape of cluster and percentage of samples in cluster having deviation from that shape.

Further, unlike Hierarchical clustering, there are techniques such as affinity propagation [19] which does not require number of clusters and clusters automatically in most optimal way. By viewing each data point as a node in a network, Brendan et al [19] devised a method that recursively transmits real-valued messages along edges of the network until a good set of exemplars and corresponding clusters emerges. The algorithm proceeds by alternating two message passing steps, to update two matrices - The “responsibility” matrix  $R$  has values  $r(i, k)$  that quantify how well-suited  $x_k$  is to serve as the exemplar for  $x_i$ , relative to other candidate exemplars for  $x_i$ .

The “availability” matrix  $A$  contains values  $a(i, k)$  that represent how “appropriate” it would be for  $x_i$  to pick  $x_k$  as its exemplar, taking into account other points' preference for  $x_k$  as an exemplar. [19]

The Euclidean distances based on angular measure can be used for distance measure in affinity propagation as well.

### 4.3 Municipal Budget Dataset of Multiple Cities

Analysis was extended to multiple cities as well at broad level, and this section presents brief overview of same.

#### 4.3.1 Dataset Used

The budget data of major cities like Mumbai, Chennai, Bengaluru, Ahmedabad, Pune, Gandhinagar, Gurgaon etc was taken. The available data was of only 4 years, so time series length was small. .

One of the approach to find similarity in functioning of cities is to merge all their account heads, and cluster them into logical groups based on similar patterns.

For e.g. account heads of Ahmedabad and Hyderabad were merged and one of the cluster formed is shown in Fig 4.10

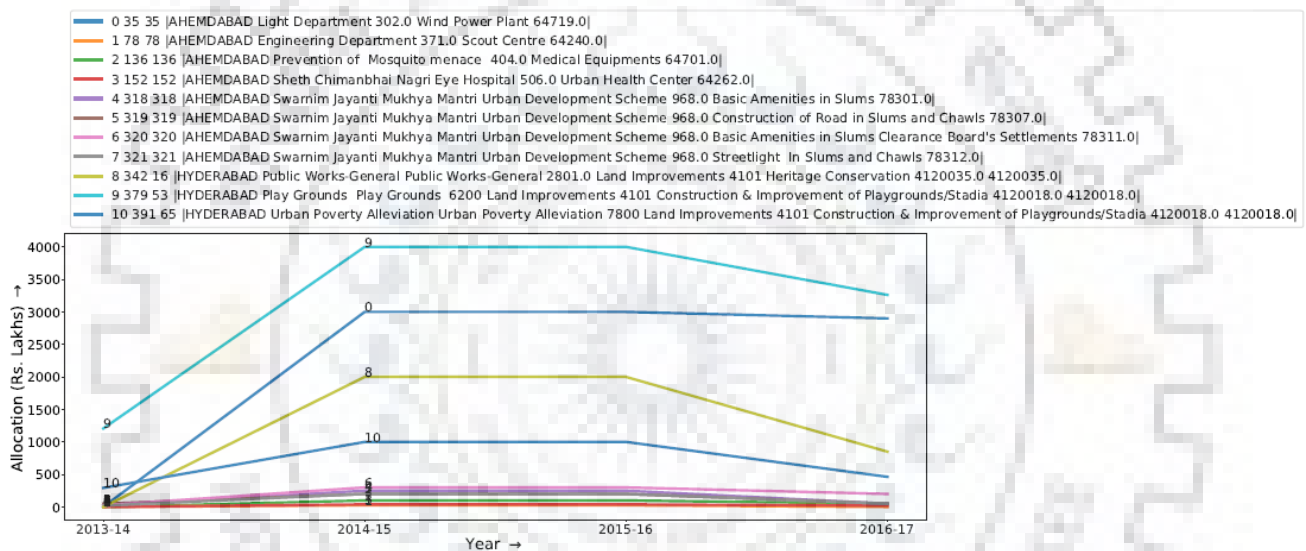


Fig 4.10 One of the cluster formed after combining all account heads of Ahmedabad and Hyderabad.

However, because the granularity at account head level is too fine, the results may not have any conclusion, as there is high chance that few of unrelated budget heads amongst large number have followed similar pattern.

So, analysis for multiple cities can be done by taking aggregate data. For e.g. department level data may be taken. One approach is taking aggregate data as per its source. Budgets are broadly divided into four categories – Revenue Receipt, Revenue Expenditure, Capital Receipt, Capital Expenditure. Revenue Expenditure estimates of multiple cities were taken as example to find pattern amongst cities. Results are shown in Fig 4.11

As can be seen, the revenue expenditure of Bhopal and Bengaluru followed a different pattern. Again, city level analysis can be done for finding what factors were there for different behaviour. Once patterns are found, the allocations due which that pattern occurred can be found through



exploratory analysis. This approach is useful for general public as well as analysts to mine anomalies, patterns in data.

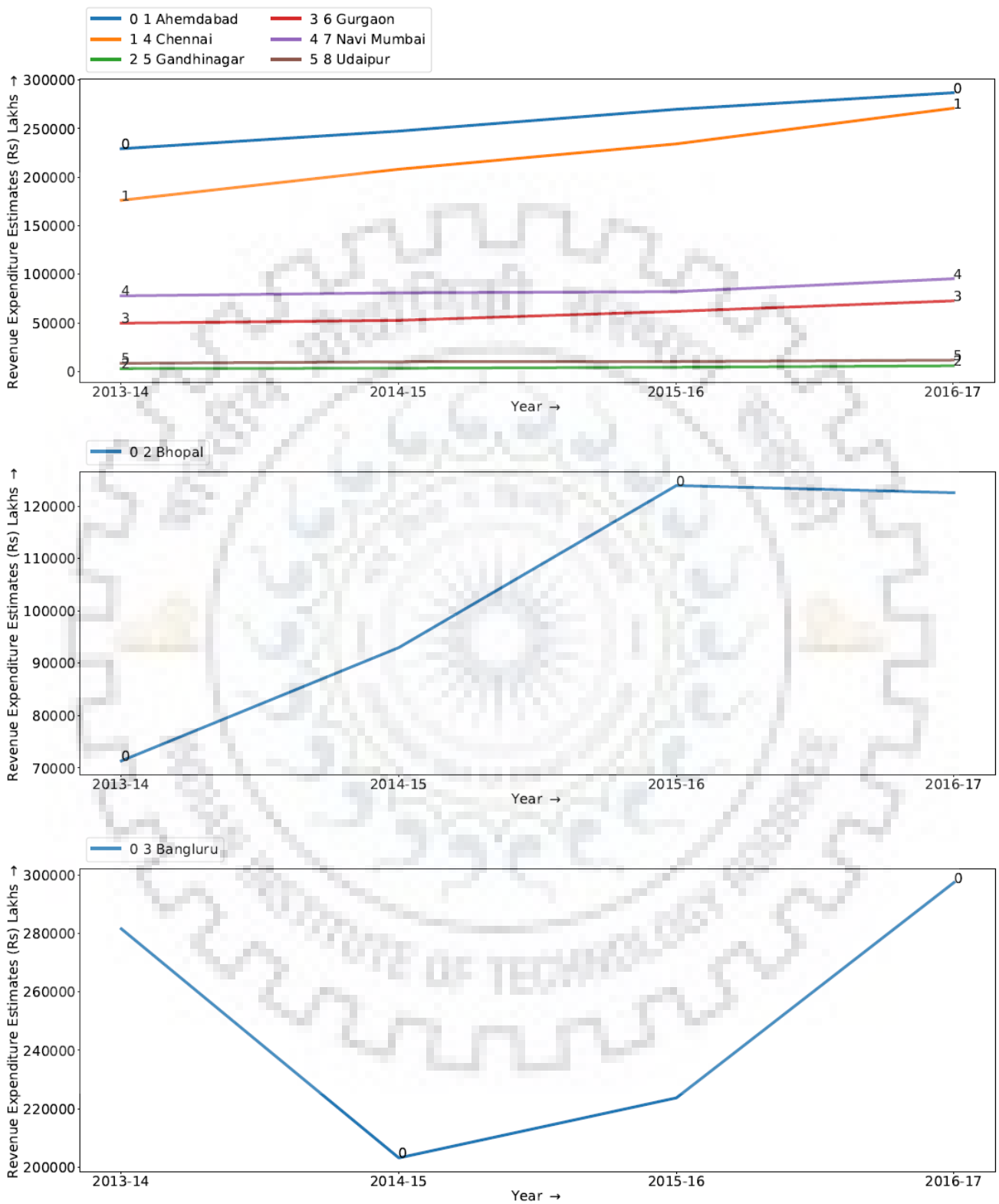


Fig 4.11 Revenue expenditure of various cities

Again, same technique can be applied (in this example) for Revenue Expenditure Estimates of Bengaluru or exploratory analysis can be done if data is less. Even though the data is of 4 years and exploratory analysis is sufficient, the technique was applied to Revenue Expenditure Estimates of Bengaluru.

Firstly, Department wise aggregation of Revenue Expenditure Estimates was done and then departments were clustered. One of the clusters is shown in Fig 4.12

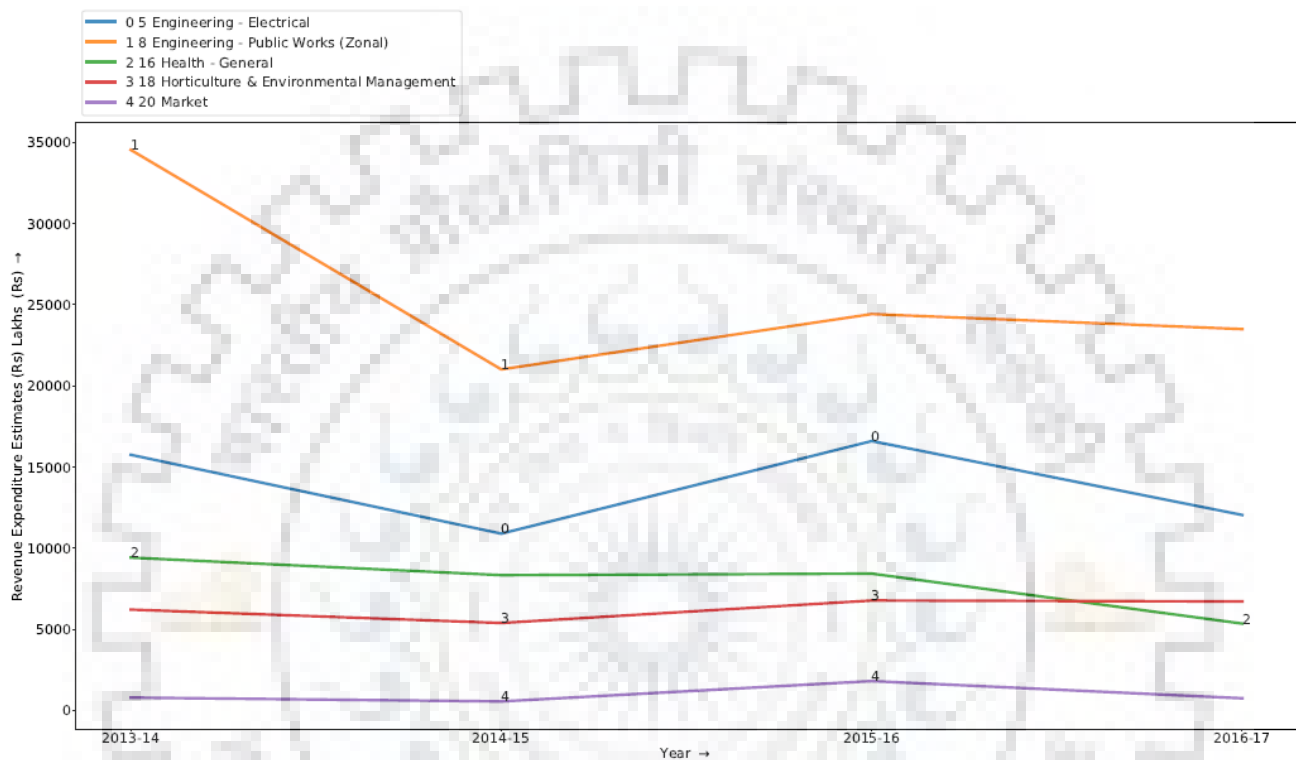


Fig 4.12 Sample Cluster formed for Department wise Aggregated Revenue Expenditure

As can be seen in Fig 4.12, amongst others, “Department Engineering - Public Works (Zonal)” showed a decrease in 2014-15. (Note that this is just a sample cluster. Many departments in other clusters also have shown a decrease in trend for 2014-15. Exact analysis can only be done by subject matter experts and after considering ground truth.)

Further Department “Engineering - Public Works (Zonal)” was again clustered using the same approach. Fig 4.13 shows one of the sample clusters.

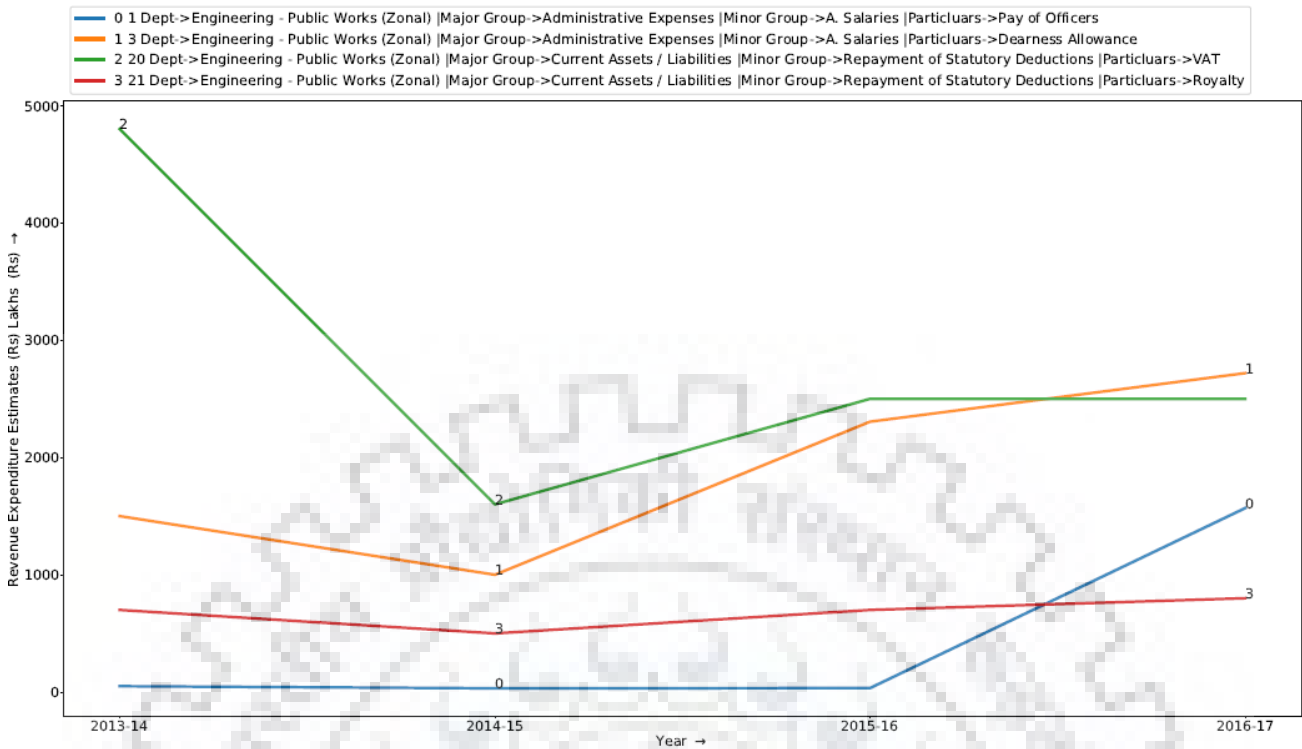


Fig 4.13 Sample Cluster formed for Department “Engineering - Public Works (Zonal)”

Similarly, clustering was also done after Revenue Expenditure Estimates were aggregated on the basis of “Minor Head”. Fig 4.14 shows a sample cluster for the same.

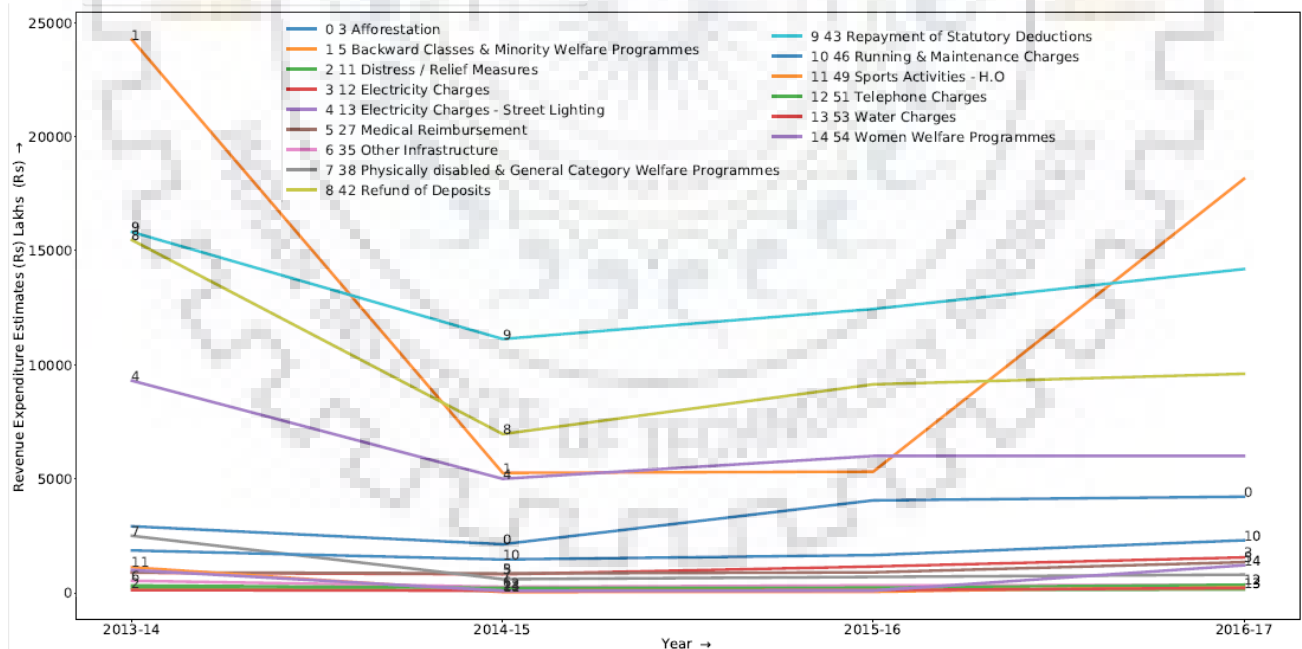


Fig 4.14 Sample Cluster for Revenue Expenditure Aggregated as per “Minor Head”

In this way, analysis can be done to any level of granularity. As stated earlier, if data is of few years, exploratory analysis is possible. However, as the data size increases, it becomes difficult to analyse it manually and the technique proposed can be applied to gain insights.

#### 4.4 Retail Store Dataset

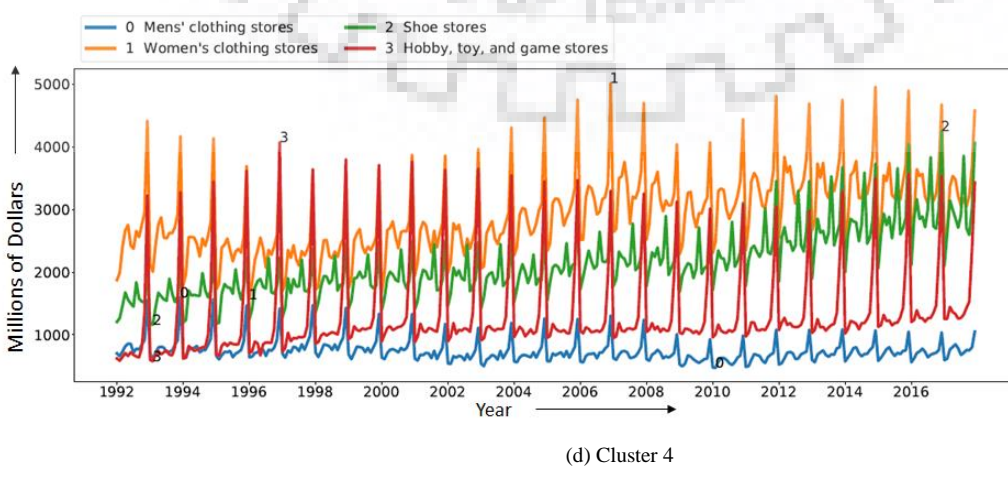
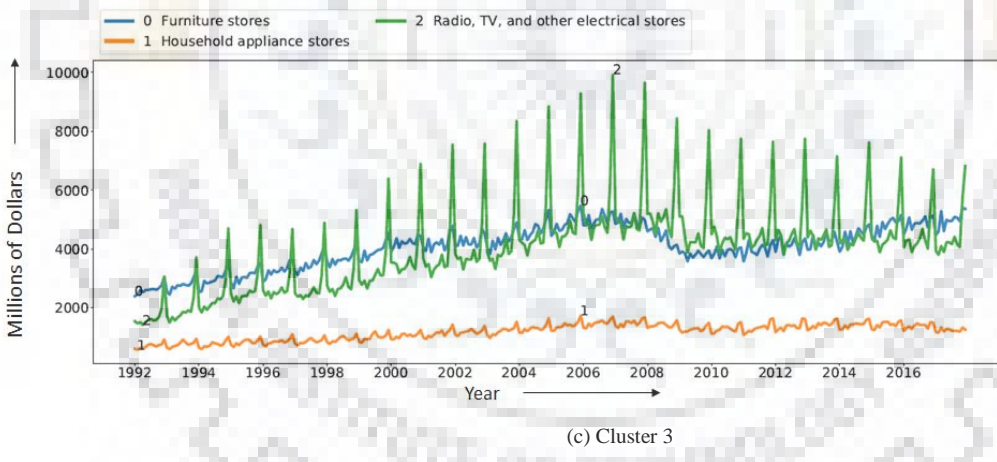
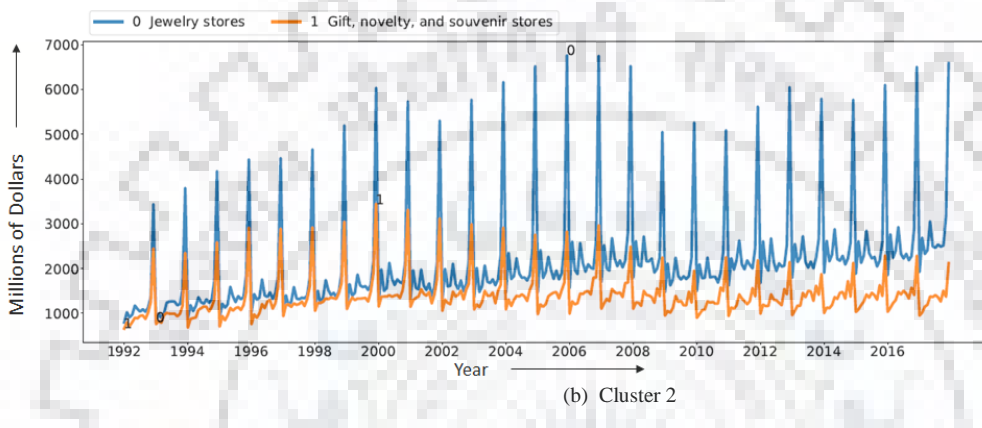
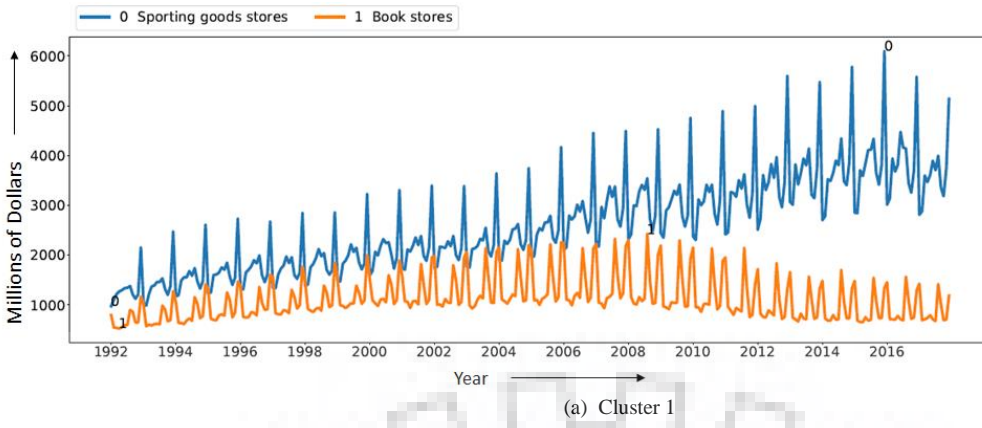
This experiment was performed on retail sales data of US stores. The stores considered for the experiment were

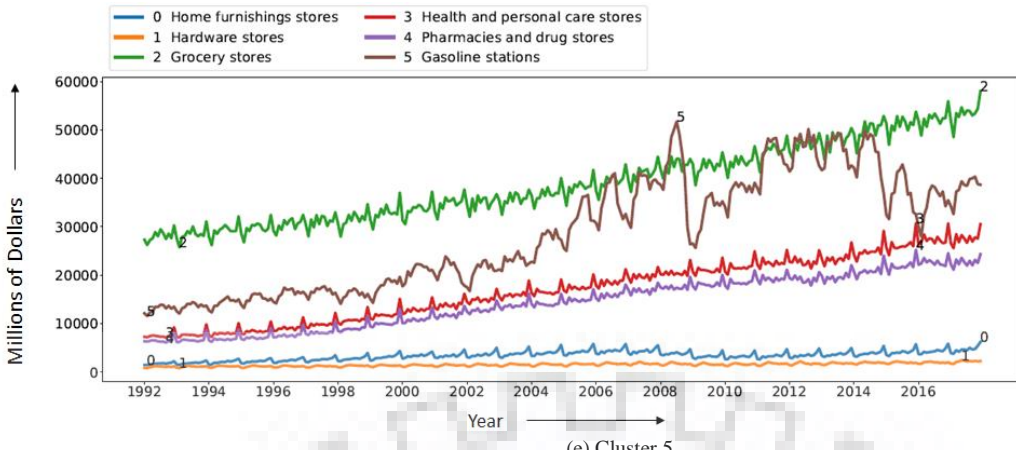
|  |                                 |
|--|---------------------------------|
| New car dealers                        | Health and personal care stores |
| Used car dealers                       | Pharmacies and drug stores      |
| Furniture stores                       | Gasoline stations               |
| Home furnishings stores                | Men's clothing stores           |
| Household appliance stores             | Women's clothing stores         |
| Office supplies and stationery stores  | Book stores                     |
| Radio, TV, and other electrical stores | Shoe stores                     |
| Hardware stores                        | Jewellery stores                |
| Gift, novelty, and souvenir stores     | Sporting goods stores           |
| Grocery stores                         | Hobby, toy, and game stores     |

For each store, the data set has monthly sales values. 132 values from January 1991 – December 2017 were considered. The weights were adjusted to moderate value so that both the factors (angular and increase decrease string-based distances) have same order of magnitude of differences (as explained in 4.1) Hierarchical Agglomerative Clustering was performed with complete linkage. 20 stores were divided into 7 clusters. Following clusters were formed

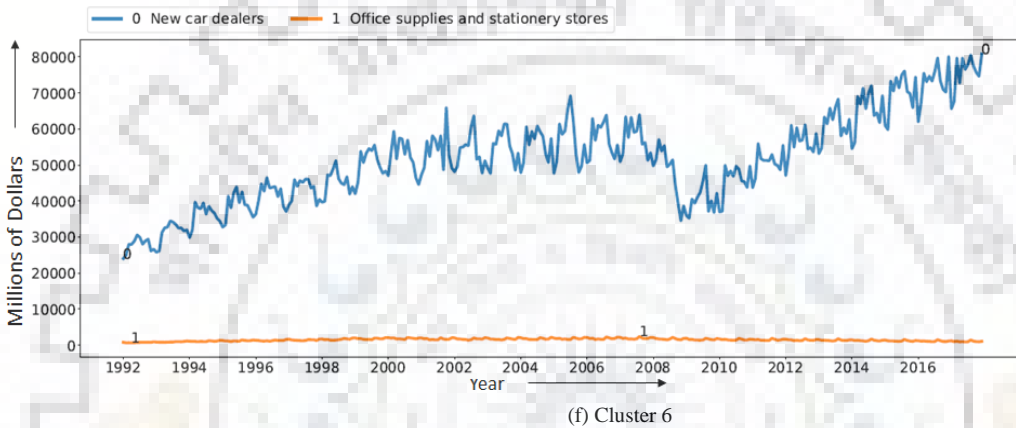
- Cluster 1 {Sporting goods stores; Book stores}
- Cluster 2 {Jewellery stores Gift, novelty, and souvenir stores}
- Cluster 3 {Furniture stores; Household appliance stores; Radio, TV, and other electrical stores},
- Cluster 4 {Mens' clothing stores; Women's clothing stores; Shoe stores; Hobby, toy, and game stores},
- Cluster 5 {Home furnishings stores; Hardware stores; Grocery stores; Health and personal care stores; Pharmacies and drug stores; Gasoline stations},
- Cluster 6 {New car dealers; Office supplies and stationery stores} and
- Cluster 7 {Used car dealers}.

As can be inferred from the results, items having similar sales pattern came in same cluster. Although the global trend of sales pattern may be different but the stores having items with similar shopping behaviour came into same cluster. For e.g. Cluster 2 depicts items which people generally buy during festival etc. when discounts are offered. Cluster 5 depicts items which are purchased regularly. Fig 4.15 depicts the plots of sales of various stores grouped in their respective clusters. It can be observed that stores within same cluster does not necessarily have same global pattern but they follow similar local pattern.

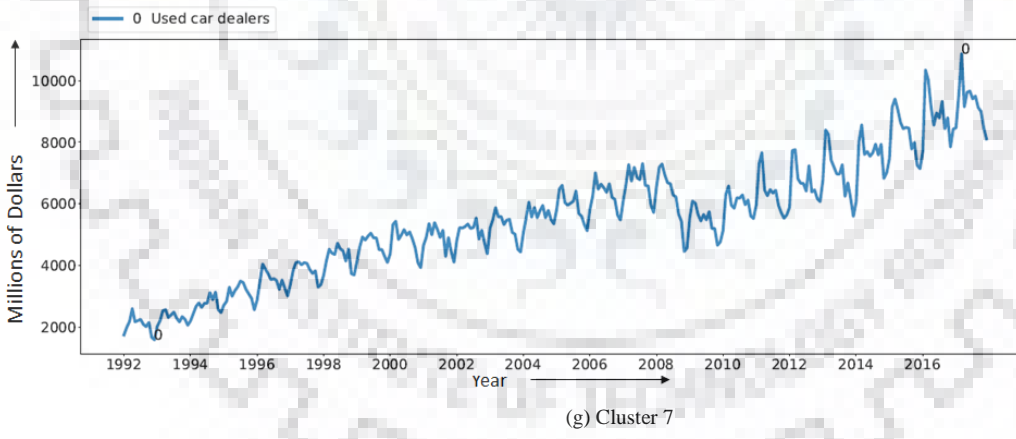




(e) Cluster 5



(f) Cluster 6



(g) Cluster 7

Fig 4.15 Clustering of Various US Retail Stores on basis of sales



## **4.5 Rainfall Dataset**

### **4.5.1 Subdivision wise**

This experiment was performed on rainfall data of India. Rainfall data of 30 subdivisions for period of 146 years was taken. The time series for each subdivision was prepared with the data points in chronological order (month-year wise). The results when subdivisions were divided into 5 clusters by the proposed approach were:

- Cluster1 {Coastal Andhra Pradesh, Rayalseema, Tamil Nadu, SouthInt.Karnataka, Kerela};
- Cluster 2 {Gangetic W.B, Orrisa, Jharkhand, Bihar, East Uttar Pr},
- Cluster 3 {West U.P. PLA, Haryana, Punjab, West Rajasthan, East Rajasthan, West Madhay P, East Madhya P, Gujrat, Saurashtra & Kucha, Vidarbha, Chhatisgarh},
- Cluster 4 {Assam, Naga.Mani.Mizo. & Trip, Sub-Hima. W. B} and
- Cluster 5 {Konkan and Goa, Madhya Maharashtra, Marathawada, Telangana, Coastal Karnataka, North Int. Karnataka}.

The results show that each cluster contains adjacent/neighbouring subdivisions which is supported by fact that adjacent regions exhibit similar climatic and rainfall pattern. The density of rainfall may vary but the variation in rainfall in neighbouring region is same. That is if a region experiences peak rainfall in month of June, its neighbouring regions will also experience the same pattern. The pattern changes slowly from one region to other.

### **4.5.2 District Wise**

The 5 years district wise rainfall data of India was taken. It was clustered and plotted on map Fig 4.16 (a). For some regions data was incomplete/unavailable and hence are not plotted.

The plot resembled the map which shows arrival period of monsoon in India Fig 4.16 (b). Since the rainfall in a region is affected by location of region and arrival of monsoon in that region, the results prove that the proposed approach can find similarities in pattern.

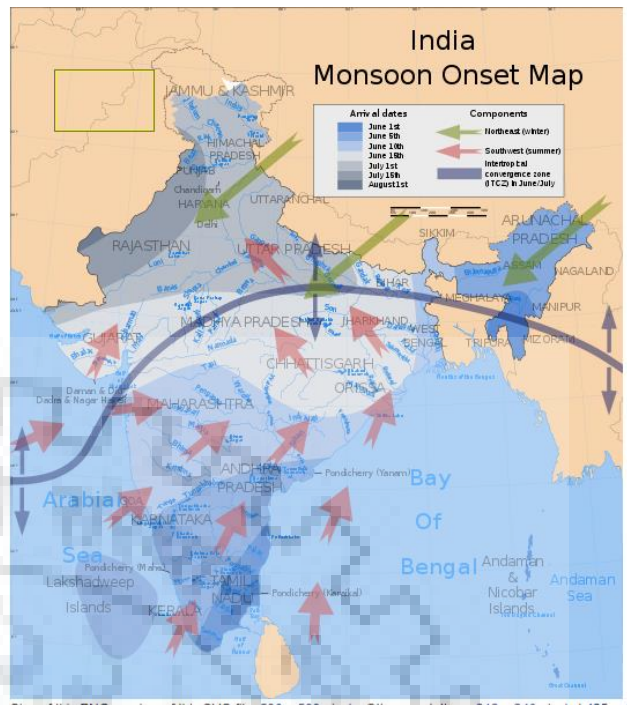
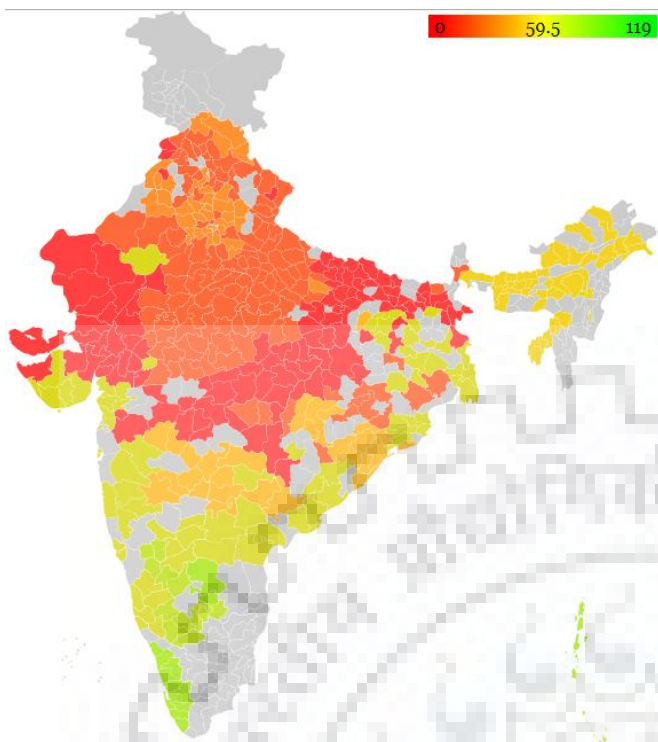


Fig 4.16(a) Clusters formed by proposed method for rainfall across districts in India

Fig 4.16(b) Monsoon arrival map of India



## 5. CONCLUSION AND FUTURE SCOPE

This work presented a generalized technique for similarity search and mining patterns in time series data. The approach uses two factors as its feature vector- angular values and Increase-Decrease Pattern. The approach was applied on variety of data. Particularly, problem of mining similar patterns in municipal budget data set lead to this approach, however, it can be applied to any time series in which local patterns need to be extracted. E.g. include market basket analysis for finding similarity in sales of various products. Although Municipal budget analysis in itself is a very broad area requiring expertise knowledge, application of the proposed technique over Municipal Data set was presented. Various other self-explanatory data sets of different domains (like shopping, whether) were also mined by applying the proposed technique.

This work can be extended in future for localized pattern mining of variety of time series data. Further as the budget data is getting digitized, the technique proposed will be useful for analysing tremendous amount of data. The statistical and exploratory analysis can be done only when data size is small, like budget data of few years. By programmes like Digital India, Government is digitizing data and soon data of several years will be available in digitized format. The method proposed will be of great importance for mining such data. And the scope lies not only in municipal budget data, but to other time series data as well. With urbanization and digitization, the data is getting generated in tremendous amount which can be mined for pattern similarity by the proposed approach.

## REFERENCES

- [1] Philippe Esling and Carlos Agon. Time series data mining. ACM Computing Surveys (CSUR) Surveys, Volume 45 Issue 1, Article No. 12. ACM New York, NY, USA, November 2012.
- [2] Dongkuan Xu and Yingjie Tian. A Comprehensive Survey of Clustering Algorithms. Annals of Data Science, Volume 2, Issue 2, pp 165–193, June 2015.
- [3] Jiawei Han, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques. Elsevier, June 2011.
- [4] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi and, Teh Ying Wah. Timeseries clustering A decade review. Information Systems, Vol. 53 Issue C, pp 16-38, May 2015.
- [5] Donald J. Berndt, and James Clifford. 1994 Using dynamic time warping to find patterns in time series. AAAI-94 Workshop on Knowledge Discovery in Databases (KDD-94), Seattle, Washington.
- [6] Jessica Lin, and Yuan Li. (2009) Finding Structural Similarity in Time Series Data Using Bag-of-Patterns Representation. In: Winslett M. (eds) Scientific and Statistical Database Management. SSDBM 2009. Lecture Notes in Computer Science, vol 5566. Springer, Berlin, Heidelberg
- [7] John Paparrizos, and Luis Gravano. 2016. k-Shape: Efficient and Accurate Clustering of Time Series. ACM SIGMOD Record, Volume 45 Issue 1, (pp.69-76)
- [8] Eamonn J. Keogh and, Michael J. Pazzani. Derivative Dynamic Time Warping. In Proceedings of the First SIAM International Conference on Data Mining, 2001
- [9] Durga Toshniwal and, R C Joshi. Using Cumulative Weighted Slopes for Clustering Time Series Data. International Transactions on Computer Science and Engineering, Vol. 20 Number 1 Pages 29-40, October 2005
- [10] Jessica Lin, Eamonn Keogh, Stefano Lonardi and, Bill Chiu. DMKD '03. A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, Pages 2-11, San Diego, California, June 2003. DMKD '03 Proceedings
- [11] Eamonn J. Keogh and, Michael J. Pazzani Scaling up dynamic time warping for data mining applications. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 285-289, Boston, Massachusetts, USA, August 2000, KDD '00
- [12] Sita Sekhar, Smita Bidarkar, 1999. Municipal Budgets in India, Comparison across Five Cities. Economic and Political Weekly Vol. 34, Issue No. 20, p.1202

- [13] Olga Parkhimovich and, Vitaly Vlasov. March 2016. Methodology for Evaluation and Rating of Open Municipal Budgets in Russia: a Research Plan. In Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance (pp. 371-373). ACM.
- [14] Bernardino Benito, María-Dolores Guillamón, Francisco Bastida, March 2015, Budget Forecast Deviations in Municipal Governments: Determinants and Implications. Australian Accounting Review, Vol. 25, Issue 1 (pp. 45-70)
- [15] Alan G. Mayper, Michael Granof, Gary Giroux, (1991) "An Analysis of Municipal Budget Variances", Accounting, Auditing & Accountability Journal, Vol. 4 Issue: 1,
- [16] <http://internationaljournals.co.in/pdf/GIIRJ/2014/March/14.pdf>
- [17] [http://pas.org.in/Portal/document/PIP%20Application/Municipal%20Finance%20Assessment\\_Paper.pdf](http://pas.org.in/Portal/document/PIP%20Application/Municipal%20Finance%20Assessment_Paper.pdf)
- [18] <https://rbidocs.rbi.org.in/rdocs/Content/PDFs/82506.pdf>
- [19] Brendan J. Frey, Delbert Dueck, 2007. Clustering by Passing Messages Between Data Points. Science, Vol. 315, Issue 5814, pp. 972-9

## LIST OF PUBLICATIONS

- A. Rathi and, D. Toshniwal. “Similarity Search for Localized Patterns in Time Series Data”, Learning and Reasoning: Principles & Applications to Everyday Spatial and Temporal Knowledge – IJCAI-ECAI 2018.  
(Communicated)

