APPLYING DATA MINING TECHNIQUES IN URBAN COMPUTING FOR SMART CITIES

A DISSERTATION

Submitted in partial fulfillment of the

requirement for the award of the degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

PREETI BANSAL

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY, ROORKEE

ROORKEE- 247667 (INDIA)

May, 2016

APPLYING DATA MINING TECHNIQUES IN URBAN COMPUTING FOR SMART CITIES

A DISSERTATION

Submitted in partial fulfillment of the

requirement for the award of the degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

PREETI BANSAL

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY, ROORKEE

ROORKEE- 247667 (INDIA)

May, 2016

CANDIDATE'S DECLARATION

I hereby declare that the work, which is being presented in the dissertation entitled "**Applying Data Mining techniques in Urban Computing for Smart Cities**" towards the partial fulfillment of the requirement for the award of the degree of **Master of Technology in Computer Science and Engineering** submitted in 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 the period from July 2015 to May 2016, under the guidance of **Dr. Durga Toshniwal, Associate Professor,** Department of Computer Science and Engineering, IIT Roorkee.

The matter 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 (**Preeti Bansal**)

CERTIFICATE

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

Date:
Place: Roorkee

(**Dr. Durga Toshniwal)**
Associate Professor
Department of Computer Science and Engineering
IIT Roorkee

ACKNOWLEDGEMENT

I would like to express my deep gratitude to my guide Dr. Durga Toshniwal, Associate Professor, Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, for her valuable suggestions, reviews and her extended support even in non-working hours. Her integrity, commitment and professional attitude has inspired me a lot.

I, hereby, express my deepest gratitude and heartiest thanks to Dr. Manoj Mishra, Head of Department, Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, for providing me the opportunity, guidance and necessary facilities. I am also thankful to all the staff members of the department of computer science for their support. I sincerely thank to my family and all friends for directly or indirectly helping me and giving me moral support that motivated me during the course of the work.

(Preeti Bansal)

ABSTRACT

To transform a city into a smart city it is important to focus on civic issues faced by the inhabitants. Civic complaints incorporate problems related to street condition, traffic, noise, water etc. Their analysis can contribute in proactive decisions to be taken by the city planners. Urban Computing is applied in many areas like transportation, environment, and security etc. but there is a need to explore more on urban planning from the perspective to analyze root cause of civic issues and reducing their concentration. In the present work, segregation of civic complaints based on different urban areas has been done and civic issues critical in an urban region are determined. For this purpose two approaches have been proposed namely Dynamic grid based clustering (DGCA) and clustering based on zip code approach (CZCA). A two phase clustering has been performed for both of the proposed approaches. The Phase 1 is different for both the approaches whereas the Phase 2 is similar. The purpose of Phase 1 is formation of spatial clusters. In DGCA, Phase 1 comprises of breaking the metropolitan area into grids representing spatial clusters. The granularity of the grid is determined by density of the civic complaints. In CZCA, Phase 1 comprises of dividing the civic complaints based on the zip codes of the region. The purpose of Phase 2 which is common for the two approaches is formation of sub-clusters based on complaint category over the spatial clusters obtained in Phase 1. These sub clusters are further analyzed to determine regions of city imitating similar complaint behaviour and finding the criticality of different complaint categories. For the purpose of experiment the real world dataset have been used for multiple metropolitan cities for USA and India. Experimental results have also been visualized to show better interpretation and compared with standard clustering algorithm and real world ground truth. The results are very promising and will help in planning strategies to improve inhabitant's satisfaction rate and consequently improving their quality of life.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Introduction and Motivation

With advanced lifestyle of inhabitants in cities many big challenges like traffic congestion, tremendous air pollution, incalculable energy consumption and overpowering noise pollution etc. are originating which needs to be tackled. New technologies are developing to transform cities into smart cities. Cities will be smarter if general amenities like water, electricity, transportation and clean air are managed in an efficient manner. Urban Computing is administrating these key issues by employing certain computing strategies like data collection, pre-processing of data, and interpretation of data and end-service provisioning [1].

The motivation for implementing these strategies in smart cities is to manage resources in order to improve the environment and also to lead up gradation from technological perspective. Urban Computing requires collaborative efforts from different fields like civil engineering, transportation, environment, economy etc. To procure the operation of smart cities, data from heterogeneous sources like city traffic system and weather forecast must be made accessible for monitoring, analysis and control. Multiple data accumulating technologies, data organization strategies, data interpretation models and unique representation methods are recursively applied on the data sensed from heterogeneous sources. The applications and components of urban computing which eases the life of inhabitants are elaborated in the next sub sections.

It can be observed that urban computing plays a significant role so it is required to explore it further. There are various urban computing techniques existing as discussed in literature review. Considering techniques meant for urban planning, it can be observed that they targeted primarily on requests referring to only single section of city operation. Hence the current research work aimed to consider the blend of requests generated from different utility services and intended to predict the underlying problems for their severe behaviour in urban area.

In the current research work various techniques are implemented to analyse the behaviour of general civic complaints reported from an urban area. Prior to description of the proposed techniques the below section introduces urban computing followed by the applications of urban computing.

1.2 Urban Computing

"Urban computing is a process of performing analytic and visualization techniques on the heterogeneous data acquired from ubiquitous technologies; so that it can contribute in improving urban planning methodologies, inhabitant's lifestyle and different city operation networks."

1.2.1 Application of Urban Computing in Smart cities

Urban Computing, being at a commencement stage, has been explored in different domains to resolve various challenges, existing in urban areas. Some of them are discussed below.

1. Urban Planning

It involves extracting the underlying problems existing in the city for e.g. identifying problems existing in transportation network. To make an efficient city planning it is required to discover the functional regions existing in a city. It can also involve dealing with daily civic amenities problems which inhabitants deal with. City planners can analyse the situations and work to unveil the problem existing behind the severity of the daily civic issues.

2. Urban Transportation

The transportation system in cities can be improved in following two aspects:

Urban Computing is used to *improve a driver's experience* [2] by providing personalized environment to driver so that it becomes easy for driver to take various decisions while driving. For e.g. Driver is given a customized path generated by incorporating factors like weather, traffic conditions and driver's habit etc. in a similar way as Google map does it. Travel time of a particular path can be estimated using historical observations, present situations like current traffic, distance covered till now with respect to time, peak hours, weather condition etc.

Urban Computing is used to enhance services [3] like predicting the most optimal taxi to pick a user in terms of economic and schedule feasibility. It includes defining time efficient strategies to assist users to quickly find a cab. The identical services used by present taxi services are Ola, Uber etc.

3. Urban Energy Consumption

Diminishing of the natural resources and rapid increasing environment pollution are a major concern and urban computing defines certain strategies to handle above problems. It involves monitoring of the amount of gas consumed as well as gas emitted throughout the city. It is done by analysing the refuelling events of taxicabs occurred at different gas stations

throughout a city. Refuelling event [4] basically denotes visit of a vehicle at a gas station to get refuelled. Analysing all such refuelling events will result into extraction of a pattern demonstrating a generic refuelling behaviour adopted by majority of customers which can be used to estimate the gas consumption of entire city. Gas consumed [5] and evolved at a particular region of urban area can also be predicted using the number of taxis traversing in that region and preventive measures can be taken to control the emission and alert the inhabitants that gas consumption had been determined in excess.

4. Urban Environment

To save the environmental conditions from getting deteriorated a technique to measure noise pollution [6] occurring due to different types of noise is applied. Noise generated due to various sources is accumulated and analysed to gather information like probability that certain set of noise will co-occur. The noise generated from various sources is categorized and analysed. In a similar manner air quality of city can also be evaluated.

5. Social Applications

Data from social media like Facebook, Twitter is used to determine similarity in user behaviour. For example, information like users check-in data, user likes etc. can indicate routine visiting pattern of citizens. Based on above information a place likely to be visited can be predicted.

6. Economy

The most appropriate locations for deploying more infrastructures and facilities like setting up of new shops, new gas stations and new amusements spots are determined by identifying the most crowded region of urban area and analysing citizen's generic visiting patterns. The optimum decision of proper deployment of resources will be reflected in economy.

1.2.2 Components of Urban Computing

Urban Computing is an interdisciplinary field, consisting of various components contributing as a comprehensive functional unit. The components are explained below -

1. Urban data collection - The city dynamics are sensed automatically by installing devices like GPS sensors mobile phones, vehicles, loop sensors at appropriate locations. The general traffic pattern and mobilizing behaviour of people can be predicted by employing humans as sensors. It includes analysing GPS traces, check-in data generated by user, tweets generated on social network intimating about user's location. Thus data is compiled from multiple heterogeneous sources, which can be noisy, and have skewed distribution

i.e. a large data related to a particular entity compared to void data for other entity. There are two ways to gather data as proposed by Yu Zheng et.al [1]

Passive crowd sensing [1] – The user generates data but is not aware that he/she has contributed in data generation.

Participatory sensing [1] – The user actively contributes in data generation e.g. human as sensors.

2. Data Pre-processing – Managing of the above gathered data in a well formed structure is performed in this step. The data structure may be in the form of trajectory or data stream. It is accomplished by techniques like data transformation, handling missing values, maintaining index based structure, pattern mining to make data more consistent and informative.
3. Data Analytics – After computing consistent data from above step, techniques like data mining, data visualization and machine learning are applied to extract the required information from the raw heterogeneous data.
4. Service providing – The extracted information from above step is delivered to end user as an application for example notifying the less congested routes to the drivers in case of transportation.

So urban computing is a hybrid system serving both city and inhabitants. It connects both real world and digital world as data collected from real-world is ingested for processing to platform like cloud and then sent to end user's mobile equipment. The applications of these components in urban computing are discussed in subsequent chapters.

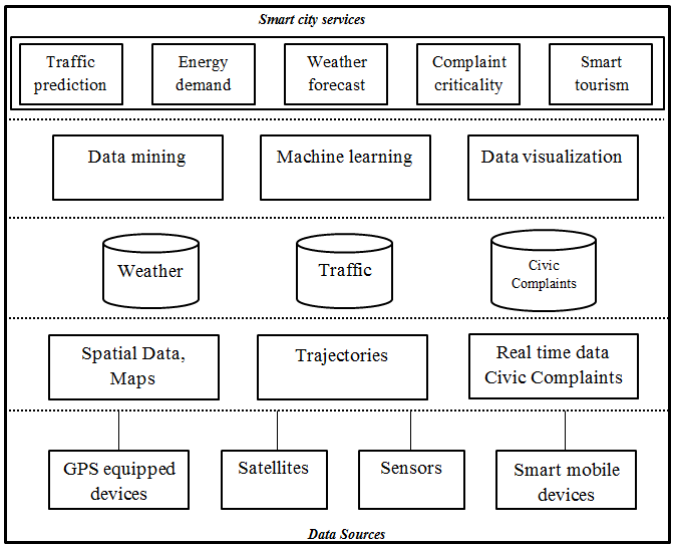


Figure 1.1: Components of Urban Computing

1.3 Problem Statement

"*To develop data mining techniques that help to track and predict the behaviour of the civic complaints in an urban area towards achieving smart city*".

Here Tracking basically indicates identifying significance of various utility services and their influence in different regions of an urban area. It can be done by applying data mining techniques that will analyse the urban data. This analysis will support in urban profiling of regions i.e. identifying which complaint type is acting critical in a region. This is how data mining techniques in urban computing will be helpful in evolving of smart cities.

The sub-problems are:

I. Data Collection and pre-processing
II. Formation of clusters based on spatial attribute using two approaches
    a. Dynamic Grid based clustering approach (DGCA)
    b. Clustering based on Zip Code approach (CZCA)
III. Categorization of clusters formed using spatial attribute
IV. Profiling of complaint category for a cluster formed on spatial attribute

Specific Contributions of the work

- In the present work specific contribution is towards analysis of the civic complaints.
- The civic complaints reported from metropolitan cities were spatially clustered using two approaches DGCA and CZCA; so that the interdependency occurring due to spatial proximity is taken into account.
- These spatial clusters were sub clustered based on complaint category and further analysed to determine the criticality of the complaint and categorizing the urban region based on complaint behaviour so that the city planners can get a summarized view unveiling the existing problems in an urban region.

1.4 Organization of the Report

The present work consists of five chapters, organized as follows Chapter 1 a brief introduction about the topic of this dissertation, the motivation behind it and the description of the problem statement. In Chapter 2, an overview of the different types of existing approaches used for urban planning is discussed along with the challenges associated with them. It concludes with identified research gaps. In Chapter 3, a description of the proposed framework catering to the needs of the problem statement is provided. Chapter 4 includes results of the current implementation. The present work end with the conclusions derived in Chapter 5.

Chapter 2 Background and Related Work

Urban Computing is an emerging field to solve multiple problems by analysing data generated from different source of urban areas. In this chapter, a brief overview of the existing state of research on the topic has been presented followed by a discussion of the various challenges faced and techniques used for solving them. In the following sub section, significance of urban computing along with the existing approaches for urban planning are discussed.

2.1 Significance of Urban Computing

Smart cities are developing rapidly with introduction of new practices and services for inhabitants. The association of the technological perspective with city planning and services is essential to facilitate growth of city. The optimum deputation of resources, analysing city dynamics and processing of heterogeneous data is required to initiate the proactive decision making. Data is generated from multiple sources including human as a sensor. The data from social media like check in records and twitter etc. is also used. The analysis serves both city as well as inhabitants. The information extracted from results will accelerate decision making process and also improve the quality of decision like identification of the regions having the requirement of more resources will lead to improvement in city functioning.

2.2 Urban Planning

Urban computing has made significant contributions in urban planning. Effective planning plays an important role to develop a smart city. The problems occurring in the smart city has to be tackled immediately and properly. To ease the task of city planners it is required to analyse various activity patterns associated with the city. There are various approaches to analyse these activities and many aspects to be considered in city planning – city construction dynamics, city areas such as residential/industrial, preferable areas to live, areas where new infrastructure is progressing, noise prone area, usually crowded regions etc.

Using the clustering algorithm the system classifies a set of objects by grouping together similar objects on the basis of some similarity measure and separates the dissimilar objects automatically. This is an unsupervised learning method. The different types of clustering are partitioning, density based, grid based, model based, hierarchical. The existing approaches are focused on finding the similarity between the patterns, grouping them into a set, analysing further and suggesting the solution to the existing problems.

2.2.1 Analysis of construction requests

Tzu-Chi Yen [7] proposed an analytic and interactive visualization system to track and predict the urban construction behaviour and the system analyses the city construction dynamics. The data of construction requests corresponding to one year duration was acquired from Taipei City, Taiwan to develop the system. The visualization system helped the users in understanding the interdependency between different construction requests along with time duration and corresponding geographical regions. K-means, clustering based approach was used to cluster the regions and to identify the relationship between different types of construction requests occurring in certain regions. The number of clusters was finalised as four (K=4) based on the analysis (K = 4 to 6). Similarity in frequency patterns of construction requests was observed over some regions which represented the construction needs of that geographical region. The system was also used as a predictive model by applying regression based modelling on historical data of one year to predict upcoming number of construction requests. The above analysis would be helpful in identifying the underlying problems for occurrence of construction requests. The analysis would also highlight such construction problems which have longer resolution duration. It would be helpful to identify the drawbacks in construction management system. The above approach could be used for another type of urban data for e.g. 311 requests and extended for real time prediction service. Moreover more urban data like traffic requests, check-in data could be incorporated to explore the system further and increase its efficiency.

2.2.2 Land-use modelling for urban planning

Vanessa Frias-Martineza et.al [8] proposed solution to a similar kind of problem which was identification of the land uses using the tweets reported from that region as the data source. The classification of the urban land use as industrial, residential, parks etc. is crucial for city planners and authorities. The classification was termed as urban zoning which meant division of the maps into certain zones. Land segmentation i.e. partitioning of land data was performed by formulating activity vectors for the land segments. Each land segment was further characterized by its usage. Average tweeting activity for the region was computed to characterize each land segment. The average tweeting activity of a particular land segment was denoted by activity vector computed as the number of land requests registered in twitter during a certain interval of time e.g. 20 minutes in a particular day and for a particular land segment. Clustering was applied on the obtained activity vectors to find the urban land use. A large number of clustering techniques could have been used like K-means, decision tree,

hierarchical clustering. But spectral clustering was used as it doesn't require any prior knowledge about the cluster as required in K-means. Spectral clustering could tackle high dimensional data by using dimensionality reduction and provide good clusters with low computation cost. Each cluster contained activity vectors of the land segments that were included in the respective clusters. An average activity vector was computed representing the tweeting activity of the cluster to analyse the type of land use associated to each cluster. Thus the system gave the understanding of traditional land uses of an urban area. Limitation of this method is that as the data was fetched from micro-blogging site like twitter, the data may be ambiguous because it defers from individual to individual how they share their thoughts. Hence there was a need to first analyze the context of tweets and then separate them accordingly.

Similarly the categorization of land use and comparison of the organizations in different urban areas of Spain was done in [9]. In this approach mobile phone records were used to define the activity profiles along time over a week. The region was divided in to cells. A Pearson correlation matrix was computed between cell activities and converted into a weighted graph and further clustered using community detection techniques. The technique was applied on different sizes of cities in Spain which was helpful in comparing different regions of Spain. The technique was helpful in identifying four different types on land uses each representing different temporal patterns.

2.2.3 Analysis of city dynamics for urban planning

Mobility patterns followed by citizens have a significant impact in city planning decisions. The city planners requires extracting information like the most crowded region of the city to take decisions such as building new infrastructures accordingly, taking precautions during disaster etc. Hence, to explore city dynamics, 18 million check-ins data was collected from foursquare, used by Justin Cranshaw et.al [10]. Check-ins data was used to cluster nearby foursquare venues which have similar users visiting pattern using spectral clustering technique. To compute this, an affinity matrix was constructed consisting of all foursquare venues and number of times a user U_i checked-in that venue. Cosine similarity was computed between the venue selected, and other m closest venues. This computation was performed for each venue V_j. Value of cosine similarity varies from -1 to 1 where -1 denotes least similarity and 1 denote highest similarity. The venues were then represented in form of nodes and a weighted graph was formed using affinity matrix. For e.g. Let V_1, V_2…..V_n be the set of

venues in foursquare. Each venue V_j would be U dimensional where U represented the number of user components. Each entry V_{ij} will denote number of times user U_i checked-in to venue V_i. Cosine similarity was computed to get social similarity between two pairs of venue. The affinity matrix was filled using the spatial distance between m closest venues. Let's assume for Venue V_j, the U dimensional vector representing number of visits made by each user U_j would be as shown in Table 2.1.

Table 2.1: Vector for a Venue representing number of visits by different check-in users

User	U_1	U_2	……	U_u
No. of Visits	12	10	……..	11

Similar vector was formed for other *n-1* venues. Then cosine similarity between each pair was calculated. Then affinity matrix of n x n dimension was computed by filling value of cosine similarity for m closest pair of venues considering each venue. An example of an affinity matrix is given in Table 2.2.

Table 2.2: Example of an affinity matrix

$$\begin{bmatrix} 12 & \cdots & 20 \\ \vdots & \ddots & \vdots \\ 11 & \cdots & 0 \end{bmatrix}$$

Then spectral clustering was applied. Spectral clustering is a top down clustering which represents the data in the form of a graph. The process of formation of cluster was treated as a graph partitioning algorithm. The clusters formed by spectral clustering had many properties like volume of a cluster, cut in cluster, which were required for partitioning of the graph. The proposed model clustered venues based on their spatial attributes as well as the from the check-ins data denoting their social proximity. As a result it revealed the local social patterns and characteristics of the city.

A similar service which identifies city dynamics using spatiotemporal data was suggested in [11]. It basically utilized the benefits of huge amount of geo-located data generated due to user's mobility. Data consisting of mobility tracks pertaining to various sports activities was used for this purpose which was collected from Nokia Sports Tracker (NST). Data consisted of information related to GPS route, distance, duration and speed corresponding to individual sport activities. Kernel density estimation was used to identify the hotspots which referred to the areas in the city which were most active for a particular sport at a certain time. Use of kernel density estimation was helpful to anticipate the data values for entire population. This approach would be helpful in identifying the active regions in a city like regions popular for biking or more crowded recreational areas so that urban planners could provide apt services in those regions. Moreover identifying such regions would be helpful in disaster management.

2.2.4 Urban planning influence on environment

It was observed that urban planning might affect environment. It could be contemplated that existence of different infrastructure, city operations and millions of people performing multiple activities leads to generation of plenty of noise. Zheng, Yu, et al. [6] proposed a solution which aimed to diagnose the origin of urban noises and factors leading their generation so that effective decisions could be made and citizens could be made aware about the existing situation. For this, 311 NYC data was collected which was processed and represented in form of a matrix as shown in Figure 2.1; indicating frequency of noise complaint in a particular region during a particular time slot. To remove the sparseness extra features like check-in records, point of interests were combined. This was basically done to bring uniformity as there was limitation of data sparseness when humans would not act as active contributors to generate data. Hence, additional features were processed in the analysis which recovered the missing instances of noise related request. Such type of analysis could help government officials to tackle noise pollution.

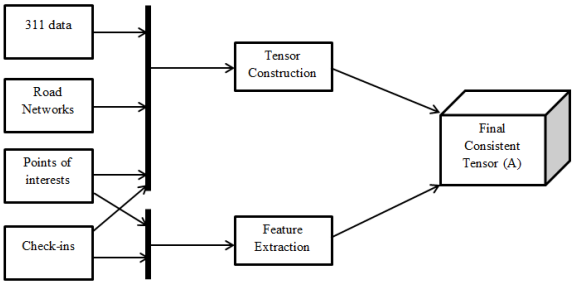


Figure 2.1 Modelling of urban noise system

2.2.5 Urban planning influence on civic needs

Ali et.al [12] proposed an approach to analyse citizen's requirements. Data about citizen's complaints was collected for Tehran city and K-Means clustering algorithm was applied to group the category of complaints using the parameters like frequency, the time interval and number of days. The tests were performed for different values of K and K=3, was considered as the optimal choice. Each cluster depicted some property like cluster 2 had more number of construction and traffic. There were at least 5 complaints in a day, with interval of time as 13. It was observed from results that there was no uniformity in occurrence of such complaints, they occurred occasionally like not during winter season. Association rule based mining was used to infer the factors that may affect the state of satisfaction as each record consisted of attributes like satisfied and unsatisfied. It was available to segregate agency based on rate of satisfaction and identify influence of time on satisfaction rate.

Another approach to analyse non-emergency calls from 311 data was done by Yilong [13]. The 311 calls analysis would be helpful in many aspects like prediction of status of the city, diagnosing the efficiency of different departments contributing in solving the 311 requests. The analysis would be helpful in prediction of frequency of calls related to different problems in a city. For this significant features were extracted from additional data sources like U.S calendar and NYC historical weather data. Linear regression models were used to identify the features which were significant and the frequency of future events. Features like day of week, snow, average temperature, temperature range, public holiday and 311 calls for last week were considered as significant. These features were combined to estimate the count of 311 calls in the future. The features were evaluated by conducting many predictions, each with absence of certain feature. Then the accuracy of feature was computed using MSE. This process lead to extraction of important features accompanied with significance of every feature. The 311 requests for last week/7 days were considered to be the most significance feature. The 311 request count was estimated separately for each complaint type. This is how the analysis would be helpful in diagnosing the city status by city planners. On studying the analysis approaches, some limitations were observed in the current research, which are enlisted as research gaps in the next section.

2.3 Research Gaps

Based on the literature review done, it was identified that to facilitate urban planning, existing work basically included analysis of noise complaints, construction requests or identification of type of land use etc. The civic complains like streetlights, road conditions and parking etc were not considered for analysis. So a need was identified to include all the requests corresponding to different utility services so that multiple questions of a city planner can be answered. Additionally, providing integrated view of how various utility services influence a region will also be helpful. In order to capture this integrated view there was a need to focus on different city problems faced by the inhabitants. An analytics and visualization system needs to be developed to analyse the critical complaints which are high in frequency and occurs recurrently so that city planners can take pre-measures and provide a prompt service to the citizens. Apart from criticality of a complaint, the system should also be able to categorize the regions based on civic complaints behavioural pattern. This comparison will be helpful to both city authorities and citizens. In the present research work; as described in the next chapter, a framework has been proposed which tries to provide a solution for such a system.

Chapter 3 Proposed work

This chapter discusses the proposed research work. The work aims to develop a platform which analyses civic complaints for proactive maintenance of smart city.

3.1 Proposed Framework

The analysis of civic complaints was carried out using two approaches. The approaches dif fer in the way the civic complaints were clustered based on spatial attribute.

3.1.1 First Proposed Approach: Dynamic Grid based clustering approach (DGCA)

In this approach the analysis of civic complaint is carried by clustering the civic complaints based on the spatial attribute using the dynamic grid based technique so it is named as Dynamic Grid based clustering approach (DGCA).

Data pre-processing block

The data consisting of complaints from multiple cities is segregated based on location attribute and the civic complaints corresponding to single city are processed further. The two steps performed in data pre-processing block are segregation based on location and de-noising of data.

Phase 1: Dynamic Grid based Clustering block

In this module, each broadly segregated set of complaint is encapsulated in a grid covering all the relevant complaints present in the concerned area.

Phase 1: Merging of Location based Clusters block

It might occur that few complaints lie on the common boundary of two adjacent bounding boxes which may affect the further analysis. It will affect the analysis because some existing clusters might have got divided between two bounding boxes. So, merging of these bounding boxes is required.

Phase 2: Categorization of location based clusters block

In this block, all the location based clusters are studied to find the similarity among clusters. For each one of the location based cluster, sub-clustering is done and categorized in six different scenarios.

Phase 2: Profiling of complaint category for a location based cluster block

It will be advantageous to identify the severe complaints in a particular region so that actions

can be taken to resolve or scale down such complaint on priority basis. Hence, the criticality of a complaint type within a location based cluster is estimated. Criticality Score is measured using temporal, density or resolution score and typical time slot factor of the complaint category.

3.1.2 Second Proposed Approach: Clustering based on Zip Code Approach (CZCA)

In this approach, civic complaints are segregated on the basis of zip codes. Actual boundaries drawn in zip code based shape files are used to separate the complaints. The approach clusters complaint records based on Zip Code, so it is named as Clustering based on Zip Code Approach (CZCA).

Chapter 4 Experimental Results

The proposed approaches are implemented using the programming language Java and SQL server 2008. The approaches are applied on complaints collected from different urban areas. The urban areas are listed in Table 4.1.

Table 4.1: Description of different urban areas dataset

S.No.	Regions for which Datasets are considered		Number of records
	Metropolitan Areas for USA		
1.	Austin	11th most populous city in United States	1,00,000
2.	Boston	Largest City in New England	33,129
3.	Chicago	3rd most populous city in United States	64,627
4.	New York region	Includes regions like Bronx, Brooklyn and many suburbs Arverne, Woodside	1,00,000
5.	San Francisco	Second most densely populated	98,877
Metropolitan Area for India			
6.	Bangalore	5th most populous urban agglomeration in India	25,986

On an average 50k records were considered for analysis. The set of civic complaints in form of text for each of the above urban area are archived in SQL Server.

4.1 Dataset description

Attributes in generalized schema for the datasets corresponding different urban areas are listed in Table 4.2. Results and observations are discussed below for the urban areas considered.

Table 4.2: Set of attributes common in all datasets

Attribute	Description
Complaint Number	It is a unique key representing a complaint record.
Created Date	Timestamp on which complaint was registered.
Complaint type	The complaint category mentioned in few words.
Complaint description	Description of the registered complaint in the form of text.
Closed Date	The timestamp when complaint record was marked as closed.
Status	Resolved/ open/ in progress/ closed
Updated date	The timestamp when complaint record was updated.
Agency	The agency responsible to resolve complaint.
City	City from which complaint was reported.
Location	Address where the complaint is registered.
Spatial Location (Latitude/Longitude)	Pair of latitude and longitude.

4.2 United States Metropolitan Cities

4.2.1 New York, USA

Data pre-processing block: The dataset comprises of the complaints collected for New York. The numbers of civic complaints recorded from July 2015 to October 2015 are 100k. The dataset comprises the complaints from cities like New York City, Bronx, Brooklyn, Staten Island, Astoria, and Arverne etc. Complaints were recorded from 45 different city suburbs and the data was separated for all the small cities using city attribute. The data was de-noised by discarding records with invalid location and blank location attribute. Table 4.3 lists some

results of data pre-processing block for New York region.

Table 4.3: Sample data snapshot after Data pre-processing for some regions pertaining to NY dataset

City	Bronx	Brooklyn	NY City	Staten Island
No. of complaints initially	15453	28455	19821	4412
No. of complaints after pre-processing	15259	28272	19906	4445

It was observed that 4 cities namely Bronx, Brooklyn, New York City, and Staten Island had complaints records more than 1500 and so phase 1 computation of these 4 cities was carried out. The remaining 41 suburban cities having number of civic complaints less than 1500 were considered as separate 41 clusters and forwarded for phase 2 analyses directly.

Results using first proposed Approach: Dynamic Grid Based Clustering Approach (DGCA)

Phase 1: Dynamic Grid Based Clustering

This process is carried separately for the 4 cities mentioned above. In dynamic grid based clustering block we first identify the grid encapsulating all complaint records. To formulate the grid we find the maximum and minimum value of latitude and longitude by querying SQL server which archives the complaints for each city. We need to select optimal ϵ epsilon value i.e. number of complaint records to be encapsulated in each grid divided. So we carried out this division for five values of ϵ (epsilon) that are 5%, 10%, 15%, 20% of the number of complaints in dataset. To select the optimal ϵ epsilon value we computed cluster validity index for all location based clusters formed by each division and selected the minimum value as shown in Figure. 4.1. It can be derived from the plot that for Bronx and Brooklyn epsilon value is shown as 10% of the total dataset and for New York City and Staten Island it is selected as 15% of the total dataset. The set of location based clusters selected corresponding to optimal epsilon value is forwarded for next step of the process i.e. merging of location based clusters. Table 4.4 enlists number of location based cluster formed based on optimal epsilon ϵ value.

Table 4.4: Number of location based cluster before merging step

City	Bronx	Brooklyn	NY City	Staten Island
No. of location based cluster formed using optimal ϵ value	22	36	25	9

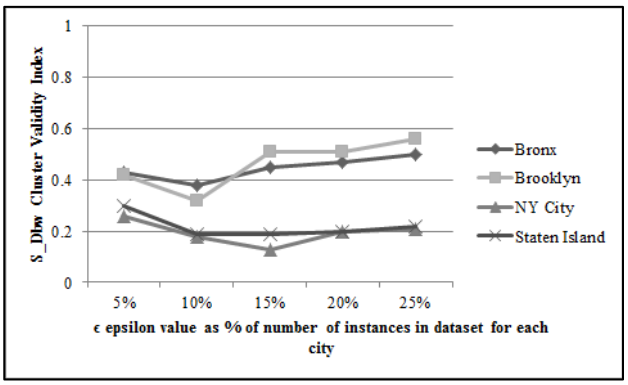


Figure 4.1: Plot of epsilon value v/s cluster validity index

Phase 1: Merging of Location based Clusters

In merging of location based clusters we calculate density for each location based cluster. For this range *R* needed to be calculated, in present work we have taken *x* value equal to 10 to calculate *R*. Table 4.5 enlists the number of final location based clusters formed after merging and the reduced *S_Dbw* cluster validity index.

Table 4.5: Result after Merging of location based cluster

City	Bronx	Brooklyn	NY City	Staten Island
No. of location based cluster formed on merging	17	31	17	7
S_Dbw cluster validity index after merging	0.31	0.3	0.10	0.17

Phase 2: Categorization of location based clusters

To categorize location based clusters formed from New York region the six scenarios mentioned in Section 2 are considered. Including the 41 suburban cities which are not considered in Phase 1 and clusters formed from 4 cities; total 113 clusters were formed. The results generated from both distance measures were compared with ground truth as shown in Table 4.6. Ground truth is calculated by representing each location based cluster with complaint type, identified to be maximum among the six scenarios. Figure 4.6, 4.7, 4.8 represent categorization of location based clusters in six scenarios represented as bounding boxes. Figure 4.9, 4.10, 4.11 represent same categorization, but the 41 cities which were not forwarded through Phase 1 are represented in form of polygon. It can be observed that visualization on map shows some overlapping. The overlapping occurs because of merging step. All location based clusters are spatially disjoint. Reason of overlapping is discussed below.

Reason of overlapping in the plot for categorization of location based cluster

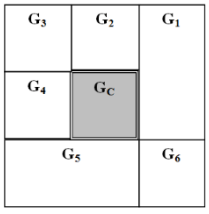


Figure 4.2:Example snapshot of grid division

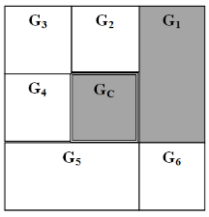


Figure 4.3: Grid on merging based on density

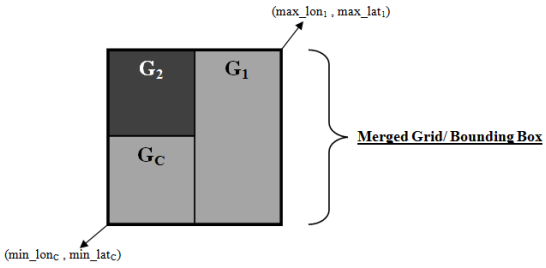


Figure 4.5: Snapshot of merged bounding box/grid

Consider the grid clustering as shown in Figure 4.2. Let G_C be the candidate cluster selected for merging. Clusters adjacent to G_C for e.g. G_1, G_2, G_3, G_4, G_5 are considered for merging. Density of each adjacent cluster is checked if it comes in range R or not. Suppose it is found that among the adjacent clusters G_1 falls in density range R. So G_1 and G_C can be merged as shown in Figure 4.3. From Figure 4.4 spatial coordinates of grid G_1, G_2, G_C can be seen. In

this case $min\_lat_C = min\_lat_2$, $max\_lon_2 = max\_lon_1$. So on merging the coordinates of bounding box for visualization results as (min_lon_C, min_lat_c, max_lon_1, max_lat_1) as shown in Figure 4.5. It can be seen that though G_2 is not merged with G_1 and G_C based on density, then too it gets visually encapsulated within the merged grid/ bounding box. Hence only the visualization plot shows overlapping but practically G_2 and merged G_1 G_C are disjoint.

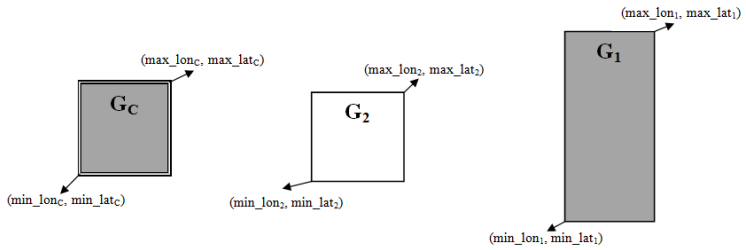


Figure 4.4: Spatial Coordinates of Grids

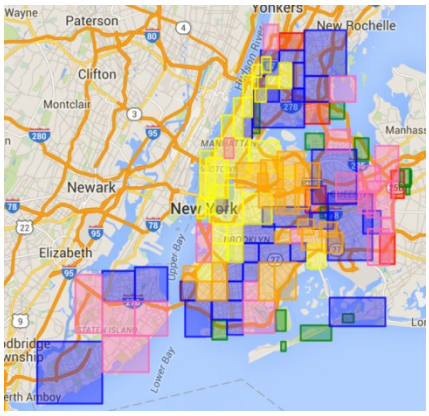


Figure 4.6: Bounding Box plot using Euclidean measure

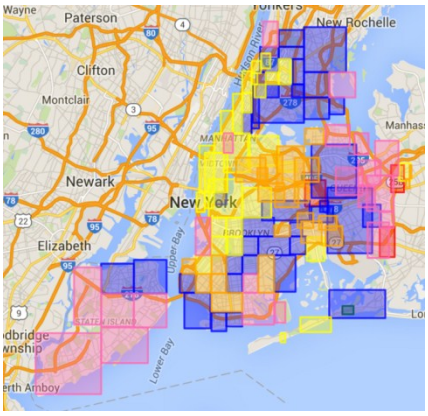


Figure 4.7: Bounding Box plot using Cosine Similarity

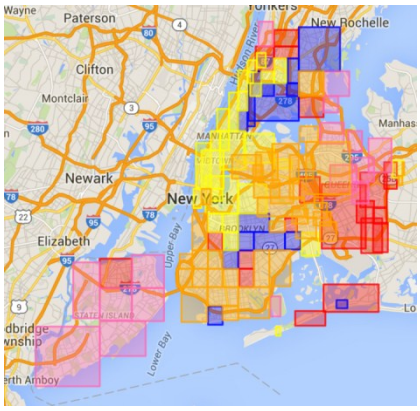


Figure 4.8: Bounding Box plot for Ground Truth

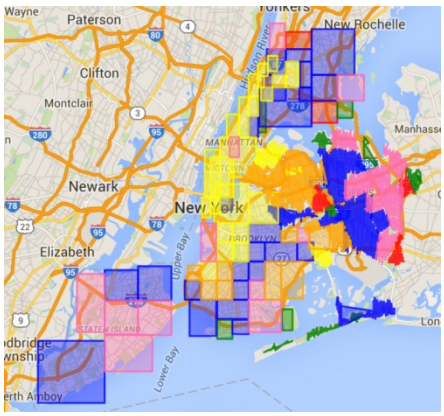


Figure 4.9: Polygon plot using Euclidean measure

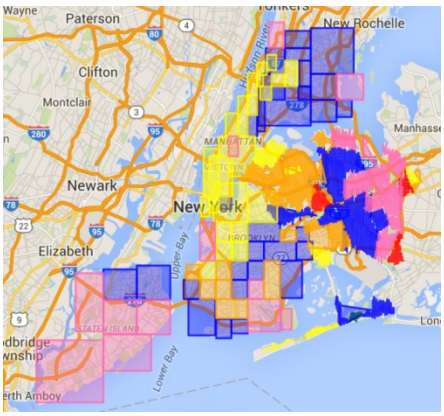


Figure 4.10: Polygon plot using Cosine Similarity

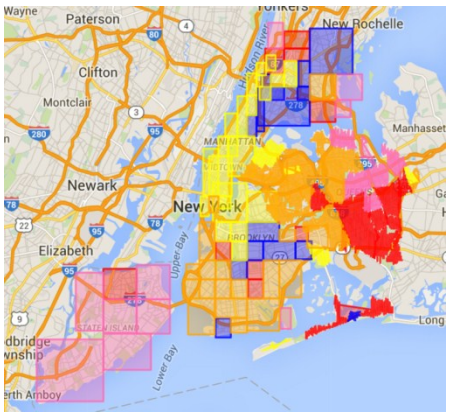


Figure 4.11: Polygon Plot for Ground Truth

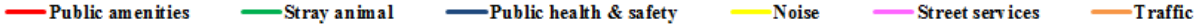


Table 4.6: Accuracy results on performing categorization

Measure used for categorization	Euclidean Distance	Cosine Similarity
Accuracy	60.17%	68%

Phase 2: Profiling of complaint category for a location based cluster

The criticality of the complaint categories in a particular location based cluster is identified in

this block. 113 location based clusters were formed after phase 1. We identify some clusters which display the highest severity in terms of criticality score for some complaint type.

Density or Resolution Score (*DRC*)

This factor is identified in two ways. If complaints are high concentrated, kernel density evaluation is used, else average resolution time period is computed and allotted weighted score according to the range. Density and Resolution Score values are listed for the location based cluster corresponding to different complaint type in Table 4.7. The *DF* and *RTF* values are equal to 1 as we are only considering the location based clusters where a particular complaint type is found to be critical. Hence *DRC* also results to be 1.0.

Table 4.7: Density or resolution score for some clusters

Critical Complaint Type	**Cluster Name**	**Complaint Type**	**Density/Resolution Score**
Park Related Complaints And Parking	Cluster LBC_{65} ("Prospect Park, Brooklyn ")	Air Pollution	0
		Construction	0.034201
		Electric	0
		Fire Alarm	0
		Noise	0.006961
		Park	**0.207856**
		Illegal Parking	**0.215147**
		Public Amenities	0.170299
		Public Health	0.010756
		Stray Animal	0.168082
		Street	0.077284
		Streetlight	0.19341
		Traffic	0.087084
		Water	0.050549
		$DF_{park} = DF_{parking} = 1.0$	
Traffic	Cluster LBC_{61} ("Dahill Road, Brooklyn")	Air Pollution	0.011333
		Construction	0.134142
		Electric	0.001769
		Fire Alarm	0
		Noise	0.141627
		Park	0.025158
		Parking	0.009147
		Public Amenities	0.097838
		Public Health	0.02587
		Stray Animal	0
		Street	0.028546
		Streetlight	0.137059
		Traffic	**0.239653**
		Water	0.019328
		$DF_{traffic} = 1.0$	
Electricity Related Complaints	Cluster LBC_{111} ("Westerleigh")	$RTF_{electric}$ = 1.0 As Average Resolution Time Period Is 0	

Temporal patterns of complaints (TF)

Table 4.8 displays some set of correlation values which increase criticality score due to co-occurrence of two complaint types. So this temporal pattern of co-occurrence is included in computation of criticality.

Criticality Score Computation

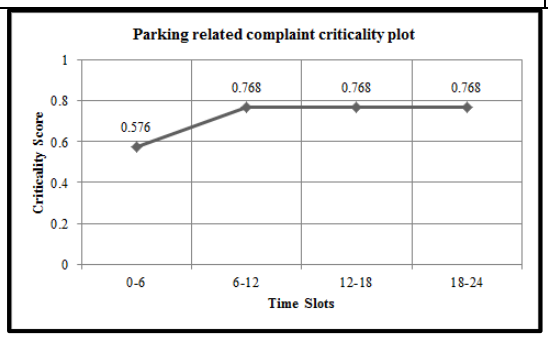
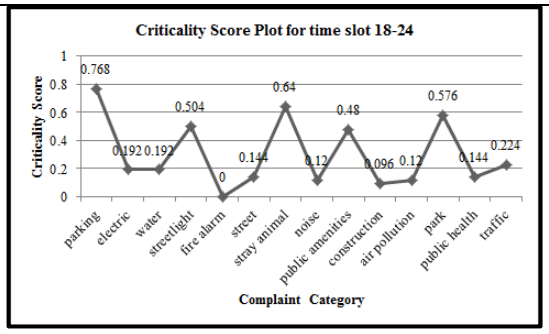
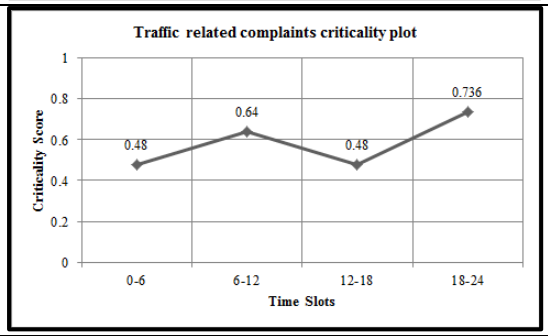
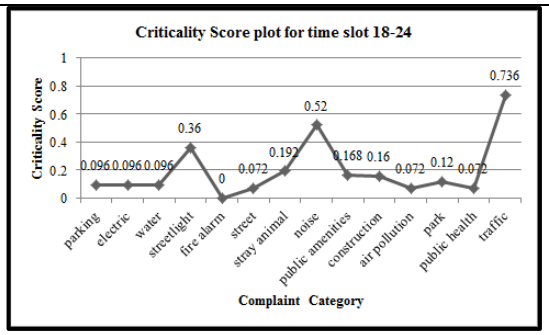
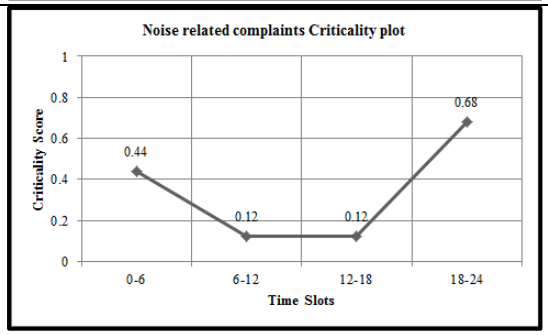
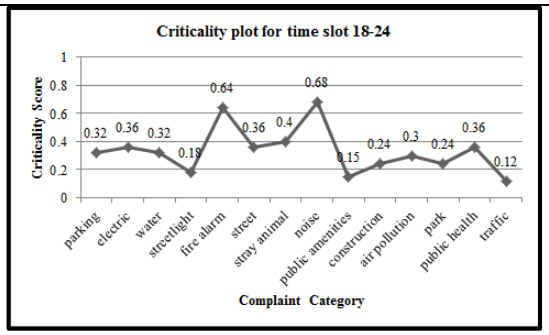
Two plots are drawn corresponding to some cluster as shown in Table 4.9. First plot represents criticality score variation of one of the identified critical complaint for four time slots. Second plot represents criticality score in the cluster for each complaint category in a particular time

slot.

Table 4.8: Correlation values for some clusters

Cluster Name	Location	Correlated Complaints		Correlation Amount
Cluster LBC_{61}	"Dahill Road, Brooklyn"	Air Pollution	Traffic	0.18
		Parking	Traffic	0.23
		Street Services	Traffic	0.27
		Streetlight	Traffic	0.91
		Noise	Traffic	0.41
Cluster LBC_{13}	"Forest Hills"	Street Services	Traffic	0.87
Cluster LBC_{23}	"Maspeth, Queens"	Construction	Noise	0.50
		Noise	Traffic	0.41
		Noise	Stray Animal	0.82
Cluster LBC_{88}	"Sheepshed Bay, Staten Island"	Public Amenities	Street Services	0.65
		Public Amenities	Park Service	0.75
		Public Amenities	Public Health & Safety	0.73
		Public Amenities	Stray Animal	0.79
		Public Amenities	Streetlight	0.99

Table 4.9: Criticality Score plots for some clusters

Cluster name	Criticality Score variation for four time slots	Criticality Score for all Complaint Categories for a selected time slot
Cluster LBC_{65} ("Prospect park, Brooklyn")		
Cluster LBC_{61} ("Dahill road, Brooklyn")		
Cluster LBC_{106} ("Time square, NYC")		

Inferences:

It can be inferred from first approach results that cosine similarity categorizes the location based clusters more accurately as compared to Euclidean measure. The criticality score plots of different location based clusters helps in ranking the areas represented by location based clusters based on the severity of the complaints faced. The criticality plots help in inferring

many observations. They are:

- Complaints related to traffic are high in time slots 6-12 and 18-24. It can be interpreted that it is because of office hours.
- Complaints like noise is highly critical during 18-24 as sensitivity towards noise disturbances might increase during night hours. Noise is even critical during 0-6 slot.
- It can be inferred that noise and traffic might be correlated as they are highly critical.
- If we consider location based cluster (Cluster LBC_{106}) near Time Square, New York and analyse its criticality plot we find complaints related to noise, fire alarm, traffic are reported to be critical.
- It can be inferred that problems like traffic, air pollution, noise, parking complaints, street services co-occur. Similarly fire alarm and complaints related to public health & safety issues co-occur.
- Cluster LBC_{65} reporting issues related to park complaints is found to be area near prospect park, Brooklyn. The public park covers the area of around 585 acre.

Second Approach: Clustering based on Zip code approach (CZCA)

Phase 1: Clustering based on geographical boundaries representing zip codes block

To cluster civic complaints based on geographical boundaries representing zip codes shape file corresponding to New York region is fetched. The shape file is than converted into SQL table consisting of polygon shape, area corresponding to polygon, coordinates of each polygon. Each polygon is formed by geographical boundary representing different zip code. 65 polygons were identified for New York region. Complaints overlapping with each polygon were clustered together leading to formation of zip code based cluster. Hence 65 zip code based cluster were formed. These zip code based cluster were then forwarded for processing in phase 2.

Phase 2: Categorization of zip code based clusters

To categorize zip code based clusters formed from New York region the six scenarios mentioned in Section 2 were considered. The results generated from both distance measures were compared with ground truth. Ground truth was simply calculated by representing each zip code based cluster with complaint type which is identified to be maximum among the six scenarios. Figure 4.12, 4.13, 4.14 represent categorization of location based clusters in six scenarios. These shape files plot were constructed using ArcMap tool. Table 4.10 shows accuracy results for two measures.

Table 4.10: Accuracy results on performing categorization

Measure used for categorization	Euclidean Distance	Cosine Similarity
Accuracy	69.23%	69.23%

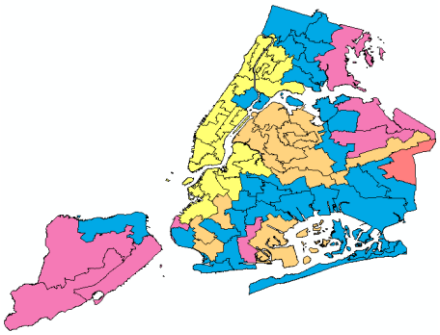


Figure 4.12: Shape files plot using Euclidean measure

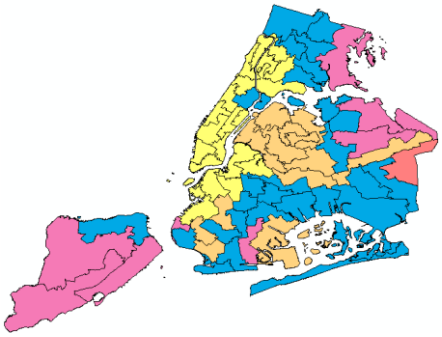


Figure 4.13: Shape files plot using Cosine Similarity

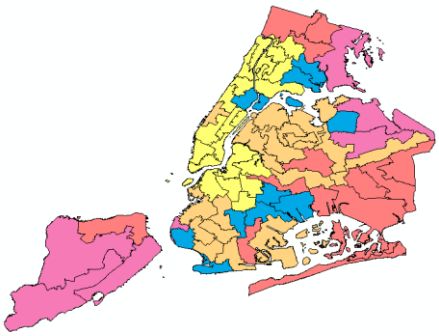
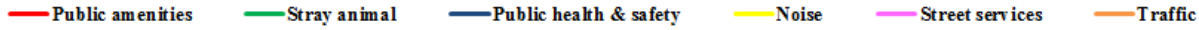


Figure 4.14: Shape file plot of Ground Truth



Phase 2: Profiling of complaint category for a zip code based cluster

The complaint categories critical in a particular zip code based cluster are identified in this block. 65 zip codes based clusters were formed after phase 1. Here again we identify clusters which display the highest severity in terms of criticality score for each complaint type.

Density or Resolution Score (*DRC*)

This factor as mentioned is identified in two ways. If complaints are highly concentrated, kernel density evaluation is used, else average resolution time period is computed and allotted weighted score according to the range. Density/Resolution Score values for each of the above zip code based cluster corresponding to each complaint type are listed in Table 4.11. The *DF* and *RTF* values are equal to 1 as we are only considering those zip code based clusters where a particular complaint type is found to be critical. Hence *DRC* also results to be 1.0.

Table 4.11: Density or resolution score for some clusters

Critical Complaint Type	Cluster Name	Complaint Type	Density/Resolution Score
Public Amenities Complaints	Cluster ZC_{15} ("Bushwick, Brooklyn")	Air Pollution	0.116327
		Construction	0.138988
		Electric	0.002483
		Fire Alarm	0
		Noise	0.038457
		Park	0.008018
		Parking	0.074979
		Public Amenities	**0.168729**
		Public Health	0.005403
		Stray Animal	0.08744
		Street	0.054316
		Streetlight	0.016242
		Traffic	0.011389
		Water	0.005283
		$DF_{public\_Amenities} = 1.0$	
Illegal Parking, Construction And Air Pollution Related Complaints	Cluster ZC_{16} ("Brownsville, Brooklyn")	**Air Pollution**	**0.157836**
		Construction	**0.130141**
		Electric	0.073923
		Public Health	0.052088
		Noise	0.029945
		Parking	**0.131155**
		$DF_{air\_Pollution} = DF_{construction} = DF_{parking} = 1.0$	

Temporal patterns of complaints (TF)

Table 4.12 displays some set of correlation values which increases criticality score.

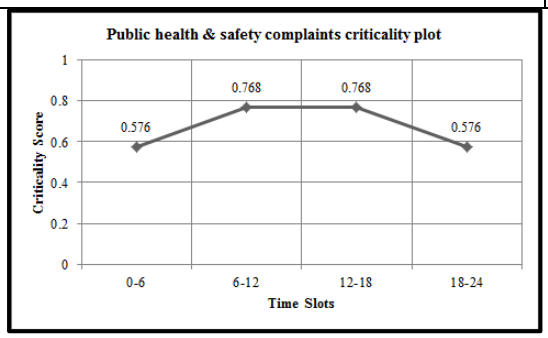
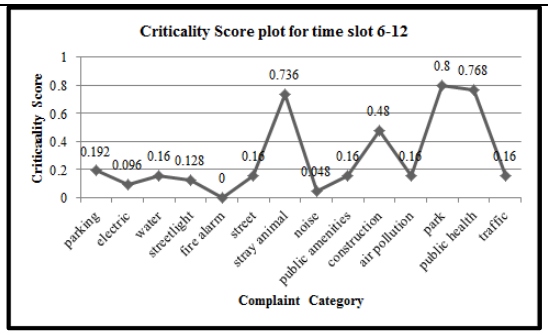
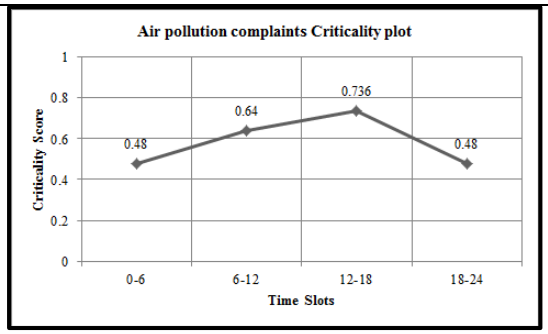
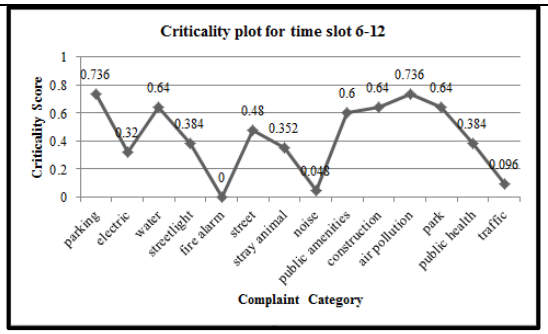
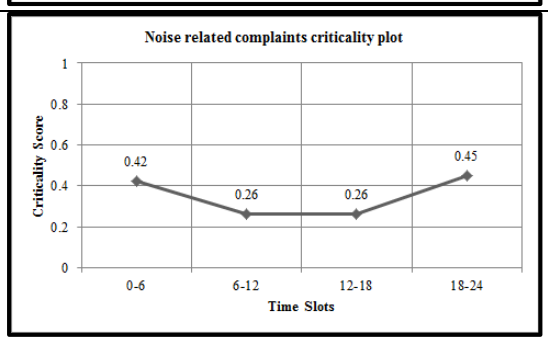
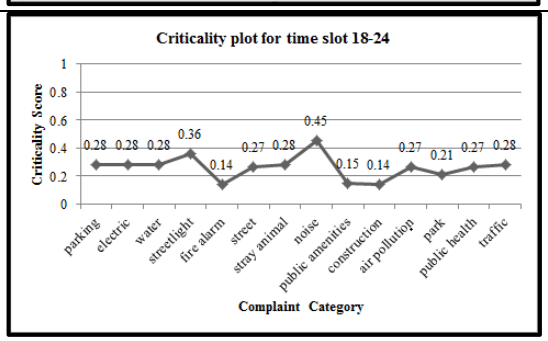
Table 4.12: Correlation values for some clusters

Cluster Name	Location	Correlated Complaints		Correlation Amount
Cluster ZC_{17}	"Bedford, Ny"	Park	Public Amenities	0.86
Cluster ZC_{63}	"Morris Park"	Water	Public Amenities	0.52
Cluster ZC_{42}	"Greenpoint"	Electricity	Public Amenities	0.51
Cluster ZC_{16}	"Brownsville"	Air Pollution	Public Health & Safety	0.59
Cluster ZC_{47}	"West Brighton"	Noise	Traffic	0.62
Cluster ZC_{19}	"Arden Avenue"	Air Pollution	Traffic	0.54
		Parking	Traffic	0.23
		Street Services	Traffic	0.85

Criticality Score Computation

Two plots are drawn corresponding to each cluster as shown in Table 4.13. First plot represents criticality score variation of one of the identified critical complaint for four time slots. Second plot represents criticality score in the cluster for each complaint category in a particular time slot.

Table 4.13: Criticality Score plots for some clusters

Cluster name	Criticality Score variation for four time slots	Criticality Score for all Complaint Categories for a selected time slot
Cluster ZC_{17} ("Bedford, NY")		
Cluster ZC_{16} ("Brownsville")		
Cluster ZC_{39} ("Times square, NY")		

Inferences:

It can be inferred from results acquired from second approach that cosine similarity categorizes the zip code based clusters with same accuracy as compared to Euclidean measure. The criticality score plots of different zip code based clusters helps in ranking the areas represented by them based on the severity of the complaints faced. The criticality plots help in inferring many observations. They are:

- Complaints related to traffic are high in time slots 6-12 and 18-24. It can be interpreted that it can be because of office hours.
- Complaints like noise is highly critical during 18-24 as sensitivity towards noise disturbances might increase during night hours. Noise is even critical during 0-6 slot.
- It can be inferred that noise and traffic might be correlated as they are highly critical. Even traffic and air pollution are also highly correlated.
- If we consider zip code based cluster (Cluster ZC_{39}) near Time Square, New York and analyse its criticality plot we find complaints related to noise, traffic are reported to be critical. Moreover other categories of complaints are also reported highlighting that it is an active region of the city where participatory sensing is performed.

4.2.2 Austin, USA

Data pre-processing block

The dataset comprises of the complaints collected for Austin city. The numbers of civic complaints recorded from May 2015 to March 2016 are 100k. Table 4.14 lists some results of data pre-processing block for Austin region.

Table 4.14: Results of Data pre-processing for Austin city

City	Austin
No. of complaints initially	1,00,000
No. of complaints after pre-processing	75,250

First Approach: Dynamic Grid Based Clustering Approach (DGCA)

Phase 1: Dynamic Grid Based Clustering

The ϵ epsilon value which gave the minimum cluster validity index was selected as shown in Figure 4.15. It can be derived from the plot that for Austin, epsilon value is shown as 10% of the total dataset. Table 4.15 enlists number of location based cluster formed based on optimal epsilon ϵ value.

Table 4.15: Number of location based cluster before merging step

City	Austin
No. of location based cluster formed based on optimal epsilon ϵ value	28

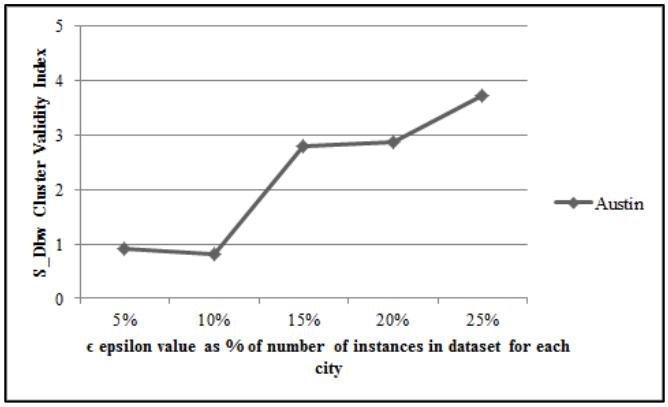


Figure 4.15: Plot of epsilon value v/s cluster validity index

Phase 1: Merging of Location based Clusters

Table 4.16 enlists the number of final location based clusters formed after merging and the reduced *S_Dbw* cluster validity index.

Table 4.16: Result after Merging of location based cluster

City	Austin
No. of location based cluster formed on merging	24
S_Dbw cluster validity index after merging	0.75

Phase 2: Categorization of location based clusters

24 location based clusters were formed. Figure 4.16, 4.17, 4.18 represent categorization of location based clusters in six scenarios each represented as bounding boxes. Table 4.17 below shows accuracy results on performing categorization using two different measures.

Table 4.17: Accuracy results on performing categorization

Measure used for categorization	Euclidean Distance	Cosine Similarity
Accuracy	95.23%	100%

Phase 2: Profiling of complaint category for a location based cluster

The criticality of the complaint categories in a particular location based cluster is identified in this block. 24 location based clusters were formed after phase 1.

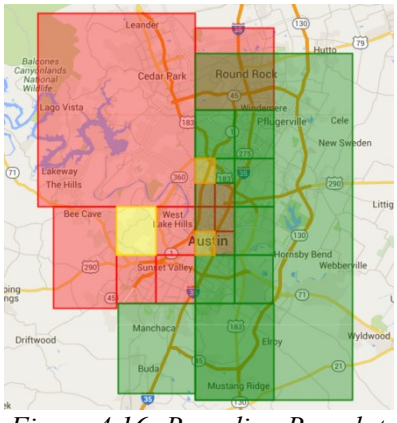


Figure 4.16: Bounding Box plot using Euclidean measure

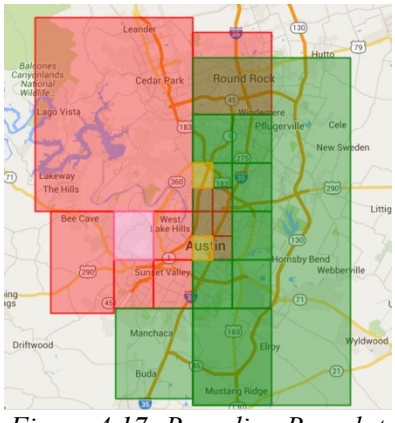


Figure 4.17: Bounding Box plot using Cosine Similarity

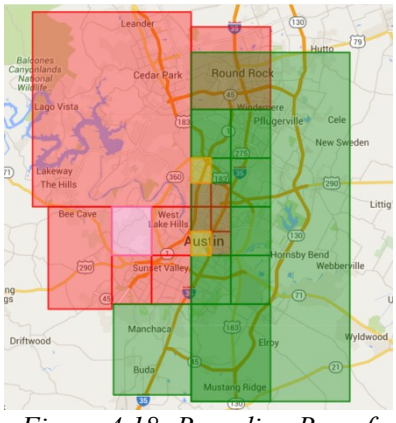
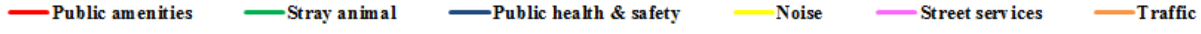


Figure 4.18: Bounding Box of Ground Truth



Density or Resolution Score (*DRC*)

Density and resolution score values for some location based cluster corresponding to the complaint type are listed in Table 4.18.

Table 4.18: Density or resolution score for some location based cluster

Critical Complaint Type	Cluster Name	Complaint Type	Density/Resolution Score
Park Related Complaints	Cluster LBC_{15} ("Guerrero Colorado Park")	**Park**	**0.641755**
		Public Amenities	0.026621
		Public Health	0.048562
		Stray Animal	0.005756
		Street	0.048013
		Streetlight	0.00848
		Traffic	0.016955
		$DF_{park} = 1.0$	
Public Amenities Complaints	Cluster LBC_9 ("Yarrabee Bend")	Park	0.08253
		Parking	0.138662
		Public Amenities	**0.22624**
		Public Health	0.146718
		Street	0.157132
		Streetlight	0.03356
		$DF_{public\ Amenities} = 1.0$	

Temporal patterns of complaints (TF)

Table 4.19 displays some set of correlation values which increases criticality score. The temporal pattern of co-occurrence is included in calculation in computation of criticality.

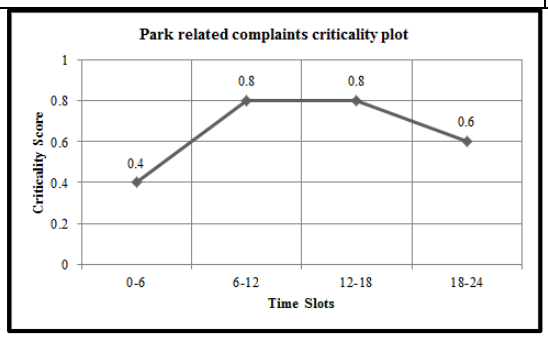
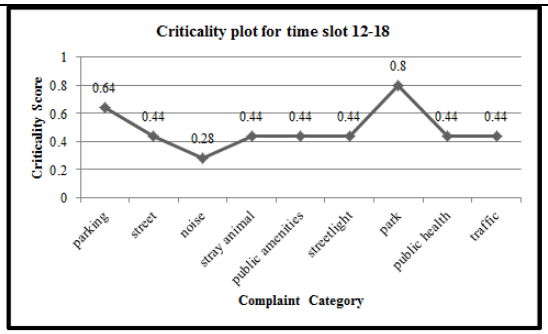
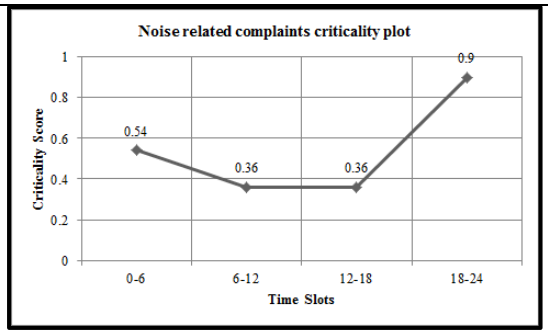
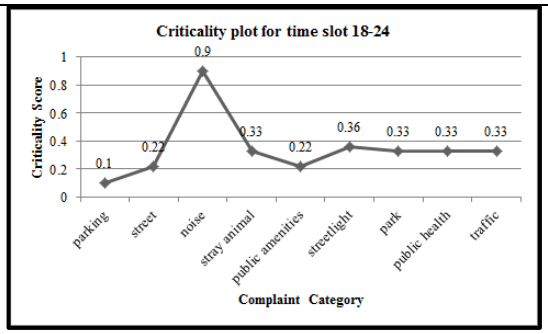
Table 4.19: Correlation values for some clusters

Cluster Name	Location	Correlated Complaint		Correlation Amount
Cluster LBC_{25}	"Howard"	Public Health & Safety	Stray Animal	0.87
Cluster LBC_{20}	"North Shoal Creek"	Parking	Traffic	0.89
		Street Services	Traffic	0.91

Criticality Score Computation

Criticality Score results for some clusters is shown in Table 4.20.

Table 4.20: Criticality Score plots for some clusters

Cluster name	Criticality Score variation for four time slots	Criticality Score for all Complaint Categories for a selected time slot
Cluster LBC_{15} ("Guerrero Colorado Park")		
Cluster LBC_{19} ("Northwest hills")		

Inferences:

It can be inferred from first approach results that cosine similarity categorizes the location based clusters more accurately as compared to Euclidean measure. The categorization shows that more complaints related to stray animals are faced in Austin. The observations inferred from criticality plots are:

- Complaints related to traffic are high in time slots 6-12, 12-18 and 18-24. It can be interpreted that it can be because of office hours.
- Complaints like traffic, illegal parking complaints and street services are highly correlated.
- Noise complaints are measured high during 0-6 and 18-24 time slots, reason can be that in these time slots people might become more sensitive towards disturbances due to noise.
- Stray animal and public health & safety complaints are highly correlated.

Second Approach: Clustering based on Zip codes approach (CZCA)

Phase 1: Clustering based on geographical boundaries representing zip codes block

To cluster civic complaints based on geographical boundaries representing zip codes shape file corresponding to Austin region is fetched. The shape file is than converted into SQL table consisting of polygon shape, area corresponding to polygon, coordinates of each polygon. Each polygon is formed by geographical boundary representing different zip code. 60 polygons were identified for Austin region. Complaints overlapping with each polygon were clustered together leading to formation of zip code based cluster. Hence 60 zip code based cluster were formed. These zip code based cluster were then forwarded for processing in phase 2.

4.2.3 Boston, USA

Data pre-processing block

The dataset comprises of the complaints collected for Boston city. The numbers of civic complaints recorded from January 2016 to March 2016 are 33,129. Table 4.21 lists some results of data pre-processing block for Boston region.

Table 4.21: Results of Data pre-processing for Boston city

City	Boston
No. of complaints initially	33,129
No. of complaints after pre-processing	27,158

First Approach: Dynamic Grid Based Clustering Approach (DGCA)

Phase 1: Dynamic Grid Based Clustering

The ϵ epsilon value which gave the minimum cluster validity index was selected as shown in

Figure 4.19. It can be derived from the plot that for Boston, epsilon value is shown as 10% of the total dataset. Table 4.22 enlists number of location based cluster formed based on optimal epsilon ϵ value.

Table 4.22: Number of location based cluster before merging step

City	Boston
No. of location based cluster formed based on optimal epsilon ϵ value	36

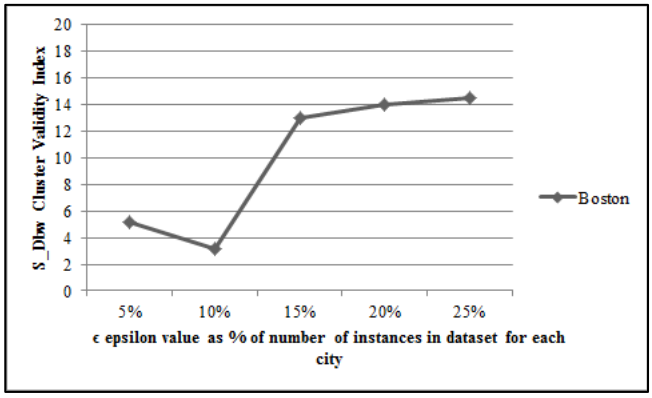


Figure 4.19: Plot of epsilon value v/s cluster validity index

Phase 1: Merging of Location based Clusters

Table 4.23 enlists the number of final location based clusters formed after merging and the reduced *S_Dbw* cluster validity index.

Table 4.23: Result after Merging of location based cluster

City	Boston
No. of location based cluster formed on merging	17
S_Dbw cluster validity index after merging	0.17

Phase 2: Categorization of location based clusters

Categorization of 24 location based clusters was done. Figure 4.20, 4.21, 4.22 represent categorization of location based clusters in six scenarios each represented as bounding boxes. Table 4.24 below shows accuracy results on performing categorization using two different measures.

Table 4.24: Accuracy results on performing categorization

Measure used for categorization	Euclidean Distance	Cosine Similarity
Accuracy	64.71%	82.35%

Phase 2: Profiling of complaint category for a location based cluster

The criticality of the complaint categories in a particular location based cluster is identified in this block. 24 location based clusters were formed after phase 1.

Density or Resolution Score (*DRC*)

Density and resolution score values for some location based cluster corresponding to the complaint type are listed in Table 4.25.

Temporal patterns of complaints (TF)

Table 4.26 displays some set of correlation values which increases criticality score.

Table 4.25: Density or resolution score for some location based clusters

Critical Complaint Type	Cluster Name	Complaint Type	Density/Resolution Score
Illegal parking complaints	Cluster LBC_2 ("West Roxbury")	Park	0.041187
		Parking	**0.45678**
		Public Amenities	0.105971
		Public Health	0.122739
		Snow	0.120645
		Stray Animal	0.250153
		Street	0.069132
		Streetlight	0.00103
		Traffic	0.028939
		$DF_{parking} = 1.0$	
Noise related complaints	Cluster LBC_{37} ("Eagle hill")	**Noise**	**0.883012**
		Park	0.069974
		Parking	0.092797
		Public Amenities	0.117632
		Public Health	0.123571
		Snow	0.077845
		Stray Animal	0.106578
		Street	0.122504
		$DF_{noise} = 1.0$, $DF_{traffic} = 1.0$	

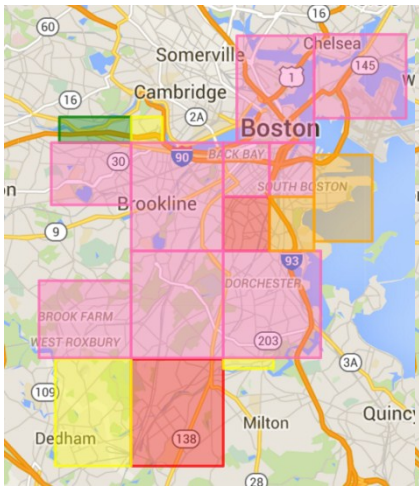


Figure 4.20: Bounding Box plot using Euclidean measure

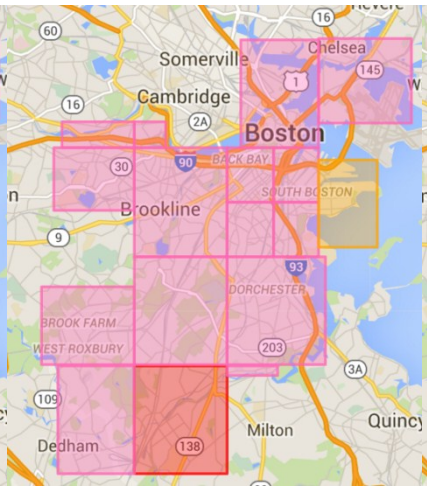


Figure 4.21: Bounding Box plot using Cosine Similarity

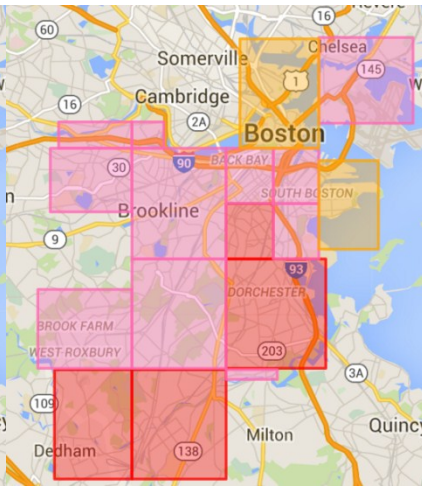


Figure 4.22: Bounding Box plot of Ground Truth

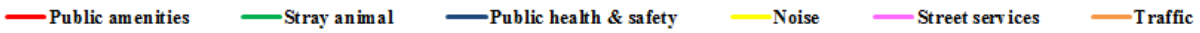


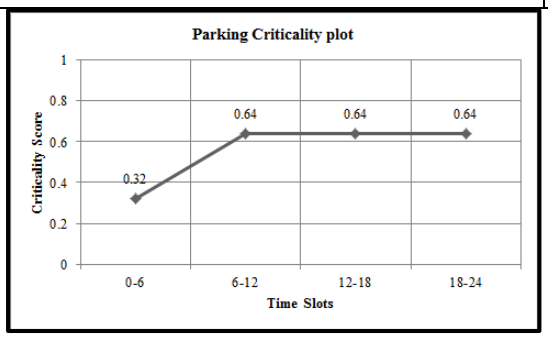
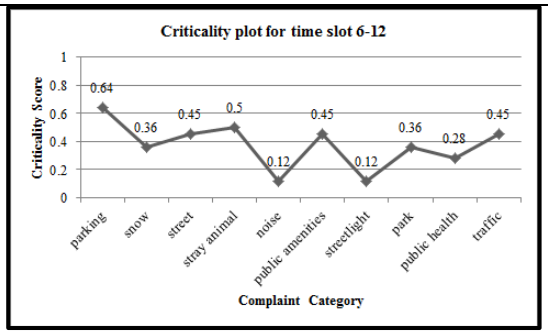
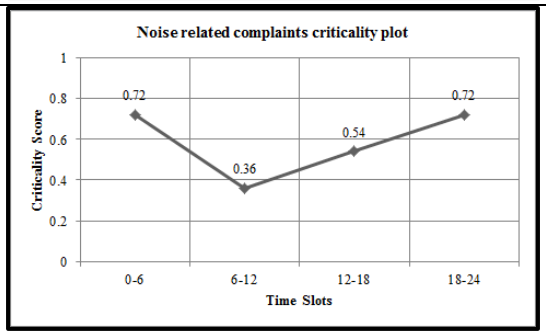
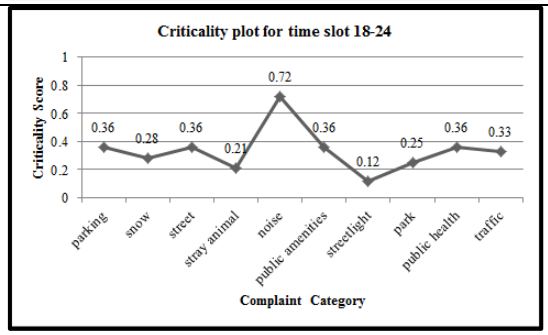
Table 4.26: Correlation values for some clusters

Cluster Name	Location	Correlated Complaint		Correlation amount
Cluster LBC_2	"West Roxbury"	Parking	Traffic	0.776
		Parking	Snow	0.84
		Parking	Public Amenities	0.90
Cluster LBC_7	"Mission hill"	Public Amenities	Park	0.98
		Public Amenities	Public Health & Safety	0.93
		Public Amenities	Street	0.91
Cluster LBC_{37}	"Eagle hill"	Noise	Park	0.71
		Noise	Traffic	0.91

Criticality Score Computation

Criticality Score results for some clusters is shown in Table 4.27.

Table 4.27: Criticality Score plots for some clusters

Cluster name	Criticality Score variation for four time slots	Criticality Score for all Complaint Categories for a selected time slot
Cluster LBC_2 ("West Roxbury")		
Cluster LBC_{37} ("Eagle hill")		

Inferences:

It can be inferred from first approach results that cosine similarity categorizes the location based clusters more accurately as compared to Euclidean measure. Majority of regions in Boston report complaints related to street services. The observations inferred from criticality plots are:

- Complaints related to illegal parking are reported high in time slots 6-12, 12-18 and 18-24.
- Noise related complaints are critical during 0-6 and 18-24 time slots.
- Complaint related to snow, traffic and public amenities are highly correlated with illegal parking complaints.
- Complaints related to public health & safety, park and street services are highly correlated with public amenities complaints.

4.2.4 Chicago, USA

Data pre-processing block

The dataset comprises of the complaints collected for Chicago city. The numbers of civic complaints recorded from January 2016 to March 2016 are 64,627. Table 4.28 lists some results of data pre-processing block for Chicago region.

Table 4.28: Results of Data pre-processing for Chicago city

City	Chicago
No. of complaints initially	64,627
No. of complaints after pre-processing	64,542

First Approach: Dynamic Grid Based Clustering Approach (DGCA)

Phase 1: Dynamic Grid Based Clustering

The ϵ epsilon value which gave the minimum cluster validity index was selected as shown in Figure 4.23. It can be derived from the plot that for Chicago, epsilon value is shown as 10% of the total dataset. Table 4.29 enlists number of location based cluster formed based on optimal epsilon ϵ value.

Table 4.29: Number of location based cluster before merging step

City	Chicago
No. of location based cluster formed based on optimal epsilon ϵ value	32

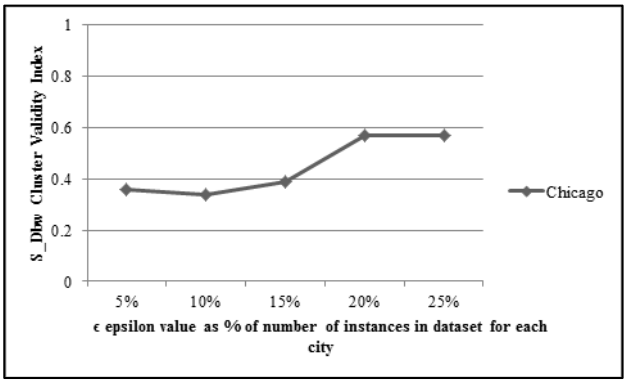


Figure 4.23: Plot of epsilon value v/s cluster validity index

Phase 1: Merging of Location based Clusters

Table 4.30 enlists the number of final location based clusters formed after merging and the reduced *S_Dbw* cluster validity index.

Table 4.30: Result after Merging of location based cluster

City	Chicago
No. of location based cluster formed on merging	25
S_Dbw cluster validity index after merging	0.30

Phase 2: Categorization of location based clusters

25 location based clusters were formed for Chicago region. Figure 4.24, 4.25, 4.26 represent categorization of location based clusters in six scenarios. Table 4.31 below shows accuracy results on performing categorization using two different measures.

Table 4.31: Accuracy results on performing categorization

Measure used for categorization	Euclidean Distance	Cosine Similarity
Accuracy	68%	100%

Phase 2: Profiling of complaint category for a location based cluster

To identify the criticality of the complaint categories in a particular location based cluster is

performed in this block. 25 location based clusters were formed after phase 1.

Density or Resolution Score (*DRC*)

Density and resolution score values for some location based cluster corresponding to the complaint type are listed in Table 4.32. The complaint dataset of Chicago consisted date of complaint only and the time stamp is not available, so analysis with typical time slot and co-occurrence was not possible.

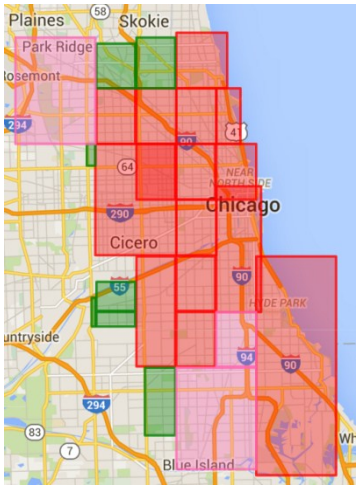


Figure 4.24: Bounding Box plot using Euclidean measure

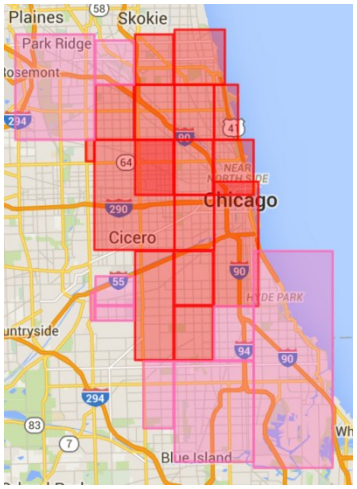


Figure 4.25: Bounding Box plot using Cosine Similarity

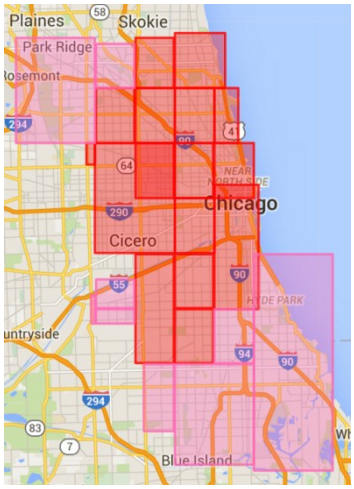


Figure 4.26: Bounding Box plot of Ground Truth

Public amenities | Stray animal | Public health & safety | Noise | Street services | Traffic

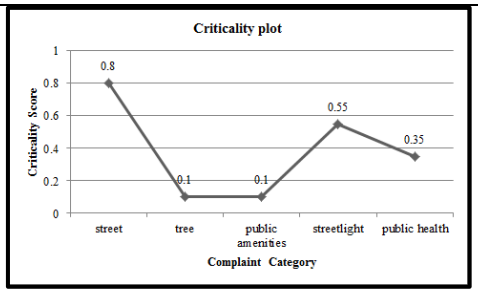
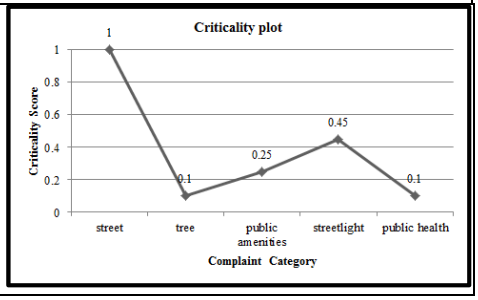
Table 4.32: Density or resolution score for some location based clusters

Critical Complaint Type	Cluster Name	Complaint Type	Density/Resolution Score
Street Services complaints	Cluster LBC_{20} ("Gage park")	Public Amenities	0.013199
		Public Health	0.001917
		Street	**0.028141**
		Streetlight	0.020818
		Tree	0.003157
		$DF_{street} = 1.0$	
Street light complaints	Cluster LBC_{14} ("Kennedy expressway")	Public Amenities	0.017495
		Public Health	0.036891
		Street	0.028897
		Streetlight	**0.061683**
		Tree	0.017316
		$DF_{street\_Light} = 1.0$	

Criticality Score Computation

Criticality Score results for some clusters is shown in Table 4.33.

Table 4.33: Criticality Score plots for some clusters

Cluster name	Criticality Score for all Complaint Categories for a selected time slot	Cluster name	Criticality Score for all Complaint Categories for a selected time slot
Cluster LBC_{14} ("Kennedy express way")		Cluster LBC_{20} ("Gage park")	

Inferences:

It can be inferred from first approach results that cosine similarity categorizes the location based clusters more accurately as compared to Euclidean measure. Majority of regions in Chicago reported complaints related to public amenities. The observations from criticality plots are:

- Complaints related to street services and streetlights are critical at the same time.
- Complaints related to public health & safety and park, tree related complaints are critical at the same time.

4.2.5 San Francisco, USA

Data pre-processing block

The dataset comprises of civic complaints collected for San Francisco city. The numbers of civic complaints recorded from December 2015 to March 2016 are 98,877. Table 4.34 lists results of data pre-processing for San Francisco region.

Table 4.34: Results of Data pre-processing for San Francisco city

City	San Francisco
No. of complaints initially	98,877
No. of complaints after pre-processing	78,333

First Approach: Dynamic Grid Based Clustering Approach (DGCA)

Phase 1: Dynamic Grid Based Clustering

The ϵ epsilon value which gave the minimum cluster validity index was selected as shown in Figure 4.27. It can be derived from the plot that for San Francisco, epsilon value is shown as 10% of the total dataset. Table 4.35 enlists number of location based cluster formed based on optimal epsilon ϵ value.

Table 4.35: Number of location based cluster before merging step

City	San Francisco
No. of location based cluster formed based on optimal epsilon ϵ value	31

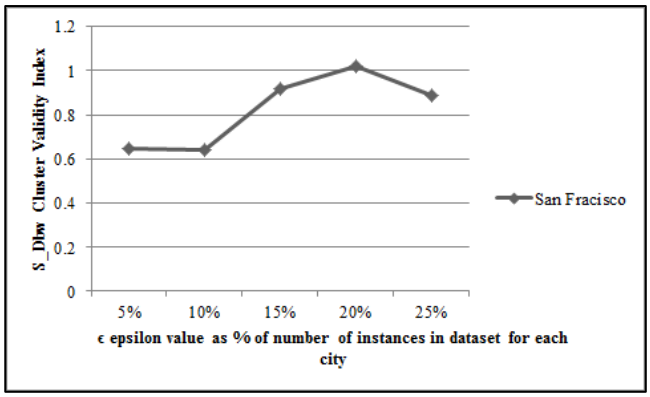


Figure 4.27: Plot of epsilon value v/s cluster validity index

Phase 1: Merging of Location based Clusters

Table 4.36 enlists the number of final location based clusters formed after merging and the reduced *S_Dbw* cluster validity index.

Table 4.36: Result after Merging of location based cluster

City	San Francisco
No. of location based cluster formed on merging	22
S_Dbw cluster validity index after merging	0.108

Phase 2: Categorization of location based clusters

Figure 4.28, 4.29, 4.30 represent categorization of location based clusters in six scenarios each represented as bounding boxes for San Francisco region. Table 4.37 below shows accuracy results on performing categorization using two different measures.

Table 4.37: Accuracy results on performing categorization

Measure used for categorization	Euclidean Distance	Cosine Similarity
Accuracy	72.72%	100%

Phase 2: Profiling of complaint category for a location based cluster

To identify the criticality of the complaint categories in a particular location based cluster is performed in this block. 22 location based clusters were formed after phase 1.

Density or Resolution Score (*DRC*)

Density and resolution score values for some location based cluster corresponding to the complaint type are listed in Table 4.38.

Temporal patterns of complaints (TF)

Table 4.39 displays some set of correlation values which increases criticality score.

Table 4.38: Density or resolution score for some location based clusters

Critical Complaint Type	Cluster Name	Complaint Type	Density/Resolution Score
Noise related complaints	Cluster LBC_{22} ("Pacific heights")	**Noise**	**0.378407**
		Park	0.009593
		Public Amenities	0.008948
		Public Health	0.026052
		Stray Animal	0
		Street	0.007831
		Streetlight	0.016115
		Traffic	0.06475
		$DF_{noise} = 1.0$	
Traffic complaints	Cluster LBC_3 ("Mission terrace")	Noise	0.0149381
		Park	0.0079561
		Public Amenities	0.0079063
		Public Health	0.0225401
		Stray Animal	0
		Street	0.0131514
		Streetlight	0.0284612
		Traffic	**0.0412983**
		$DF_{traffic} = 1.0$	

Table 4.39: Correlation values for some clusters

Cluster Name	Location	Correlated Complaints		Correlation value
Cluster LBC_{18}	"Chinatown"	Park	Public Amenities	0.94
Cluster LBC_{10}	"The Castro"	Public Amenities	Public Health & Safety	0.93
		Public Amenities	Street Service	0.99
Cluster LBC_{22}	"Pacific Heights"	Noise	Traffic	0.89
		Noise	Park	0.50

Cluster LBC_3	"Mission Terrace"	Street Services Streetlight	Traffic Traffic	0.97 0.27

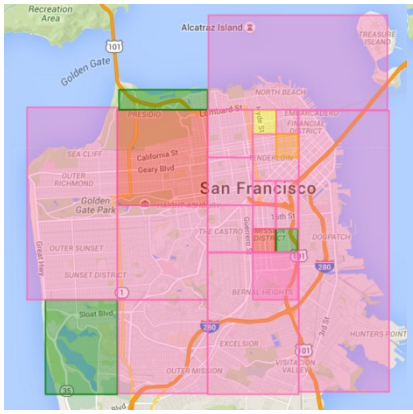


Figure 4.28: Bounding Box plot using Euclidean measure

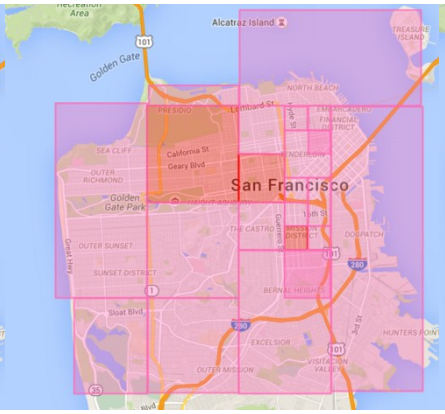


Figure 4.29: Bounding Box plot using Cosine Similarity

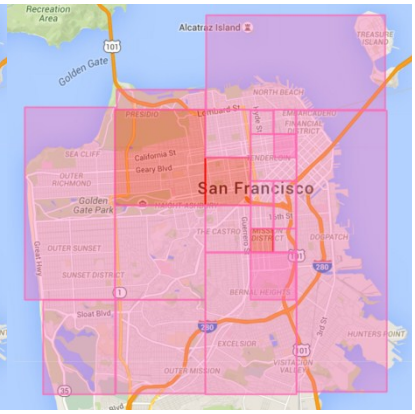


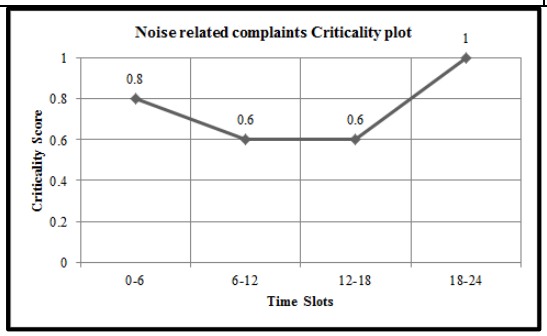
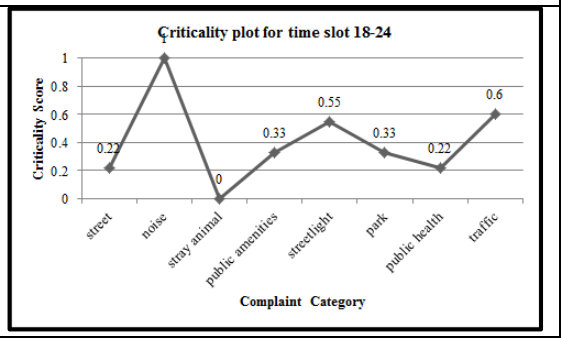
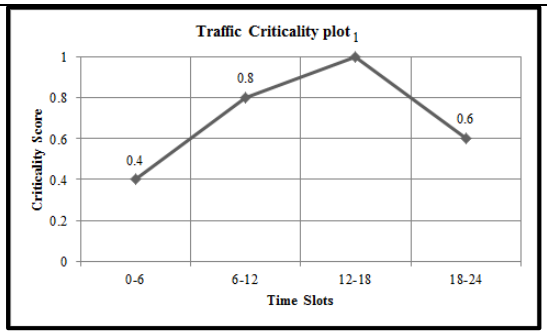
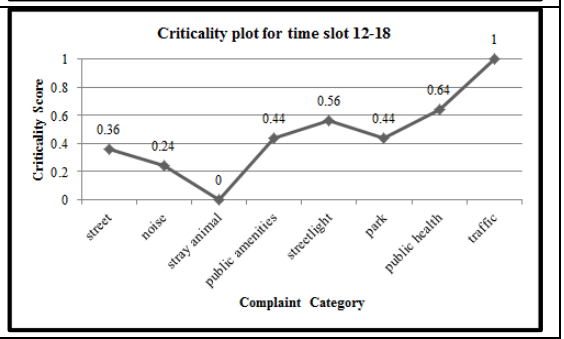
Figure 4.30: Bounding Box plot of Ground Truth

Public amenities | Stray animal | Public health & safety | Noise | Street services | Traffic

Criticality Score Computation

Criticality Score results for some clusters is shown in Table 4.40.

Table 4.40: Criticality Score plots for some clusters

Cluster name	Criticality Score variation for four time slots	Criticality Score for all Complaint Categories for a selected time slot
Cluster LBC_{22} ("Pacific Heights")		
Cluster LBC_3 ("Mission terrace")		

Inferences:

It can be inferred from first approach results that cosine similarity categorizes the location based clusters more accurately as compared to Euclidean measure. Majority of regions in San Francisco post complaints related to inefficient street services. The observations made from criticality plot are:

- Complaints related to public amenities, park co-occur.
- Complaints related to noise co-occur with traffic related complaints.

- Noise related complaints report to be critical in 0-6 and 18-24 time slots.
- Public amenities are measured critical along with public health & safety and inefficient street services complaints.

4.3 Indian Metropolitan City: Bangalore, India

Data pre-processing block

The dataset comprises of the complaints collected for Bangalore city. The numbers of civic complaints recorded from July 2015 to December 2015 are 25,986. The data was de-noised by discarding records with invalid location and blank location attribute as done in case of New York dataset. Hence data pre-processing block has same procedure as opted for New York dataset. Table 4.41 lists some results of data pre-processing block for Bangalore region.

Table 4.41: Results of Data pre-processing for Bangalore city

City	Bangalore
No. of complaints initially	25,986
No. of complaints after pre-processing	18,712

First Approach: Dynamic Grid Based Clustering Approach (DGCA)

Phase 1: Dynamic Grid Based Clustering

To select the optimal ϵ epsilon value cluster validity index for all location based clusters formed by each division was computed and the value which gives the minimum value was selected as shown in Figure 4.31. It can be derived from the plot that for Bangalore epsilon value is shown as 10% of the total dataset. The set of location based clusters selected corresponding to optimal epsilon value is forwarded for next step of the process i.e. merging of location based clusters. Table 4.42 enlists number of location based cluster formed based on optimal epsilon ϵ value.

Table 4.42: Number of location based cluster before merging step

City	Bangalore
No. of location based cluster formed based on optimal epsilon ϵ value	24

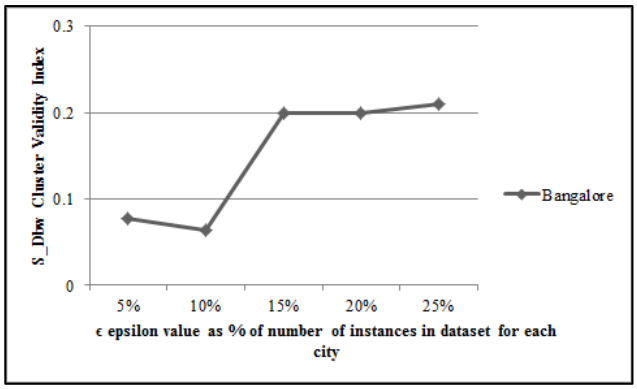


Figure 4.31: Plot of epsilon value v/s cluster validity index

Phase 1: Merging of Location based Clusters

In merging of location based clusters we calculate density for each location based cluster. Table 4.43 enlists the number of final location based clusters formed after merging and the reduced S_Dbw cluster validity index.

Table 4.43: Result after Merging of location based cluster

City	Bangalore
No. of location based cluster formed on merging	21
S_Dbw cluster validity index after merging	0.05

Phase 2: Categorization of location based clusters

To categorize location based clusters formed from Bangalore region the six scenarios mentioned in Section 2 were only considered. 21 location based clusters were formed. The results generated from both distance measures were compared with ground truth. Ground truth was simply calculated by representing each location based cluster with complaint type which is identified to be maximum among the six scenarios. Figure 4.32, 4.33, 4.34 represent categorization of location based clusters in six scenarios each represented as bounding boxes. Table 4.44 below shows accuracy results on performing categorization using two different measures.

Table 4.44: Accuracy results on performing categorization

Measure used for categorization	Euclidean Distance	Cosine Similarity
Accuracy	76.19%	90.47%

Phase 2: Profiling of complaint category for a location based cluster

The criticality of the complaint categories in a particular location based cluster is identified in this block. 21 location based clusters were formed after phase 1.

Density or Resolution Score (*DRC*)

Density and resolution score values for some location based cluster are listed in Table 4.60.

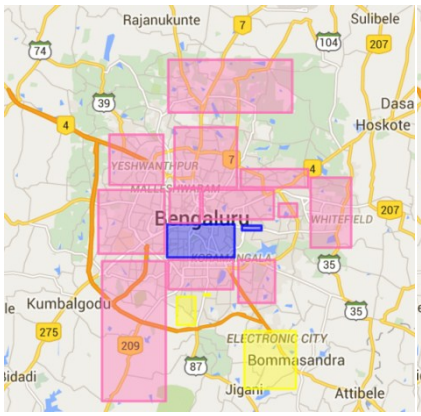


Figure 4.32: Bounding Box plot using Euclidean measure

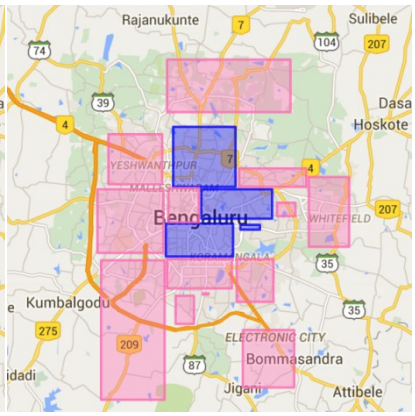


Figure 4.33: Bounding Box plot using Cosine Similarity

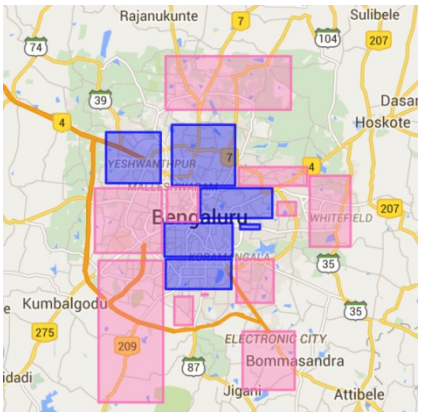


Figure 4.34: Bounding Box plot of Ground Truth

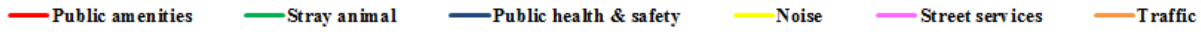


Table 4.45: Density or resolution score for some location based clusters

Critical Complaint Type	Cluster Name	Complaint Type	Density/Resolution Score
Potholes, Damaged roads complaints	Cluster LBC_8 ("Bannerghatta Road")	Electricity problem	0
		Flood road	0.841982
		Garbage epidemic disease	2.002947
		Heritage sites	0
		Noise pollution	0
		Parking	0.030173
		Pothole damage roads	**10.17664**
		Property and land related	0.255345
		Public amenities and sanitation	0.255345
		Sewage drain	1.43
		Stray animal	1.42
		Streetlights	0.5
		Traffic	0.15
		$DF_{pothole} = 1.0$	
Garbage, epidemic disease complaints	Cluster LBC_{18} ("Banaswadi Road")	Air pollution	0
		Defacement problem	0.112593
		Dengue mosquito	0.030976
		Electricity problem	0.146452
		Flood road	0.096504
		Garbage epidemic disease	**0.151535**
		Heritage sites	0
		Noise pollution	0.001039
		Parking	0.034649
		Pothole damage roads	**0.109935**
		Property and land related	0.054236
		Public amenities and sanitation	0.059379
		Public transport bus shelter	0.051053
		Sewage drain	0.053262
		Storm water drains	0.093002
		Stray animal	0.061047
		Streetlights	0.075573
		Traffic	0.04675
		Tree parks and playground	0.035061
		Unsafe region	0.117723
		Water pollution	0.024723
		Water related problem	0.062212

Temporal patterns of complaints (TF)

Table 4.46 displays some set of correlation values which increases criticality score. The temporal pattern of co-occurrence is included in computation of criticality.

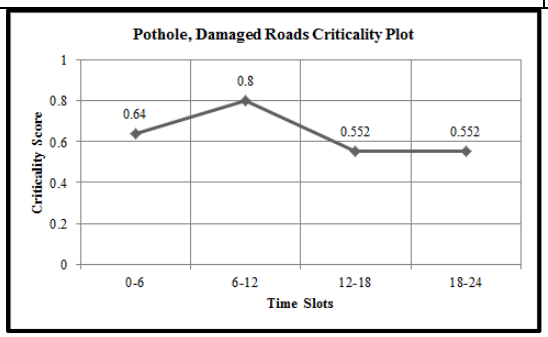
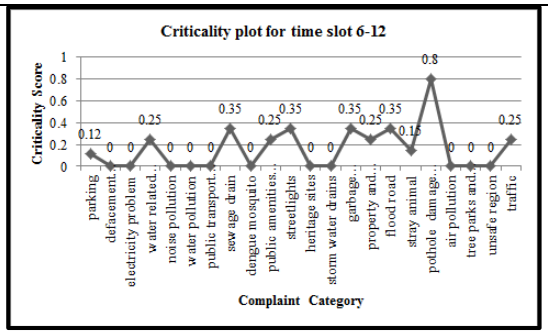
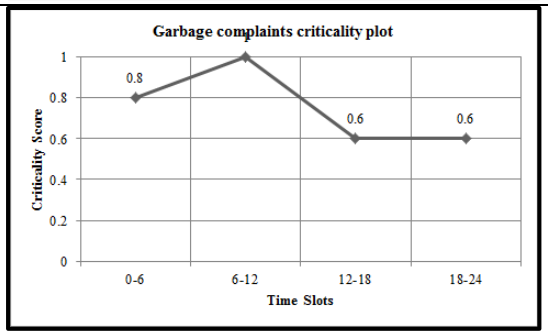
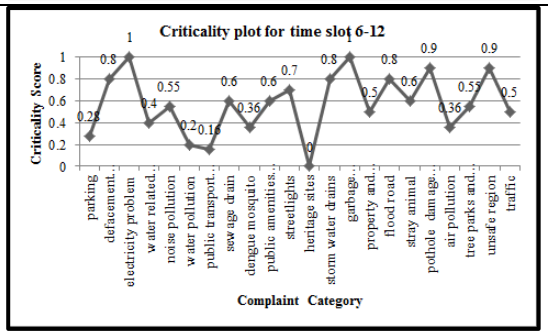
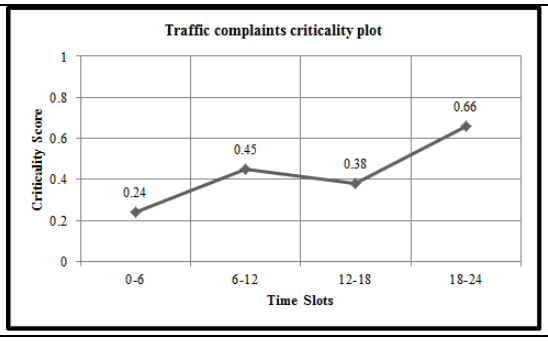
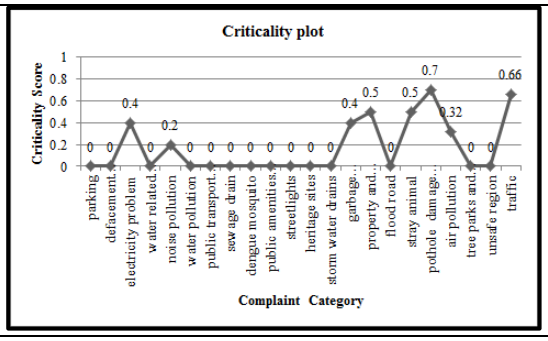
Table 4.46: Correlation values for some clusters

Cluster Name	Location	Correlated Complaint		Correlation amount
Cluster LBC_{18}	"Banaswadi Road"	Dengue Mosquito	Garbage Epidemic Disease	0.76
		Garbage Epidemic Disease	Public Amenities & Sanitation	0.96
		Garbage	Epidemic Disease Sewage Drain	0.95
Cluster LBC_8	"Bannerghatta Road"	Air Pollution	Pothole, Damaged Roads	0.88
		Flood Road	Pothole, Damaged Roads	0.86
		Pothole, Damaged Roads	Traffic	0.20

Criticality Score Computation

Criticality Score results for some clusters is shown in Table 4.47.

Table 4.47: Criticality Score plots for some clusters

Cluster name	Criticality Score variation for four time slots	Criticality Score for all Complaint Categories for a selected time slot
Cluster LBC_8 ("Bannerghatta Road")		
Cluster LBC_{18} ("Banaswadi Road")		
Cluster LBC_{16} ("Marathalli")		

Inferences:

It can be inferred from first approach results that cosine similarity categorizes the location based clusters more accurately as compared to Euclidean measure. The criticality score plots of different location based clusters helps in ranking the areas represented by location based clusters based on the severity of the complaints faced. The criticality plots helps in inferring many observations. They are:

- Complaints related to traffic are high in time slots 6-12 and 18-24. It can be interpreted that it can be because of office hours.
- Complaints like garbage, public amenities and sanitation, dengue mosquito are highly correlated.
- It can be inferred that noise and traffic might be correlated as they are highly critical.
- If we consider location based cluster (Cluster LBC_{16}) near Bangalore, India and analyse its criticality plot we find complaints related to noise, air pollution, traffic are reported to be critical.

Second Proposed Approach: Clustering based on Zip codes approach (CZCA)

Phase 1: Clustering based on geographical boundaries representing zip codes block

To cluster civic complaints based on geographical boundaries representing zip codes shape file corresponding to Bangalore region is fetched. The shape file is than converted into SQL table consisting of polygon shape, area corresponding to polygon, coordinates of each polygon. Each polygon is formed by geographical boundary representing different zip code. 136 polygons were identified for Bangalore region. Complaints overlapping with each polygon were clustered together leading to formation of zip code based cluster. Hence 136 zip code based cluster were formed. These zip code based cluster were then forwarded for processing in phase 2.

Phase 2: Categorization of zip code based clusters

To categorize zip code based clusters formed from Bangalore region the six scenarios mentioned in Section 2 were considered. The results generated from both distance measures were compared with ground truth. Ground truth was simply calculated by representing each zip code based cluster with complaint type which is identified to be maximum among the six scenarios. Figure 4.35, 4.36, 4.37 represent categorization of location based clusters in six scenarios. Table 4.48 shows accuracy results for two measures.

Table 4.48: Accuracy results on performing categorization

Measure used for categorization	Euclidean Distance	Cosine Similarity
Accuracy	46.32%	72.79%

Phase 2: Profiling of complaint category for a zip code based cluster

Zip code based clustering is performed in this block to identify the criticality of complaint category. The results for next steps are displayed for few clusters only.

Density or Resolution Score (*DRC*)

This factor as mentioned is identified in two ways. If complaints are highly concentrated, kernel density is evaluated, else average resolution time period is computed and allotted weighted score according to the range. Density/Resolution Score values are listed for each of the above zip code based cluster corresponding to each complaint type is given in Table 4.49.

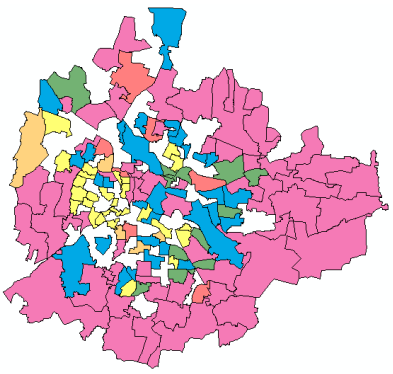


Figure 4.35: Shape file plot using Euclidean measure

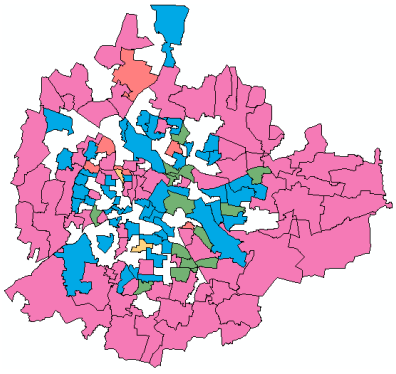


Figure 4.36: Shape file plot using Cosine Similarity

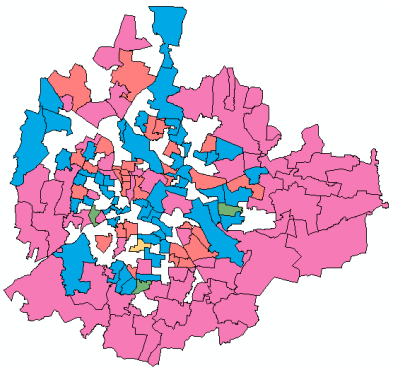


Figure 4.37: Shape file plot of Ground Truth

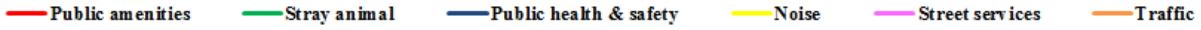


Table 4.49: Density or resolution score for some location based clusters

Critical Complaint Type	Cluster Name	Complaint Type	Density/Resolution Score
Air pollution complaints, Noise related complaints	Cluster ZC_{117} ("Vijayanagar")	$RTF_{air\_pollution}$ = *1.0* Average resolution time period =0	
Dengue mosquito complaints	Cluster ZC_{17} ("Kodigehalli")	$RTF_{dengue = 1.0}$ Average resolution time period= 0	
Tree, parks related complaints	Cluster ZC_{175} ("Bilekahalli")	Air Pollution	3.222909
		Dengue Mosquito	0.0182102
		Property And Land Related	0.0021322
		Public Amenities And Sanitation	0.9492472
		Public Transport Bus Shelter	0.5081652
		Sewage Drain	0.1731325
		Storm Water Drains	1.0255819
		Tree Parks	**1.86**
		DF_{tree}= *1.0*	

Temporal patterns of complaints (TF)

Table 4.50 displays some set of correlation values which increases criticality score. So this temporal pattern of co-occurrence is included in computation of criticality.

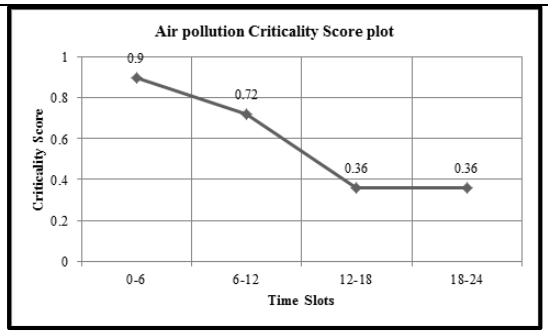
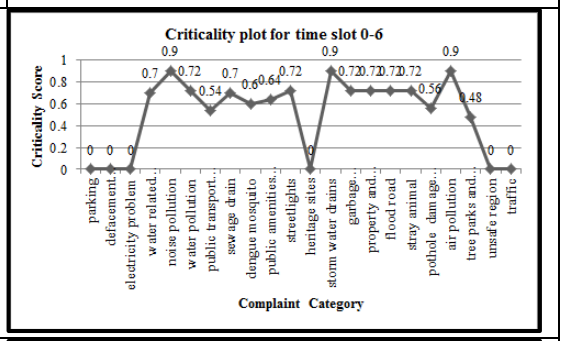
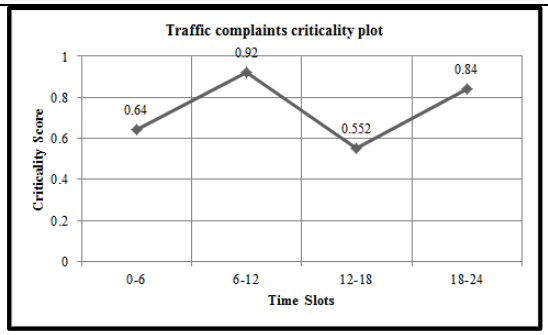
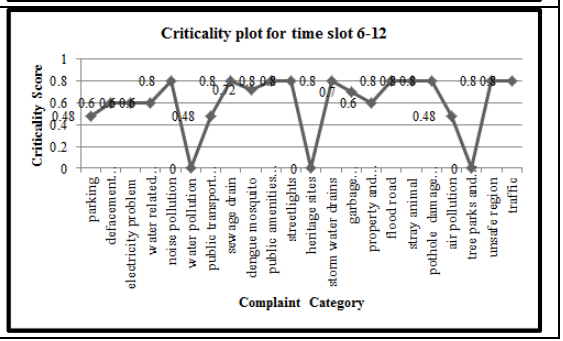
Table 4.50: Correlation values for some clusters

Cluster Name	Location	Correlated Complaint		Correlation amount
Cluster ZC_{17}	"Kodigehalli"	Dengue Mosquito	Flood Road	0.58
		Garbage	Dengue Mosquito	0.96
		Dengue Mosquito	Public Amenities	0.69
		Dengue Mosquito	Sewage Drain	0.84
Cluster ZC_{45}	"Rajamahal Guttahalli"	Pothole	Air Pollution	0.76
		Pothole	Public Amenities	0.90
Cluster ZC_{109}	"Rajaji Nagar"	Flood Road	Traffic	0.78
		Pothole	Traffic	0.27
		Public Amenities	Traffic	0.81
		Street Light	Traffic	0.94
Cluster ZC_{138}	"Siddapura"	Flood Road	Sewage Drain	0.901
Cluster ZC_{117}	"Vijayanagar"	Air Pollution	Water Pollution	0.57
		Flood Road	Water Pollution	0.57
		Garbage	Water Pollution	0.87
		Public Amenities	Water Pollution	0.58
		Sewage Drain	Water Pollution	0.09
		Storm Water	Water Pollution	0.58

Criticality Score Computation

Criticality Score results for some clusters is shown in Table 4.51.

Table 4.51: Criticality Score plots for some clusters

Cluster name	Criticality Score variation for four time slots	Criticality Score for all Complaint Categories for a selected time slot
Cluster ZC_{117} ("Vijaynagar")		
Cluster ZC_{197} ("Marathalli")		

Inferences:

It can be inferred from second approach results that cosine similarity categorizes the zip code based clusters with more accuracy as compared to Euclidean measure. The criticality score plots of different zip code based clusters helps in ranking the areas represented by them based on the severity of the complaints faced. The criticality plots helps in inferring many observations. They are:

- Complaints related to traffic are high in time slots 6-12 and 18-24. It can be interpreted that it can be because of office hours.
- Complaints related to potholes, streetlight and public amenities are correlated with traffic complaints.
- Complaints related flood road, public amenities, sewage drain and garbage are correlated with complaints related to dengue mosquito reporting unhealthy conditions.
- Cluster ZC_{197} represent region of Marathalli, Bangalore which is crowded area and considered one of the traffic prone area. It can be observed that the cluster reports traffic complaint, air pollution complaint, noise related complaints as critical.
- Cluster ZC_{117} marks air pollution as a critical complaint. It is observed that it covers the region of Vijaynagar, Bangalore where Hi-tech Rubber industry and many other industries are located.

4.4 Comparison with ground truth

4.4.1 Comparison of Phase 1 of DGCA with ground truth

As the data was already tagged consisting of details like broad location attribute, county/borough details i.e. geographical regions, community board. So DGCA results were compared with these tagged data and gave 78.83% accuracy. The classical grid based clustering GRIDCLUS gave 80.92% accuracy on an average. Hence the results are comparable.

(a) *(b)*

Figure 4.38 Spatial Clustering plots using (a) DGCA (b) GRIDCLUS

4.4.2 Comparison of Phase 2 with real world ground truth

1. New York region

- "Bushwick, Brooklyn" (Cluster ZC_{15}) suffered complaints related to public amenities.
- "Brownsville, NY" reported complaints related to illegal parking (Cluster ZC_{16})


The New York Times

N.Y. / Region

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

NEIGHBORHOOD REPORT: BUSHWICK

NEIGHBORHOOD REPORT: BUSHWICK; Bats, Cats and Boomboxes To Quench a Modern Plague

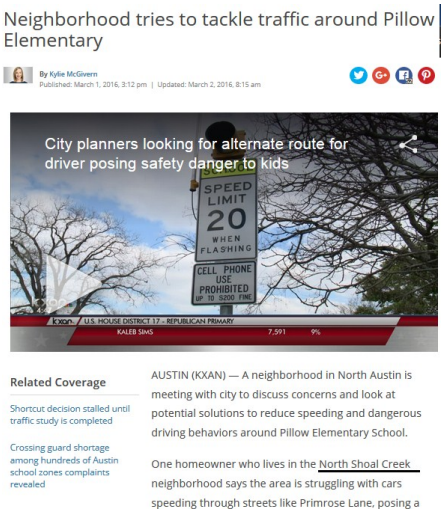
By TARA BAHRAMPOUR
Published: June 25, 2000

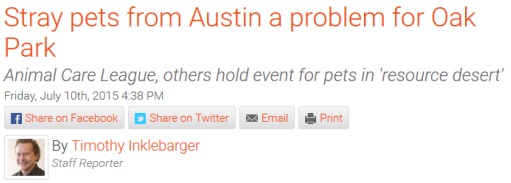
As animal-rights advocates protested recently in Manhattan against the killing of rats on the CBS show "Survivor," Bushwick residents were filming their own rat movie.

FACEBOOK
TWITTER
GOOGLE+

2. Austin, USA

- Stray animal issue in Austin and traffic issue near "North Shoal Creek" (ClusterLBC_{20})



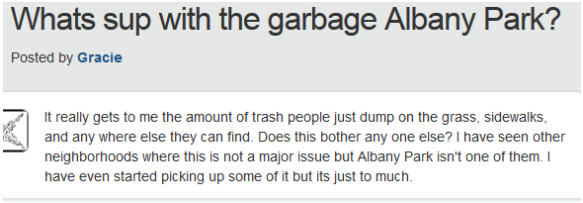


3. Boston

- "Eagle Hill" (ClusterLBC$_{37}$) suffers from noise related complaints.
- "West Roxbury" (ClusterLBC$_{2}$) suffer from complaints related to streetlights

4. Chicago

- "Albany Park" (ClusterZC$_{13}$) suffers from complaints related to public amenities.
- "Lakeview area" (ClusterZC$_{15}$) suffers from complaints related to streetlights.



5. San Francisco, USA

- Area near "Haight" (Cluster ZC$_{21}$) suffers from noise related complaints.

- Area near “Chinatown” (Cluster ZC_{13}) suffers from public amenities complaints (graffitti).

S.F. faces increasing pressure to clean sidewalks as city grows

It's next to impossible to find a truly bucolic, peaceful neighborhood in San Francisco, but some areas are undoubtedly noisier than others. Real estate website Trulia took a stab at figuring out which neighborhoods are the loudest by compiling police data on noise complaints around the city for the past five years or so. This method isn't perfect, as they admit—serial complainers, reporting biases, and population could be factors skewing the data. But there are enough facts there to give us a pretty good idea of where people are complaining about noise the most.

Not surprisingly, **the Tenderloin, the Haight, and North Beach are hot spots** for lots of sound. There's a clear line of noise complaints stretching down Sixth Street, and the Mission is also covered by a high decibel zone. Some smaller audio issues appear in surprising places, including a dot that pops up in the Presidio.

It's taggers vs. artists in San Francisco's graffiti war

By Richard Ehrlich, for CNN
Updated 1602 GMT (2302 HKT) September 3, 2014

13 photos: San Francisco graffiti war

Tagged mural – Mats Stromberg (pictured) completed his "Giant Selfie" mural in San Francisco's Mission District in 2013, but blobs of silver-colored letters outlined in red, spelling the name "Blake," smothered much of it in July.

6. Bangalore

- Area near Banaswadi ClusterLBC_{18} suffers from garbage issue.
- Area near Kodigehalli ClusterZC_{17} suffers from dengue mosquito problem.

GARBAGE ISSUE ON BANASWADI MAIN ROAD

By: Somanna M B, 2016-01-24 18:01:17.0 Tweet

Incomplete gutters roads and increasing in mosquito problems

Open Tagged in BWSSB, Health & Sanitation

Hello, Year back BBMP people started to replace concrete gutters and gibe for drainage pipe replacements. The entire area was excavated, this work yet to be completed. From past 6 months no work is in progress and gutter construction work was partially completed. This is leading to drain water storage and mosquito problem, garbage storage at entire area. Please look into this issue in priority. Area: BBMP Ward no.8, Balaji layout, Kodigehalli, byataranapura, yelhanka hoblie, Bangalore North.

5. Conclusion and Future work

5.1 Conclusion

The report has addressed the problem of performing analysis on the civic complaints raised by the public. Actual geographical boundaries are also included in the analysis leading to a different view of the analysis. The analysis will be helpful for city planners to take proactive decisions. It can be concluded after performing the civic complaints analysis; that the two approaches i.e. DGCA and CZCA performed almost similarly. Resources can be allocated by considering the criticality of the complaint type. The criticality index also comprises the factor of timestamp of complaint. Hence the actions can be influenced based on the time of the day. This will help in efficient relocation of resources. A cumulative criticality index is also computed for each city corresponding to different complaint category to compare the different cities. The analysis also considered sub regions of a city denoted as location based cluster or zip code based cluster. This will help city planners to analyse the problem at the ground level and take measures to resolve them. The analysis of criticality can be helpful in developing the alert systems which sends the appropriate alerts to the authorized agency for the prompt action.

The categorization of sub regions within a city helps in analysing the frequency patterns of different complaint category. The prioritization can be done for each sub regions based on the categorization done for the scenario. It is concluded that Cosine similarity is much more accurate than Euclidian to perform categorization for both the approaches.

5.2 Future work

In this report analysis of civic complaints from different data source is done. The analysis includes finding the criticality of complaint in terms of density, co-occurrence with respect to time and resolution time period. There can be improvements in computing criticality so that more parameters are considered in computing criticality. These techniques can further be extended to perform predictive analysis. Moreover new techniques can be explored to perform analysis on other attributes of a complaint.

References

[1] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, “Urban computing: Concepts, methodologies, and applications,” ACM Trans. Intell. Syst. Technol., vol. 5, no. 3, pp. 38:1–38:55, Sep. 2014

[2] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, “Tdrive: Driving directions based on taxi trajectories,” in Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, ser. GIS ’10. New York, NY, USA: ACM, 2010, pp. 99–108.

[3] S. Ma, Y. Zheng, and O. Wolfson, “T-share: A large-scale dynamic taxi ridesharing service,” in Data Engineering (ICDE), 2013 IEEE 29th International Conference on, April 2013, pp. 410–421.

[4] F. Zhang, D. Wilkie, Y. Zheng, and X. Xie, “Sensing the pulse of urban refueling behavior,” in Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ser. UbiComp ’13. ACM, 2013, pp. 13–22.

[5] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu, “Inferring gas consumption and pollution emission of vehicles throughout a city,” in Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD ’14. ACM, 2014, pp. 1027–1036

[6] Y. Zheng, T. Liu, Y. Wang, Y. Zhu, Y. Liu, and E. Chang, “Diagnosing new york city’s noises with ubiquitous data,” in Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ser. UbiComp ’14. New York, NY, USA: ACM, 2014, pp. 715–725.

[7] T.-C. Yen, T.-Y. Lin, C.-Y. Yeh, H.-P. Hsieh, and C.-T. Li, “An interactive visualization system to analyze and predict urban construction dynamics,” in SIGKDD International Workshop on Urban Computing, ser. (UrbComp 15. Sydney, Australia: ACM, 2015.

[8] V. Frias-Martinez and E. Frias-Martinez, “Spectral clustering for sensing urban land use using twitter activity,” Eng. Appl. Artif. Intell., vol. 35, pp. 237–245, Oct. 2014.

[9] M. Lenormand, M. Picornell, O. G. Cant ´u-Ros, T. Louail, R. Herranz, M. Barthelemy, E. Fr´ıas-Mart´ınez, M. San Miguel, and J. J. Ramasco, “Comparing and modelling land use organization in cities,” Royal Society Open Science, vol. 2, no. 12, 2015.

[10] J. Cranshaw, R. Schwartz, J. I. Hong, and N. Sadeh, “The livehoods project: Utilizing social media to understand the dynamics of a city,” in International AAAI Conference on Weblogs and Social Media, 2012, p. 58

[11] Ferrari, Laura, and Marco Mamei. "Discovering city dynamics through sports tracking applications." Computer 44.12 (2011): 63-68.

[12] Ahmadvand, B. Bidgoli, and E. Akhondzadeh, “A hybrid data mining model for effective citizen relationship management: A case study on tehran municipality,” in e-Education, e-Business, e-Management, and eLearning, 2010. IC4E ’10. International Conference on, Jan 2010, pp. 277–281.

[13] Zha, Yilong Frank, and Manuela Veloso. "Profiling and Prediction of Non-Emergency Calls in NYC." Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014.

[14] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Elsevier, 2011.

[15] Leg´any, S. Juh´asz, and A. Babos, “Cluster validity measurement techniques,” in Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, ser. AIKED’06. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2006, pp. 388–393.

[16] http://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/how-kernel-density-works.htm

List of Publications

[1] Preeti Bansal and Dr. Durga Toshniwal , "Analysing Civic Complaints for Proactive Maintenance in Smart City",15th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2016), June 26 - 29, 2016, Okayama, Japan (Accepted on 18th April 2016)