CARNEGIE MELLON UNIVERSITY

School of Architecture
College of Fine Arts

Thesis

Submitted in Partial Fulfillment of the requirements for the degree of

Master of Science in Computational Design

TITLE:

A New Web-Based Interactive Data Visualization Tool for Geo-Located Data Clustering and Its Applications in Design

AUTHOR:

Yuqiao Tang

ACCEPTED BY ADVISORY COMMITTEE:

[redacted]		
Professor NAME	Principal Advisor	DATE
[redacted]		5.8.2018
Professor NAME	Advisor	DATE

A New Web-Based Interactive Data Visualization Tool for Geo-Located Data Clustering and Its Applications in Design

Yuqiao Tang

Advisory Committee:
Professor Daniel Cardoso Llach
Professor Stephen Quick

Abstract

Data clustering is an important technique for data visualization and statistical data analysis. Using data clustering technique for geo-located data analysis and visualization has been shown to be useful in many practical domains. Although there are numerous studies for data clustering algorithms, the implementations and applications based on these algorithms are rare and inadequate, especially for web-based data visualization services.

In this research, a new web-based interactive data clustering prototype is proposed for geo-located data visualization. The necessary web-based infrastructure for geo-located data analysis and visualization is designed and implemented at first. And then the new interactive data clustering prototype using multi-dimensional K-Means algorithm is implemented on the existing infrastructure.

In order to assess whether the new prototype can contribute meaningfully to architectural design and urban studies, qualified users with urban study or architectural backgrounds are invited to use the prototype and then make their comments and evaluations. To make it clear how this new prototype might apply in real world settings, an application scenario in which the user might interact with the prototype is provided and analyzed.

Keywords

Data Visualization, Clustering, K-Means, Online Platform

Acknowledgement

Throughout the completion of my thesis, I have been supported, helped and advised by many friends, colleagues and outstanding professors. I'd like to thank you all for your time and effort in helping me through my graduate study.

Most importantly, I'd like to thank my committee members, professor Daniel Cardoso Llach and professor Stephen Quick for your valuable comments and suggestions. I was greatly inspired by your advice and it helped me greatly in improving the quality of this thesis.

Additionally, I'd like to thank Chen Xu for working with me on the development of the platform infrastructure and George Zhu for being my technical consultant. Also, I'd like to thank the five anonymous users who evaluated the clustering platform and offered precious feedbacks.

In the end, I'd like to thank my parents for supporting me through my graduate study at Carnegie Mellon University. Love you!

Table of Contents

CHAPTER 1 : INTRODUCTION 6

1.1 BACKGROUND 6
1.2 RESEARCH QUESTION AND GOALS 7
1.3 THESIS OUTLINE 7

CHAPTER 2 : LITERATURE REVIEW 9

2.1 LAND USE CLUSTERING 9
2.2 DEMOGRAPHIC CLUSTERING 13
2.3 CRIME INFO CLUSTERING 17
2.4 NATURAL DISASTER CLUSTERING 20
2.5 EXISTING DATA CLUSTERING TOOLS 22
2.6 LIMITATIONS OF PREVIOUS RESEARCH 25

CHAPTER 3 : HYPOTHESIS 26

CHAPTER 4 : METHOD 27

4.1 WEB-BASED DV PLATFORM DESIGN AND IMPLEMENTATION 27
4.1.1 System Architecture 27
4.1.2 Core Functionality 28
4.1.3 Supporting Functionality 30
4.2 NEW DATA CLUSTERING PROTOTYPE 32
4.2.1 Observation-Weighted K-Means Algorithm 32
4.2.2 New Feature: Cluster Analysis 34
4.2.3 New Feature: Data Sharing 36
4.2.4 Platform as a Learning Tool 37
4.3 SYSTEM OVERVIEW 38
4.3.1 Design Choices and Tradeoffs 38
4.3.2 Tests and Deployment 39

CHAPTER 5 : APPLICATION SCENARIOS 40

5.1 CITY OF PITTSBURGH MARKET VALUE ANALYSIS 40
5.1.1 Using Geographic Information as Features 41
5.1.2 Sale Price & Assessed Condition of the Property 42
5.1.3 Sale Price & Percent of Residential Vacancy 43
5.1.4 Sale Price & Density of Housing Units 44
5.1.5 Sale Price & Index of Distress 44
5.1.6 Comprehensive Analysis for Sale Price 45
5.1.7 Conclusion for 2016 MVA Data Clustering 47

CHAPTER 6 : USER EVALUATION 48

6.1 INSTITUTIONAL REVIEW BOARD 48
6.2 PROCEDURES OF THE USER EVALUATION 48
6.2.1 Pre-Test Questionnaire for Background Identification 48
6.2.2 Feature Demonstration 50

6.2.3 User Participation *50*
6.2.4 After-Test Questionnaire for Feedback Gathering *50*
6.3 FEEDBACKS AND DISCUSSIONS 52
CHAPTER 7 : CONTRIBUTIONS AND FUTURE WORK 53
7.1 RESEARCH CONTRIBUTIONS 53
7.2 LIMITATIONS AND FUTURE WORK 54

Chapter 1 : Introduction

1.1 Background

Data clustering, which indicates 'the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are most similar (in some sense) to each other than to those in other groups (clusters)'(Tryon, 1939), is an important technique for data visualization and statistical data analysis. It has been studied for many years and has been used in many fields, including data classification, machine learning, pattern recognition, computer graphics, image analysis and data visualization.

Using data clustering technique for geo-located data analysis and visualization has been shown to be useful in many practical domains, such as urban planning, city zoning and diagnosis, real estate development, architectural design, social science and economic analysis, historical climate and natural disaster studies.

Although there are many studies for data clustering algorithms, the implementations and applications based on these algorithms are rare and inadequate, especially for web-based data visualization services. Although some existing web-based data visualization platforms, such as Carto and Mapbox, provide 'data clustering' feature for users, they only provide one basic algorithm, the k-means algorithm, to support datasets with the specific schema. They also do not accept large datasets for data clustering feature.

Many researchers, especially those with urban planning backgrounds, use ArcGIS for visualizing and clustering geo-located datasets. Although ArcGIS is a powerful tool in many aspects, it requires many professional skills while in the same time it is not interactive enough for new users. In other words, ArcGIS is too 'cold' a tool for many studies analyzing geo-located datasets.

Some researchers created their own tool for data clustering and visualization for their projects using programming languages such as JavaScript and Python. However, it required the researchers to not only have the professional skills in their own fields, but also the strong skills in programming and database manipulation. It sure is a time-consuming task, and it is never a choice for those beginners, most junior students and researchers who just don't have enough programming skills.

To provide a better tool in geo-located data clustering and visualization, this research is dedicated to designing and developing a new prototype and conducting its application scenarios for a variety of applications.

1.2 Research Question and Goals

In this thesis, a new web-based interactive data clustering prototype is proposed for geo-located data visualization. The necessary web-based infrastructure for geo-located data analysis and visualization is designed and implemented at first. And then the new interactive data clustering prototype using multi-dimensional K-Means algorithm is implemented on the existing infrastructure.

Considering the significant advantages of interaction, it is believed the new prototype can benefit studies and projects in many practical domains where geo-located datasets are used.

In order to assess whether the new prototype can contribute meaningfully to architectural design and urban studies, qualified users with urban study or architectural backgrounds are invited to use the prototype and then make their comments and evaluations. To make it clear how this new prototype might apply in real world settings, an application scenario in which the user might interact with the prototype are provided and analyzed.

1.3 Thesis Outline

In chapter two, previous work which used data clustering techniques on many different geo-locate datasets is introduced. Previous researchers have used data clustering technique on geo-located data analysis and visualization in urban studies, city zoning, site analysis etc. Different datasets, including census data, air quality data, crime information, social media data and natural disaster data are clustered in the studies for digging more information about the cities.

After explaining the hypothesis and validation method of this research, chapter four introduced the features of the new prototype and all the architectural decisions. The design and implementation method of the web-based infrastructure for geo-located data analysis and visualization is introduced. The architecture of the system, including the backend, frontend and the deployment is documented. The core functionality and the supporting functionality of the web-based platform are introduced, and the design tradeoffs of the system are discussed.

In chapter five, two application scenarios are provided in order for the users to understand how the prototype might apply in real world urban planning work. The two application scenarios are the market value analysis in the City of Pittsburgh, and the ethnic clustering analysis in the United States by county. The prototype is not limited to these categories and users can use it to explore any clustering problem that he cares about as well.

In chapter six, the preparations and procedures for user evaluation are presented. For users to evaluate the new prototype, qualified users with urban study or architectural backgrounds are invited to use the prototype and then make their comments and evaluations. The preparations of this user evaluation experiments include questionnaire

design, data pool establishment and applying for IRB approval. The user evaluation experiment has four procedures introduced in the second section of this chapter. And the feedbacks and evaluations are gathered and discussed in the last section of this chapter.

In chapter seven the results and contributions are presented at the first place, while the limitations and the future work of this research are discussed as well.

Chapter 2 : Literature Review

Theoretical research on data clustering algorithm has been studied for many years. Clustering algorithms can be categorized based on their cluster model, objective function, target question. There is no objectively "correct" data clustering algorithm. As it was noted, "clustering is in the eye of the beholder" and "an algorithm that is designed for one kind of model will generally fail on a data set that contains a radically different kind of model." (Estivill-Castro & Vladimir, 2002)

Previous researchers have used data clustering technique geo-located data analysis and visualization in urban studies, city zoning, site analysis etc. Different datasets, including census data, air quality data, crime information, social media data, natural disaster data, etc. are clustered in the studies for digging more information about the cities.

2.1 Land Use Clustering

Evaluating the actual urban land use has been an important concern for urban planning. The traditional methods to evaluate the actual use of urban lands required on site data collection, including data collection through direct observation or data collection from questionnaires that attempt to 'capture how citizens interact with their urban environment' (Frias-Martinez & Frias-Martinez, 2014). These traditional methods have many limitations, including the resiliency of citizens to provide information that might cause privacy concerns, the cost of using questionnaires that could highly limit the frequency with the information collection (Yin et al., 2011). To evaluate the actual urban land use in a more efficient and accurate way, researchers had done studies using data clustering algorithms on citizens' digital footprints, including mobile devices, social media activities etc., to identify all possible land uses using spectral clustering.

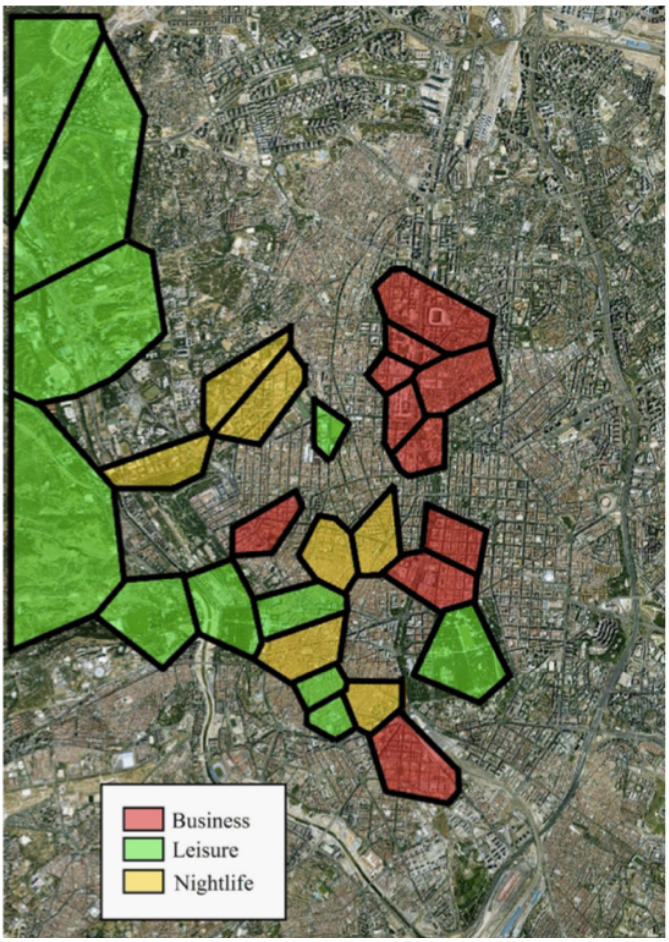
In the article *'Spectral Clustering for Sensing Urban Land Use Using Twitter Activity',* the authors used the data from Twitter activity as the sensor of interactions between individuals and their environment. The authors proposed a new technique which used the unsupervised learning algorithm on urban areas to cluster geographical regions and automatically identify the actual land use of the regions with similar tweeting activity patterns. To validate this new technique, the authors presented case studies on three different urban environments: Manhattan (NYC), London (UK) and Madrid (Spain) using the local Twitter activity datasets and the official land use information provided by the city planning departments. By using Twitter Streaming API provided by Twitter Inc. since 2013, the researchers gathered real time geo-located tweets which provided the real time information needed to identify the actual urban land uses (Frias-Martinez & Frias-Martinez, 2014).

The following figures used in this work showed the physical layout of business, nightlife and leisure clusters in three different urban environments. Areas not marked with any color indicated the residential land uses. The figure (a) showed the view of

physical layout of Manhattan, while figure (b) showed the physical layout of Madrid and figure (c) showed the physical layout of London.



(a)



(b)



(c)

Fig.2.1.1 Physical layouts of business, nightlife, leisure and residential land use in Manhattan, Madrid and London (Frias-Martinez & Frias-Martinez, 2014)

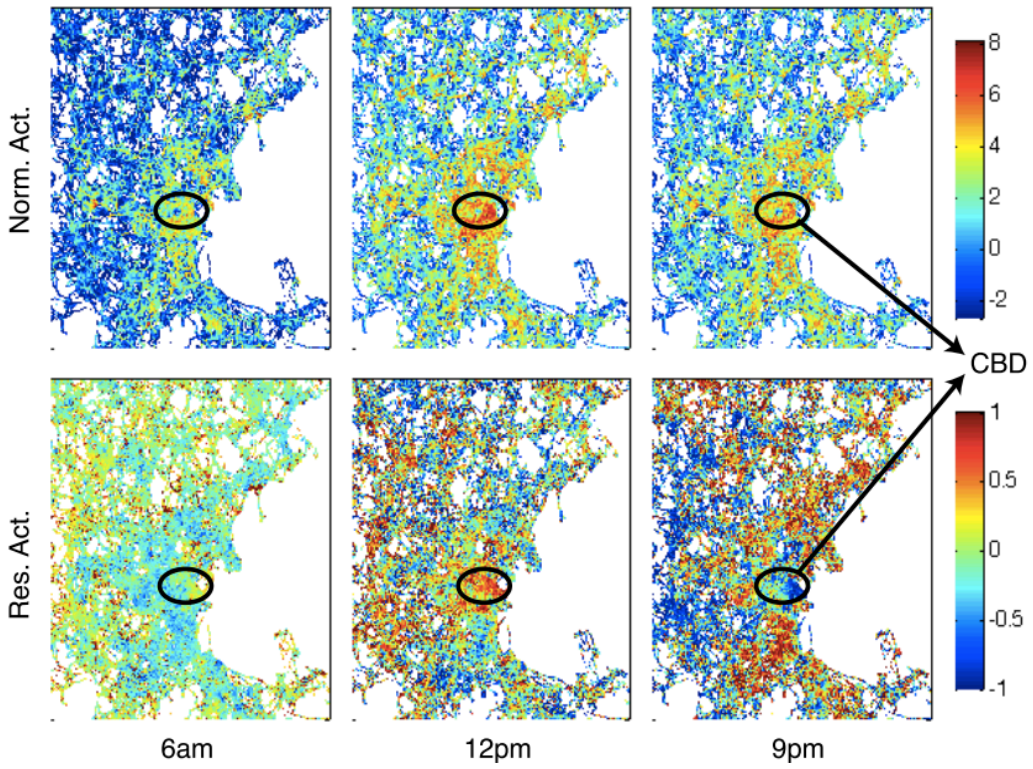
The following table used in this research showed the percentage of the overlap between official land use information provided by the urban planning departments and Twitter land uses for Manhattan, London and Madrid. These results indicated that using spectral clustering on geo-located Twitter activities can be a powerful technique for identifying the actual urban land uses.

Table.2.1.1 The percentage of overlap between official and Twitter land uses (Frias-Martinez & Frias-Martinez, 2014)

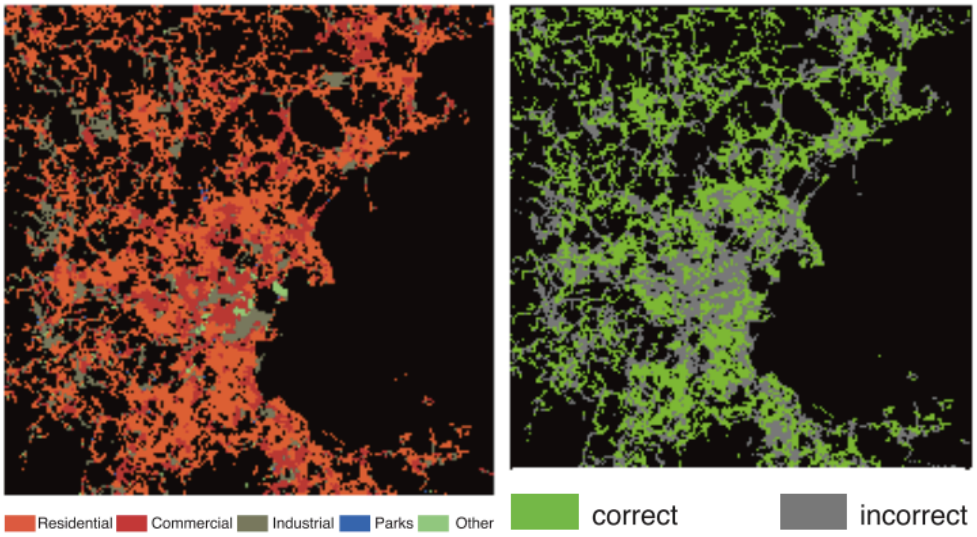
Official land use	**Twitter land use**				
	Business	Residential	Nightlife	Leisure& Weekend	Industrial
London					
Non-domestic buildings	**61%**	9%	3%	2%	25%
1t1Domestic Buildings	9%	**56%**	23%	6%	6%
Greenspace & paths	8%	11%	7%	**72%**	2%
Manhattan					
Commercial	**81%**	12%	3%	2%	–
Residential	7%	**68%**	19%	4%	–
Industry	13%	77%	**0%**	6%	–
Park & recreation	6%	7%	6%	**81%**	–
Madrid					
Commercial & business	**69%**	25%	4%	2%	–
Residential	11%	**61%**	18%	10%	–
Industry	58%	33%	**3%**	6%	–
Greenspace	7%	16%	6%	**71%**	–

In the article *'Inferring land use from mobile phone activity'*, the authors used the ubiquitous mobile sensors including personal GPS devices and mobile phones instead of the costly survey to collect massive amounts of data on urban system. The researchers used these data on conducting research on the spectral distribution of citizens in urban environments. By using these dynamic data generated by mobile phone users, the researchers identified the relationship between actual urban land use and the dynamic population. The authors then used the unsupervised machine learning classification algorithm to identify clusters of regions with similar zoned uses and mobile phone activity patters. To validate the capability of using mobile phone data to identify the actual urban land use, the researchers made comparisons between the classification results and the official zoning regulation for Boston area (Toole, Ulm, González, & Bauer, 2012).

The following figures used in this work showed the spatial distribution of mobile phone activity and the zoning maps for Boston area. Figure (a) showed the spatial distribution of 'absolute and residual phone activity over the course of a day'. Figure (b) showed the zoning map as predicted from mobile phone data using the random forest classification algorithm. Figure (c) showed the spatial distribution of where the algorithm predicted the actual land use correctly and where it predicted wrong.



(a)



(b) (c)

Fig.2.1.2 Spatial distribution of mobile phone activity and the zoning maps for Boston area (Toole et al., 2012)

These results indicated that using spectral clustering on geo-located mobile phone data is a powerful technique for delivering useful information on the actual urban land uses which can supplement the official zoning regulation (Toole et al., 2012).

2.2 Demographic Clustering

Demographic clustering is a topic of considerable academic interest focused by urban planners, real estate developers and social scientist to study and evaluate the urban development and social science. The studies vary from 'Quality of Life' in certain neighborhood, 'Spatial Segregation of Ethnic Minorities' in a region, to 'Identifying the Areas with Homogeneity' in communities.

In the article *'Legible Cities: Focus-Dependent Multi-Resolution Visualization of Urban Relationships',* Chang developed a unique tool, *UrbanVis*, based on a clustering algorithm to provide physical and informational display depicting abstract data such as land use and demographic information, etc. that are easy to understand from all levels of abstraction. The tool was developed with the goal of opening up the possibilities for the potential users to explore and re-examine the existing understandings of their target cities.(Chang, Wessel, Kosara, Sauda, & Ribarsky, 2007)

The following figure used in this work showed the building model aggregation used by *UrbanVis*. The figure (a) showed the view of downtown Charlotte with its surrounding regions, figure (b) was the enlarged version of the selected yellow box region from figure (a), figure (c) used a pixel tolerance rate of 50, figure (d) used a pixel tolerance rate of 500.

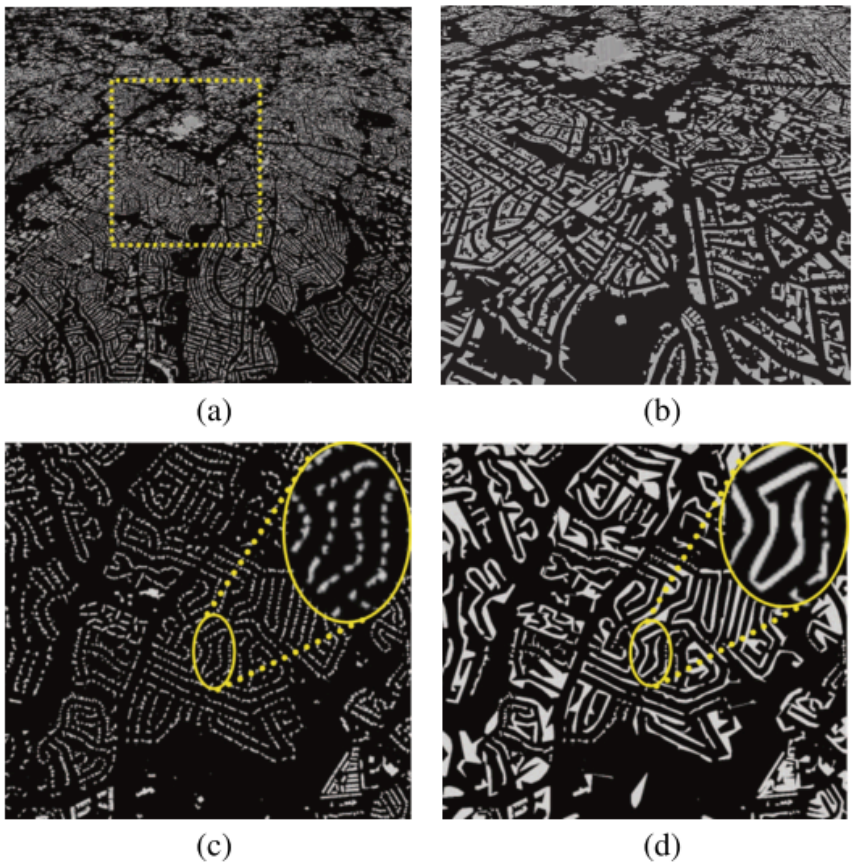


Fig.2.2.1 The building model aggregation used by UrbanVis (Chang et al., 2007)

In this work, the authors used the demographic data, which covers various categories including ethnicity, citizenship, job status, income and housing statistics, taken from the United States Census 2000 for Charlotte in North Carolina for two application scenarios. By using the census data for the county of Mecklenburg, they were able to conduct part of the 'Charlotte Neighborhood Quality of Life Study' to look for communities with low levels of income in order to figure out the possible improvements to these communities through urban planning methods.

The following figure was used in this article to illustrate this application scenario. It showed the communities with high Hispanic population near downtown area.



Fig.2.2.2 The communities with high Hispanic population near downtown area (Chang et al., 2007)

Another scenario the authors provided in their work is to identify areas with homogeneity in demographics. This topic is important for real estate industry. A real estate developer in Harris Associates pointed that identifying areas with homogeneity was often crucial when new concepts or radical issues in urban planning were introduced in a community. Based on his experience, new concepts are more likely to be accepted in communities with high homogeneity in demographics because people share similar background there, while different demographic groups often provide different opinions based on their differences in perspectives in communities with high heterogeneity. (Chang et al., 2007)

The following figure was used in this article to illustrate the above application scenario. The top figure showed the amount of homogeneity in Davidson area, while the bottom figure showed the amount of homogeneity in downtown Charlotte area.

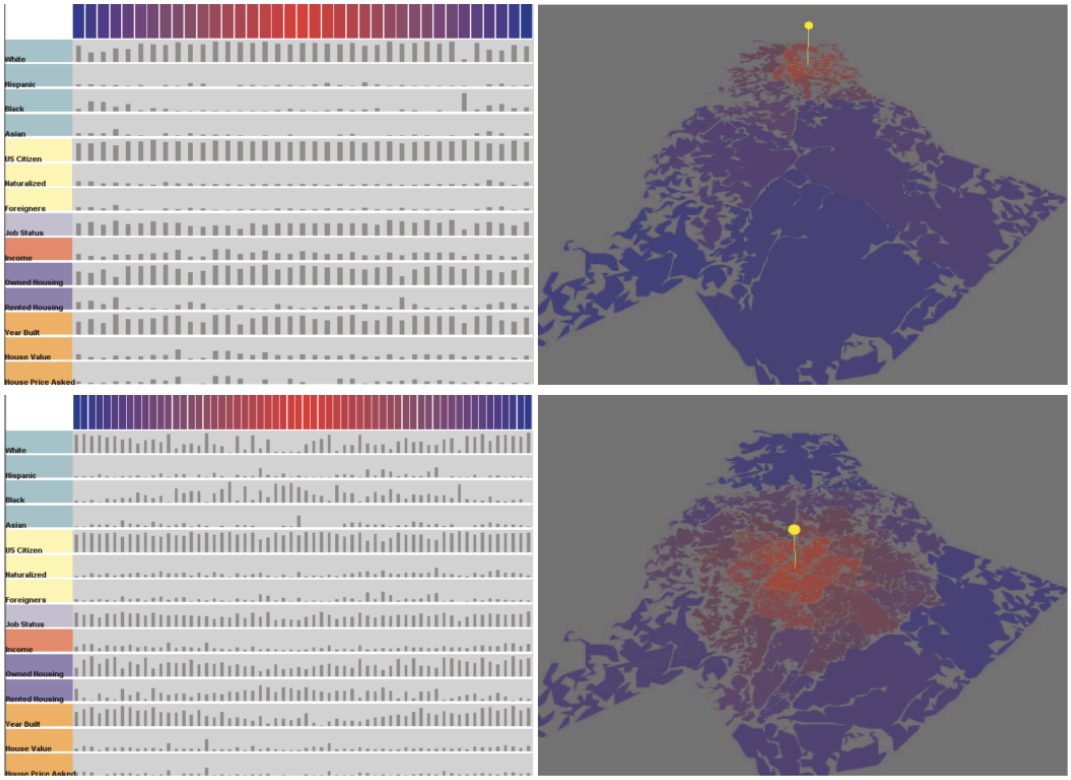
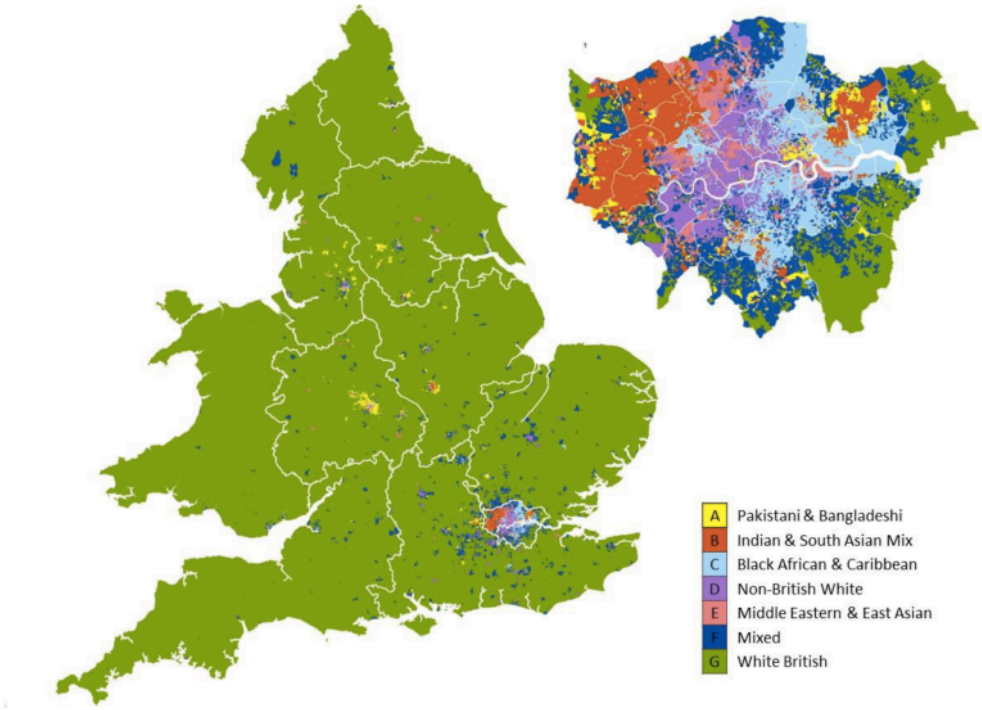


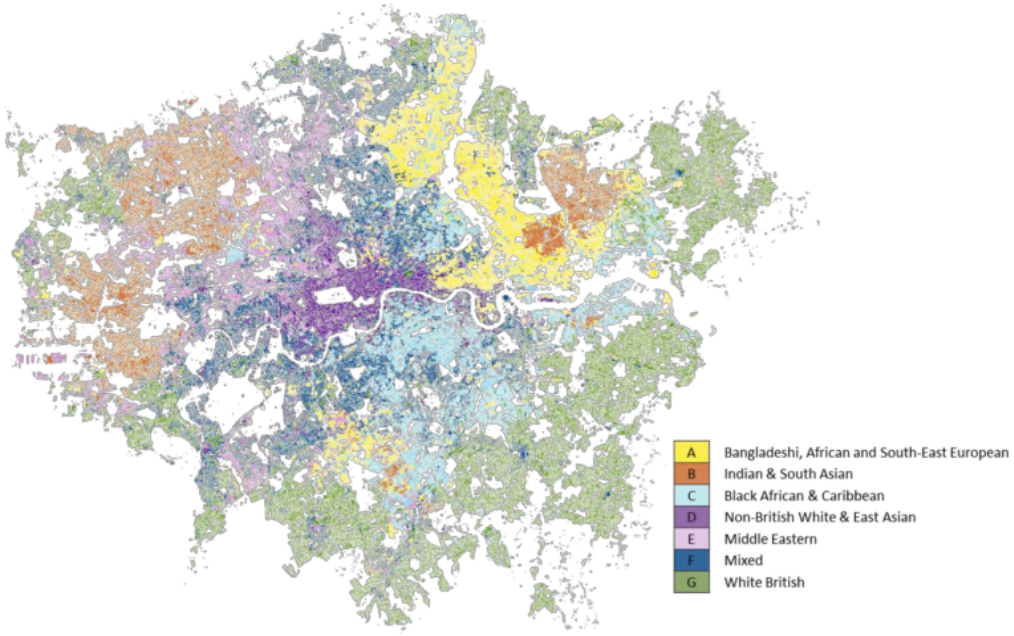
Fig.2.2.3 The differences in the amount of homogeneity between Downtown Charlotte and Davidson area (Chang et al., 2007)

In the article *'Identifying the Major Traits of Ethnic Clustering in England and Wales from the 2011 Census'*, the authors used the 2011 Census data from England and Wales to identify the 'spatial segregation of ethnic minorities' and the 'major spatial variations in ethnic composition' between communities in England and Wales. The researchers believed that the demographic clustering for small communities can be beneficial to identify the local composition of cultural groups, which can improve the local service provision such as shopping facilities, grocery store provision for minority groups in Britain who have distinctive food consumption habits due to their cultural origins. (Uskul & Platt, 2014) (Lansley, Wei, & Rains, 2015)

The following figures were used in this article to demonstrate the results. Figure (a) is used to show the 'Cultural, ethnic and linguistic output area classification for England and Wales' in 2011. Figure (b) is used to show the 'Dasymetric map of the London specific CELOAC', while CELOAC indicating Cultural, Ethnic and Linguistic Output Area Classification. Figure (c) is used to show the 'distance of each OA to the cluster center in England and Wales (left) and in London (right)', while OA indicating output area.


A Pakistani & Bangladeshi
B Indian & South Asian Mix
C Black African & Caribbean
D Non-British White
E Middle Eastern & East Asian
F Mixed
G White British

(a)


A Bangladeshi, African and South-East European
B Indian & South Asian
C Black African & Caribbean
D Non-British White & East Asian
E Middle Eastern
F Mixed
G White British

(b)



(c)

Fig.2.2.4 The figures used in the article for demonstrations

The authors used K-means clustering algorithm in their work. It is an iterative top down approach which tries to 'reduce the sum distance between each data point and their respective cluster center' (Harris R, Sleight P, 2005). This well-known clustering algorithm can also be used in high dimensions, and it will be introduced with more details in the 'Method' chapter of this thesis.

2.3 Crime Info Clustering

Crime rate of neighborhood is one of the major concerns for people ranging from city dwellers to professional urban planners. Frequency of crime occurrences of neighborhoods is a strong indicator of economic development, safety, and level of happiness of urban residents. Although in recent years, the development of transport and communication system has expanded the range of criminal activities, regional location still has a powerful impact on the occurrence of crimes.

The study of crime-based city zoning has always been an active topic in urban planning. The development of data mining techniques has brought up a new angle in addressing this question. In addition to academic studies, data mining techniques for crimes also presents itself as a new tool for law enforcement department to efficiently analyze the growing volumes of crime data. (Chen et al., 2004)

In India, Thota et al. tried to model crime zones using the criminal dataset from India's National Crime Records Bureau. They used cluster technique of k-means algorithm to get zoning information. The output from the algorithm is used to create a color-coded map that displays an overall crime profile of all states in India. The authors argues that this custom map will help law enforcement department to 'take

additional preventive measures to combat against crime and plan advanced investigation strategies'. (Thota et al., 2017)

The following figure was presented by Thota et al. as a result of cluster analysis. States in India were clustered into three categories, each representing a certain risk level.

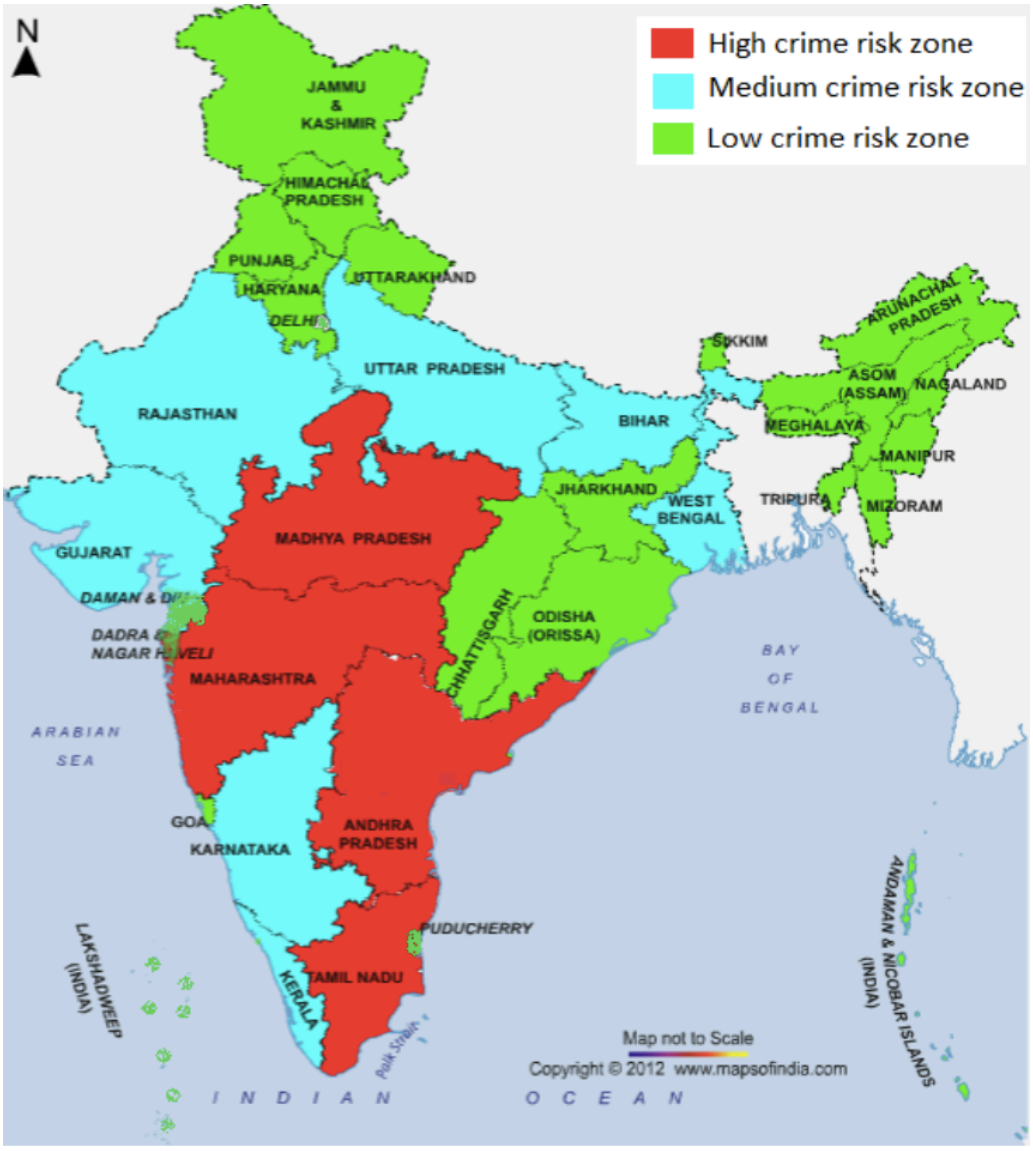


Fig.2.3.1 Color-coded crime info map of India created from cluster analysis (Thota et al., 2017)

Despite urban zoning studies, cluster algorithm also serves as an analytical tool in automatically profiling criminals based on patterns of crimes. (Chen et al., 2004) Suspects who conduct crimes in similar ways or individuals from a particular gang can be identified automatically using cluster algorithms and other data-mining techniques. Driven by the needs to automate crime investigation, Chen et al. developed a data mining framework that incorporates many different types of algorithms including neural network, clustering, social network analysis, etc. Each of algorithms is used for

a different goal of analysis ranging from pattern visualization, prediction, association and entity extraction.

The following figure was used by Chen et al. to illustrate a use case of their framework. It shows 16 target gang members were identified by network analysis from a crime network of 164 criminals in the Tucson Police Department database.

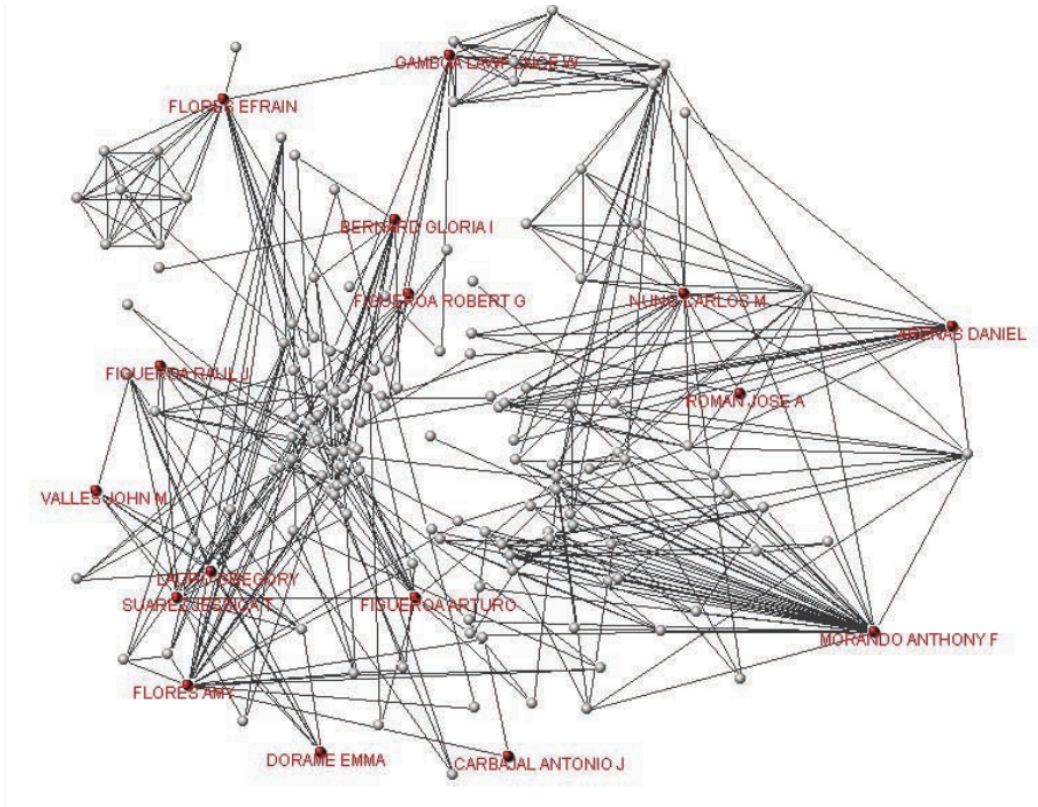


Fig.2.3.2 Crime network analysis (Chen et al., 2004)

2.4 Natural Disaster Clustering

Cluster analysis of geo-located data can also be used in zoning of natural disasters. Occurrences of certain natural disasters, such as earthquake, volcanic eruptions and lightning strikes can have significant correlation with geographical locations. In the analysis of natural disasters, the clustering algorithm is often used as a classification tool to generate severity indicators of different geographical zones. It is often claimed that these indicators will help disaster management authorities predict and take preventive measures in the event of natural disasters.

Gao et al. combined data of lightning disasters, observational thunderstorms of meteorological stations and regional population and economic indicators to analyze vulnerability risks of different zones in Hainan province, China.(GAO, MENG, & LAO, 2013) Gao et al. used a hierarchical cluster analysis method, Q-Cluster, to classify counties of Hainan province into multiple lightning vulnerability risk zones. The authors argue that their new approach is more adaptive than the traditional five-level classification technique, as it can set different number of clusters and mathematically validate classification results. They used a metric of Comprehensive Vulnerability Indicator which combines lightning frequency, lightning severity, economic loss and human casualties. It is argued that this classification metric is more accurate than standard empirical metric, which lacks the possibility to be validated.

Seismic zoning is also an area where cluster analysis is widely used. As the occurrence of earthquakes is highly related to the movements of earth's crustal plates, cluster analysis based on geo-located data is a very reasonable choice.

The problem of seismic zoning in some bounded area has been a widely studied problem in many researches(Markušić & Herak, 1998), many of which try to incorporate algorithmic approaches into traditional zoning techniques (Morales-Esteban, Martínez-Álvarez, Scitovski, & Scitovski, 2014; Scitovski, 2018).

Specifically, Scitovski et al. developed a new clustering algorithm, Rough-DBSCAN, that is possible to recognize non-convex shapes. This feature is very desirable as many clustering algorithms such as K-Means operates on Euclidean distances, hence only provide convex shapes as clustering outcome. Scitovski et al. applied the clustering algorithm to the problem of earthquake zoning in a wider area of the Republic of Croatia. (Scitovski, 2018).

Upon retrieving the dataset, Scitovski first passed the dataset through a normalization step to prepare it for clustering. This step can be essential in various other application scenarios where numerical ranges of data varies greatly across different features. The following figure shows an initial clustering of 15 clusters of earthquake data.

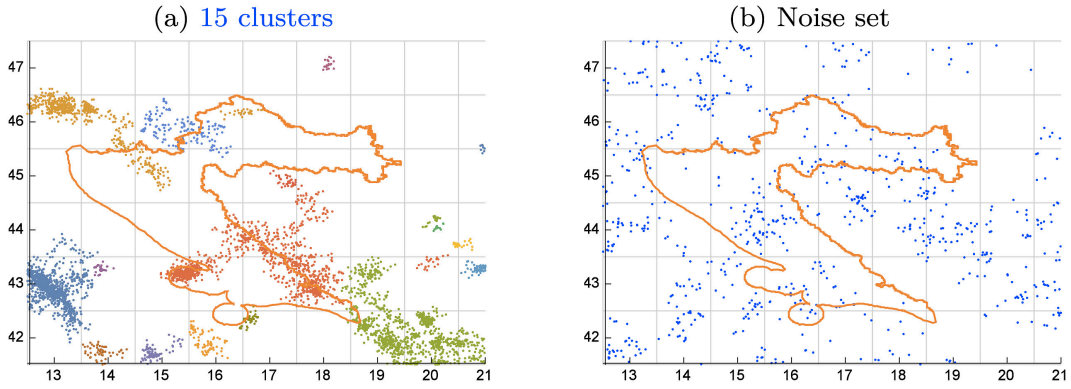


Fig.2.4.1 The results of the Rough-DBSCAN algorithm (Scitovski, 2018)

The author also applied post-processing of data on clustering output to let the result fit into a seismic zoning context. In the following figure, the author identified six significant earthquake zones for the Republic of Croatia. $\Pi_1 - \Pi_6$ on the following figure represent South Croatia, the vicinity of Zagreb, North Italy and Istria, Central Italy, Montenegro and North Albania and North Bosnia, respectively.

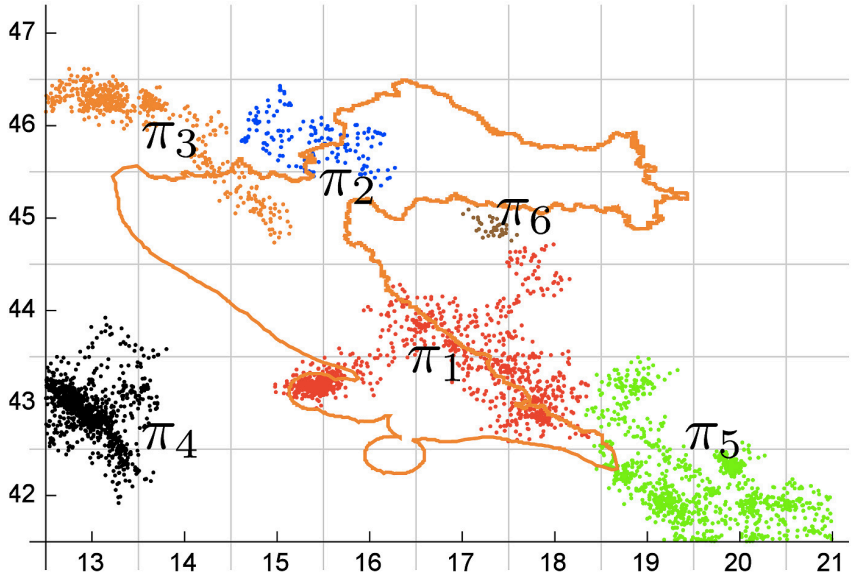


Fig.2.4.2 Significant earthquake zones for the Republic of Croatia (Scitovski, 2018)

2.5 Existing Data Clustering Tools

Cluster analysis is a very common data analysis technique that is supported by many web-based data visualization platforms and APIs. This chapter will review three of the most popular online data visualization tools with a focus on their clustering feature.

The following two figures shows how clustering is performed by Carto, a well-known data visualization platform. The US 2010 Census Populations by Zip Code (Data.gov, 2010) dataset is used for this demo. Although it successfully clusters data points into 16 color-coded clusters, there is severe limitations for this analytical feature: it only supports clustering by geographic location and can only specify a maximum of 16 clusters.

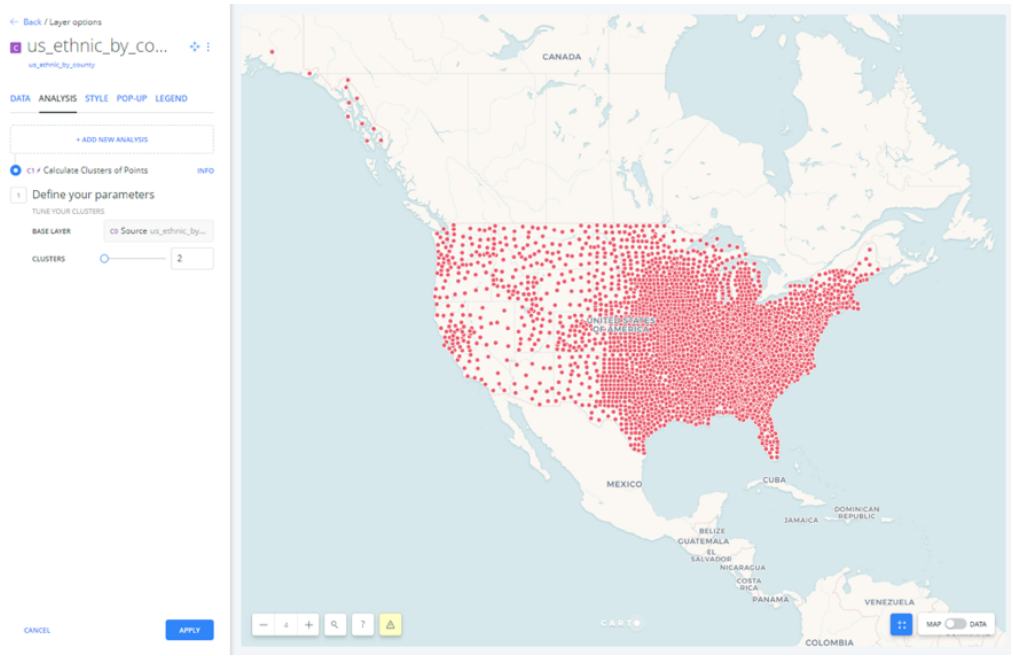


Fig.2.5.1 Clustering configuration page for Carto

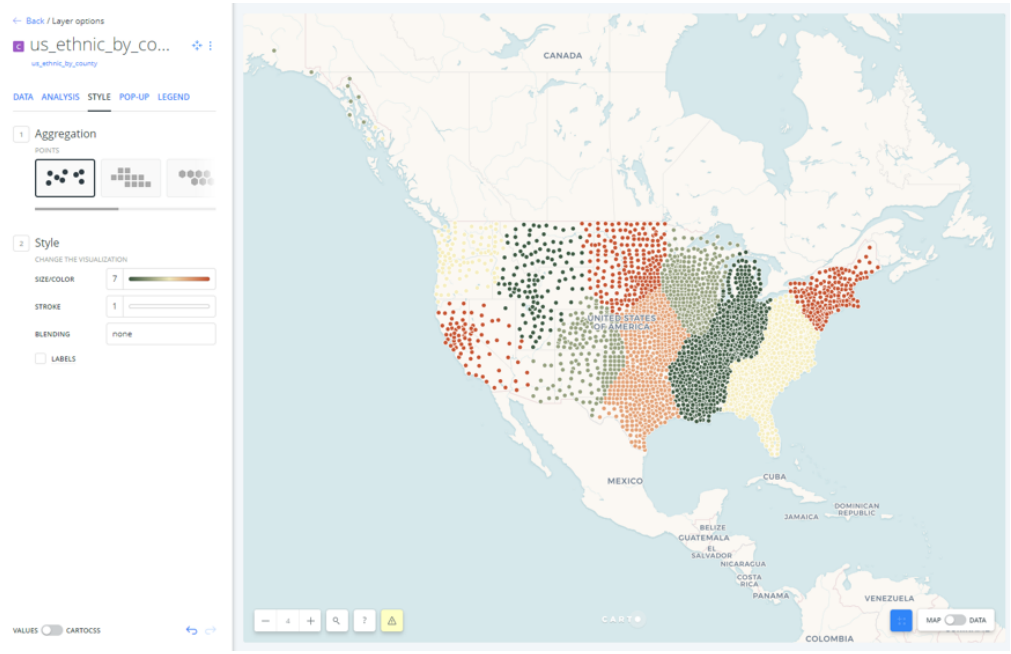


Fig.2.5.2 Clustering result on Carto

ArcGIS, another widely used data visualization platform that targets professional urban designers and researchers, also supports cluster analysis on data. The following figure shows how ArcGIS aggregates clusters into larger circles. Similar to Carto, the clustering feature is also limited to geographic locations. There is no limit on number of clusters in ArcGIS.

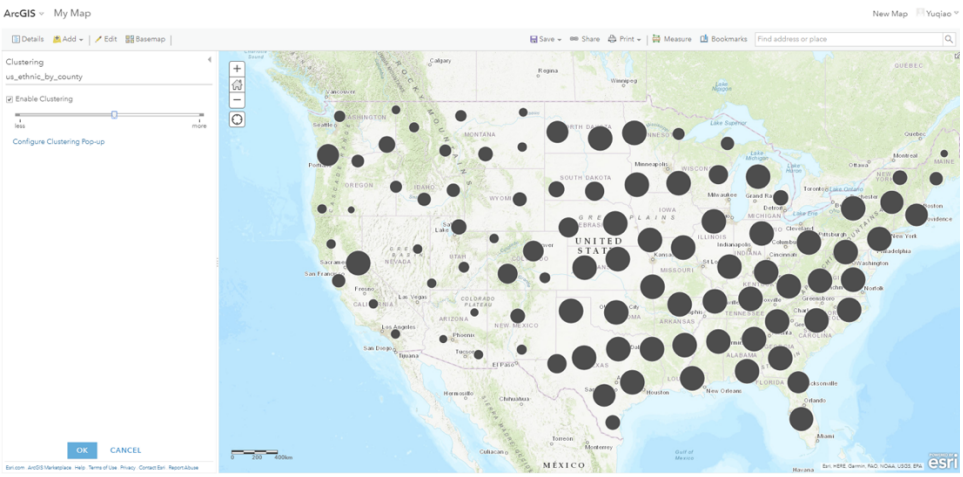


Fig.2.5.3 Clustering result represented in circles by ArcGIS

Mapbox, a web-based data visualization programming library, offers native support on clustering. Similar to ArcGIS, it renders aggregated clusters as larger circles. It offers the option of distinguishing clusters by colors and number of points included. However, as it is a programming interface, it can only be accessed using the JavaScript API, which is not friendly to designers who lack programming skills. Furthermore, like the previous two platforms, it only supports clustering by geographic locations.

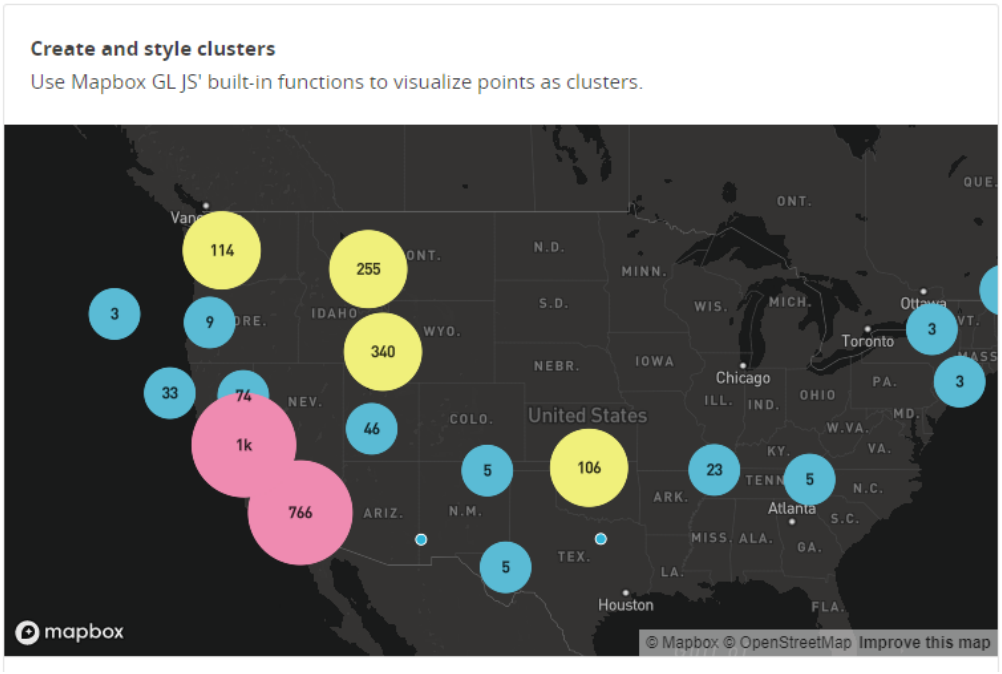


Fig.2.5.4 Clustering feature in the Mapbox API

2.6 Limitations of Previous Research

Although the previous research made great contributions in many fields, the limitations cannot be ignored.

Many researchers, especially those with urban planning backgrounds, used ArcGIS for visualizing and clustering geo-located datasets. Although ArcGIS is a powerful and convenient tool in many aspects, conducting clustering algorithm on geo-located datasets required the users to be proficient in the programmable interface of ArcGIS. Even by using the programmable interface of ArcGIS, there are still many limitations on data clustering algorithms. Some algorithms are not the adaptable, and it is not interactive enough when using ArcGIS as the tool for clustering.

Some researchers conducted their own tools for their projects. On the one hand, developing the specified software for the study required the researchers to not only contain the professional skills in the own field, but also the strong skills in programming and database manipulation. For those beginners, students, or researchers who just don't have enough programming skills, this is obviously not a choice. On the other hand, even for people who has the ability to conducting their own tools, developing a specified tool for one project can be a time-consuming task, and probably not worthy if the tool is not generic enough to be used in other related studies.

For geo-located data visualization and analysis, on-line platforms such as Carto and Mapbox are both generic and interactive enough for most datasets. However, when targeting at the 'Data Clustering' feature, the existing on-line platforms are still in level zero where everything is just too naïve for most of the urban studies to be conducted on. The existing on-line data platforms uses the most basic data clustering algorithm which makes 'distance' the only character that matters, and they have strict limitations on the data schema which filters many datasets out. Although these platforms have significant advantages on genericity and interaction, while in the same time they require few human efforts, the existing on-line platforms are not adequate for most research scenarios.

Chapter 3 : Hypothesis

In this research, a new web-based interactive data clustering prototype is proposed for geo-located data analysis and visualization. Considering its significant advantages on interaction and the ease of use, it is believed that developing and introducing an interactive web-based clustering prototype can benefit studies and projects in urban studies and architectural design process where geo-located datasets are used for data clustering.

To validate the hypothesis, the necessary web-based infrastructure for geo-located data analysis and visualization is designed and implemented at first. And then the new interactive data clustering prototype using multi-dimensional k-means algorithm is implemented on the existing infrastructure.

In order to assess whether the new prototype can contribute meaningfully to architectural design and urban studies, qualified users with urban study or architectural backgrounds are invited to use the prototype and then make their comments and evaluations. To make it clear how this new prototype might apply in real world settings, an application scenario in which the user might interact with the prototype are provided and analyzed.

Chapter 4 : Method

In this chapter, the design and implementation method of the web-based infrastructure for geo-located data analysis and visualization is introduced. The architecture of the system, including the backend, frontend and the deployment is documented. The core functionality and the supporting functionality of the web-based platform are introduced, and the design tradeoffs of the system are discussed.

The new data clustering prototype built on the web-based infrastructure is then introduced with the multi-dimensional K-Means algorithm which is used for unsupervised data classification. New features, including 'Cluster analysis' and 'Data sharing', are documented then. The possibility and potential values of using the new data clustering platform as a learning tool is also discussed in this chapter.

Last but not least, the design choice and tradeoffs of the whole system are discussed in section 'System Overview'. Test scenarios and AWS deployment of the system is also presented in this chapter.

4.1 Web-based DV platform design and implementation

To further verify how web-based data visualization tools can facilitate the design process, an online data visualization platform for geo-located dataset is designed and implemented as the infrastructure of the new data clustering prototype. The motivation is to let data talk in an explicit and beautiful way, enabling visualization of different datasets in a 2D map space.

The platform is also intended as a learning site for data visualization beginners, who can view and learn from data and visualization results shared by other users.

4.1.1 System Architecture

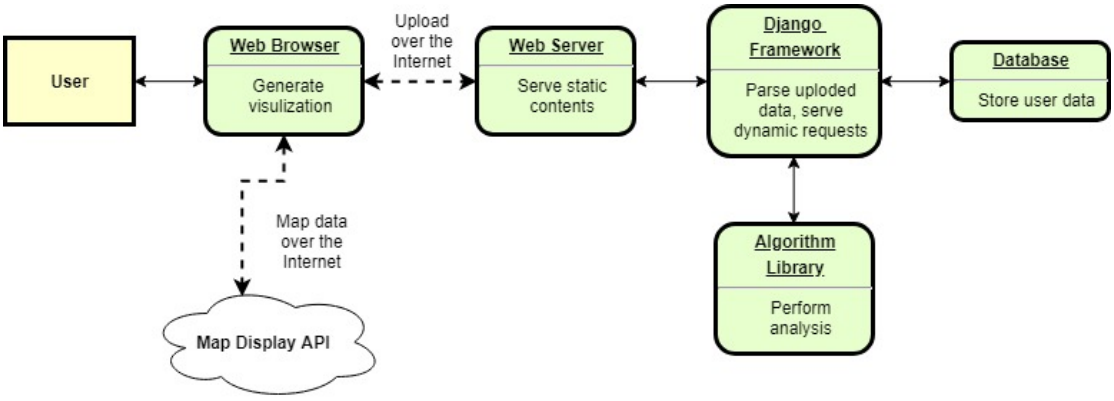


Fig.4.1.1 Platform Architecture

Like an ordinary web application, the data visualization platform follows a multi-layer system architecture. Different layers are responsible for different functionality in this platform.

(I) Web Browser - Frontend

During the process of visualizing data, the web browser on a user's computer is responsible for displaying all kinds of contents for the user, contents that include input boxes, buttons, data points and the map. JavaScript (Severance, 2012) is used to handle dynamic contents such as interactively adding and removing desired features for clustering, zooming in/out of the map, and drawing data points on the map. The Mapbox API (Mapbox, n.d.)is used for generating 2D maps and displaying data points. While interacting with the map, the web browser will constantly send requests to the MapBox API for related map data.

(II) Web Server

The web server handles all requests coming from a user (client). It immediately sends back static contents requested by user (such as pictures, HTML files) and properly forwards dynamic requests (such as a cluster analysis request) to the backend web framework. In this platform, the Apache HTTP Server (Apache Software Foundation, n.d.)is used.

(III) Web Framework and Database - Backend

Dynamic requests are handled by the web framework. In this platform, the Django (Django Software Foundation, n.d.)framework is used.

On upload, the framework parses users' dataset, store it into a database and returns a GeoJson((IETF), 2016) formatted dataset to the user, which is later used by the web browser to render data points.

When an analysis request is received, the web framework will first retrieve the user's uploaded data from the database, and then call the multi-dimensional K-Means clustering routine to get clustering result. Finally, the user's dataset, together with computed cluster information is sent to user's web browser for display. For faster performance and reliability, the SK-Learn (Pedregosa et al., 2011) library is used for clustering computation.

4.1.2 Core Functionality

To properly display a user uploaded dataset properly on the 2D map and perform cluster analysis, several procedures are involved.

(I) Upload/View Existing Dataset

On entering the platform, users are expected to choose from one of the following actions, upload a new dataset or view an existing map. Uploaded files are expected to be a CSV (Comma-Separated Values) file that contains columns explicitly named as “longitude” and “latitude”. Users are free to assign multiple hashtags to the data which can later be used as classification criteria.

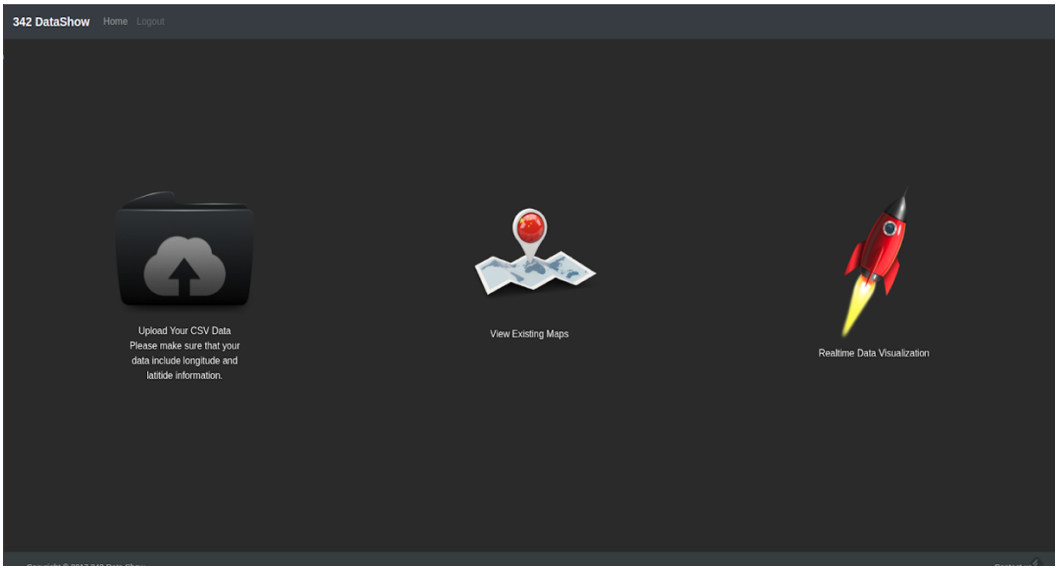


Fig.4.1.2 Choose Action Page

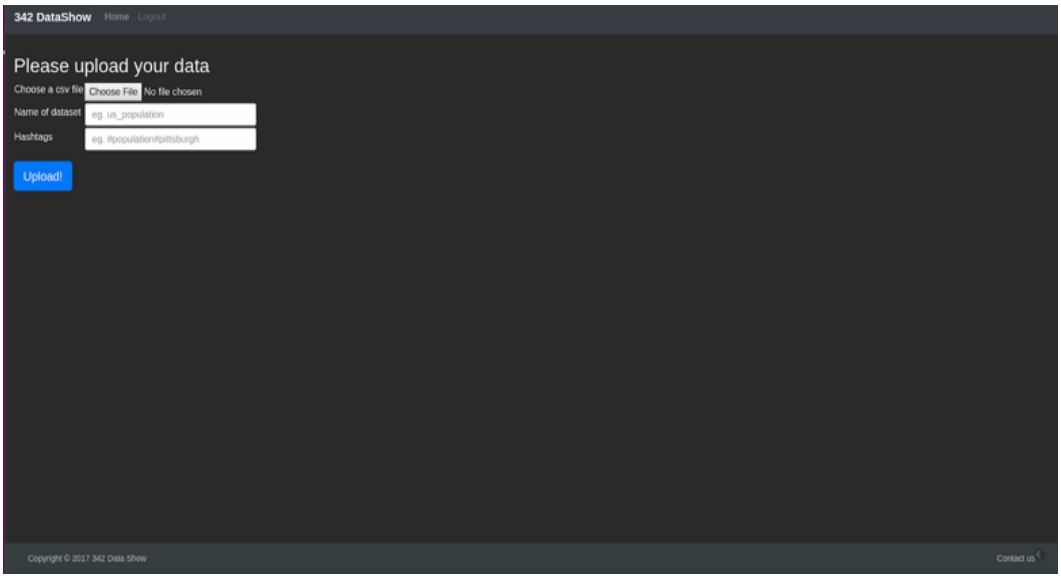


Fig.4.1.3 Upload Data Page

(II) View Map/Change Style

After uploaded data is checked and stored on server, it is ready for view on the 2D map. When data is displayed on the map located on the right side of the

web browser, users are free to directly interact with the map area with actions one may interact with an ordinary online map application. Supported actions include dragging, zooming in/out, etc.

Users can specify the display style of data points to be either circle or square. Display color, point size and stroke width can also be modified using the selection menu and the slider on the left. Also, users can specify the blending mode for the data points.

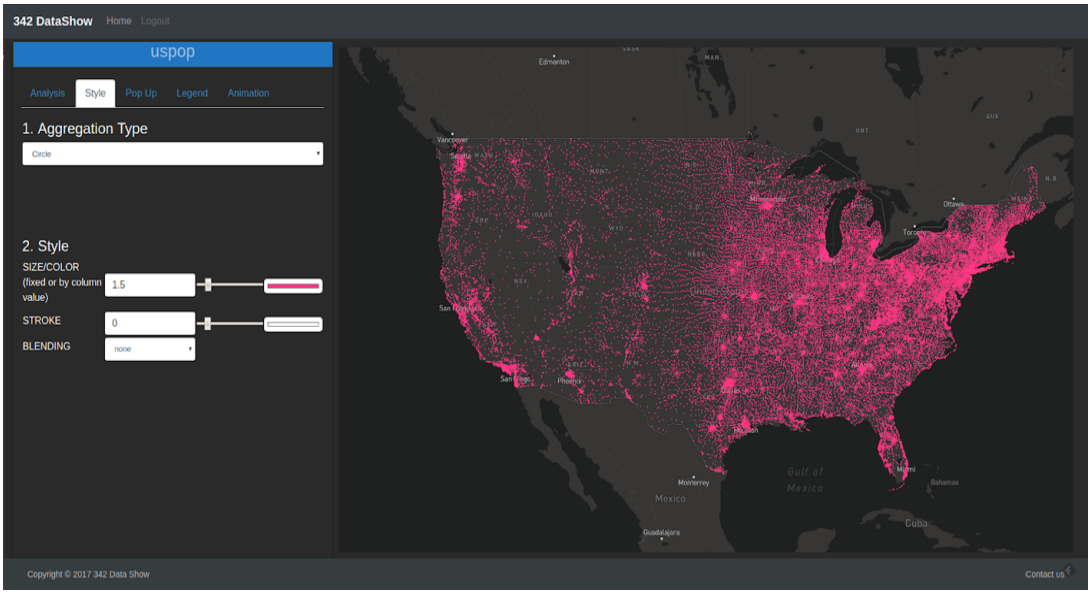


Fig.4.1.4 Change Style Page

4.1.3 Supporting Functionality

To facilitate sharing of analysis, additional supporting features are implemented. These features manage user profiles and sharing of analytic results.

(I) Registration/Login/Logout

Users are required to sign up for an account on the platform in order to use the various data visualization features. This mechanism is to ensure different users will not intervene during visualization process.

(II) Pop-up Function

To further facilitate understanding of the dataset, a pop-up function is added to the visualization platform. After switching to the pop-up tab, users can add desired items to be displayed on the pop-up balloon, specify a style and click on any data point on the map to view the actual data inside a point.

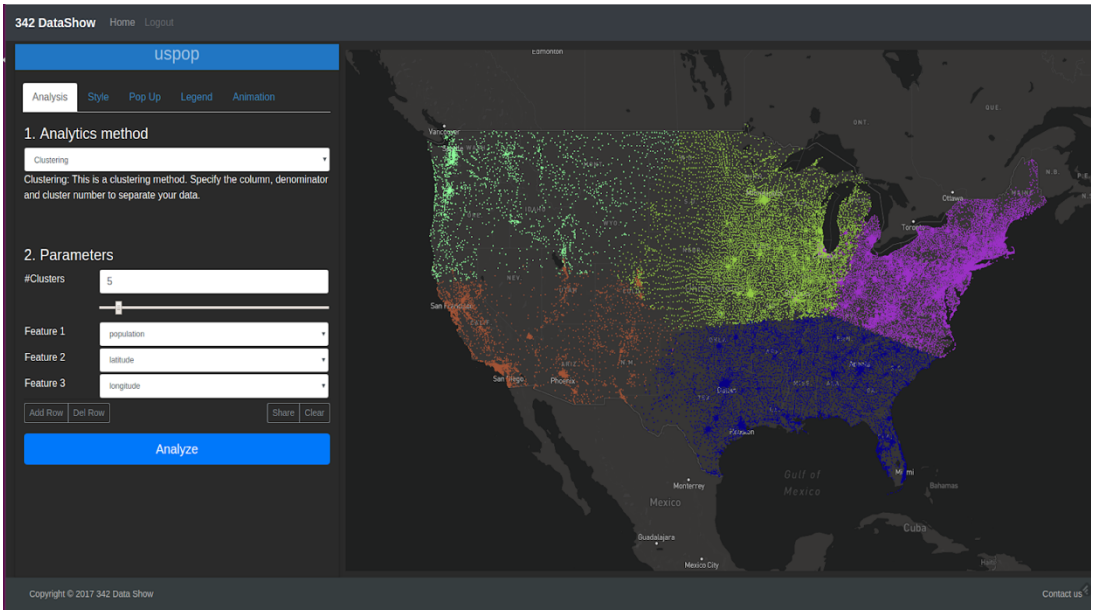


Fig.4.1.5 Cluster Analysis Module

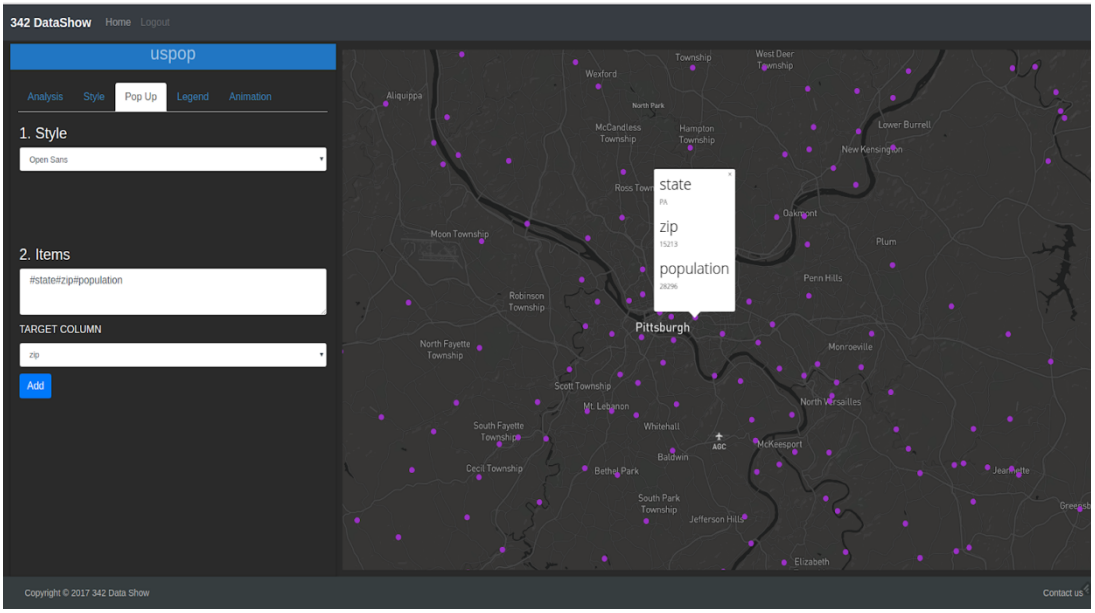


Fig.4.1.6 Pop-up Function

4.2 New Data Clustering Prototype

4.2.1 Observation-Weighted K-Means Algorithm

In addition to the well-known k-means algorithm (Hartigan et al., 1979), where each entry is considered of equal weights, it is possible to consider each data point to be of different importance. This extension of the original k-means algorithm, which is known as the Observation-Weighted K-Means algorithm was introduced in various occasions (Sparks, 1973) (Burkardt, 2011) (Ackerman et al., 2011).

This algorithm allows certain fields of the data points to be considered as weight values, which might represent a measure of importance, a frequency count, etc. During the clustering process, I choose the weighted centroid of all points in a cluster to be the representative Z(J) of this cluster. (Burkardt, 2011)

Z(J) = Sum (all X(I) in cluster J) W(I) * X(I) / Sum (all X(I) in cluster J) W(I)

The objective energy function to be minimized is defined as:

E(J) = Sum (all clusters) Sum (all X(I) in cluster J) W(I) * || X(I) - Z(J) ||^2

(I) Clustering of Weighted Geo-Located Dataset

Using the observation-weighted k-means algorithm described above, I extended the clustering module to accept user-defined weight column that is used for cluster analysis. This column can be any numeric field. If weight field is not assigned, every entry will be considered to have equal weight. As usual, the x and y coordinates are set to longitude and latitude by default, and user can specify a desired number of clusters.

(II) Speeding up Clustering for Unweighted Geo-Located Dataset

Using observation-weighted k-means for clustering opens up an opportunity to speed up the cluster analysis for unweighted dataset. As k-means algorithm has a complexity of O (I * M * N), where M is the size of data, N is the number of desired clusters, and I is the number of iterations before converging. Therefore, I can use this weighted k-means algorithm to speed up cluster analysis. I can achieve this by taking the following steps:

- First aggregate individual data points into groups using frequency count as weight values. One possible approach is to aggregate by coordinate blocks and use the center of mass to represent each group. This is called the pre-processing step. (As shown below)

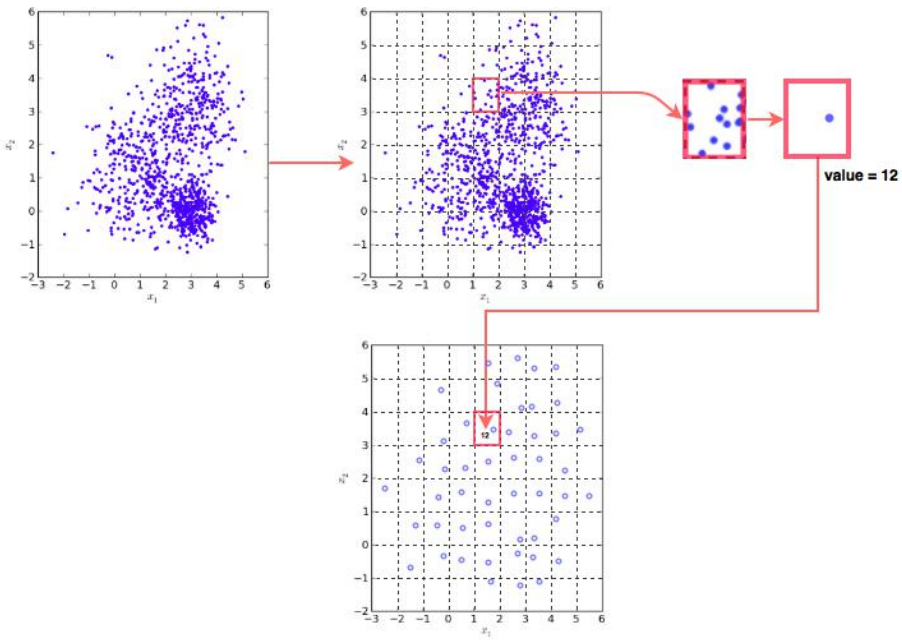


Fig.4.2.1 Data Aggregation

- Run observation-weighted k-means algorithm on aggregated data to get cluster centers and assignment of points to clusters.
- Use the mapping from original data to aggregate data to figure out cluster assignments of the original data points. This is called the post-processing step.
- Display the clustering results on the data visualization platform.

Similar approaches for speeding up the speed the clustering algorithms were brought up in many literatures such as (Davidson et al. 2003) (Salman et al., 2011). It is noticeable that through the steps above, the k-means algorithm can run on a far smaller dataset that can have a significant influence on performance.

(III) Integrated Clustering Module for Unweighted and Weighted Geo-Located Data

In order to provide the users with a unified interface for both weighted and unweighted data, the following logic is used to determine whether pre-process is necessary on the input dataset.

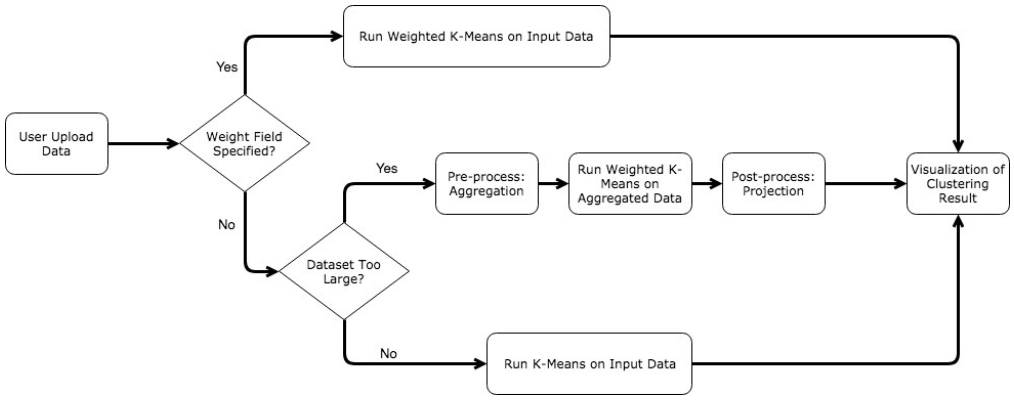


Fig.4.2.2 Workflow for Integrated Clustering Module

(IV) Multi-Dimensional K-Means

To support adding/deleting features dynamically, each feature is projected as one dimension in the feature space. The numerical ranges of different features can be very different from one another, resulting in a non-universal influence factor for each feature. To eliminate this effect, all features will be regularized to the range of 0-100 before clustering.

4.2.2 New Feature: Cluster Analysis

A multi-feature cluster analysis module is embedded in the data visualization platform. This module is more powerful than cluster analysis modules in any current commercial web-based data visualization platforms as it provides a very comprehensive clustering function.

After switching to the analysis tab and choosing the clustering method, users are free to specify different parameters for cluster analysis. Unlike most commercial platforms where clustering can only be calculated based on location data, this platform offers a more generic set of options for cluster analysis. It allows users to specify an arbitrary number of features used in clustering. The features do not have to be location-specific and can be interactively added or removed using the Add Row/Del Row buttons. Furthermore, there is no restriction on the number of clusters desired.

Because we are dealing with data uploaded by non-professional users, there is no guarantee that the location of the data points can be associated to any pre-defined geographical regions, e.g., countries, states/provinces or counties. Considering this fact, two visualizing approaches for clusters are provided:

(I) View clusters as colored data points

One simple approach is to show different clusters as different colors of data points. After clustering is computed, each data point will have a label representing the cluster it is assigned to. This label, which is essentially an

integer, is then mapped to a color that is used to fill the circle representing this data point. Clustering result will be presented as circles with different colors on the map.

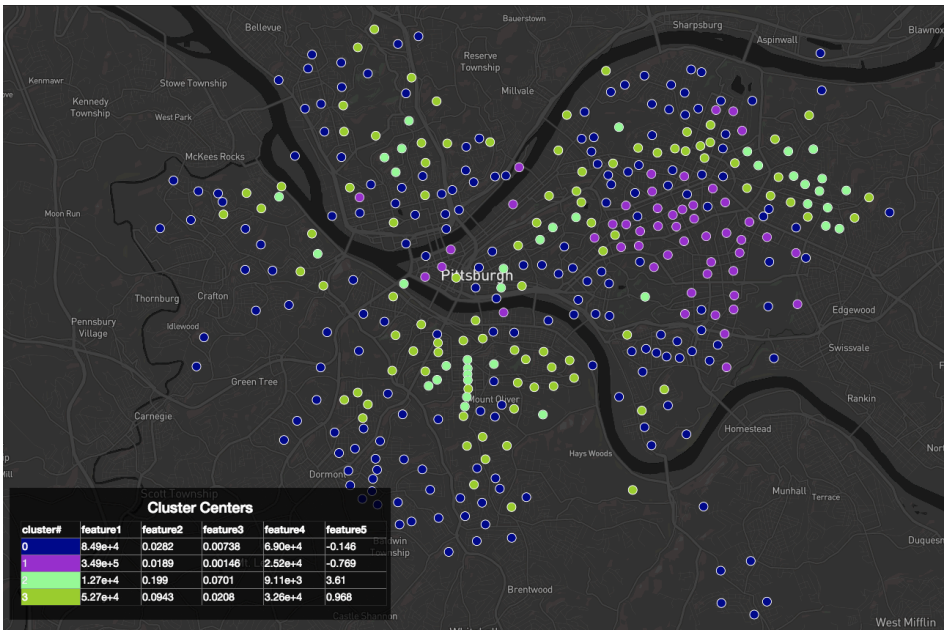


Fig.4.2.3 Point Display of the Clustering Result

(II) View clusters as colored Voronoi regions

Another visualization approach is to show data points as Voronoi polygons. All data points are used to generate a Voronoi tessellation of the region of interest, and each polygon represents the area of influence of the corresponding data point. Each Voronoi polygon is then colored by its cluster assignment. In this way, instead of scattering points across the map, the area of interest is well segmented into different colored regions. In this way, it is easier for users to get a high-level view of clustering result in this region. The SciPy library (SciPy.org, 2016) is used to compute Voronoi cells from data points.

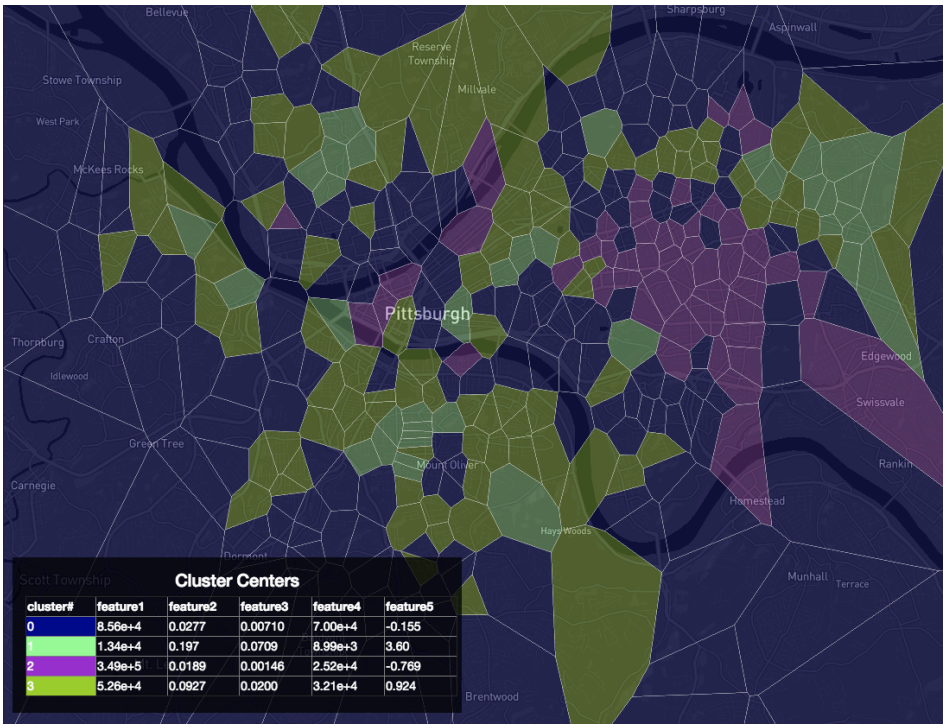


Fig.4.2.4 Regional Display of the Clustering Result

(III) Detailed cluster information display

Mapping from colors to clusters is shown on the bottom left corner of the map area. For each cluster, the center (averaged values of each feature) is listed to help users dig deep into understanding what the clustering really means. The cluster center is a representative of all points in this cluster. All points in the cluster is therefore considered to have relatively the same feature values to the center. Story-telling is therefore enabled by providing this detailed cluster information.

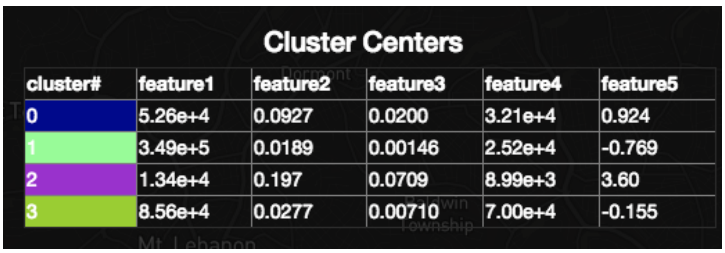


Fig.4.2.5 Detailed Cluster Information Display of the Clustering Result

4.2.3 New Feature: Data Sharing

After setting up proper parameters for data analysis and acquiring desired results, a user can choose to share the map with other users using this platform. Shared maps can be accessed via the “choose existing maps” function in the action selection page. Shared maps will be organized by their hashtags to help new users quickly find their desired data.



Fig.4.2.6 Maps Organized by Hashtags

4.2.4 Platform as a Learning Tool

With the data sharing feature, the datasets will be categorized into different directories based on the hashtags of each uploaded dataset. If the user chooses to share the dataset and the analysis he or she has done with the other users, the dataset and the analysis parameters will be stored in related categories inside the 'existing maps' folder.

When other users who have interests in the same filed or the same location, he or she can go to the specified category inside the 'existing maps' folder to see what others have done. The user might find some useful information to help him (or her) better understand a region or a topic. The user might get inspired on others' datasets and want to use the same data that others' have chosen to share. The user might learn a better way to analysis the datasets from other users' analysis parameters, and thus can do better on his (or her) own datasets.

For example, the user who is interested in the market value of Pittsburgh might find there are existing maps of this city using many helpful datasets, such as 'crime info', 'air quality', 'education level', etc. And the user might find it helpful to get a better understanding about his own topic after exploring the data resource that was already there shared by the previous users.

This new data clustering platform is designed for the users to share their data resource and their ways of thinking with other users. By sharing the information, people can learn from the platform and gain a better understanding in their fields of interests.

4.3 System Overview

4.3.1 Design Choices and Tradeoffs

When developing a web-based data visualization platform, many factors should be taken into consideration to present a user-friendly application. Two key factors to be considered are performance and user interaction. Discussed below are some of the important design choices considered in the development of this platform.

(I) Where does computation happen?

For clustering analysis, the most computation intensive part is computing k-means. Based on the architecture of the platform, there are two options for where k-means is computed:

- K-means can be computed on the server side, by the Django framework. Results will be sent via the internet to user's web browser.
- K-means can be computed on the client side. The algorithm can be implemented in JavaScript to be run by user's web browser. Results are immediately available to user.

Between these two options, the first one is finally chosen. The main consideration is that, running computation intensive procedures on client side can be unreliable. User may experience high CPU consumption while using the platform, which is a highly undesirable artifact. If k-means is implemented on server side, we can use any available library that is computationally efficient and highly reliable. Therefore, k-means is designed to be computed on the server side.

(II) How are data stored?

To store user uploaded data, a backend database is used. Database schema, therefore, is another important choice to be decided. The options are as follows:

- For each individual dataset, a different table will be created with schema reflecting the columns of the uploaded CSV file.
- All datasets are pre-processed into the GeoJson format and stored into a single table.

The first option preserves flexibility for future analytical queries, while the second option is optimized entirely for visualization purposes. Both approaches were implemented. However, after testing, the first option is found to be very slow on large input datasets, because it requires a great number of database manipulations. The second option is eventually chosen.

4.3.2 Tests and Deployment

The entire platform is deployed on an Amazon AWS virtual machine, which can be openly accessed via the internet.

To test the basic functionality of the platform, two correctness tests are conducted, using the 2010 Census Populations by Zip Code (Data.gov, 2010) dataset.

- Population clustering with one feature (population value).
- Population clustering by location (two features, longitude and latitude).

The first test case can be verified by manually examining clustered results. Data points with similar population values should be clustered together.

The second test case can be verified by examining the border of each cluster. If the function is implemented correctly, the borders should exhibit a Voronoi pattern.

Chapter 5 : Application Scenarios

Urban planning has focused largely on the use of social, economic, and political factors in evaluating urban growth and development (Garvin & Garvin, 2002). To make it clear how this new prototype might apply in real world settings, an application scenario in which the user might interact with the prototype is provided and analyzed. However, the prototype is not limited to these categories and users can use it to explore any clustering problem that he cares about.

5.1 City of Pittsburgh Market Value Analysis

In 2016, the Urban Redevelopment Authority (URA) of Pittsburgh has completed the 2016 Market Value Analysis (MVA) for the City of Pittsburgh, which offered 'a unique tool for characterizing markets' based on many parameters for evaluating the residential real estate market (URA, 2016).

The geographic unit is the census block group. Each data point represents a 2010 Census Block. And the column 'geo_boundary' represents the census block ID.

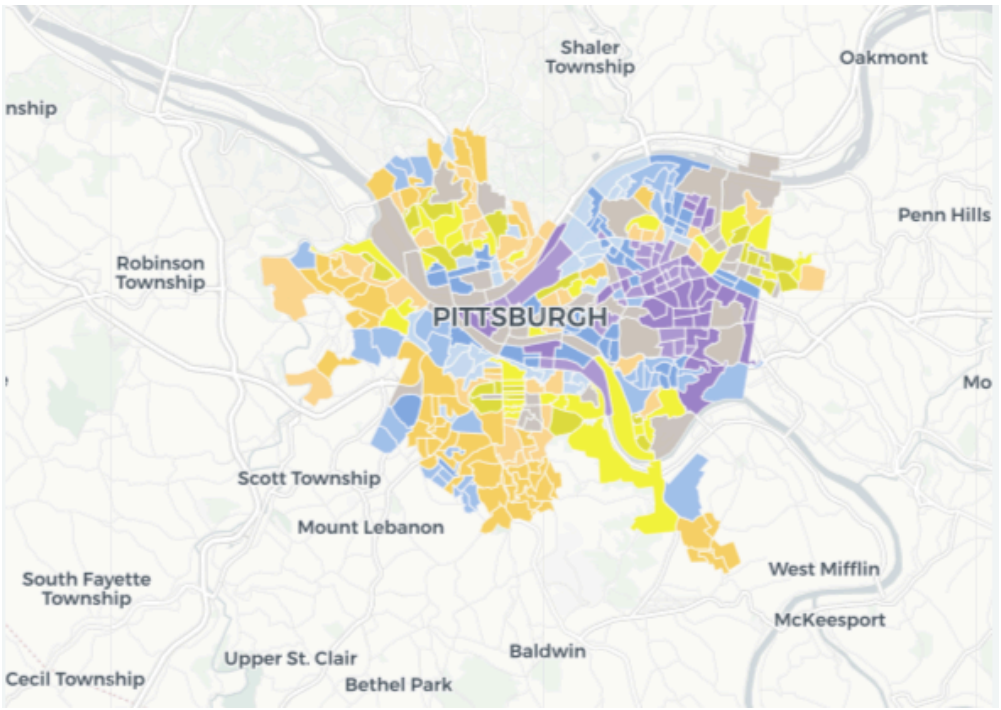


Fig.5.1.1 Demo visualization from URA showing each census block (URA, 2016)

URA believes that by analyzing the MVA data of Pittsburgh, it would benefit the community revitalization, and identify the areas which are the highest demand markets while in the same time 'the areas of greatest distress', and 'the various markets types between'(URA, 2016).

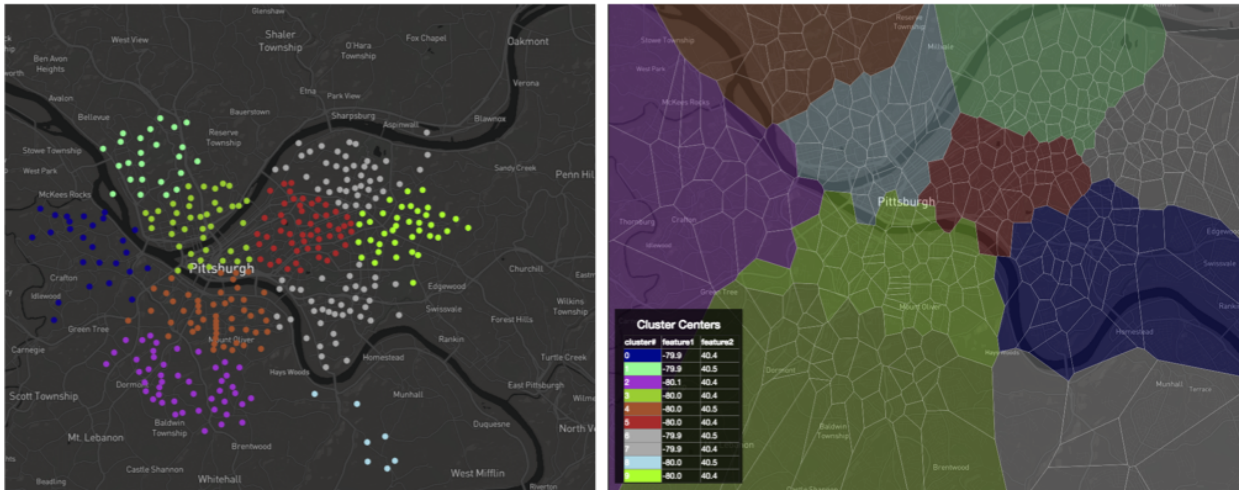
The MVA datasets contain longitude and latitude information for each residential property listed in each row. It also contains the 'Median Residential Sales Price'

between July 2013 and June 2016, the ‘Variance in the Sales Price’ between July 2013 and June 2016, the ‘Percent of Housing Units with Home Owner Occupants’, the ‘Percent of Properties with Assessed Condition Poor or Worse’, the ‘Percent Residential Vacancy as of 2015’, the ‘Index of Distress’ which indicates the composite index of housing age, condition and vacancy, the ‘Percent of Rental Units that Received Housing Choice Voucher or other Subsidy’ and the ‘Density of Housing Units per Square Mile’.

Using the new data clustering prototype, it is simple to do the clustering analysis interactively and efficiently on the MVA datasets to ‘form groups of areas which are similar’ based on the MVA descriptors noted above. By picking different descriptors each time, different clustering results were showed for MVA analysis.

5.1.1 Using Geographic Information as Features

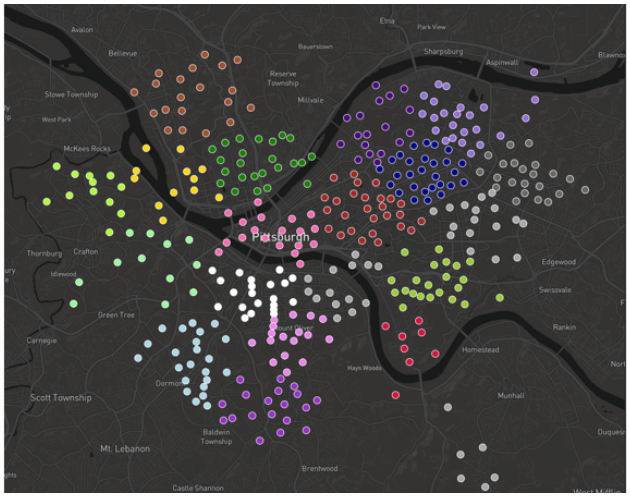
The following figures are the clustering results with two characterized features. Feature one is the ‘Longitude’ of each residential property, while feature two is the ‘Latitude’. The data is clustered into ten groups. Figure (a) is the point display of the result, while figure (b) is the regional display of the same result. The result shows that the MVA datasets can be clustered into n blocks based on the geographic information for each residential real estate property. In this case, n is ten, and the city is zoned into ten large communities accordingly for future discussion.



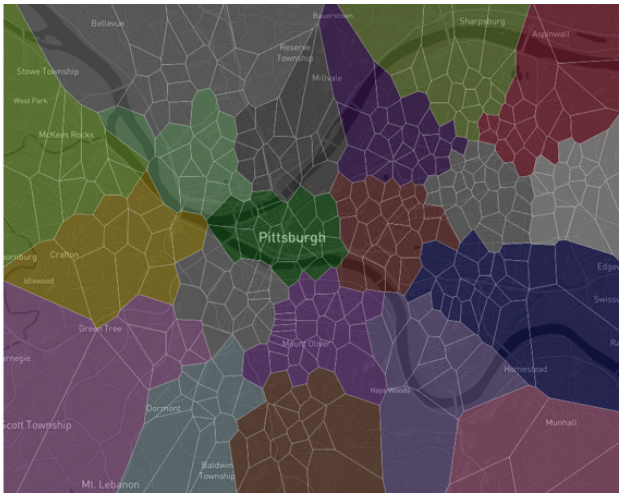
(a) (b)

Fig.5.1.2 Ten Blocks’ Clustering Results for ‘Longitude’ and ‘Latitude’

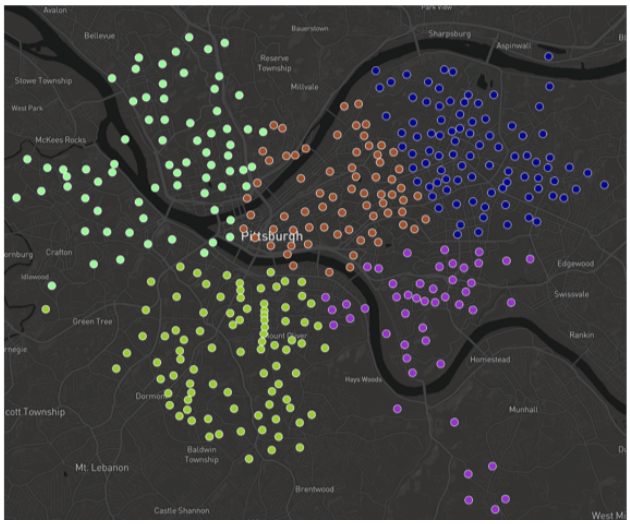
The city can also be zoned into more blocks if smaller communities are needed, or less blocks if it is better for clarification. The following figure (a) and (b) are results for clustering the datasets into 20 groups, while figure (c) and (d) are results for clustering the datasets into 5 groups. And the maximum number of groups is 100 for this system.



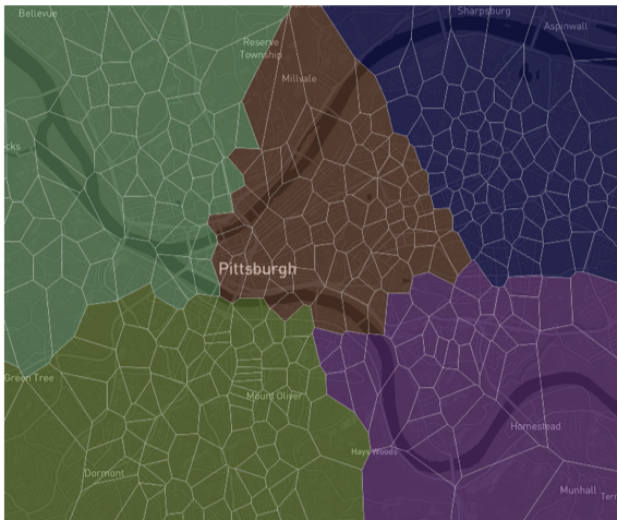
(a)



(b)



(c)



(d)

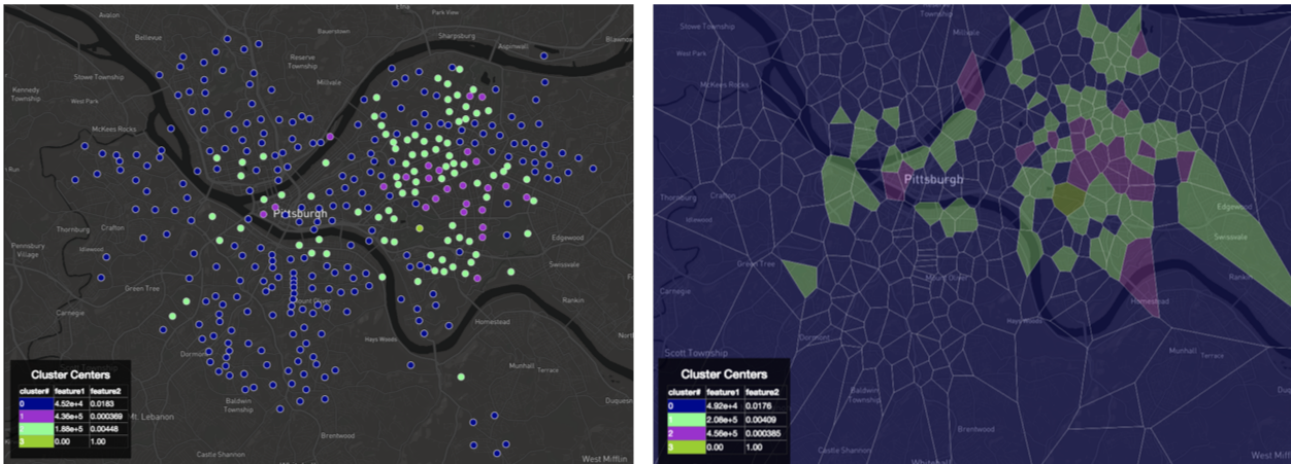
Fig.5.1.3 Clustering Results for 'Longitude' and 'Latitude', (a) and (b) are clustered into 20 groups, while (c) and (d) are clustered into 5 groups

5.1.2 Sale Price & Assessed Condition of the Property

The following figures are the clustering results with two characterized features. Feature one is the 'Median Residential Sale Price' of each residential property, while feature two is the 'Percent of Properties with Assessed Condition Poor or Worse'. The data is clustered into four groups. Figure (a) is the point display of the result, while figure (b) is the regional display of the same result.

The result shows that the condition of the property is highly related to the sale price. The pink areas, which includes S*hadyside, Squirrel Hill North* and *Squirrel Hill* have the lowest percent of properties with poor condition, and the highest average median residential sale price of $436,000. The green areas have the second lowest percent of properties with poor condition, and the second highest average median residential sale price of $188,000. The blue areas where the residential properties are more likely

to be accessed with poor or worse condition, have much cheaper median sale price with the average median sale price of $45,200.



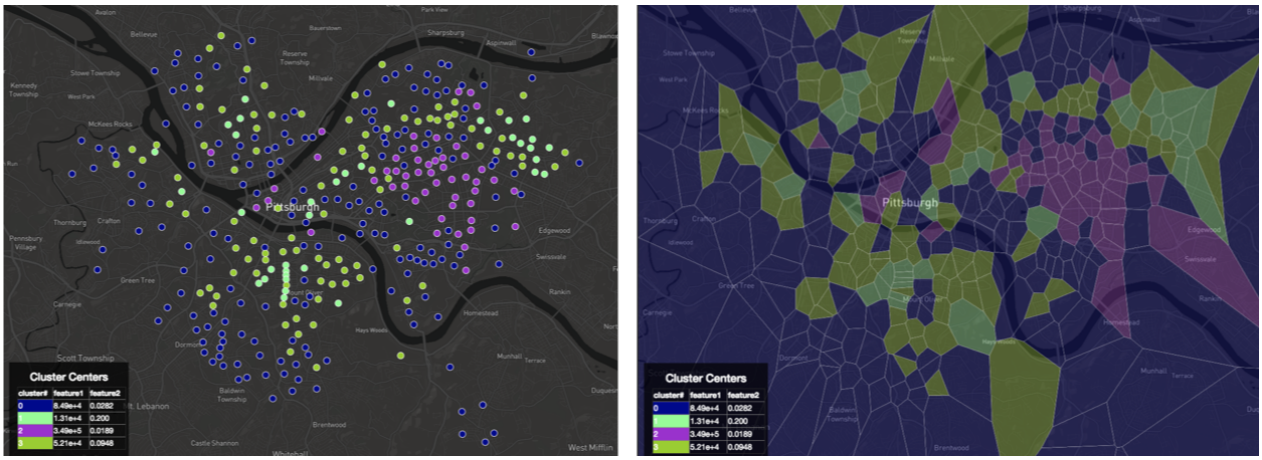
(a) (b)

Fig.5.1.4 Clustering Results for ‘Median Residential Sale Price’ and ‘Percent of Properties with Assessed Condition Poor or Worse’

5.1.3 Sale Price & Percent of Residential Vacancy

The following figures are the clustering results with two characterized features. Feature one is the ‘Median Residential Sale Price’ of each residential property, while feature two is the ‘Percent of Residential Vacancy’. The data is clustered into four groups. Figure (a) is the point display of the result, while figure (b) is the regional display of the same result.

The result shows that the pink areas are the most expensive communities with the average median residential sale price of $349,000, and the percent of residential vacancy is only 1.9%. These communities including *Shadyside, Squirrel Hill North, Squirrel Hill and Squirrel Hill South* are popular and more expensive in the City of Pittsburgh.



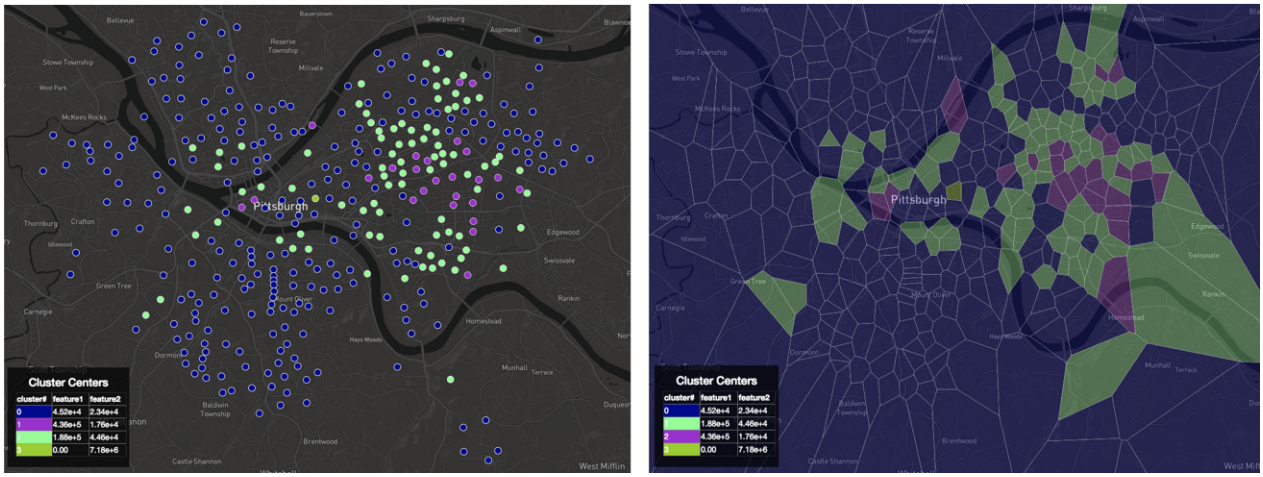
(a) (b)

Fig.5.1.5 Clustering Results for ‘Median Residential Sale Price’ and ‘Percent of Residential Vacancy’

5.1.4 Sale Price & Density of Housing Units

The following figures are the clustering results with two characterized features. Feature one is the ‘Median Residential Sale Price’ of each residential property, while feature two is the ‘Density of Housing Units per Square Mile’. The data is clustered into four groups. Figure (a) is the point display of the result, while figure (b) is the regional display of the same result.

The result shows that the density of housing units is highly related to the sale price. The pink areas, which includes S*hadyside, Squirrel Hill North* and *Squirrel Hill* have the lowest density of housing units, and the highest average median residential sale price of $436,000.



(a) (b)

Fig.5.1.6 Clustering Results for ‘Median Residential Sale Price’ and ‘Density of Housing Units per Square Mile’

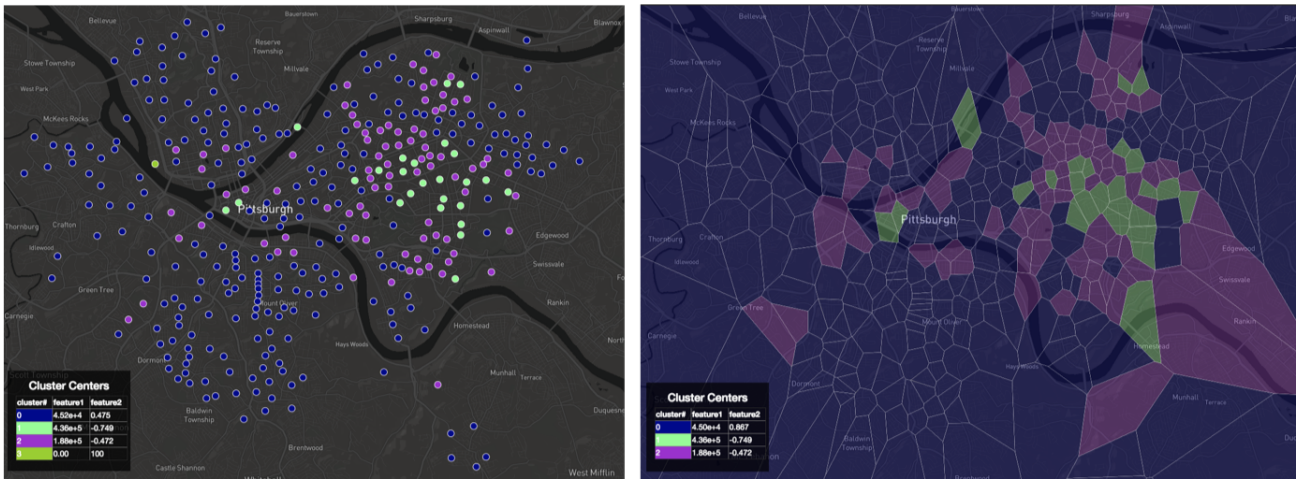
5.1.5 Sale Price & Index of Distress

The following figures are the clustering results with two characterized features. Feature one is the ‘Median Residential Sale Price’ of each residential property, while feature two is the ‘Index of Distress’. The data is clustered into four groups. Figure (a) is the point display of the result, while figure (b) is the regional display of the same result.

The ‘Index of Distress’ indicator is a composite index of housing age, condition and vacancy. Higher values are more distressed and lower values are less distressed. The result shows that the index of distress is highly related to the sale price.

The green areas, which includes S*hadyside, Squirrel Hill North* and *Squirrel Hill* are the least distressed, and the highest average median residential sale price of $436,000. The pink areas have the second lowest index of distress, and the second highest

average median residential sale price of $188,000. The blue areas are the most distressed areas, have much cheaper median sale price with the average median sale price of $45,200.

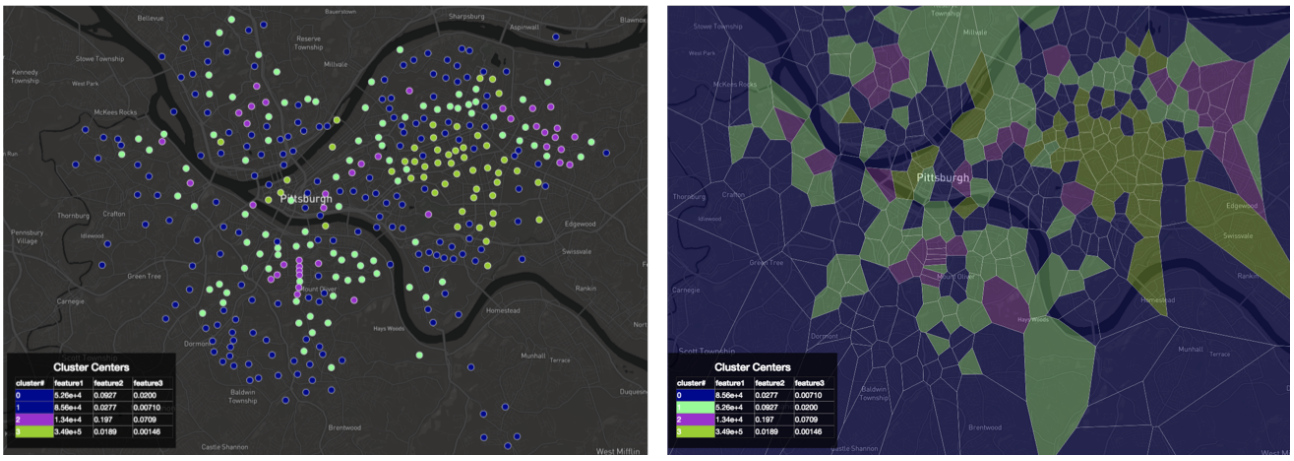


(a) (b)

Fig.5.1.7 Clustering Results for 'Median Residential Sale Price' and 'Index of Distress'

5.1.6 Comprehensive Analysis for Sale Price

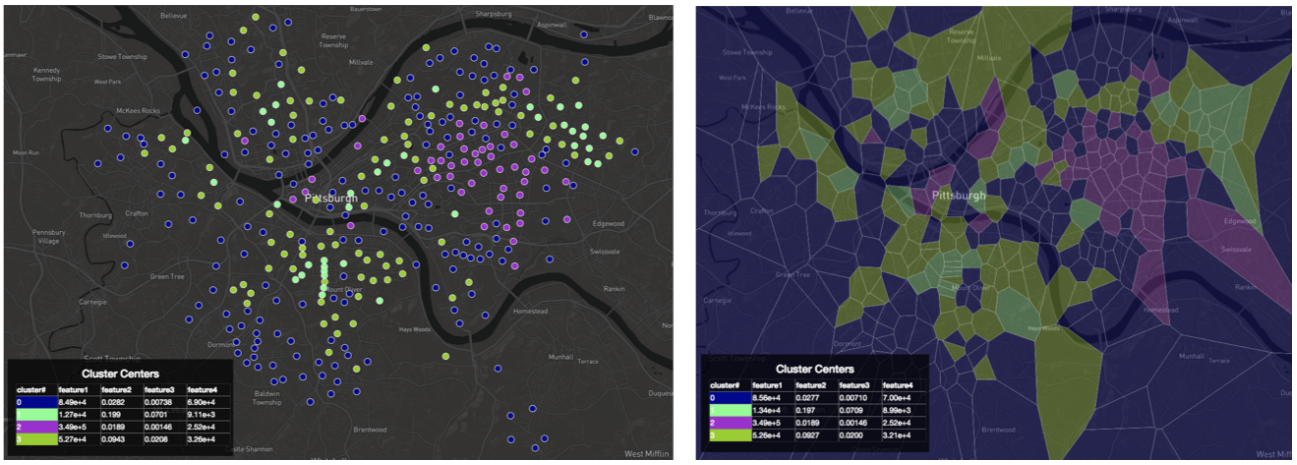
The following figures are the clustering results with three characterized features. Feature one is the 'Median Residential Sale Price' of each residential property, while feature two is the 'Percent of Residential Vacancy' and feature three is 'Percent of Properties with Assessed Condition Poor or Worse'. The data is clustered into four groups. Figure (a) is the point display of the result, while figure (b) is the regional display of the same result.



(a) (b)

Fig.5.1.8 Clustering Results for 'Median Residential Sale Price', 'Percent of Residential Vacancy' and 'Percent of Properties with Assessed Condition Poor or Worse'

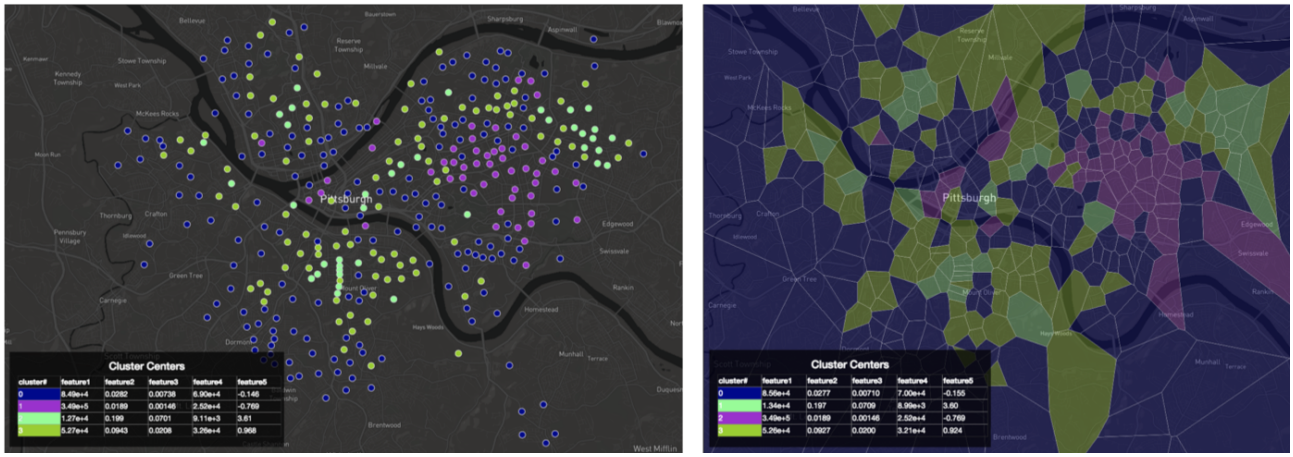
The following figures are the clustering results with four characterized features. Feature one is the ‘Median Residential Sale Price’ of each residential property, while feature two is the ‘Percent of Residential Vacancy’, feature three is ‘Percent of Properties with Assessed Condition Poor or Worse’ and feature four is ‘Density of Housing Units per Square Mile’. The data is clustered into four groups. Figure (a) is the point display of the result, while figure (b) is the regional display of the same result.



(a) (b)

Fig.5.1.9 Clustering Results for ‘Median Residential Sale Price’, ‘Percent of Residential Vacancy’, ‘Percent of Properties with Assessed Condition Poor or Worse’ and ‘Density of Housing Units per Square Mile’

The following figures are the clustering results with five characterized features. Feature one is the ‘Median Residential Sale Price’ of each residential property, while feature two is the ‘Percent of Residential Vacancy’, feature three is ‘Percent of Properties with Assessed Condition Poor or Worse’, feature four is ‘Density of Housing Units per Square Mile’ and feature five is ‘Index of Distress’.



(a) (b)

Fig.5.1.10 Clustering Results for ‘Median Residential Sale Price’, ‘Percent of Residential Vacancy’, ‘Percent of Properties with Assessed Condition Poor or Worse’, ‘Density of Housing Units per Square Mile’ and ‘Index of Distress’

5.1.7 Conclusion for 2016 MVA Data Clustering

The data clustering results from the comprehensive analysis of the 2016 MVA show that 'Percent of Residential Vacancy', 'Percent of Properties with Assessed Condition Poor or Worse', 'Density of Housing Units per Square Mile' and 'Index of Distress' can have significant affection on the residential sales price in the City of Pittsburgh. Several communities which includes S*hadyside, Squirrel Hill North, Squirrel Hill* and *Squirrel Hill South* are key residential communities of the City of Pittsburgh. These communities have a very low percent of residential vacancy, while in the same time have better housing condition and less density of housing units, and the areas are the least distressed ones in the Pittsburgh. Because of these significant advantages, these communities are believed to be most popular ones in the real estate market with highest average median residential sale price.

And the coldest residential communities of the city have the 19.7 percent of vacant residential units. 7.1 percent of the properties in these communities are assessed to have poor or worse condition. These communities are also more distressed than all the other communities in Pittsburgh. Thus, the residential properties in these communities are the cheapest with the average median residential sale price of $13,400.

By conducting the data clustering analysis of the MVA data of Pittsburgh, it is believed that people can get a better insight into *'the variation of the market strength and weakness within and between the traditional neighborhood boundaries'* (URA, 2016). And the results shown in this chapter can be instructive about the potential market changes, and the direction of the market investigation.

Chapter 6 : User Evaluation

For users to evaluate the new prototype, qualified users with urban study or architectural backgrounds are invited to use the prototype and then make their comments and evaluations. The preparations of this user evaluation experiments include questionnaire design, data pool establishment and applying for IRB approval. The user evaluation experiment has four procedures introduced in the second section of this chapter. And the feedbacks and evaluations are gathered and discussed in the last section of this chapter.

6.1 Institutional Review Board

To conduct the user evaluation experiment, the approval from the institutional review board is needed. The material, including the 'Consent Form', 'Pre-test Questionnaire', 'Feedback Questionnaire' and other documentations, was submitted and revised in early-March 2018. And the case was approval in mid-April 2018.

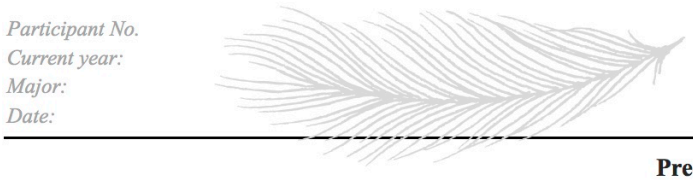
6.2 Procedures of the User Evaluation

There are four steps of the user evaluation experiment. First of all, a pre-test questionnaire is used to identify the backgrounds of the participant. Then, qualified users will be invited to test the new prototype. After the key features being demonstrated to the participant, the participant will use the prototype by himself/herself using the any dataset from the data pool. An after-test questionnaire is used at last to gather the feedbacks and the evaluations from the participant.

6.2.1 Pre-Test Questionnaire for Background Identification

The pre-test questionnaire identifies the academic backgrounds of the participants in their proficiency levels with urban/architectural studies, their proficiency levels with database manipulation, ArcGIS and computer programming. It also asks questions about their interests in using a 'non-threshold (for people with data analysis/architectural/urban design background or experience)' web-based platform for site analysis and urban studies, and their dependency on data visualization tools during the research process. This questionnaire also asks about the reason why the participants prefer to use, or not to use the existing tools for geo-located data analysis and visualization.

The following figure is the pre-test questionnaire for the participants to complete before the user evaluation experiments.



Participant No.
Current year:
Major:
Date:

Pre-test Questionnaire

– User Evaluation on A New Web-Based Interactive Data Clustering Prototype

- Have you had any experiences with architectural/urban studies?
  [ ] Yes [ ] No

- Have you used programmable interface of ArcGIS before?
  [ ] Yes [ ] No

  If yes, how do you describe your competence of using it?
  [ ] Excellent [ ] Skilled [ ] Average [ ] Novice

- Have you used any database language before?
  [ ] Yes [ ] No

  If yes, how do you describe your competence of using it?
  [ ] Excellent [ ] Skilled [ ] Average [ ] Novice

- Have you used any online data visualization platform (Carto, Mapbox etc.) before?
  [ ] Yes [ ] No

  If yes, did you use the platform for architectural/urban/design studies?
  [ ] Yes [ ] No

  If yes, based on your experience, what is the most **pleasant feature** of the platform(s) you used?

  If yes, based on your experience, what is the most **unpleasant feature** of the platform(s) you used?

– User Evaluation on A New Web-Based Interactive Data Clustering Prototype

Fig.6.2.1 The Pre-Test Questionnaire for Background Identification

Qualified participants are selected based on this questionnaire. The participant, who is major in or familiar with architecture design or urban planning, and is unable to use or unfamiliar with database languages and database management, while in the same time does not contain enough programming skills to develop his own data visualization tools for urban planning or site analysis, and is not comfortable with the programmable interface of ArcGIS, is qualified for the user evaluation experiment, and will be invited to continue on the next steps.

6.2.2 Feature Demonstration

The key features of the new web-based interactive data clustering prototype are demonstrated for the qualified participants. The features include data uploading, data clustering analysis, feature adding and deleting, different ways of result display and the resource sharing with other users.

6.2.3 User Participation

The participants are invited to use the new prototype by themselves on any datasets from the data pool. The datasets include census data, market value analysis data and air pollution data. The participants can use these datasets for ethnic clustering, market value analysis and air pollution analysis.

6.2.4 After-Test Questionnaire for Feedback Gathering

After using the new prototype for data clustering by themselves, each participant is invited to complete the feedback questionnaire. The questionnaire is used to gather the evaluations of the usefulness and week points of the prototype. The suggestions and the potential values of the prototype are gathered and discussed during the user evaluation experiment as well.

The following figure is the pre-test questionnaire for the participants to complete before the user evaluation experiments.

Participant No.

Date:

Feedback Questionnaire

– User Evaluation on A New Web-Based Interactive Data Clustering Prototype

- Do you think this prototype is useful for data clustering?
  [ ] Yes [ ] No

- Do you think this prototype is useful for architectural/urban/design studies?
  [ ] Yes [ ] No

- Are you willing to use this prototype in your future architectural/urban/design studies?
  [ ] Yes [ ] No

  Why or why not?

- Which part of this prototype you find **most impressive**? Why?

- Which part of this prototype you think **needs improvement** the most? Why?

- What **potential value(s)** you think this prototype might have in the future?

– User Evaluation on A New Web-Based Interactive Data Clustering Prototype

Fig.6.2.2 The After-Test Questionnaire for Gathering Feedbacks

6.3 Feedbacks and Discussions

Five participants were invited to use the new prototype by themselves on any datasets from the data pool. The results showed that all of the five participants believed the prototype was useful for data clustering, architectural design as well as urban studies. All of the five participants would like to use this prototype in their future architectural design and urban studies. Participants described the prototype as 'pretty straightforward', 'institutive and fast', 'easy to use', 'better user interface than the existing online system' and 'very helpful for those without enough coding skills'. One of the participants commented the prototype as 'it is very useful in site analysis, and I can get an overall understanding of different characteristics of the communities I cared about'.

As for the most impressive part of this prototype, two participants think the prototype runs faster than existing on-line data visualization platforms, three participants like the multiple ways of the display especially the 'regional display mode', and one participant think the user interface is clear, user-friendly and well designed. Two participants think the detailed analysis algorithm on different clusters is very impressive and helpful. And one participant believes the data sharing feature is very impressive and make it easier for users to communicate.

The participants also mentioned the features they think need improvement in the feedback questionnaire. Their comments of improvement include the 'ways of display', 'legend settings' and the 'zoom in difficulties'. Most of these features are improved based on the advice to get a better user experience.

The participants list some potential values of the prototype. Three participants think the data sharing feature is a good idea for designers to share their thoughts with others. One participant believes the prototype can be integrated as part of the existing online platform. Two participants think it is valuable and handy for site analysis and urban studies.

Chapter 7 : Contributions and Future Work

7.1 Research Contributions

This thesis aimed to design and introduce a web-based interactive data clustering prototype for geo-located data analysis and visualization to benefit studies and projects in many practical domains where geo-located datasets are used for data clustering. The prototype is proved to be generic, interactive, non-threshold (for people with data analysis/architectural/urban design background or experience) and user-friendly.

The data clustering platform provided in this research is also helpful and valuable for the users to share their data resource and their ways of thinking with other users. With the data sharing feature, the datasets will be categorized into different directories based on the hashtags of each uploaded dataset. By sharing the information, people can learn from the platform and gain a better understanding in their fields of interests.

By conducting the user evaluation experiment, the usefulness, potential values and the future work were discussed among the potential users and the researchers. The limitations and future direction of this research are better understood and emphasized.

7.2 Limitations and Future Work

(I) Creation of Recommender System

The data clustering prototype has the data sharing feature which allows the users to share their resource and view the existing maps. For the users to learn from others' work more efficiently, a recommender system should be created in the future to recommend existing maps with similar features (including the locations and fields of interests) automatically to all the users uploading their own datasets. This recommender system is predictive service based on the users interests and will make the platform more intellectual and user-friendly.

(II) Improving Data Clustering Guidelines

Some users might be the first time to do data clustering analysis for geo-located datasets. They can have problems in data clustering settings and storytelling based on the results. By providing more detailed and tutorial guidelines for data clustering and sample analysis for the users to follow in the future, it would benefit users who just start their 'data clustering tour' to get a better user experience.

(III) Enriching the Variety of Data Display

The data clustering prototype has two ways of data display, the point display and the regional display. It might be not enough for ambitious users who want the results to be shown in a more personalized way. As pointed out by users with a professional urban design background, a data visualization platform will better suit professional needs if it can support regional display of census blocks. One of the major features to support is importing census block group IDs, which is an widely adopted method of displaying census data.
By enriching the variety of data display in the future, the prototype can be guaranteed to be more valuable for users with stricter display requirements. Furthermore, by making the front-end of the data display feature open-sourced, ambitious users can contribute to the community and share their personalized ways of data display with all the other users. It will give the users more methods to choose from and more freedom to improve the results by themselves.

Bibliography

(IETF), I. E. T. F. (2016). *The GeoJSON Format (RFC 7946)*. Internet Engineering Task Force (IETF). Retrieved from https://tools.ietf.org/html/rfc7946

Apache Software Foundation. (n.d.). *About the Apache HTTP Server Project*. Apache Software Foundation. Retrieved from http://httpd.apache.org/ABOUT_APACHE.html

Chang, R., Wessel, G., Kosara, R., Sauda, E., & Ribarsky, W. (2007). Legible cities: Focus-dependent multi-resolution visualization of urban relationships. *IEEE Transactions on Visualization and Computer Graphics*, *13*(6), 1169–1175. http://doi.org/10.1109/TVCG.2007.70574

Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., & Chau, M. (2004). Crime data mining: a general framework and some examples. *Computer*, *37*(4), 50–56. http://doi.org/10.1109/MC.2004.1297301

Data.gov. (2010). 2010 Census Populations by Zip Code. Retrieved April 6, 2018, from https://catalog.data.gov/dataset/2010-census-populations-by-zip-code

Django Software Foundation. (n.d.). The Web framework for perfectionists with deadlines | Django. Retrieved April 6, 2018, from https://www.djangoproject.com/

Estivill-Castro, V., & Vladimir. (2002). Why so many clustering algorithms. *ACM SIGKDD Explorations Newsletter*, *4*(1), 65–75. http://doi.org/10.1145/568574.568575

Frias-Martinez, V., & Frias-Martinez, E. (2014). Spectral clustering for sensing urban land use using Twitter activity. *Engineering Applications of Artificial Intelligence*, *35*, 237–245. http://doi.org/10.1016/j.engappai.2014.06.019

GAO, Y., MENG, X., & LAO, X. (2013). Cluster analysis-based vulnerability risk zoning of lightning disaster in Hainan Island [J]. *Journal of Natural Disasters*, *1*, 24.

Garvin, A., & Garvin, A. (2002). *The American city: What works, what doesn't*. McGraw-Hill New York, New York;

Harris R, Sleight P, W. R. (2005). Geodemographics: Neighbourhood Targeting and GIS. *Chichester: John Wiley and Sons*.

Lansley, G., Wei, Y., & Rains, T. (2015). Identifying the Major Traits of Ethnic Clustering in England and Wales from the 2011 Census. *The Consumer Data Research Centre Working Paper Series*, *1*.

Mapbox. (n.d.). Mapbox. Retrieved April 6, 2018, from https://www.mapbox.com/

Markušić, S., & Herak, M. (1998). Seismic Zoning of Croatia. *Natural Hazards*, *18*(3), 269–285. http://doi.org/10.1023/A:1026484815539

Morales-Esteban, A., Martínez-Álvarez, F., Scitovski, S., & Scitovski, R. (2014). A fast partitioning algorithm using adaptive Mahalanobis clustering with application to seismic

zoning. *Computers & Geosciences*, *73*, 132–141. http://doi.org/10.1016/J.CAGEO.2014.09.003

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. Retrieved from http://jmlr.org/papers/v12/pedregosa11a.html

SciPy.org. (2016). scipy.spatial.Voronoi — SciPy v0.18.1 Reference Guide. Retrieved April 16, 2018, from https://docs.scipy.org/doc/scipy-0.18.1/reference/generated/scipy.spatial.Voronoi.html

Scitovski, S. (2018). A density-based clustering algorithm for earthquake zoning. *Computers and Geosciences*, *110*, 90–95. http://doi.org/10.1016/j.cageo.2017.08.014

Severance, C. (2012). JavaScript: Designing a Language in 10 Days. *Computer*, *45*(2), 7–8. http://doi.org/10.1109/MC.2012.57

Thota, L. S., Alalyan, M., Khalid, A. O. A., Fathima, F., Changalasetty, S. B., & Shiblee, M. (2017). Cluster based zoning of crime info. In *2017 2nd International Conference on Anti-Cyber Crimes, ICACC 2017* (pp. 87–92). IEEE. http://doi.org/10.1109/Anti-Cybercrime.2017.7905269

Toole, J. L., Ulm, M., González, M. C., & Bauer, D. (2012). Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing - UrbComp '12* (p. 1). New York, New York, USA: ACM Press. http://doi.org/10.1145/2346496.2346498

Tryon, R. C. (1939). Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality. *Edwards Brothers.*

URA. (2016). Market Value Analysis of the City of Pittsburgh.

Uskul, A. K., & Platt, L. (2014). A note on maintenance of ethnic origin diet and healthy eating in understanding society. Retrieved from https://www.econstor.eu/handle/10419/91673

Yin, J., Yin, Z., Zhong, H., Xu, S., Hu, X., Wang, J., … Hu, X. (2011). Monitoring urban expansion and land use/land cover changes of Shanghai metropolitan area during the transitional economy (1979–2009) in China. *Environ Monit Assess*, *177*(177). http://doi.org/10.1007/s10661-010-1660-8