Extracting Meaningful Insights on City and Zone Levels Utilizing US Open Government Data

Samaa Gazzaz[†]

Computer Science Department, King Abdulaziz University, Jeddah, Saudi Arabia, sgazzaz@kau.edu.sa

Praveen Rao

Department of Computer Science and Electrical Engineering, University of Missouri-Kansas City, Kansas City, Missouri, USA, raopr@umkc.edu

[†] This work was done while the author was affiliated with the University of Missouri-Kansas City.

ABSTRACT

It is estimated that merely 4% of the world's population reside on US soil. Remarkably, 43% of the entire population of prominent websites are hosted in the United States (Fig. 1). Even though most data content on the Web is unstructured, the US government has had big contributions in producing and actively releasing structured datasets related to different fields such as health, education, safety and finance.

Aforementioned datasets are referred to as Open Government Data (OGD) and are aimed at increasing the structured data pool in conjunction with promoting government transparency and accountability. In this paper, we present a new system "OGDXplor" which processes raw OGD through a well-defined procedure leveraging machine learning algorithms and produces meaningful insights.

The novelty of this work is encompassed by the collective approach utilized in developing the system and tackling challenges. First by addressing arising challenges due to data being collected and aggregated from heterogeneous sources that otherwise would have been impossible to acquire as a comprehensive unit. moreover, classification and comparisons are drawn on a much finer level that we refer to as zone level. Zones are the areas encompassed and defined by zip codes and are seldomly used in classifying and extracting insights as presented here. OGDXplor facilitates comparing and classifying zones located in different cities or zones within an individual city.

The system is presented to end-users as a web application allowing users to elect zones and features relevant to their use case. Results are presented in both chart and map formats which aids the decision-making process.

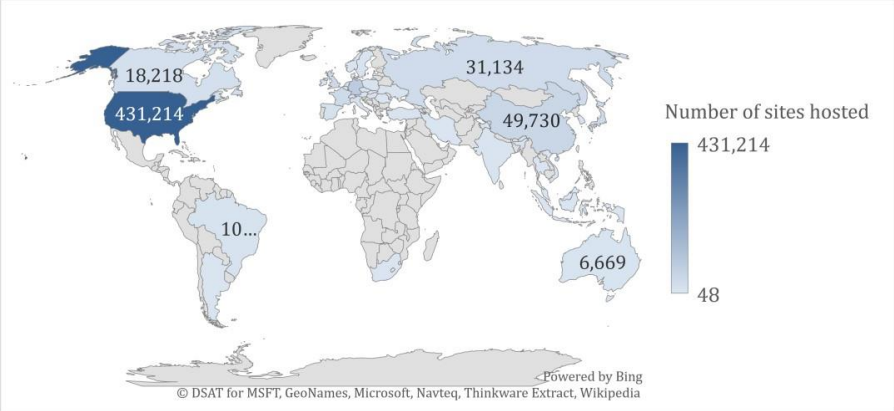

Figure 1: Top 100 web hosting countries with respect to the top 1 million influential websites.

CCS CONCEPTS

• **Information systems~Clustering** • **Information systems~Web applications** • **Information systems~Clustering and classification** • Information systems~Data cleaning • Mathematics of computing~Dimensionality reduction • Mathematics of computing~Cluster analysis • Software and its engineering~Process validation • Software and its engineering~Use cases

KEYWORDS

Data science, open data, clustering, feature selection.

1 Introduction

Publishing Open Government Data (OGD) has been a growing subject of interest among governments within the last decade. The US is considered the largest influential country in producing data on the Web (followed by Germany hosting only 8%)[14]. It is estimated that the use of OGD when developing applications and services can yield $3 trillion in income across global economy [16]. This would yield better decision-making, trend-recognition and prediction of future events [16].

Providing access to OGD promotes the involvement of citizens within their low and high-level governments. In addition, it affirms the transparency and accountability of said governments. Nonetheless, when it comes to OGD, the datasets are usually burdened with weaknesses whether in regards to its completion, accuracy or conforming to a unified form of publication. The release of such datasets does not guarantee the availability of means to interpret, visualize and analyze this data. Although federal and state governments within the United States contribute extensively to the body of open data [20], the datasets remain inadequate in facilitating critical decision-making processes. Moreover, due to the existence of various data sources and the absence of a regulating body for open government data, accuracy and conformation challenges arise.

In this paper, we present a new approach exploiting OGD and machine learning algorithms aiming at producing a user-friendly application OGDXplor enabling users to gain insights from the datasets. First, data is collected from multiple sources and dataset weaknesses are addressed. Next, irrelevant features are eliminated, and data is aggregated and clustered based on relevant features. Finally, the system features include clustering zones and cities based on relevant features acquired from feature selection. Results are presented to the user in easy-to-follow formats aiding decision making and insight extraction. Those insights will enable users to learn more about different zones and cities comparing data related to health, education, safety, and more.

The paper is organized as follows, in section 2 we discuss the related work and the different challenges faced when dealing with OGD. In section 3, our approach in tackling those challenges, system flow and methods used are discussed. In section 4, evaluation of OGDXplor's accuracy and results of the system are presented. Finally, in section 5, we conclude the paper.

2 Related Work and Challenges

2.1 Related Work

Even though the open data concept is relatively immature [20], there has been an abundant number of research applications based on open data sources. Moreover, the movement towards utilizing OGD expanded when hundreds of national and local governments started releasing OGD portals [20]. In [21], researchers discuss utilizing "open-access satellite data" in the field of biodiversity research. Open data is transformed by applying different techniques and extracting meaningful information from raw input. In addition, numerous research opted to gathering datasets from a variety of commercial "non-government" sources optimizing the benefits of analysis and visualization of data [10][12]. Nonetheless, most of the literature is fixated at explaining the open data initiative, its advantages and disadvantages, and how beneficial it could be if adapted in the right manner [9][6][8][16][2]. On the other hand, we seldom come across a system that is built on heterogeneous OGD gathered from diverse government agencies and structured into a meaningful system.

One project to be highlighted is Data USA [5] in which researchers gathered OGD and implemented a visualization system for extracting facts about areas in the US. Data USA utilizes a collection of datasets from varying government sources to create one comprehensive website that delivers a user-friendly application where the use of the data is optimized [1].

Although it introduced a solution to the existing problem with multi-source open datasets [8], Data USA does not provide pattern recognition in similarities between multiple cities/zones, future possible occurrences, and recommended actions for decision-makers. In OGDXplor on the other hand, we introduce a system model

that utilizes both heterogeneous OGD and machine learning techniques in extracting meaningful insights, recognizing patterns among cities and zones and facilitating the decision-making process.
This submission version of your paper should not have headers or footers, Baldassare (2000). It should remain in a one-column format—please do not alter any of the styles or margins.

2.2 Challenges

OGD portals offer huge potential when it comes to insightful understanding of the trends behind the data enabling an informed decision process. Unfortunately, while obtaining and processing raw OGD, several challenges arise with respect to the data collection process, understanding the meaning of the data in question, and processing heterogeneous data through the same pipeline. These challenges are summarized in figure 2. In the following sections, we focus on discussing the most commonly faced challenges and their implications.

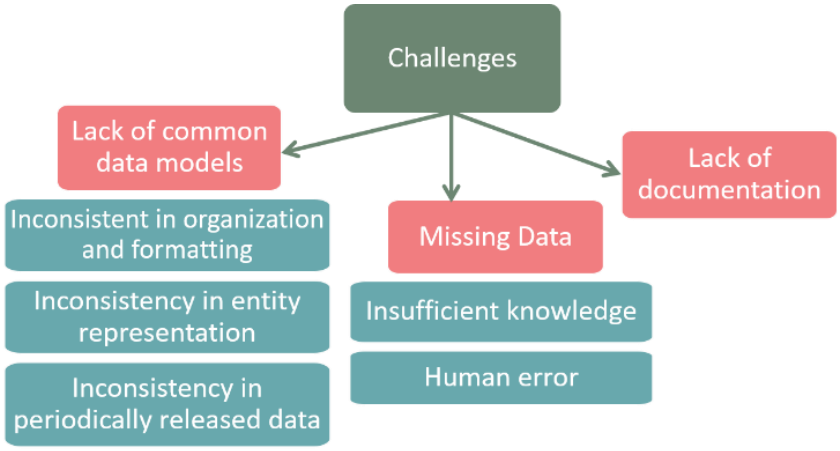

Figure 2: Challenges associated with collecting OGD especially from heterogeneous sources.

2.2.1 Lack of Common Data Models.

One of the most occurring challenges is the lack of common data models. As a result of the multi-source data collection process, data models are recognized to be very inconsistent from source to source. For example, when collecting crime data released by Los Angeles city and Chicago city data portals, we identify the inconsistency in organization and formatting. Data is labeled differently and organized based on different features and properties. Both datasets cannot be combined seamlessly without manual alteration.
Moreover, inconsistency in entity representation is another common challenge. We can look at the crime datasets from both Kansas City and Chicago to immediately recognize that inconsistency. Even though both datasets are concerned with public safety and crime information, we notice the vast difference of entity (represented by a row) interpretation in those datasets. In the Chicago crime data, each entity involves information about the case number, primary type, data and description of a crime. In this case, that information implies that each entity represents a crime. On the other hand, in Opendata KC, each entity is described by information such as involvement, race, sex and age, which in return implies that each entity represents a person involved with a crime (whether a suspect or a victim).
Finally, each source has different release period for each dataset where, for example, one source could release data every year grouped by month while the other releases every quarter grouped by location.

2.2.2 Missing Data.

Another frequently occurring problem is missing data. Whether it is the result of insufficient knowledge about a specific feature or simply human error when entering data, missing information does not only impair the full understanding of the information provided, but also hinders the ability to infer and predict future trends in an unbiased fashion. Insufficient information can occur while generating or entering the data. For example, there are cases where zip code information was not attainable when entering data resulting in entries

such as 99999, 00000 or XXXXX as the zip code value. In addition, incomplete knowledge when the data was being generated results in leaving out attribute values that appear as missing data in published dataset.

3 Our Approach

Initially, datasets are collected from heterogeneous sources such as local governments and privately-owned businesses. In order to address datasets' issues, it is crucial to first handle each dataset separately. During which, we address missing data and the abundance of irrelevant features using feature extraction. Afterwards, datasets are collectively aggregated and merged to generate a coherent dataset that describes a wide range of features. Finally, clustering the cities and zipcode areas (i.e. zones) based on similarities depending on features that can be specified by end users. An overview of OGDXplor is visualized in figure 3. In this section, we discuss in depth, the process and the approach.

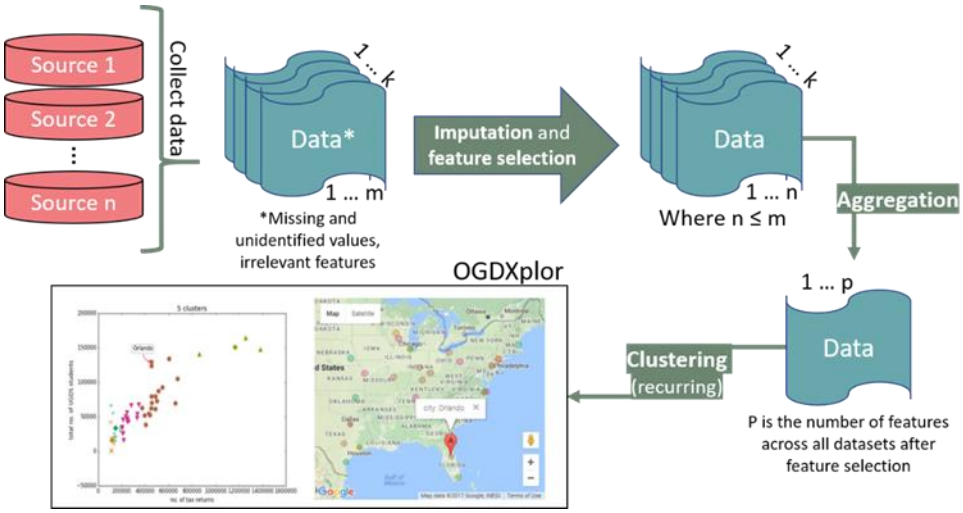

Figure 3: OGDXplor system overview.

3.1 Data Collection and Preparation

In order to establish a solid foundation of the system, datasets are collected from OGD portals with good documentation and consistency in release periods. These collected datasets (Table 1) play a great role providing the system with feature that will be useful to end users. Datasets collected from the department of energy (https://energy.gov/) mainly contained information regarding the rates of utility companies within proximity of a city or a zip code. Those included investor and non-investor owned companies in addition to the service type and commercial/industrial/residential rates.

Moreover, datasets collected from the department of education (https://www.ed.gov/) were comprehensive nationwide statistics about colleges and universities in the US. That information includes more than 7800 colleges and encompasses more than 40 attributes regarding each college. From the department of agriculture (https://www.agriculture.gov/), we collected datasets which included information regarding areas and nearby farmer's markets nationally. This information can be an important factor in many decisions such as area to live or start a local produce market or restaurant.

The department of treasury (https://www.treasury.gov/) provides valuable information regarding taxes filed by tax payers nationwide. This information includes counts of all individually/joint filed taxes, number of dependents, in addition to other data all mapped to zip code areas in the US. Finally, datasets collected from the department of defense (https://www.defense.gov/) included information about the residency of military personnel within the US. This dataset also provides information about age/gender/racial demographics of the enlisted recruits. All collected data is in tabular form. Data is transformed into JSON format for its flexibility preparing for aggregation based on zip codes.

Table 1: Data sources and brief description of content.

Data source	Time period	Number of attributes	Number of entries	Brief Description of content
Department of Energy	2014-2015	9	34k+	Information regarding rates of utility companies
Department of Education	2013-2015	40+	7800+	Extensive statistics about nationwide colleges
Department of Agriculture	2013	2	440k+	Listing of areas and nearby farmer's market
Department of Treasury	2013	100+	27k+	Taxes filed nationwide and filing information
Department of Defense	2010	30+	940+	Defense military recruits enlisted

3.2 Feature Selection

Feature selection is defined as the election of the attributes that most closely represent the whole dataset fairly, even when other attributes are missing. Usually, feature selection is used for dimensionality reduction and pattern recognition in a dataset distribution [22].

The most prominent technique for dimensionality reduction is Principal Component Analysis (PCA), where the resultant features are the outcome of the mapping to the lower level space [22]. On the other hand, our intentions in this application are different since we aim to select a subset of the existing features rather than find a mapping to a new lower dimension. Principal Feature Analysis (PFA) [22] is an adaptation of PCA that allows the retention of previously existing features even after the reduction of dimensionality. As the first step of PFA, the covariance matrix is calculated from the original dataset such that each entry in the resulting matrix is defined as follows:

$$\rho_{ij} = \frac{E[x_i x_j]}{E[x_i^2]E[x_j^2]}$$

Next, we compute the principal components as in PCA and the eigenvalues of the covariance matrix. The retained variability must be established before choosing the subspace dimension. Then, we cluster the data using K-means and use the Euclidean distance to decide where each data point resides. Finally, for each cluster, obtain the corresponding feature that closely represent that cluster and consider this feature as a Principal Feature. The resulting is a list of the most relevant features.

3.3 Data Imputation

There is extensive research in the area of data imputation, and we can categorize data imputation techniques to: mean substitution, regression and K-Nearest Neighbor imputation. In mean substitution, we calculate the mean of all the values in the same feature and impute the result value in all missing cells. This technique is the fastest, but it imposes risk of introducing bias. Regression imputation utilizes the trend analysis of existing values and predicts the missing value based on the trend. This technique becomes expensive as the size of the dataset increases. In addition, it is mostly used to impute datasets that are missing values in a single feature. K-Nearest Neighbor (KNN) technique only considers k entities out of the whole dataset in imputing the missing value. Those k entities are usually chosen based on similarity to the entity with the missing value. Next, the values in the k entities are averaged, resulting in the imputed value.

For the purposes of this research, we utilize KNN as it does not introduce the kind of bias that mean substitution introduces, nor is computationally expensive as regression. KNN algorithm can be generally used in multiple applications such as estimation, classification and imputation [18]. In the case of imputation, the choice of the number of nearest neighbors to consider is very critical. As a rule of thumb, it is preferred to consider $k=\sqrt{n}$ where n is the number of entities in the dataset [18]. Considering $\sqrt{n}$ entities as the nearest neighbors to reference when imputing missing data ensures that we only consider entities that are similar to the entity whose missing field we are trying to impute.

3.4 Clustering

Since the beginning, our goal was to deliver a system that enables users to compare and differentiate cities and zones upon features of their selection. In order to provide that ability to distinguish between the different areas, clustering is utilized where areas are grouped based on similarity. Clustering is perfect for our dataset since it is used as part of unsupervised learning. In order to cluster, we need to select the "optimal" number of clusters desired. Choosing the optimal k is a broad research area where multiple techniques have been developed. The most famous yet is the Gap statistic [19]. In this approach, they utilize the within-cluster dispersion to decide the estimated number of clusters from a clustering algorithm's results [19]. Where Dr is the sum of all data points in a cluster and Wk is the within-cluster sum of squares around the center of the cluster, we calculate Gap statistic as:

$$D_r = \sum_{i,i' \in C_r} d_{ii'}, \qquad W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

$$Gap_n(k) = E_n^*\{\log(W_k)\} - \log(W_k)$$

We show how the Gap statistic optimizes the number of clusters in the following example. To cluster zones in El Paso, Texas, based on two features: number of tax returns and number of dependents, we first calculate the Gap statistic. Estimating the number of clusters k over k = 2, 3, 4, 5, and 6, the result is shown on the left in figure 4 where the Gap value is high when k=3. As sown on the right in figure 4, clustering over 3 groups gives a clear boundary to each group of zones within El Paso.

After figuring out the best value for k for a specific configuration, we start the clustering process via k-means algorithm. In here, we employ the Lloyd's algorithm which implements k-means iteratively to converges to local minimum in lowest amount of time:

$$C_k = \{x_n : ||x_n - \mu_k|| \leq all\ ||x_n - \mu_l||\}$$

$$\mu_k = \frac{1}{C_k} \sum_{x_n \in C_k} x_n$$

The notation denotes that each cluster C_k is a set of points x_n such that the distance from a mean is minimized. The symbol μ_k represents the mean of cluster k.

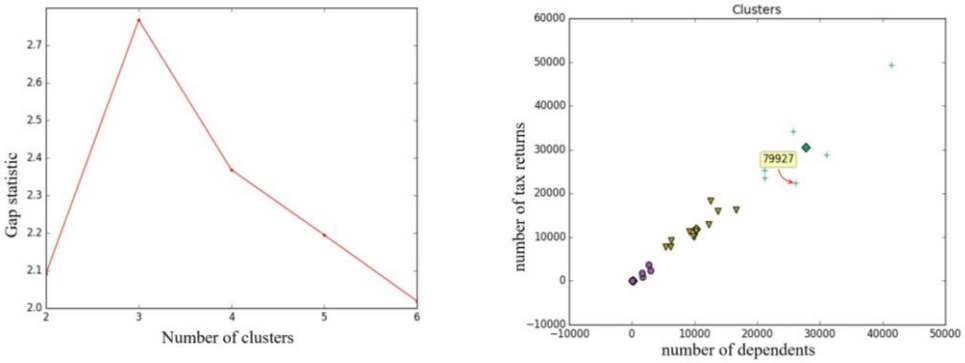

Figure 4: Gap statistics results for k = 2, 3, 4, 5, 6 (left) and Sample date clustered based on number of tax returns and dependents (right).

4 Evaluation and Results

Cluster analysis is a broad science concerned with evaluating clustering technique and evaluating the optimal number of clusters. When evaluating clustering validity, three validation criteria can be explored. External criteria, which consider a pre-specified structure when evaluating outcomes of a clustering algorithm [17].

Relative criteria, which evaluate a clustering algorithm's results by comparting them to results from other clustering algorithms [17]. Internal criteria, evaluating the outcomes of a clustering based on a calculated value involving entities in the dataset within the evaluation process [17]. We will utilize internal criteria for cluster validity.

For internal criteria, there are two main features that are considered when validating: compactness and separation. Compactness refers to ensuring the minimization of the distance among data points within a single cluster (e.g. variance can be used to calculate compactness) [3]. Separation criteria favors higher distances between cluster centers (i.e. distinct cluster assignments) [3]. We can calculate the separation among two clusters by measuring the distance between: the closest data points, the furthest data points, or the centers of the two clusters. This is referred to as single linkage, complete linkage, and comparison of centroids, respectively [3]. In order to evaluate internal criteria, multiple validation indices were introduced to evaluate the compactness and separation levels of clusters.

In order to evaluate our clustering approach, involving the use of the Gap statistic as input to k-means clustering, we compare the recommended number of clusters provided by the Gap statistic with other internal criteria-validation indices. The goal of this evaluation is to detect how accurate the clustering is when the number of clusters is determined by the Gap statistic. This is done by comparing the estimated number of clusters with the results from the following indices: Silhouette index [15] [13], Calinski-Harabasz index [11] [4], Dunn index [11] [7] and Davis-Bouldin index [11] [13].

The experiment is divided into two categories: clustering over features elected by feature selection, and clustering over randomly chosen features. In the first, features are the most relevant representation of data; while in the second, selection is simulating user activity. The hope is to recognize that our approach performs well under both circumstances.

The results of the experiments are shown in table 2 and figure 5. In the figure, results of validating clustering over randomly selected features are indicated in orange and feature selection results are in blue. Values presented are averaged out based on cross validation. Data was clustered into 2, 3, 4, 5, and 6 clusters respectively, where each time the value of the index is calculated. The results shown indicate a pattern of conformation among the indices. For example, when features are randomly collected, the recommended number of clusters by 3 of the 4 indices is 5 clusters. This value is the same value resulting from utilizing the Gap index in our approach. On the other hand, there is a slight misalignment between the validity metrics recommendation of optimal number of clusters when clustering using features selected by feature selection and the Gap statistic recommendation. In this case, the value recommended by the Gap statistic is 3 where none of the other indices recommend that value; however, the values recommended are close. Considering the results of our experiments, we conclude that the clustering technique and approach of the system is performing efficiently and yielding sufficient clustering.

Table 2: Validation of clustering using internal validity indices vs Gap statistic.

Validity metrics	Feature selection		Features selected randomly	
	Number of clusters	Best value	Number of clusters	Best value
Silhouette index	4	0.927	5	0.819
Calinski-Harabasz index	6	2099.5	6	2328.2
Dunn index	4	0.345	5	0.168
Davis-Bouldin index	4	0.344	5	0.161
Gap statistic	3	2.775	5	1.407

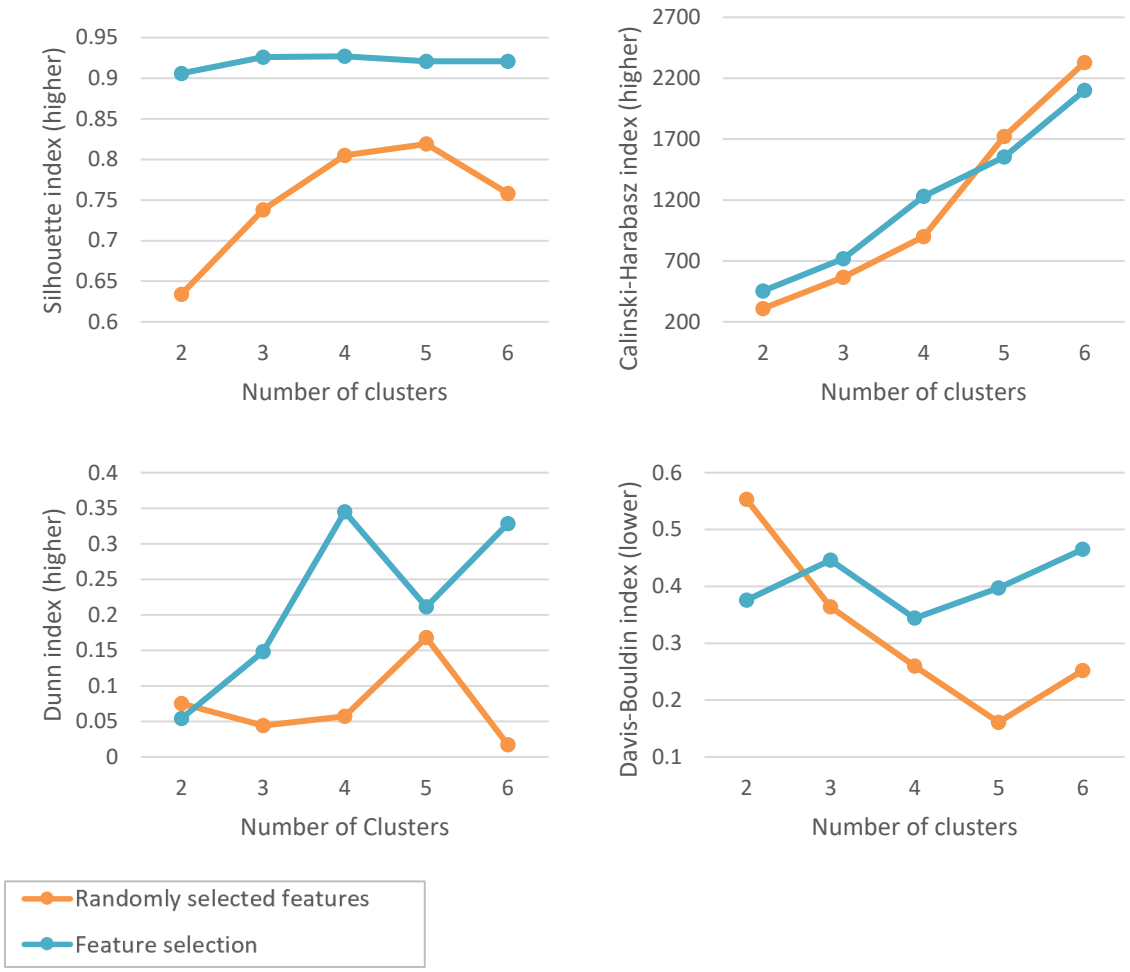

Figure 5: Cluster validation results using Silhouette index, Dunn index, Davis-Bouldin index and Calinski-Harbaz index.

4.1 Application Use Case Example

A prospective college student is interested in attending a college in one of the largest 100 cities in the US but would like to know the ratio of college students to the general populations in these cities. Moreover, they would be interested in knowing how Orlando, FL ranks in that comparison and what are similar cities to Orlando considering these properties.

Through OGDXplor, the student is able to provide the city name then specify the features she is interested in. Next, the system clusters the cities over the selected features and produce these two views to the user (shown in figure 6).

In this case, the system identified five clusters considering the total number of undergraduate students and the number of tax returns (representing the general population). In figure 6 (left), we see that the city of Orlando is located somewhat in the middle cluster while Miami is located in the upper-right cluster.

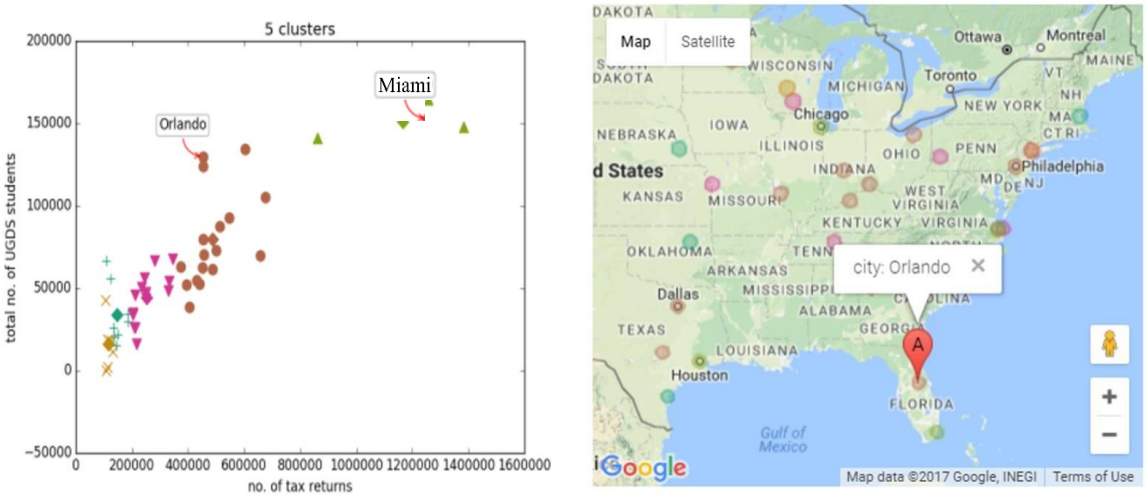

Figure 6: System result when clustering over number of tax returns and total number of undergraduate students.

5 Conclusion

In this paper we presented OGDXplor, a system that utilizes open government data and machine learning in producing visualization of the cities and zones within the US. Regardless of the dataset's weaknesses, we were able to tackle the challenges and chosen zip code information to be the distinctive key to each area. Feature selection enables the retention of the most relevant features, thus saving time and ensuring relevant results to the user. Finally, clustering was a big part of our approach in addition to utilizing Gap statistic to estimate the best possible number of clusters. In the end, the user can choose features they want to compare, choose the area and view charts and the map showing the grouping and comparing results. This system is helpful for multiple applications and of great help in the decision-making process.

REFERENCES

[1] Ann Perrin and César Hidalgo. 2016. Data USA: The Most Comprehensive Visualizations of U.S. Public Data. (April 2016). Retrieved April 3, 2017 from https://www.media.mit.edu/sponsorship/getting-value/collaborations/datausa

[2] Barbara Ubaldi. 2013. Open government data: towards empirical analysis of open government data initiatives. OECD Working Papers on Public Governance 22. OECD Publishing, Paris, France.

[3] Gordon Linoff and Michael Berry. 2011. Data Mining Techniques For Marketing, Sales and Customer Support. (3rd. ed.). John Wiley & Sons, Inc., USA.

[4] Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods 3.1 (1974), 1-27.

[5] Data USA. 2016. Retrieved from https://datausa.io

[6] Felipe Gonzalez-Zapata and Richard Heeks. 2015. The Multiple Meanings of Open Government Data: Understanding Different Stakeholders and Their Perspectives. Government Information Quarterly 32, 4 (October 2015), 441–452.

[7] James Bezdek and Nikhil Pal. 1995. Cluster validation with generalized Dunn's indices. Proceedings of the Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems, Dunedin, 1995, pp. 190-193.

[8] James Hendler, Jeanne Holm, Chris Musialek, and George Thomas. 2012. US government linked open data: semantic.data.gov. IEEE Intelligent Systems 27, 3 (May 2012), 25-31.

[9] Judie Attard, Fabrizio Orlandi, Simon Scerri, and Sören Auer. 2015. A systematic review of open government data initiatives. Government Information Quarterly 32, 4 (October 2015), 399–418.

[10] Klaus Ackermann, Eduardo Reyes, Sue He, Thomas Keller, Paul van der Boor, and Romana Khan. 2016. Designing policy recommendations to reduce home abandonment in Mexico. In Proceedings of

the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York, NY, USA, 13-20.

[11] Ujjwal Maulik and Sanghamitra Bandyopadhyay. 2002. Performance evaluation of some clustering Aagorithms and validity indices." IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 12, (December 2002), 1650–1654.

[12] Muhammad R. Khan and Joshua E. Blumenstock. 2016. Predictors without borders: behavioral modeling of product adoption in three developing countries. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York, NY, USA, 145-154.

[13] Petrovic, Slobodan. 2006. A comparison between the silhouette index and the Davies-Bouldin index in labelling ids clusters. Proceedings of the 11th Nordic Workshop on Secure IT Systems (NordSec 06). Sweden. (Oct 20, 2006).

[14] Pingdom. 2012. The US Hosts 43% of the World's Top 1 Million Websites. (July 2012). Retrieved April 2, 2017 from http://royal.pingdom.com/2012/07/02/united-stateshosts-43-percent-worlds-top-1-million-websites/

[15] Peter J. Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, (1987), 53–65.

[16] Sharon S. Dawes, Lyudmila Vidiasova, and Olga Parkhimovich. 2016. Planning and designing open government data programs: an ecosystem approach. Government Information Quarterly 33, 1 (January 2016), 15–27.

[17] Sergios Theodoridis and Konstantinos Koutroumbas. 2008. Pattern Recognition (4th ed.). Academic Press.

[18] Saravanan Thirumuruganathan. A Detailed Introduction to K-Nearest Neighbor
(KNN) Algorithm. (May 2010). Retrieved July 19, 2017 from
https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introductionto-k-nearest-neighbor-knn-algorithm/

[19] Robert Tibshirani, Guenther Walther and Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63, 2, (November 2000), 411–423.

[20] Tim Davies. 2013. Open Data Barometer - 2013 Global Report. (October 31, 2013), 24-35. Open Data Institute, World Wide Web Foundation. Retrieved April 3, 2017 from
http://www.cocoaconnect.org/sites/default/files/publication/Open-Data-Barometer-2013Global-Report.pdf

[21] Woody Turner, Carlo Rondinini, Nathalie Pettorelli, Brice Mora, Allison Leidner, Zoltan Szantoi, Graeme Buchanan, Stefan Dech, John Dwyer, Martin Herold, Lian Koh, Peter Leimgruber, Hannes Taubenboeck, Martin Wegmann, Martin Wikelski, and Curtis Woodcock. 2015. Free and open-Access satellite data are key to biodiversity conservation. Biological Conservation 182 (February 2015), 173–176.

[22] Yijuan Lu, Ira Cohen, Xiang Sean Zhou, and Qi Tian. 2007. Feature selection using principal feature analysis. In Proceedings of the 15th ACM international conference on Multimedia (MM '07). ACM, New York, NY, USA, 301-304.