ZUHA AHMAD

DS-670: Capstone: Big Data & Business Analytics

Dr. Jaume

22 April 2017

**A Journey to Smart Cities:**

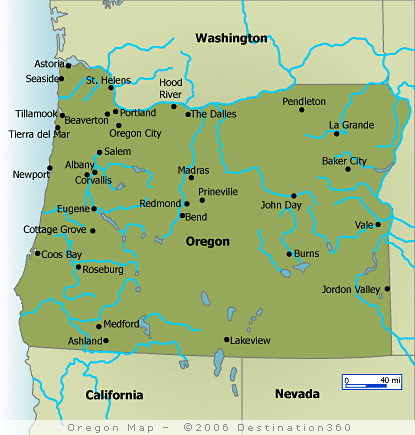
**Exploring Crime Analytics Across Portland, Oregon Using K-Means Clustering**

**CONTRIBUTION**

Crime will remain to be erratic and irregular in many societies. However, with the power of analytics, we can hope to project the estimated outcome of future crimes and implement strategies to decrease all crimes, if not, some. The objective of the project is to conduct crime forecasting by performing a regression analysis, specifically a K-means clustering, based in the city of Portland, Oregon in the United States of America. The notion is to create a methodology to visualize the crimes in Portland and then further analyze the areas that are impactful to the community. By doing so, we can detect which areas are significant and how it can improve under law enforcement. The strategy is to use programming tools such as Zeppelin notebook to import and organize the comma separated values (CSV). Afterwards, R, will be used to generate a superlative algorithm and it will showcase findings through data visualization with an application called Shiny from *RStudio*. The dataset will be utilized is derived from the United States National Institute of Justice (USNIJ). Although, there are contributing variables, the main variables that will be considered over in this dataset are accidents, burglary, and shootings, will be taken into consideration. This would help to facilitate practical crime prevention solutions that correspond to specific times and places in Oregon.

**STATE-OF-THE-ART**

According to the Oxford Dictionary, a crime is *an action or omission which constitutes an offence and is punishable by law*. A crime can occur at any place at any time which can be done discretely, unless the individual is caught in the act. Since there is always a possibility that crime will occur in the future by a series of chances, it would be practical and sensible to collect and review numerous amounts of information with data (i.e. qualitative, quantitative). Steps can be taken further as to create data visualization such as plotting, marking, and navigating “hotspots” of crime locations in maps and dashboards. The important aspect, above all else, is to predict the outcome and likelihood of the crime occurring in a different time, place, region or interestingly, repeated crime in the same time, place, and region. This leads to the emergence of crime analytics where overall, law enforcements analyze and effectively respond to crime patterns, series and trends by enabling data sharing, pattern analysis, predictive analytics, crime mapping and reporting. Consequently, this will establish law enforcement and police departments into action by constructing crime prevention strategies, raising public safety awareness, and feasibly modifying laws to governmental (city, state, federal) policies.

**BACKGROUND**

Oregon is a [state](https://en.wikipedia.org/wiki/U.S._state) located in the [Pacific Northwest](https://en.wikipedia.org/wiki/Pacific_Northwest) region on the [West coast](https://en.wikipedia.org/wiki/West_Coast_of_the_United_States) of the [United States](https://en.wikipedia.org/wiki/United_States). It is one of the most geographically diverse states in the United States, marked by volcanoes, abundant bodies of water, dense evergreen and mixed forests, as well as high deserts and semi-arid shrub lands (Jewell & McRae, 2014). Oregon's economy is mainly powered by various forms of agriculture, fishing, and hydroelectric power due to its landscapes and waterways. Portland is a port and the largest city in the state of Oregon, USA. The city covers 145 square miles and had an estimated population of 632,309 in 2015, making it the 26th most populous city in the United States (U.S. Census Bureau). According to the U.S. Census Bureau, Oregon's population as of 2015 is 4,028,977 with the Metropolitan Statistical Area (MSA) being 2,389,228. This leaves roughly 60% of Oregon's population residing within the metropolitan area. A dense population such as Portland would more than likely have high crime rates. Portland crime statistics indicate that crime is overall decreasing such as violent crime and property. Although improvements are being made, high crime rate indexes that occurred in 2012 for instance happened recently and no matter the predictions, future crimes remain to be random and changeable.

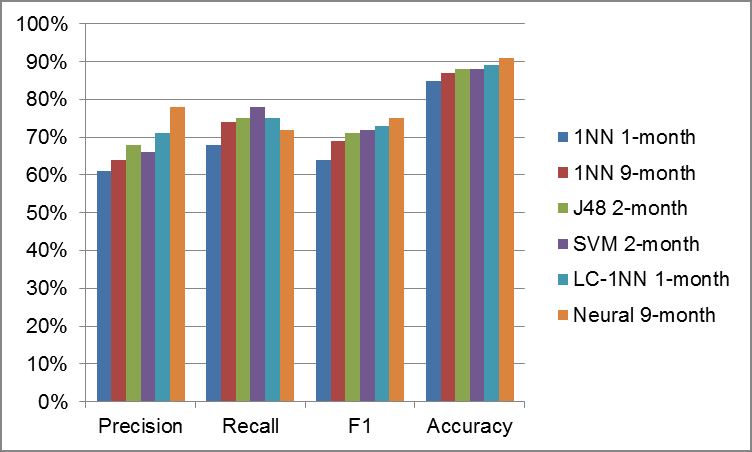
**WORK BY COMPETITORS**

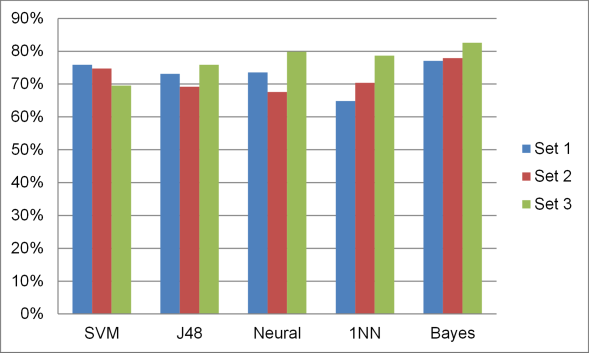
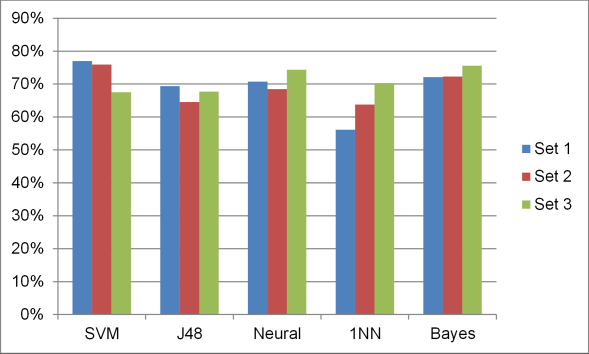
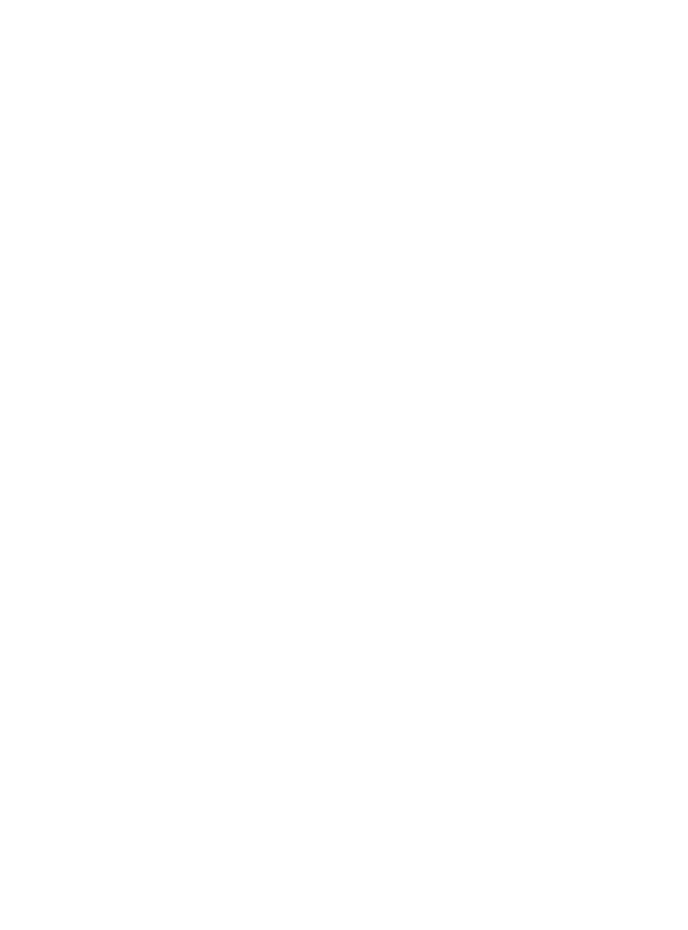
Yu, C. H., Ward, M. W., Morabito, M., & Ding, W. (2011, December). Crime forecasting using data mining techniques. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on* (pp. 779-786). IEEE.

The competitor’s article that I have chosen is called ‘*Crime Forecasting Using Data Mining Techniques’*. I find the article to be compelling and applicable towards my research of crime forecasting. The competitor states that crime is ‘neither systemic nor entirely random’, but it can be classically ‘unpredictable’ which happens to be the case when combating with crime. There are several aims that the competitors had wanted to accomplish. The goal of their project was to explore a methodology for reliably predicting the location, time, and/or likelihood of future residential burglary. The dataset was derived from the United States National Institute for Justice (NIJ). Although the crimes were located in the Northeast, due to the sensitivity of the data; the name of their city was not specifically mentioned. First, they discussed how to generate architected data sets from original crime records. The architected data sets contain the aggregated counts of different types of crimes and related events as categorized by the city’s police department. Second, they had an ensemble of data mining classification techniques (i.e. SVM, J48, 1NN, Neural Networks, Naïve Bayes) that were chosen to perform the crime forecasting. Finally, they analyzed which classification approach is potentially the best method for predicting whether residential burglary will happen. In their results, they had listed out three findings having to do with spatial knowledge. First, there was success of the simple 1NN classifier modified with location constraint. It turned out that finding the most similar circumstance within the same neighborhood proved more effective than finding it within the entire city. Second, even though Neural Network is an ideal classifier, the Naive Bayes classifier had yielded better results in the F1 graph which is the total weight of precision and recall. Finally, the 24-by-20 grid data showed success measures that were consistently higher when using lower resolution data set which is due to each grid cell exhibiting a broader spatial knowledge.

**Best Overall Classifier Performance**

Accuracy 9-month Leave-One-Month-Out (LOMO)





F1 9-month Leave-One-Month-Out (LOMO)

Figure 1. Overall classification results. Best performance of classifier using different training sets

Figure 2. Accuracy and F1 scores using SVM, J48, Neural, 1NN, and Bayes.

**DATA**

The datasets are based on the locations listed in calls-for-service (CFS) records provided by the Portland Police Bureau (PPB) for the period of March 1, 2012 and had updated through February 28, 2017. However, I had collected the dataset from March 1, 2012 to 2016. The data is from the United States National Institute of Justice and it is based on a challenge called ‘Real-Time Crime Forecasting Challenge’ In this table, they have included crime categories, code (which had been re-categorized as “other”), and translation. They also had excluded data due to sensitive information and focus.

| Table 1: USNIJ – Crime Category Definitions – CFS Code and Translation | | |
| --- | --- | --- |
| CFS Category | Code | Translation |
| Burglary |  |  |
| BURGP | BURGLARY – PRIORITY \*H |
| PROWLP | PROWLER |
| Street Crime |  |  |
| ASSLTP | ASSAULT –PRIORITY *Note: This code initially was listed erroneously​ as "ASSLTT"* |
| ASSLTW | ASSAULT WITH WEAPON \*H |
| DISTP | DISTURBANCE – PRIORITY |
| DISTW | DISTURBANCE – WITH WEAPON \*H |
| GANG | GANG RELATED |
| ROBP | ROBBERY – PRIORITY \*H |
| ROBW | ROBBERY – WITH WEAPON \*H |
| SHOOTW | SHOOTING – WITH WEAPON \*H |
| SHOTS | SHOTS FIRED |
| STABW | STABBING WITH WEAPON \*H |
| THRETP | ​THREAT - PRIORITY |
| THRETW​ | ​THREAT - WITH WEAPON \*H |
| VICE | VICE-DRUGS, LIQUOR, PROSTITUTION, GAMBLING |
| Theft of Auto |  |  |
| RSTLN | ROLLING STOLEN \*H |
| VEHREC | VEHICLE RECOVERED |
| VEHSTP | VEHICLE STOLEN – PRIORITY |
| All CFS | This category includes all CFS including those in the above categories. | |

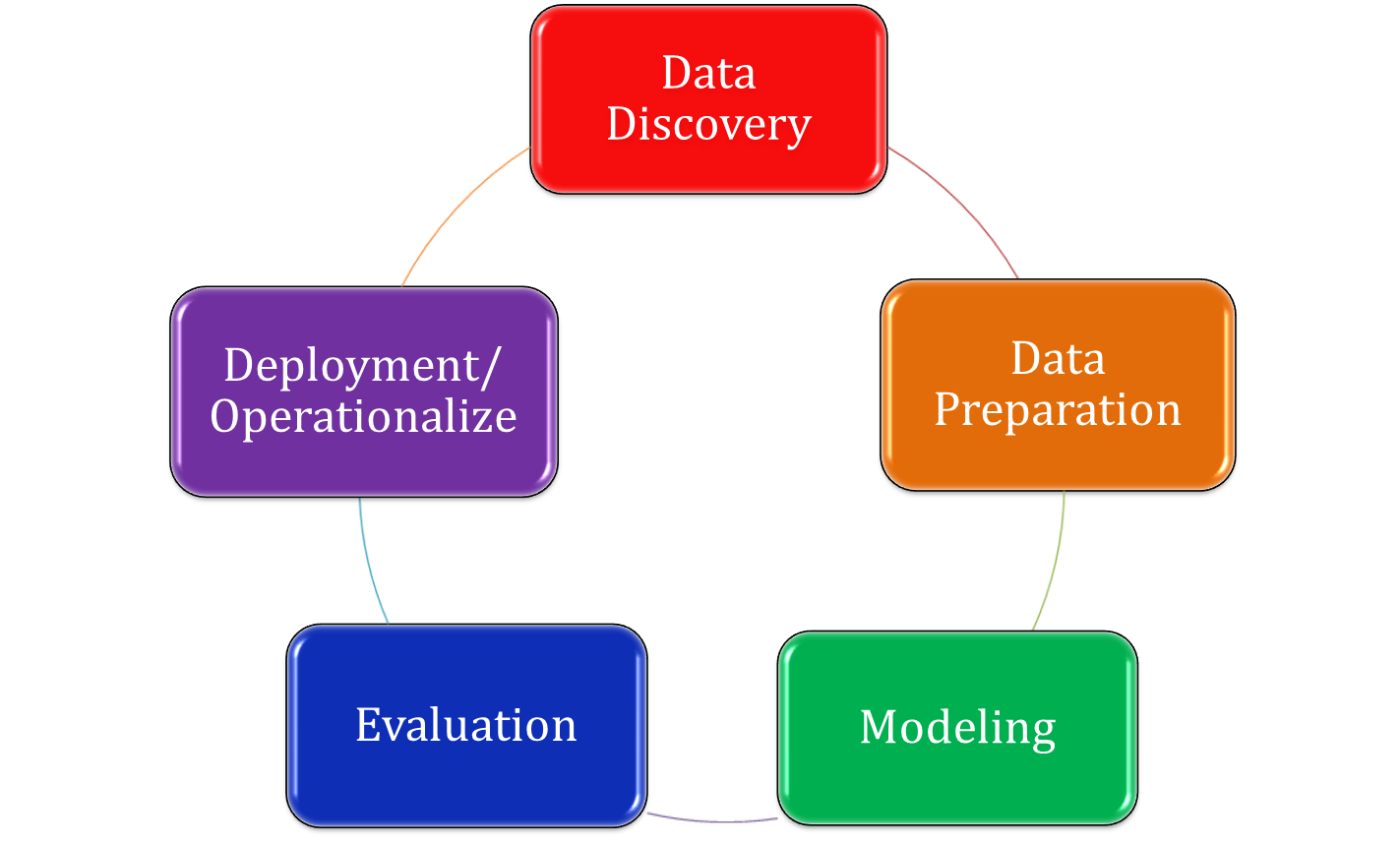
**EXCLUSIONS**

The following CFS is excluded from the data set: administrative, arson, crisis, death, domestic disturbance/violence, juvenile offenses, kidnapping, k9 explosive sweep, missing person, rape, restraining order, sex offense, stalking, and suicide.

The following addresses are also excluded: 737 SE 106th Ave (East Precinct), 449 NE Emerson St (North Precinct), 2801 N Gantenbein Ave (Legacy Emanuel Hospital), 4804 NE Glisan St (Providence Hospital), 1111 SW 2nd Ave (Central Precinct), 10300 SE Main St (Adventist Medical Center), 1014 NW 22nd Ave (Legacy Good Samaritan Medical Center), 41 NE Grand Ave (Detox Center), 3181 SW Sam Jackson Park Rd (OHSU Hospital), 10123 SE Market St (Adventist Medical Center), 12240 NE Glisan (Multnomah County Sheriff’s Office), and 3303 SW Bond Ave (OHSU Hospital).

**METHOD**

United States National Institute of Justice (USNIJ)

****

R Shiny Application

Zeppelin Notebook

R Programming & Shiny Application

R Programming (*RStudio*)

**APACHE ZEPELLIN**

The first software that was implemented into the data preparation phase was the Apache Zeppelin. The Apache Zeppelin is a web based notebook that enables interactive data analytics. In this application, we used a multi-featured notebook which incorporates the theme of data ingestion, discovery, analytics, as well as visualization & collaboration. Apache Zeppelin provides built-in Apache Spark integration and supports multiple back-end language. We used the pyspark interpreter in the Zeppelin notebook. This will allow the data to be imported and loaded which processes it much quicker. This was convenient for the progression of data cleaning, extraction, and manipulation.

**R PROGRAMMING LANGUAGE**

After the Zeppelin notebook was applied, the following software that was used was R. R is a programming language, which is an open source platform and a software environment for statistical computing and graphics. It is a popular programming language which is widely used among researchers, statisticians, and even data miners for developing statistical software and data analysis. R has its own libraries that can be used for installation and implements a variety of statistical and graphical techniques. The RStudio environment was utilized and three sets of R scripts (i.e. ui.r, server.r, global.r) were being executed to run an application called R Shiny. The statistical codes and techniques that we applied was clustering. Some statistical examples include classical statistical tests, linear and nonlinear modeling, time-series analysis, classification, clustering, and many more. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. The packages that were part of the process were libraries of (dplyr), (DT), (ggmap), (ggplot2), (googleVis), (leaflet), (RColorBrewer), (readr), (shiny), (shinydashboard), (shinythemes), and (tidyr).

**R SHINY (INTERACTIVE)**

The R Shiny is a useful web application where you can show data visualization. This is a used with a combination of a web browser and *RStudio*. Since, this is an interactive shiny application, it allows us to choose the crime variables that are needed to be analyzed and forecasted. The R shiny was created to visualize the city map of Portland, Oregon. It has a customized widget which helps to navigate the crime with latitude and longitude coordinates provided by the dataset. Overall, to forecast this application, we can change the date, months, even years in accordance to several crime variables. Since we are utilizing clustering algorithm, we can look at the time, trends, and the quantity of crimes in different spots of the city.

**CLUSTERING**

The principal meaning of clustering refers to the process of making a group of abstract objects into classes of similar objects. The important aspects of clustering are that a cluster of data objects can be treated as one group. By performing cluster analysis, we would have to partition the set of data into groups based on data similarity and then assign the labels to the groups.

Clustering methods can be classified into the following categories:

* Partitioning Method
* Hierarchical Method
* Density-based Method
* Grid-Based Method
* Model-Based Method
* Constraint-based Method

Functions of Cluster Analysis:

* Used in many applications such as market research, pattern recognition, data analysis, and image processing
* Help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns
* Used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations
* Helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location
* Clustering also helps in classifying documents on the web for information discovery
* Used in outlier detection applications such as detection of credit card fraud
* As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster

Clustering in Data Mining:

* **Scalability** − Highly scalable clustering algorithms to deal with large databases.
* **Ability to deal with different kinds of attributes** − Capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
* **Discovery of clusters with attribute shape** − Capable of detecting clusters of arbitrary shape.
* **High dimensionality** − Handle low-dimensional data and high dimensional space
* **Ability to deal with noisy data** − Contain noisy, missing or erroneous data Some algorithms are sensitive to such data and may lead to poor quality clusters.
* **Interpretability** − Should be interpretable, comprehensible, and usable

**K-MEANS CLUSTERING**

K-means is one of the easiest unsupervised learning calculations that take care of the notable grouping issue (MacQueen, 1967). The method takes after a straightforward and simple approach to group a given informational collection through a specific number of bunches settled from the earlier. The primary thought is to characterize k centroids, one for each bunch. These centroids ought to be put shrewdly in view of various area causes distinctive outcome. Along these lines, the better decision is to place them however much as could reasonably be expected far from each other. The following stride is to take each direct having a place toward a given informational index and partner it to the closest centroid. At the point when no point is pending, the initial step is finished and an early gathering age is finished. Now we have to re-ascertain k new centroids as barycenters of the groups coming about because of the past stride. After we have these k new centroids, another coupling must be done between similar informational collection focuses and the closest new centroid. A circle has been created. Accordingly, of this circle we may see that the k centroids change their area well-ordered until no more changes are finished. In other words, centroids do not move any more. Lastly, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function is as followed:

http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/images/image009.gif ,

where,  
                           *‘||xi- vj||’* is the Euclidean distance between *xi* and *vj.*

*‘ci’* is the number of data points in *ith* cluster

*‘c’* is the number of cluster centers.

The algorithm is comprised of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.

Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

**REFERENCES**

Arietta, S. M., Efros, A. A., Ramamoorthi, R., & Agrawala, M. (2014). City forensics: Using visual elements to predict non-visual city attributes. *IEEE transactions on visualization and computer graphics*, *20*(12), 2624-2633.

Ayday, E., Delgosha, F., & Fekri, F. (2012). Data authenticity and availability in multihop wireless sensor networks. *ACM Transactions on Sensor Networks (TOSN)*, *8*(2), 10.

Gubbi, Jayavardhana, et al. "Internet of Things (IoT): A vision, architectural elements, and future directions." Future generation computer systems 29.7 (2013): 1645-1660.

Hatton, W., Zhao, J., Gorantla, M. B., Chae, J., Ahlbrand, B., Xu, H., ... & Ko, S. (2015, October). Visual analytics for detecting communication patterns. In Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on (pp. 137-138). IEEE.

J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297

Jin, J., Gubbi, J., Marusic, S., & Palaniswami, M. (2014). An information framework for creating a smart city through internet of things. IEEE Internet of Things Journal, 1(2), 112-121.

Kamisar, Y. (1972). How to Use, Abuse--and Fight Back with--Crime Statistics. Okla. L. Rev., 25, 239.

Kostakos, V., O'Neill, E., Penn, A., Roussos, G., & Papadongonas, D. (2010). Brief encounters: Sensing, modeling and visualizing urban mobility and copresence networks. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *17*(1), 2.

Morenoff, J. D., Sampson, R. J., & Raudenbush, S. W. (2001). Neighborhood inequality, collective efficacy, and the spatial dynamics of urban violence. Criminology, 39(3), 517-558.

Nakaya, T., & Yano, K. (2010). Visualising Crime Clusters in a Space‐time Cube: An Exploratory Data‐analysis Approach Using Space‐time Kernel Density Estimation and Scan Statistics. Transactions in GIS, 14(3), 223-239.

Perera, C., Liu, C. H., & Jayawardena, S. (2015). The emerging internet of things marketplace from an industrial perspective: A survey. *IEEE Transactions on Emerging Topics in Computing*, *3*(4), 585-598.

"Population Estimates". United States Census Bureau.

Ruggieri, S., Pedreschi, D., & Turini, F. (2010). Data mining for discrimination discovery. ACM Transactions on Knowledge Discovery from Data (TKDD), 4(2), 9.

Sas, C., & Neustaedter, C. (2017). Exploring DIY practices of complex home technologies. *ACM Transactions on Computer-Human Interaction (TOCHI)*.

Shahidehpour, M., Li, Z., Bahramirad, S., & Khodaei, A. (2016). Optimizing Traffic Signal Settings in Smart Cities. IEEE Transactions on Smart Grid.

Wang, D., Kaplan, L., & Abdelzaher, T. F. (2014). Maximum likelihood analysis of conflicting observations in social sensing. *ACM Transactions on Sensor Networks (ToSN)*, *10*(2), 30.

Wang, F. (2005). Job access and homicide patterns in Chicago: An analysis at multiple geographic levels based on scale-space theory. *Journal of Quantitative Criminology*, *21*(2), 195-217.

Wang, F. Y. (2015). Scanning the issue and beyond: Transportation and mobility transformation for smart cities. IEEE Transactions on Intelligent Transportation Systems, 16(2), 525-533.

Wang, Y. C., & Chen, G. W. (2017). Efficient Data Gathering and Estimation for Metropolitan Air Quality Monitoring by Using Vehicular Sensor Networks. *IEEE Transactions on Vehicular Technology*.

Zheng, Y., Capra, L., Wolfson, O., & Yang, H. (2014). Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(3), 38.

Yu, C. H., Ward, M. W., Morabito, M., & Ding, W. (2011, December). Crime forecasting using data mining techniques. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on* (pp. 779-786). IEEE.

Zanella, A., Bui, N., Castellani, A., Vangelista, L., & Zorzi, M. (2014). Internet of things for smart cities. *IEEE Internet of Things journal*, *1*(1), 22-32.

(2007). Organized Crime in Oregon. *Trends in Organized Crime, 10*(3), pp. 50-75. doi:10.1007/s12117-007-9005-x

"2010 Census profiles: Oregon cities alphabetically M-P" (PDF). Portland State University Population Research Center.