Visual Analytics in the Public Sector: An Analysis on Diversities and SimilarIties of London’s Wards

Daqing Chen

School of Engineering, London South Bank University, London SE1 0AA, UK

Beatriz Martin Sanz

News UK, London SE1 9GF, UK

Erlong Zhao

School of Information Science and Technology, Northwest University, Xian 710127, China

ABSTRACT

In this paper, an analysis is presented on the diverse and common characteristics in different geographical areas across London’s wards with respect to certain social, economic, and welfare measures. 18 data sets from different sources are used in the study. The principal component analysis and the *k*-means cluster analysis have been applied by using SAS Enterprise Guide and Miner. Visual analytics has been implemented with Tableau to identify patterns and correlations among various measures. It has been found that a geographical distance or proximity does not necessarily indicate a significant difference or similarity between different areas on a given social and economic measure. The work suggests that collaborative management across all the London’s council boroughs is meaningful.

KEYWORDS

Visual analytics, data mining, principal component analysis, *k*-means clustering analysis, inequality.

1. INTRODUCTION

For more than two decades, a large amount of data from the public sector has been collected and accumulated to a massive scale. The data is rich involving various social and economic measures. Exploring and analyzing such as a data asset has been evidently crucial for the central and local governments to make tangible informed decisions in their strategic management processes (Syvajarvi and Stenvall, 2010).

As one of the most unique, multiple-cultural cities in the world and the global business and finance hub, London has posed a great challenge to the local authorities in terms of how to make London’s social and business environment stabilized, healthy, and sustainable. Inequality and poverty in London, for instance, is a long-established problem since the capital has been recognized as the most unequal city in the UK (Trust for London, 2015). This research presents a case study of identifying and analyzing the diverse characteristics in Greater London, aiming to provide some insight to help the central and local governments better understand the capital from multiple perspectives.

Geographically, London is currently divided into 33 Borough Councils. Each council constitutes several smaller areas known as wards, representing different electoral districts. There are 629 wards in total in London. These wards share some common features; however, each ward may its own characteristics as well. As such, completely different and diverse wards might be found within a borough. Understanding the diversity of London’s wards and the major causes for the diversity will help the local authorities adopt appropriate strategies and policies to tackle problems.

18 data sets from different sources are used in the study. All the data is publicly accessible, and the data can be categorized into several groups by measure including demographics, economics, quality of life, crime, and accidents and emergencies. The principal component analysis (PCA) and the *k*-means clustering analysis have been employed in this study. Detailed analysis on different ward groups is provided along with a visualized presentation on a polygon map.

The reminder of this paper is organized as follows. Section 2 provides a brief literature review on the relevant works and findings. Section 3 gives a detailed account about the data collection and pre-processing, and how the PCA and the *k*-means cluster analysis have been conducted. The diverse characteristics and the common features of the wards are identified. Further in Section 4, a correlation analysis is presented to explore possible causes for the diversities and similarities. Finally, in Section 5, the main findings of this research are discussed with recommendations for future work.

1. backgRound and relevant works

Due to its political, social and economic importance, nationally and globally alike, London has always been a great research focus. Many case studies are available in the literature with regard to what impact that various social, economic, and welfare characteristics can have on the quality of living in London, such as, inequality, poverty, and housing affordability, to name just a few.

Boyne and Cole (1998) analysed the structure of London’s local government and its evolution in the history for over a period of 150 years. Based on the Survey of Londoner’s Living Standards in 1987, Harloe (1992) examined the association of certain social characteristics with housing inequalities in London. The research established that economic position, household structure, gender, and ethnic identity of the Londoners have shown having the strongest association with different housing circumstances, and the housing affordability concerned a wide range of people, not only the poorest classes.

Aldridge *et. al.* (2015) examined the characteristics of London’s poverty in their report for Trust for London. They measured the role of different indicators relating to inequality, housing, employment, and education in all London’s boroughs. On the other hand, some studies have concentrated on a particular London borough. Watt (2003) studied the impact caused by the labour market restructuring in the late 1990’s on the employment circumstances of the local tenants in Camden. Arbaci and Rae (2012) conducted an analysis on mixed-tenure neighbourhoods in order to understand if the social and tenure-mixing policies have helped with alleviating deprivation effectively or had no positive effect at all. Using quantitative and qualitative longitudinal analyses, they concluded that diversification of housing tenure had positive effect on mitigating deprivation in London. Kirkbride *et. al.* (2014) studied whether the social deprivation and inequalities in East London were associated with the emergence of non-affective psychotic disorders. Hamnett and Butler (2011) have also placed their research emphasis on East London, studying how the distance from school has caused educational inequalities in the area. They argued that increasingly geography has been becoming the basis for rationing access to some forms of welfare including allocating secondary school places. The study on the NEET issues in London boroughs by Chen *et. al.* (2016) suggested that the median property price could be considered a simple and seemingly accurate indicator of areas likely to suffer from NEET.

Interestingly, Green (2012) applied a different approach to investigate the inequalities around London by choosing the London underground map to depict the distinct deprivation scores and provide a picture of a divided city, with areas wealthier than others.

In summary, much of the relevant research shown in the literature has been conducted at a borough level using multiple measures and exploratory data analysis techniques.

1. data collection, pre-processing and analysis

In this research we place our emphasis on identifying diversities and similarities across London’s wards based on multiple measures to provide a detailed analysis at an appropriate granular level. SAS Enterprise Guide, SAS Enterprise Miner and Tableau have been used as the main tools in this study.

A group of data sets of year 2014 has been collected from different data repositories as shown in Table 1. Each of the data sets involves a single measure only and therefore represents a variable in the analysis.

A target data set has been created by integrating the original data sets, in which, there are 629 rows, each corresponding to a particular ward.

Table 1. Data sets and sources (Note: All measures are numerical data type)

|  |  |  |  |
| --- | --- | --- | --- |
| Measure | Description | Measure | Description |
| Theme I: Demographics | | Theme IV: Crime | |
| Population | London population | Crime | Number of total crimes |
| Births | Number of births | Deliberate Fires | Number of all deliberate fire incidents recorded by the London Fire & Emergency Planning Authority |
| Deaths | Number of deaths | Assault Incidents | Number of assaults attended by the London Ambulance Service. |
| Theme II: Economics | | Weapon Injuries | Number of weapon injuries attended by the London Ambulance Service |
| Incapacity Benefit | Number of claimants | Drugs | Number of drug crimes in the Metropolitan Police Area. |
| Income Support | Number of claimants | Theme V: Accidents and emergencies | |
| Employment & Support Allowance | Number of claimants | Ambulance Attendance | Number of all incidents attended by the London Ambulance Service |
| Jobseekers Allowance | Number of claimants | Road Casualties | Number of road causalities |
| Houses Sold | Number of houses sold | Binge Drinking | Numbers of incidents where someone suffering from an alcohol related illness |
| House Price | Average house price in sterling |  |  |
| Theme III: Quality of life | |  |  |
| Public Transport Accessibility (PTA) | A score represents an accurate accessibility, where 8 is the highest level of accessibility. |  |  |
| Data sources: In general: <http://data.london.gov.uk>. Theme I: <https://data.london.gov.uk/dataset/birth-and-death-rates-ward>. Theme II: <https://data.london.gov.uk/dataset/average-house-prices>. Theme III: <https://data.gov.uk/dataset/17136b47-d5b4-4df7-9e88-b9653b34a060/crime-rates-in-the-metropolitan-policearea-by-ward>. Theme IV: <https://data.london.gov.uk/dataset/crime-rates-metropolitan-police-area-ward>.  Theme V: <https://data.london.gov.uk/dataset/ward-profiles-and-atlas>. | | | |
|  | | | |

The analysis started with the PCA for dimensionality reduction of the target data set. The first three principal components have been recognized significant and the cumulative value of the eigenvalues of these components was 75%. Accordingly, each ward instance in the data set has been transformed and represented in the 3-dimensional principal component space. This transformation is essential, especially when the data set under consideration contains many variables and/or many of the variables are categorial data type with many distinct values.

Using the transformed data set, the *k*-means cluster analysis has been further performed with 5 centroids and range normalization for variable normalization. Note that outliers in the data set, if any, wouldn’t be removed due to the context of the analysis that each instance is related to a specific ward. As such, a very small-sized cluster would be expected from the cluster analysis.

The statistics summary of the resultant clusters is given in Table 2. Each ward has been assigned to one of the 5 clusters uniquely. The clusters are also visualized on a polygon map as illustrated in Figure 1, where, as an example, the corresponding wards contained in Cluster 5 are highlighted.

For comparison purposes, the following discussions are based on the relative frequencies of the values in Table 1, i.e., the ratio of a given measure value to the total population in a cluster.

Table 2. Statistics summary of clusters (Note: OBS stands for number of observations)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Cluster 1 (OBS: 2) | | | | Cluster 2 (OBS: 222) | | | |
| Mean | Std Dev | Min | Max | Mean | Std Dev | Min | Max |
| Population | 11460 | 511 | 11099 | 11821 | 15116 | 1992 | 11024 | 22100 |
| Births | 87 | 1 | 86 | 87 | 247 | 63 | 97 | 465 |
| Deaths | 48 | 11 | 40 | 55 | 72 | 19 | 38 | 133 |
| Incapacity Benefit | 103 | 11 | 95 | 110 | 131 | 28 | 75 | 235 |
| Income Support | 115 | 14 | 105 | 125 | 310 | 83 | 135 | 575 |
| Employment Support | 440 | 120 | 355 | 525 | 596 | 127 | 320 | 975 |
| Jobseekers Allowance | 203 | 88 | 140 | 265 | 369 | 103 | 175 | 805 |
| House Price | 1986482 | 537463 | 1606439 | 2366526 | 401366 | 194921 | 191601 | 1601476 |
| House Sold | 231 | 106 | 156 | 306 | 166 | 66 | 37 | 431 |
| PTA | 8 | 0 | 8 | 8 | 4 | 1 | 2 | 8 |
| Crime | 12419 | 1315 | 11489 | 13348 | 1427 | 570 | 699 | 4233 |
| Deliberate Fires | 6 | 1 | 5 | 6 | 9 | 8 | 0 | 67 |
| Assault Incidents | 403 | 92 | 338 | 468 | 64 | 28 | 15 | 197 |
| Weapon Injuries | 13 | 1 | 12 | 13 | 7 | 3 | 0 | 17 |
| Drugs | 575 | 112 | 495 | 654 | 99 | 79 | 21 | 754 |
| Ambulance Attendance | 7845 | 1291 | 6932 | 8758 | 2009 | 504 | 958 | 4030 |
| Road Casualties | 340 | 86 | 279 | 401 | 59 | 29 | 4 | 174 |
| Binge Drinking | 1144 | 187 | 1012 | 1276 | 63 | 51 | 6 | 348 |
|  |  |  |  |  |  |  |  |  |
| Variable | Cluster 3 (OBS: 33) | | | | Cluster 4 (OBS: 177) | | | |
| Mean | Std Dev | Min | Max | Mean | Std Dev | Min | Max |
| Population | 17477 | 3213 | 11667 | 24567 | 11056 | 1710 | 5283 | 15423 |
| Births | 260 | 65 | 109 | 367 | 155 | 48 | 54 | 324 |
| Deaths | 120 | 38 | 45 | 219 | 55 | 16 | 16 | 91 |
| Incapacity Benefit | 96 | 33 | 35 | 170 | 69 | 30 | 5 | 160 |
| Income Support | 216 | 102 | 45 | 440 | 134 | 71 | 0 | 335 |
| Employment Support | 420 | 163 | 110 | 815 | 307 | 132 | 20 | 671 |
| Jobseekers Allowance | 223 | 101 | 75 | 470 | 156 | 75 | 10 | 395 |
| House Price | 404228 | 150746 | 219095 | 862243 | 763924 | 663197 | 184410 | 5257978 |
| House Sold | 334 | 144 | 97 | 897 | 151 | 48 | 53 | 303 |
| PTA | 3 | 1 | 2 | 6 | 4 | 2 | 1 | 8 |
| Crime | 1728 | 1097 | 562 | 5371 | 796 | 440 | 231 | 3627 |
| Deliberate Fires | 9 | 8 | 2 | 46 | 3 | 3 | 0 | 20 |
| Assault Incidents | 60 | 46 | 10 | 200 | 24 | 14 | 4 | 91 |
| Weapon Injuries | 5 | 4 | 0 | 17 | 3 | 2 | 0 | 11 |
| Drugs | 96 | 99 | 21 | 464 | 42 | 34 | 6 | 286 |
| Ambulance Attendance | 2602 | 1433 | 1508 | 9465 | 1181 | 333 | 617 | 3155 |
| Road Casualties | 74 | 37 | 24 | 169 | 37 | 23 | 7 | 159 |
| Binge Drinking | 84 | 80 | 16 | 315 | 32 | 24 | 4 | 209 |
|  |  |  |  |  |  | | | |
| Variable | Cluster 5 (OBS: 195) | | | |
| Mean | Std Dev | Min | Max |
| Population | 13423 | 2153 | 7702 | 18066 |
| Births | 186 | 56 | 82 | 379 |
| Deaths | 90 | 26 | 21 | 166 |
| Incapacity Benefit | 74 | 31 | 15 | 185 |
| Income Support | 146 | 75 | 30 | 485 |
| Employment Support | 296 | 116 | 100 | 705 |
| Jobseekers Allowance | 158 | 74 | 40 | 370 |
| House Price | 431669 | 186606 | 192270 | 1330065 |
| House Sold | 217 | 69 | 71 | 486 |
| PTA | 3 | 1 | 1 | 8 |
| Crime | 804 | 327 | 281 | 2167 |
| Deliberate Fires | 5 | 5 | 0 | 28 |
| Assault Incidents | 25 | 14 | 2 | 77 |
| Weapon Injuries | 2 | 2 | 0 | 11 |
| Drugs | 38 | 26 | 6 | 154 |
| Ambulance Attendance | 1490 | 355 | 719 | 2687 |
| Road Casualties | 39 | 20 | 6 | 128 |
| Binge Drinking | 29 | 20 | 3 | 115 |

As shown in Figure 1, wards located in different boroughs can have some common characteristics although they are quite far one from the other geographically. For instance, the wards in Cluster 5 have the lowest crime rate (5.88%) across all the 5 clusters and this is coincided with the lowest rates of assault incidents, drugs relevant problems, and road causalities (0.19%, 0.28%, and 0.28%, respectively). On the other hand, a borough can have quite diverse wards although geographically those wards are close to each other. For example, Southwark Borough has wards belonging to Clusters 2, 4, and 5 as shown in Figure 2, where Cluster 2 has a much higher crime rate (9.44%) and higher rates of assault incidents, drugs relevant problems, and road causalities (0.44%, 0.62%, and 0.45%, respectively).

1. area of attention

In this Section we examine the patterns of correlations among certain variables to provide indications for possible causes of diversities and similarities. It is worth noting that correlation not necessarily indicates a causation.

|  |  |
| --- | --- |
| A close up of a map  Description generated with high confidence  Figure 1. Examples of wards geographically located in different boroughs are grouped in a cluster. | A close up of a map  Description generated with high confidence  Figure 2. Example of a borough can have diverse wards that belong to different clusters. |
| A close up of text on a white background  Description generated with high confidence  Figure 3. Variation in different clusters. | A close up of text on a white background  Description generated with high confidence  Figure 4. Variation of average home price vs. average number of homes sold. |

Unemployment related issues such as crime is always a major concern for the central and local governments and authorities. Checking with the clusters established, it can be seen that a higher employment support is usually coincided with a higher crime. However, there are some specific areas where a much higher ratio of crime to employment support can be evidenced. In general, all the groups have a relatively concentrated value in terms of the ratios of crime to employment support, drugs related problems, and assault incidents except for Cluster 3. This cluster only contains 33 wards and has the highest variation with all these ratios. The wards particularly concerned attributing to the variation are Stratford and New Town, Heathrow Villages, and Fairfield as shown in Figure 3. These wards have a much higher variation against the trend line compared with the other wards. In comparison, Cluster 3 has a very low variation with the ratios as indicated in Figure 3, where a bigger circle indicates a higher crime rate.

House price and the number of houses sold are always an important economic indicator reflecting collectively the popularity of an area, quality of life, social development and inequality, etc. In the clusters created, Cluster 4 has many wards that have a high ratio of average house price value to the average number of houses sold. On the other hand, Cluster 3 has many wards that have a comparably low ratio value for average house prices vs. houses sold, as displayed in Figure 4. Considering the patterns identified in Figures 3 and 4, it becomes evident that there is a clear and strong correlation among employment support, crime, house price and houses sold. Similarly, a correlation analysis can be performed with other measures and concerns.

For areas that have some common characteristics as reflected by certain measure values, it is advisable that probably similar strategies and policies could be adopted in practice for the local governments in their management processes. Therefore, an effective mechanism for collaborations and communications across all the boughs should be established to support regular and collective reviews on the wards.

1. conclusion and future work

In this paper, the diverse characteristics and common features across London’s wards with respect to certain measures have been analyzed. It has been shown that a geographical distance or proximity does not necessarily indicate a significant difference or similarity between different areas with regards to a given social and economic measure. A council borough usually constitutes many diverse wards as reflected by certain measures. On the other hand, wards from different boroughs can have very similar social and economic features. This suggests that collaborative management across all the London’s council boroughs is meaningful. The local authorities need to consider how to apply similar strategies to the wards that share some common features. A strong correlation among the measures has been established which may suggest implicitly causes for the diversities and similarities.

Future research includes analysis at a further refined granular level, the LSOA level (known as Lower Layer Super Output Areas). In addition, in order to examine how diversities and similarities have evolved over time, data from different years should be collected and analysed.

REFERENCES

Aldridge, H., Born, T.B., Tinson, A., and MacInnes, T., 2015. London’s Poverty Profile. [*https://www.npi.org.uk/files/5714/4533/2889/LPP\_2015\_report.pdf*](https://www.npi.org.uk/files/5714/4533/2889/LPP_2015_report.pdf)

Arbaci, S. and Rae, I., 2012. Mixed-tenure Neighbourhoods in London: Policy myth or effective device to alleviate deprivation? *International Journal of Urban and Regional Research*, Vol. 37, No. 2, pp. 451–479.

Boyne, G.A., Cole, M., 1998. Revolution, Evolution and Local Government Structure: An Empirical Analysis of London. *Urban Studies*, Volume: 35, Issue: 4, pp.751-768.

Chen, D., Asaolu, B., and Qin, C., 2016. Big Data Analytics in the Public Sector: A Case Study of NEET Analysis for the London Boroughs, Proceeding of International Conference on Big Data Analytics. *Proceedings of Data Mining and Computational Intelligence*. Funchal, Madeira, Portugal, pp.171-178.

Green, M.A., 2012. Mapping Inequality in London: A different approach. *The Cartographic Journal*, Vol. 49, No. 3, pp. 247–255.

Hamnett, C. and Butler, T., 2011. ‘Geography matters’: the role distance plays in reproducing educational inequality in East London, *Institute of British Geographers Transactions*, Vol. 36, No. 4, pp. 479-500.

Harloe, M., 1992. Housing inequality and social structure in London. *Housing Studies*, Vol. 7, No. 3, pp. 189–204.

Kirkbride, J.B., Jones, P.B., Ullrich, S., and Coid, J.W., 2014. Social deprivation, inequality, and the neighborhood-level incidence of psychotic syndromes in East London. *Schizophrenia Bulletin*, Vol. 40, No. 1, pp.169–180.

Syvajarvi, A. and Stenvall, J., 2010. *Data Mining in Public and Private Sectors: Organizational and Government Applications*, IGI Global, Hershay, USA.

Trust for London, 2015. London’s Poverty Profile, Inequality Indicators. [*http://www.londonspovertyprofile.org.uk/indicators/topics/inequality/*](http://www.londonspovertyprofile.org.uk/indicators/topics/inequality/)*.*

Watt, P., 2013. Urban marginality and labour market restructuring: Local authority tenants and employment in an inner London borough. *Urban Studies*, Vol. 40, No. 9, pp. 1769–1789.