BIG DATA ANALYTICS IN THE PUBLIC SECTOR: A CASE STUDY OF NEET ANALYSIS FOR THE LONDON BOROUGHS

Daqing Chen[1], Babatunde Asaolu[1] and Chao Qin[2]
[1]*Division of Computer Science and Informatics, School of Engineering, London South Bank University*
103 Borough Road, London SE1 0AA, UK
[2]*School of Electronics and Information, Northwestern Polytechnical University*
127 West Youyi Road, Xi'an 710072, China

ABSTRACT

For decades, the issue of young people who are aged 16 -18 and not in employment, education, or training (NEET) has been a major concern for governments and local authorities. In this paper, a big data perspective is taken to examine the NEET issue in order to highlight factors that are correlated with NEET, the negative consequences and causes of NEET, and potential solutions to NEET. The NEET dataset about the 33 London Boroughs has been considered along with other seven datasets relating to population, crime offences, benefits claimants, median property price, active businesses, immigrants, and conception under 18. All the datasets are public-accessible and comprise a data collection of the same period of time from 2009 to 2013. Each of them represents a particular measure. Hierarchical variable clustering, *k*-means clustering, and correlation analysis have been conducted using SAS Enterprise Miner and Tableau in this work. These tools enable us to analyse the problem in a multi-dimensional, hierarchical, integrated, and longitudinal way. The research has demonstrated that a) The NEET issue is much more severe in outer London than in inner London; b) The main factors correlated with NEET vary from inner London to outer London; c) Each of the measures considered has a certain correlation strength with the NEET rate, and amongst them, median property price is a simple and seemingly accurate indicator of areas likely to suffer from NEET and thus to take appropriate precautions in order to reduce the likelihood of further increases in NEET; and d) The London Boroughs can be grouped based on similarities in terms of a set of given measures, and the memberships of the groups remain stable.

KEYWORDS

Big data analytics, NEET analysis, *k*-means clustering, Hierarchical clustering, SAS Enterprise Miner, Tableau.

1. INTRODUCTION

For decades, the issue of young people who are aged 16 -18 and not in employment, education, or training (NEET) has been a major concern for governments and local authorities (Stoten 2014, Woolford 2012). In London particularly, NEET remains a problem within the London Boroughs although many courses of action have been taken and met with varying degrees of success (GLA 2007).

It has been evident that the costs of young people who are NEET can be high, with long-term consequences, not only to the individual, but also to the entire society and the economy as a whole. Therefore, various qualitative and quantitative approaches and techniques have been employed in the attempt to better understand the NEET phenomenon, analyse and identify the main factors that are attributed to NEET, and develop measures to address NEET accordingly (Bymer and Parsons 2002, Stoten 2014, Woolford 2012, Britton, Gregg, MacMillan and Mitchell 2011, Egan M., Daly M., Delaney L 2015). Due to the social and economic complexity of the NEET problem, it becomes essential that multiple factors should be taken into account in the research to reflect the nature of the problem, that is, how various factors are correlated with, and potentially have collectively affected, the existence of the NEET problem.

In this paper, a big data perspective is taken to examine the NEET issue in order to highlight factors that are correlated with NEET, the negative consequences and causes of NEET, and potential solutions to NEET. The NEET dataset about the 33 London Boroughs has been considered along with other seven datasets relating to population, crime offences, benefits claimants, median property price, active businesses, immigrants, and conception under 18. All the datasets are public-accessible and comprise a data collection of the same period of time from 2009 to 2013. Each of them represents a particular measure. Hierarchical variable clustering, *k*-means clustering, and correlation analysis have been conducted using SAS Enterprise Miner and Tableau in this work. These tools enable us to analyse the problem in a multi-dimensional, hierarchical, integrated, and longitudinal way.

The reminder of this paper is organized as follows. In Section II a brief discussion on the relevant work on NEET is given, and the approach adopted to the problem in this research is highlighted. In Section III, the datasets to be considered in this research are described in detail. Further the relevant activities for data pre-processing are discussed in order to deal with data quality issues and to get the data fit for the required analysis. Detailed analysis is given in Section IV, upon which, the main findings from the analysis are discussed in Section V. Finally, the concluding remarks and suggestions for future work are provided in Section VI.

2. RELATED WORK

The issue of NEET has been a concern for governments and local authorities since it was first identified as a problem (Stoten 2014). Various studies have been carried out to identify the characteristics and the causes of NEET. The Youth Cohort Study (Wolford 2012) identified several factors relating to NEET:

- Having no formal qualification;
- Being excluded from school in years 10 or 11;
- Having a disability or health problem;
- Constant truancy in year 11;
- Parent are not in full-time employment;
- Living separately from parents; and
- Looking after own children.

In addition, the study has proposed a more creative approach, known as LIFT, to tackle the issue, including attendance, motivation, behavior, and post-16 participation.

Stoten (2014) explored the problem of NEET in the Cleveland area by looking at what features define NEETs, what would be the best responses in tackling NEET across the region. Stoten designed two plans to address the issue. The first one was a reduction strategy plan, which brought together a consortium of organizations that worked closely together in tackling the issue. It was found that the reduction strategy plan actually had some success and reduced the local NEET rate from 12.8% to 10.2%. The second plan was to raise the school leaving age, and the key to this plan was to minimize young people dropping out post 16.

The Greater London Authority (GLA) performed a study on what were the best practices that stopped young people from being NEET in London, and further recommended various ways to help reduce and prevent NEET in London (GLA 2007).

In the research undertaken by Bymer and Parsons (2002), there were two fundamental questions that the research was concerned and intended to find answers to: 1) Are NEETs simply a group who have failed to do well in school or are there other factors which set them on a route with little or no opportunity; and 2) Is being NEET just a temporary stage in life due to disadvantages and failures or does being NEET itself constitute as a condition that makes it hard to adjust to adult life. A longitudinal dataset from the British birth cohort study in the 1970s was used and a logistic regression model was employed in the research. The dataset was collected from a variety of sources, including interviews with teachers, parents, self-completed questionnaires and tests. Young males and females were assessed separately as variables for predicting who can be classified as NEET. Bymer and Parsons concluded that, there is little doubt that poor educational qualifications have an impact on being branded as NEET. Lack of parental interest in children's education, labour market, and teen-age pregnancy are also important factors to consider.

All these insightful studies have identified several key factors that could help explain the NEET phenomenon and address the NEET problem in their social environment; however, due to the social and economic complexity of the NEET problem, there are other possible factors, such as crime offences and benefits claimants, that should be explored in order to provide a complete account of the problem. This research intends to further expand on some of the issues raised in these studies, and use big data approach to examine various factors that are correlated with and possibly attributed to NEET.

3. DATA CONSIDERATION

3.1 The Datasets

Various open datasets from the public sector are available. These datasets are usually spatially and temporally related, comprehensive, and are accumulated year on year. As an enriched data repository for analytics, these open datasets provide different measures on various aspects of our society.

In this study a number of open datasets are selected. In addition to the NEET dataset itself, the following seven datasets about the 33 London Boroughs have been chosen. These datasets are considered correlated with NEET in some implicit or explicit ways by their nature:

- Number of crime offences;
- Population of male and female aged 16 - 18;
- Number of immigrants;
- Number of people claiming benefits;
- Number of active businesses;
- Median house price; and
- Number of conceptions under 18.

The data sources and the relevant URLs are shown Table 1. The 33 London Boroughs are listed in Table 2 along with the Local Authority District (LAD) codes. These Boroughs can be grouped as inner and outer London groups. Note that only the Borough level's data is considered in this paper, although other geographical levels' data is available, for instance, LSOA and Ward levels. All the datasets comprise a data collection of the same period of time from 2009 to 2013.

3.2 Data Pre-processing

No quality data, no quality analysis outcomes. Data pre-processing is essential in any data analytics project. Typical data pre-processing tasks include dealing with data quality issues (such as missing values, outliers, and noisy data), and normalizing and transforming the original data to make it fit the required analysis and modelling. In this research, the main activity performed for data pre-processing involves, for each of the eight datasets, transforming all the absolute values (numbers) into the corresponding percentages per Borough. This ensures that meaningful comparison can be made across all the London Boroughs based on a given measure. For example, transforming the number of NEETs of each Borough into a corresponding percentage out of the total number of NEETs in all London Boroughs allows us to compare each Borough's performance with the others based on the measure of the NEET rate. In the following Sections, we use the NEET rate and the number of NEETs interchangeable, and similarly to all the other 7 factors (measures).

All the transformed datasets were further integrated into a fact table which contains the following 12 fields (measures): Borough, Year, Male_Population, Female_Population, Crime_Offences, Active_Businesses, Benefits_Claimants, Immigrants, Conception, NEET, Inner/Outer_London, and Median_House_Price. The fact table was then uploaded into Tableau and SAS Enterprise Miner for analysis.

Table 1. Data Collection

Dataset name	Website name and URL
Number of NEETs	The Greater London Authority data store http://data.london.gov.uk/dataset/young-people-not-employment-education-or-training-borough
Number of crime offences	The Metropolitans Police http://maps.met.police.uk/tables.htm
Population of male and female aged 16 - 18	The Greater London Authority data store http://data.london.gov.uk/dataset/office-national-statistics-ons-population-estimates-borough
Number of immigrants	The Greater London Authority data store (collected by The Office of National statistics (ONS)) http://data.london.gov.uk/dataset/national-insurance-number-registrations-overseas-nationals-borough

Number of people claiming benefits	The government website GOV.uk https://www.gov.uk/government/statistics/census-output-area-data-on-workless-benefit-claimants-YEAR*-london * To be substituted with the year required, e. g., 2009.
Number of active businesses	The Greater London Authority data store http://data.london.gov.uk/dataset/business-demographics-and-survival-rates-borough
Median house price	The Greater London Authority data store http://data.london.gov.uk/dataset/average-house-prices-borough
Number of conceptions under 18	The Greater London Authority data store http://data.london.gov.uk/dataset/teenage-conceptions-borough

Table 2. List of the London Boroughs

LAD Code	LAD Name	Inner/Outer London	LAD Code	LAD Name	Inner/Outer London
E09000001	City of London	Inner	E09000018	Hounslow	Outer
E09000002	Barking and Dagenham	Outer	E09000019	Islington	Inner
E09000003	Barnet	Outer	E09000020	Kensington and Chelsea	Inner
E09000004	Bexley	Outer	E09000021	Kingston upon Thames	Outer
E09000005	Brent	Outer	E09000022	Lambeth	Inner
E09000006	Bromley	Outer	E09000023	Lewisham	Inner
E09000007	Camden	Inner	E09000024	Merton	Outer
E09000008	Croydon	Outer	E09000025	Newham	Inner
E09000009	Ealing	Outer	E09000026	Redbridge	Outer
E09000010	Enfield	Outer	E09000027	Richmond upon Thames	Outer
E09000011	Greenwich	Outer	E09000028	Southwark	Inner
E09000012	Hackney	Inner	E09000029	Sutton	Outer
E09000013	Hammersmith and Fulham	Inner	E09000030	Tower Hamlets	Inner
E09000014	Haringey	Inner	E09000031	Waltham Forest	Outer
E09000015	Harrow	Outer	E09000032	Wandsworth	Inner
E09000016	Havering	Outer	E09000033	Westminster	Inner
E09000017	Hillingdon	Outer			

4. ANALYSIS AND DISCUSSION

4.1 Correlation

Although NEET is correlated with multiple factors and some of them may be considered influential (attributing) factors, each factor involved may have a different degree of correlation strength with NEET. Determining the degree of correlation strength of each factor is important in NEET analysis as it helps identify potential influential factors and highlight factors that have similar degree of potential influence. It is particularly useful if a very high number of factors to be considered in the analysis.

In order to investigate the degree of correlation strength, a hierarchical clustering of the variables and their correlation with the NEET rate were conducted in SAS Enterprise Miner (Cerrito 2006, Aggarwal and Kosian 2011) (using the Variable Clustering node and the StatExplore node). The results are shown in Figure 1 and Figure 2.

As illustrated in Figure 1, the factors under consideration can be grouped into three clusters, labelled as CLUS1, CLUS2, and CLUS3, respectively, in a hierarchical way. CLUS1 includes 5 variables: NEET, Male_Population, Female_Population, Conception, and Median_House_Price. The variables in this cluster are correlated with the NEET rate closely. CLUS2 has 3 variables: Immigrants, Benefits_Claimants, and Crime_Offences, and these variables have relatively weak correlation with the NEET rate. CLUS3 only has one variable Active_Businesses, and it has the weakest correlation with the NEET rate.

Figure 2 provides consistent results that the correlation strength of each variable with the NEET rate can be shown in a descending order as Male_Population, Female_Population, Conception, Crime_Offences, Benefits_Claimants, Immigrants, Active_Businesses, and Median_House_Price. Among them, Male_Population, Female_Population, and Conception are the most positively correlated variables, and Median_House_Price is the most negatively correlated variable.

The correlation strength was further examined by inner and outer London groups, and it has been found that the correlation strength of each variable with the NEET rate varies with Boroughs in inner and outer London. For the inner London group, Median_House_Price, Immigrants, and Active_Businesses are more significantly correlated with the NEET rate than in the outer London group. In addition, the outer London group has a much higher NEET rate (43% higher) than the inner London group.

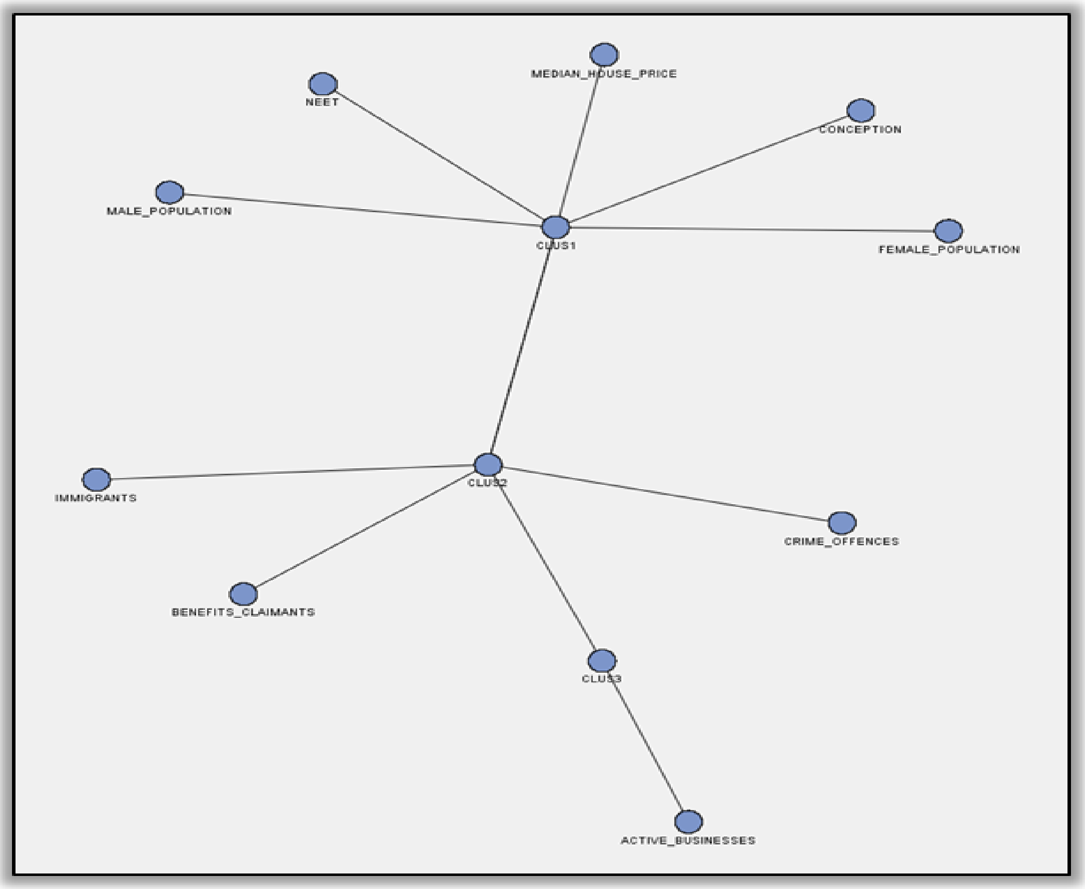

Figure 1. Hierarchical clustering of the variables

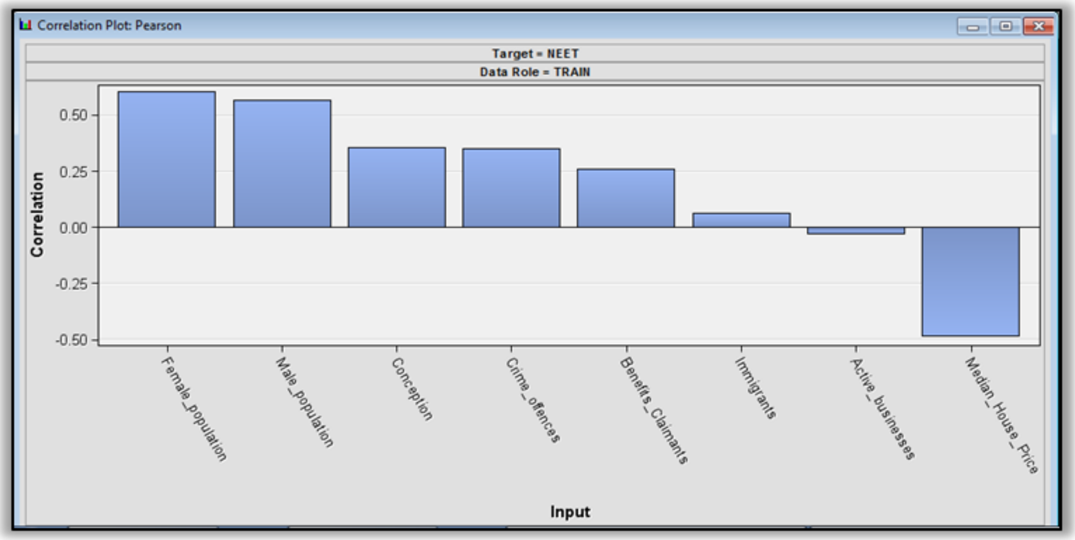

Figure 2. Correlation of the variables with the NEET rate

4.2 Similarity across London Boroughs

Based on the factors under consideration, it is interesting to see if there are any similarities across the London Boroughs in relation to the factors, e. g., which Boroughs have similar crime rate and NEET rate. This is of practical importance from managerial perspective in tackling the NEET issue - Boroughs that are under a similar situation may apply similar policies and strategies if a similarity can be established. Therefore, one Borough's experience in addressing the NEET problem can be helpful for other similar Boroughs to generalize.

In order to identify similarities across the London Boroughs, the *k*-means clustering was performed in SAS Enterprise Miner (Cerrito 2006) (using the Cluster node), in which the Euclidean distance was used as the similarity measure. The number of clusters (i.e., centroids) was set to 3, as it is always advisable to use as fewer clusters as possible to group as many samples as possible, as long as meaningful clusters can be established within a certain context. Note that the *k*-means clustering is a type of partitional clustering, which means that after the clustering, each sample (i.e., each Borough in our case) will be assigned to a particular cluster only. Furthermore, the clustering results were imported into Tableau to visualize the clusters established with their members associated.

Two factors were considered in the *k*-means clustering: NEET and Median_House_Price. This is based on the correlation analysis discussed in the previous Section that Median_House_Price is the most negatively correlated factor with NEET. In addition, the clustering was conducted using four years data 2010, 2011, 2012, and 2013 separately in order to identify if the similarity and cluster memberships have changed over time. The results are shown in Fig. 3. It becomes evident from the *k*-means clustering results that there are indeed some clear similarities across the London Boroughs. For instance, these 7 Boroughs: Camden, Harrow, Kingston upon Thames, Hammersmith and Fulham, Richmond upon Thames, and Wansworth, and Westminster usually had similar NEET rate and Median_House_Price rate, i.e., low NEET rate and high Median_House_Price rate. In comparison, Boroughs like Barking and Dagenham, Bromley, Croydon, and Enfield usually had a high NEET rate with a low Median_House_Price rate. The memberships of the clusters in general remain stable over the 4 years although there are some variances.

4.3 Discussion

From the analysis results illustrated in the previous Sections, we have the following remarks:

a) Each factor under consideration has a certain degree of correlation strength with the NEET rate across all the London Boroughs. The degree of correlation strength varies with different factors and with the inner and outer London groups as well.

b) Amongst all the 8 factors, Median_Property_Price rate was the most negatively correlated factor with the NEET rate. In other words, this factor may be used as an indicator of areas that are likely to suffer from NEET and thus to take appropriate precautions in order to reduce the likelihood of further increases in NEET. Note that, though, this doesn’t mean the NEET problem has been caused by low property prices in an area.

c) Boroughs across London do have similarities based on a given set of factors. As such, the London Boroughs can be segmented into several meaningful groups, each containing a number of Boroughs that have similar situation and issues. In addition, this segmentation remains stable over the period of time considered. This finding could be beneficial to Boroughs in the same segment in that they can work together sharing experiences and adopting similar strategies in tackling the NEET issue.

d) The similarity analysis can be further conducted at other geographical levels, such as Ward and LSOA levels. Usually from the local authority’s perspective, analysis at LSOA level is of more practical importance. Note that the amount of data to be handled is big if data at all those levels is used, since there are roughly 649 Wards and 4642 LSOA codes across the London Boroughs.

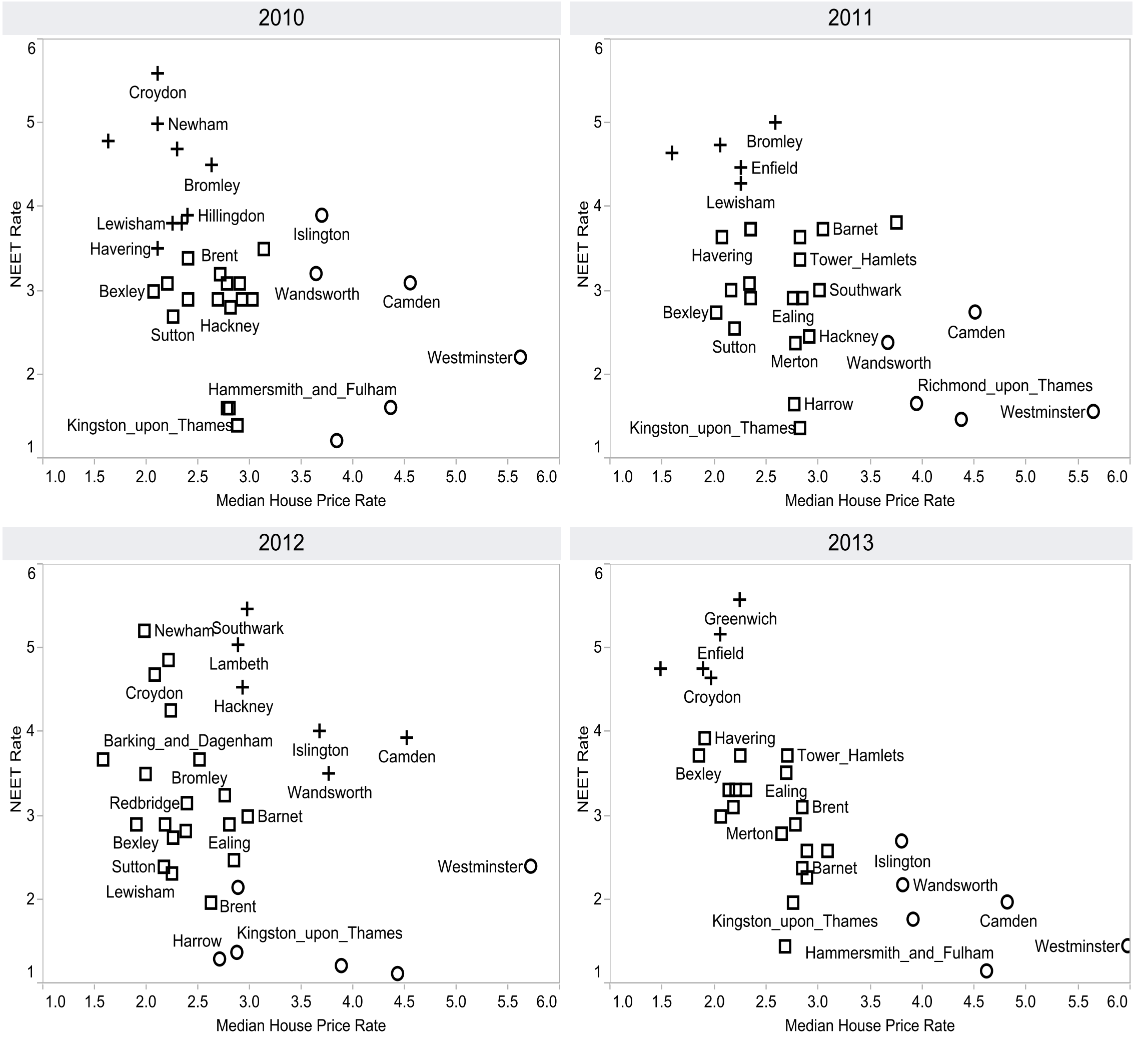

Figure 3. Similarity identified across the London Boroughs by using *k*-means cluster analysis based on two factors (variables). Data of 4 years was used separately. 3 clusters were considered, labelled by the signs plus, square, and circle, respectively

5. CONCLUSION AND FUTURE WORK

The NEET issue is complex and involves many factors to be considered. In this paper, a big data approach has been adopted to analyse the problem by taking into account as many factors as possible that are implicitly or explicitly relevant. The main advantage to use this approach is that it potentially can reveal valuable insight which may not be uncovered if only considering some of the factors. It is an integrated, holistic approach. It has been demonstrated that: a) The NEET issue is much more severe in outer London than in inner London; b) The main factors correlated or potentially attributing to NEET vary from inner London to outer London; c) Each of the measures considered has a certain correlation strength with the NEET rate, and amongst them, median property price is a simple and seemingly accurate indicator of areas likely to suffer from NEET and thus to take appropriate precautions in order to reduce the likelihood of further increases in NEET; and d) the London Boroughs can be grouped based on similarities in terms of a set of given measures, and the memberships of the groups remain stable.

We intend to include more factors in the future and to work jointly with the local authorities in London. Also, other geographical levels' data will be considered including Ward and LSOA levels. In the longer term, a data platform should be developed in order to accommodate more data in a dynamical and substantial way, and scalable clustering algorithms need to be explored. In addition, appropriate APIs should be implemented for retrieving data effectively from various sources on the Internet.

ACKNOWLEDGEMENT

The authors would like to thank the seven anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper.

REFERENCES

Aggarwal V., and Kosian S., 2011. Feature Selection and Dimension Reduction Techniques in SAS®, Northeast SAS Users Group 2011 Proceedings. Available at: http://www.lexjansen.com/nesug/nesug11/sa/sa08.pdf.

Britton J., Gregg P., MacMillan L. and Mitchell S., 2011. The early bird…preventing young people from becoming a NEET statistic. Available at: http://www.bristol.ac.uk/media-library/sites/cmpo/migrated/documents/earlybirdcmpo.pdf

Bynner, J. and Parsons, S., 2002. Social Exclusion and the Transition from School to Work: The Case of Young People Not in Education, Employment, or Training (NEET). *Journal of Vocational Behavior*, Vol. 60, No. 2, pp. 289-309.

Cerrito, P., 2006, *Introduction to Data Mining Using SAS Enterprise Miner*. SAS Institute.

Egan M., Daly M., Delaney L. 2015. Childhood psychological distress and youth unemployment: Evidence from two British cohort studies. *Social Science & Medicine*, Vol. 124, pp. 11-17.

Greater London Authority, 2007. Toolkit for London Extended Schools – working with young people NEET, Available at: http://legacy.london.gov.uk/mayor/children/docs/neet-toolkit.pdf

Stoten, D., 2014. NEETs: a case study in addressing the issues relating to disengaged youth in East Cleveland, *Education + Training*, Vol. 56, No. 5, pp. 467-480.

Woolford, C., 2012. A North East Pupil Referral Unit's Response to the Challenge of NEETs, in: A North East Pupil Referral Unit's Response to the Challenge of NEETs, in: Emerald Group Publishing Limited, pp. 223-236.