Use of Visual Analytics to Study the Impact of Socio-Economic Factors on Wellbeing in Citizens across London Wards in 2011

Ryan Nazareth
Department of Computer Science, City, University of London

1.MOTIVATION AND RESEARCH QUESTIONS

Measuring the wellbeing of a population, is important to understand what factors affect individuals in their daily lives. In addition, poor wellbeing can lead to a number of mental health problems, which can result in unnecessary visits to the hospital and a severe reduction in the quality of life and inability to carry out daily tasks with full efficiency. Local authorities in London rely on wellbeing data in order to target resources, and with local authorities currently gaining more responsibilities from government, this is of increasing importance [1]. The insights about wellbeing could help provide a solid evidence base for informed local policy-making, and the distribution of regeneration funds. In addition, it could identify causes behind an improvement in wellbeing in certain areas, which could help in policy decision.

For this purpose, a dataset from the London Datastore [2] was chosen which contains a wealth of information for investigating which indicators contribute most to wellbeing (output variable) across 625 London wards. These wellbeing scores present a combined measure of wellbeing indicators of the resident population based on different factors such as health, childhood obesity, incapacity benefits claimant rate, economic security, safety, education, children, families, transport, environment and happiness. Each of these indicators is subdivided into further categories resulting in 76 attributes in the dataset. For the majority of this study, only the 2011 wellbeing scores were used as this would give a larger variable pool after merging with the 'census 2011' and 'access to open space' datasets [2]. Since most counts were represented as a percentage of the total population in the dataset, we did not need to normalize by population. However, a z-score standardization was applied to scale the range of the variables before Geographically Weighted Regression (GWR) modelling. Our aim is to identify the most discriminating and generalisable socio-economic variables that appear to drive spatial differences and then compare the results to existing theories published by the Office for National Statistics (ONS) [3] and other government and public health organisations. These lead us to the following research questions:

- How did wellbeing scores vary across London in 2011 and how did they compare to 2010 and 2012?
- Which factors had the most impact on wellbeing scores across London in 2011 and can they be used to explain geographic differences in wellbeing across London?

2.TASK AND APPROACH

The data was first visualised in Tableau [4] to build simple bar charts and spatial maps to understand the dataset structure and variables in more detail. Data Analysis was then performed in the R programming language [5] using a number of packages for carry out numerous tasks: importing raw data in different formats (readr, readxl, rgdal), tidying and wrangling the data (tidyr, dplyr), modelling (GWmodel) [6] and visualisation (ggplot2, tmap, Rcolorbrewer). Following merging and cleaning of the datasets, 20 variables were produced to characterise the 12 indicators described in section 1. The analytical tasks suitable for answering the research questions in section 1 are listed below. For each task the visual and computational techniques which will be used have been described with a justification in each case.

- Exploring spatial variation in wellbeing scores in 2011 across different London Wards. How does this compare to scores in previous and future years (from the 2009 to 2013 period)?

We will use a chloropeth map with a divergent colour scheme to explore spatial differences in wellbeing attributes. This helps to distinguish wards which have high wellbeing scores from wards with low wellbeing scores. Using the 2011 year as the baseline, the percentage change in wellbeing from this year compared to previous years (2009, 2010) and future years (2012, 2013) will also be explored

- Which socio economic variables have the biggest impact on well-being in London in 2011?

Using correlation matrix and Pearson correlation plots, we can investigate the relationship between multiple variables. This will tell us which independent variables vary linearly with the output variable and give us some initial indication of collinearity between different variables. Problems in collinearity can be further investigated by building a multi-regression model of the chosen variables and using the Variance Inflation Factor (VIF) technique [7]. This measures the degree of inflation of variance of regression coefficients compared to when the independent variables are not linearly related. VIF scores less than 5 indicate low level of collinearity whilst VIF values greater or equal to 10 indicate that the predictors are highly correlated.

- Does the relationship between wellbeing and indicators exhibit any non-stationarity? How can we model this effectively?

Geographically Weighted Summary Statistics (GWSS) [8] can be used to investigate the geographically weighted versions of descriptive statistics. Geographically weighted mean and standard deviation will then be visualised on a chloropeth map. Use of computational techniques like multivariate regression modelling will then allow a more formal investigation of possible explanatory variables. This includes a global model and geographically weighted regression (GWR) model [9] using the variables selected previously from VIF and correlation matrix analysis. Spatial chloropeth maps of residuals for global and geographically weighted models will infer if GWR provides a better model as it accounts for spatial variation in variables.

- Can this non-stationarity in coefficients produce obvious spatial clusters which can be attributed to local environment changes in the given regions in London?

Here we will use a computational technique called k-means clustering [10] using the geographically weighted regression coefficients as features, to investigate any interesting spatial pattern in clusters which can be attributed to local environmental changes. The data will be scaled so that no single variable is weighted more due to its distribution. A range of cluster numbers (k) will be selected manually. The optimal solution will be met when the selected number of clusters produces the lowest withinness score i.e. total sum of squares of each cluster. The number of iterations will also be increased to check for any improvement in the clustering withinness scores.

3.ANALYTICAL STEPS

To answer our first analytical question in section 2, we can visually inspect how 2011 wellbeing scores vary spatially across London wards by plotting the raw scores on a Choropleth map as in figure 1a). To differentiate between negative and positive wellbeing scores, a diverging colour scheme from red to green is used. Figure 1b and c show a change in wellbeing scores in 2010 and 2012, calculated relative to wellbeing scores in 2011. Wellbeing scores in 2012 got progressively worse in wards outside Central London compared to 2011. In 2010, a larger proportion of wards in London had better wellbeing scores relative to 2011, with few wards on the outskirts and central London showing worse wellbeing scores.

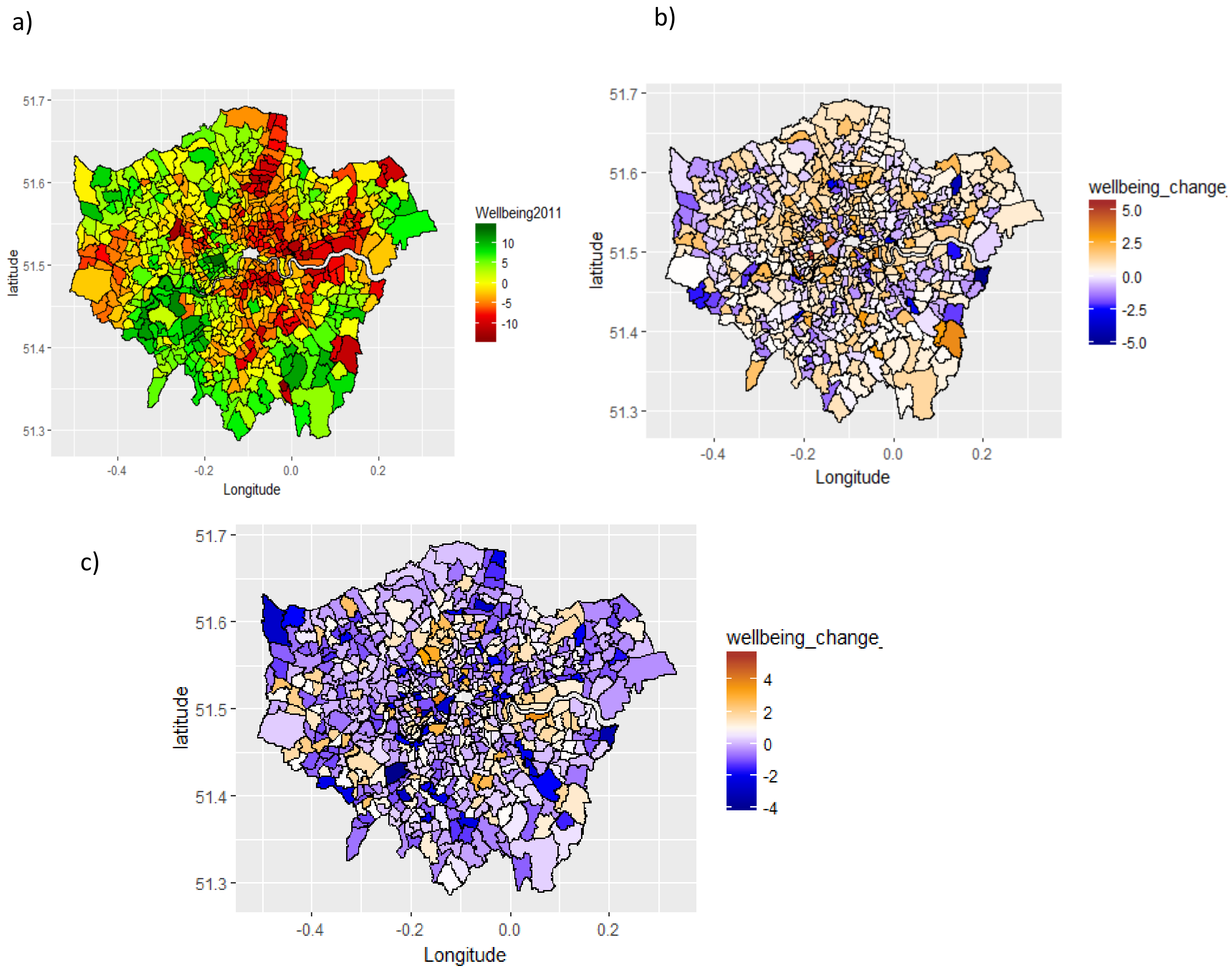


Figure 1: Choloropeth maps showing a) wards with positive(green) and negative(red) wellbeing scores. b) changes in wellbeing in 2010 relative to 2011 c) changes in wellbeing in 2012 relative to 2011

The investigate our second analytical question, scatter plots and Pearson Correlation coefficients were used to investigate each variable's relationship with wellbeing. Figure 2 shows an example of such scatterplots. Good health and unemployment show positive (correlation:0.65) and strong negative correlation (correlation: -0.86) with wellbeing but surprisingly good access to nature showed no significant correlation to wellbeing. Variables with poor correlation were removed from subsequent analysis. The only exception being the happiness score (correlation:0.3), as publications from the ONS deemed this factor to be very important for determining subjective wellbeing.

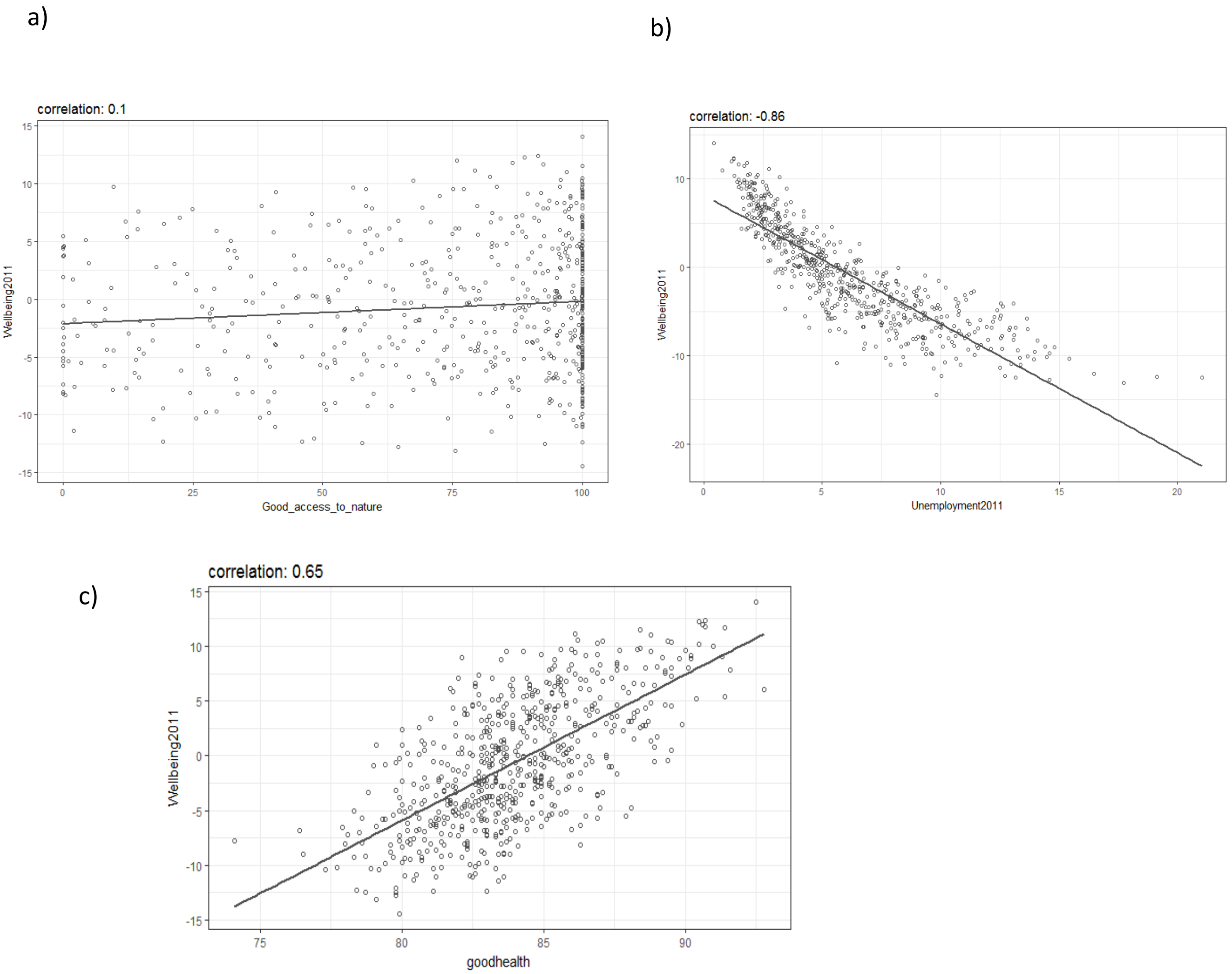


Figure 2. Scatter plots with trend line and correlation coefficients for wellbeing 2011 scores and explanatory variables a) good health, b) unemployment and c) good access to nature.

The correlation between the selected variables was investigated further by computing a correlation matrix as in figure 3. We can see that variables which correlate strongly with wellbeing also correlate strongly with many other variables. This problem of collinearity was investigated further by computing VIF scores [7,11]. Table 1 summarises the VIF scores generated for each of these variables and how these scores reduced after removing four variables. This results in final 10 variables which will be used for the modelling tasks.

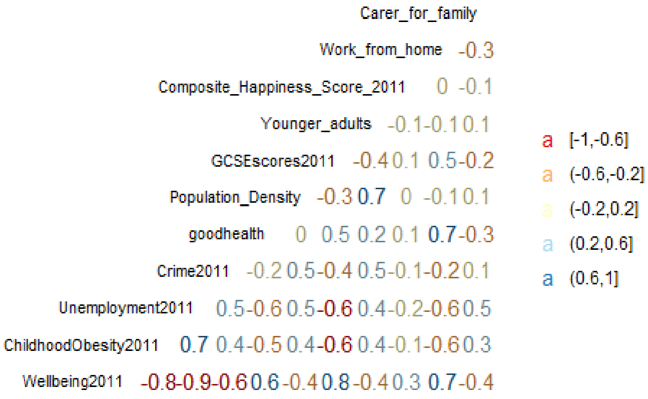


Figure 3: Correlation matrix of explanatory variables and wellbeing

Variable	***VIF_scores before***	***VIF_scores after***
School_Absence	*2.2*	-
Life_Expectancy	*2.5*	-
Childhood_obesity	*2.4*	*2.1*
Population Density	*3.1*	*2.4*
Happiness_Score	*1.1*	*1.1*
Crime	*1.6*	*1.6*
GCSE_scores	*2.1*	*1.8*
good_health	*5.4*	*3.4*
Unemployment	*5.2*	*3.6*
Professionals	*10.5*	-
Younger_adults	*5.5*	*3.4*
Work_from_home	*5.3*	*2.2*
Carer_for_family	*1.7*	*1.3*
dependent_children	*8.1*	-

Table 1: VIF scores before and after manipulating variables

Using this smaller pool of variables, a refined multivariate regression model was fitted (R^2= 0.92, p < 0.01). Figure 4 shows the residuals plotted on a chloropeth map, showing some clusters of red and blue residuals in certain parts of London (north west and centre). Since we can see some spatial auto correlation of residuals, this suggests that another model is required which takes into account the spatial variation [9,11]. A fundamental element in GW modelling is that it quantifies the spatial relationship or spatial dependency between the observed variables and will hence be a good choice for the next steps in the analysis.

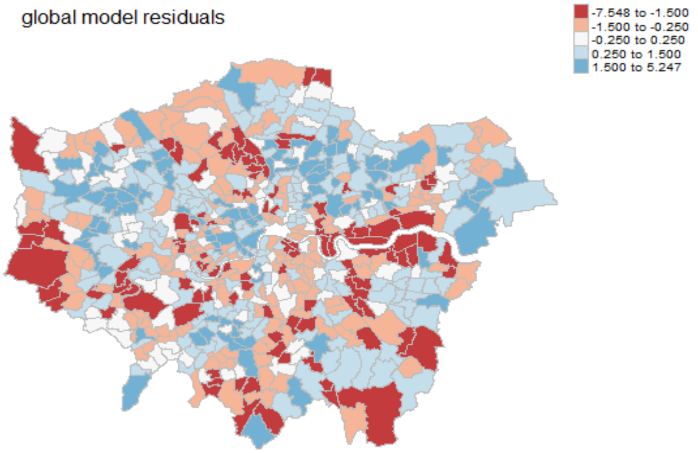
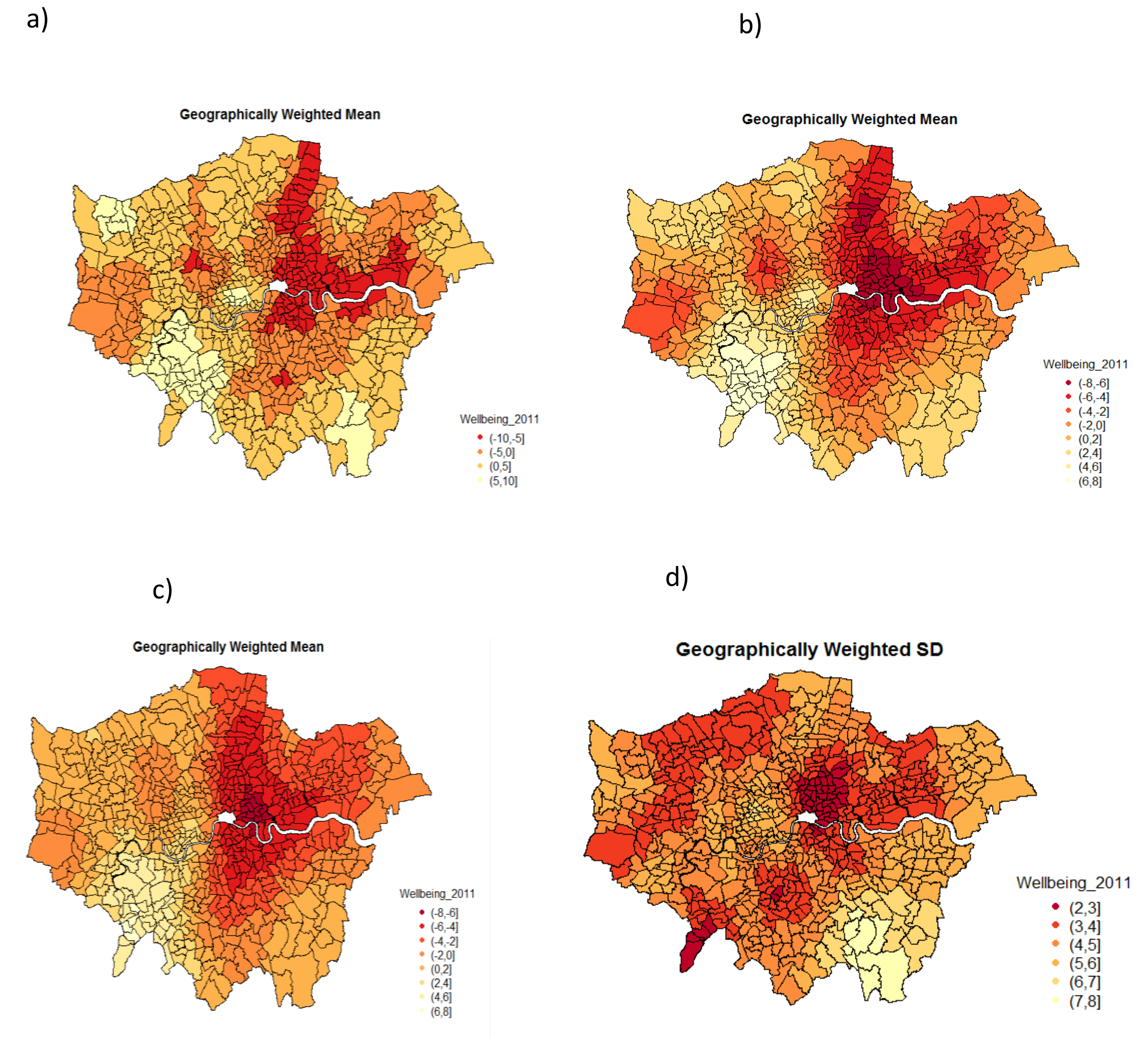


Figure 4. Spatial residual variation for Global regression model

Modelling of the spatial relationships was achieved through a weighting function with Euclidean distance measure and a bisquare kernel function. The GWmodel package in R provides an option for automated bandwidth tuning for GWR, but not for GWSS [6,8]. Alternatively, the bandwidth can be manually chosen as shown in the example in figure 5(a-c). The smoothing effect increases with increasing bandwidth. The optimal bandwidth of 50 local points produced a dark red cluster in north and east-central side of London, a bright red pattern running from north to south and a bright yellow cluster in the south west. The GW standard deviation (figure 5d) shows a high variation in wellbeing, concentrated in the south east. The GW correlation coefficients for unemployment, crime and good health (figures 5e-g) against wellbeing show that the strength of the relationships between these variables varies spatially. Unemployment and crime are both negatively correlated with wellbeing, but crime is more strongly correlated in the south-east wards and unemployment more so in the south, east and some wards in the north-west. Good health is positively correlated in most parts of London except for some wards in the east (red cluster).



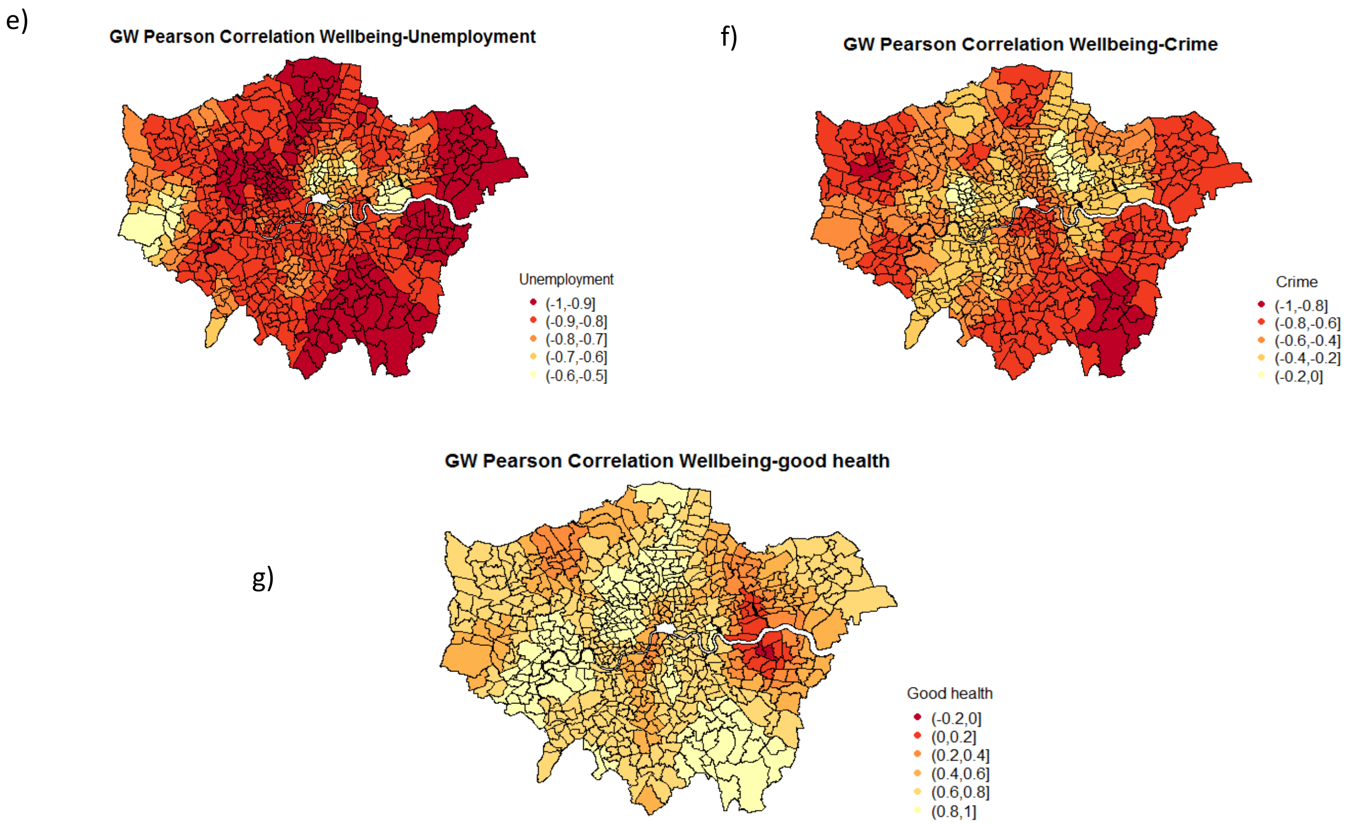


Figure 5. Chloropeth maps with GW mean wellbeing scores (bandwidths: 20(a), 50 (b), 100 (c)), Standard deviation (d) GW correlation coefficient maps for Unemployment (e) Crime (f) good health (g) against wellbeing

Figure 6 shows the residuals of the GWR model results for manually optimised bandwidth (50) and auto tuned bandwidth (175). The results of manual tuning produced much smaller residuals across the whole of London compared to auto tuned bandwidth, where large negative residuals (dark red) can be seen in a number of wards. The results of auto tuning produced similar pattern to the global model (figure4).

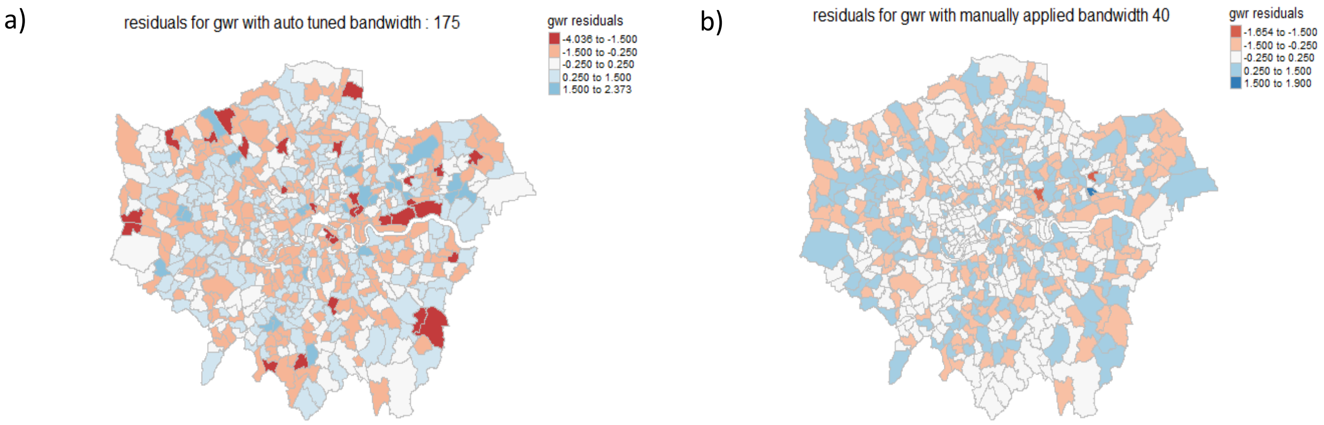


Figure 6. GWR model residuals for a) auto-`tuned bandwidth (175 local points) b) manually tuned bandwidth (50 local points)

For the final analytical task, we explore how the explanatory variables vary spatially through k-means clustering on the geographically weighted regression coefficients [10]. This will allow identification of wards that share similar combinations of relationships. A range of k values were used to generate clustering solutions. An example of clustering output using 3,4 and 5 clusters is shown in figure 7. We can see that using 3 clusters produces the most discernible cluster solution. The pattern of clusters however is not well defined and inconclusive with two clustered groups of wards (green and purple) mainly concentrated on the outskirts (and some parts of Central London) and another group of wards concentrated in Central London and some parts of the outskirts.

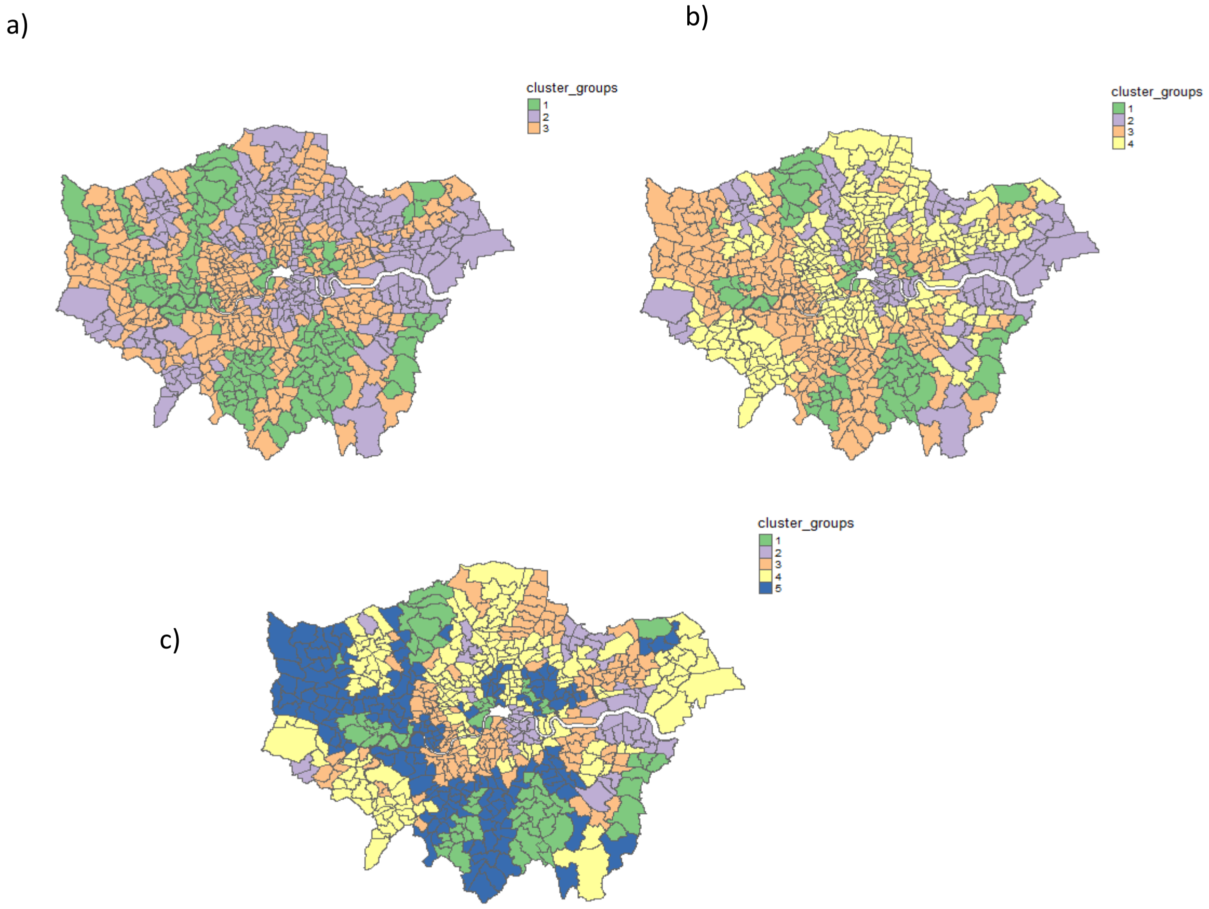


Figure 7. a) Chloropeth maps of k-means clustering solutions for a) k= 3 b) k= 4 c) k =5

4.FINDINGS

The highest wellbeing scores were found in wards concentrated in the south west of London (Twickenham, Richmond park, Brentford, Ealing and Acton Central, Putney, Wimbledon, Chelsea etc.). This can be attributed to high levels of affluence in these areas, coupled with high levels of social cohesion. Concentrations of lower wellness scores (red coloured wards) were clustered around north and north-east London and some parts of the south and east London (Wards in between Oval, Crystal Palace, Woolwich, Lewisham). Additional clusters were also seen in more deprived areas in part of north-west London (Dollis Hill, Willesden) and along the outskirts on the west (Hayes and Harlington). The worst scores (dark red) being in wards of Upper Edmonton, Tottenham Hale, Bow East/West which have been known to be impoverished areas. Wellbeing scores got worse in many parts of

London (more concentrated in the outskirts) in 2012 compared to 2011. In contrast, wellbeing was better in most wards around London (except for a few on the outskirts) in 2010 compared to 2011. This agrees with some of the findings in the press regarding problems with wellbeing getting progressively worse with every year. Economic and political events during that period may well have been a driving force behind these observations e.g. UK general election in 2010, recession in 2010.

The scatter plots and correlation matrix confirm that Unemployment correlated most strongly with wellbeing along with long term sick and disabled. Surprisingly access to nature showed little or no correlated with wellbeing. Given the distribution of wellbeing scores across London, one would expect there to be a positive correlation as wards on the outskirts with large greenspaces and parks displayed higher wellbeing scores. Similarly, other explanatory variables like access to transport deemed important for wellbeing, also showed no relationship. The possible reasons for this will be discussed in the next section.

The relationship between wellbeing and indicators exhibited non-stationarity. Residual plots from GWR and global model revealed that the spatial relationships between explanatory variables and wellbeing were best modelled using GWR and manually tuning the bandwidth to 40 local points (rather than using the automated method).

The subsequent use of clustering on the GWR coefficients produced some distinct clusters on the outskirts of London but no significant pattern which could be used to investigate local environmental factors which affected the non-stationarity of coefficients. This will be discussed more in the next section.

5.CRITICAL REFECTION

Visual Analytics approaches to answering research questions

The findings showed that certain variables were more discriminatory than others e.g. those relating to income status and long term illness. However, by including other variables which were less discriminating, like safety, childhood obesity, family size, it was possible to build a more robust model of the spatial relationship between wellbeing and factors affecting it. Since we were interested in wellbeing at the population rather than individual level, attributes were chosen, aggregated and normalised to relate to a population rather than an individual. Since, we were interested in studying wellbeing at a population level, chloropeth maps were used for a number of tasks in section 3. These include: the analysis of spatial variation of wellbeing in 2011 and comparing changed across 3 years, analysing model residuals and for studying how relationships between variables and wellbeing varied spatially. The issue of collinearity for building the models and carrying out clustering analysis was accounted for by generating a correlation matrix and inspecting how the VIF scores of the fitted model coefficients change when variables are added, removed or merged. These are widely used techniques in the literature [7]. A standard partition based technique: k-means clustering, was used to condense the high dimensional dataset into clusters based on similar patterns within groups. Another option was to use a technique called Hierarchical clustering where partitions can be visualised through a tree structure (dendogram) and does not require the number of clusters as input. However, k-means clustering is known to produce much denser clusters [10]. Furthermore, the optimal cluster solution was decided by inspecting the within group sum of squares for each cluster solution. Alternatively, silhouette plots [12] could have been used for validation of cluster membership. The stopping criterion for producing each cluster solution was determined by manually increasing the number of iterations in the k-means function in R until no changes in cluster assignments was observable.

Implications of the findings and limitations

Managing policy design which requires us to measure wellbeing in different populations in different areas that may be affected by policy [13]. The chloropeth maps revealed an interesting pattern in geographically weighted mean wellbeing scores across London wards where wellbeing scores were lower in certain wards closer to north-central and east-central part of London, and higher in the south-west. The fact that the UK was plunged into a recession towards the autumn of 2010, may have been one of the driving factors behind the general worsening of wellbeing scores from 2010-2011 and 2011-2012. This also suggest that the policies in place during those years was not sufficient as populations in more areas of London were also expressing lower wellbeing. In contrast to results from the commonly used global regression model, geographically weighted regression showed that certain variables increased or decreased wellbeing scores only in certain areas. This is important as it would allow the local authorities to build a strong case for certain programmes e.g. unemployment [1,13].

This study has a number of limitations which need to be made clear. The data in this report has been aggregated to ward level and would hide variation which is more local. Data availability for small areas is far more limited. Furthermore, wellbeing score metrics computed on the London Datastore [2] were only available at ward level, which limits use of more local variables even if they were available. The indicators on the London Datastore were used as a basis for choosing the variables. More specific factors like loneliness and mental health/depression (indices which are difficult to quantify) can also have a drastic effect on wellbeing. Given the scope and time frame of this study, it was not possible to carry out a thorough literature review to generate a larger pool of variables to start with, which could have improved the model. The outputs from clustering analysis is generally very difficult to interpret especially when clusters are not limited to distinct regions on the map. Furthermore, interpreting spatial patterns in factors affecting a complex topic like wellbeing requires significant domain expertise in public health policy. However, further analysis could have been carried on the geographically weighted correlation coefficients for each variable to paint a clearer picture. A dimensionality reduction technique like principal component analysis [14] prior to running k-means clustering could have been another option, although issues can arise when interpreting the two transformed dimensions.

Generalisability to other domains

The visual and computational techniques used in this report are applicable to a number of domains ranging from politics, social science, health, agriculture/forestry, climate etc. However, the applicability of certain techniques are problem specific. Since most of the analysis in the study was carried out on spatial data and at population level, it was sensible to use chloropeth maps and GWR to investigate the spatial patterns in more detail. This may not necessarily be suitable for studies where spatial data is not available or studies with different aims or other factors invalidating model assumptions e.g. fitting a time series model or suing a space time cube would be more appropriate for visualising changes in individual house prices over time, as house prices can change very rapidly over very short distances and will not be appropriate more fitting a GWR model. The choice of variables in this study was made easier because of the information from publications available from the London Datastore and Office for National Statistics website [1,2,3,13]. In other topics which may not be so publicly scrutinised and still in the very early phases of research, it may not clear which variables affect the overall outcome. Future work would be to collect a wider set of variables to better characterise certain indicators like transport and access to greenspace which may then produce a stronger correlation with wellbeing scores and reveal some potentially interesting patterns. It would also be interesting to extend this study to the whole of the UK and carry out a deeper analysis into patterns in other years.

6.REFERENCES

[1] Spence, A., Powell, M. and Self, A., 2011. Developing a Framework for Understanding and Measuring National Well-being. Office for National Statistics at http://www. ons. gov. uk/ons/guide-method/userguidance/well-being/publications/previous-publications/index. html (accessed 17 October 2013).

[2] London Datastore: Official site providing free access to ward well-being data, Access to Public Open Space and Nature by Ward data and census data. from the Greater London Authority. Datasets used:
https://data.london.gov.uk/dataset/london-ward-well-being-scores
https://data.london.gov.uk/census/data/
https://data.london.gov.uk/dataset/access-public-open-space-and-nature-ward

[3] Beaumont, J., 2011. Measuring national well-being: Discussion paper on domains and measures. Newport: Office for National Statistics.

[4] Murray, D.G., 2013. Tableau Your Data!: Fast and Easy Visual Analysis with Tableau Software. John Wiley & Sons.

[5] R Core Team., 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[6] Gollini, I., Lu, B., Charlton, M., Brunsdon, C. and Harris, P. (2015), GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models. Journal of Statistical Software, 63(17):1-50.

[7] O'brien, R.M., 2007. A caution regarding rules of thumb for variance inflation factors. Quality & Quantity, 41(5), pp.673-690.

[8] Brunsdon, C., Fotheringham, A.S. and Charlton, M., 2002. Geographically weighted summary statistics—a framework for localised exploratory data analysis. Computers, Environment and Urban Systems, 26(6), pp.501-524.

[9] Fotheringham, A.S., Brunsdon, C. and Charlton, M., 2003. Geographically weighted regression: the analysis of spatially varying relationships. John Wiley & Sons.

[10] Jain, A.K., 2010. Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), pp.651-666.

[11] Wheeler, D. and Tiefelsdorf, M., 2005. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. Journal of Geographical Systems, 7(2), pp.161-187.

[12] Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, pp.53-65.

[13] Hicks, S., Tinkler, L. and Allin, P., 2013. Measuring subjective well-being and its potential role in policy: Perspectives from the UK office for national statistics. Social Indicators Research, 114(1), pp.73-86.

[14] Charlton, M., Brunsdon, C., Demsar, U., Harris, P. and Fotheringham, S., 2010. Principal components analysis: from global to local.