Rory Hurley
INM433 Report

Socio-demographics of Diabetes Prevalence in London

1. Introduction, Motivation & Data

Diabetes is a medical condition where the body is unable to regulate blood sugar levels, causing blood pressure to become too high over a long time. This can be extremely damaging to a person's health, with short term effects such as fatigue, vomiting and collapsing, and long-term effects including cardiovascular disease, strokes, and loss of sight *(Kitabchi et al., 2009)*. Since 1996, the number of people diagnosed with diabetes in the UK rose from 1.4 million to 3.5 million, and is predicted to rise to 5 million by 2025 *(Diabetes.co.uk)*. Type 2 diabetes is avoidable and can be prevented through proper nutrition, and studies have found links between socio-demographic indicators and diabetes prevalence. A study from 1997 found Type 2 diabetes is more than six times more common in people of South Asian descent and up to three times more common among people of African and African-Caribbean origin, whilst Gaskin et al., (2014) concluded that individual poverty increased the odds of having diabetes for both white and black people.

This project combines data on diabetes prevalence and socio-demographic indicators and apply visual analytic methods with the aim of understanding the spatial patterns of diabetes prevalence in London and the drivers behind it. The project uses ward-level estimates on the percentage of people in England based on the GP practices that its residents attend. This data has its limitations in that, some divergence between areas served by an individual GP practice is likely to be lost due to the fact that not all people visiting a GP are the same. Despite this, it is a reliable and the most detailed estimate based on the data available. This data is combined with data compiled from the 2011 census and estimates from some years since then. The data is reliable as much of it is as up to date as possible, and compiled from official sources.

2. Research Questions

1. RQ1 Does diabetes prevalence vary spatially in London? Where are the 'hot and cold' spots?

2. RQ2 What are the most influential socio-demographic factors that are linked to diabetes prevalence, and can these be modelled to explain it?

3. RQ3 Are the socio-demographic factors that affect diabetes the same across London?

3. Tasks and Approach

3.1. Diabetes Spatial Variation

The first analytical task will be to describe how diabetes prevalence varies across London. Choropleth mapping will be used to gain a holistic overview of diabetes in London, and identify some patterns, such as regional variation some suspected pockets or areas with high diabetes prevalence. This visual technique will allow for a second task to be addressed - to relate different areas of London. Computational methods will be applied by using spatial statistics. Firstly, by using a Global Moran's I test, as used by Fu et al., (2014) to evaluate whether the data is exhibiting statistically significant evidence of clustering; more reliable than assessing the choropleth map. A Getis-Ord Gi* statistic will then be applied to the data. This computational method will identify statistically significant spatial clusters of high diabetes prevalence (hot spots) and low diabetes prevalence (cold spots). The output of the computational analysis will enable choropleth mapping of hot and cold spots, and thus enable the third analytical task to be carried out – to locate specific areas of high and low diabetes prevalence.

3.2. Socio-demographic Factors of Diabetes Prevalence

The census data used in this project contains socio-demographic data for each London ward in the study. There are 64 features for each ward. It is unlikely that every one of these indicators will be a factor affecting diabetes prevalence, so feature selection will need to be carried out to reduce dimensionality, and allow for computational models to be built. The first task will be to locate or identify the most important features that affect diabetes prevalence. To begin, all the features will be visualised using a correlation matrix. A random forest will be employed as the computational technique which will reduce the dimensionality of the census data, keeping the key of socio-demographic variables related to diabetes, which will then be visualised using an interactive dashboard. A second task will be carried out - to relate these factors to diabetes prevalence. Each census ward will be visualised on a scatter plot with a colour scale reflecting its diabetes prevalence, and linear regression will be used to compute the degree of relation between each independent variable and the target variable. The plots will be visually enhanced to gain more insight into spatial patterns colouring points by borough. Task 3 which will be to compare the visual output of this with previous outputs. A global linear model will be computed using the features and the diabetes prevalence figures. The residual errors between the global model and each ward will be visualised using choropleth mapping, allowing for patterns to be assessed.

3.3. Spatial Variation of Socio-demographic Factors

From the analytical steps taken to address RQ1 and RQ2, we will have outputs that will help us address RQ3. RQ2 shows we can fit a multivariate regression model, RQ1 outputs indicate that diabetes prevalence varies spatially. Therefore, geographically weighted regression will be applied. This computational technique will allow geography to be considered, and assess whether and how the explainable variables vary spatially across London. The visual output from this will be choropleth map for each variable, with each ward coloured according to its GW correlation coefficient. Finally, hierarchical clustering analysis will be applied to cluster areas based on their GW coefficients. The output of these will be visualised in the form of a choropleth map and adjacent feature-cluster density plots.

4. Analytical Steps

4.1. Diabetes Spatial Variation

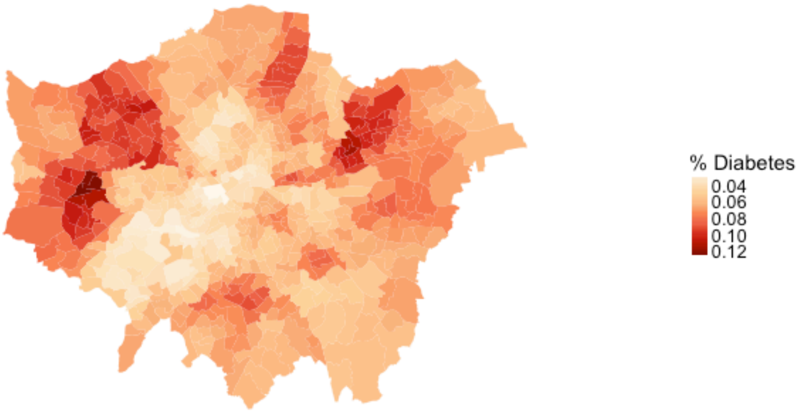
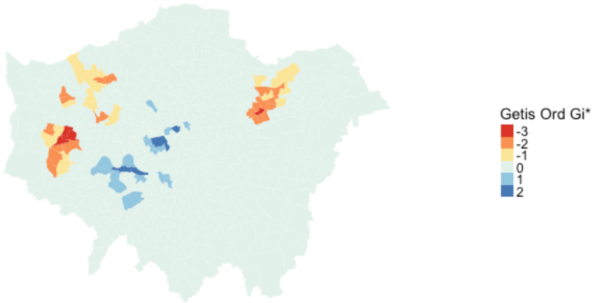


Figure 1- Diabetes Prevalence in London

Choropleth mapping (Figure 1) shows diabetes prevalence varies from 4% of the local population to as high as 12% in some areas, suggesting that diabetes does not appear to be consistent across London. Areas in central to Southwest London appear to exhibit low levels of diabetes prevalence whereas there appears to be high rates of diabetes in areas of East and West London. Visual assessment of the data indicates some level of autocorrelation as values do appear to be clustered in areas. Global Moran's I, a test to confirm whether there is statistically significant evidence diabetes clustering in London, was computed. The test used 'Inverse distance' as a parameter to conceptualise spatial relationships, which assigns higher weights to closer locations, in accordance with Tobler's first law of geography (1970). 'Euclidean Distance' as the distance metric, which means the algorithm uses straight line distance between areas. The result of the test was that there is a strong clustering pattern of diabetes London at the 99% confidence level. The test was repeated using 'Countiguity' as the distance metric which had similar results.

This visual analytic process proved that there was clustering and spatial autocorrelation, but not where these areas were. The Getis-Ord Gi* statistic was computed to find statistically significant hot and cold spots of diabetes. The test was initialised using parameters 'Inverse Distance' and 'Euclidean Distance'. The visual output of the computations can be seen in Figure 2a. This confirms suspicions of spatial clustering of high diabetes prevalence in West and East London, and Southwest London showed clusters of cold spots, indicated by the blue areas. Figure 2b is the output of the same test run with 'Contiguous Edges' as the distance metric. This exhibited similar patterns, but covered larger areas. This parameter also highlighted some other pockets in South London and North London.



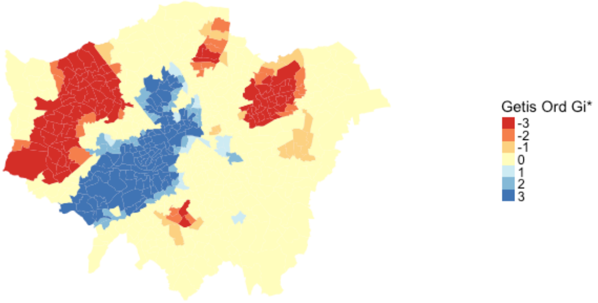


Figure 2a- Visualisation of Getis-Ord Gi Statistic Results Getis-Ord Gi* Statistic Results (Inverse distance & Euclidean Distance)*
Figure 2b- Visualisation of Getis-Ord Gi Statistic Results (Inverse Distance & Contiguous Edges)*

4.2. Data Modelling

Outputs from 4.1. coupled with domain knowledge indicates that there could be specific socio-demographic factors that influence diabetes prevalence. Census data was used to explain what the drivers are behind these patterns. The 64 variables were visualised on a correlation plot for initial visual exploratory analysis, shown in Figure 3.

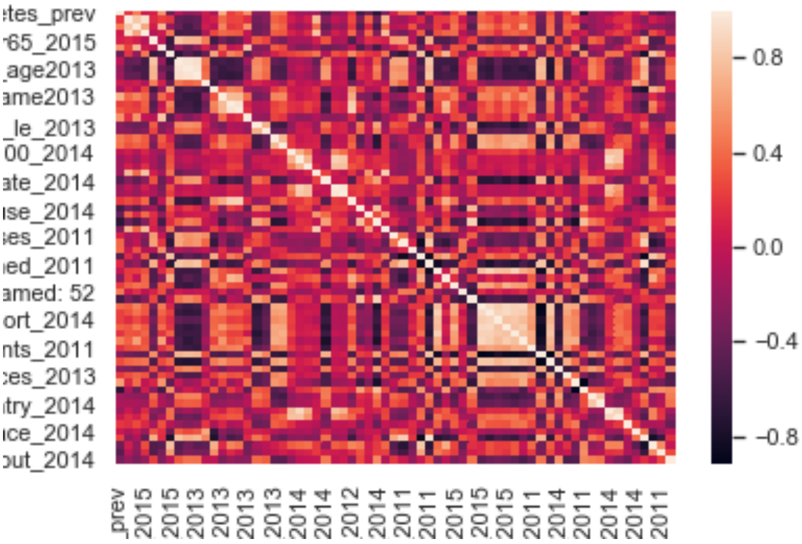
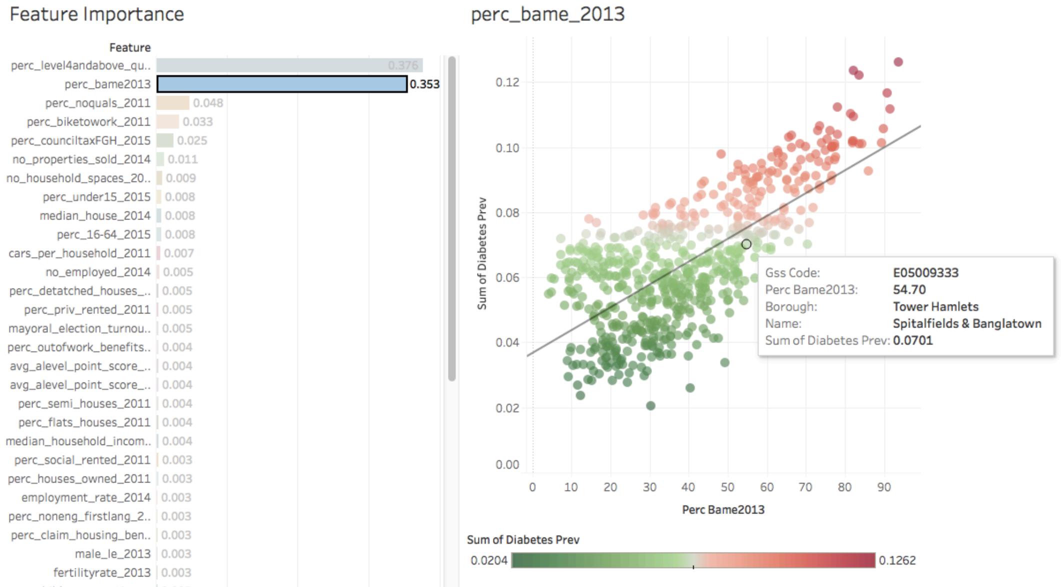


Figure 3- Census Variables Correlation Matrix Heatmap

The plot informs us that there is a mixture of strong positively and negatively correlated variables that could be used for modelling, but the plot does not allow for easy interpretation. Features were selected computationally via a random forest model on the 64 independent variables, with diabetes prevalence as the dependent variable. From this model, the relative importance of each feature could be derived; the output of which was displayed as an interactive bar chart in Tableau. This approach meant that and each feature could be clicked on from the bar chart and users could not only compare relative feature importance, but relate the feature to its effect on the target variable using the scatter plot split view (Figure 4).


Feature Importance
Feature
perc_level4andabove_qu.. 0.376
perc_bame2013 0.353
perc_noquals_2011 0.048
perc_biketowork_2011 0.033
perc_counciltaxFGH_2015 0.025
no_properties_sold_2014 0.011
no_household_spaces_20.. 0.009
perc_under15_2015 0.008
median_house_2014 0.008
perc_16-64_2015 0.008
cars_per_household_2011 0.007
no_employed_2014 0.005
perc_detatched_houses_.. 0.005
perc_priv_rented_2011 0.005
mayoral_election_turnou.. 0.005
perc_outofwork_benefits.. 0.004
avg_alevel_point_score_.. 0.004
avg_alevel_point_score_.. 0.004
perc_semi_houses_2011 0.004
perc_flats_houses_2011 0.004
median_household_incom.. 0.004
perc_social_rented_2011 0.003
perc_houses_owned_2011 0.003
employment_rate_2014 0.003
perc_noneng_firstlang_2.. 0.003
perc_claim_housing_ben.. 0.003
male_le_2013 0.003
fertilityrate_2013 0.003
perc_bame_2013
Sum of Diabetes Prev
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
50
60
70
80
90
Perc Bame2013
Sum of Diabetes Prev
0.0204
0.1262

Gss Code: E05009333
Perc Bame2013: 54.70
Borough: Tower Hamlets
Name: Spitalfields & Banglatown
Sum of Diabetes Prev: 0.0701

Figure 4- Dashboard allowing for features selected by random forest to be interactively selected and correlated with diabetes prevalence.

The top 8 features from the random forest were then compared simultaneously to assess the correlations, and were visualised on adjacent scatter plots. This allowed for easy comparison between factors that correlate with diabetes prevalence. The points (wards) were coloured according to its borough, as this would inform the user about clustering within the scatter graphs, and perhaps uncover some non-obvious patterns in the data. An example of this can be seen in Figure 5, where wards Barking and Dagenham tend to show signs of clustering.

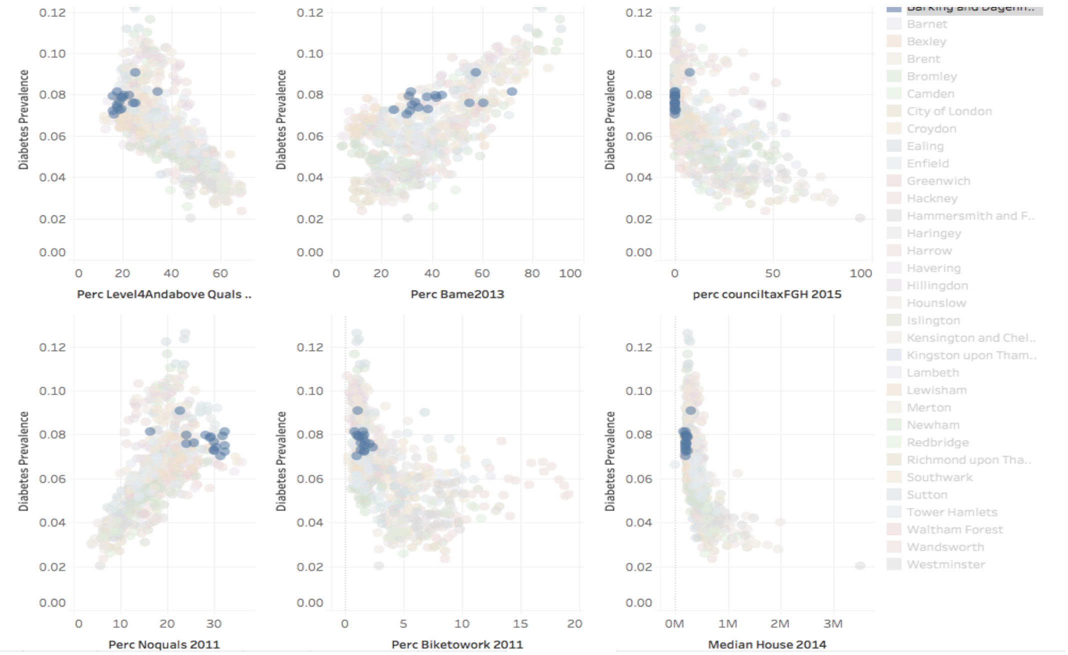


Figure 5- Dashboard with scatter plots coloured according to borough

Following these computational and visual steps, the evidence suggested that the selected variables could be used to explain diabetes prevalence. Linear regression models were decided on as a suitable computational method to help explain the phenomenon. The best 6 variables were selected that could be used in the final model to avoid complexity. To account for model bias from collinearity, Variance Inflation Factor (VIF) was calculated on the remaining variables. 'Median House Price (2014)' was dropped due to a high VIF score, likely due to its association with 'Percentage of Council Tax FGH Band' column. House price was dropped as there is high variability across London, and not necessarily a reflection of socio-economic conditions. The remaining 5 variables were kept as they had low VIF scores, and provided a good range of socio-demographic variables. Linear models were applied to the data and residual error could be calculated for each ward. The computational analysis is visualised on the plots on Figure 6.

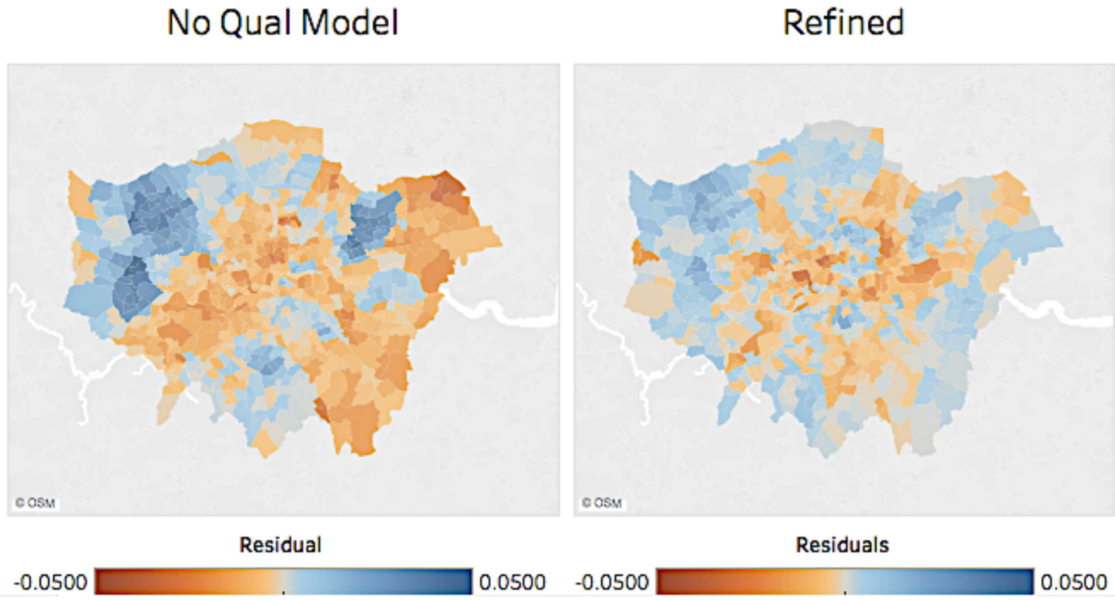


Figure 6- Visualisation of model residuals betweem univariate modelling (no qualifications) and refined multivariate model.

Figure 6 shows a univariate linear model for no qualification and diabetes, there are clear signs of autocorrelation in some areas. Areas with negative residuals (brown) show that the univariate model predicting diabetes as a linear function of the local population with no qualification, overestimates the diabetes prevalence in these locations. When the variables are combined into a refined, multivariate linear model, the regional variation in residual differences are significantly reduced, although evidence of spatial autocorrelation remained.

4.3. GWR

Geographically weighted regression was used to account for spatial variation in the multivariate model. GW summary statistics were computed to support understanding of how correlations between diabetes prevalence and the explanatory variables vary spatially. The correlation coefficients were visualised on choropleth maps in Figure 7.

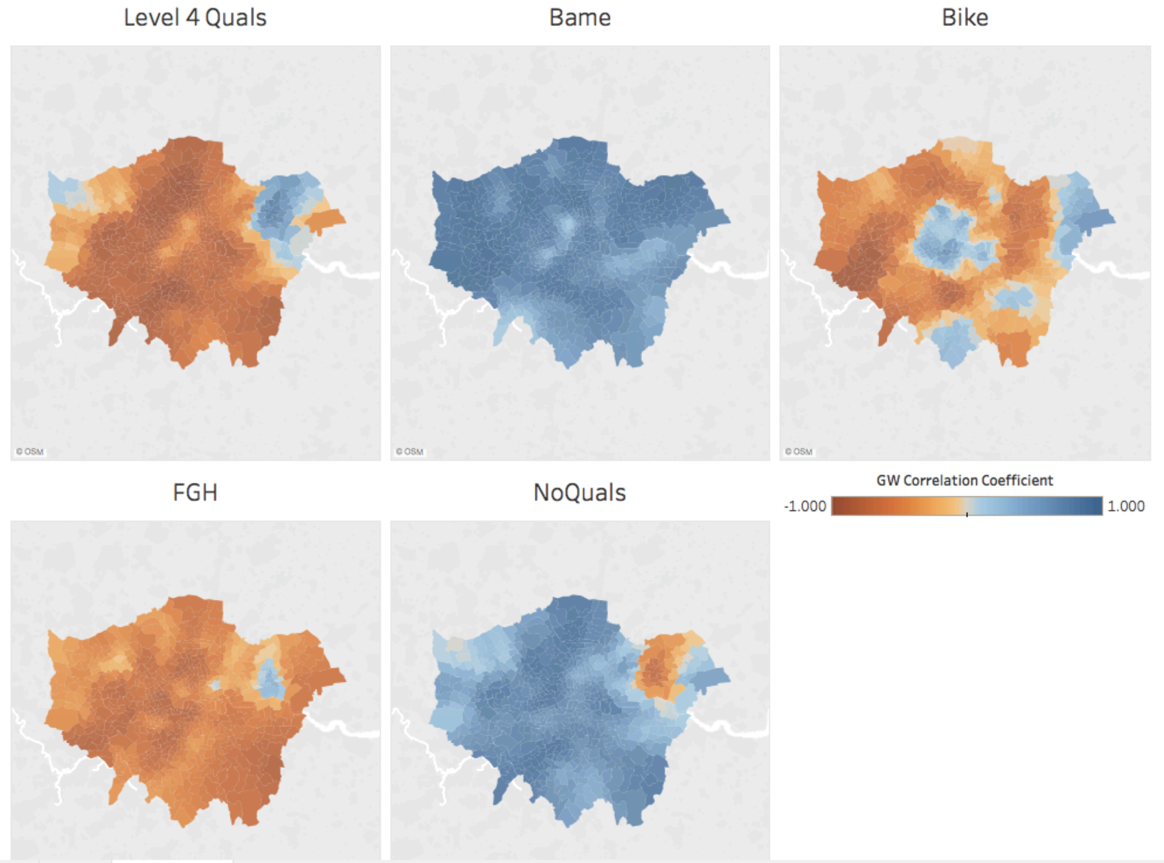


Figure 7- Choropleths illustrating geographically weighted coefficients for each variable

Brown areas in the figures indicate negative correlation with diabetes prevalence, and blue areas positive correlation. The figure suggests a key link between education and diabetes. This is consistent over most of London, except for a pocket in East London, where the relationship is the opposite. Proportion of FHG council tax band properties in wards is negatively correlated with diabetes. This is understandable as, generally, areas like these with higher cost properties will potentially mean a more affluent, educated population and therefore less likely to get diabetes. This is with the exception of Newham, the blue area in the east.

Drawing comparisons between maps is difficult and can be subjective, so wards were summarised by their GW correlation coefficient and hierarchical cluster analysis (HCA) used to identify groups of wards with similar combinations of relationship. The algorithm was tested with 3-6 clusters and dendrogram assessed for each, before it was decided that 4 would be an optimal number of clusters to use. Collinearity was also assessed as before, but little collinearity was present.

5. Results

5.1. Diabetes Spatial Variation

Following visual assessment of the diabetes choropleth, the Moran's I test with parameters 'Inverse Distance' and 'Euclidean Distance' gave results that there was evidence of spatial clustering of diabetes prevalence at the 99% confidence level. The Getis-Ord Gi* test showed that there were specific wards that were 'hot spots' of diabetes prevalence. These were primarily located in Ealing and Newham. Ealing contained 3 wards that were determined to show hotspots at the 99% confidence level, whereas Newham contained 1.

There were cold spots located in Richmond and Kensington & Chelsea in southwest London at 99% confidence level. These findings successfully address RQ1, proving that diabetes prevalence varies across London and identified the hot and cold spots.

5.2. Socio-demographic Factors of Diabetes

Random forest and linear regression found that the main explainable variables were related to socio-demographics. Further visual analysis of these results indicated that there were some exceptions, and patterns were highlighted such as clustering of wards in some boroughs indicating that there was evidence for spatial variation. Using a multivariate model reduced overall residual error compare to univariate linear regression, but there was still some evidence of spatial clustering of residual errors and autocorrelation. The findings were successful in answering RQ2, showing that proportion of population with degrees, percentage of houses in tax band FGH, and percentage of people cycling to work were generally negatively correlated with diabetes prevalence. Proportion of BAME residents and proportion of population with no qualifications were both positively correlated to diabetes prevalence. Multivariate modelling produced better results than univariate, but autocorrelation indicated spatial variation.

5.2. Spatial Variation of Socio-demographic Factors

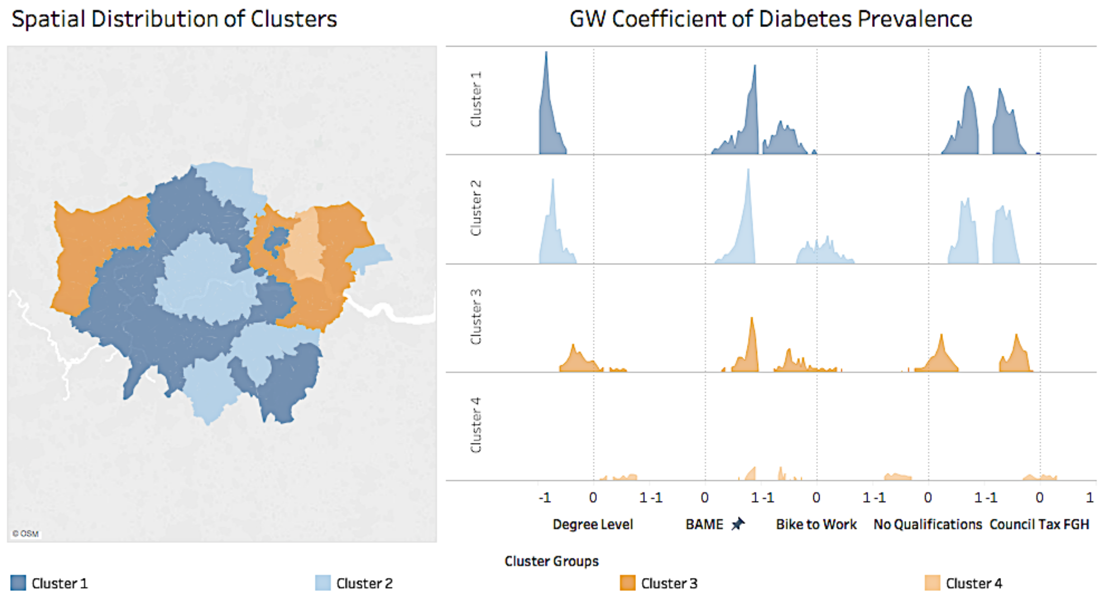


Figure 8- Choropleth of wards clustered by GW coefficeints. Cluster group denisity responses to variables illustrated adjacent.

The maps confirm that confirm that high BAME correlations are strongly positively correlated with diabetes prevalence, this pattern is consistent across the whole of London, but varies in strength. Populations with no qualification were also mostly positively correlated across London. In contrast, degree education and diabetes prevalence was negatively correlated over most of London, as too was proportion of houses in higher tax bands. Interestingly, for all of ‘No qualifications’, ‘FGH tax band’ and ‘Degree educated’, the area in north East London around Newham exhibits the opposite trend. The ‘Bike to Work’ variable is interesting. The negatively correlated suburbs contrasts with the positive coefficient in inner London. The results show that in relation to RQ3, the variables affecting diabetes prevalence is generally consistent across London, with some exceptions, notably the area in north east London.

6. Critical Reflection

The most important finding was that diabetes is clearly linked with socio-demographic indicators in London. This won’t come as much as a surprise for healthcare professionals; numerous studies such as that done by Gaskin et al., (2014) and others have made the connection between diabetes and poverty and ethnicity. However, the extent to which this case specifically in London has not been studied. The project has highlighted correlations between several factors, mainly linked to ethnicity, education level and affluence, so specific education schemes could be set up to target those that fall into the most ‘at risk’ groups as part of healthcare and social policy plans. These groups include the less affluent, but more clearly BAME communities who, from analysis, are more vulnerable to the disease. Whilst the models created in this project were done so in an explanatory context, the model could potentially be applied to other areas of the UK to find at-risk populations.

Overall, the complimentary nature of the visual and analytic approaches worked well at progressively developing the project, and answering the research questions. The project started with a choropleth map of ward level prevalence in London. This gave a good overview of diabetes in the city, highlighting some general trends and potentially high areas, but it was hard to draw conclusions from. Applying computational geostatistics not only allowed for clustering to be definitively confirmed, but the Getis-Ord Gi* statistic meant that the statistically significant hot and cold spots could be assessed, and hypotheses could be formulated as a result of some basic comparisons between these areas. For example, Kensington and Newham have different socio-demographic profiles, which suggested these could be a factor. It is important to appreciate that parameterisation and neighbourhood size can affect the outcomes of these statistics, so it was a good decision to run the algorithms with different parameterisation.

Using a heatmap to illustrate variable correlations was not very effective for finding important variables in this case. This was due to the high dimensionality of the data, as it became heavy on cognitive load and near impossible to make any sense of what variables would be useful for analysis. Random forest feature selection worked well, although it must be acknowledged that this is not always an effective means of selecting important features as results can vary on each iteration, although this was not much of an issue for this project. The visualisation of the output features from this analytic task was very effective. The interactive nature of the bar graph and dynamic split view made for an effective way of interrogating variable to diabetes correlation and an enjoyable user experience. The combination of computation and graphs worked well in a visual analytic context and could certainly be extended to other domains and projects.

The visualisation of independent and dependant variable correlations allowed for interactivity and detailed exploratory analysis, finding anomalies, patterns and identifying regional variation. Subsequent modelling and visualisation of residuals was effective. It demonstrated the limitations of using univariate linear models in situations where the situation is nuanced in a socio-demographic context. Formulating a multivariate model worked much better at reducing overall residual error between the figures, showed by visualisation, and highlighted that there was evidence of autocorrelation and confirmed the need for geographically weighted models. Overall the number of variables was ideal, as it prevented the model from being uninterpretable, but perhaps could have benefitted from combining the varies into broader categories, for example 'education' rather than 'no qualifications' and 'degree educated', although this was not a significant drawback.

The analysis highlighted that there was a need for geographically weighted regression to be carried out. The local models provided a more detailed understanding of processes on a local level whilst enabling more general observations to be carried out. Clustering the areas was a good choice to generalise areas, however in my opinion did not work effectively in the case of clusters 1 and 2 as it is hard to generalise complex socio-demographic differences over a large city well. However, Cluster 4 was identified as being anomalous to the rest of the city, and did highlight the areas of high diabetes prevalence in the east and west of the city, and the social makeup of these clusters in the adjacent GW coefficient density plots. The visual analytic methods were overall useful for answering Research Question 3.

The workflow and visual analytic approaches used in this project would be useful for other domains. The workflow of finding overall patterns, finding factors that influence patterns and then modelling the variables is a good strategy for any task related to data analytics in various domains and industries. In particular, the interactive visualisations would be beneficial for exploring data in any topic, as it allows the user to 'get to know' the data well and uncover patterns that otherwise would have gone unfound. In this example, however, the level of detail in the findings was greatly dependant on the data used. The UK census is generally very broad and gives an encompassing dataset, which many countries where there is a need for improved healthcare research don't have. It is worth noting at this stage that the census data dates back as far as 2011 which, although still entirely valid to use, is not as ideal as using recent data. In this instance the data was deemed relevant enough, but socio-demographics in cities can change quickly and this would need to be considered at the start of any project. Future work could be to see how the models made in this project would scale to other areas of the UK; cities, rural areas and nationally.

References

Kitabchi, A., Umpierrez, G., Miles, J. and Fisher, J. (2009) 'Hyperglycemic Crises in Adult Patients With Diabetes', *Diabetes Care,* 32(7): pp.1335–1343.

Diabetes.co.uk (2018) *Diabetes Prevalence.*
Available at: https://www.diabetes.co.uk/diabetes-prevalence.html. (Accessed: 19/12/2018)

Nazroo, JY. (1997) 'The health of Britain's ethnic minorities: findings from a national survey' London. Policy Studies Institute.

Gaskin, D., Thorpe, R., McGinty, E., Bower, K., Rohde, C., Young, J., LaVeist, T. and Dubay, L. (2014) 'Disparities in Diabetes: The Nexus of Race, Poverty, and Place', *Am J Public Health,* 104(11): pp.2147–2155.

Fu, W., Jiang, P., Zhou, G. and Zhao, K. (2014) 'Using Moran's I and GIS to study the spatial pattern of forest litter', *Biogeosciences, 11 pp.2401–2409*

Tobler, W. (1970) A Computer Movie Simulating Urban Growth in the Detroit Region, *Economic Geography*, Vol. 46, pp. 234-240.

Diabetes Data: https://commonslibrary.parliament.uk/social-policy/health/diabetes-in-england-where-are-the-hotspots/

Census Data: https://data.london.gov.uk/dataset/ward-profiles-and-atlas