Simple Geo-Spatial Analysis using R-Shiny

ISSS608 Group 11

Abstract—Geo-spatial analytics is a growing discipline providing important analytics in a wide range of applications. Tesco Grocery 1.0 provided an opportunity to enhance traditional global linear regression and analytics with geographically weighted clustering and regression methods due to the existence of spatial autocorrelation. A R-Shiny application was developed following a data analytics workflow process of exploratory single and bivariate data analysis, exploratory spatial data analysis, geographically weighted multivariate cluster analysis and regression. The extensive use of geo-visualisations and availability of user-selectable options allow users to uncover hidden patterns in an effective and pleasing way.

Authors – LI Junyi Darren, Muhammad Jufri Bin RAMLI, TEO Lip Peng Raymond

1 Introduction

Summarised grocery data from in-store purchases of 411 Tesco shops in the Greater London are made available at 4 different administrative area granularities, together with their respective population and health statistics. We combined the aspatial attribute data with geo-spatial polygons of the administrative areas and developed a Shiny application to provide interactive analysis of the data using traditional and Geographically Weighted (GW) methods.

The analysis is performed, notably through four sections:

- Exploratory Data Analysis (EDA)
- Exploratory Spatial Data Analysis (ESDA)
- Clustering (Hierarchical, GeoSpatial, Skater Clustering)
- Geographically weighted regression (GWR)

2 Motivation of the application

Despite the importance of studying food consumption at scale, there is little data about what people eat over long periods of time. The recent availability of this dataset provides us with an opportunity to investigate the nutrients and energy consumption of Greater London residents through Tesco store sales as an approximation of the population.

Geospatial analysis has gained much traction in recent years. Visualisation of the data using additional geo-mapping methods can bring out underlying relationships, variations, cluster groupings of spatial nature. In addition, there is an opportunity to apply GW regression techniques to explain variations of child obesity and adult diabetes and to compare the results against global linear regression.

Lastly, commercial analytics software has progressed well in terms of functionalities and usability at a premium to users. Using free open-source tools as R requires user to code which may not be possible for most users. We aspire to bridge the gap by providing a free application that offers the functionalities and usability of commercial tools in a free open-source tool without requiring users to code or understand R.

3 Review and critic on past works

SGSAS drew much inspirations from two previous works by students of Singapore Management University (SMU).

In How Healthy is Your Neighbour [2019], synchronised graphs provided excellent side by side comparisons of indicators. The subsetting of the data into smaller administrative districts allow closer analysis of the data. However, the calculations of the local indicators of spatial autocorrelation were done at each administrative district level which may introduce biasness due to the small sample sizes. As mentioned by the authors, other clustering models were not explored in the interests of time. The analysis may also benefit from adding geographically weighted models and tests to select and confirm spatial relationships. Minor cosmetic enhancement can be done (eg, colouring of the dendrogram lines according to the cluster group) to enhance the visual cues. Overall, the clear and neat design provided good interactivity and visualisations.

Corn: The A-maize-ing Crop [2018] has a clear work process for geographically weighted modelling. Additional parameter inputs and test indicators can be added to enhance model calibration and result display. As mentioned by the authors, spatial multicollinearity tests using GW Principle Component Analysis for variable reduction and principle component composition were not done due to time constraints.

Aiello, L.M., Quercia, D., Schifanella, R. et al. (2020) [1] performed linear regression on the data to derive two main explanatory variables of energy-carbs and $H_{\text{nutrients-calories}}$ for diabetes in Greater London. However, to achieve this, they had to filter data with representativeness > 0.1 to account for the areas having fewer Tesco shops relative to the population of that area. They further performed hierarchical regression by grouping wards into their respective administrative areas, and obtained a Bayesian R^2 of 0.78, suggesting that accounting for the hierarchical structures in the data increases the amount of explained variability.

There were few visualisations of the data and analysis, using mainly tables, due to the number-focused nature of the research. The results suggested possibilities of spatial biasness which may favour a GW model to cater to local variations of the model for a better explanation of the target variable.

This research provided opportunities to enhance on the above. The team aspires to incorporate the best of LISA and Cluster Analysis with GW modelling to present a more comprehensive workflow, and to use the developed application on the Tesco dataset to enhance the explanatory powers. Further developments on synchronisation of the views (e.g. bivariate to map linking), view filtering, parameter controls and result visualisations shall be explored to make the tool useful.

4 Design framework

The application uses the free and open-source R language that offers a thriving programming environment for statistical and graphical analysis. Our 4 design considerations for development:

- Performing calculations/logic within R for reproducibility.
- Using standard R packages on the Comprehensive R Archive Network (CRAN) for supportability.
- Using Shiny to webify the codes for simplicity; and
- Providing options and visualisations for interactivity.

The typical data analytics workflow process of data preparation, interactive exploratory data analysis, analysis and modelling, and insights discovery are made available in the application. The application provides common statistical methods and additional geo-spatial analytical methods to uncover geo-spatial autocorrelation and biasness.

4.1 Data Preparation

All data preparations are performed using R. Tesco provided 4 datasets at the LAD, Ward, MSOA and LSOA levels. Each dataset contains summarised purchases in weight and volume, their nutrient values, and population in the same format. The data is joined with child obesity data from Public Health England and area boundaries shape files from Office for National Statistics. The resulting objects are saved in R binary data files to facilitate ease of loading in shinyapps.io.

4.2 Exploratory Data Analysis

Our application allows interactive bivariate and correlation analysis to be performed under the exploratory data analysis (EDA) tab.

A scatterplot with linked map allows users to select points on the scatterplot to view their location of the map denoted by their polygon outlines.

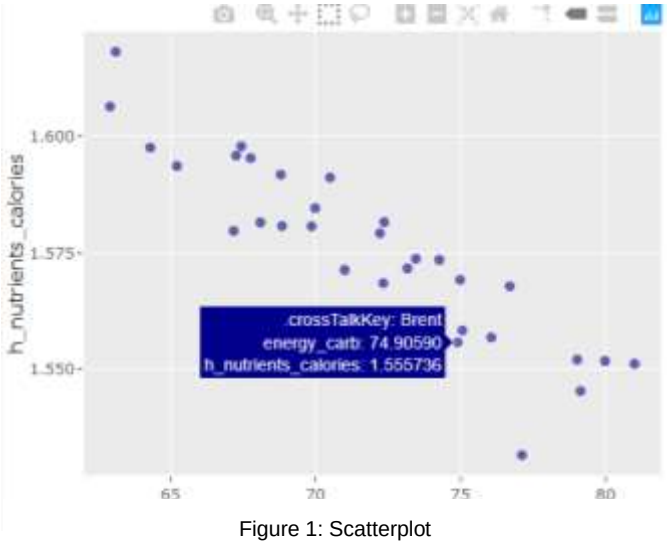

Figure 1: Scatterplot

A correlation matrix allows users to select measures in real-time to view their correlation.

4.3 Exploratory Spatial Data Analysis

Spatial clustering of geographically referenced attribute is performed in this section. We detect clusters and/or outliers using Local Indicator of Spatial Association (LISA). LISA are statistics that evaluate the existence of clusters in the spatial arrangement of a given variable.

This form of exploratory spatial data analysis (ESDA) investigates the location component of our dataset explicitly instead of looking at relationship between variables and how they affect each other in EDA. ESDA considers the values of the same variable in the neighbourhood.

Two maps are linked to show relationships between the selected variables and for more effective analysis.

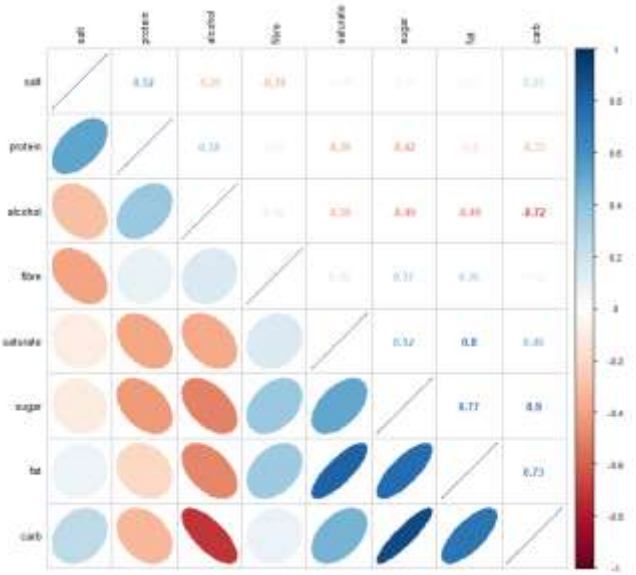

Figure 2: Correlation Matrix

Figure 3: Map Polygon Outline

4.3.1 Distance & neighbours calculations

A spatial weigh matrix is derived by building a neighbours list based on regions with contiguous boundaries (i.e. sharing one or more boundary points).

Figure 4: Contiguity map of Greater London

Our application allows the selection of the following contiguity matrix to define neighbours.

- Contiguity Queen
- Contiguity Rook
- K Nearest Neighbours
- IDW Queen
- IDW Rook

Contiguity Queen defines any place within the selected boundary neighbours, while contiguity rook only includes shared borders. The queen's neighbours will also include neighbours that do not share any boundary.

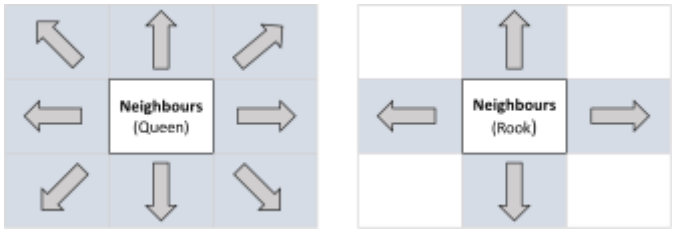

Figure 5: Queen and Rook boundaries

A sample of a region's neighbours defined by contiguity queen is shown.

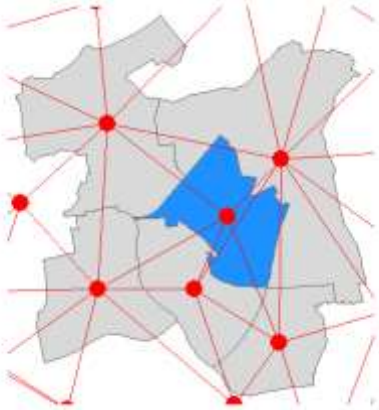

Figure 6: Queen Boundaries

When using K Nearest Neighbours, the K number can be set by the user using a slider. In inverse distance weighted (IDW) interpolation, the difference with the nearest neighbours approach is that points that are further away get less weight in predicting a value in a location.

4.3.2 LISA computation and interpretation

A SpatialPolygonDataFrame is created prior to mapping the LISA clusters. In the below LISA cluster, there are 5 different clusters. Regions. The LISA cluster map shows the significant locations color coded by type of spatial autocorrelation.

The variable of interests is first centered around its mean.

$$DV = V_{interest} - mean\ (V_{interest})$$

This is followed by centering the local Moran's around the mean (C_mI). A statistical significance level for the local Moran, selectable by the user, is also used.

Each LISA Cluster is then paired according to the following combination.

LISA Cluster	DV	C_mI
Insignificant	Non-significant	
Low-low	<0	<0
Low-High	<0	>0
High-Low	>0	<0
High-High	>0	>0

The local Moran's I values are mapped and linked to the p-values map for more effective interpretation.

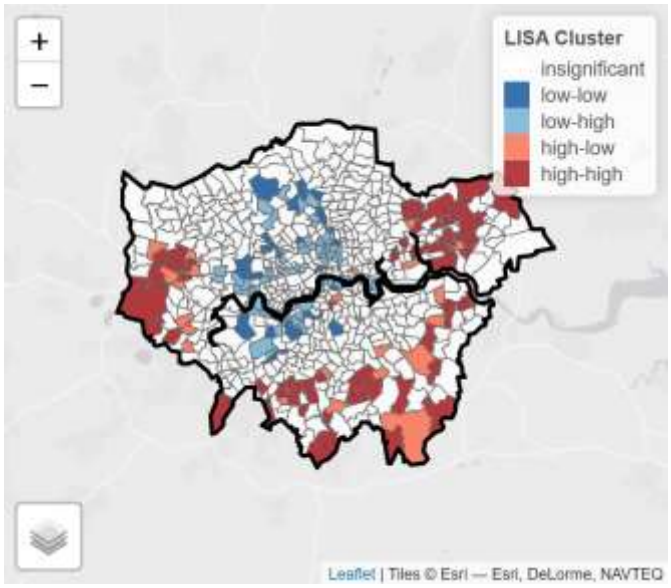

Figure 7: LISA Clusters

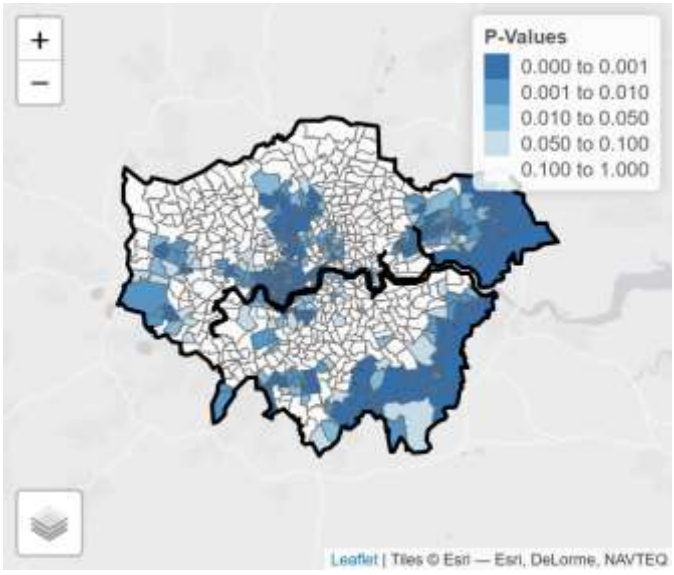

Figure 8: P-Values

4.4 Cluster Analysis

In clustering, it is a common practice to delineate the market or planning area into homogeneous regions by using multivariate data. In SGSAS, we are interested to delineate Greater London into homogeneous areas by using multiple nutrients measures, namely: fat, saturate, salt, sugar, protein, carb, fibre and alcohol.

There are three methods of clustering used in the tool. The first is agglomerative hierarchical clustering based on the measures, without any geospatial constraints. The second uses Ward-like hierarchical clustering algorithm including spatial/geographical constraints. The third explicitly considers the contiguity constraints in the clustering process.

4.4.1 Selecting cluster size

The fundamental of clustering is choosing the cluster size. In SGSAS, it is able to suggest to user what is the optimal cluster size using the average Silhouette method It is a way to measure how close each point in a cluster is to the points in its neighboring clusters. Silhouette values lies in the range of [-1, 1]. A value of +1 indicates that the sample is far away from its neighboring cluster and very close to the cluster its assigned. Similarly, value of -1 indicates that the point is close to its neighboring cluster than to the cluster its assigned. A value of 0 means it's at the boundary of the distance between the two cluster. Value of +1 is idea and -1 is least preferred. Hence, higher the value better is the cluster configuration.

4.4.2 Selecting clustering method

As mentioned in section 4.4, we will be using 3 methods.

In our hierarchical clustering, we will be using agglomerative clustering. It is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. The algorithm starts by treating each object as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. The result is a tree-based representation of the objects, named *dendrogram*. The dendrogram is also color-coded with the number of clusters selected.

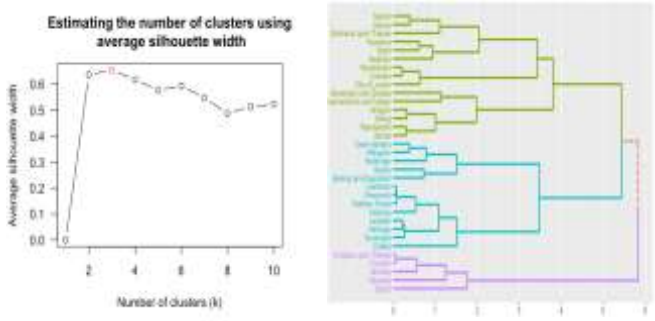

Figure 9: Clusters estimates with Dendrogram

In our geometrically constraint clustering method, it implements a Ward-like hierarchical clustering algorithm including soft contiguity constraints. This algorithm takes as input two dissimilarity matrices D0 and D1 and a mixing parameter alpha between 0 and 1. The dissimilarities can be non-Euclidean, and the weights of the observations can be non-uniform. The first matrix gives the dissimilarities in the "feature space" (socio-demographic variables or grey levels for instance). The second matrix gives the dissimilarities in the "constraint" space. The mixing parameter alpha sets the importance of the constraint in the clustering procedure.

The parameter alpha (the weight of this convex combination) controls the weight of the constraint in the quality of the solutions. When alpha increases, the homogeneity calculated with D0 decreases whereas the homogeneity calculated with D1 increases.

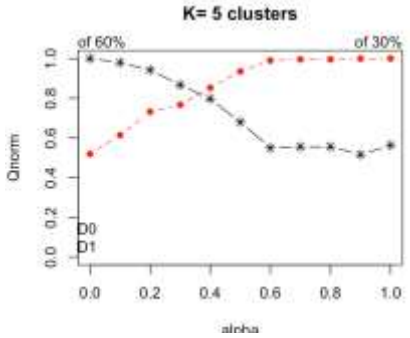

Figure10: Homogeneity criterion

In skater method, it is an approach that explicitly takes into account the contiguity constraints in the clustering process and is based on the algorithm outlined in Assuncao et al. (2006). The algorithm carries out a *pruning* of the *minimum spanning tree* created from the spatial weights matrix for the observations. The *weights* in the spatial weights matrix correspond to the pair-wise dissimilarity between observations, which become the edges in the graph representation of the weights (the observations are the nodes).

4.5 Geographically Weighted Regression

A Geographically Weighted (GW) model fits situations when spatial data is poorly described by the global model, and for some locations, a localised calibration provides a better description. It uses a moving window weighting technique, where localised models are found at target locations. Parameters and outputs of a GW model are mapped to provide a useful exploratory tool for statistical analysis.

SGSAS utilises GWmodel R package to provide basic GW regression fits, diagnostics and visualisations with enhanced kernel weighting, distance metric, bandwidth selection capabilities.

A fundamental element in GW modelling is the spatial weighting function (Fotheringhamet al. 2002) that quantifies the spatial relationship or spatial dependency between the observed variables. Here $W(u_i, v_i)$ is a n x n (with n the number of observations) diagonal matrix denoting the geographical weighting of each observation point for model calibration point i at location (u_i, v_i). We have a different diagonal matrix for each model calibration point. There are three key elements in building this weighting matrix: (i) the type of distance, (ii) the kernel function and (iii) its bandwidth.

4.5.1 Selecting the distance

Distances are calculated using the power of Minkowski distance - p. It includes Euclidean distance with p = 2 and Manhattan distance with p = 1.

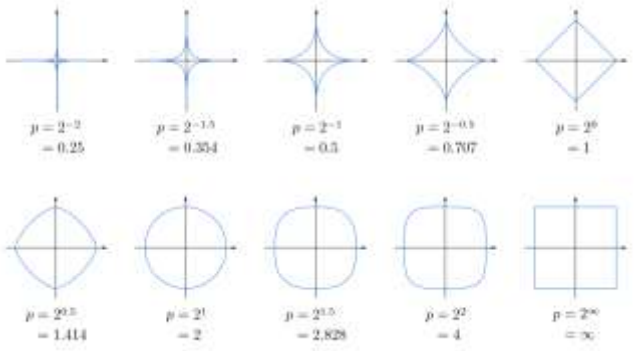

Figure 11: SGSAS provides choice of Euclidean or Manhattan distances

4.5.2 Selecting the kernel

The Gaussian and exponential kernels are continuous functions of the distance between two observation points (or an observation and calibration point). The weights are maximum (equal to 1) for an observation at a GW model calibration point and decreases according to a Gaussian or exponential curve as the distances between observation and calibration points increase.

The box-car kernel is a simple discontinuous function that excludes observations that are further than some distance b from the GW model calibration point. This is equivalent to setting their weights to zero at such distances. The bi-square and tri-cube kernels are similarly discontinuous, giving null weights to observations with a distance greater than b. However, unlike a box-car kernel, they provide weights that decrease as the distance between observation and calibration points increase, up until the distance b. Thus, these are both distance-decay weighting kernels with a cut-off distance.

SGSAS allows selection of all these kernels.

Global Model $$w_{ij} = 1$$

Gaussian $$w_{ij} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right)$$

Exponential $$w_{ij} = \exp\left(-\frac{|d_{ij}|}{b}\right)$$

Box-car $$w_{ij} = \begin{cases} 1 & \text{if } |d_{ij}| < b, \\ 0 & \text{otherwise} \end{cases}$$

Bi-square $$w_{ij} = \begin{cases} (1-(d_{ij}/b)^2)^2 & \text{if } |d_{ij}| < b, \\ 0 & \text{otherwise} \end{cases}$$

Tri-cube $$w_{ij} = \begin{cases} (1-(|d_{ij}|/b)^3)^3 & \text{if } |d_{ij}| < b, \\ 0 & \text{otherwise} \end{cases}$$

4.5.3 Bandwidth Selection

The key controlling parameter in all kernel functions is the bandwidth b. Bandwidths can be specified either as a fixed distance in metres or as an adaptive distance in number of neighbours. A fixed bandwidth suits regular sized areas whilst an adaptive bandwidth suits highly irregular sized area. Adaptive bandwidths ensure sufficient (and constant) local neighbouring areas for each local calibration of the GW model.

Bandwidths for GW models can be found via automated procedures by minimising the cross-validation (CV) or Akaike information criterion (AIC). Further calibration is possible through user-specified values to suit local context of the data.

5 Application Insights

5.1 LISA Spatial Autocorrelations

At the ward level, using a confidence level of 90% for energy-carbs and $H_{\text{nutrients-calories}}$, the ESDA pane. There is spatial correlation denoted by the colours on both mapped variables.

The areas marked by LISA also shows very similar areas or regions where spatial correlation is present. However, they are opposite of each other. While showing the spatial autocorrelations, the p-values are also shown on the second linked map on our application when the LISA clusters are formed.

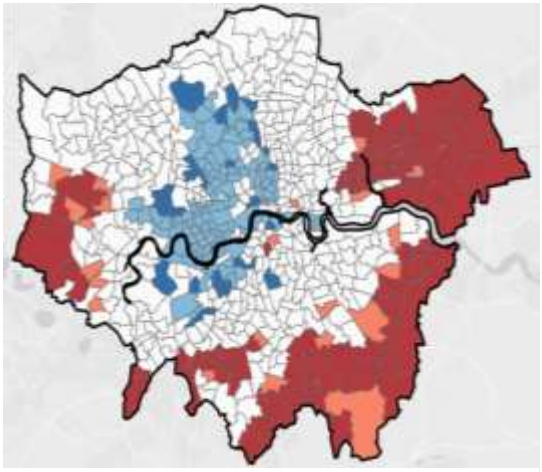

Figure 12: Energy_Carb

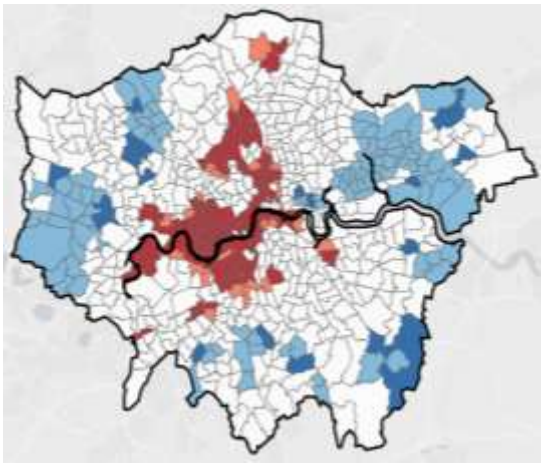

Figure 13: H_Nutrients_Calories

5.2 Geographical Constraint Clustering

Clustering with just multivariate measures for geospatial data will make the cluster looks scattered on a make. However, with geo spatial constraint in place, it will delineate the LAD into homogeneous regions.

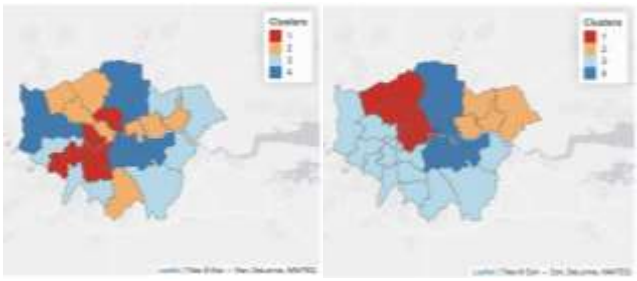

Figure 14: Scattered and Homogeneous regions

Additionally, we can get more insights into the cluster by looking at the parallel coordinate which tells us if they skewed towards a certain measure for a cluster.

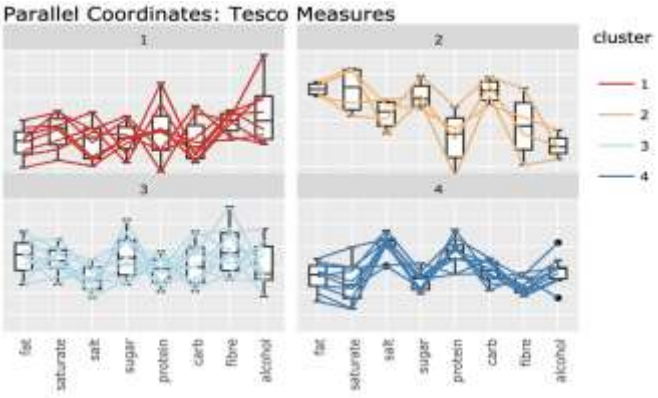

Figure 15: Parallel Coordinate Plots

5.3 Geographically Weighted Regression

The linear regression results by Aiello, L.M., Quercia, D., Schifanella, R. et al. (2020) were verified. The two highest correlated factors, energy-carbs and $H_{nutrients\text{-}calories}$, have a high explanatory power (R2 = 0.56) for the dependent variable, obesity prevalence. Similarly, adding demographics and store penetration control variables of average age, % of female residents, density of residents in the area, and number of transactions only raised the R2 marginally to 0.61.

With GW model, the R2 with the control variables is 0.825 using adaptive bandwidth with 19 neighbours. Switching to fixed bandwidth of 1986 metres raised the R2 to 0.876. However, using just energy-carbs and $H_{nutrients\text{-}calories}$ already produced high R2 of 0.871, which is way above that of using global linear regression, showing that localised model explained the variation better.

It is worth pointing out that diabetes prevalence (2016) is only available at the Ward level. Also, Ward boundaries have undergone many changes over time, resulting in many non-matches when the diabetes prevalence data is joined with spatial data. Such non-matching data is excluded.

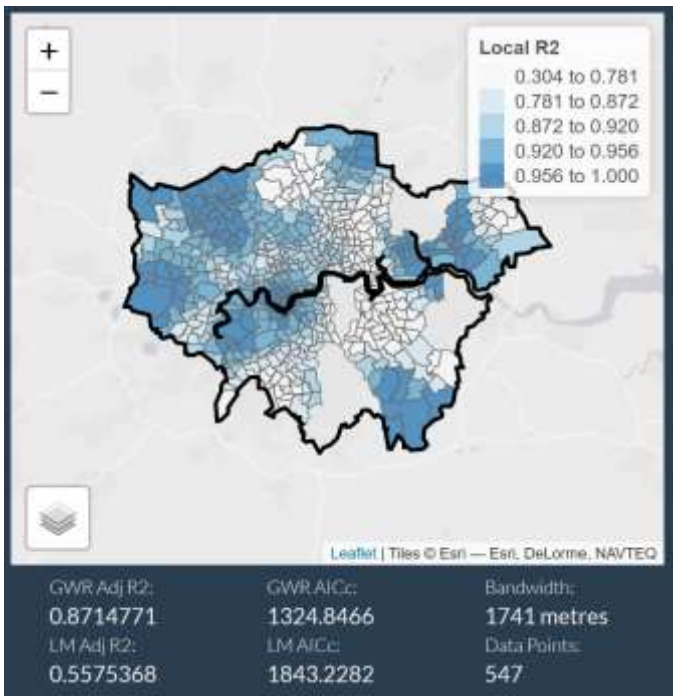

Figure 16: GWR Local R2 of energy-carb and $H_{nutrients\text{-}calories}$ for diabetes

6 Future Work

Principal Component Analysis (PCA) is commonly used to explain the covariance of a multivariate dataset using only a few components. GW PCA accounts for spatial heterogeneity of the multivariate data, assessing how data dimensionality varies spatially and how the original variables influence each spatially varying component. It will be useful to look at combination of variables to find out those highly correlated in terms of spatial locations and how are they distributed (closeness/proximity) using GW PCA.

SGSAS was built with the Tesco Grocery 1.0 dataset as a use case. Throughout the application, many selections and arguments have been exposed as user configurable options in order to remove coding required of users. The methods presented are equally applicable to other geo-spatial datasets. Hence, the application will benefit by incorporating a data load and processing function to accommodate different datasets and maps.

Acknowledgments

The authors wish to thank Professor Kam Tin Seong of Singapore Management University for providing extensive guidance and support on this project.

References

[1] Aiello, L.M., Quercia, D., Schifanella, R. et al. Tesco Grocery 1.0, a large-scale dataset of grocery purchases in London. Sci Data 7, 57 (2020)

[2] https://doi.org/10.1038/s41597-020-0397-7

[3] Aiello, Luca Maria; Schifanella, Rossano; Quercia, Daniele; Del Prete, Lucia (2020): Tesco Grocery 1.0. figshare. Collection. https://doi.org/10.6084/m9.figshare.c.4769354.v2

[4] Metadata record for: Tesco Grocery 1.0, a large-scale dataset of grocery purchases in London. (2020). Retrieved 26 April 2020, from https://springernature.figshare.com/articles/Metadata_record_for_Tesco_Grocery_1_0_a_large-scale_dataset_of_grocery_purchases_in_London/11799765

[5] Aiello, L., Schifanella, R., Quercia, D., & Del Prete, L. (2019). Large-scale and high-resolution analysis of food purchases and health outcomes. EPJ Data Science, 8(1). doi: 10.1140/epjds/s13688-019-0191-y

[6] Guide to Presenting Statistics for Super Output Areas (June 2018). (2020). Retrieved 26 April 2020, from https://geoportal.statistics.gov.uk/datasets/guide-to-presenting-statistics-for-super-output-areas-june-2018

[7] Obesity Data and Tools :: Public Health England Obesity Knowledge and Intelligence team. (2020). Retrieved 26 April 2020, from https://webarchive.nationalarchives.gov.uk/20170110165409/https://www.noo.org.uk/visualisation

[8] Greater London. (2019). Retrieved 26 April 2020, from https://en.wikipedia.org/wiki/Greater_London

[9] Regions (December 2019) Boundaries EN BGC. (2020). Retrieved 26 April 2020, from https://geoportal.statistics.gov.uk/datasets/regions-december-2019-boundaries-en-bgc

[10] Local Authority Districts (December 2019) Boundaries UK BGC. (2020). Retrieved 26 April 2020, from https://geoportal.statistics.gov.uk/datasets/local-authority-districts-december-2019-boundaries-uk-bgc

[11] Wards (December 2019) Boundaries EW BGC. (2020). Retrieved 26 April 2020, from https://geoportal.statistics.gov.uk/datasets/wards-december-2019-boundaries-ew-bgc

[12] Middle Layer Super Output Areas (December 2011) Boundaries EW BGC. (2020). Retrieved 26 April 2020, from https://geoportal.statistics.gov.uk/datasets/middle-layer-super-output-areas-december-2011-boundaries-ew-bgc

[13] Lower Layer Super Output Areas (December 2011) Boundaries EW BGC. (2020). Retrieved 26 April 2020, from https://geoportal.statistics.gov.uk/datasets/lower-layer-super-output-areas-december-2011-boundaries-ew-bgc

[14] How Geographically Weighted Regression (GWR) works—ArcGIS Pro | Documentation. (2020). Retrieved 26 April 2020, from https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/how-geographicallyweightedregression-works.htm

[15] SAGE Reference - Exploratory Spatial Data Analysis. (2020). Retrieved 26 April 2020, from https://sk.sagepub.

[16] /reference/geography/n406.xml

[17] SAGE Reference - Exploratory Spatial Data Analysis (ESDA). (2020). Retrieved 26 April 2020, from https://sk.sagepub.com/reference/geoinfoscience/n64.xml

[18] Gollini, I., Lu, B., Charlton, M., Brunsdon, C., & Harris, P. (2013). GWmodel: an R Package for Exploring Spatial Heterogeneity using Geographically Weighted Models. Retrieved 26 April 2020, from https://arxiv.org/abs/1306.0413

[19] Lu, B., Harris, P., Charlton, M., & Brunsdon, C. (2013). The GWmodel R package: Further Topics for Exploring Spatial Heterogeneity using Geographically Weighted Models. Retrieved 26 April 2020, from https://arxiv.org/abs/1312.2753

[20] Murakami, D., Tsutsumida, N., Yoshida, T., Nakaya, T., & Lu, B. (2019). Scalable GWR: A linear-time algorithm for large-scale geographically weighted regression with polynomial kernels. Retrieved 26 April 2020, from https://arxiv.org/abs/1905.00266

[21] The Minkowski approach for choosing the distance metric in geographically weighted regression (2020). Retrieved 26 April 2020, from http://mural.maynoothuniversity.ie/7850/1/MC_Minkowski.pdf