CITY UNIVERSITY LONDON

MSc in Data Science

Project Report

academic year 2014/2015

Effects of bandwidth settings in Geographically Weighted Regression over different density patterns

Andrea Sportelli
Supervised by Cagatay Turkay

ABSTRACT:

This dissertation investigates the effects of Geographically Weighted Regression (GWR) bandwidth settings over different density patterns. This is motivated by the current void in research around the relation between GWR and geographical patterns and by the shortage of detailed studies on the effects of different GWR parameters settings in general. The goal of the work is to provide evidence to inform academic practise and to compare GWR by adding the geographical dimension as a variable. The previous research has focused almost exclusively on refining GWR and adding new features and new parameters without fully developing an understanding of its behaviour in different circumstances. The most important contribution of this study is a detailed analysis of the different GWR bandwidths and kernel settings over data with different geographical distribution. This was achieved through the comparison of three case studies. The findings from the research show that density pattern plays a major role on GWR's outcome. Specifically, the adaptive bandwidths should be employed when the density pattern of the data is not uniform, while fixed bandwidth should be preferred when data is composed by a multitude of clusters.

Keywords:

Geographically Weighted Regression, Kernel function, Bandwidth, Spatial analysis.

TABLE OF CONTENT

1. INTRODUCTION

Local spatial regression techniques were originally elaborated in the work by Cleveland and Delvin (1998). The "Geographically Weighted Regression" (GWR) (Brunsdon et al.1996), was a spatial regression technique that stood out as an innovative approach. Developed to test the null hypothesis that regression constants are static without exception, this technique became rapidly a popular choice for research over spatial non stationary relations (Fotheringham and Brunsdon, 1999. Lloyd and Shuttleworth, 2005. Yu, 2006. TU and XIA, 2008. Cho, Lambert and Chen, 2010). Compared to a conventional Ordinary Least Squares Regression, GWR has the ability of generating different estimate parameters as well as other statistics such as t-values for a multitude of points (often conceding with the data point) across that data surface. These are calculated taking into account a set of neighbours giving them decreasing influence over the local coefficient according to their distance from the regression point. Thanks to its ability of modelling varying relations across space and through a graphical representation in the form of a map, GWR together with the other geographically weighted techniques inspired by it, proved itself extraordinarily powerful at summarising complex spatially varying relations (Cho et al., 2009). Notwithstanding, despite the popularity gained by GWR, scholars lament the absence of clear-cut properties in the many settings available in GWR. Specifically, the behaviour of different GWR bandwidth settings is a topic that has not gained as much academic attention as hoped (Nakaya et al. 2005). While many studies used GWR as their main technique on a variety of field giving further credit to its capabilities, fewer studies focused on the technique itself and to the comparison of the different available settings. Indeed, results in the form of reported graphical representations are missing due to the prohibitive amount of maps needed in order to compare all possible variations that this model can take over a single dataset. This lack of guidelines was recognised as a problem when researches did not find answers to those questions that arose while the GWR was set up (Yu, 2006). Despite the many studies (next section) that have been targeted on expanding the functionalities of GWR, a comprehensive analysis on the effects of different parameter settings is still absent. In particular, it has been completely overlooked how the GWR behaves over different density patterns.

The purpose of this analysis is to find evidence to inform academic practice. This will be achieved by determining how the choice of certain bandwidth settings influence the outcome of a GWR. The main focus is therefore on the interplay and trade-offs between parameters and data characteristics with a greater attention to the different density patterns to which GWR is applied. Looking at the scope of the study, the outcome of the GWR is evaluated in terms of estimate parameters and t-values. While estimate parameter is the coefficient that has received most of the attention (next

section), t-values have not been studied with the same degree of detail. Their inclusion in this work attempts to reduce the research void on this subtopic. Moving to the bandwidth settings specifically, these will be the bandwidth kernels Gaussian, Boxcar and Exponential, bandwidth types: adaptive / fixed and the bandwidth size through auxiliary graphs (tool and evaluation section). These settings will be compared against each other over three datasets with different density patterns. One uniform, one with varying levels of densities with scattered outer regions together with denser areas, and finally one structured as a multitude of small clusters. The research question addressed, is: "*What are the effects of bandwidth settings in geographical weighted regression over different density patterns?"* The general objective is to develop guidelines for the application of GWR in accordance with dataset characteristics. The analysis is to constitute an asset for all the researchers that have faced problems while using GWR, such as a lack of an informative and comprehensive source when it comes to choosing the bandwidth kernel settings. Especially for scholars not directly involved in geo-statistical analysis, like social sciences, will find this work beneficial. Furthermore, the resulting maps from GWR often come as didactic visual aids for policymakers at all levels of administration (Cho et al., 2009). Decision-making, based on maps produced by informed researchers, will contribute to the betterment of society as a whole. This will be accomplished by informing analysts with the different effects and trade-offs of their settings. The results deriving from the experiments will be interpreted and explained according to parameter properties and assumptions. The substantial result will be a documented set of maps and graphs combined with reports that depict the local parameters variations using different bandwidth sizes, kernel functions and adaptive/fixed types.

The structure of this study is the following: It begins with a literature review on the research made on GWR. Where it is identified by the current state-of-the-art knowledge of the technique in several areas. Limitations and gaps identified in this section provide the motivations of this work. The second section describes in details the methodology adopted by GWR. The description is concentrated on the computation of coefficients and on the weighting scheme determined by the bandwidth functions. The third section gives an account of evaluation and tools used during the analysis. The following sections are dedicated to an in-depth analysis of the bandwidth parameter effects over the data proposed. This starts off with a descriptive analysis of the properties of these datasets highlighting their geographical pattern. Subsequently, each of three dataset chosen is used to test various GWR parameters. Lastly, the discussion section, is used to compare results and findings of each dataset returned in the analysis.

2. LITERATURE REVIEW

A naive method to account for geographical variability and test if relations are stationary over space, is to divide the data surface into sub regions and compute an independent regression on each subset. This would produce a set of regression coefficients for each sub region that may be plotted and then visually investigated. This approach carries several problems. The division in sub regions may result in statistical issues when one sub region contains too few data points to reach statistical significance or there might be a problem of class imbalance not present in the whole data set. As we assume that coefficients may not be stationary on the whole map, this may also be true within the sub region not making clear what should be the correct partition. Finally, most spatial variations are continuous. By dividing arbitrarily a map (i.e. using pre-defined administrative boundaries) the real underlying nature of the spatially varying relation cannot be observed. These shortcomings triggered the need for a better technique able to model non-stationary relations (Fotheringham, Brunsdon and Charlton, 2002).

The Moving Window Regression (MWR) was a first attempt to address the limitations formerly expressed. A region (theoretically of any shape, usually squared) of fixed size is defined around each regression point (usually coinciding with data points). Coefficients are computed taking into account only those data points within the region. The process is then repeated for each regression coordinate. By overlapping the regression regions, there is a component of continuity in the model that is not affected by arbitrary divisions. Despite that MWR is still a discontinuous technique, we can interpret the region definition of the MWR as a weighting scheme that assigns a weight of 1 for those data points inside the region, while data points outside, are given 0. This weighting scheme introduces by itself a source of discontinuity making the results depending on the window size (Fotheringham, Brunsdon and Charlton, 2002).

Geographically weighted regression (GWR) was thought as an improvement to MWR by introducing a continuous weighting scheme defined by a distance decay function. More weight is allocated to data points close to the regression location than those further away. The introduction of this system allows GWR to model data beyond the restriction of sub region division since all (even though data points far away can be still be given a weight of 0) are used to calibrate the regression at any specific regression (detailed description of GWR in the methodology section). The weighted function is therefore crucial to GWR and its study has covered a great share of works based on geographically varying relations (Lloyd and Shuttleworth, 2005). In particular, the bandwidth kernel was extensively examined. Research has been mainly focused on three areas: Firstly, the

effects of the size of the bandwidth over local coefficients. Secondly, determining the optimal size of the bandwidth has led scholars to investigate the outcome of different validation/selection techniques used in setting this parameter. Thirdly, the kernel function can assume two forms: fixed or adaptive. The choice of one type over another can provide significantly different results. What is more, besides the bandwidth kernel, there are also other aspects of GWR that have been investigated, resulting in new alternatives to the basic model. The use of the distance function that determines the proximity between a data point and the regression point has been investigated to provide an alternative to the simple Euclidean distance. Also data visualisation techniques and visual encodings have been studied to summarise the information provided by GWR in a concise manner.

Regarding the effects of increasing and decreasing the size of the bandwidth, there is a general agreement that effects of different values need to be visually interpreted in order to be assessed. As a result, GWR is generally based on the human computer interaction loop, since spatial varying relations need a human interpretation that cannot be fully expressed by a coefficient (Brunsdon et al. 1996, Lloyd and Shuttleworth, 2005, Mennis, 2006). Notwithstanding, researchers have yielded empirical recurrent results deemed informative in the awareness of trade-offs and determination of the direction of the change. GWR, build on small bandwidth, are believed to provide less stable local coefficients given the few data points used to make inference. Yet, small bandwidth are able to unveil interesting local anomalies that may be masked out by large bandwidth. While larger bandwidths provide smoother data surfaces, increasing further the size, leads to poor model fitting as well as undermining the logic behind GWR, making itself become similar to a global model (Lloyd and Shuttleworth, 2005).

Methods used to determine the optimal bandwidth size do not rely only on visualisation. Model validation/selection techniques have been widely applied to GWR in order to determine the bandwidth value that best minimises residual spatial dependence. Cross Validation scores was found the most used method to set the optimal bandwidth size. Nevertheless, this method is not free from drawbacks (Farber and Páez, 2007). In fact, outliers may negatively affect Cross Validation performance. The Cross Validation algorithm (detailed description in methodology) recursively computes its score for each bandwidth size returning the one that minimised the score. However, this bandwidth optimisation procedure might be biased by a few data points. Looking at Cross Validation Scores decomposed in terms of individual observations, Faber and Páez (2007) found that the goodness-of-fit may be maximised by smaller bandwidth than those resulting from the Cross Validation algorithm. Another common choice of model validation/selection used in GWR to

find the optimal bandwidth size is the Akaike Information Criterion (TU and XIA, 2008). As in Cross Validation, the bandwidth that minimises the score is the one that provides a model that better represents the spatially varying relation within the data. Moreover, the Akaike Information Criterion has the advantage over the Cross Validation to provide whther a GWR provides a better fit than a global model by taking into account also the different degrees of freedom of the two models. However, comparisons between the performance of the Cross Validation and Akaike Information Criterion did not show significant variations (Fotheringham, Brunsdon and Charlton, 2002). Finally, model validation/selection techniques were developed to account for the GWR issues caused by residual spatial dependence. Scholars (Wheeler and Tiefelsdorf, 2005) argued that spatially varying patterns suggested by GWR may not be representative of the underlying structure of the data but only artefacts resulting from the model calibration. A hybrid model that sees the use of Cross Validation scores combined with Lagrange Multiplier test statistics was developed by Cho, Lambert and Chen (2010). In a comparative study against Cross Validation trade-offs between the two models emerged. The hybrid model was found able to better deal with spatial autocorrelation yielding smaller bandwidths compared to Cross Validation scores. Although, the cost of reduced spatial autocorrelation is paid off by a higher rate of extreme coefficients.

Despite that kernel functions have been the parameters studied the most, also the distance measurement used in GWR calibrations have received considerable attention. As a default parameter, Euclidean distance was applied to the majority of research papers using GWR. Notwithstanding, due to the simplicity of Euclidean distance that calculates the proximity between two points as a straight line in a bi-dimensional space as in the case of GWR, research was made in order to test if this parameter was always the best option to choose. In fact, two points in a geographical space might be divided by rivers, mountains, or other natural obstacles, whereas railways and roads may significantly increase the reachability of geographically distant areas. In a study on local variation between price and floor area over London houses, Lu et al. (2011) applied a non Euclidean distance measure in their analysis. This was replaced by a network distance in which London streets were considered as edges of a graph and Houses as nodes. Despite the improvement in adjusted R square was negligible, estimate parameters were widely affected by the reachability expressed by the network distance. Following this study, Lu, Charlton and Harris (2012) compared Euclidean distance and Manhattan distance performances over simulated datasets. The results outlined that the model using Manhattan distance had a significantly better performance with respect to the Euclidean model. Precisely, estimate parameter accuracy was much greater.

Also visual encoding techniques have received notable attention by research on GWR. The main advantage of these techniques compared to a global model is to produce results that can be mapped

and visually interpreted by the analyst through a process of human computer interaction as described by Munzner and Maguire (2014) . By mapping the results of GWR the interpretation of the spatial context together with the know characteristics of the study area is facilitated (Goodchild and Janelle, 2004). Despite its crucial role, there are several challenges to deal with, when GWR maps are created. The first issue concerns the data classification technique used to divide data into classes. This procedure has the advantage of increasing the readability of the map over implementations in which continuous variables are represented as shades of colour. The most common approach is the equal step in which data range is chosen by simply dividing it in classes of equal extent. However, this method should be confined only to data having a uniform distribution. When data follows other distributions, let us say normal distribution, the equal step approach might generate misleading results inasmuch the central classes will have much more observations than those at the ends. Other methods have been adopted to improve visualisation in this regard. For instance, data was divided into corresponding statistics, such as quartiles or optimal methods for within class homogeneity maximisation were applied (Cromley, 1996). In addition, data classification of the t-value should account for its significance. The visual encoding should take into account threshold values that distinguish between t-values that have reached statistical significance and t-values that have not. If the visual encoding has one class containing both significant and not significant t-values, it is not possible to visually distinguish those coefficients that are significant and those that are not (Evans, 1977). Finally, maps should present the spatial distribution of the parameters together with the distribution of significance. This is fundamental to extrapolate meaningful results from the visual interpretation. Mennis (2006) developed a bivariate choropleth mapping encoding able to map estimate parameters coefficients and distinguish between highly significant, significant and not significant areas at the same time. This approach has the advantage of reducing the number of maps by half and facilitate the distinction between significant and insignificant coefficients.

To summarise, the studies carried on GWR have covered several different areas from model selection techiniques used for determining the optimal bandwithd size to visual encodings techniques employed to improve redability and interaction. However variations and trend-offs have been covered by scholars, there seems to be still room for exploration. No study was found dealing with the issue of varying density pattern. This omission together with the need of guidelines in setting GWR models constitutes the justification of this work aimed to fill this gap in the literature.

3. METHODOLOGY

3.1. ESTIMATE PARAMETER AND T-VALUES

Differently from a global regression model:

(1) $$y_i = \beta_0 + \sum_{i=1}^{k} \beta_k \, x_{ik} + \varepsilon_i$$

GWR can be seen as an extension to the conventional framework in which local estimate parameters are modelled under the equation:

(2) $$y_i = \beta_0(u_i v_i) + \sum_{i=1}^{k} \beta_k(u_i v_i) \, x_{ik} + \varepsilon_i$$

Where $(u_i v_i)$ represents the coordinates of the *ith* data point in space (i.e. latitude and longitude) while $\beta k(u_i v_i)$models the relation between y and x at this location in space. Therefore, parameter values are not fixed but are expressed as a continuous surface. Each $\beta k(u_i v_i)$ represents a measurement of this surface at a certain point denoting spatial variability across the surface. This equation opens the possibility to measure this spatial variability and to model its effect.

However, calibrating the GWR involves a trade-off between bias and standard error. In order to capture the spatial variability in one point in space, the estimate parameter is computed on a subset of datapoint within a certain distance (next section). If the variation in the data is spatially consistent, it is expected that points close in space will return similar coefficients. By moving across space and calculating new coefficients, however, these will be computed on new subset of data. The size of the subset determines the standard error of the estimate parameter. The greater the subset the lower is the standard error. However, large subsets are more likely to introduce bias since a large subset could introduce in the computation some data points belonging to another "region" whose values are significantly different from those expressed by data points close to the location under investigation. By increasing the subset size it is possible that its edges trespass a region of different coefficients.

This effect is reduced by following the main assumption of GWR which is that points that are closer to coordinates of the point under investigation are more likely to have similar coefficients than those points further away. Therefore, when the equation measures the relationship around each location *i* this has to take into account a weighting scheme (next section) that allows increasing influence in proximity to the location *i*.

This study is focused on two coefficients that are computed in GWR:
The first one is the estimate parameter. The conventional estimate parameter used in a global ordinary least squares regression

(3) $$\hat{\beta} = (X^T X)^{-1} X^T Y$$

is replaced with

(4) $$\hat{\beta}(u_i v_i) = \left(X^T W(u_i v_i) X\right)^{-1} X^T W(u_i v_i) Y$$

where $\hat{\beta}$ is the estimate of β while W $(u_i v_i)$ is a *n* by *n* matrix whose diagonal elements represent the geographical weighting of the data point inside the window at location *i*. This equation presents a weighted least squares estimator with a varying weight matrix determined by a kernel function rather than a fixed one. This let the model to represent the relation at a specific point in space without however ignoring the surroundings completely.

The second coefficient is the t-value. In order to compute the t-value it is necessary to obtain the local standard error. This accounts for the variations in the data used to compute the estimates parameters. Estimates parameters might be derived from regions with different data points densities. A small number of data points together with low weights given by sparseness, provide misleading results. Local standard error is crucial to account for this deficiency and to determine statistical significance indicators such as t-values and confidence intervals. This is computed as follows (Fortheringham, Brunsdon, Charlton, 2002):
By rewriting the local estimate parameter equation as:

(5) $$\beta(u_i v_i) = \mathrm{C} y$$

where,

(6) $$\mathrm{C} = \left(X^T W(u_i v_i) X\right)^{-1} X^T W(u_i v_i)$$

It is possible to calculate the variance of the estimate parameter as:

(7) $$Var\,[\beta(u_i v_i)] = \mathrm{C}\mathrm{C}^T\,\sigma^2$$

where σ^2 represents the normalised residual sum of squares given by the local regression. This is given by:

(8) $$\sigma^2 = \sum_{i=1}^{n} \frac{(y_i - y_i)}{(n - 2v1 + v2\ )}$$

where,

(9a) $$v_1 = tr(S)$$ (9b) $$v2 = tr(S^T S)$$

S is a matrix known as the hat matrix. The values of $\hat{y}$ are mapped on to y according to:

(10) $$y = Sy$$

Each row in the S matrix, denoted r_i, is given by:

(11) $$r_i = X_i\left(X^T W(u_i v_i) X\right)^{-1} X^T W(u_i v_i)$$

After having obtained the variance of each parameter estimate according to equation (Var), the local standard errors are derived from:

(12) $$SE(\beta_i)\ sqrt\,[Var\,(\beta_i)]$$

where,

(13) $$\beta_i = \beta(u_i v_i)$$

Finally, as in the conventional least square regression, t-values are then computed by dividing the estimate parameter at each location by the square error at the same location:

(14) $$t = \frac{\beta}{SE(\beta)}$$

3.2 KERNEL FUNCTION AND WEIGHTINH SCHEME

Both previous statistics required the determination of $W(u_i v_i)$. (This can be defined as a weight matrix tailored to coordinates $(u_i v_i)$. The elements of the matrix are set to assign more weight to those data points whose coordinates ui vi are closer to the coordinates of the location i for which coefficients are calculated. Each location $u_i\, v_i$ requires the determination of a new matrix:

(15)
$$W(u_i v_i) = \begin{bmatrix} W(u_i v_i)^1 & 0 & 0 \\ 0 & W(u_i v_i)^2 & 0 \\ 0 & 0 & W(u_i v_i)^n \end{bmatrix}$$

where $W(u_i v_i)$ is the weight given to the data point at $u_i v_i$. As a result, the W matrix weights are the result of a kernel function of the distance (usually Euclidean distance) of a data point from the coordinates at which statistics for the location *i* are computed.

The conventional Global OLS framework can be seen as model where the weighting scheme allocates to each data point the same weight without making distinction between spatial locations. In order to differentiate the importance given to data points according to distance, several kernel functions have been employed to determine rates at which the weight given to neighbour observations decreases. A simple kernel, called Boxcar (figure 1), represents a weighting scheme that assigns a weight of 1 to data points within a certain bandwidth (distance from location i) and 0 to those data points outside the bandwidth.

(16a) $W_{ii} = 1 \; if \; d_{ii} < d$ (16b) $W_{ii} = 0 \; otherwise$

On one hand, this approach represents a step further from a global model, it is subject to the problem of discontinuity. As the "window" moves on the data surface, the coefficients estimated might change drastically as data points are included or excluded from the subset. Even though sudden changes in local parameters may be genuinely representative of the underlying relations in the data when using this approach, these changes may as well be artefacts introduced by the arrangements of the data points making the Boxcar kernel an unpopular choice. However, scholars do not completely agree. In fact, in a study conducted by Siegmund and Worsley (1995) box shaped kernel provided smoother results surfaces than other kernels.

One of the most popular used kernels in GWR is the Gaussian (Fortheringham, Brunsdon and Charlton, 2002). this has the advantage of reducing the weight given to neighbour observation as a continuous function of the distance from location *i*.

(17) $$W_{ii} = exp\left[-\frac{1}{2}\left(d\frac{dii}{b}\right)^2\right]$$

A data point is given the weight of 1 if its coordinates coincide to those of the location *i*. Any other position will assign a decreasing weight according to a Gaussian curve as the distance between the data point and the location *i* increases. In this approach the inclusion of a data point in the calibration of the model is fractional. For instance, a data point can be given a weight of 0.5 instead of either 1 or 0 contributing only half weight in the calibration. Together with the Gaussian kernel other shapes have been regularly used in GWR (figure 1). By changing the kernel it is possible to differentiate the exponential decay profile which in turn yields estimates and other statistics that vary more or less rapidly over space.

Another kernel function often used is the Bisquare function (Brunsdon et al 1998a, Bivand and Brunstad, 2015):

(18a) $$W_{ii} = exp\left[-\frac{1}{2}\left(d\frac{dii}{b}\right)^2\right]^2 \; if \; d_{ii} < b$$

(18b) $$W_{ii} = 0 \; otherwise$$

This has the advantage of describing a near-Gausssian function up to bandwidth *b* from the location *i* assigning a fractional weight accordingly, and than 0 weight beyond that distance.

Very similar is the tricube kernel which resembles quite closely the bisquare kernel by assigning higher weights to close observations and lower weights to those further away Lu et al. (2015).

(19a) $$W_{ii} = exp\left[-\frac{1}{2}\left(d\frac{dii}{b}\right)^3\right]^3 \; if \; d_{ii} < b$$ (19b) $$W_{ii} = 0 \; otherwise$$

Finally, the Exponential kernel has the steepest shape with a weight decay function that assigns very little weights to the neighbour observations even though this are very close to the calibration point.

(20) $$W_{ii} = exp\left[-\left(d\frac{dii}{b}\right)\right] \; if \; d_{ii} < b$$

Since the effects of Gaussian kernel, Boxcar and Tricube have been reported as similar (Lu et al., 2015) this study will be focused on only the Gaussian compared to Boxcar and Exponential.

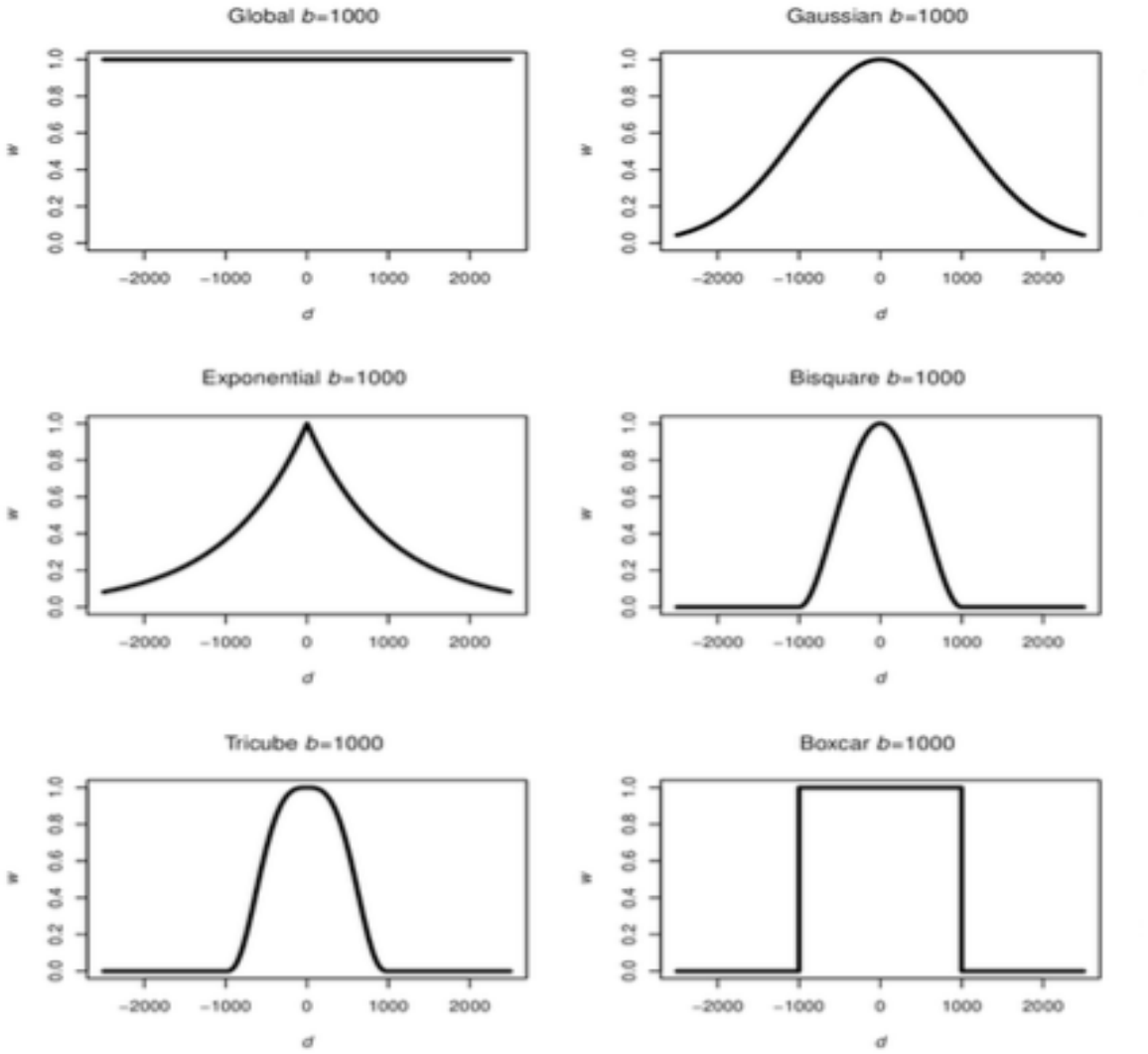


Figure 1 - compares the different shapes of kernel functions (Lu et al.,2015).

3.3. CALIBRATION

In GWR the effects of the kernel functions are highly dependent on the bandwidth size. As the bandwidth tends to infinity, the results of the GWR will resemble those of a conventional OLS regression with little if no variation. On the other hand, as the bandwidth decreases, the variance in the relations will increase. However, a too small bandwidth will prevent to model any relation since the regression will be performed on a subset too small to reach statistical significance.

The bandwidth size can be either fixed or adaptive. When fixed, this corresponds to a fixed distance from the location *i*. As a consequence, the number of neighbours of each location *i* varies according to the different density pattern in the data surface. Coefficients computed in denser areas will be based on more data points than those in sparse areas and vice versa. When adaptive, a number of neighbours is set and the bandwidth size adapts itself to include the number specified. A downside of the adaptive kernel is that if the number of neighbours is set relatively high and the location *i* belongs to a very sparse region, this will be influenced by data points far away and, as a result, it would return distort representation of the underlying relations in the data.

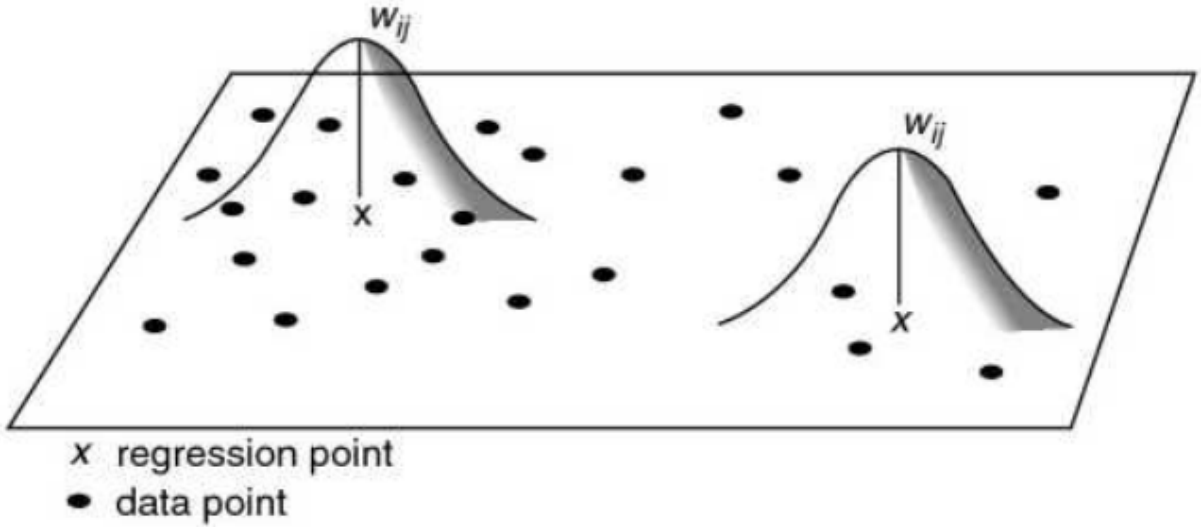


Figure 2 – representing fixed bandwidth. The number of data points used to compute coefficients at each regression point varies according to the density of the region (Fortheringham, Brunsdon and Charlton, 2002).

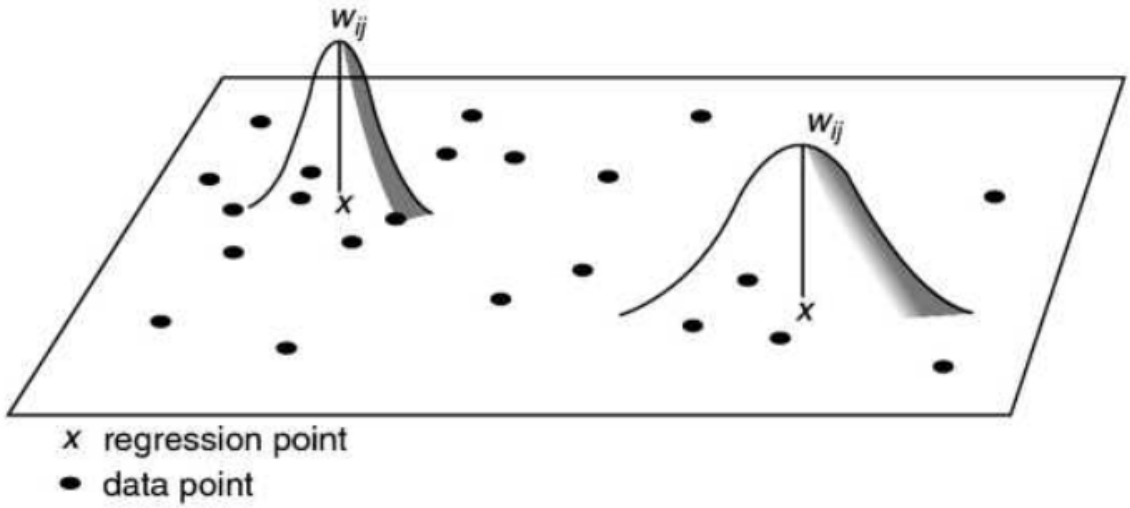


Figure 3: representing adaptive bandwidth. The number of data points used to compute the coefficients is always the same for each regression point (Fortheringham, Brunsdon and Charlton, 2002).

Determining the best bandwidth size is not an easy task. GWR is a visual technique that requires the human computer interaction loop in order to be set correctly since only by producing several maps it is possible to determine the goodness of parameter settings. However, model calibration techniques have been regularly used to determine what would have been an optimal size.

A naive solution to find the optimal bandwidth value ***b*** would be minimising the sum of square errors

(21) $$z = \sum_{i=1}^{n} [y_i - \hat{y}_i (b)]^2$$

where $\hat{y}_i(b)$ is the estimate value of y_i given bandwidth b. Bk*(uivi)* needs to be estimated for each data point to find $y_i - \hat{y}_i(b)$. The problem arises if we consider that, in order to minimise the error, the equation will shrink the bandwidth to such an extent to include only the data points closest or coinciding to y_i that is the actual value y_i. The sum of square errors might return even to 0 if the locations at which the regressions are computed correspond to the data points. This is of course not a helpful strategy inasmuch the statistical significance of the relations would not be sufficient and parameter coefficients would fluctuate wildly across the data surface.

Cross-Validation (CV) represents a solution to this problem.

(21) $$CV = \sum_{i=1}^{n} [y_i - \hat{y}_{\neq i}(b)]^2$$

Where $y_{\neq i}(b)$ is the estimated value of y_i according to the kernel bandwidth b.

CV calibrates the model similarly to the sum of square errors with the exception that it omits the data point corresponding to the location i. This allows the calibration to be performed on data points around i and not only i itself.

Finally, a second method for model calibration has got more popular in GWR. This is the Akaike Information Criterion (AIC) which provides a trade-off between degrees of freedom and goodness-of-fit (Hurvich et al. 1998).

(22)

$$AIC_c = 2n \log_e(\hat{\sigma}) + n \log_e(2\pi) + n\left\{\frac{n + \mathrm{tr}(S)}{n - 2 - \mathrm{tr}(S)}\right\}$$

where $\hat{\Sigma}$ is the estimate of the standard deviation, n is the size of the sample and $tr(S)$ is a function of the bandwidth representing the trace of the hat matrix. Compared to CV it has the advantage of taking into account differences in degrees of freedom. Notwithstanding, in general there is no significant difference between a bandwidth obtained through CV and a bandwidth obtained through AIC (Guo 2008). As Faber and Páez (2007) pointed out, among several studies using GWR, the most used calibration method was CV, used 54.7% of the time. AIC was the second most used with 20.3% of research papers adopting it, while 14.1% used predefined bandwidth and 10.9% did not mention how the bandwidth was selected.

4. EVALUATION AND TOOLS

4.1. Evaluation

The main focus of this study is on the different outcomes in local estimate parameters and local t-values caused by the following parameters:

Bandwidth type:	Adaptive/Fixed.
Bandwidth kernel function:	Gaussian, Exponential, Boxcar.
Bandwidth size:	Optimal CV Bandwidth / continuos

Since most of the studies used either Gaussian or Bisquare and the effects of these kernels were reported as similar (Brundson, Fortheringham and Charlton, 2002), the Gaussian was chosen as it is the most popular. This is compared to Exponential and Boxcar, which have not been extensively used.

Together with the change in parameters, the GWR will be applied to different density patterns as described in the following section. By density pattern it is meant the change in quantity of observations across the geographical area. As stated in the literature review, a comparison between behaviours of GWRs over different density patterns is missing. These can be evaluated both comparing the same bandwidth setting over different density patterns as well as comparing different bandwidth settings over the same density pattern.

Regarding the evaluation, GWR is a visual analytical technique that for its evaluation mainly depends on the visual interpretation (Mennis 2006, Brundson, Fortheringham and Charlton, 1998) based on a human computer interaction loop where the effects of changing parameters are visually evaluated by the analyst (Munzner and Maguire, 2014). This is the main tool of evaluation used for the analysis since it was not found any coefficient or test able to represent the goodness of the model other than visually interpret the maps produced. However, the limitation of this evaluation approach is given by the procedure itself that results extremely time consuming if variation of the bandwidth size are included. Each variable in a dataset generates two maps (Estimate parameter, t-values). Maps then have to be drawn again for each different parameter setting, that can be the bandwidth type: adaptive or fixed, or the kernel function: Gaussian, Exponential, Boxcar. Each variable in the dataset would produce then six maps. If detailed description of variations of the bandwidth size would be applied the number of maps to visualise would become prohibitive. This problem is addressed by drawing only maps for the best CV value. To compensate this lack of scope, graphs have been used to as a complementary tool of evaluation. These graphs show how the variation of bandwidth sizes affects estimate parameters and t-values comparing the three kernels in

the same graph keeping the same bandwidth type (adaptive or fixed). On the x-axis it is shown the bandwidth size by using a double reference. In fact, Exponential and Gaussian returned similar bandwidth values that could have been compared on the same scale, while Boxcar had much greater values. In order to provide an overview on how these parameter settings behave once moving away from the best CV bandwidth value, graphs where produced centering the upper reference at the value between the optimal CV bandwidth for Exponential and Gaussian, while the second reference was centred on the optimal bandwidth for Boxcar. The y-axis instead represents in turn the mean and the standard deviation of estimate parameters and t-values. These graphs together with the maps generated over the optimal CV bandwidth, returned a combination of in depth analysis (maps) and extensive analysis (graphs).

4.2. Tools

The programme languages used in this analysis are R and Python. R was chosen due to the package support that geographical information systems community dedicated to this software. Several GWR packages were available:

GWmodel (Lu et al., 2015), spgwr (Bivand and Yu, 2015), gwrr (Wheeler, 2015), McSpatial(McMillen, 2015). The paper by Lu et al. (2015) provides a good comparison table between these packages:

	GWmodel	**spgwr**	**gwrr**	**McSpatial**
Kernel functions:	Box-car, Bi-square, Tri-cube, Gaussian and Exponential	Bi-square, Tri-cube and Gaussian	Gaussian and Exponential	Rectangular, Triangular, Epanechnikov, Bi-square, Tri-cube, Tri-weight and Gaussian
Adaptive bandwidth?	Yes	Yes with `gwr.adapt`	No	No
Fixed bandwidth?	Yes	Yes	Yes	Yes
Spatial distance metrics:	Euclidean, Great Circle and Minkowski	Euclidean and Great Circle	Euclidean	Euclidean and Great Circle
Functions for weights matrix computation?	Yes with `gw.dist`	Yes with `gwr.bisquare`, `gwr.tricube` and `gwr.gauss`	No	Yes with `makew`

Table 1

The choice was the GWmodel package. This offered the advantage of choosing both adaptive and fixed bandwidth together with a wide variety of kernel functions. Moreover, features offered by other packages were not significant to the scope of this study.

Regarding Python, this was mainly used for the descriptive analysis of the data and the global model. Several libraries such as Numpy and Pandas make this programme language an agile solution for this task (Mckinney, 2012).

The choice of colour schemes used in this research were informed by ColorBrewer, an online mapping tool for choosing colour schemes for choropleth maps (Harrower and Brewer, 2003).

5. DATA DESCRIPITIVE ANALYSIS AND GLOBAL MODEL

5.1. GEORGIA CENSUS DATA SET DESCRIPTIVE ANALYSIS

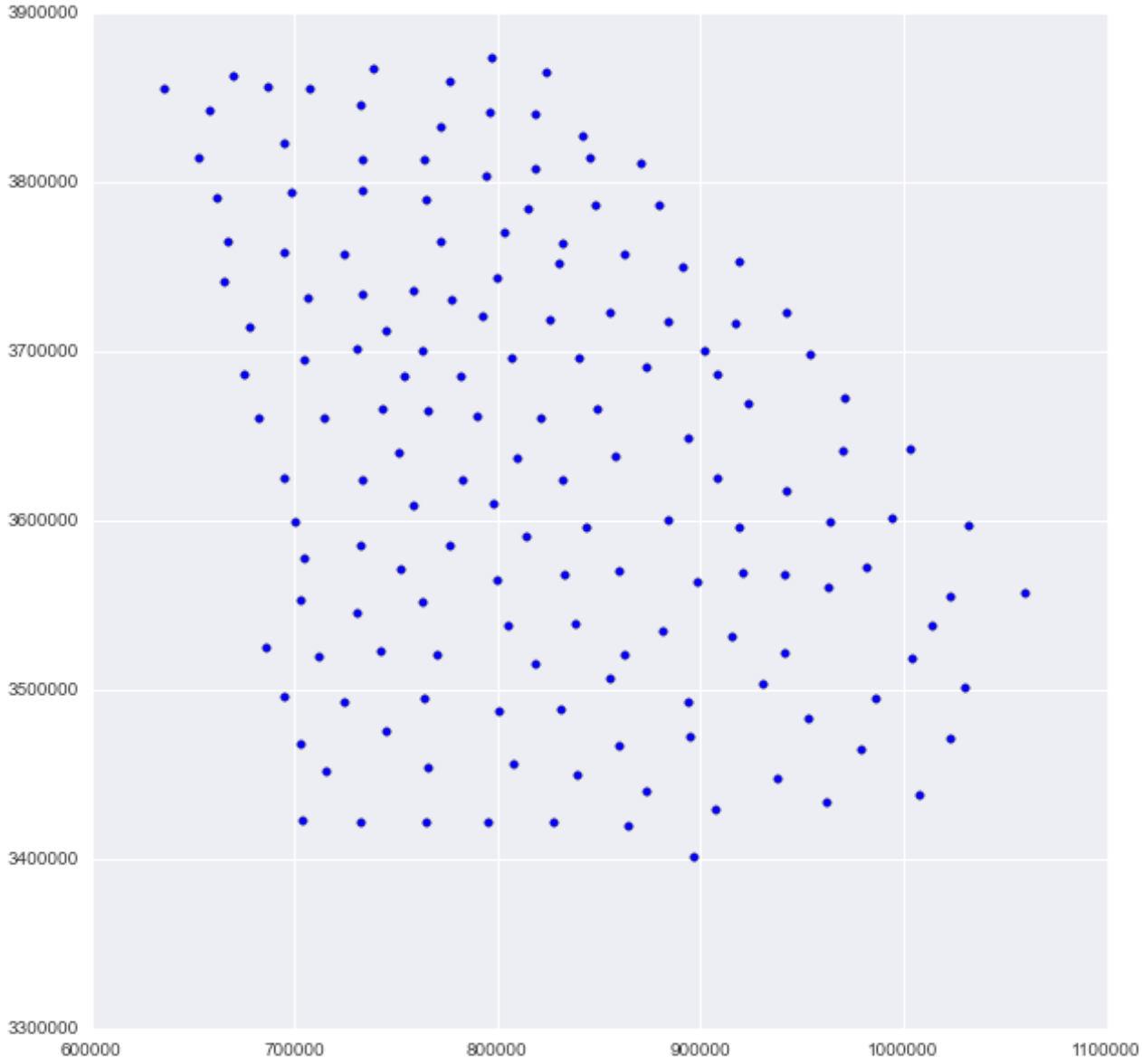


Figure 4

The first dataset chosen for this study is the Georgia census dataset. Data comes from Fotheringham, Brunsdon and Charlton (2002). The density pattern (figure 4) of the Georgia counties shows how the observations are uniformly distributed over space. This is arguably an optimal starting point for analysing the effects of bandwidth parameter since the geographical dimension should not play as a variable. In fact, different bandwidth kernels might have advantages or disadvantages when observations have a geographical pattern. Here it will be possible to compare settings over a plain ground to get how different shapes, more than fixed or adaptive, affect the results of a GWR.

Data represents census data for the US state of Georgia in 1990. Variables represent mainly social indicators such as education and society composition. Looking in details:

TotPop90 County population in 1990

PctRural County population percentage employed as rural workers

PctBach County population percentage with a bachelors degree

PctEld County population percentage aged at least 65

PctFB County population percentage of residents born outside the US

PctPov County population percentage living below the poverty line

PctBlack County population percentage of Black people.

The variables total population, percentage of bachelors holders and percentage of foreigners are right skewed while percentage of rural workers is left skewed. These variables describe a society mainly based on agricultural economy where people with high education are few and most of the work force is employed in the agricultural sector. The variables percentage of black, percentage of poverty and percentage of elderly resemble more a normal distribution.

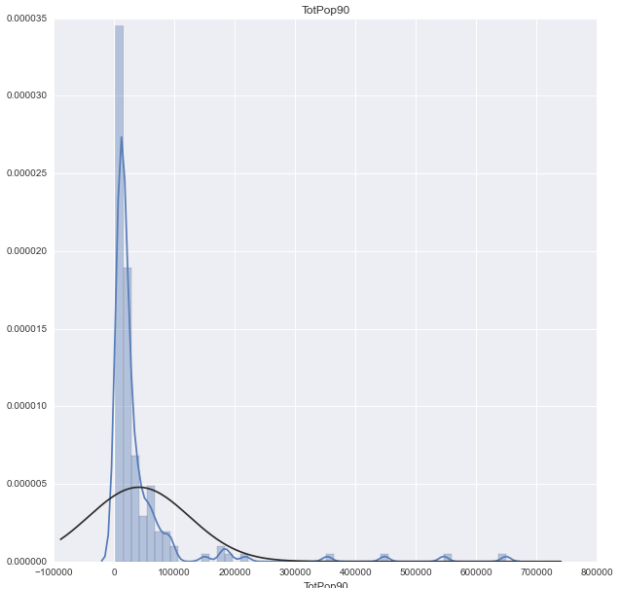


Figure 5

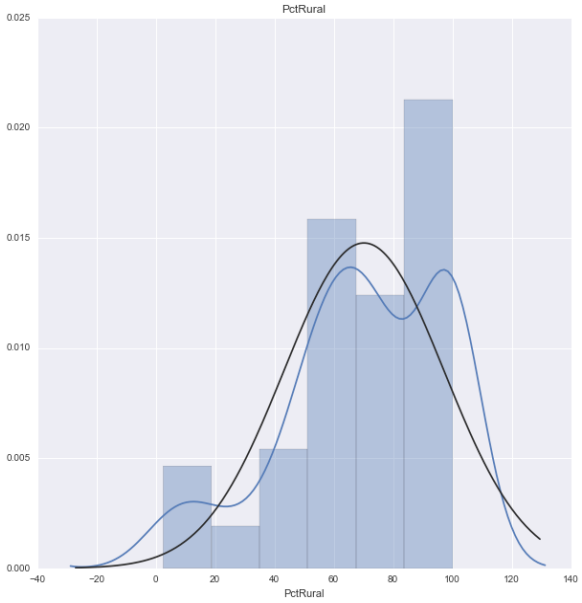


Figure 6

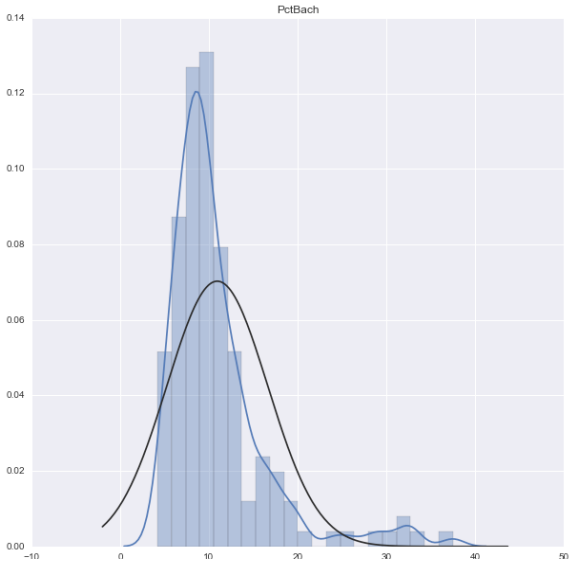


Figure 7

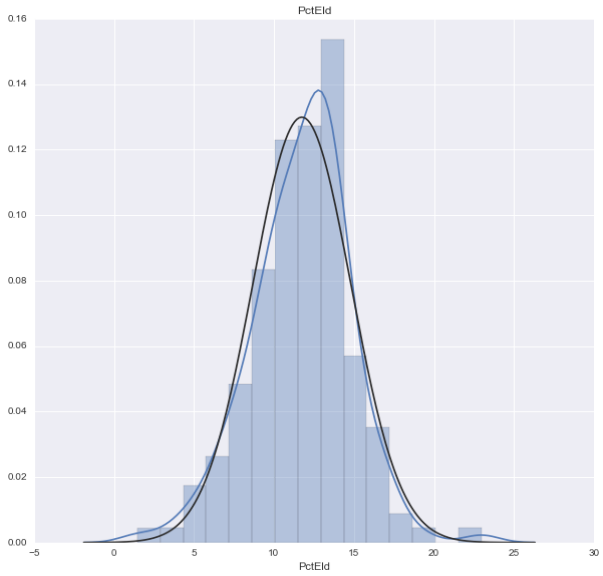


Figure 8

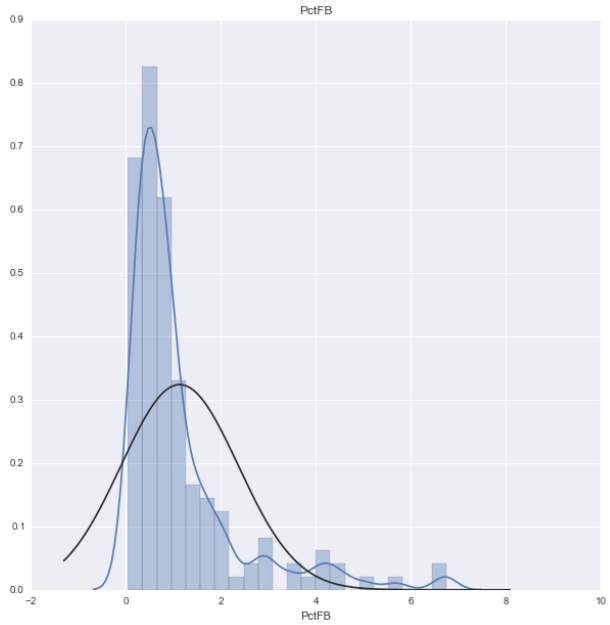


Figure 9

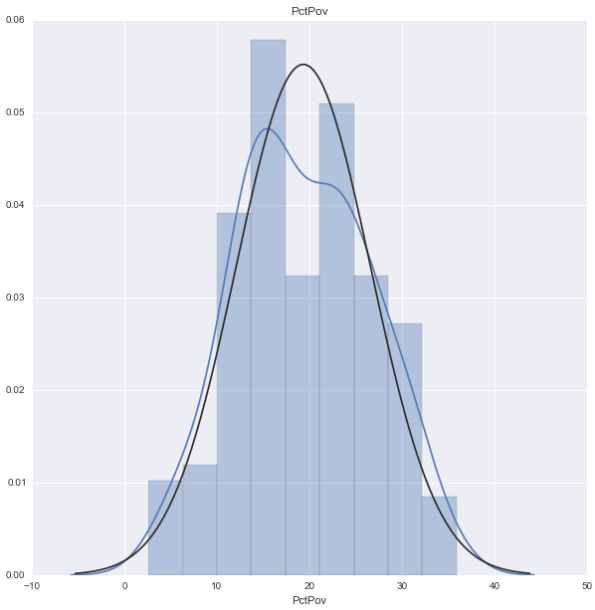


Figure10

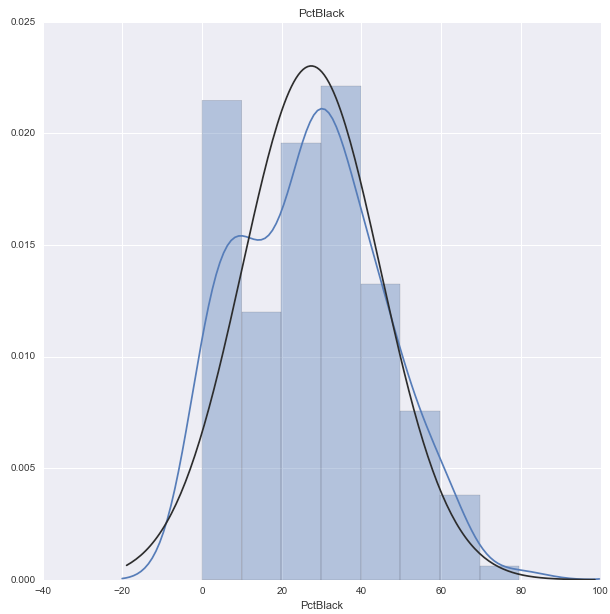


Figure 11

Looking at the summary statistics (table 2), the extreme high levels of peakiness expressed by the kurtosis in the total population together with the left skewness describe a society in which most of the population lives in small state and the four bounces in the left tail of the distribution are probably representing the main cities. In fact, the average percentage of population employed in the field is around 70%. Black people constitute as well a significant portion of society with an average of 27%

	TotPop90	PctRural	PctBach	PctEld	PctFB	PctPov	PctBlack
count	159	159	159	159	159	159	159
mean	40743.4969	70.1798742	10.9471698	11.7406289	1.13113208	19.3408805	27.3930818
std	83663.7652	27.0993736	5.69703244	3.08094114	1.23629447	7.25333807	17.3822894
min	1915	2.5	4.2	1.46	0.04	2.6	0
25%	9219.5	54.7	7.6	9.81	0.415	14.05	11.75
50%	16934	72.3	9.4	12.07	0.72	18.6	27.64
75%	36058	100	12	13.7	1.265	24.65	40.065
max	648951	100	37.5	22.96	6.74	35.9	79.64
kurtosis	28.3755851	-0.126072	6.03883105	1.14628627	6.10383755	-0.6653047	-0.5496494
Skewness	5.04580346	-0.7218126	2.25790866	-0.1831831	2.39233989	0.07773102	0.24264205

Table 2

The relation between variables is described by a Pearson correlation test (table 3). Poverty is greatly affected by the percentage of black people (0.736) and percentage of elderly (0.568). Apparently, percentage of black and percentage of rural are not correlated (-0.013). This gives insights on the high poverty level for the black community since society is mainly based on agriculture. The percentage of black and the percentage of elderly are also correlated with each other (0.297) rising some multicollinearity.

	TotPop90	PctRural	PctBach	PctEld	PctFB	PctPov	PctBlack
TotPop90	1	-0.6041953	0.71115683	-0.3452795	0.61149536	-0.31199	-0.0129644
PctRural	-0.6041953	1	-0.6188556	0.39031726	-0.5467835	0.17420183	-0.0690503
PctBach	0.71115683	-0.6188556	1	-0.4584957	0.67194669	-0.4016181	-0.1092925
PctEld	-0.3452795	0.39031726	-0.4584957	1	-0.4826286	0.56818326	0.29716672
PctFB	0.61149536	-0.5467835	0.67194669	-0.4826286	1	-0.3287093	-0.1120165
PctPov	-0.31199	0.17420183	-0.4016181	0.56818326	-0.3287093	1	0.73563771
PctBlack	-0.0129644	-0.0690503	-0.1092925	0.29716672	-0.1120165	0.73563771	1

Table 3

OLS Regression Results

Dep. Variable:	PctPov	R-squared:	0.719
Model:	OLS	Adj. R-squared:	0.708
Method:	Least Squares	F-statistic:	64.78
Date:	Mon, 31 Aug 2015	Prob (F-statistic):	2.04e-39
Time:	12:15:39	Log-Likelihood:	-439.28
No. Observations:	159	AIC:	892.6
Df Residuals:	152	BIC:	914.0
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	7.2571	2.290	3.169	0.002	2.732	11.782
TotPop90	-1.396e-05	5.69e-06	-2.452	0.015	-2.52e-05	-2.71e-06
PctRural	-0.0175	0.016	-1.090	0.277	-0.049	0.014
PctBach	-0.2000	0.091	-2.208	0.029	-0.379	-0.021
PctEld	0.7250	0.125	5.813	0.000	0.479	0.971
PctFB	0.3430	0.369	0.930	0.354	-0.386	1.072
PctBlack	0.2616	0.019	13.523	0.000	0.223	0.300

Omnibus:	7.871	Durbin-Watson:	2.096
Prob(Omnibus):	0.020	Jarque-Bera (JB):	7.653
Skew:	0.474	Prob(JB):	0.0218
Kurtosis:	3.507	Cond. No.	6.85e+05

Table 4

The results of the global model (table 4) with least squares regression reported that percentage of the elderly is the variable with the greatest estimate coefficient (0.725) highlighting the social hardship of the ageing part of the pupation. While the percentage of the elderly is the variable with the strongest coefficient, the percentage of black people has the highest significance with a t-value of 13.523. This reveals how detrimental conditions for the black community was a consistent phenomenon.

5.2 DUBLIN VOTERS TURN-OUT DATA

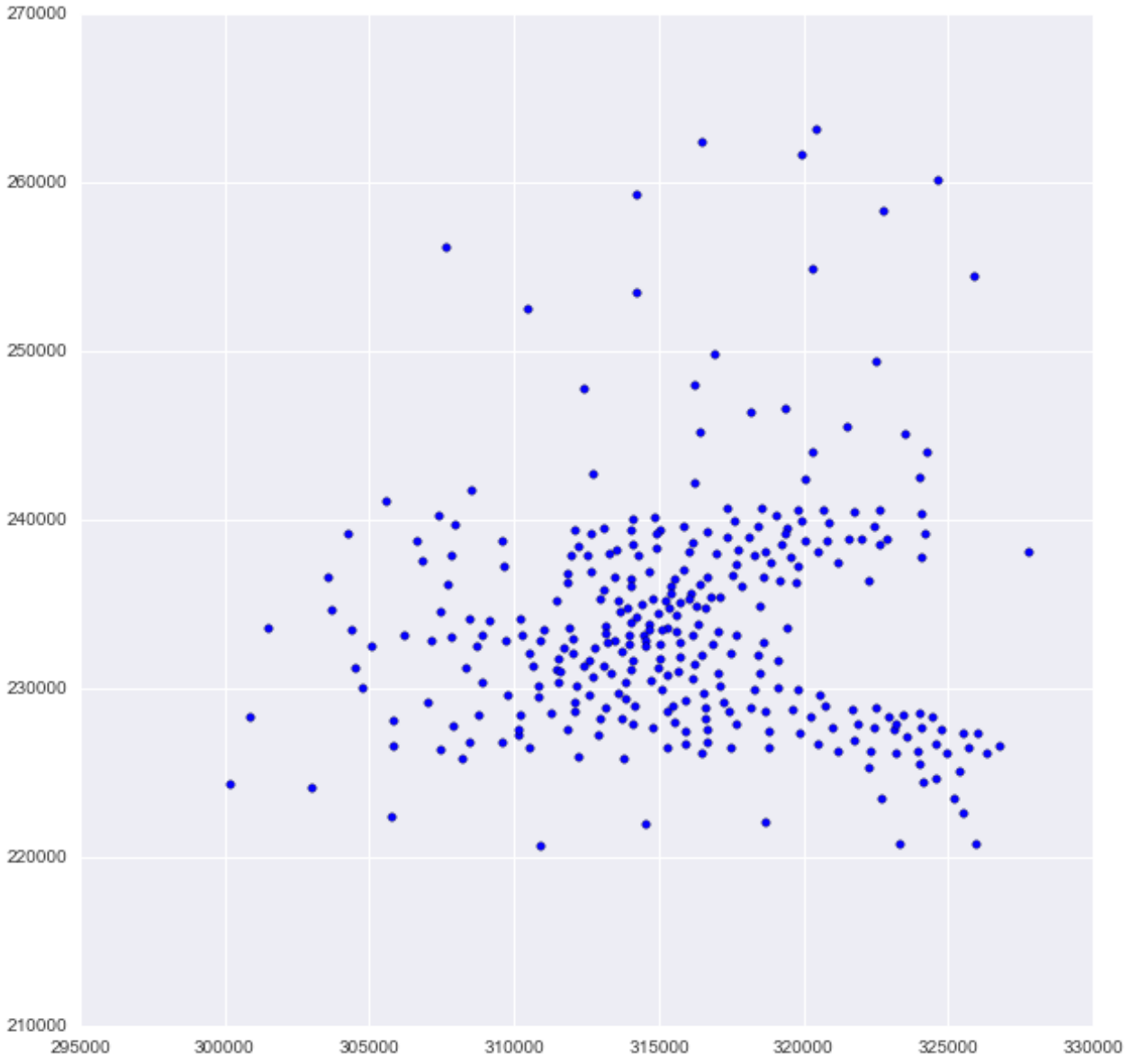


Figure 12

An interesting geographical pattern to test several bandwidth settings is represented by the voter turnout data in Greater Dublin data for the 2004 elections and the 2004 census. Data was taken from the studies of Kavanagh, A. (2006) and Harris et al. (2011). The density pattern (figure 12) of this dataset presents a dense area in the lower half of the map while the South border and the North West region is quite sparse. A varying density pattern could unveil what is the stability of the models based on different settings. When fixed, an Exponential kernel might be better at dealing with a denser region due to its small window. In fact, this could give a tailored coefficient calibrated only with close neighbours avoiding the loss of definition. On the other hand, a small window size might not be ideal when in sparser areas the support for the significance might be low and the coefficients might experience great variation.

The dataset is a collection of social indicators such as age, social class and level of education with percentage of turnout in 2004 as target variable. More specifically the indicators are:

DiffAdd percentage of the citizens in each district who are one-year migrants (i.e. moved to a different address 1 year ago)
LARent percentage of the citizens in each district who are local authority renters
SC1 percentage of the citizens in each district who are social class one (high social class)
Unempl percentage of the citizens in each district who are unemployed
LowEduc percentage of the citizens in each district who are with little formal education
Age18_24 percentage of the citizens in each district who are age group 18-24
Age25_44 percentage of the citizens in each district who are age group 25-44
Age45_64 percentage of the citizens in each district who are age group 45-64
GenEl2004 percentage of the citizens in each district ED who voted in 2004 election

Looking at their distributions, the variables GenEL2004, Age45-64 and Age18-24 follow a normal distribution whereas all the other variables are right skewed. Dublin appears to be a city where levels of unemployment may reach considerable high levels in some districts looking at the right tail of the percentage of unemployed distribution. What is more, local authority renters appear to be mostly confined in one district that reaches 100% which underlines a not uniform distribution of the phenomenon. On the other hand, citizens deemed as social class one are more uniformly distributed compared to unemployment and local renters as to indicate that the upper class lives across most parts of the town and not only in one area.

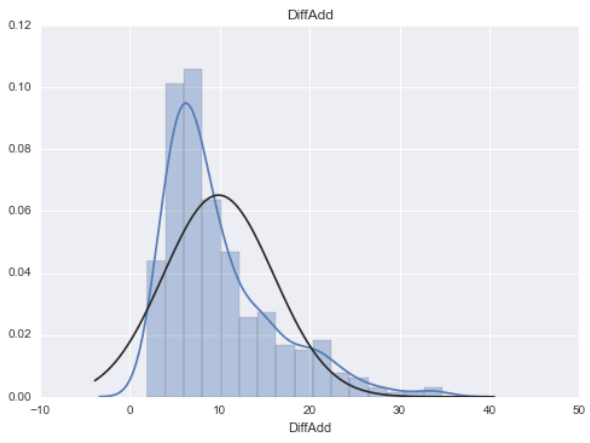


Figure 13

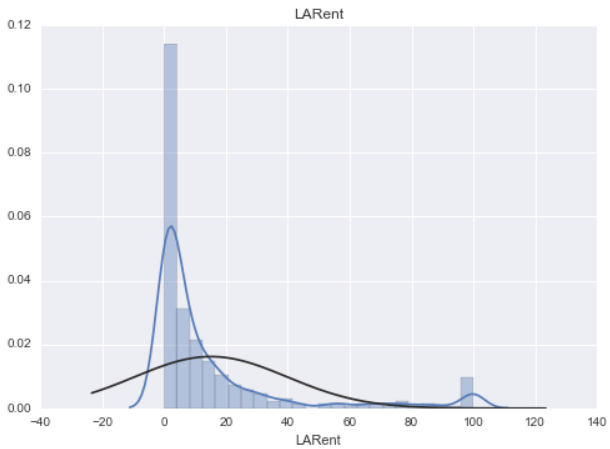


Figure 14

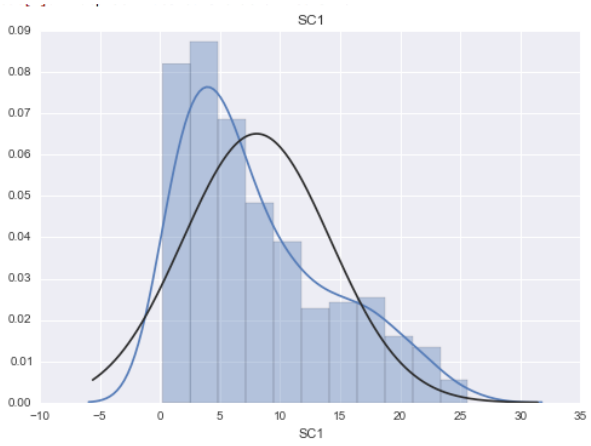


Figure 15

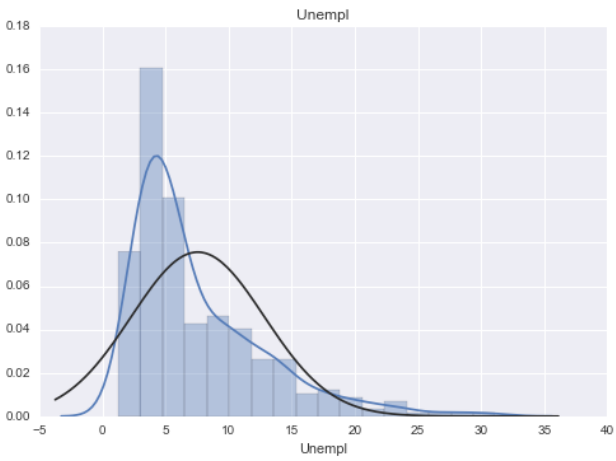


Figure 16

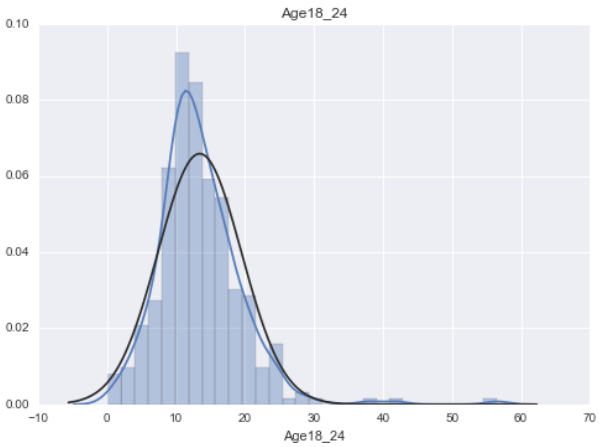


Figure 17

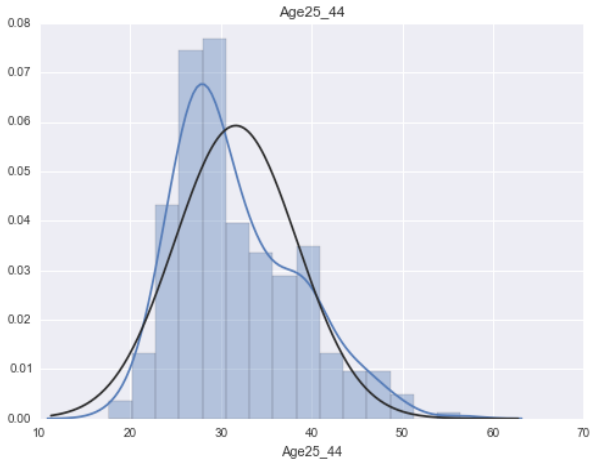


Figure 18

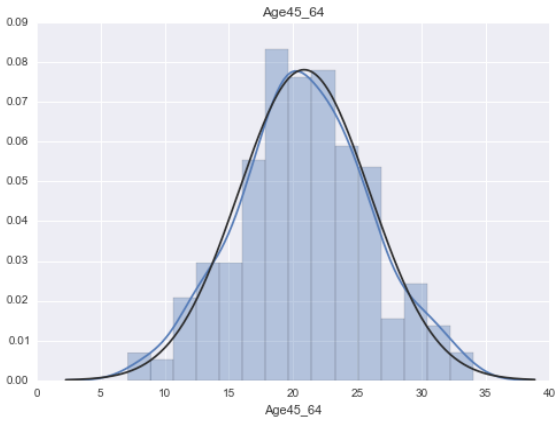


Figure 19

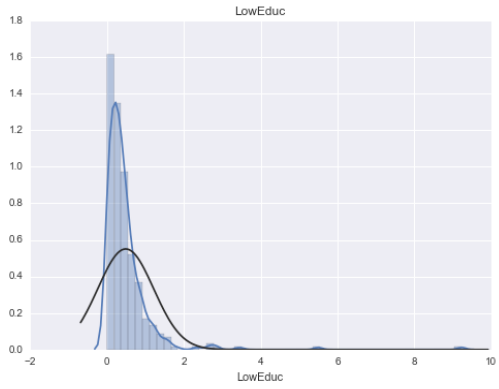


Figure 20

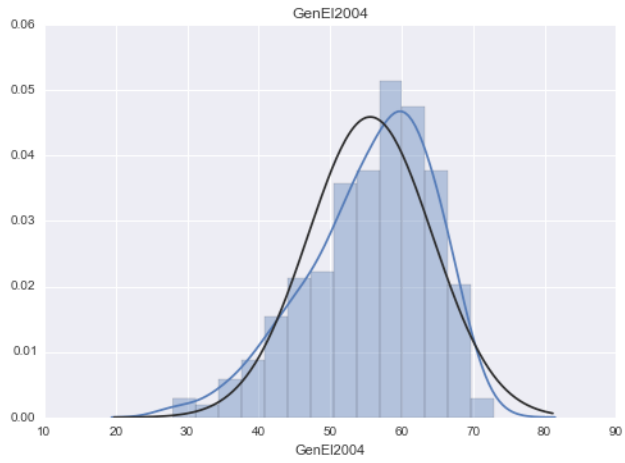


Figure 21

The summary statistics (table 5) gives insights on how percentage of low education with its extreme high kurtosis coefficient (72.043) and a mean of 0.4882 represents a negligible issue. Age 25-44 constitute the highest proportions and therefore the most crucial for the elections results. Percentage of unemployment is averagely under control but can reach very high levels in some areas (max:31.14). Similar trend is found in the variable percentage of one year migrant with a mean of 9.86 and a max of 34.74. Finally, percentage of local renters are a considerable part of the society and can reach 100% in a determine district.

	DiffAdd	LARent	SC1	Unempl	LowEduc	Age18_24	Age25_44	Age45_64	GenEl2004
count	322	322	322	322	322	322	322	322	322
mean	9.86342607	15.1705327	8.05127351	7.56303654	0.48824993	13.4328069	31.6466419	20.8653071	55.6142519
std	6.13345995	24.689051	6.14297413	5.28130045	0.72595679	6.0638089	6.74204556	5.11903173	8.70805929
min	1.904506	0	0.220204	1.257862	0	0.144092	17.633302	7.117797	27.984596
25%	5.48478425	1.11326575	3.2315635	3.9656325	0.156817	9.88645525	26.6871818	17.6864973	50.453984
50%	7.699046	4.836823	6.2096485	5.7526705	0.3323515	12.753749	29.991837	20.6839115	57.054095
75%	12.6034205	15.997051	11.6888748	10.0466488	0.5856355	16.560587	36.12597	24.124119	61.9857758
max	34.741144	100	25.629229	31.136364	9.235127	56.548176	56.404095	34.013605	72.914241
kurtosis	1.84386947	4.7032552	-0.2402466	2.72765927	72.0430317	8.98831875	0.11657145	-0.1451194	0.11881972
skewness	1.40890125	2.34870558	0.8358726	1.60213462	7.09601881	1.76847732	0.74591313	0.01773552	-0.6955535

Table 5

The relation between variables is described by a Pearson correlation test (table 6). The most influential variables with respect to voter turnout are percentages of local renters and unemployment that affect negatively the chance of voting with respectively -0.68 and -0.682 correlation coefficients. These two variables are strongly correlated (0.668) underlying the same feature. People aged between 45 and 64 seems to be the most interested in the elections (coefficient of 0.4849) as well as the upper class (coefficient of 0.352).

	DiffAdd	LARent	SC1	Unempl	LowEduc	Age18_24	Age25_44	Age45_64	GenEl2004
DiffAdd	1	0.27576301	0.37229879	0.00577997	-0.0318577	0.33530041	0.70306243	-0.5612839	-0.3082693
LARent	0.27576301	1	-0.2922633	0.66877617	0.16758975	0.25243277	0.3124497	-0.4626929	-0.6806665
SC1	0.37229879	-0.2922633	1	-0.5915679	-0.2727821	-0.0329514	0.0907555	0.08930325	0.35172646
Unempl	0.00577997	0.66877617	-0.5915679	1	0.28281831	0.11338083	0.13174128	-0.3742686	-0.6822627
LowEduc	-0.0318577	0.16758975	-0.2727821	0.28281831	1	-0.000127	0.02830716	-0.0723868	-0.1978619
Age18_24	0.33530041	0.25243277	-0.0329514	0.11338083	-0.000127	1	0.12619751	-0.2115455	-0.2597295
Age25_44	0.70306243	0.3124497	0.0907555	0.13174128	0.02830716	0.12619751	1	-0.69323	-0.4268253
Age45_64	-0.5612839	-0.4626929	0.08930325	-0.3742686	-0.0723868	-0.2115455	-0.69323	1	0.48362078
GenEl2004	-0.3082693	-0.6806665	0.35172646	-0.6822627	-0.1978619	-0.2597295	-0.4268253	0.48362078	1

Table 6

OLS Regression Results

Dep. Variable:	GenEl2004	R-squared:	0.638
Model:	OLS	Adj. R-squared:	0.629
Method:	Least Squares	F-statistic:	69.03
Date:	Sun, 30 Aug 2015	Prob (F-statistic):	1.35e-64
Time:	12:28:34	Log-Likelihood:	-989.58
No. Observations:	322	AIC:	1997.
Df Residuals:	313	BIC:	2031.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	77.7047	3.939	19.726	0.000	69.954	85.455
DiffAdd	-0.0858	0.086	-0.999	0.319	-0.255	0.083
LARent	-0.0940	0.018	-5.326	0.000	-0.129	-0.059
SC1	0.0864	0.071	1.219	0.224	-0.053	0.226
Unempl	-0.7216	0.094	-7.687	0.000	-0.906	-0.537
LowEduc	-0.1307	0.430	-0.304	0.761	-0.977	0.716
Age18_24	-0.1399	0.055	-2.554	0.011	-0.248	-0.032
Age25_44	-0.3536	0.075	-4.747	0.000	-0.500	-0.207
Age45_64	-0.0920	0.090	-1.020	0.309	-0.270	0.086

Omnibus:	25.034	Durbin-Watson:	1.728
Prob(Omnibus):	0.000	Jarque-Bera (JB):	37.927
Skew:	-0.525	Prob(JB):	5.81e-09
Kurtosis:	4.313	Cond. No.	628.

Table 7

The global model obtained through ordinary least squares regression confirms the analysis so far. Percentages of low education does not reach any statistical significance. Despite the strong t-value, percentage of local renters have a very weak estimate parameter indicating that this variable did not play a crucial role in determining voters turnout. On the other hand, percentage of unemployment seems to be the most substantial variable with the strongest estimate parameter (- 0.72) and the

strongest t-value (- 7.69) indicating how the lack of opportunity generated mistrust towards the society. Apparently, the highest levels of dissatisfaction towards policy-makers was expressed by the people aged between 25 and 44. This was negatively correlated with an estimate parameter of -0.35 and t-value of -4.75.

5.3. LONDON HOUSE PRICES DESCRIPTIVE ANALYSIS

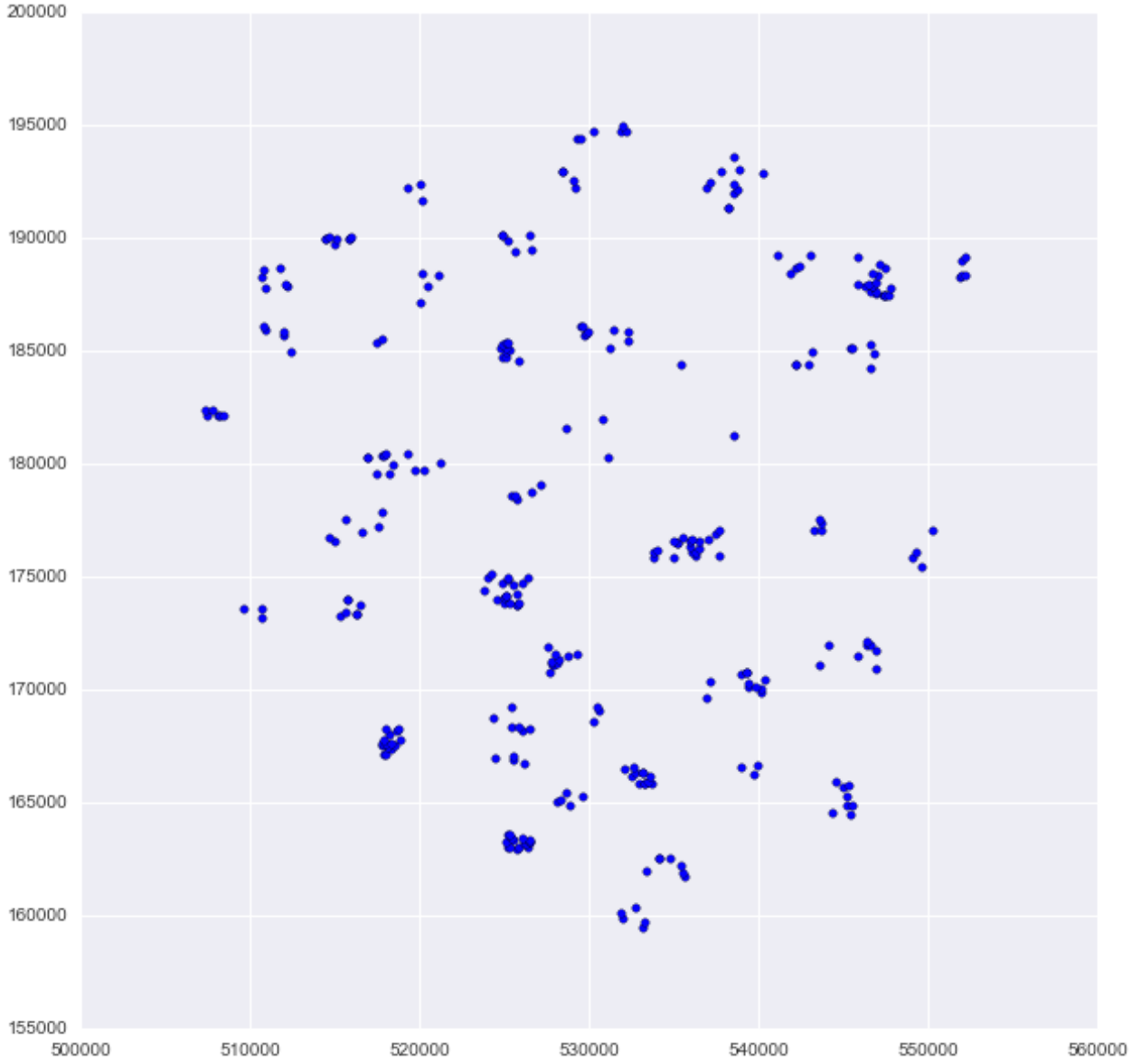


Figure 22

Data comes from the original study by Fotheringham, Brundson and Charlton (2002) and Charlton et al. (2014). The density pattern (figure 22) of this data shows a multitude of clusters, as well as isolated observations. There are no vast areas of greater density since clusters appear scattered over the map. An exception are the borders in the South East area that seem to have relatively higher concentration of observations within a single cluster compared to the others. It will be interesting to observe the different behaviour of bandwidth types when dealing with these clusters. An adaptive bandwidth during the calibration for one observation of that cluster may use a member of another cluster far away if the number of neighbours to use exceeds the members of the cluster. Similarly a fixed bandwidth with a large window, as in the case of boxcar for instance may fall in the same risk.

Also, the area immediately South from the centre is quite sparse. Due to its proximity with other more numerous clusters this may easily change their coefficients according to bandwidth settings.

Data represents London house prices in 2001. Variables are a mixture of structural and dimensional indicators such as floor size and type of bridling as well as census social indicators (percentage of professional and unemployed). More specifically:

PURCHASE the purchase price of the property
FLOORSZ floor area of the property in square metres
TYPEDETCH Boolean: 1 if the property is detached (i.e. it is a stand-alone house), 0 otherwise
TPSEMIDTCH Boolean :1 if the property is semi detached, 0 otherwise
TPSEMIDTCH Boolean: 1 if the property is in a terrace of similar houses (commonly referred to as a 'row house' in the USA), 0 otherwise
TYPEBNGLW Boolean: if the property is a bungalow (i.e. it has only one floor), 0 otherwise
TYPEFLAT Boolean: 1 if the property is a flat (or 'apartment' in the USA), 0 otherwise
BLDPWW1 Boolean: 1 if the property was built prior to 1914, 0 otherwise
BLDPOSTW Boolean: 1 if the property was built between 1940 and 1959, 0 otherwise
BLD60S a numeric vector, 1 if the property was built between 1960 and 1969, 0 otherwise
BLD70S a numeric vector, 1 if the property was built between 1970 and 1979, 0 otherwise
BLD80S a numeric vector, 1 if the property was built between 1980 and 1989, 0 otherwise
BLD90S a numeric vector, 1 if the property was built between 1990 and 2000, 0 otherwise
BATH2 a numeric vector, 1 if the property has more than 2 bathrooms, 0 otherwise
GARAGE a numeric vector,1 if the house has a garage, 0 otherwise
CENTHEAT a numeric vector, 1 if the house has central heating, 0 otherwise
BEDS2 a numeric vector, 1 if the property has more than 2 bedrooms, 0 otherwise
UNEMPLOY a numeric vector, the rate of unemployment in the census ward in which the house is located
PROF anumericvector,the proportion of the work force in professional or managerial occupations in the census ward in which the house is located

Moving to the variables distributions, the histograms for the continuous variables show how floor size, unemployment and purchase price are right skewed while professional tend more to a normal distribution. As expected, very large houses with high prices have the tendency to be an exception. Regarding the social indicators, the rate of unemployment is generally low while the level of professionals is much greater.

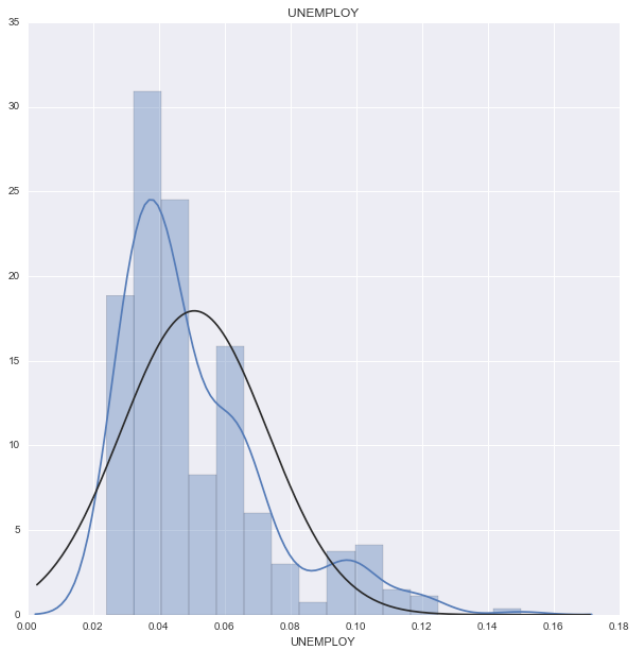


Figure 23

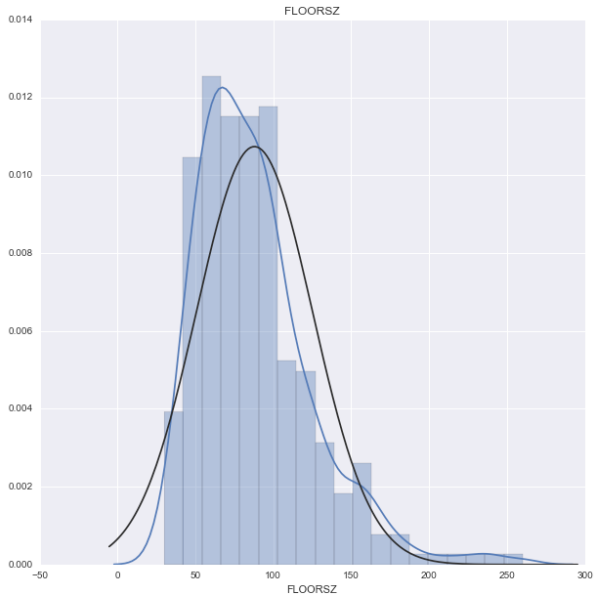


Figure 24

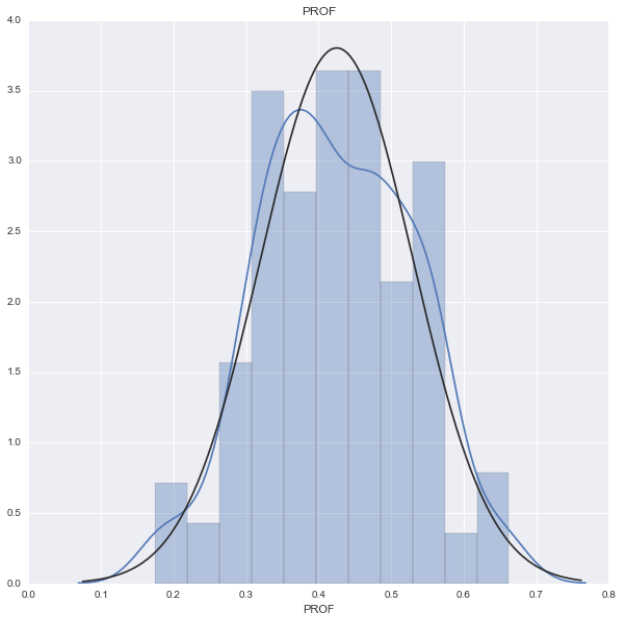


Figure 25

Looking at the summary statistics (table 8), the high levels of skewness and kurtosis of the purchase price suggest how prices are mostly contained in a range but the high level of right skewness shows that it is possible to go also far above this range. Floor size is much more contained in this regard with much lower variation both in terms of standard deviation and shape. The unemployment rate is generally much lower than levels of professionals living in the area. However, this can vary more with respect to its mean and reach significantly higher proportions in some areas.

	PURCHASE	FLOORSZ	UNEMPLOY	PROF
count	316	316	316	316
mean	150879.794	87.9240506	0.05080435	0.42487805
std	75399.5578	37.2269819	0.02226601	0.10511368
min	45000	30	0.02408854	0.17441655
25%	99998.75	61	0.03416149	0.34656944
50%	132000	80.5	0.04312389	0.41791488
75%	179962.5	104.25	0.0632284	0.50353635
max	567500	260	0.14997581	0.66231343
kurtosis	6.10469345	2.73743369	1.98067157	-0.4549512
Skewness	1.99733869	1.36983723	1.42786503	-0.0170848

Table 8

	PURCHASE	FLOORSZ	TYPEDETCH	TPSEMIDTCH	TYPETRRD	TYPEBNGLW	TYPEFLAT	BLDPWW1	BLDPOSTW	BLD60S
PURCHASE	1	0.69719225	0.30741164	0.30589873	0.04731533	-0.0019061	-0.3579457	0.11155131	-0.1170463	-0.0722659
FLOORSZ	0.69719225	1	0.30532545	0.38228387	0.26725644	-0.0332787	-0.6068853	-0.043241	-0.0258979	-0.018363
TYPEDETCH	0.30741164	0.30532545	1	-0.0746167	-0.1158535	0.30293796	-0.1901725	-0.0971575	-0.0529862	0.04465998
TPSEMIDTCH	0.30589873	0.38228387	-0.0746167	1	-0.2645251	-0.0467358	-0.4342156	-0.1394628	0.14496792	-0.0921424
TYPETRRD	0.04731533	0.26725644	-0.1158535	-0.2645251	1	-0.0725642	-0.6741838	0.03053957	-0.0846107	-0.0447855
TYPEBNGLW	-0.0019061	-0.0332787	0.30293796	-0.0467358	-0.0725642	1	-0.1191135	-0.0844781	-0.0331876	0.10784594
TYPEFLAT	-0.3579457	-0.6068853	-0.1901725	-0.4342156	-0.6741838	-0.1191135	1	0.11422883	-0.0031206	0.06318969
BLDPWW1	0.11155131	-0.043241	-0.0971575	-0.1394628	0.03053957	-0.0844781	0.11422883	1	-0.2186828	-0.1665534
BLDPOSTW	-0.1170463	-0.0258979	-0.0529862	0.14496792	-0.0846107	-0.0331876	-0.0031206	-0.2186828	1	-0.0654314
BLD60S	-0.0722659	-0.018363	0.04465998	-0.0921424	-0.0447855	0.10784594	0.06318969	-0.1665534	-0.0654314	1
BLD70S	-0.0403827	0.01747039	0.05390861	-0.056449	0.0022765	0.05203568	0.00284767	-0.2590634	-0.1017744	-0.0775135
BLD80S	-0.2292717	-0.2559935	-0.061731	-0.1409486	-0.0593963	-0.0386649	0.17954924	-0.2547739	-0.1000892	-0.07623
BLD90S	-0.0944479	-0.1020323	-0.0374446	-0.0403177	-0.0625991	-0.0234533	0.10117543	-0.1545402	-0.0607119	-0.0462395
BATH2	0.3661559	0.39282562	0.13447467	0.09232853	0.04530224	-0.0353049	-0.1496101	-0.0467622	-0.0501348	-0.0172357
BEDS2	0.34661735	0.53541478	0.10437072	0.23830678	0.3539192	-0.065372	-0.5049156	-0.1639157	0.14214846	0.02577696
GARAGE1	0.2183739	0.37461652	0.17965128	0.34494603	-0.0500916	-0.0675793	-0.2534628	-0.4302964	0.01155014	0.13730324
CENTHEAT	0.16285753	0.16717382	0.0627703	0.11436412	0.04268441	0.03931584	-0.1460076	0.04598679	-0.011734	-0.0185424
UNEMPLOY	-0.1023594	-0.0736855	-0.1500499	-0.166928	0.06333263	-0.0392069	0.11348507	0.33125955	-0.0634143	-0.0530922
PROF	0.38209262	-0.0298063	0.09559876	0.04813679	-0.3356872	-0.0181659	0.23842863	0.10298525	-0.1269677	-0.0277436
BLDINTW	0.21559995	0.28349219	0.13857328	0.26338752	0.1081979	0.06106565	-0.3254642	-0.4452996	-0.1749382	-0.1332366

Table 9

	BLD70S	BLD80S	BLD90S	BATH2	BEDS2	GARAGE1	CENTHEAT	UNEMPLOY	PROF	BLDINTW
PURCHASE	-0.0403827	-0.2292717	-0.0944479	0.3661559	0.34661735	0.2183739	0.16285753	-0.1023594	0.38209262	0.21559995
FLOORSZ	0.01747039	-0.2559935	-0.1020323	0.39282562	0.53541478	0.37461652	0.16717382	-0.0736855	-0.0298063	0.28349219
TYPEDETCH	0.05390861	-0.061731	-0.0374446	0.13447467	0.10437072	0.17965128	0.0627703	-0.1500499	0.09559876	0.13857328
TPSEMIDTCH	-0.056449	-0.1409486	-0.0403177	0.09232853	0.23830678	0.34494603	0.11436412	-0.166928	0.04813679	0.26338752
TYPETRRD	0.0022765	-0.0593963	-0.0625991	0.04530224	0.3539192	-0.0500916	0.04268441	0.06333263	-0.3356872	0.1081979
TYPEBNGLW	0.05203568	-0.0386649	-0.0234533	-0.0353049	-0.065372	-0.0675793	0.03931584	-0.0392069	-0.0181659	0.06106565
TYPEFLAT	0.00284767	0.17954924	0.10117543	-0.1496101	-0.5049156	-0.2534628	-0.1460076	0.11348507	0.23842863	-0.3254642
BLDPWW1	-0.2590634	-0.2547739	-0.1545402	-0.0467622	-0.1639157	-0.4302964	0.04598679	0.33125955	0.10298525	-0.4452996
BLDPOSTW	-0.1017744	-0.1000892	-0.0607119	-0.0501348	0.14214846	0.01155014	-0.011734	-0.0634143	-0.1269677	-0.1749382
BLD60S	-0.0775135	-0.07623	-0.0462395	-0.0172357	0.02577696	0.13730324	-0.0185424	-0.0530922	-0.0277436	-0.1332366
BLD70S	1	-0.118571	-0.0719226	-0.0004549	0.0825474	0.23367872	0.05465165	-0.0996669	0.11485549	-0.2072411
BLD80S	-0.118571	1	-0.0707317	-0.0336479	-0.2091015	-0.0392142	-0.1485837	-0.0832415	-0.0314619	-0.2038097
BLD90S	-0.0719226	-0.0707317	1	-0.0085167	-0.2115798	-0.0150093	-0.0823367	0.09039075	-0.080758	-0.1236265
BATH2	-0.0004549	-0.0336479	-0.0085167	1	0.12858612	0.16816168	0.0723299	0.01039903	0.0325979	0.11755302
BEDS2	0.0825474	-0.2091015	-0.2115798	0.12858612	1	0.21173495	0.20047226	-0.1759944	-0.0195494	0.26155493
GARAGE1	0.23367872	-0.0392142	-0.0150093	0.16816168	0.21173495	1	0.13762222	-0.27388	0.00917428	0.26469828
CENTHEAT	0.05465165	-0.1485837	-0.0823367	0.0723299	0.20047226	0.13762222	1	-0.0106815	0.05514975	0.0680033
UNEMPLOY	-0.0996669	-0.0832415	0.09039075	0.01039903	-0.1759944	-0.27388	-0.0106815	1	-0.4750331	-0.209
PROF	0.11485549	-0.0314619	-0.080758	0.0325979	-0.0195494	0.00917428	0.05514975	-0.4750331	1	-0.0434444
BLDINTW	-0.2072411	-0.2038097	-0.1236265	0.11755302	0.26155493	0.26469828	0.0680033	-0.209	-0.0434444	1

Table 10

The Pearson correlation test (table 9 and 10) reveals that purchase price is greatly correlated with floor size (0.697) and professionals (0.382). Also structural features such as two bathrooms (0.366) and type flat (0.358) have a predictive power over price. Regarding multicollinearity, professionals and purchase price, the two most substantial indicators, appear to have no correlation (-0.03).

The results of a global model using an ordinary least squares regressions (table 11) shows the following results:

OLS Regression Results

Dep. Variable:	PURCHASE	R-squared:	0.705
Model:	OLS	Adj. R-squared:	0.687
Method:	Least Squares	F-statistic:	39.42
Date:	Mon, 31 Aug 2015	Prob (F-statistic):	8.97e-68
Time:	13:49:52	Log-Likelihood:	-3803.9
No. Observations:	316	AIC:	7646.
Df Residuals:	297	BIC:	7717.
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	-1.217e+05	4.45e+04	-2.734	0.007	-2.09e+05	-3.41e+04
FLOORSZ	1179.8012	103.626	11.385	0.000	975.866	1383.736
TYPEDETCH	2.932e+04	4.39e+04	0.668	0.505	-5.71e+04	1.16e+05
TPSEMIDTCH	1481.0283	4.72e+04	0.031	0.975	-9.13e+04	9.43e+04
TYPETRRD	-6514.1405	4.7e+04	-0.139	0.890	-9.91e+04	8.6e+04
TYPEBNGLW	6534.0419	3.44e+04	0.190	0.850	-6.12e+04	7.42e+04
TYPEFLAT	-1.782e+04	4.64e+04	-0.384	0.701	-1.09e+05	7.34e+04
BLDPWW1	-7406.3612	9101.498	-0.814	0.416	-2.53e+04	1.05e+04
BLDPOSTW	-1.956e+04	1.02e+04	-1.927	0.055	-3.95e+04	416.042
BLD60S	-2.534e+04	1.13e+04	-2.242	0.026	-4.76e+04	-3101.629
BLD70S	-3.426e+04	9579.364	-3.576	0.000	-5.31e+04	-1.54e+04
BLD80S	-1.674e+04	9546.033	-1.754	0.081	-3.55e+04	2047.191
BLD90S	-1.128e+04	1.28e+04	-0.882	0.378	-3.65e+04	1.39e+04
BATH2	2.221e+04	9264.889	2.398	0.017	3980.805	4.04e+04
BEDS2	2710.4967	7526.124	0.360	0.719	-1.21e+04	1.75e+04
GARAGE1	2701.6922	7111.878	0.380	0.704	-1.13e+04	1.67e+04
CENTHEAT	1338.3264	8011.810	0.167	0.867	-1.44e+04	1.71e+04
UNEMPLOY	6.575e+05	1.44e+05	4.572	0.000	3.74e+05	9.41e+05
PROF	3.593e+05	3.07e+04	11.715	0.000	2.99e+05	4.2e+05
BLDINTW	-7069.5367	8728.553	-0.810	0.419	-2.42e+04	1.01e+04

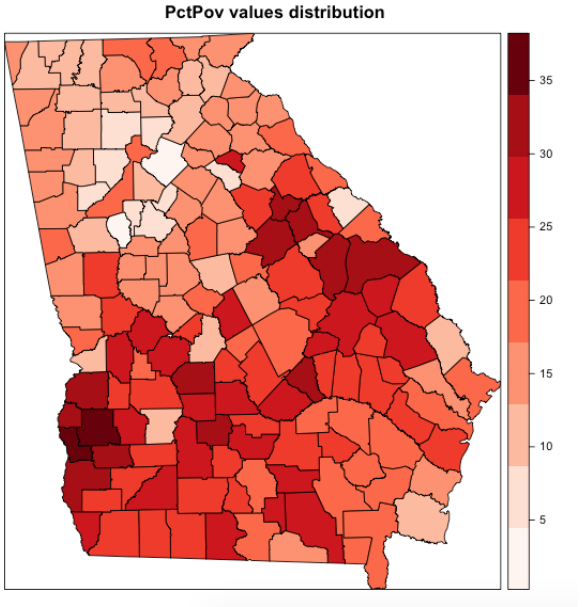
Omnibus:	94.608	Durbin-Watson:	1.596
Prob(Omnibus):	0.000	Jarque-Bera (JB):	421.711
Skew:	1.187	Prob(JB):	2.67e-92
Kurtosis:	8.137	Cond. No.	6.17e+17

Table 11

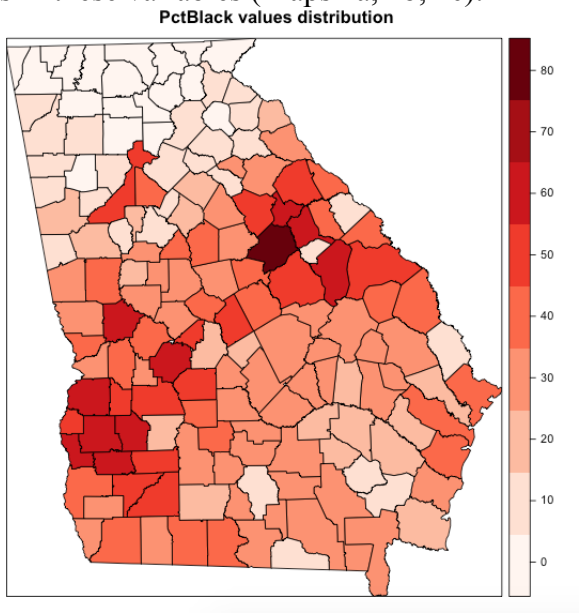
As expected, floor size and professionals are two most substantials variables in the model with 99.9% significance and 1189.8 and 3.593e+05 estimate parameters respectively. Twelve variables do not reach statistical significance. These are mostly structural variables suggesting that the geographical position of the house affects its price more than its typology.

6. GEORGIA CENSUS DATA ANALYSIS

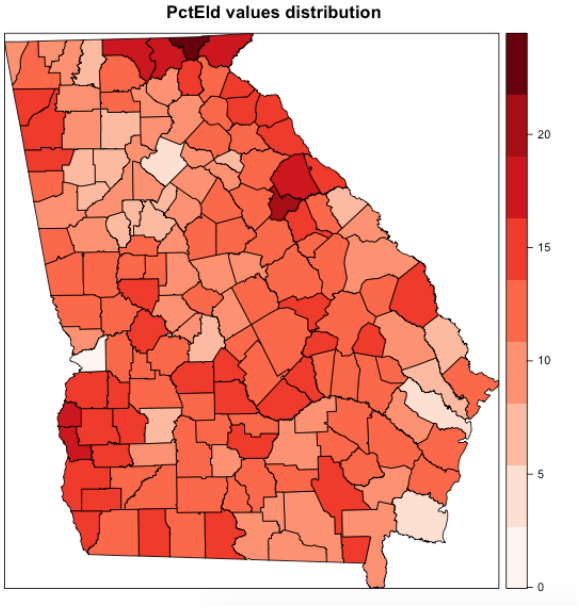
The first dataset under investigation is the Georgia census data set. The density pattern that could be described as uniform makes this a good starting point since the evenly distributed observations should describe the behaviour of the different GWR settings when varying levels of densities do not apply as a variable. Looking at the results from the data description section, the variables that turned out to be more substantial over the dependent variable percentage of poverty in the OLS regression are the percentage of black people which has a strong t-value (over 13) and an estimate parameter of 0.26 and the percentage of elderly with t-value of 5.813 and estimate parameter of 0.725, making it less significant than the percentage of black people but more strongly correlated. Looking at the geographical distribution of the values in these variables (Maps 1a, 1b, 1c):



Map 1 a



Map 1b



Map 1 c

The percentage of poverty has high values in the southern part of the state with a slightly better condition near the border with Florida. Looking at the percentage of elderly, this is mainly evenly spread across the map with higher values on the top near the border with Tennessee. Finally, the percentage of Black people is very low in the north with highest values in the areas close to Madison in the North East and Columbus in the South West.

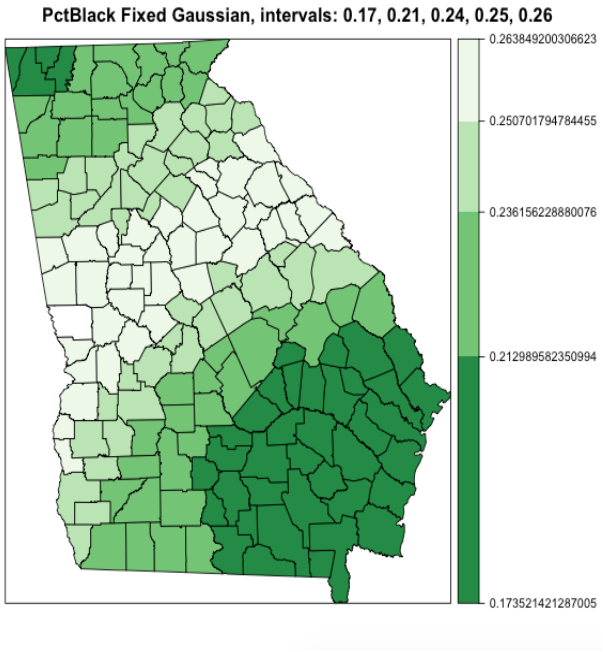
6.1. PERCENTAGE OF BLACK PEOPLE

The first set of maps (map1a, 1b, 1c) shows varying degrees of spatial correlation. Overall, the percentage of black people has stronger positive correlation near the area in correspondence of cities highly populated by Afro Americans compared to the North West and the South East.

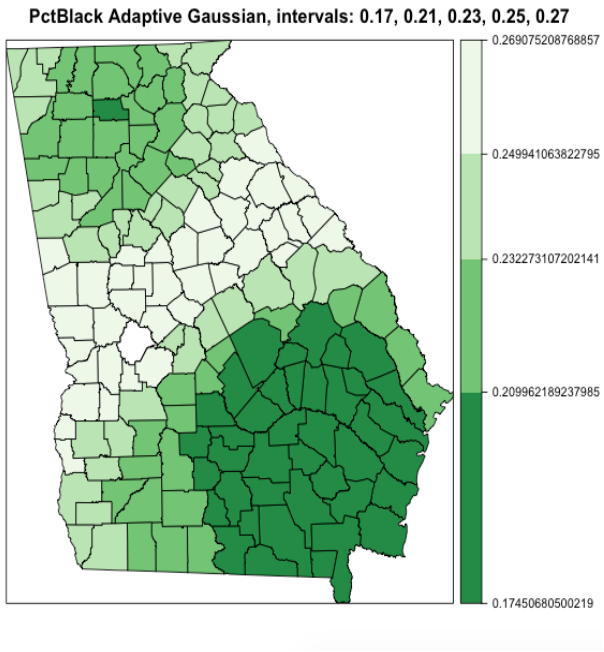
Looking at the first two maps, Gaussian fixed and Gaussian adaptive (maps 2 a, 2b), it is possible to notice how the choice of adaptive or fixed bandwidth, despite returning a similar overall trend, differs greatly at the edges of the map. In fact, the North West corner that in the map with Gaussian fixed bandwidth was regarded as low correlation has instead an above average estimate in the map with the adaptive bandwidth. Comparing the two maps with the geographical values distribution (Maps 1a, 1b) we can notice how the north has both a low percentage of black people as well as a low percentage of poverty, suggesting the same trend as the global model (+ 0.2). The Gaussian map drawn with fixed bandwidth seems not to capture this latent correlation compared to the same map with adaptive bandwidth which creates a pole of low correlation where percentage of black people is high but percentage of poverty is relatively low. Compared to Gaussian, the Exponential bandwidth (maps 2c, 2d) due to its steeper shape has a smaller bandwidth. The smaller radius, or fewer neighbours, used to determine the estimate parameters shows a division between the transversal high correlation area suggested by the Gaussian kernel. The zone between Madison and Columbus does not show the same strong correlation as the areas of the two cities. Apparently, the smaller bandwidth reduces the level of smoothness of the map. Turning to the Boxcar (2e, 2f), the much higher bandwidth value and the same weight applied to all observations returned an approximative description where the strongest correlation is given to the area West from Madison totally ignoring the correlation around Columbus.

While the difference between Adaptive and Fixed bandwidth in estimate parameters was limited to variations at the edge of the map, the results for t-values present a much stronger variation. In the case of Exponential kernel (maps 3c, 3d), high levels of significance correspond to high estimate parameters giving stronger significance to the area around Madison. The pattern shown by the fixed

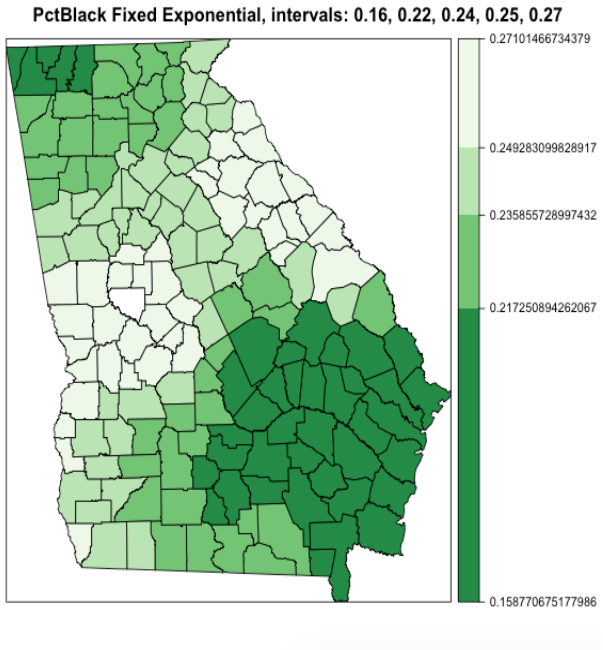
bandwidth is different. High significance is concentrated towards the centre of the map. Unsurprisingly, observations on the edge are less likely to gain high significance due to the fewer number of neighbours captured by the bandwidth. This is confirmed by the trend of Gaussian and Boxcar kernels (maps 3a, 3b, 3f) that, by having a larger bandwidth, return greater areas of high significance. Although the variation in the maps with fixed bandwidth is not great, the variation in the adaptive maps depicts more dissimilar patterns. In particular, the Boxcar adaptive map loses the high significance trend over the boarder presented by the two other kernels. The only possible distinction is between a North area with high significance and a south area with low significance. All areas are above the significance threshold of 1.96. What is more, while the highest t-values levels expressed by Fixed or Adaptive maps are similar, the lowest t-values are generally higher for adaptive maps underlying further the problem of the significance for observations at the edges of the map.



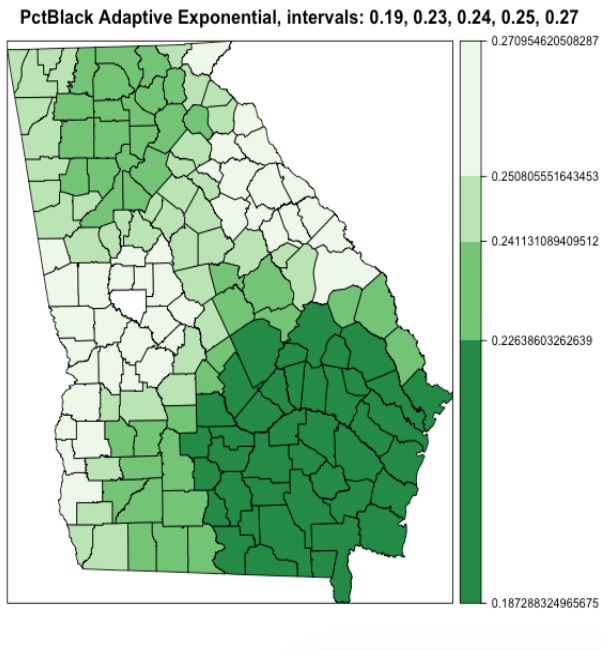
Map 2 a



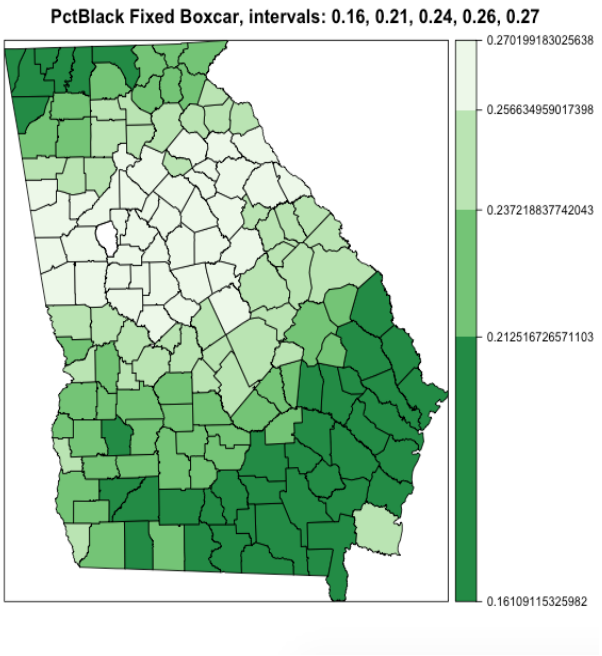
Map 2 b



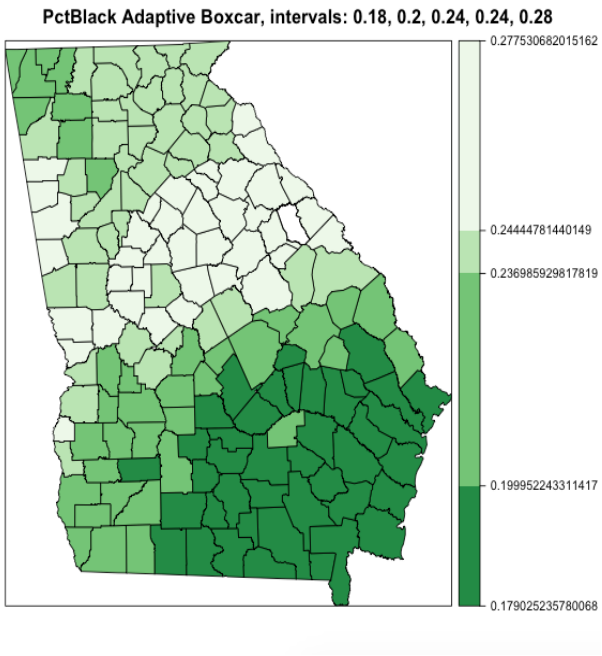
Map 2 c



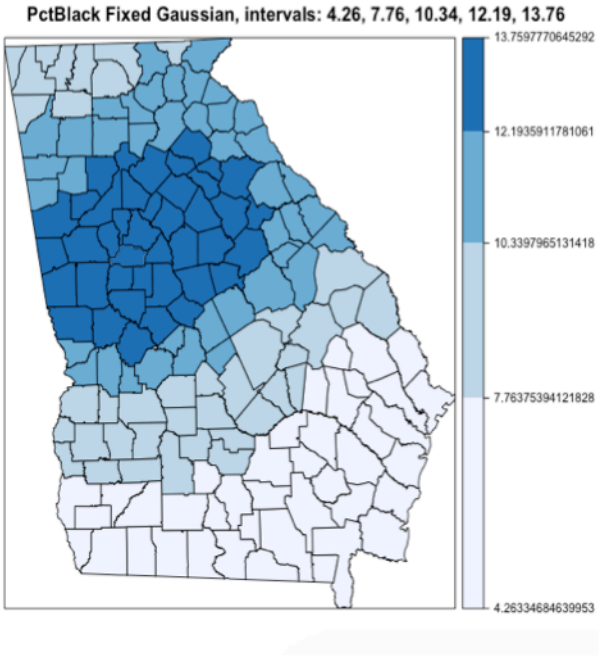
Map 2 d



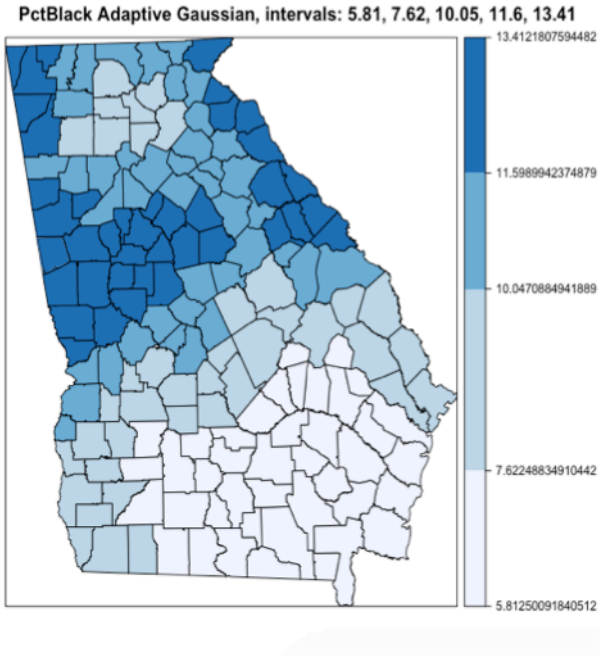
Map 2 e



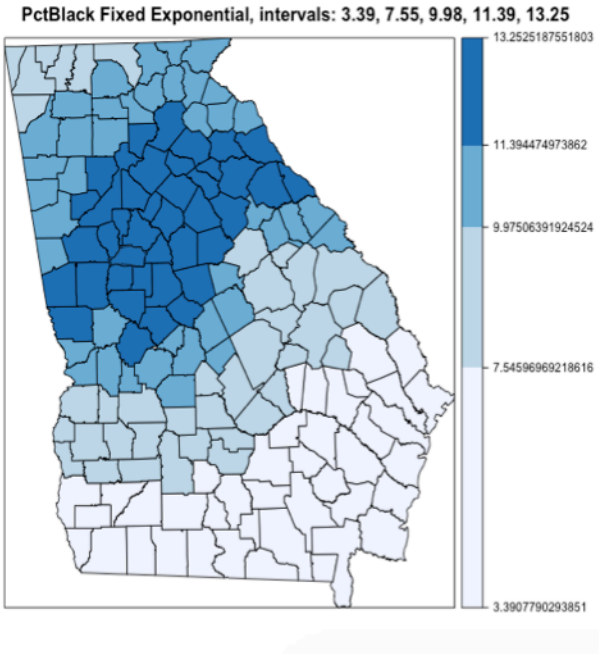
Map 2 f



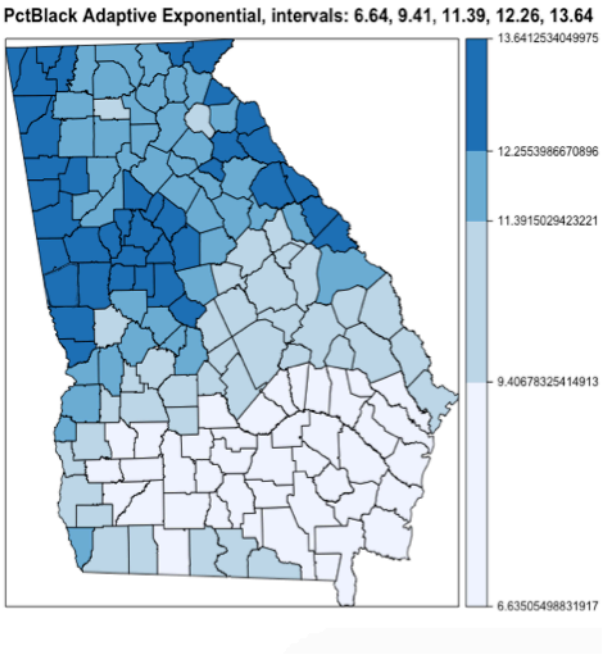
Map 3 a



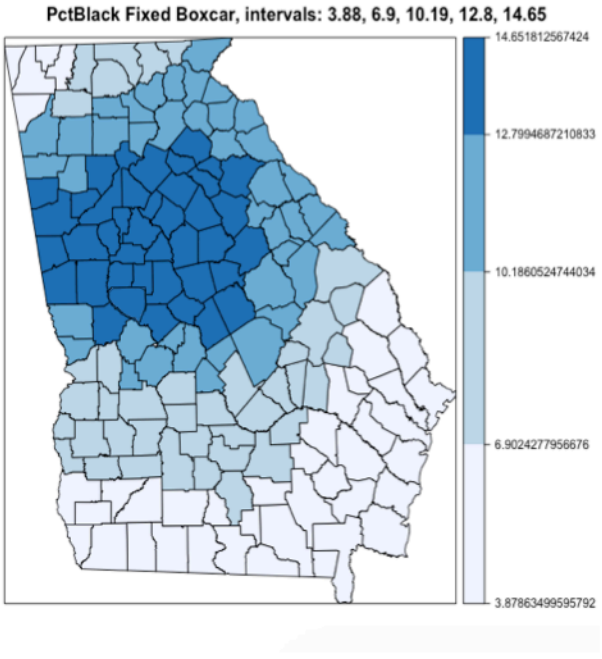
Map 3 b



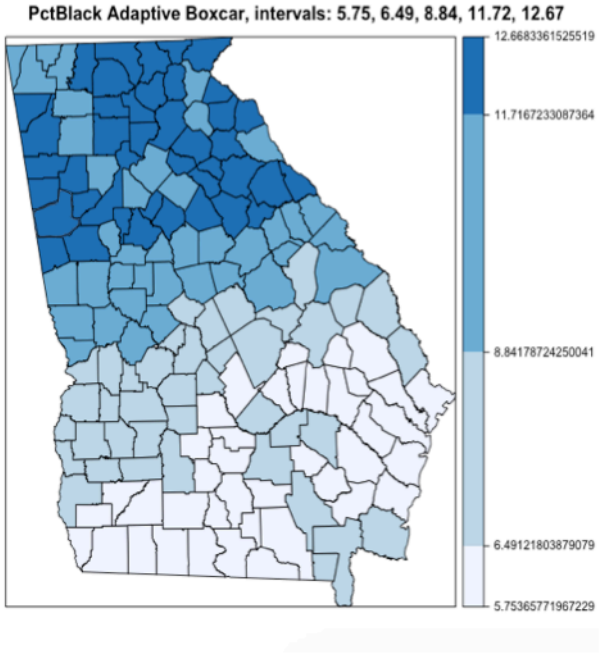
Map 3 c



Map 3 d



Map 3 e



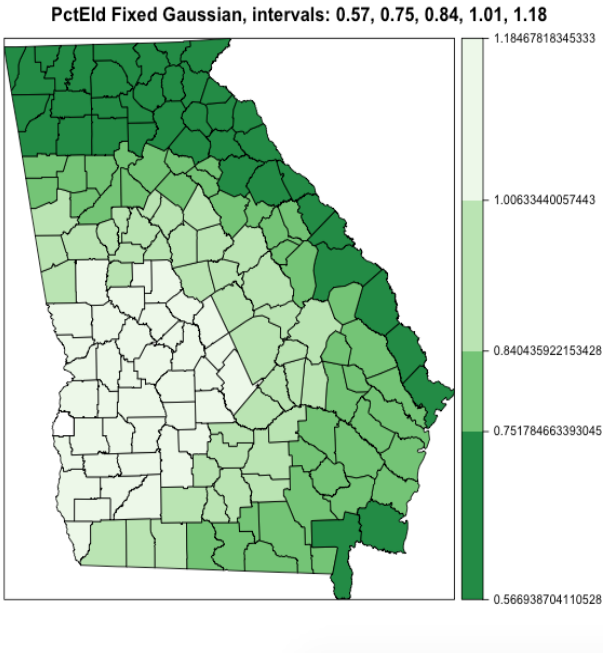
Map 3 f

6.2.PERCENTAGE OF ELDERLY

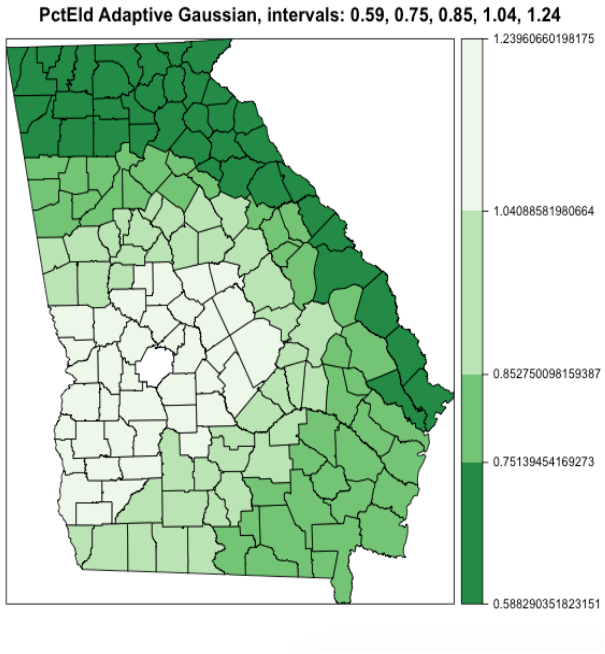
The GWRs computed on this explanatory variable, mostly indicate a stronger correlation in the lower Eastern region of the State with low estimate parameters in the opposite border and the Northern edge.

The values of both Percentage of Elderly and Percentage of Poverty present an evenly distribution across the map. Percentage of Poverty has mainly two weak poles, one near the South West close to Columbus and one centre East corresponding to Madison (map 1a). On the other hand, the variable percentage of elderly (map1c) follows weakly this two poles with a third one in the North at the border with Tennessee. Gaussian and Exponential kernels give credit only to the area in the South West close to Columbus. The maps (maps 4a, 4b, 4c, 4d) in both fixed and adaptive versions consider this area as the main focus of the correlation between the two variables gradually reducing the magnitude of the estimate parameters while moving away from this epicentre. Differently, the Boxcar maps (maps 4e, 4f) show stronger correlations in the Centre East zone. Even though the pole is not fully represented, the estimate coefficients give to that area a stronger correlation especially when using the adaptive kernel. However, looking at the distribution of the coefficients, Boxcar maps assign around half of the observations to the first class and in the case of the Adaptive bandwidth, the small number of observations in the weakest class and overall significantly higher coefficients than the other kernels, represent a very strong correlation that is not representative of the geographical value distribution maps. If the South West areas are those with the highest levels of poverty, the same cannot be said for the percentage of elderly. This is a case of a variable with spatial correlation, percentage of poverty, compared to a variable that is spatial stationary, percentage of elderly. Apparently, the geographical pattern of percentage of poverty takes over the other variable diminishing the validity of the model.

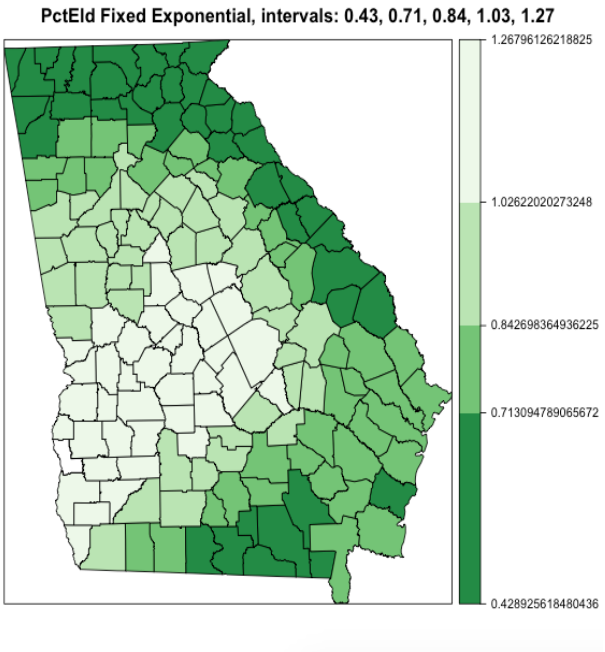
Looking at the significance, similarly to the previous variables, the fixed bandwidth tend to concentrate in the areas with high t-values at the centre of the State. In particular, the small bandwidth of the Exponential kernel returns a pole of high significance only at the centre of the map (map 5c). By allowing the bandwidth to adapt, zones of high significance tend to correspond more to the zones of strong correlation. Turning to the adaptive Boxcar (map 5f), this follows completely the value distribution of Percentage of poverty with a very strong significance over the areas with high poverty rate.



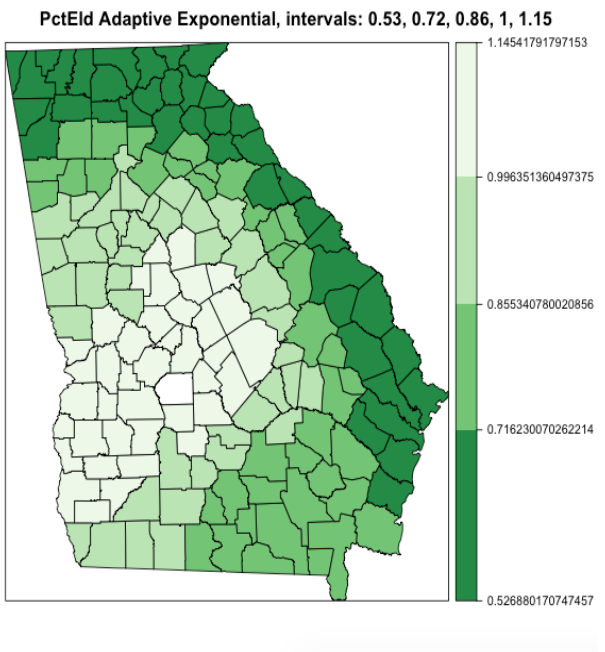
Map 4 a



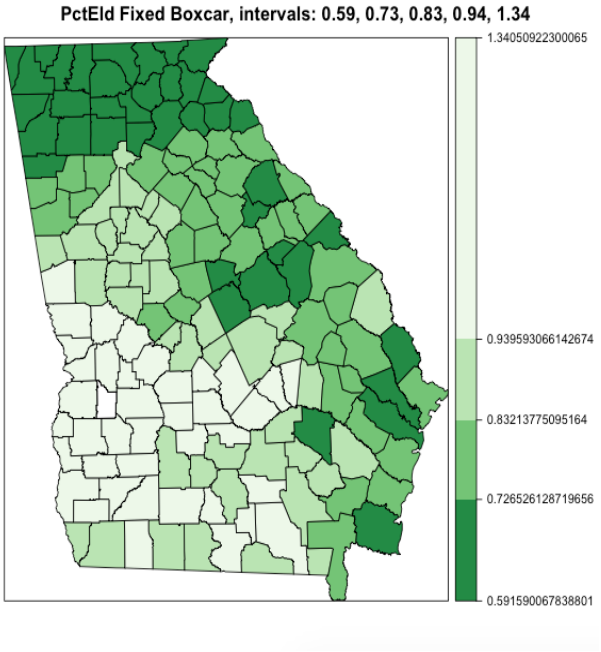
Map 4 b



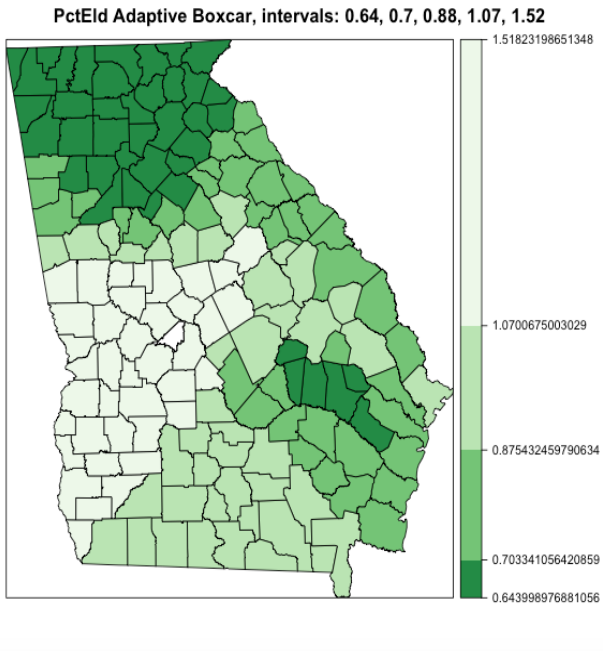
Map 4 c



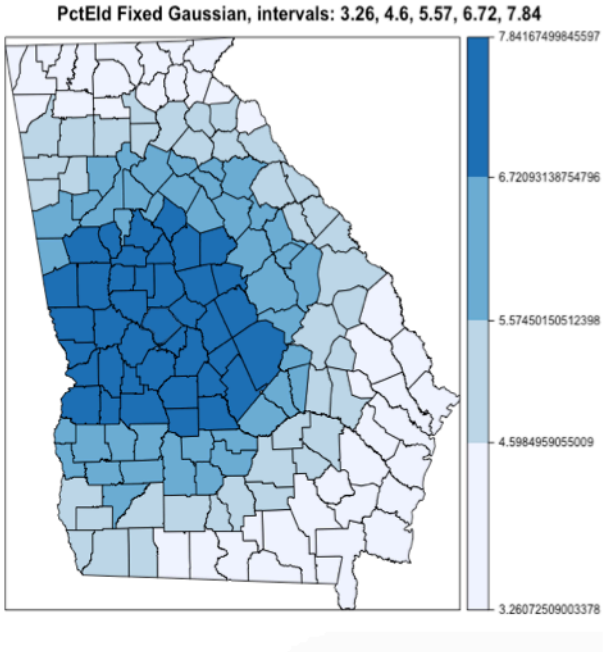
Map 4 d



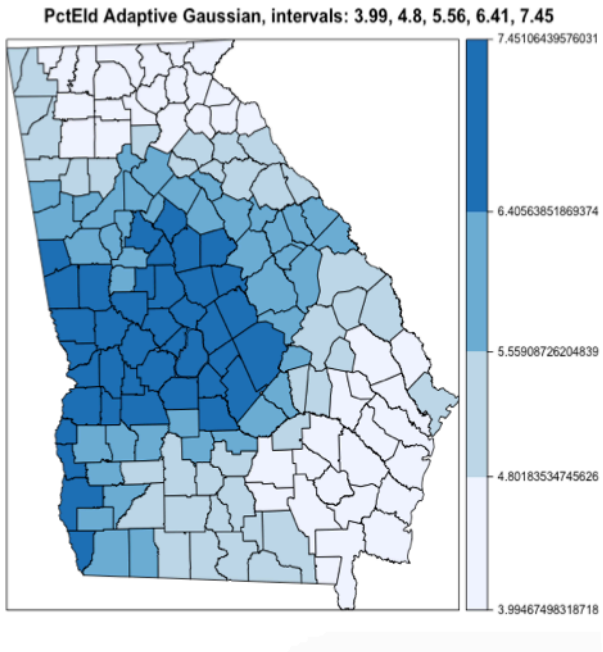
Map 4 e



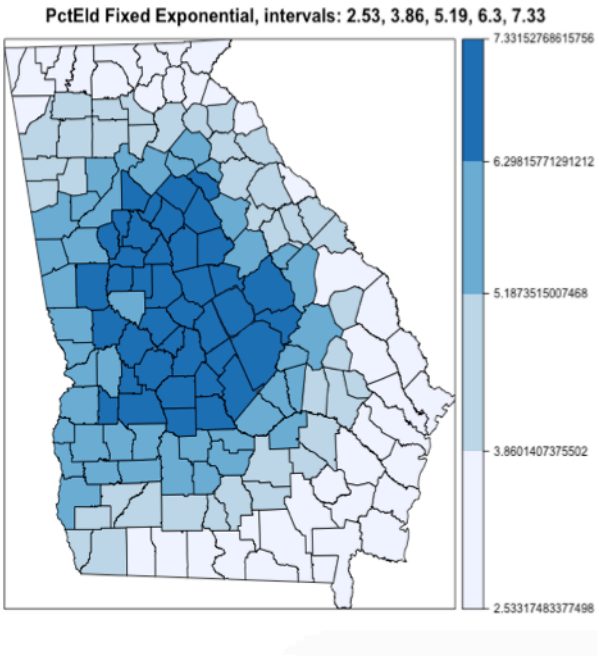
Map 4 f



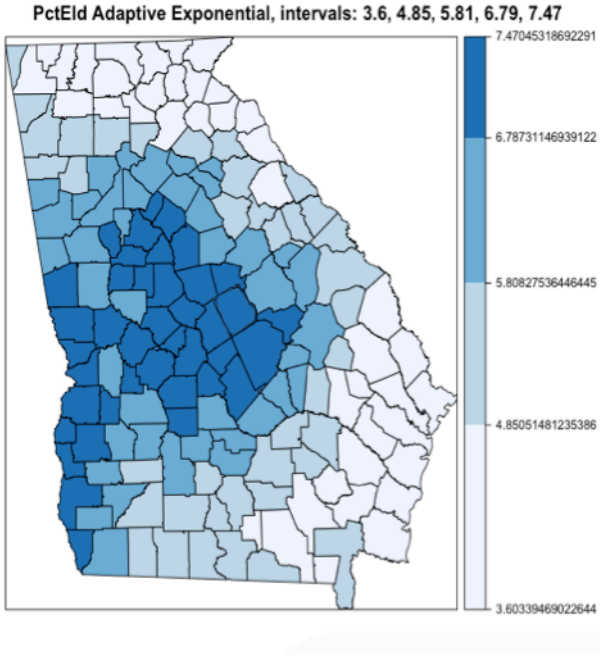
Map 5 a



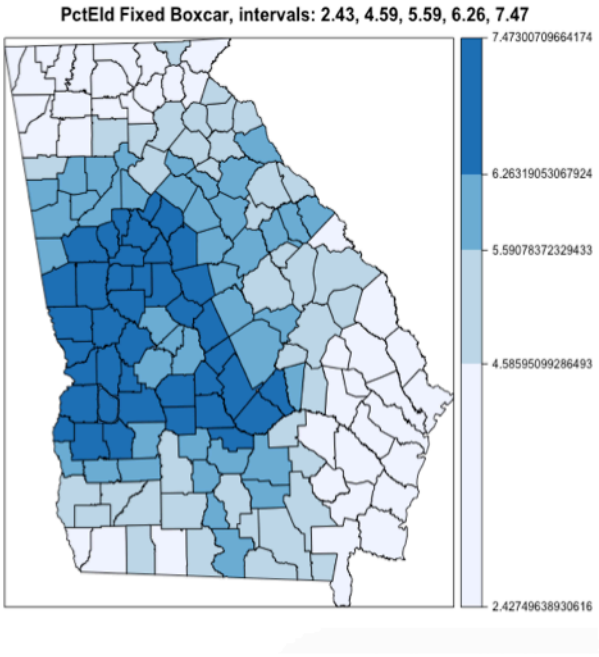
Map 5 b



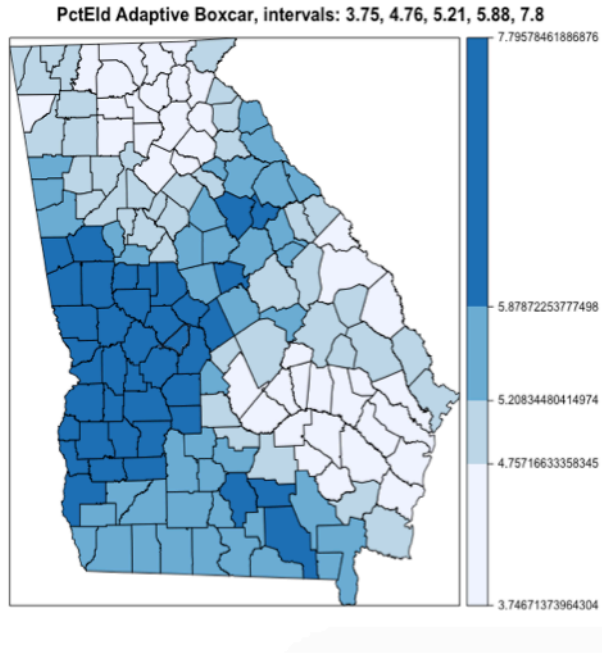
Map 5 c



Map 5 d



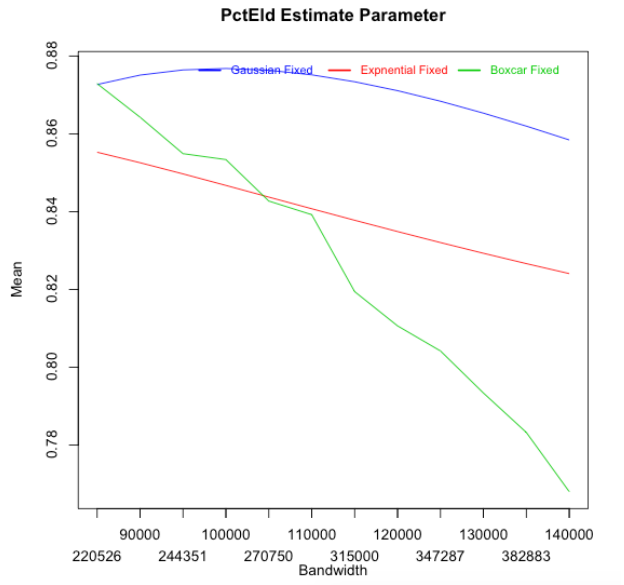
Map 5 e



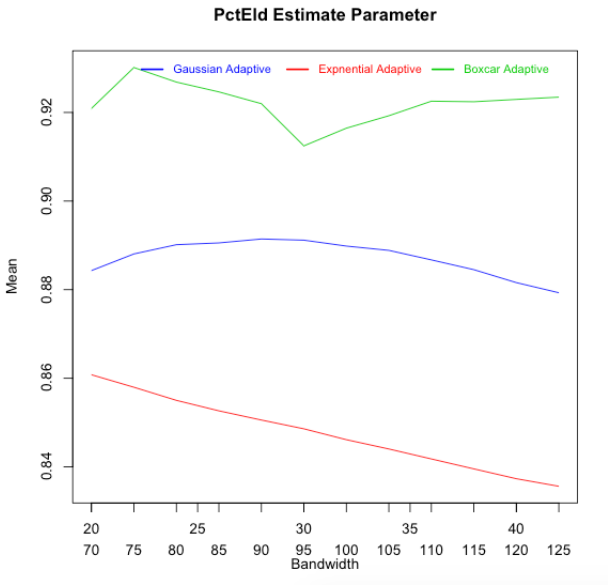
Map 5 f

6.3. GEORGIA BANDWIDTH SIZE VARIATION

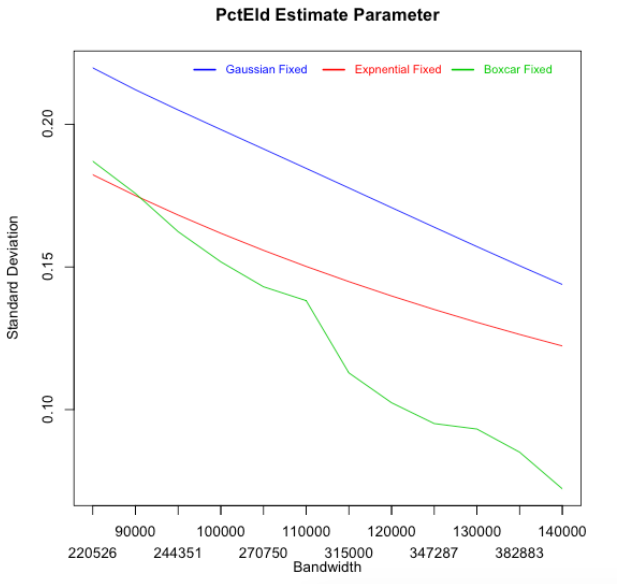
Looking at the bandwidth variation maps, it is evident how the boxcar bandwidth has a less predictable behaviour even when the density pattern is uniform. In fact, a certain degree of consistence is shown by Exponential and Gaussian kernels. Both fixed and adaptive maps for both variables show a slightly higher mean for the Exponential as well as a lower standard deviation as to the estimate parameter. Consistence is shown also by the t-values where a similar trend is shown. However, looking at the standard deviation of the t-values, the difference between Gaussian and Exponential seem to increase over greater levels of bandwidth for the percentage of black people and decrease for the percentage of the elderly. Turning to the Boxcar, its weighting scheme that lets any new observation added to the calibration affect the coefficients in the same measure of the observations already inside the bandwidth, makes the results produced by this kernel much more unpredictable compared to the others. In fact, variations from the optimal bandwidth and switching from adaptive to fixed can have a significant impact. Comparing the mean of the estimate parameters of the two values, Boxcar kernels do not follow a specific pattern returning coefficients that can differ greatly from the other kernels, as in the case of the estimate for percentage of elderly in the fixed map or have very similar trends as in the case of the estimate parameter for percentage of black people. T-values show interesting insights into the nature of the Boxcar kernel. While the rate at which the mean of the t-values grows, is somehow comparable to the other kernels, the rate at which the standard deviation decreases is apparently much faster once it reaches its peak as to suggest that the local coefficients calibrated over great bandwidth values, together with a uniform scheme, tend to be based on similar confidence intervals reducing the variation. The drop in variation seems to be close to the optimal CV bandwidth for this kernel. As for the percentage of black people in the adaptive versions and for the percentage of elderly in the fixed version, this seems to show a line that resembles a normal distribution. The other graphs show only a partial part of this pseudo-normal distributed variation.



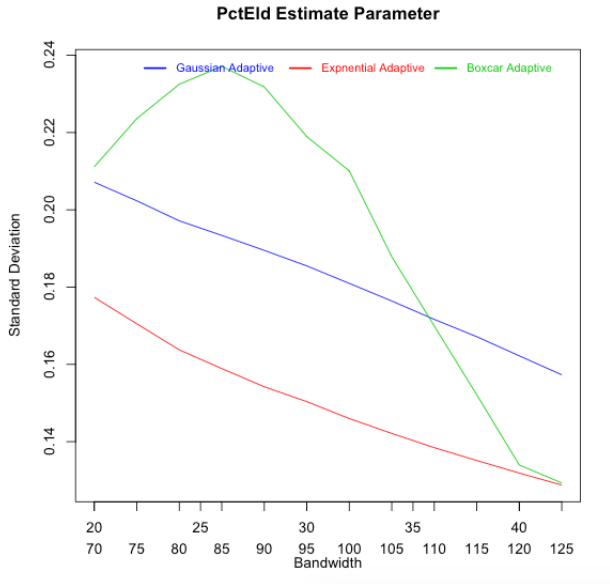
Graph 1 a



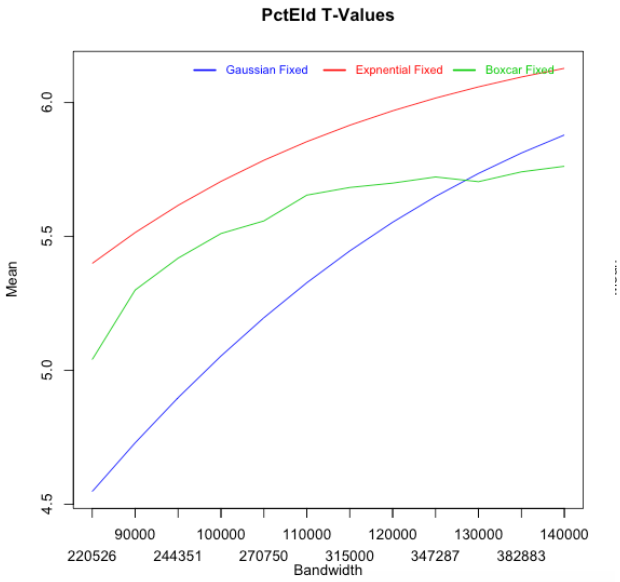
Graph 1 b



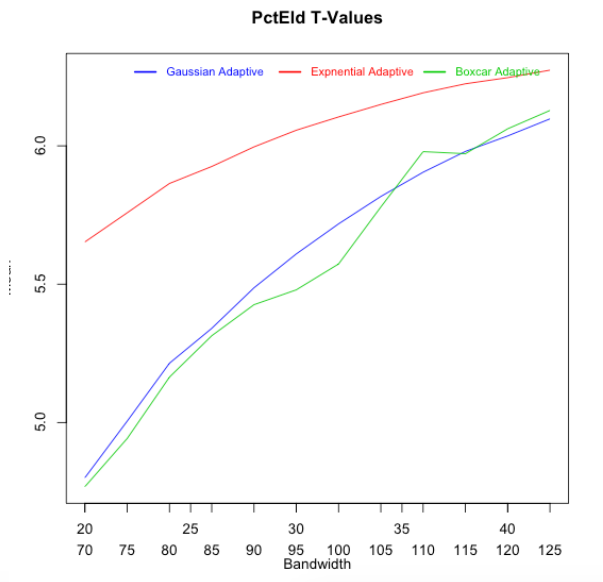
Graph 1 c



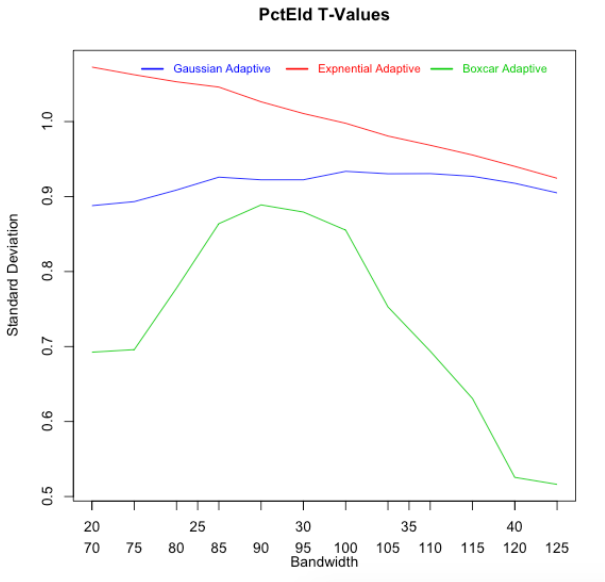
Graph 1 d



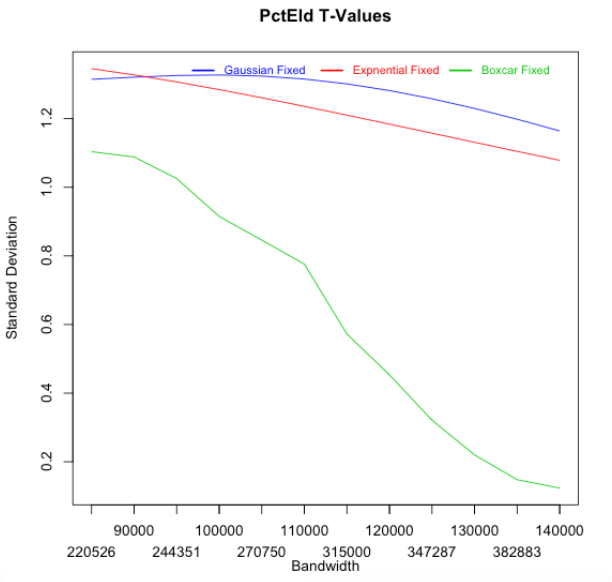
Graph 1 e



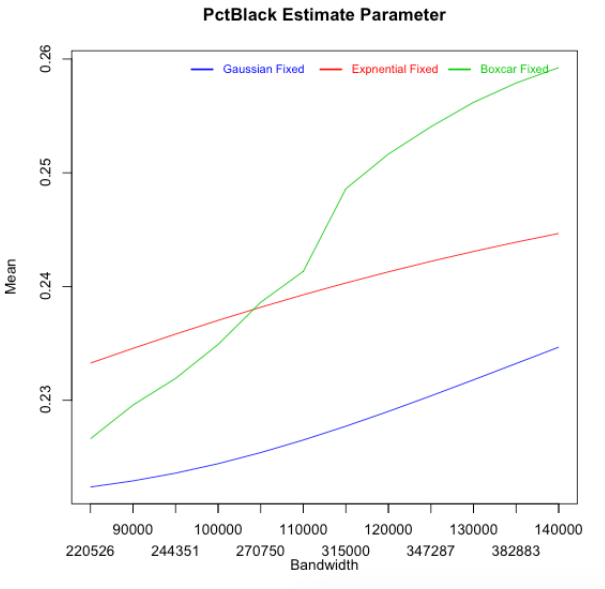
Graph 1 f



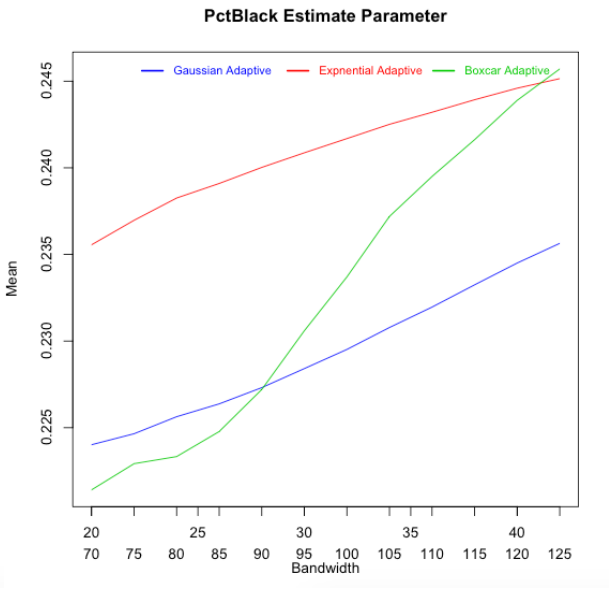
Graph 1 g



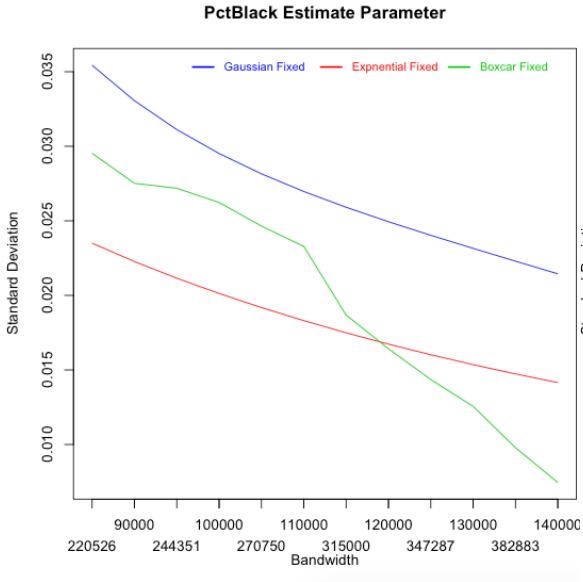
Graph 1 h



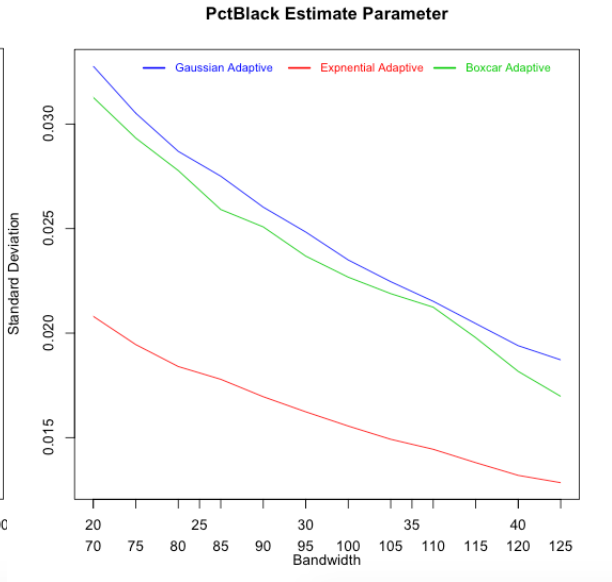
Graph 2 a



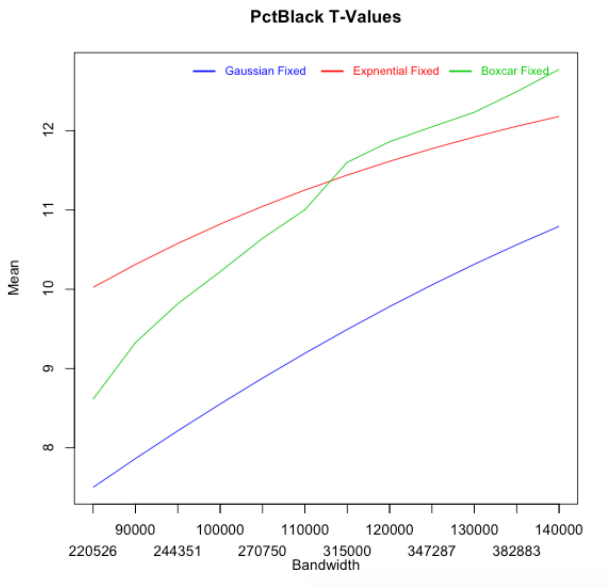
Graph 2 b



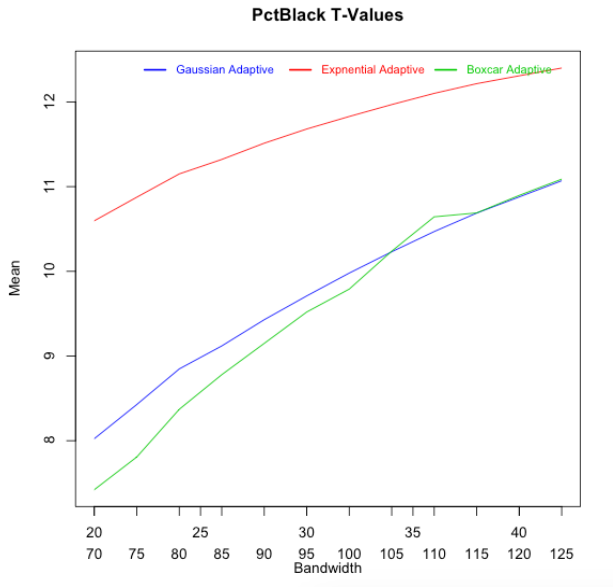
Graph 2 c



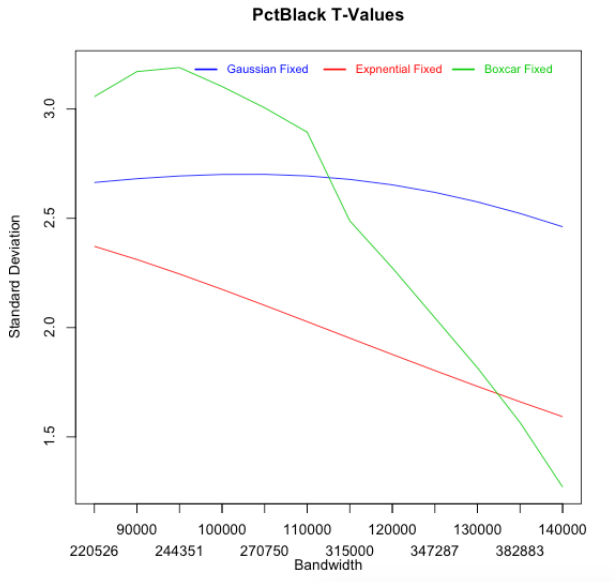
Graph 2 d



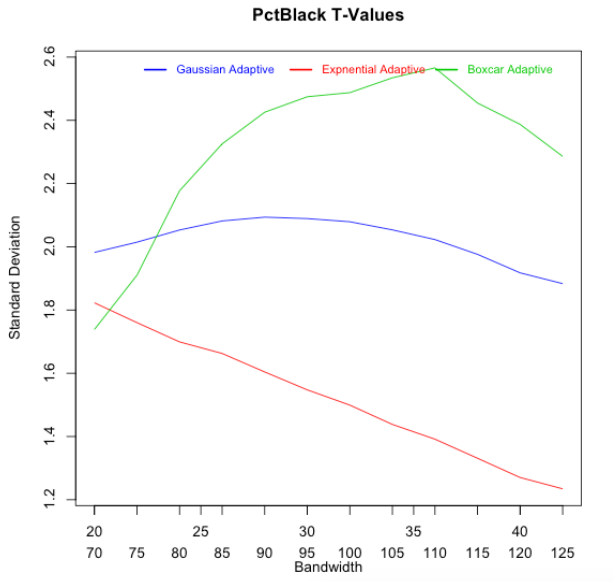
Graph 2 e



Graph 2 f



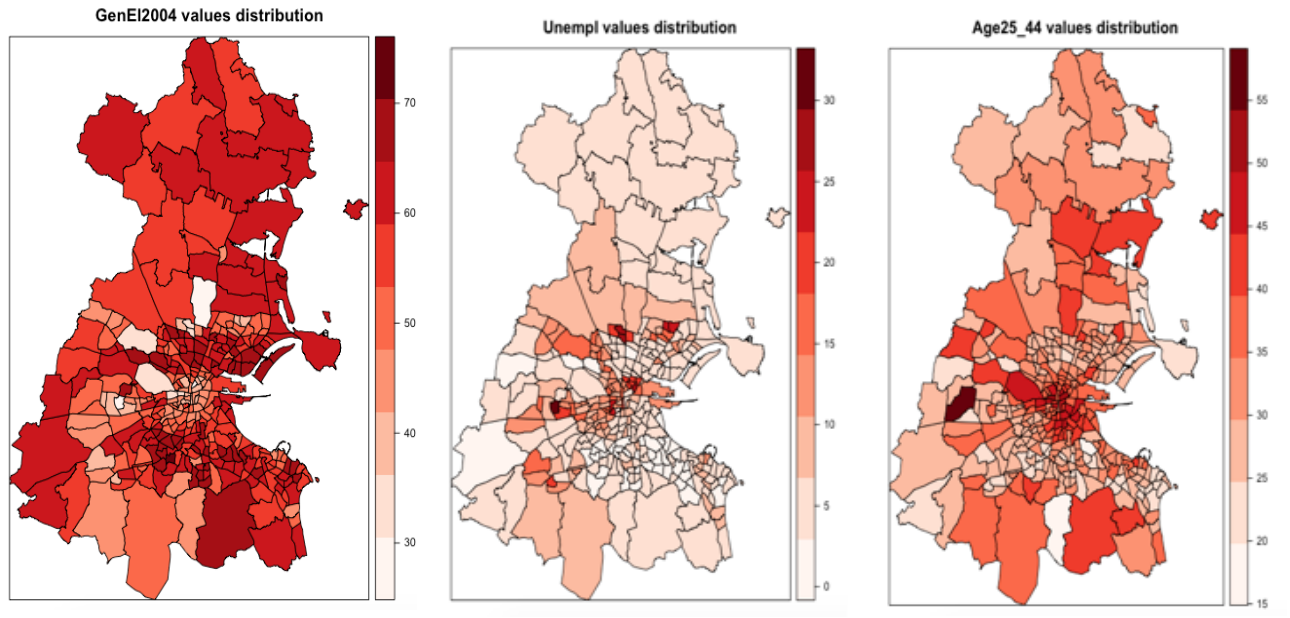
Graph 2 g



Graph 2 h

7. DUBLIN VOTER TURNOUT ANALYSIS

The second dataset used in this study is the Dublin Voter turnout. Conversely to the Georgia census dataset, this dataset has a varying density pattern with a much denser area in the lower centre of the map compared to the rest. By looking at results from the data description section, the followng variables were chosen to be analysed. Percentage of unemployment which was the most substantial variable of the global model with an estimate parameter of -0.7216 and a t-values of -7.687 and percentage of age 25-44, second strongest estimate parameter at -0.3536 with a t-value of -4.747. The variable has a slightly skewed distribution tending to normal.



Map 6 a Map 6 b Map 6 c

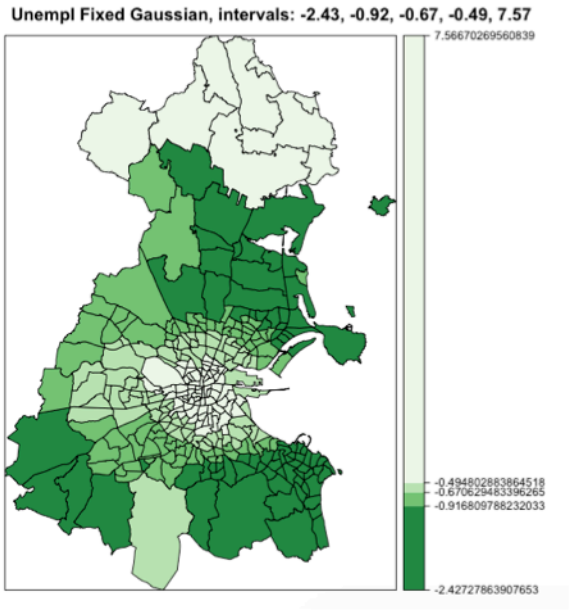
The geographical distribution of the values in these variables (Maps 6a, 6b, 6c) shows how the values of the observations of dependent variable are mostly scattered across the map with a lower percentage of turnout in the centre. Contrary, the percentage of unemployment is higher in some areas of the centre expanding towards North and West. Regarding percentage of age 25-44 , this appears to be high in the centre of the denser area lower around it and high again moving in the outer districts with the exception of the North.

7.1. UNEMPLOYMENT

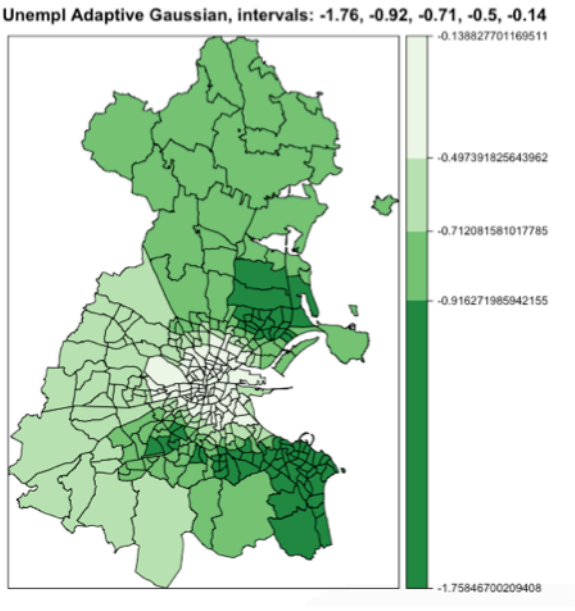
Nearly all results show a higher estimate coefficients in the centre of the Dublin for percentage of unemployed over turnout. However, outside the centre results present quite different representations.

The first striking effect of the bandwidth parameter is the difference in class boundaries in estimate parameters for fixed and adaptive bandwidth. In fact, fixed bandwidth tends to yield class boundaries for the highest class (third quartile - maximum value) much larger than the adaptive counterpart. Estimate parameters are therefore expanded to include several values far away from the means. Whereas, the class boundaries of the maps made through adaptive bandwidth return larger interval values for the lowest class (minimum value - first quartile). Fixed bandwidth maps seem to return values with a greater number of extreme values. Although, this does not apply to Boxcar kernel (map 7e). The large bandwidth given by the cross validation score (30195) gives estimate parameters between -0.76 and -0.51 making it resemble more a global model.

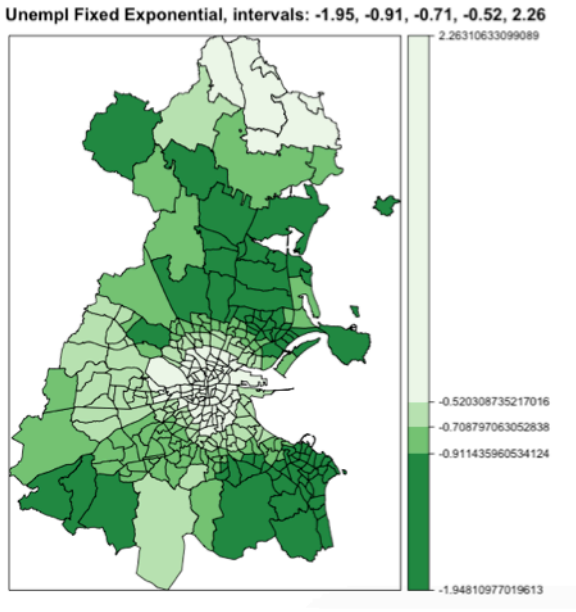
Results from the fixed Gaussian and Exponential bandwidth both capture the low correlation in the centre. In fact, centre areas are associated with low levels of unemployment as well as low levels of turnout. Compared to the global model that describes this relation as negatively correlated, both maps give to estimate parameters in these areas values that can have a positive sign (-0.49/7.57 for Gaussian and -0.52/2.26 for Exponential). Adaptive bandwidth maps with their less extreme coefficients do not confirm the change of sign, although, they maintain low coefficients for this area. Looking at areas of strong correlation instead, the areas South from the Centre and North from the centre and the North end, are represented differently by the adaptive and fixed bandwidth. The North presents low values of unemployment and high values of turnout. This is well depicted by the adaptive bandwidth maps that give to it relatively strong negative correlation. Here the density pattern plays a crucial role. In fact, the North area is much sparser compared to the rest of the map. Turning to the t-values, it is possible to notice how in the fixed bandwidth maps, sparse areas have low significance values. By being calibrated on much fewer observations, most of these areas are statistically insignificant or less significant compared to denser zones. The adaptive maps on the other hand, allocate high level of significance on sparse areas without being affected by the density pattern. Finally, the centre of the city appears to be less significant than other areas. Here the combination of high values and low values in both variables do not grant solid ground for any inference.



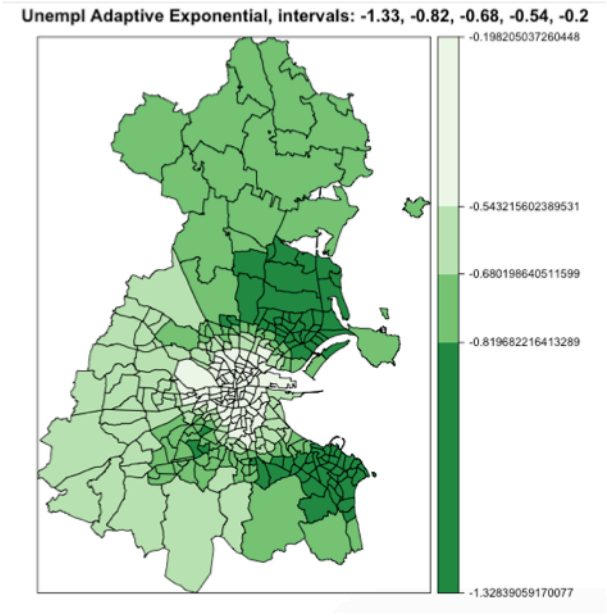
Map 7 a



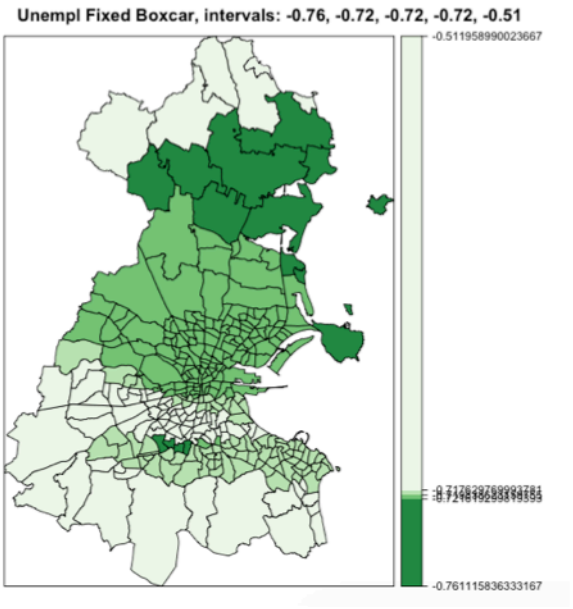
Map 7 b



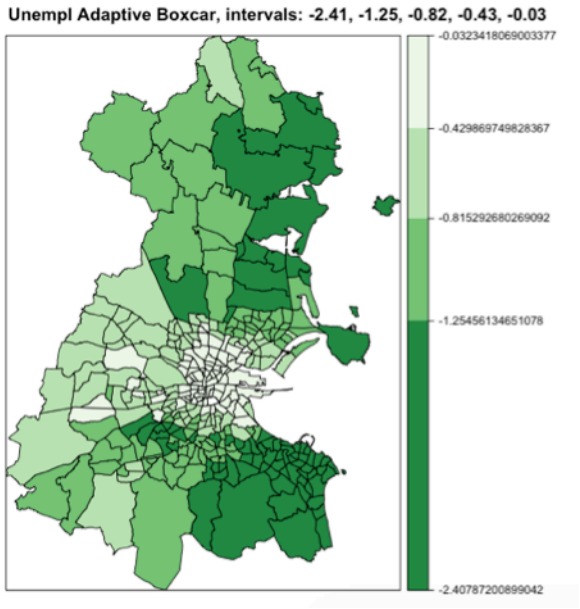
Map 7 c



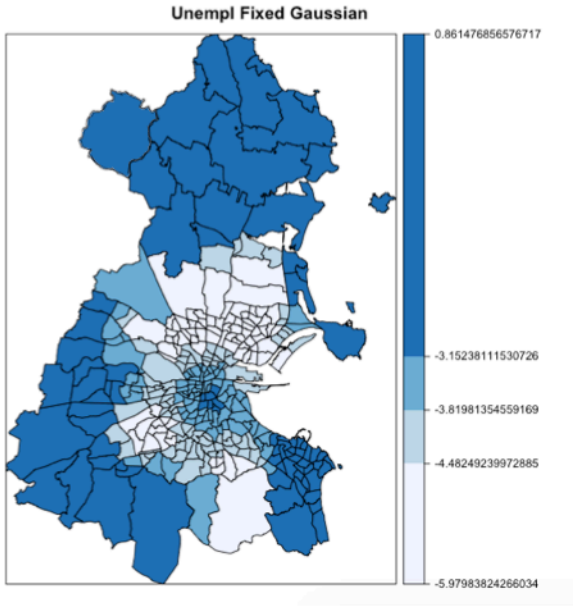
Map 7 d



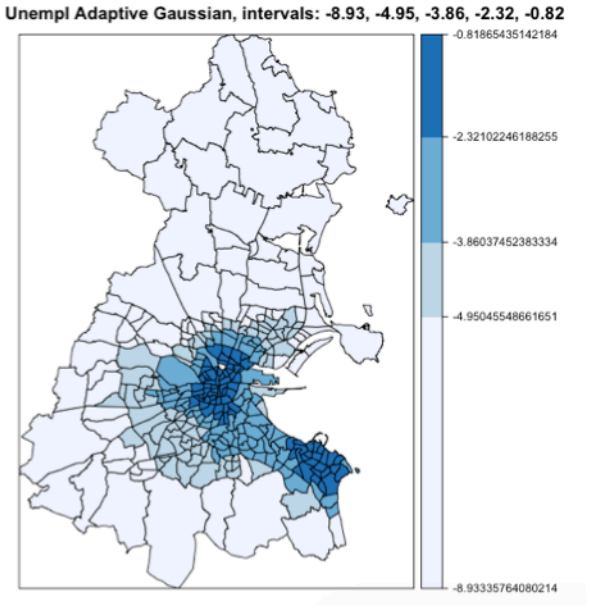
Map 7 e



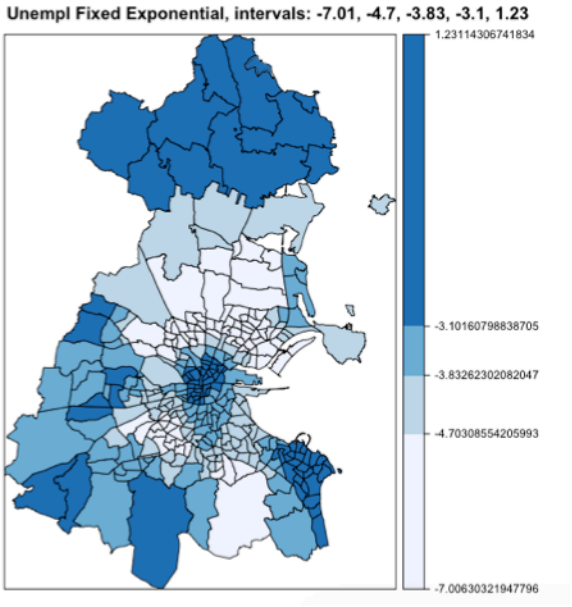
Map 7 f



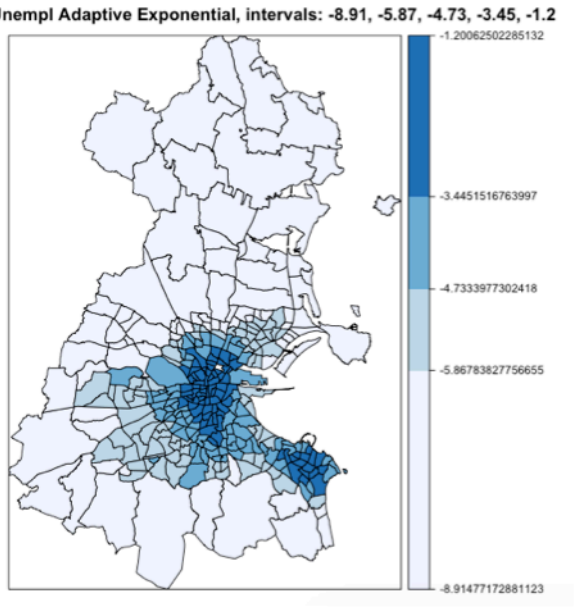
Map 8 a



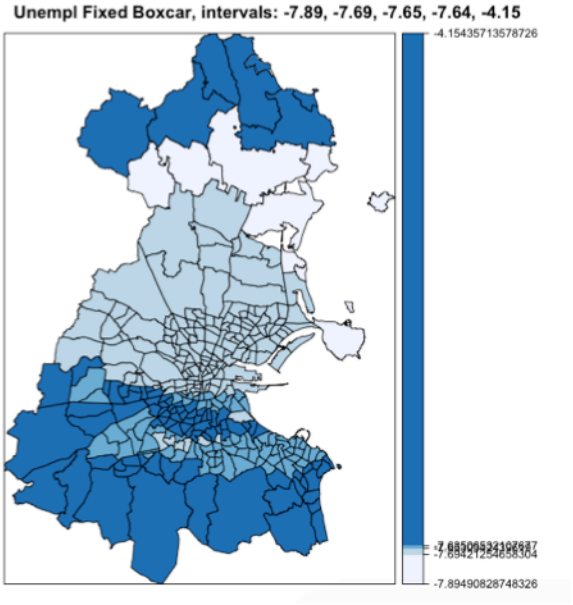
Map 8 b



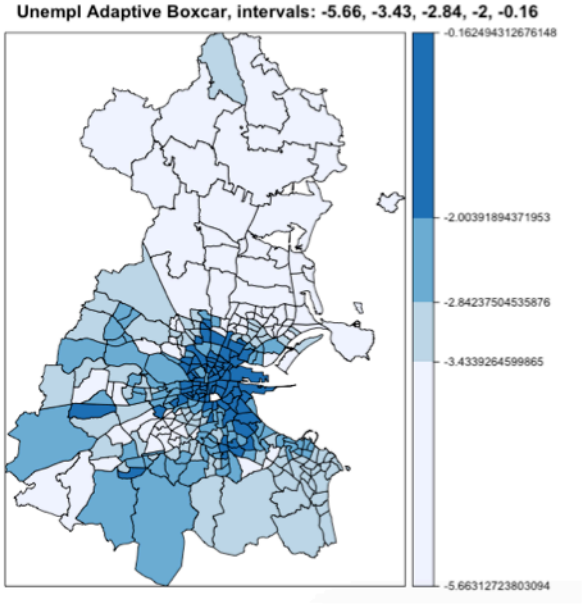
Map 8 c



Map 8 d



Map 8 e



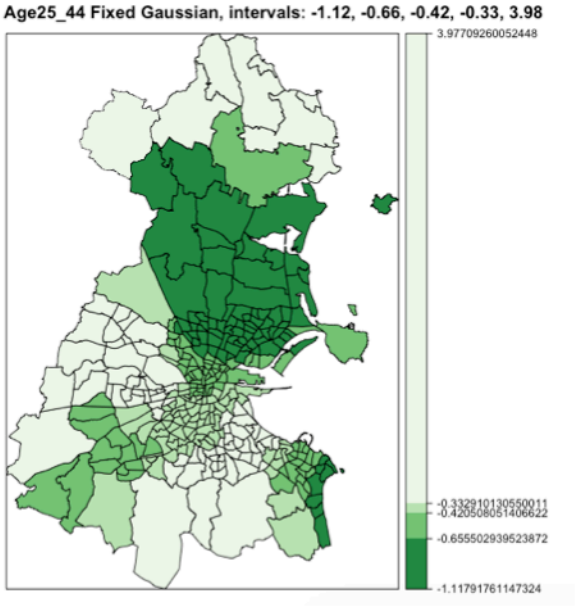
Map 8 f

7.2. Ages 25 - 44

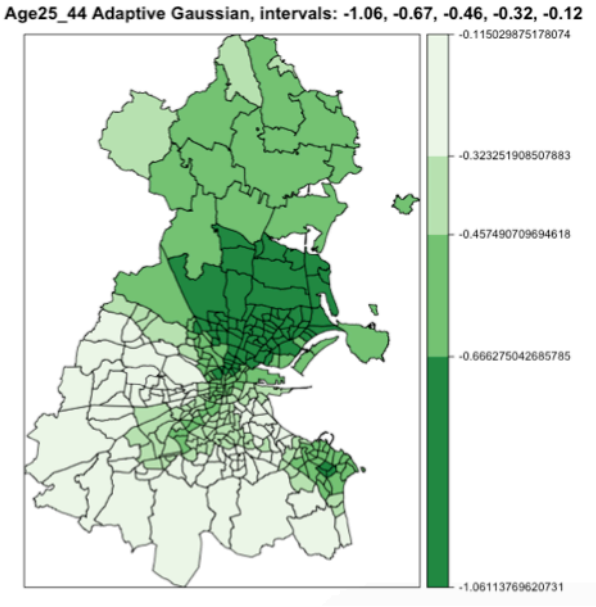
Correlation between turnout and age 25-44 appears to be stronger in the centre North of Dublin with two more poles, one over the south east area and one in the South West area. The value distribution confirms high levels of turnout in areas with a low percentage of age 25-44 supporting the negative correlation.

Similarly to the percentage of unemployment, class boundaries for fixed bandwidth maps have a noticeable degree of extreme values compared to adaptive bandwidth maps. It is interesting to notice how the fixed Boxcar map (map 9e) has a high estimate variance from the third quartile onward instead of the class, than voters within the first quartile. However, the large bandwidth did affect enormously the model giving indications on how Boxcar kernel should not be applied to this kind of density pattern. Moving to the other kernels, the behaviours of Gaussian and Exponential produce mostly the same result with a slighter underestimation of the two secondary poles in the South West and South East by the Exponential kernel in the fixed bandwidth maps (map 9c). On the other hand, results for the adaptive are almost identical. While the difference between kernel function does not play a crucial role, the difference between adaptive and fixed shows a different result in line with the ones of the percentage of unemployment.

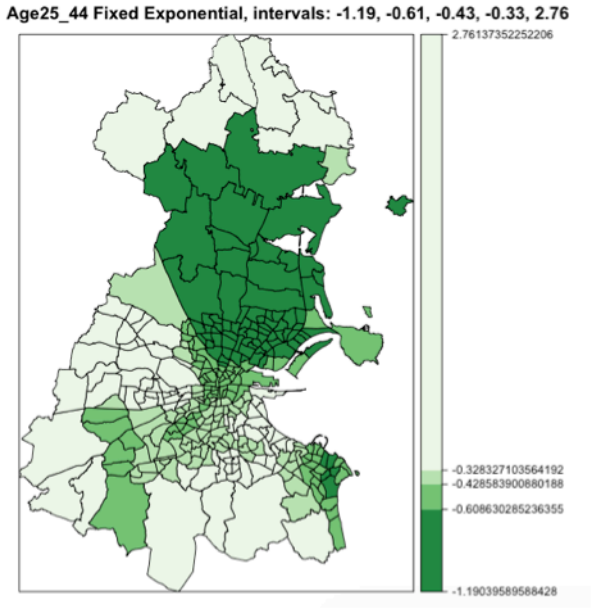
The fixed maps underestimate the coefficients in the North of the map. The value distribution would suggest strong negative correlation between percentage of age25-44 and turnout. However, not only this has low or positive coefficients but hardly reaches statistical significance. Maps made with adaptive bandwidth assign negative correlation to these areas with high levels of significance resulting in more accurate description of the latent relation.Notwithstanding, fixed bandwidth maps are able to show also the correlation of the third pole in the South West area. Despite the lack of support given by the t-values stronger correlations in these areas are well presented in the Gaussian and Exponential fixed maps (maps 10a, 10c). Finally, even though the fixed Boxcar (maps 9e, 10e) did mostly resemble a global model in its fixed version, the adaptive map (map 10e, 10f) presents some advantages compared to the others. All the three poles are successfully depicted as strongly correlated. Despite the overestimation of all the Northern area, the other two poles in the South West and South East are indeed represented and supported by mostly significant t-values. This result is in line with the previous variable, providing outcomes comparable with the other kernels.



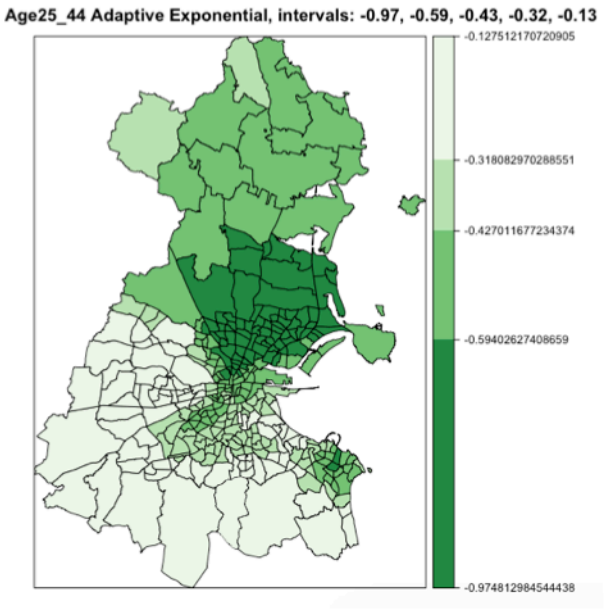
Map 9 a



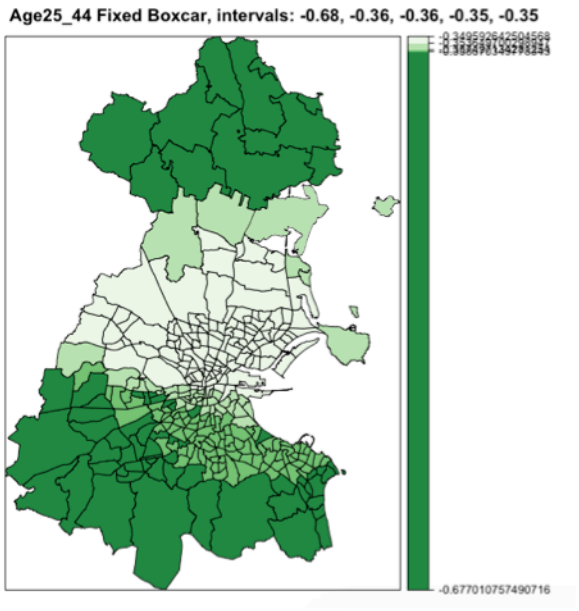
Map 9 b



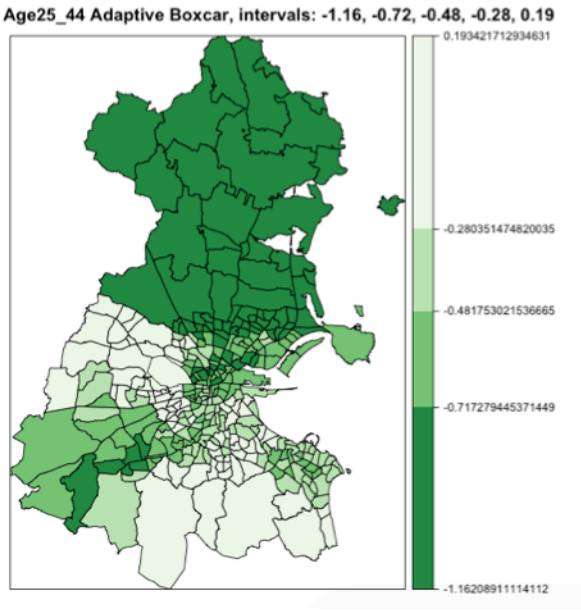
Map 9 c



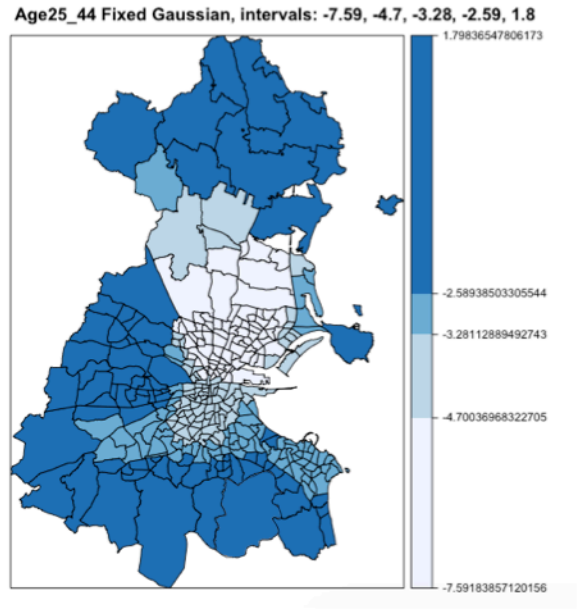
Map 9 d



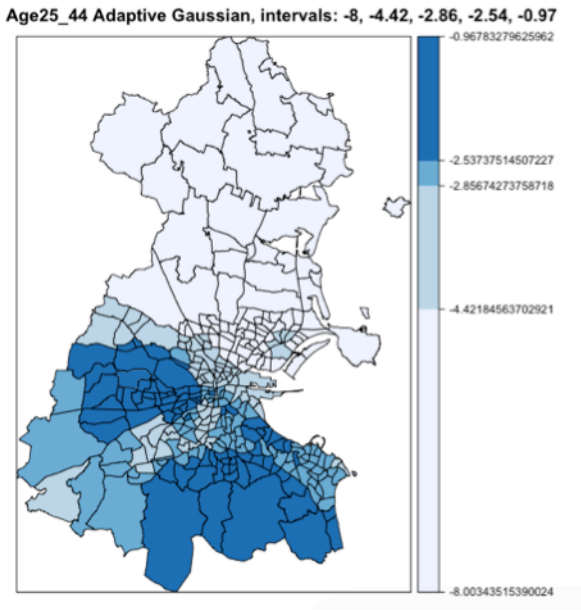
Map 9 e



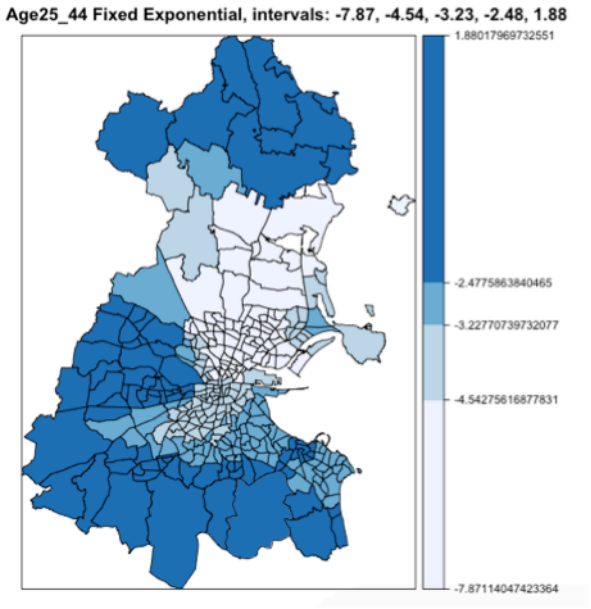
Map 9 f



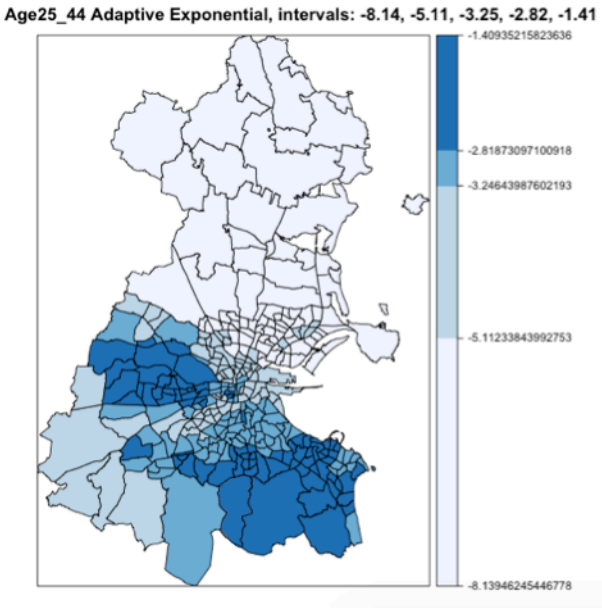
Map 10 a



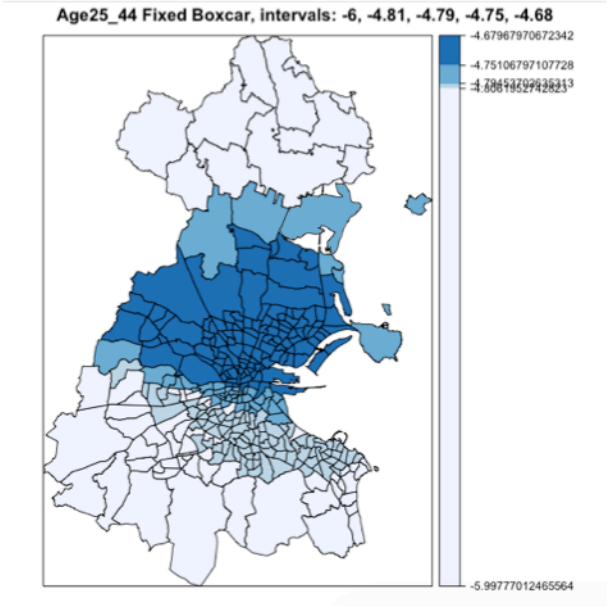
Map 10 b



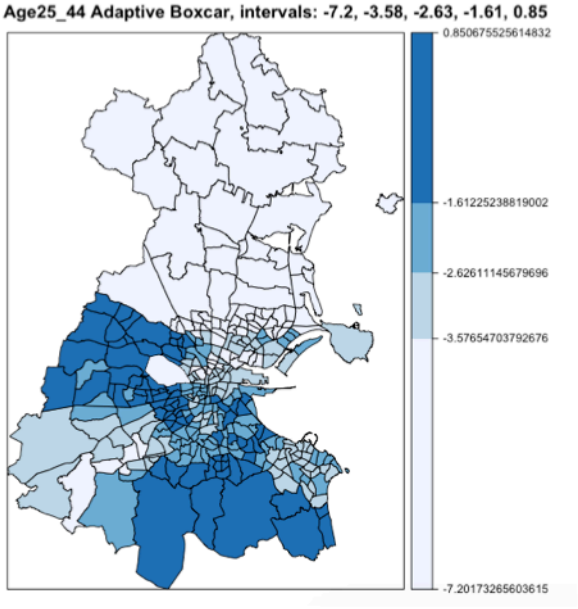
Map 10 c



Map 10 d



Map 10 e

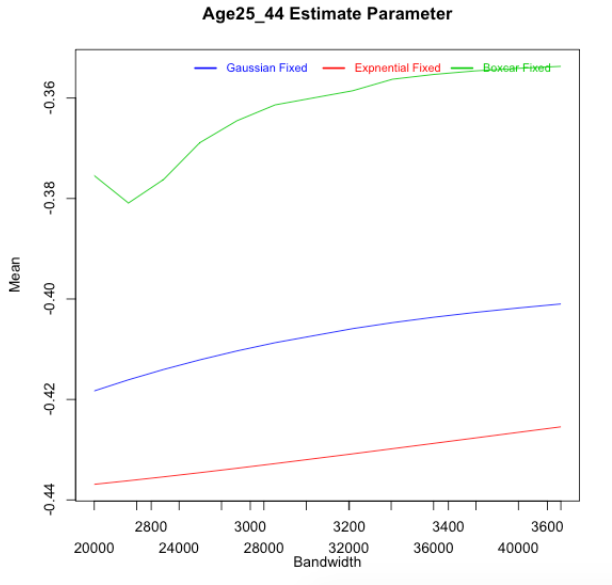


Map 10 f

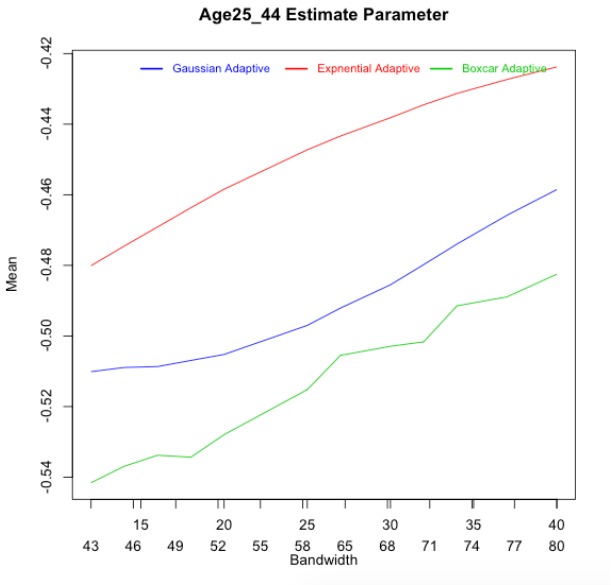
7.3.DUBLIN VOTER TURNOUT BANDWIDTH VARIATION

Similarly to the Uniform density pattern data, Estimate parameters for both Gaussian and Exponential kernels move towards lower values when the bandwidth increases. However, variations in these coefficients are lower for the Exponential kernel. Comparing fixed bandwidth to adaptive, the standard deviation of the mean seem to decrease significantly faster on the adaptive maps than in the fixed maps. Coefficients based on small bandwidth seem to be affected by a greater variability when these are adaptive rather than fixed. Particularly true for the Exponential kernel that shows very little signs of variation in both the percentage of aged 25-44 and the percentage of unemployment when fixed. Moving to the t-values, the difference between fixed and adaptive bandwidth is even stronger. T-values in the adaptive maps are significantly lower when based on few observations. Notwithstanding, this can reach greater significance than the fixed counterparts, when new observations are added to the coefficient computation. Although, despite beginning with very high variation in t-values and decreasing quickly over time, these seem to be higher, at least for the percentage of unemployment than those expressed by the fixed maps. This may be similar to the cases underlining a trade-off between a smooth t-values map and high significance levels t-values map.

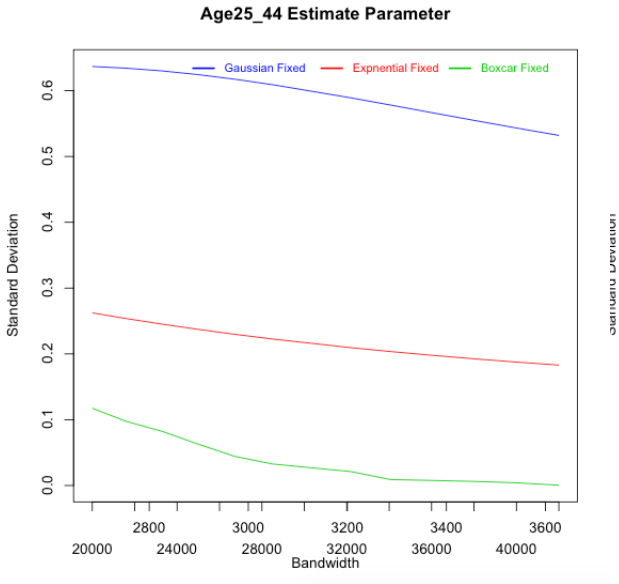
Also for this density pattern the Boxcar kernel seems to have a hardly predictable outcome. The normal shaped t-value standard deviation trend that appears in the uniform density pattern of the Georgia data seems to have disappeared. However, another interesting pattern can be observed. In fact, the optimal CV bandwidth value of the fixed bandwidth gives already very little variation to the coefficients due to its much larger value compare to Gaussian and Exponential. On the other hand, the value of the adaptive bandwidth has a much smaller difference compared to those of the other kernels. This gives to the Boxcar results, in the adaptive maps, much greater variation, thus, distinguishing it from a global model. This comes at a cost of much lower significance in the t-values for both variables.



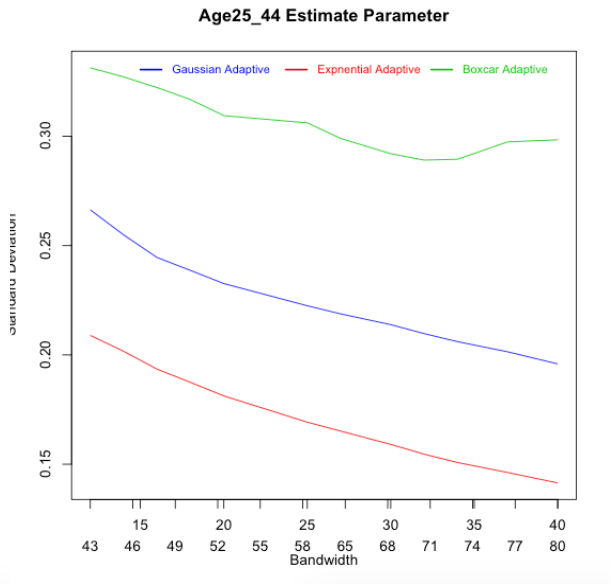
Graph 3 a



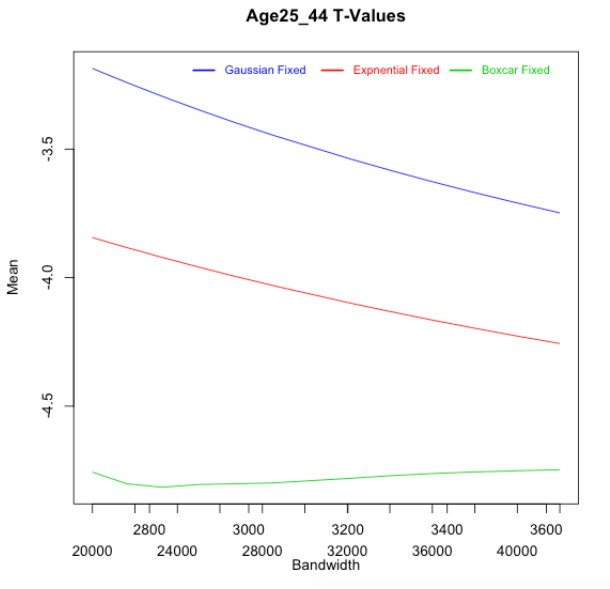
Graph 3 b



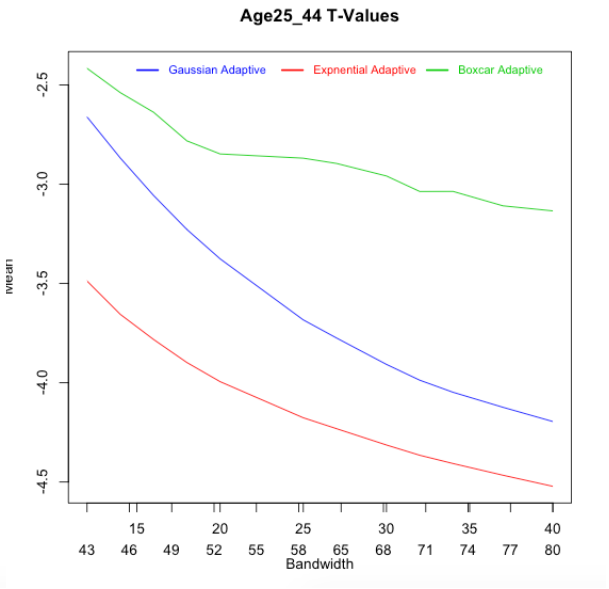
Graph 3 c



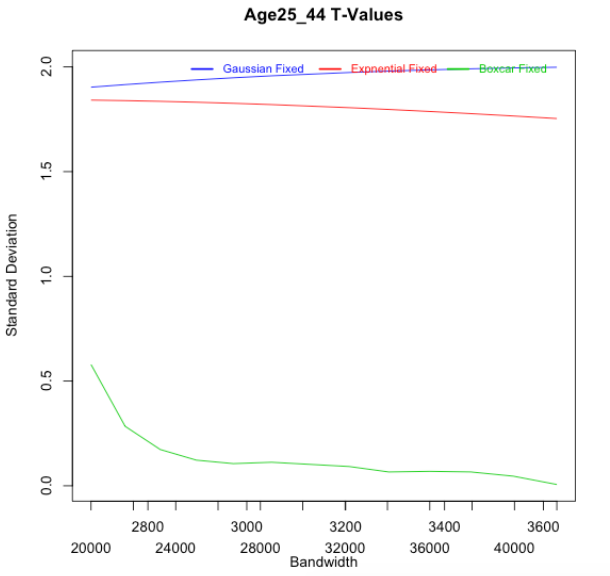
Graph 3 d



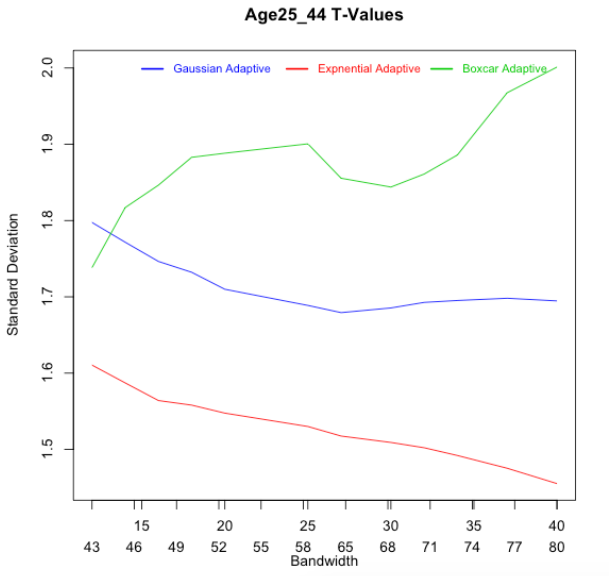
Graph 3 e



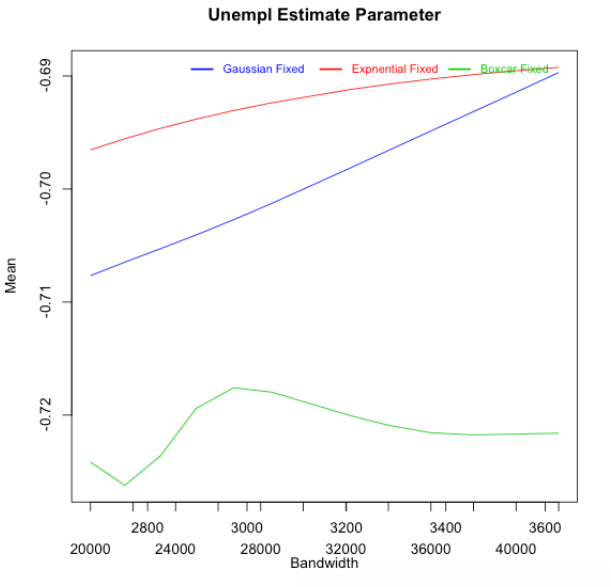
Graph 3 f



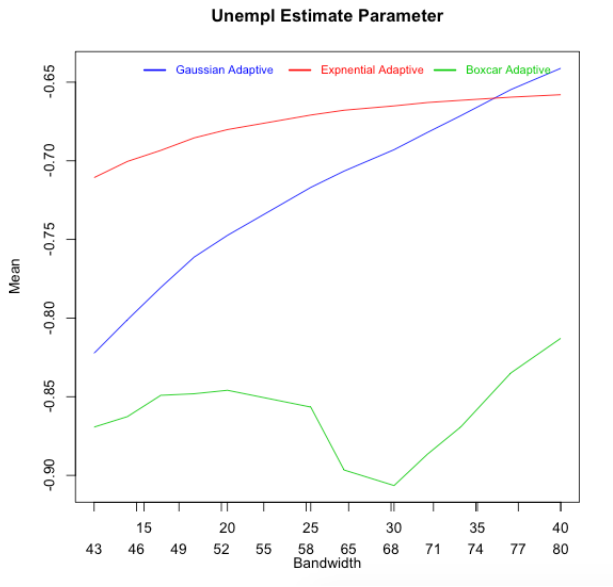
Graph 3 g



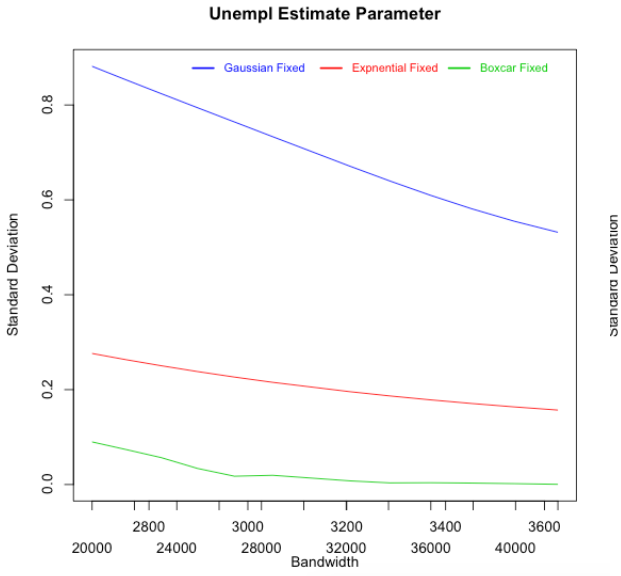
Graph 3 h



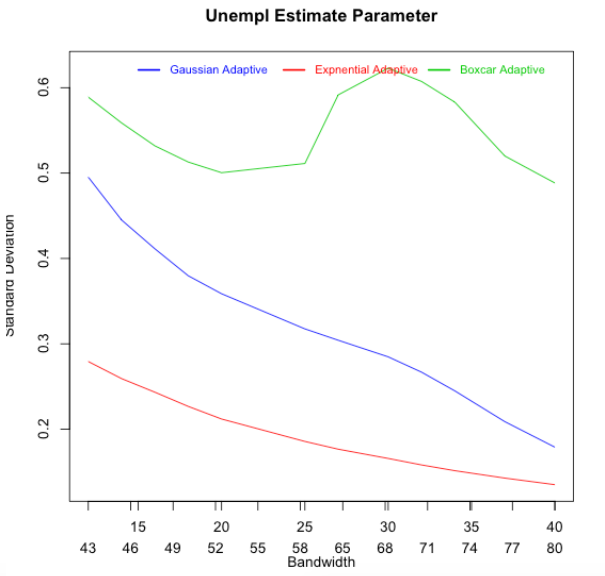
Graph 4 a



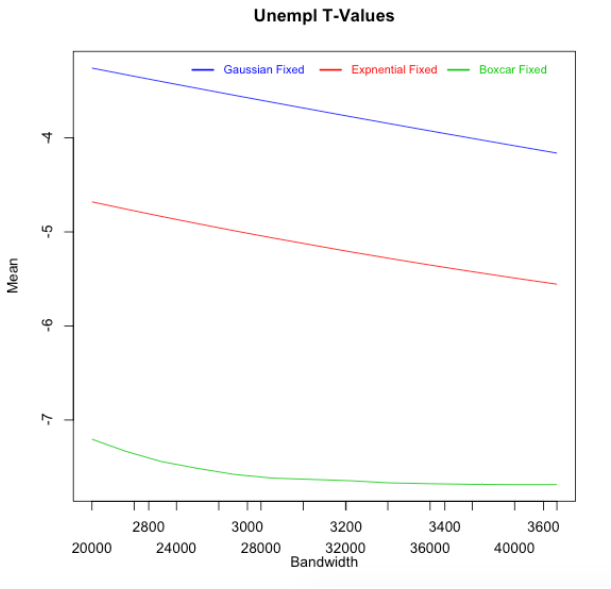
Graph 4 b



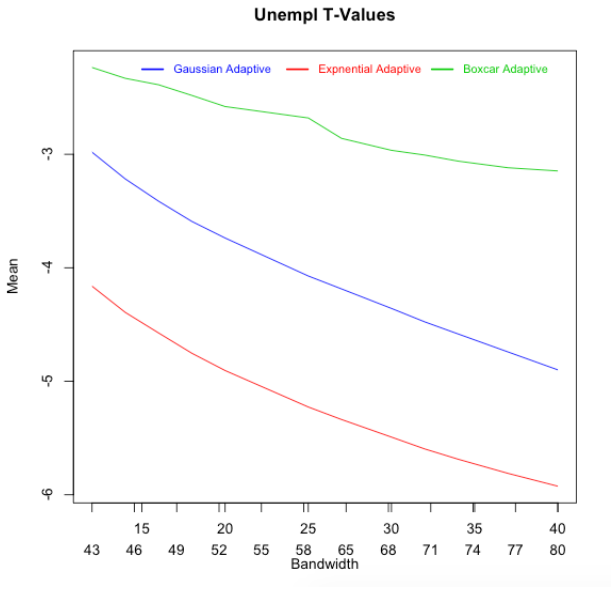
Graph 4 c



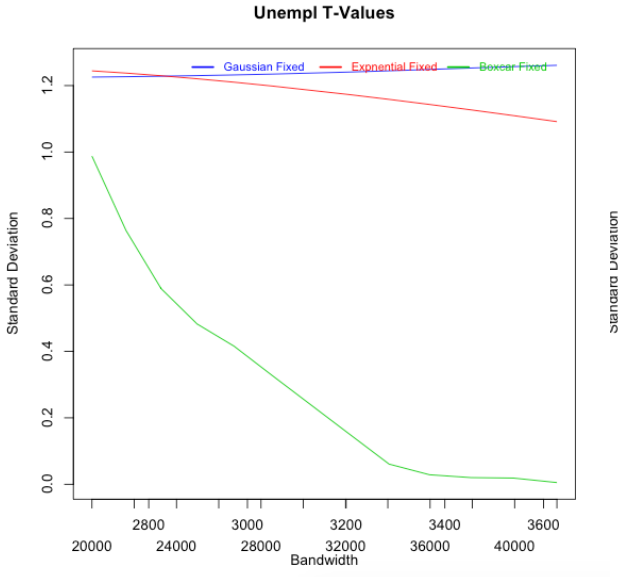
Graph 4 d



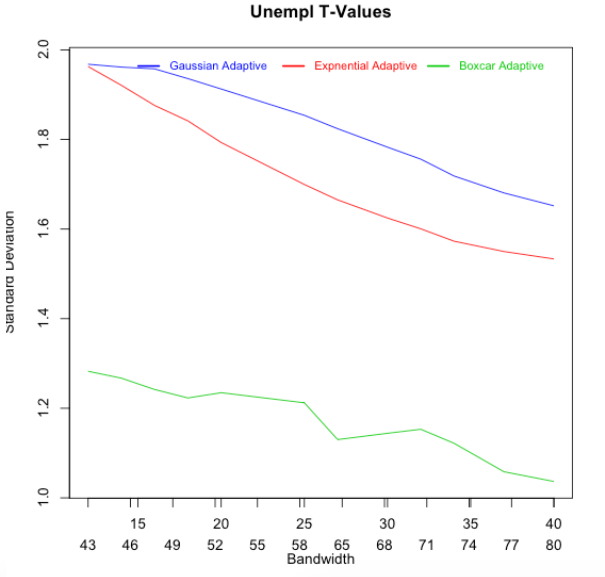
Graph 4 e



Graph 4 f



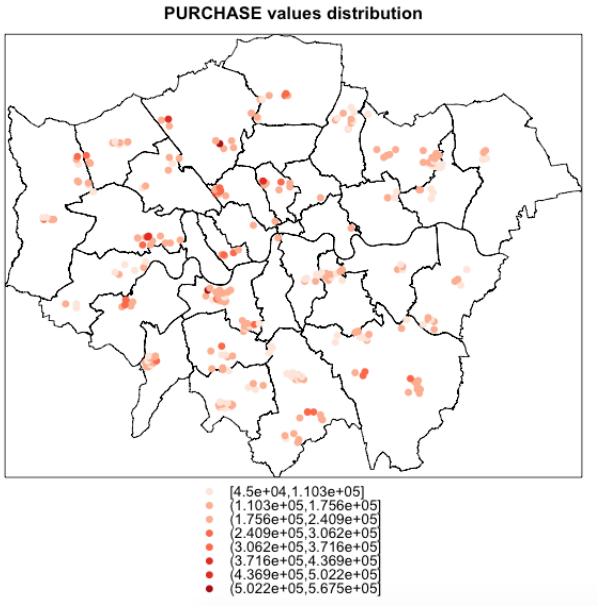
Grpah 4 g



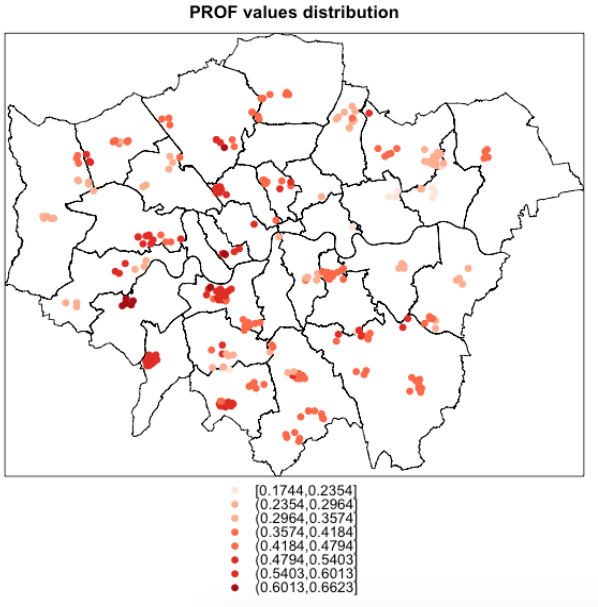
Graph 4 h

8. LONDON HOUSES PRICES

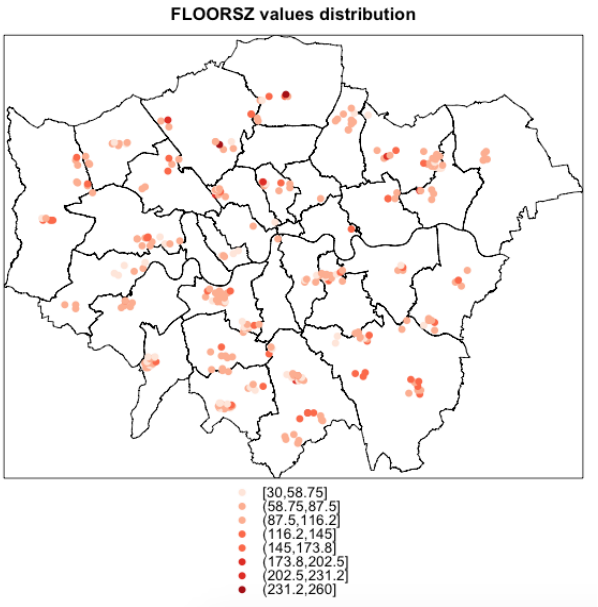
The last dataset chosen is the London house prices dataset. Contrary to the previous two sets of data this is a spatial point data frame. As reported in the descriptive section, the density pattern presents a multitude of several little clusters. The variables chosen to investigate the bandwidth parameters over these geographical patterns are percentages of professionals and floor size. The former is the variable with the most significant t-value (11.715) and a very strong estimate parameter of 3593e+05. The latter is the variable with the second most significant t-value (11.385) with an estimate parameter of 1179.08.



Map 11 a



Map 11 b



Map 11 c

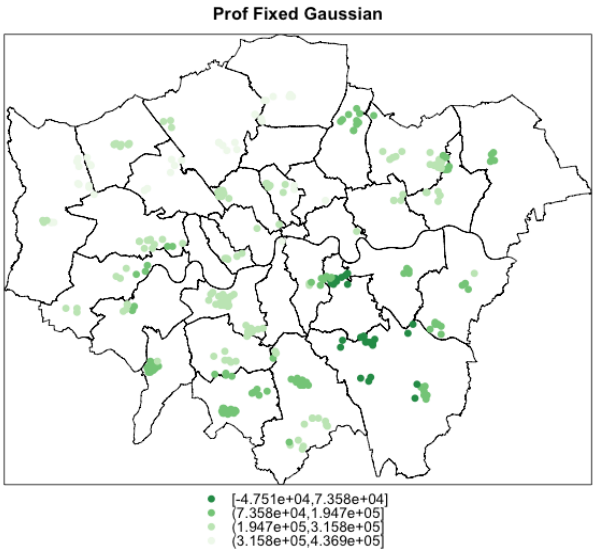
The geographical value distribution of the data shows how purchase price is at the lowest in the East part of London with the Centre North West area as the most expensive. The percentage of professionals seems to be concentrated in Middle West area with a cluster in the South over the areas corresponding to Sutton and Kingston upon Thames. Regarding floor size, this tends to be more scattered with relatively larger values in the North in boroughs such as Barnet and Enfield.

8.1. PROFESSIONALS

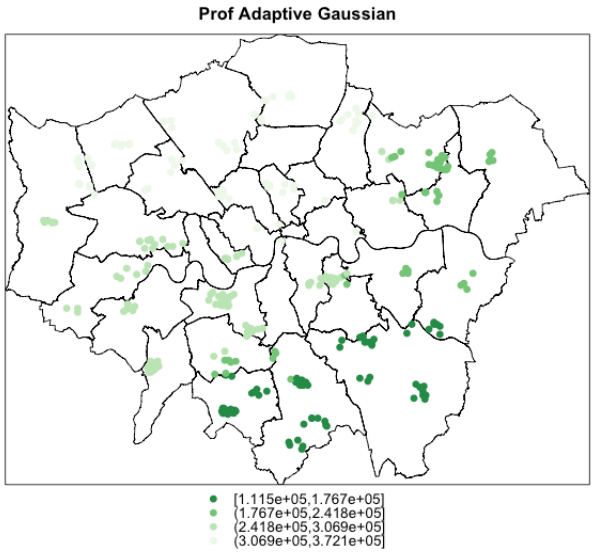
Maps for the explanatory variable percentage of professionals indicate that this positively affects the house prices mainly in the North West part of London. As shown in the geographical value distribution maps (maps 11a, 11b, 11c) East London tends to have a lower percentage of professionals compared to the West, while house prices are generally higher in the North East.

GWR results over this density pattern are highly influenced by the choice between adaptive and fixed bandwidth. Looking at the two Gaussian maps (maps 12a, 12b), while the North West shows strong correlation in both maps, the situation of the South is quite different. Clusters in Sutton, Croydon and part of Bromley were assigned high estimate coefficients by the fixed gaussian, while the adaptive version returned low estimates. The values distribution maps reported that the area of Sutton has a high number of professionals and low prices. Similar trends even though less marked, applied to Kingston upon Thames and Croydon. Although, Kinston upon Thames is given strong positive coefficient in both maps, the latent pattern seems to be described better by the fixed version. Despite the fact that the model expresses slightly stronger estimate parameters for these boroughs, the t-values of the fixed Gaussian (maps 13a, 13b) give low, no significance to these areas. Whereas, the adaptive version allocates a minimum t-value of 2.101 which is above the requirement for 95% confidence interval (1.96). The lack of statistical significance for such relation is arguably an advantage for the fixed gaussian that avoids taking into account an apparently misleading result. Notwithstanding, all maps generated by fixed bandwidth show high levels of significance only towards the centre of the map. Contrary to the adaptive bandwidth, the number of observations used to calibrate and compute the estimate parameters and the t-values varies enormously for isolate observations when a fixed radius is applied. For instance, the area of Bromley presents quite sparse clusters that in all fixed bandwidth maps are below the 90% confidence interval standard, while the upper boundary of the lowest class in all the adaptive maps is always above the 99% confidence. This is due to the nature of the adaptive bandwidth that ignores the geographical distance and assigns to the calibration of each regression point the same number of neighbours reducing the chance of low significance areas around the borders.

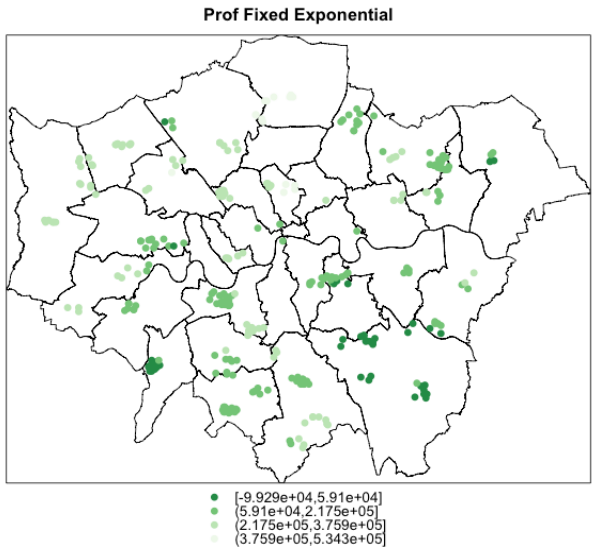
While the Boxcar kernel has a similar outcome compared to the Gaussian, the Exponential kernel due to its steeper shape and smaller bandwidth, behaved differently. Observing the estimate coefficients in the borough of Kingston upon Thames in the fixed map (map 12e) it is possible to notice that these are very low, as supported by the values distributions. The isolate geographical positions of those observations together with the small bandwidth given by the Exponential kernel, lets this area show its true relation to the dependent variable, allocating also negative coefficients to the area. Also the t-values do not support the relation (map 13e). The Exponential kernel in its fixed version appears to respect also the relations giving high significance and estimate coefficient to the Middle North West parts of London and no significance to the isolated parts in the South Eastern areas.



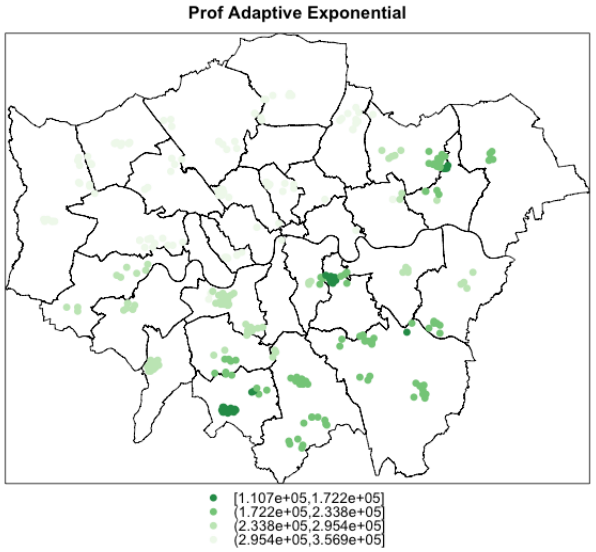
Map 12 a



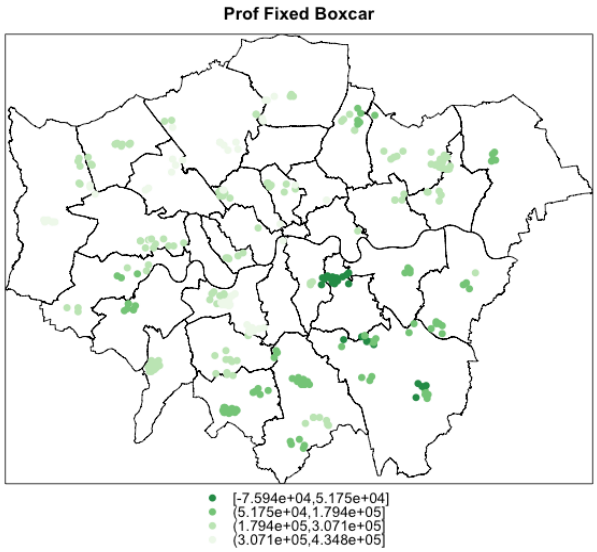
Map 12 b



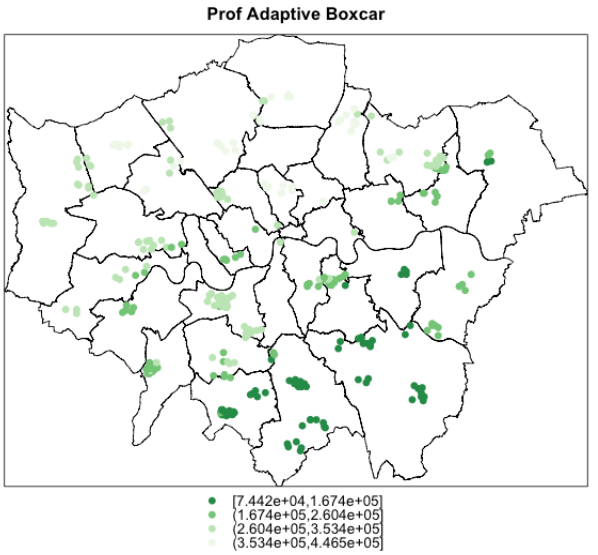
Map 12 c



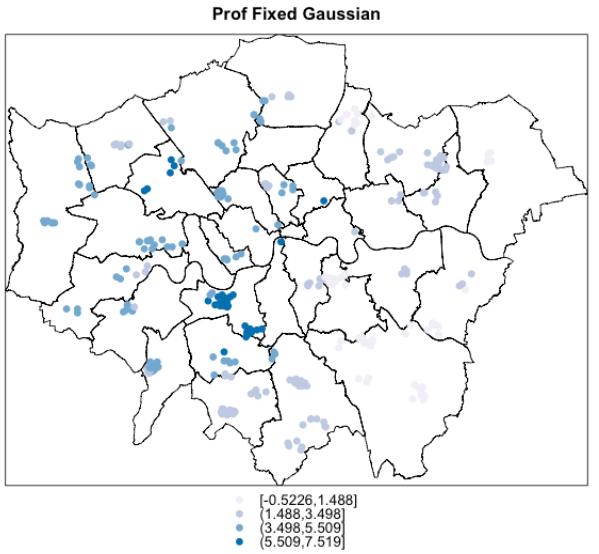
Map 12 d



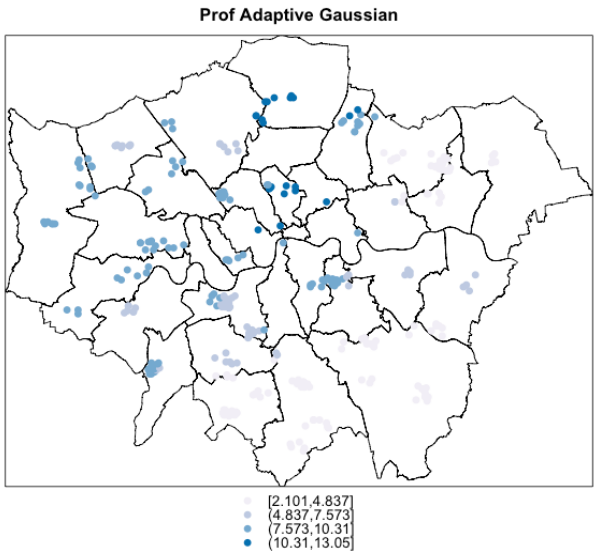
Map 12 e



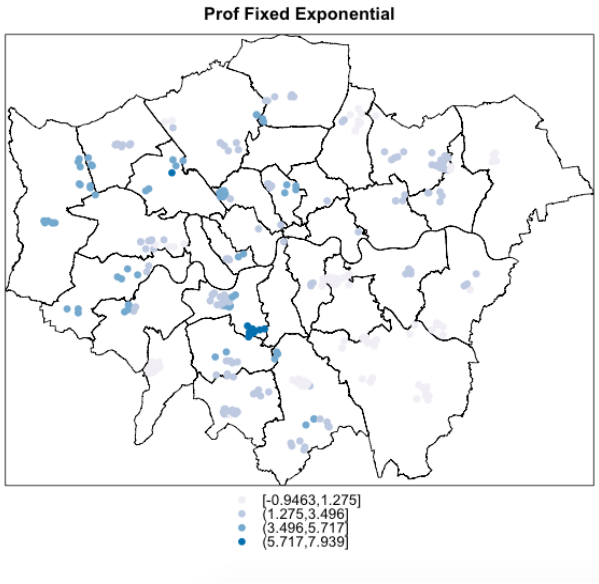
Map 12 f



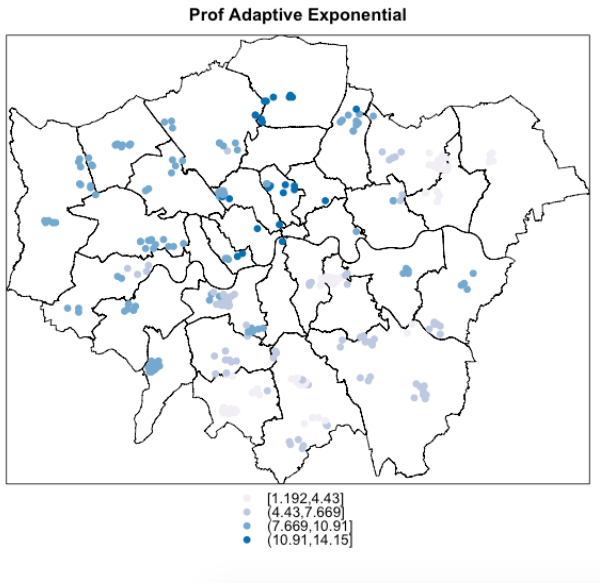
Map 13 a



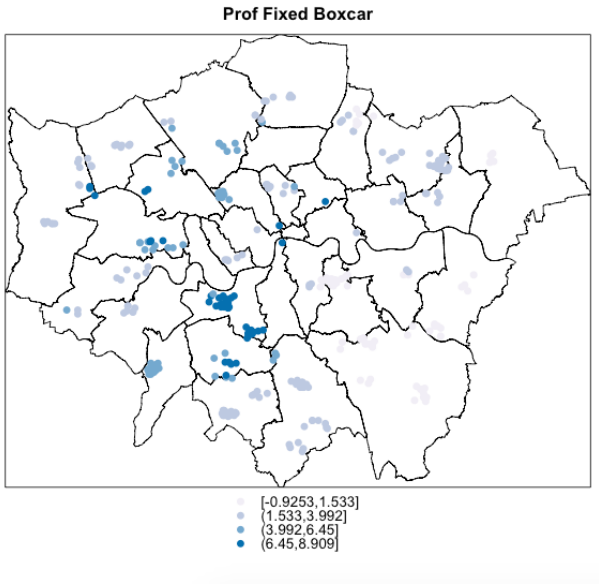
Map 13 b



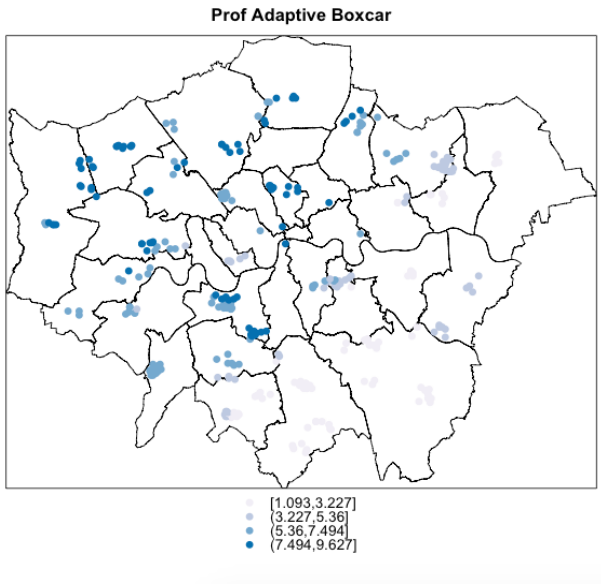
Map 13 c



Map 13 d



Map 13 e



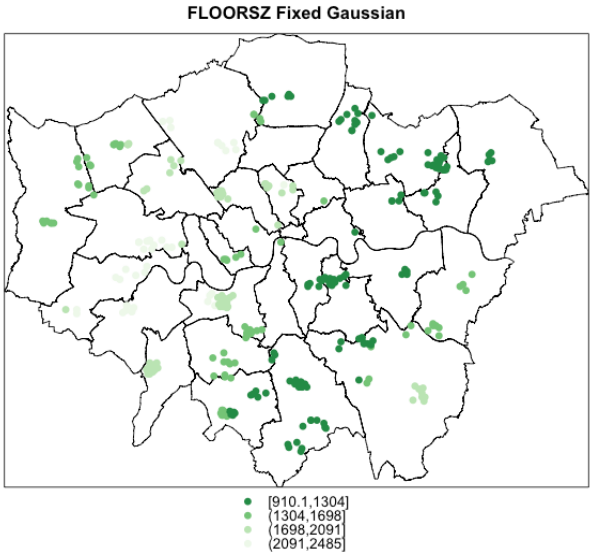
Map 13 f

8.2. FLOOR SIZE

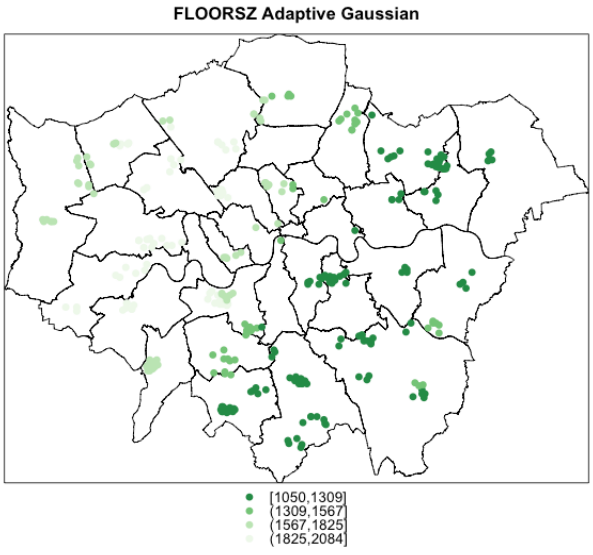
The relation between floor size and house prices appears stronger in the Central West part of London with low strength in the South and Middle Eastern boroughs. Areas such as Bromley and Bexley present some clusters of stronger correlation while the borough of Croydon varies depending on the parameters used.

Looking at the t-values, similarly to the previous explanatory variable, maps with fixed bandwidth have the regression points with the lowest t-values on the borders. Despite all observations reaching statistical significance, minimum t-value is 2.658 in the fixed Exponential map (map 15c) above the 99% significance threshold of 2.576. These have lower value compared to those in the adaptive maps especially in the North West boroughs. Turning to the adaptive maps, t-values for Gaussian and Exponential (maps 15b, 15d) tend to be similar compared to the Boxcar kernel (map 15f) that assigns low t-values to the whole East. The Boxcar kernel does not model the relation between floor price and purchase prices and isolated clusters in Bexley and Bromley. The large number of neighbours used to calibrate the t-values, 72 out of 361 observations, does not allow the map to show details. Gaussian and Exponential adaptive maps are able to represent the low relation that is expressed by the value distribution in the North East of London. Here, while house prices seem to be lower, the floor size is not particularly smaller compared to other areas. Yet, the Boxcar kernel assigns high values of regression coefficient to these zones.

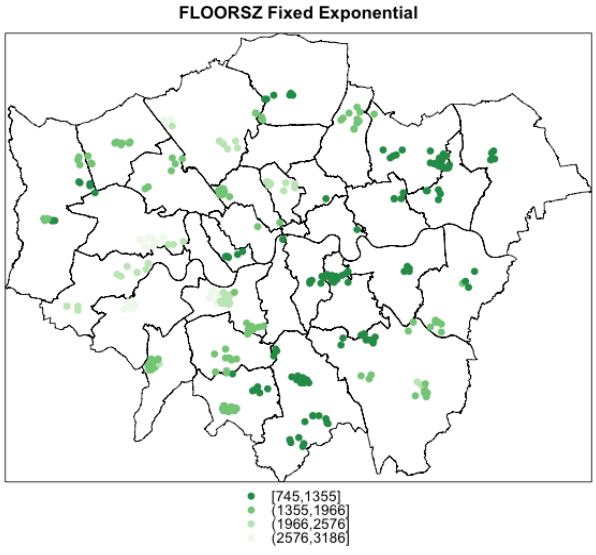
Moving to the fixed maps, in defiance of significantly different bandwidth sizes (Exponential: 1783, Gaussian: 3515, Boxcar: 8163) the difference in estimate parameters is very minimal. Only the Exponential kernel (map 14c) gives weaker regression coefficients to those observations in the North West boroughs. All the Middle Eastern section from North to South is given smaller coefficients in all the three maps and the Centre West and South East is given lower values suggesting that the shape of the kernel does not have much influence when observations are either all tighter within a cluster or far apart.



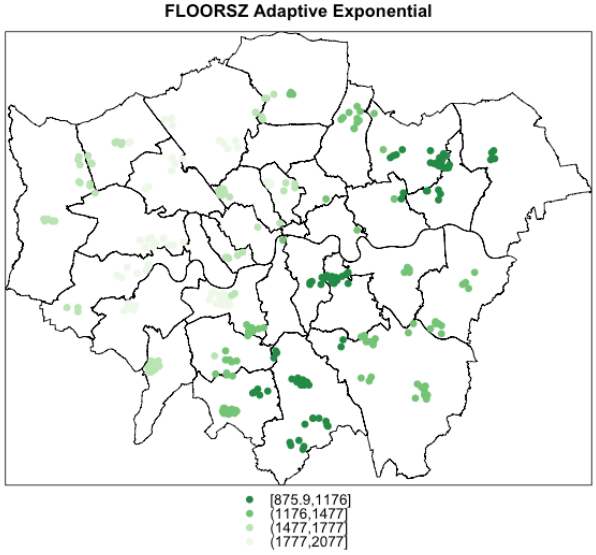
Map 14 a



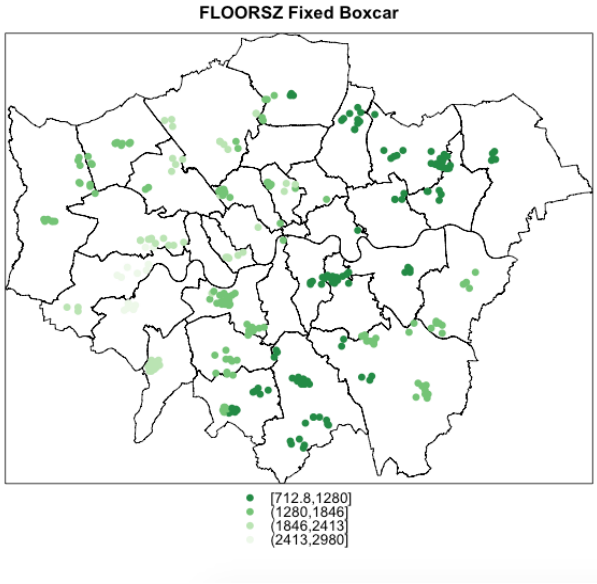
Map 14 b



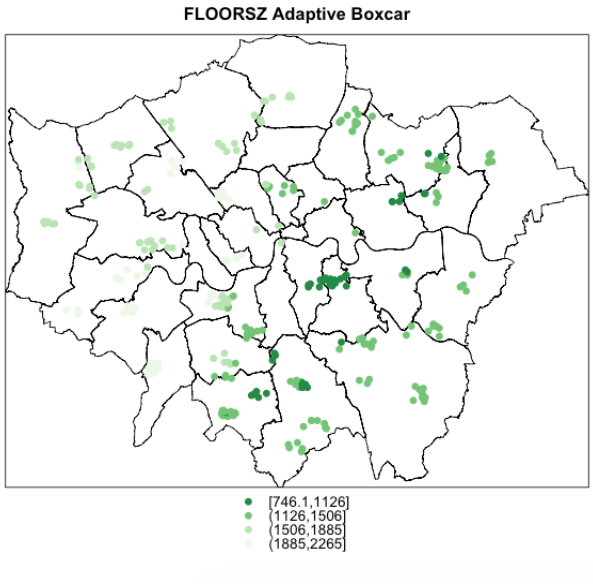
Map 14 c



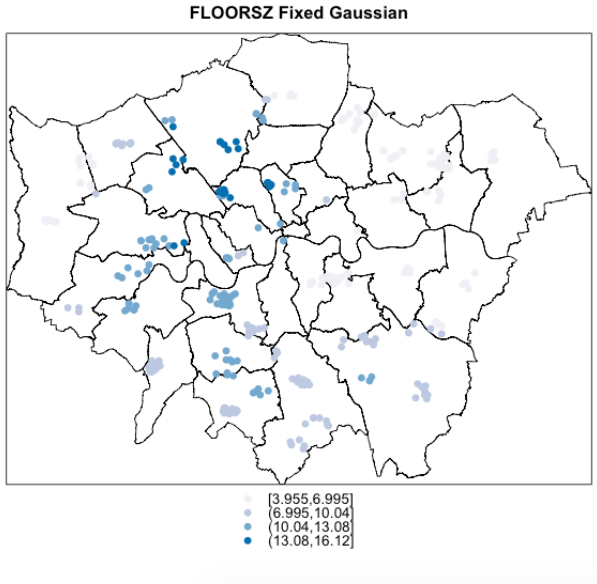
Map 14 d



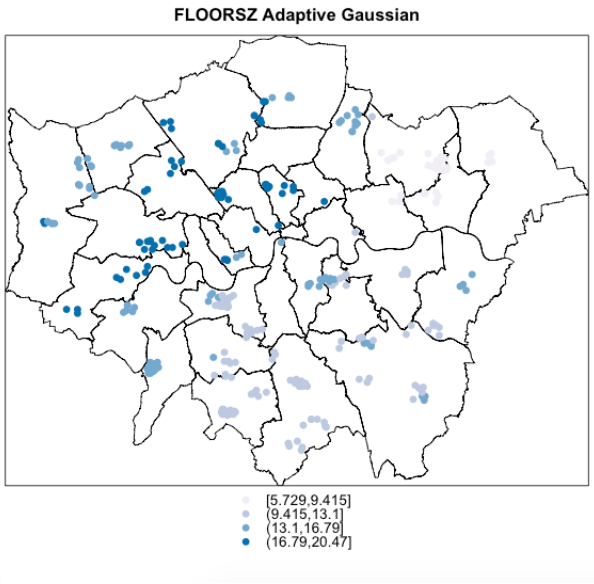
Map 14 e



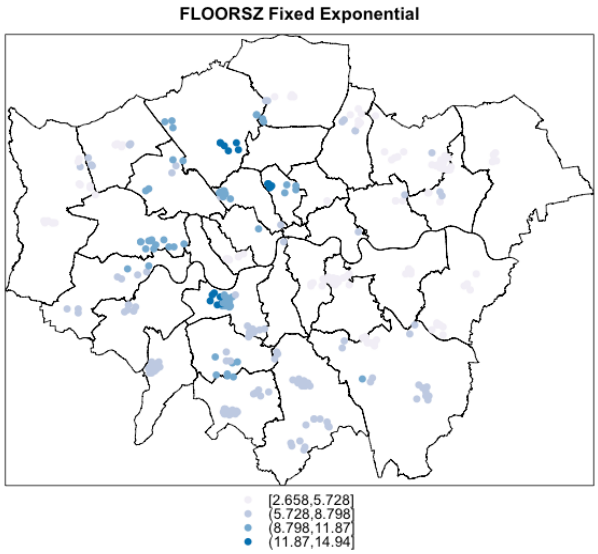
Map 14 f



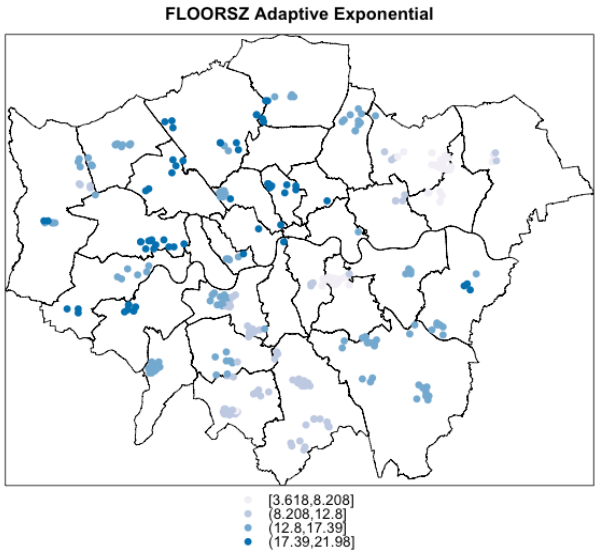
Map 15 a



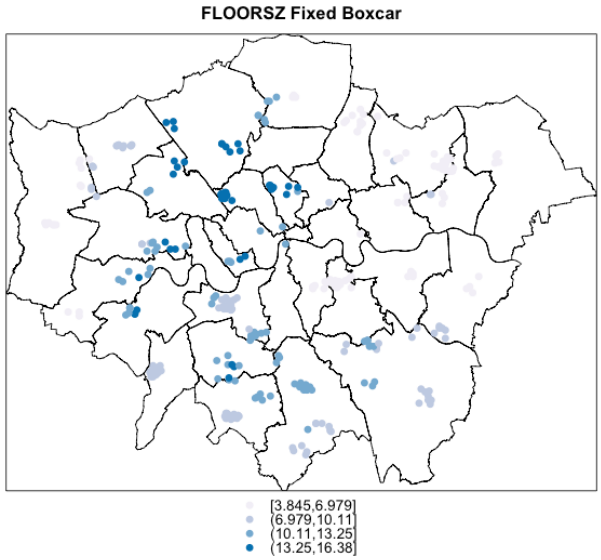
Map 15 b



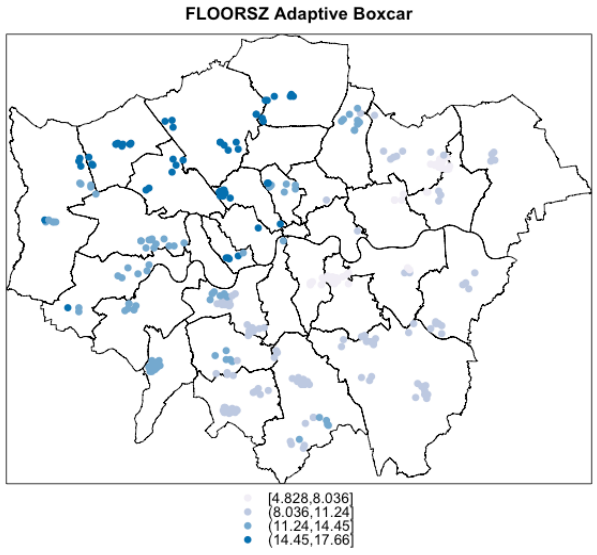
Map 15 c



Map 15 d



Map 15 e

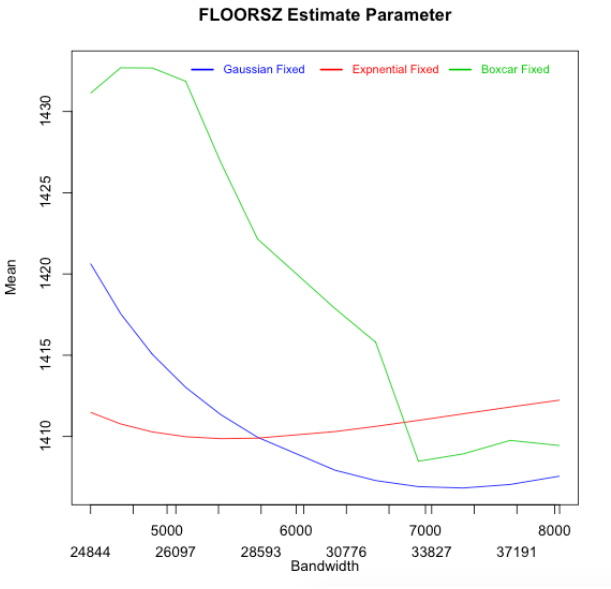


Map 15 f

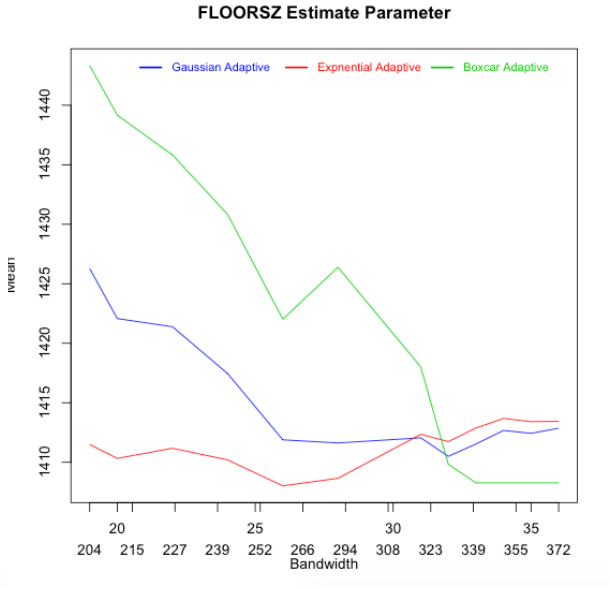
8.3. LONDON HOUSES PRICES BANDWIDTH VARIATION

The final dataset with the multiple clusters contrary to the previous ones has a surprisingly consistent trend for the boxcar kernel when it is switched from fixed to adaptive. Regularities are present in both variables for both estimate parameters and t-values. The Boxcar estimate coefficients are generally higher compared to Gaussian or Exponential with a tendency to decrease quicker and reach comparable values when the bandwidth gets larger. Standard deviation for both estimate and t-values is always significantly lower than the other kernels, highlighting a model that expresses little variation. Also, the significance expressed by the t-values has much higher levels with a consistent gap across variable and bandwidth type.

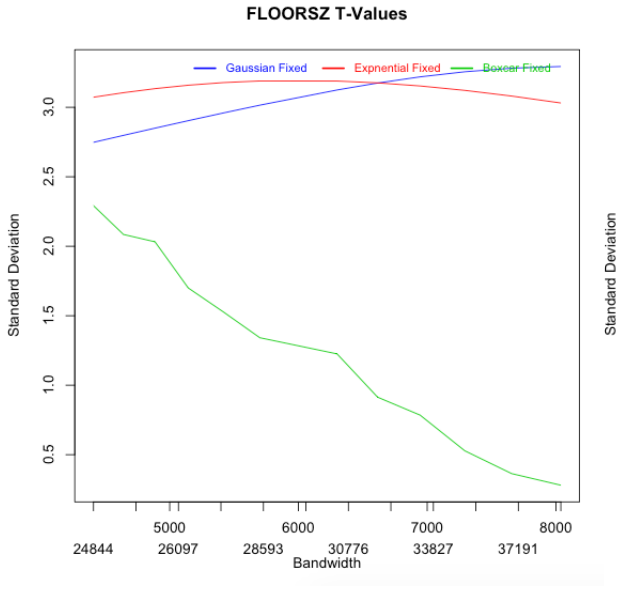
Turning to Gaussian and Exponential, differences in adaptive and fixed kernel are more deeply marked with this density pattern. Looking at the estimate parameters, it is evident how the variation in the fixed bandwidth is much more gentle than those expressed by the adaptive one. A little change in the number of neighbours can have a strong impact over the mean value of the estimate parameters. Moreover, after a certain bandwidth size, the difference between Gaussian and Exponential (and Boxcar looking at Estimate for percentage of professionals) seem to reduce consistently. As well as the variation in the t-values suffers from the inclusion or exclusion of a few neighbours. Indeed, changes in the significance distribution are much more sudden in the adaptive versions for both explanatory variables. While Exponential and Gaussian show a gentle variation in the fixed maps, the Boxcar kernel is affected by sudden changes in both versions putting forward for consideration its unpredictable nature strongly affected by any variation in the inclusion/exclusion of observations during the computations of its estimate parameters and t-values.



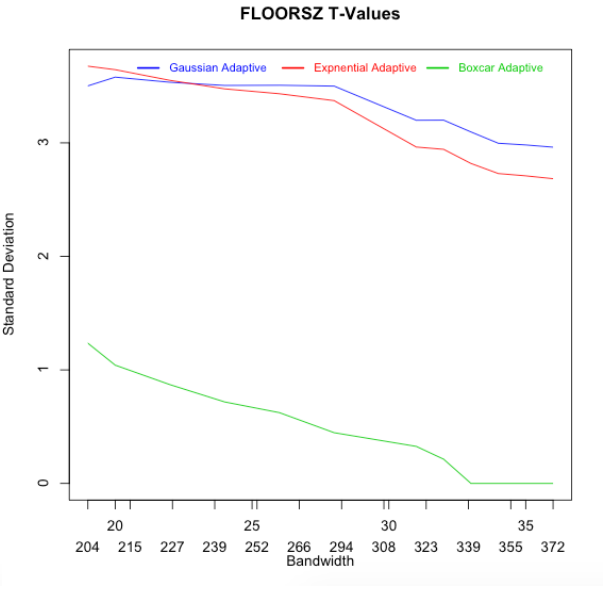
Graph 5 a



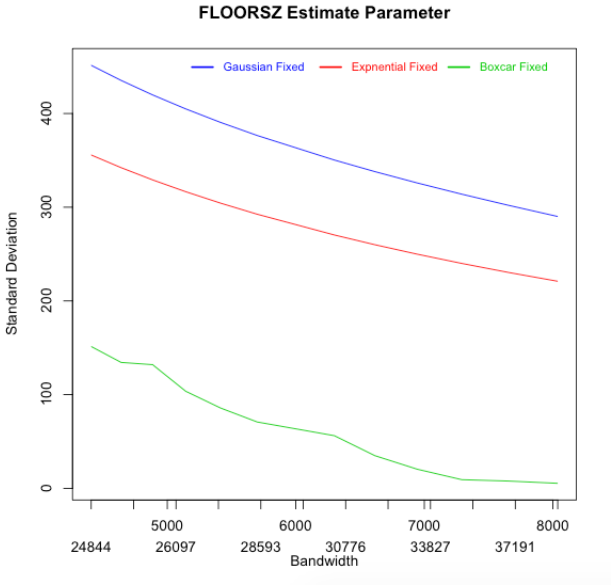
Graph 5 b



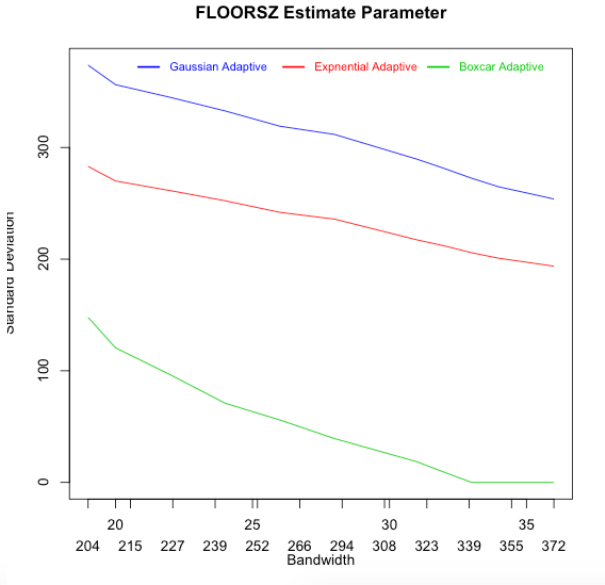
Graph 5 c



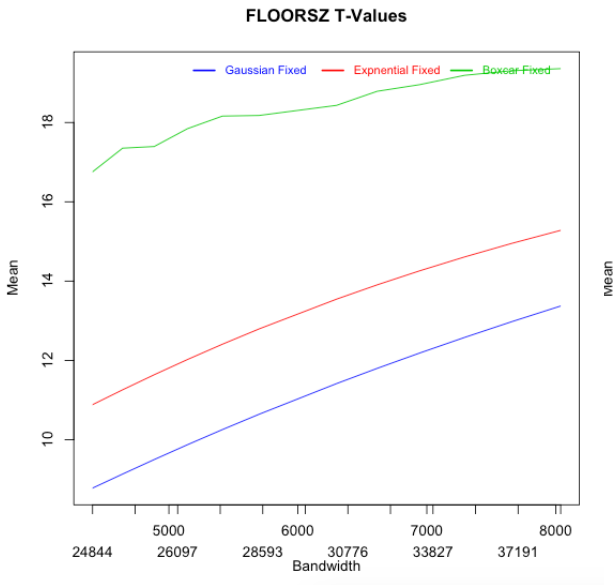
Graph 5 d



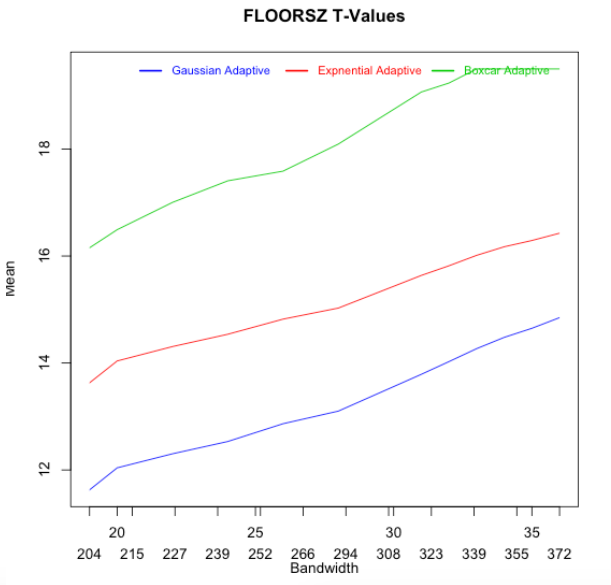
Graph 5 e



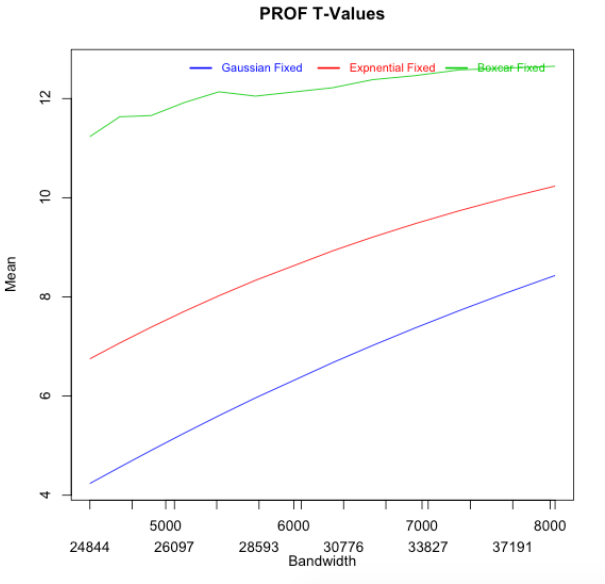
Graph 5 f



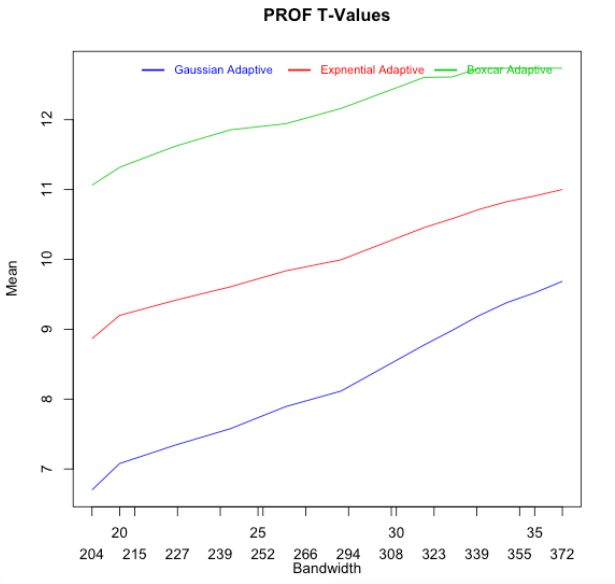
Graph 5 g



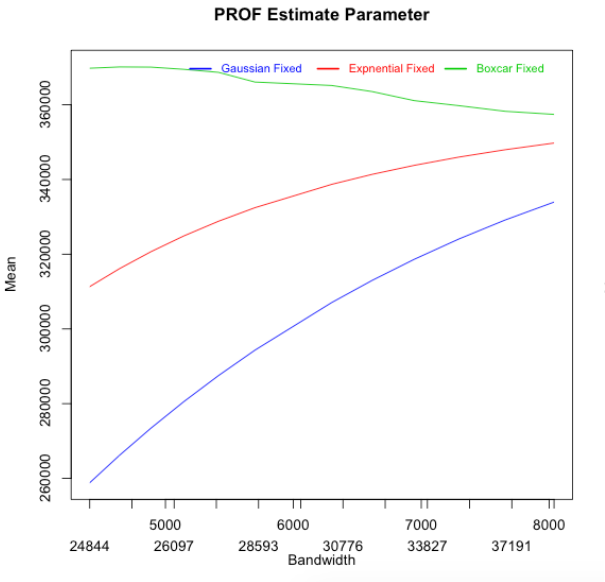
Graph 5 h



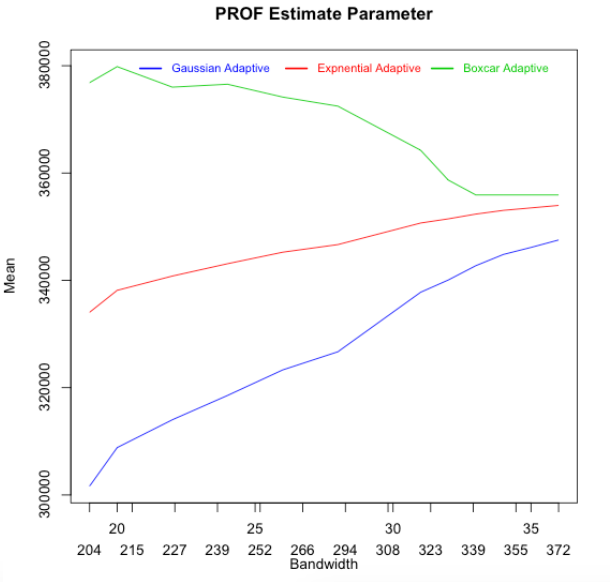
Graph 6 a



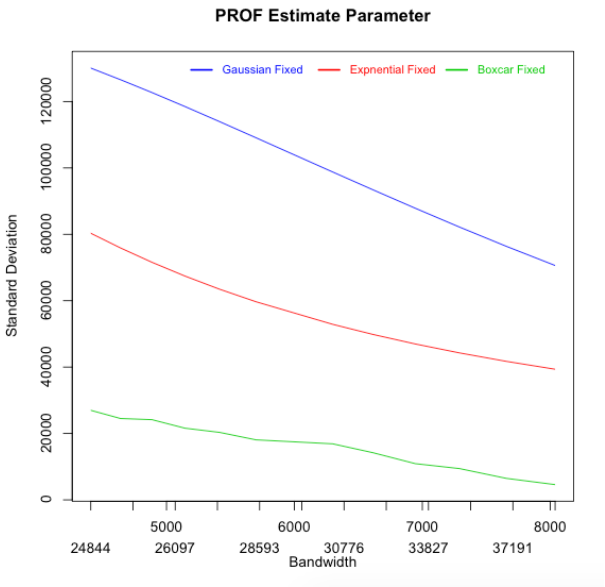
Graph 6 b



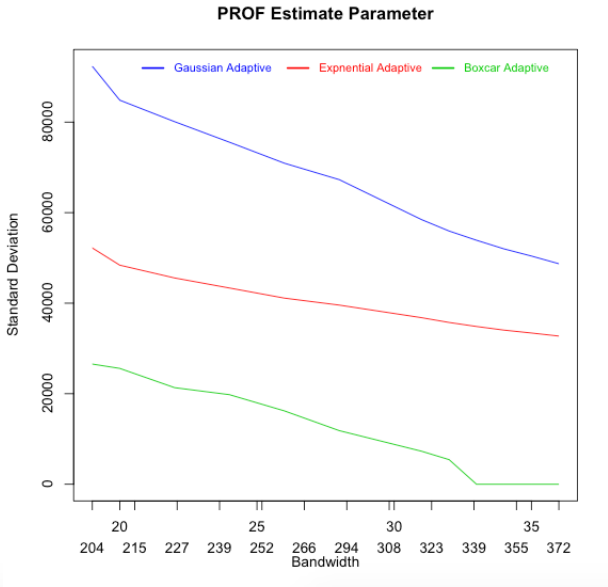
Graph 6 c



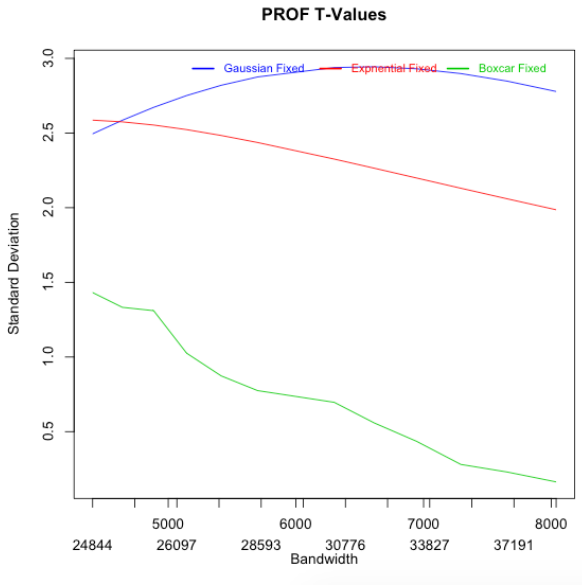
Graph 6 d



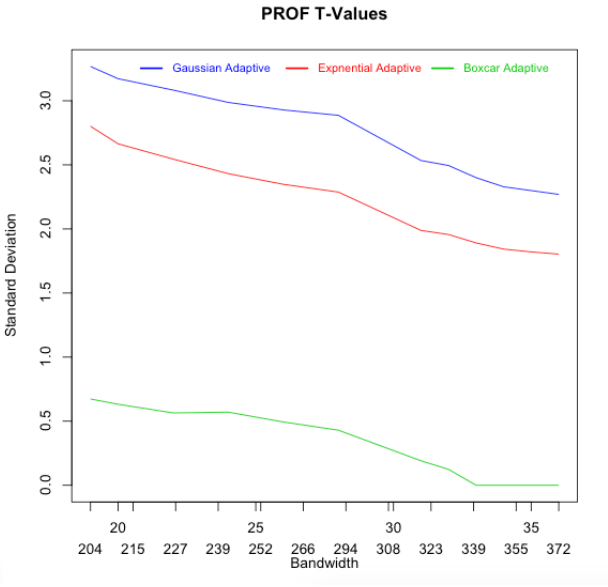
Graph 6 e



Graph 6 f



Graph 6 g



Graph 6 h

9. DISCUSSION

The analysis carried in the Georgia dataset, gave an insight over the bandwidth settings where the density pattern does not play as a variable. The first regularity that can be drawn, regards the different behaviours of fixed and adaptive bandwidth, when these are used to compute estimate parameters and t-values around the edges of the maps. The adaptive bandwidth has shown to have an advantage over the fixed, in producing coefficients more representative of the underlying data distribution, when these are computed near the borders or the corners. Thanks to the nature of the adaptive calibration, regardless the position of the observation upon which the coefficients are computed, the number of neighbours used to determine estimate parameter and t-values will always be the same for any location. This was demonstrated in the maps representing the spatial relation between percentage of black people and percentage of poverty. The North West corner in the fixed maps was indeed considered an area with particularly low estimate coefficients, while this was not reflected by the geographical distribution of the values. Also the t-values were affected by the switching between adaptive and fixed. As shown by the two variables, locations near the edges tend to have smaller significance than those in inner regions. The t-values produced by the adaptive maps follow the estimate parameters more closely compared to the fixed versions. This was particularly true when comparing the difference between high estimate parameters and high t-values for the area around Columbus in the South West. In spite of strong estimate parameters returned by both fixed and adaptive models, only the adaptive t-values gave high significance to the area, further confirming the poor performance of fixed GWRs when these have to deal with observations at the border/corner of the map. Moving to the Dublin Voter dataset, here the varying levels of density patterns portrayed supplementary difference between fixed and adaptive bandwidth. The first striking element to be considered is the different distribution of estimate parameters that the two typology of maps modelled. Fixed GWRs were strongly affected by the varying different pattern to the point that the two middle classes of estimates (first quartile-median, median-third quartile) had much smaller intervals compared to the first class (minimum-first quartile) and last class (third quartile-maximum value). The number of outliers was therefore predominant in fixed bandwidth maps especially in sparse regions. With the exception of the Boxcar kernel in percentage of age 25-44, the North of Dublin in the fixed maps appeared populated by extreme positive and high coefficients not representative of the negative correlation expressed by the geographical value distribution.

T-values had a similar behaviour showing complete different results for the two bandwidths. Fixed GWRs give very low significance to the sparse areas that are furthermore at the edges of the map.

For instance, in the percentage of unemployments maps, the fixed bandwidths GWRs concentrated most of the significance around the centre of the map, while the adaptive counterpart add the opposite behaviour. Also the percentage of ages 25-44 maps were characterised by strong differences. The North of Dublin and part of the South West area were the locations where the adaptive bandwidth gave the strongest t-values. Due to their spareness these were allocated with the lowest class of values in the fixed variants despite the fact that geographical distribution seemed to suggest correlations between the two variables in that zone. The density pattern represented by the London house prices dataset gives more credit to the fixed bandwidth as a viable setting in GWR. Looking at percentage of professionals, high values for this variable are not followed by high prices for these boroughs. Despite the fact that fixed bandwidth gives slightly stronger estimate coefficients to theses areas, the adaptive bandwidth strongly overestimates the t-values for these boroughs. The fixed model gave low to no significance to these locations while the adaptive had no observation below the 95% threshold of significance. A possible explanation might be given by thinking on how the density pattern and the two typologies of bandwidth behaves. A regression point confutes with adaptive bandwidth, which in this circumstance is highly affected by the overall observation present in the nearby clusters. An optimal CV adaptive bandwidth can have as the number of neighbours a larger number than the observations in a cluster. This makes the coefficient computed for one cluster to be based on observations that are also from different clusters. Contrary to the Dublin voter dataset where there was just one large cluster in the centre and sparse zones around it, here each small cluster may represent a separate environment with its own relation. The fixed bandwidth value used was apparently small enough to let each cluster be isolated without being biased by the trend of its surroundings. The few observations did not provide statistical evidence that the relation was significant and therefore these areas that reported a misleading relation were discarded. The same could not be said for the adaptive bandwidth that reported the relation in these areas a significant result.

Moving to the comparison between different kernel shapes, The Georgia dataset provided little evidence of any significant difference between Gaussian and Exponential kernels. This was more marked in the first set of maps depicting the relation between poverty and percentage of blacks. The steeper shape of the Exponential interrupted the continuous high estimate parameter zone between Madison (North East) and Columbus (South West). The same maps both fixed and adaptive with Gaussian kernel returned a continuous zone of high estimate parameter. Looking at the geographical distribution of the two variables, both poverty and percentage of black people have two strong poles near Madison and Columbus but different less correlated values between the two areas. The exponential seemed to be the best choice to unveil the latent relation thanks to its more precise

outcome. On the other hand, the Boxcar kernel produced a more approximated result that showed only strong estimates in the area around Madison. Notwithstanding, this kernel was still able to yield some insights able to model the relation better than a global model. What is more, the difference in t-values presented very little differences over the t-values with a degree of highly significant locations proportional to the size of the bandwidth. For instance, both the fixed Exponential maps for both percentage of elderly and percentage of black had generally lower values around the West border compared to Gaussian and Boxcar. Overall, the outcome provided did not show great differences for Gaussian and Exponential. Looking at how the kernels vary according to optimal CV bandwidth size, the gentle variation expressed by Gaussian and Exponential, together with the same directions, further reduce the difference between these two kernels. Adversely, the Boxcar, due to its larger bandwidth and different weighting scheme differs greatly. The only partial form of consistence was found in the t-values. While the rate at which the mean of these coefficients grows seems to be comparable with the other shapes the rate at which the standard deviation decreases when the bandwidth gets larger than the optimal CV value is much faster. This suggest that the Boxcar kernel GWR with high bandwidth values resembles global models in as much as little to no variation is expressed in the coefficients questioning their utility, since their primary function, describing local variation, is compromised. This was further proved by the Dublin Voter dataset. The Fixed Boxcar maps for both the percentage of unemployment and the percentage of age25-44 had no significant difference from a global model. Estimate parameters were included in the intervals -0.51/-0.76 and -0.68/-0.35 for the percentage of unemployment and percentage of age25-44, while the t-values where in the intervals -4.68/- 6 the former and -4.15/-7.89 the latter. Notwithstanding the combination Boxcar-adaptive gave a better result with wider variations in terms of both estimate parameters and t-values for both maps. Also for this dataset the difference between Gaussian and Exponential was very little. No significant difference was found in both variables neither for adaptive nor for fixed and also in the variations of bandwidth sizes. Also for this density pattern the Boxcar kernel does not show signs of consistency in its behaviour over different bandwidth values.

The London house prices dataset depicts a stronger difference between Exponential and the other two kernels than the usual. In the first set of maps representing the relation between percentage of professionals and purchase price, it is possible to notice how the fixed map made with Exponential kernel gave results significantly different for at least the borough of Kensington upon Thames. In fact, the Exponential kernel allocates very low values to this isolated cluster as well as little statistical significance as expressed by the t-values. This relation was supported by the geographical distribution of the values. The Exponential kernel in its fixed variant was the only combination able

to represent the real underlying relation between dependent and independent variable. The steep shape together with a generally smaller bandwidth, provided a good model able to tailor its coefficients to the small clusters of which the data was made contrary to Boxcar and Gaussian kernels, that, with their weighting scheme did not let to achieve such precision in a situation where a detailed map was more necessary than a smoother one. Finally, this density pattern shows the least predictable changes over bandwidth sizes experienced in this analysis for Gaussian and Exponential, while Boxcar seemed to have some regularities. In contrast with the other two datasets, variations in the means of estimate parameters is much greater when considering the adaptive kernels. This is particularly shown by the adaptive estimate parameters variations for floor size. With small bandwidth values, the difference between the Gaussian and Exponential shows a significant gap in the values of estimate parameters. Small increments of the number of neighbours used to calculate these values represent sudden changes in the mean of estimate parameters arguably representing members of different clusters affecting the values. When the number of neighbours is large enough, all three kernels converge to similar results, causing each kernel to lose its peculiar properties.

10. CONCLUSIONS

After having analysed different settings of bandwidths parameters over datasets with different structures and characteristics, it is possible to draw some conclusions regarding the effects of the parameters choices to help analysts in the calibration of their models. Under uniform density patterns, the adaptive bandwidth seems to have an advantage over the fixed one when important relations between variables are at the borders of the map. The fewer data points used to calibrate estimate parameters and in t-values tend to underestimate the strength and the significance of the relation when this is not in an inner portion of the map. Conversely, the adaptive version does not suffer from this limitation. In particular, the Exponential kernel thanks to its steep shape, was able to provide accurate results even when in the adaptive version. Notwithstanding, differences between Gaussian and Exponential were minimal over this data. On the other hand, the use of the Boxcar kernel does not seem to provide any advantage that can off-set its low level of accuracy and therefore should not be chosen when dealing with uniformed distributed data. The much larger bandwidth size and the weighting scheme do not favour the main advantage of the GWR which is returning local coefficients that describe how the relation changes across space. This is mainly due to the violation of the main assumption: Closer observations have greater influence on the coefficients of the regression point than those far away. The application of the Boxcar kernel over other data was not more successful. The results of the Dublin Voter dataset confirmed its resemblance to a global model due to its inability expressing a wide range of variation. The London house prices data showed performances almost comparable to the Gaussian kernel but overall inferior in terms of accuracy in describing the latent relation. With the Boxcar out as reliable parameter setting, the choice is narrowed to Gaussian and Exponential on one hand and adaptive or fixed bandwidth on the other hand, and their interplay between density patterns and adaptive or fixed bandwidth.

Starting with the choice of either adaptive or fixed bandwidth, the density pattern plays a considerable role in determining the coefficient values obtained by the two different kernels typologies. The Dublin Voter data with its high density centre and sparse outskirts provided a difficult test for the fixed bandwidth. In fact, the number of outliers produced by this setting was significantly higher regardless the kernel used. What is more, the t-values in the sparse regions were constantly lower in terms of significance. By keeping a fixed radius and applying the kernel weighting scheme according to this, the model that results form it is as one size fits all that cannot deal with a data surface with heterogeneous data density. The model selection carried out through optimal CV favoured the centre area giving to most of the significance. If an important relation was

present in the outer circle of the city this could have not been modelled efficiently. The adaptive bandwidth did not show such unfitness. By increasing and decreasing the bandwidth size according to the density around the regression point, the model was able to return high t-values and within range estimate parameters also for those areas that were not represented as integral part of the relation in the fixed maps. The ability to overcome the geographical distance while considering only a set number of closest neighbours gives to the adaptive setting a considerable advantage that should not be underestimated.

Nevertheless, in spite of better results in the Dublin Voter data and Georgia data both in terms of performance on the outer area of the map, the adaptive bandwidth should not be a default solution always working as default parameter. If it is true that sparse regions should be included in model in the sense that also these should be able to reach high levels of significance, it is true as well that in some circumstances, the distance in a map should not be seen as an obstacle to overcome but as specific feature to take into account. London with its peculiar nature of agglomerate of smaller urban centres and villages results a series of several little clusters making the London house prices data a good example were the geographical distance matters. The disadvantage of the adaptive bandwidth in this case was given by its impossibility of respecting the given geographical pattern of the data. Small clusters composed by very few observations often got the inclusion of far away data points in the computation of their coefficients. The result was a biased map that could not avoid to follow the general trend even in areas where the relation as depicted by the variable geographical distribution was inferably different. Conversely, the fixed bandwidth respected the geographical pattern of the data maintaining isolated those clusters in particularly far away from the centre. The fixed maps in this situation provided a much more detailed description of the underlying relation than the global model with several areas that showed much weaker estimate parameters than the overall trend with an even more specific distinction between areas that reached statistical significance and those which were below the threshold. The fixed bandwidth seemed to describe the particular relation of each isolate cluster regardless the overall. This was also the dataset in which the change in bandwidth size was particularly different for the two bandwidth. The behaviour of the adaptive bandwidth over increases of size demonstrated how at each small increment the estimate parameters would have considerably changed suggesting that the model is unstable since a little variation of one parameter may return significantly different results.

Finally, differences between the two kernels, Gaussian and Exponential were less marked. Both kernels returned similar results for most of the maps that were produced. Overall, Gaussian kernels produced smoother maps while the Exponential kernels were more precise in segmenting the areas

around interesting spots. Variations from the optimal CV bandwidth followed similar trend for both kernels further highlighting their similitude also over size changes. The density pattern of the London house prices data presented a situation in which the difference between Gaussian and Exponential was stronger. In fact, the Exponential kernel was better able to isolate the little clusters reducing at minimum the influence of neighbour data points from other clusters suggesting that this is the kernel to use when precision has to take over smoothness.

In the light of the analysis made, the findings of this study can be summarised as follows. The coefficients returned by a Boxcar kernel are hardly significantly better than a global model. Its coefficients both in terms of significance and estimate parameters express little variation across space and underestimate any form of smaller local relation that goes against the general trend. Gaussian and Exponential kernels do not present string differences. The Gaussian kernel should be preferred when identifying contiguous areas of similar coefficients and smoothness are a priority. On the other hand, for a more detailed representation of the underlying relation, the Exponential kernel can produce coefficients much more sensible to small local variations. Finally, the most important recommendation that this study can provide, on how to set bandwidth parameters in GWR, regards the relation between density pattern and bandwidth type. Adaptive bandwidth should be chosen when the data points present one area denser than the rest of the map. Thanks to its adaptive nature coefficients computed on regression points in sparse regions can reach as well level of significance as the coefficients in the denser areas. When the density pattern resembles more a series of different clusters and geographical distance makes sense, the fixed bandwidth should be chosen. This has the advantage of reducing at minimum the influence of geographically distant observations letting local variations exist.

Limitations of this work pertained mainly to two areas. First, however different and able to provide various insights over the behaviour of GWR, the density patterns that the data used in this work expressed certainly it is not exhaustive of all possible patterns that a study based on GWR may encounter. Other georeferenced datasets might provide further evidence to inform practise. Second, only one technique was considered. The GWR is just one technique of a whole family of geographically weighted techniques such as the geographically weighted principal component analysis or geographically weighted discriminant analysis. Allegedly, different bandwidth settings can have a strong impact also on these techniques. Future works should be focused on the comparison of different bandwidth parameters in GWR and other geographically weighted techniques as well as over new and different density patterns. In particular, research on geographical data where the actual distance between data points is reflective of a diminished

evenness between remote areas may provide supplementary clues over the results generated by either adaptive or fixed bandwidth.

11. References

Bivand, R. and Yu, D. (2015). *Package â€˜spgwrâ€™*. 1st ed. [ebook] Available at: https://cran.r-project.org/web/packages/spgwr/spgwr.pdf [Accessed 1 Sep. 2015].

Brunsdon, C., Fotheringham, A. and Charlton, M. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 28(4), pp.281-298.

Cho, S., Lambert, D. and Chen, Z. (2010). Geographically weighted regression bandwidth selection and spatial autocorrelation: an empirical example using Chinese agriculture data. *Applied Economics Letters*, 17(8), pp.767-772.

Cho, S., Lambert, D., Kim, S. and Jung, S. (2009). Extreme Coefficients in Geographically Weighted Regression and Their Effects on Mapping. *GIScience & Remote Sensing*, 46(3), pp.273-288.

Cleveland, W. and Devlin, S. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403), pp.596-610.

Cromley, R. (1996). A comparison of optimal classification strategies for choroplethic displays of spatially aggregated data. *International journal of geographical information systems*, 10(4), pp.405-424.

Evans, I. (1977). The Selection of Class Intervals. *Transactions of the Institute of British Geographers*, 2(1), p.98.

Farber, S. and Paez, A. (2007). A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. *Journal of Geographical Systems*, 9(4), pp.371-396.

Fotheringham, A., Charlton, M. and Brunsdon, C. (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environ. Plann. A*, 30(11), pp.1905-1927.

Fortheringham, A., Brunsdon, C. and Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester: Wiley.

Goodchild, M. and Janelle, D. (2004). *Spatially integrated social science*. New York: Oxford University Press.

Guo, L., Ma, Z. and Zhang, L. (2008). Comparison of bandwidth selection in application of geographically weighted regression: a case study. *Can. J. For. Res.*, 38(9), pp.2526-2534.

Harris P, Brunsdon C, Charlton M (2011) Geographically weighted principal components analysis. *International Journal of Geographical Information Science* 25 (10):1717-1736

Harrower, M. and Brewer, C. (2003). ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal*, 40(1), pp.27-37.

Hurvich, C. (1998). A crossvalidatory AIC for hard wavelet thresholding in spatially adaptive function estimation. *Biometrika*, 85(3), pp.701-710.

Kavanagh A (2006) Turnout or turned off? Electoral participation in Dublin in the early 21st Century. *Journal of Irish Urban Studies* 3(2):1-24

Li, S., Zhao, Z., Miaomiao, X. and Wang, Y. (2010). Investigating spatial non-stationary and scale-dependent relationships between urban surface temperature and environmental factors using geographically weighted regression. *Environmental Modelling & Software*, 25(12), pp.1789-1800.

Lloyd, C. and Shuttleworth, I. (2005). Analysing commuting using local regression techniques: scale, sensitivity, and geographical patterning. *Environ. Plann. A*, 37(1), pp.81-103.

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

Lu, B, Charlton, M, Harris, P, Fotheringham, AS (2014) Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data. *International Journal of Geographical Information Science* 28(4): 660-681

Lu, B., Charlton, M. and Fotheringhama, A. (2011). Geographically Weighted Regression Using a Non-Euclidean Distance Metric with a Study on London House Price Data. *Procedia Environmental Sciences*, 7, pp.92-97.

Lu, B., Charlton, M. and Harris, P. (2012). Geographically Weighted Regression Using a Non-Euclidean Distance Metric with Simulation DataBinbin. *International Conference on Agro-Geoinformatics*, pp.267-270.

Lu, B., Harris, P., Charlton, M., Brundson, C. and Nakaya, T. (2015). [online] Available at:

https://cran.r-project.org/web/packages/GWmodel/GWmodel.pdf [Accessed 20 Sep. 2015].

Mckinney, W. (2012). *Python for Data Analysis*. Sebastopol: O'Reilly Media.

Mennis, J. (2006). Mapping the Results of Geographically Weighted Regression. *The Cartographic Journal*, 43(2), pp.171-179.

Munzner, T. and Maguire, E. (2014). *Visualization analysis & design*. CRC press.

Nakaya, T., Fotheringham, A., Brunsdon, C. and Charlton, M. (2005). Geographically weighted Poisson regression for disease association mapping. *Statist. Med.*, 24(17), pp.2695-2717.

Siegmund, D. and Worsley, K. (1995). Testing for a Signal with Unknown Location and Scale in a Stationary Gaussian Random Field. *Ann. Statist.*, 23(2), pp.608-639.

TU, J. and XIA, Z. (2008). Examining spatially varying relationships between land use and water quality using geographically weighted regression I: Model design and evaluation. *Science of The Total Environment*, 407(1), pp.358-378.

Wheeler, D. (2015). *CRAN - Package gwrr*. [online] Cran.r-project.org. Available at: https://cran.r-project.org/web/packages/gwrr/index.html [Accessed 20 Sep. 2015].

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.

Yu, D. (2006). Spatially varying development mechanisms in the Greater Beijing Area: a geographically weighted regression investigation. *The Annals of Regional Science*, 40(1), pp.173-190.