Point Allocation Inside Polygons and GWR: An Experimental Analysis with Survey Data

Eduardo de Rezende Francisco[1,2], Peter Whigham[2], Antoni Moore[2]

[1]Escola de Administração de Empresas de São Paulo,
Fundação Getulio Vargas, São Paulo, Brazil
Phone: +64 3 479-7391 Fax: +64 3 479-8311
Email: eduardo.francisco@aes.com

[2]Spatial Information Research Centre
University of Otago. Dunedin, New Zealand
Phone: +64 3 479-7391 Fax: +64 3 479-8311
Email: pwhigham@infoscience.otago.ac.nz
Email: amoore@infoscience.otago.ac.nz

Presented at SIRC 2007 – The 19th Annual Colloquium of the Spatial Information Research Centre
University of Otago, Dunedin, New Zealand
December 6th-7th 2007

ABSTRACT

The aim of this paper is to analyse different alternative implementations for a problem defined as "point allocation inside polygons" for Geographically Weighted Regression (GWR). The problem involves situations where the precise location of each observation is not known - just its district, municipality or region, i.e. a polygon geographical location. However, associated data were available that could potentially allows point placement of observations. These analyses were applied in a Income predicting model based on electricity consumption from a survey for a power distribution company in Sao Paulo, Brazil. Completely spatially random allocation and allocations based on spatial distributions of population (universe) and of the independent variable (electricity consumption) were utilized. Results showing the coefficients of determination (R^2) suggest that a more realistic measure of the relationship between these two constructs could be evaluated.

***Keywords and phrases*:** GWR, point allocation inside polygons, point sampling, income, electricity consumption

1.0 INTRODUCTION

The accurate localisation (whilst maintaining privacy) of people or households responding to surveys is a natural aspiration, leading to potentially more powerful models and therefore new insights in the interpretation of survey results. Diverse efforts to achieve these localisations have been done in Survey, GIS-T and Geomarketing areas through intensive data collection. However, in many cases point-based data is not available and therefore several problems remain unsolved that may dramatically reduce the production of such accurate information.

It is very common nowadays to collect "spatially enabled" survey data. The motivation has come from several disciplines, including the universe of marketing and social sciences and the subsequent improvements that spatial statistics allow in the interpretation, measurement of relationships and prediction. However, it is also very common that the data related to these subjects is not quite suitable for spatial analysis. Many forms of data collection do not make available adequate location information (in accuracy or precision) – these situations produce data whose precise location of each observation is not known – making available just information about the region where the observation is located (district, region, postal code, municipality, etc).

This situation could be viewed as the opposite of the well known "Modifiable Areal Unit Problem" (Openshaw & Taylor 1979; Jelinski & Jianguo 1996), which describes the effect on the observed spatial relationships of data due to scaling and zonation. For the situation described here, rather then having a set of point data that can be aggregated in a variety of ways, we are given a fixed zonation with associated data, and want to place the data within the zone in a meaningful way.

New techniques of spatial statistics such as spatial regression models potentially benefit from methods that extend polygon-based data into point allocated data within polygons. Spatial Auto-regressive (SAR) models usually adopt distance weights (in the auto-regressive ρWy term), which are internally based on the distance between the centroids of the polygons involved in the predicting model. Geographically Weighted Regression (GWR) (Brunsdon, Fotheringham & Charlton 1998) relies on the computed distance between observations (treated directly as points) to calculate the weights for each local sample regression and, most importantly, to indicate the most appropriate local sample size, based on an Akaike Information Criterion (AIC) minimisation or cross-validation minimisation.

This paper investigates approaches to point allocation inside polygons using an empirical study by applying GWR models on a specific survey of the Brazilian power distribution sector. The regression models are constructed to predict household income with only one independent variable: the monthly billed residential electricity consumption (or simply energy or electricity consumption).

Section 2 will describe the methodology, including the objectives of the study and the four alternative point-allocation methods that are examined in this work. Section 3 will present the results and discuss the implications of the approaches, and Section 4 concludes with suggestions for on going work.

2.0 METHODOLOGY

2.1 Object: ABRADEE's Survey in the city of São Paulo, Brazil

The dataset utilized in these analyses was the traditional ABRADEE (Association of Brazilian Power Distribution Companies)'s Survey of Customer Satisfaction. This survey is applied annually for residential customers of all Brazilian power distribution companies of which there are sixty-four. Interviews are made for selected households, with the target person being the head of family. Dozens of questions related to satisfaction are applied, and some demographic aspects are obtained: household income, head of family's age and educational level, monthly billed electricity consumption of the household and some others.

The most detailed location information of the household collected in the interview is the DISTRICT where the household is located. Addresses, postal codes, census sectors or meshblocks were not considered.

The ABRADEE's Survey for the AES Eletropaulo, one of the most important Brazilian power distribution companies, were utilized – in particular, the survey applied in the city of São Paulo (in the Brazilian Southeast region) for the year of 2004.

2.2 Alternative Point Allocation Models

As the district of each interview is the only locational data recorded, and GWR considers points as the basic spatial unit of the observation, the simplest way to proceed is associate the district's centroid for each interview. In this way, many interviews are associated with the same "location".

Looking at this peculiarity, the issue would appear to be the unit of observation. The units are households, but we don't have any means of geocoding their location except to the centroid of the district in which they are associated. This means that any households in the same polygon would effectively be stacked one on top of the other at the centroid. This would result in a weight of 1 for the local sample regressions for each of these stacked points, while any households in adjacent polygons would receive a smaller weight, but again the weights would all be equal.

It seems that a centroid GWR should not produce realistic results (or at least results that are little better than a global regression) due to similar weights being given to data that may well be spatially dispersed and therefore have intrinsically different neighbourhood influences. Since this is the most naïve approach to allocating the point data it can be viewed as the null model – it would be expected that any model that can produce a more realistic spatial allocation of points within each polygon should improve the overall spatial regression performance compared with this null model.

To handle this main issue, at this point of the analysis, four alternatives of point distribution were considered. Some knowledge about the sampling planning of the survey would certainly support the implementation and improvement in quality of the alternatives, or perhaps even suggest new approaches to be included. Nevertheless, the sampling planning of most surveys are rarely known by the "users" of the survey (mainly in the Social Sciences), and for this reason the authors have decided not to consider alternatives based on specific survey metadata.

The alternative implementations were produced in the statistical environment R 2.5.1 , using the spatial packages (extensions): MAPTOOLS, SPLANCS and SPATSTAT.2.2.1.

2.2.1 Alternative 1: Completely Spatially Random Points in the Polygons

This is the simplest and most intuitive alternative to be considered. For each district sampled for this survey, we selected the number of interviews *(n)* and applied the generation of *n* points randomly in the district's polygon (actually, we generate the points inside the bounding box of the polygon, and extract those IN the polygon, repeating this process until the n interviews were positioned). Figure 1 shows an example of this implementation.

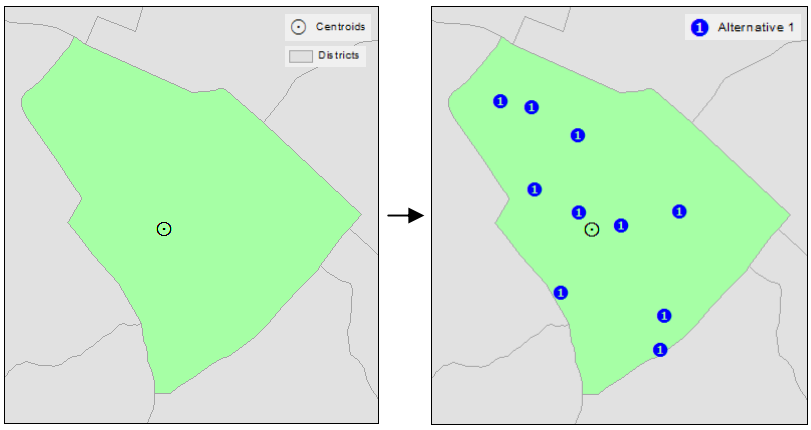

Figure 1: An example of Alternative 1 (generation of completely spatially random points within polygons)

2.2.2 Alternative 2: Generation based on Density of Households

For this alternative, information about density of households in the city of São Paulo was used as a surrogate for likelihood of survey location. This information was obtained from AES Eletropaulo, which is the only power distribution company that covers the study area. Hence, every electrified household is a customer of this company. Using a grid of 100 squared meters, we computed the number of residential customers (e.g. households) per cell per district. Based on this grid we generated a random point pattern containing *n* independent, identically distributed random points with the density of households' grid as the specified distribution (common probability density), considering that you have *n* interviews in the district.

Since the population distribution information is usually published by Census Bureaus or Agencies (in census sectors or meshblocks) the availability of this form of data is common and would normally be used in this case. The availability of density of residential customers from the power distribution company seemed a more accurate and useful alternative for this case in study, as the ABRADEE's survey universe is from the same cohort. Of course, this approach could use some alternative secondary source if required. Figure 2 shows an example of this implementation.

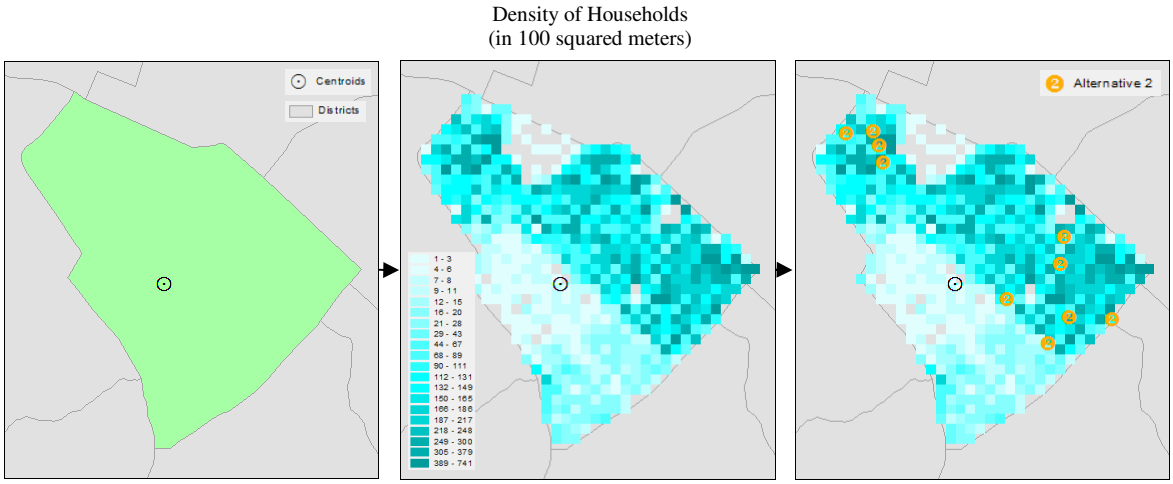

Figure 2: Example of Alternative 2 (generation of random points based on density of households)

2.2.3 Alternatives 3 and 4: Generation based on Probability of Fitness for Energy Consumption

Alternatives 3 and 4 are very similar in concept. They both are based on the distribution of energy consumption in the city of São Paulo, according to the customers of AES Eletropaulo. Alternative 3 generates a grid of fitness for electricity consumption, from the computation of the average of electricity consumption per cell in the grid. Then, similarly to Alternative 2, we generate a random point pattern of n independent points, distributed according to this surface.

Alternative 4, on the other hand, does not distribute points based on a spatial probability space. This alternative SELECTS one customer from the list of residential customers from the power distribution company. For each district and for each interview, we obtain the value of energy consumption and generate a proportional measure of fitness for energy consumption - using the list of customers' consumption, not a grid of average values. Finally, a uniform random number between 0 and 1 is generated and used to associate the location of the customer in the random index position for this interview. The assumption here is that since we have a large set of sampled values with location and we want to find a probabilistic placement of each specific individual to the list of known point values. Of course, the true location is not known (and in fact the same individual may not be represented in both surveys) so a probabilistic placement, biased towards a likely match, is the appropriate approach. The method is based on a fitness proportional model of selection, often used in evolutionary computation algorithms (Back, Fogel & Michalewicz 1997; Yao 1999).

A simplified description of the algorithm for Alternative 4 follows:

```
SurveyData ← List of Energy Consumptions of n households (collected in the interviews)
for each District in List of Districts
{
    PowerData ← List of Energy Consumptions of N residential customers in the district
    for each S in SurveyData
    {
        # q stores the absolute difference in consumption between S and each customer in PowerData
        q ← abs( PowerData - S )
        # convert q to a maximised value
        q ← (max( q ) + 1 ) - q
        # construct proportional measure of fitness for energy consumption
        q ← q / sum( q )
        # sample a completely random number in [0,1]
        i ← random(0,1)
        # get the
        P(q[i]) ← id of customer in PowerData located in the proportional position of q
        # associate S to this customer
        location( S ) ← location( P(q[i]) )
```

```
        }
    }
```

Figures 3 and 4 show examples of this implementation. Notice that some points sampled in the example of Figure 3 are located in cells with low density of population. This is because there is a small (but non-zero) probability for these points to be chosen – the alternatives implemented provide two approaches to determine the location of point data based on the spatial probability density defining the likely data distribution. Of course, these alternatives could be improved by increasing the accuracy of the underlying distribution data.

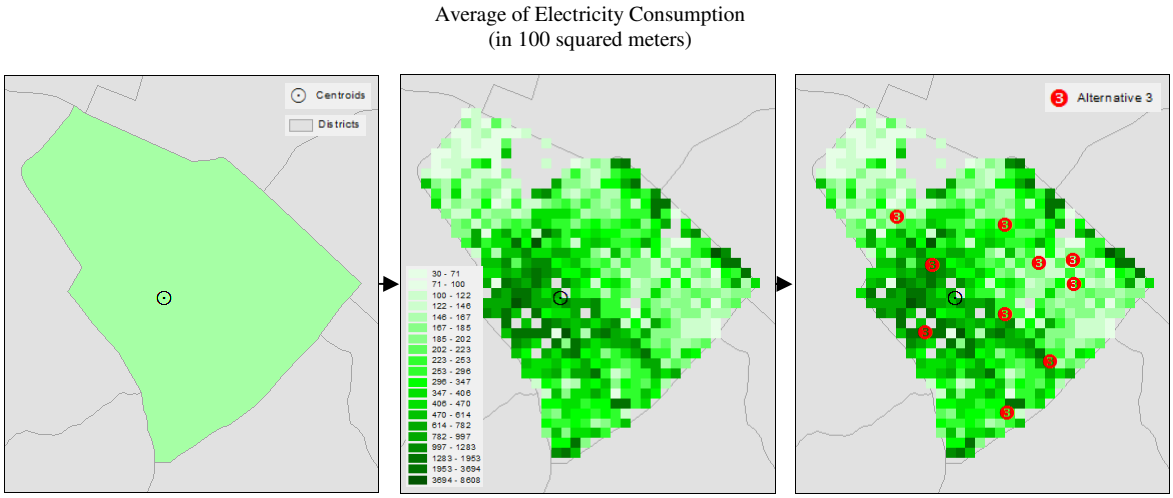

Figure 3: Example of Alternative 3 (generation of random points based on a surface of electricity consumption)

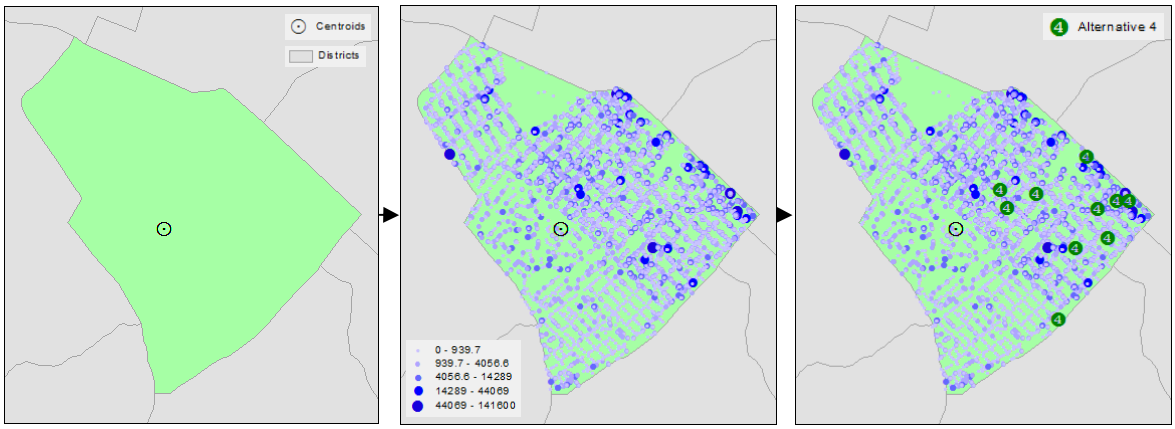

Figure 4: Example of Alternative 4 (selection of customer's location based on a probability of fitness for energy consumption)

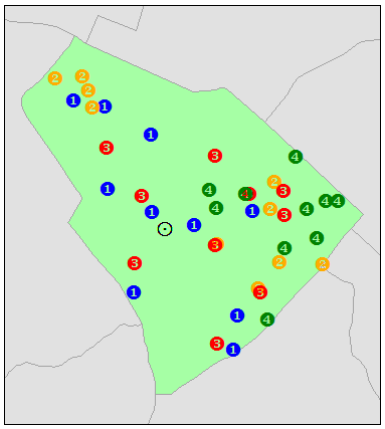

Figure 5: Overlap for Examples of 4 Alternatives

2.3 Implementation of GWR using Survey Data: original and alternatives

The alternative approaches to point-location allocation were tested by generating 1,000 iterations of GWR income-predicting model based on energy consumption for each alternative, for two minimisation approaches. Hence, 2,000 iterations for each alternative were computed: 1,000 considering the AIC minimisation local sample size suggested for each GWR model, and 1,000 considering one unique local sample size (suggested by the AIC minimisation of the original GWR) for all GWR models.

We compared the resulting performance for each alternative with the original GWR model (the null model that used the districts' centroids for point placement) and with the traditional aspatial Linear Model (LM). We named these original models as "GWR centr" and "LM centr", respectively.

In addition, we computed the average of income and of energy consumption per district and used this data set (with just one observation per district) as a simplified aggregation model with each data point associated with the centroid of the district. The linear regression and GWR models were also applied to these aggregated datasets. These models are referred to as "LM aggreg" and "GWR aggreg", respectively. Results and analyses are described in the next section.

3.0 RESULTS

3.1 Original GWR (district's centroids)

The ABRADEE's survey applied in the city of São Paulo in 2004 had 662 valid respondents. Income and Energy Consumption were collected as continuous variables - in R$ ("reais", Brazilian currency) and in kWh, respectively. Seventy-five (75) districts were sampled for this survey. Figure 6 shows the map of 96 districts of São Paulo (in gray), highlighting (in green) the 75 districts sampled for the ABRADEE's survey.

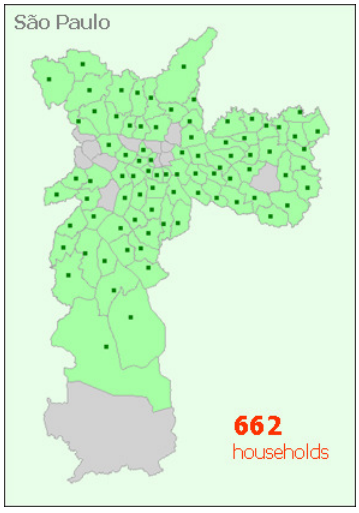

Figure 6: Map of São Paulo districts selected for ABRADEE's survey

3.2 Predictive response for GWR using each Alternative

Figures 7 and 8 summarizes the main point of this paper, and shows the dispersion of R^2 (coefficient of determination) for each alternative, through a box-plot of 1,000 computed iterations.

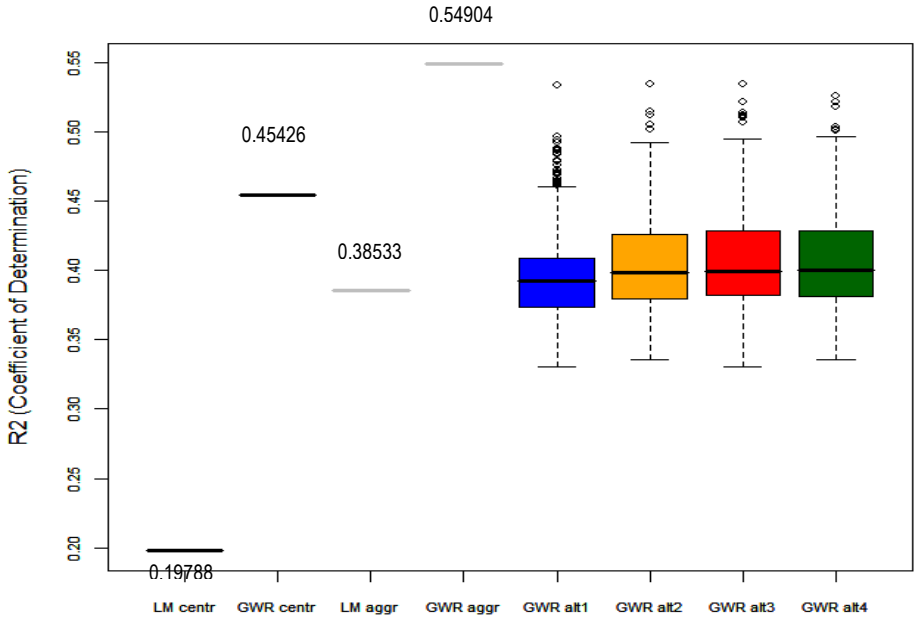

Figure 7: R^2 (and dispersion of $R^{2)}$ for GWR over 1,000 iteractions per alternative for each different regression applied in the ABRADEE's Survey.
Local Sample Size for GWR was determined by AIC minimisation for each iteraction.

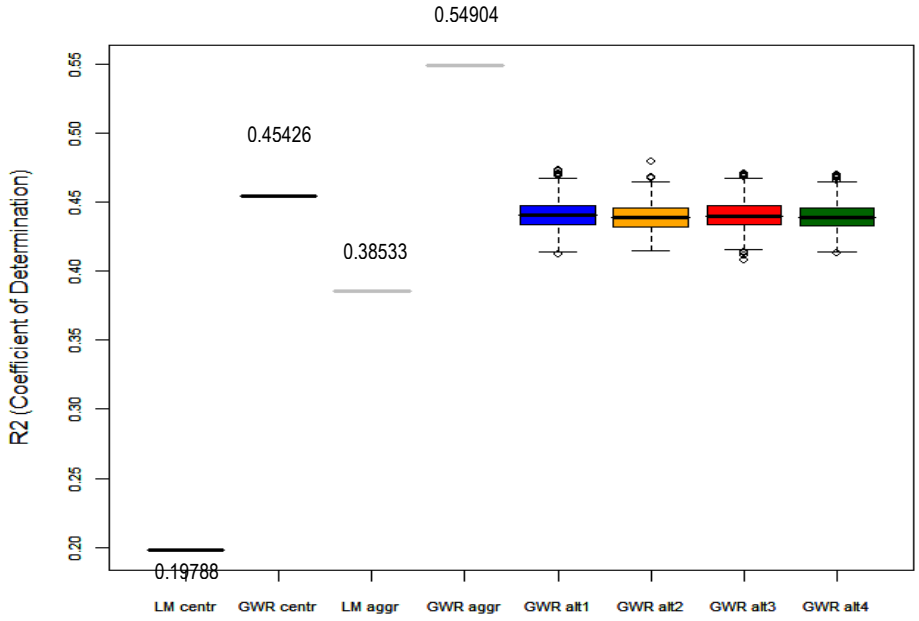

Figure 8: R^2 (and dispersion of $R^{2)}$ for GWR over 1,000 iteractions per alternative for each different regression applied in the ABRADEE's Survey
Local sample size = 0.0725
(suggested by AIC minimisation for GWR using centroids) in each iteraction.

Table 1 shows descriptive statistics of each alternative resulting from the GWR model.

Table 1: Descriptive Statistics of GWR results for 1,000 iterations of Alternatives 1 to 4

CONSIDERING different Local Sample Sizes (based on AIC minimisation)				
	Alternative 1	Alternative 2	Alternative 3	Alternative 4
R^2 – Coefficient of Determination				
Mean	0.39444	0.40373	0.40483	0.40549
Median	0.39210	0.39813	0.39936	0.39982
Standard Deviation	0.03068	0.03268	0.03341	0.03321
Bandwidth (Local Sample Size [k nearest neighbours]) – in percentage of total observations				
Mean	0.11899	0.10697	0.10792	0.10559
Median	0.11478	0.10726	0.10727	0.10572
Standard Deviation	0.02905	0.02726	0.02845	0.02802

CONSIDERING the same Local Sample Size (0.0725)				
	Alternative 1	Alternative 2	Alternative 3	Alternative 4
R^2 – Coefficient of Determination				
Mean	0.44027	0.43903	0.44080	0.43922
Median	0.44015	0.43868	0.43999	0.43863
Standard Deviation	0.01017	0.00962	0.01014	0.00964

Examining Table 1 for the different local sample sizes (from AIC minimisation) it is clear that there is little difference between any of the approaches, with the R^2 correlation coefficient for every alternative being in the range 0.39 - 0.40. Alternative 1 was the worst in magnitude (mean and median) with alternative 4 being the best. The dispersion (standard deviation) of R^2 was quite similar for all alternatives, with alternative 1 having the least variance and alternative 3 having the highest.

The variation of bandwidth between alternatives was basically opposite to the previous results, although once again all are very similar. The highest one was alternative 1 and the lowest one was alternative 4. Coefficient of determination and bandwidth were very highly correlated, which suggests a very important role for AIC minimisation. Figure 9 shows the scatter plot and the correlation of this relationship.

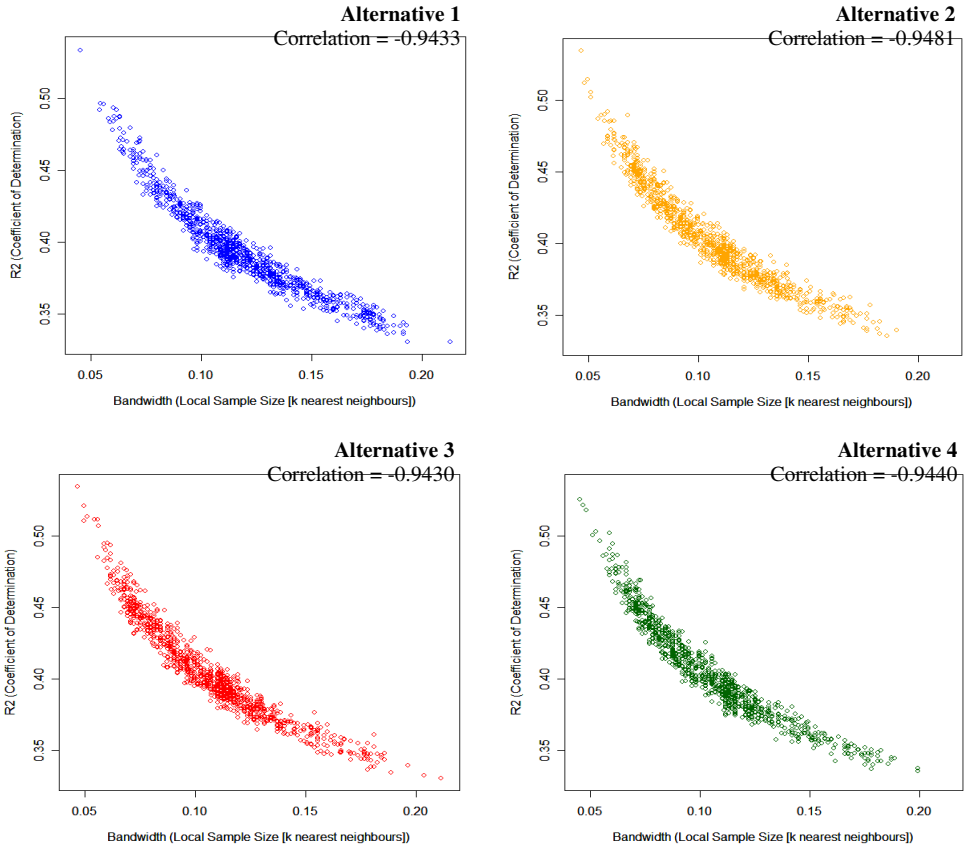

Figure X: Scatter-Plot and correlation of bandwidth (local sample size) and R^2 for 1,000 iterations for each point distribution allocation

Results for R^2 with the same bandwidth for every iteration in all alternatives is around 0.44 (mean and median) which is very similar to the centroids' GWR model results. The important role AIC minimisation has in this context for local models is shown by the variation in results when points are allocated with each model. The alternatives have been implemented so that bandwidth changes with each different point allocation in the space and this potentially affects the smoothness of the global model. Surprisingly for the authors, the resulting alternatives had lower R^2 than the simple centroid model, even though it was felt that the point-allocation was adding additional information to the data and therefore should have resulted in improvements in model prediction. This counter-intuitive result clearly must be further explored to understand in what ways the model behaviour and contribution to R^2 are interacting, in a similar manner to previous research into cross-validation properties in GWR (Farber & Paez 2007).

5.0 CONCLUSIONS

Although the R^2 for centroid placement of the point datasets is marginally (but significantly) higher, this result should be viewed with suspicion; we would regard the pseudo random point allocation results (the alternatives) as being the most realistic attempts to place the point data and therefore should result in more appropriate (and therefore improved prediction) models.

The implementation of the alternatives suggests the minimisation of AIC (Akaike Information Criterion) for every iteration, as AIC has an important role for the local sample regressions for GWR. The GWR models applied to the ABRADEE's survey led to a significant improvement to the explanation of variability for the income-predicting model based on energy consumption. The results of R^2 changed from 0.1978 to 0.4542 when the centroids were used as the point allocation of every interview which occurred in the districts.

However, R^2 measured for four alternatives of reallocation of points (households interviewed) inside districts' polygons showed R^2 of around 0.40, even though they should have resulted in potentially a better or more realistic relationship.

Future work is required to investigate the role of local bandwidth and its influence in the establishment of the point pattern itself and the R^2 and parameters' distribution results. The use of some important summary statistics for point patterns (e.g. F, G, K, J functions) should also be incorporated to help understand how the alternative methods behave. For example, it may be that the bandwidth discovered by the AIC minimisation is far larger then the neighbourhood distances within a district. This would result in the placement within a district not being a significant factor in the construction of the GWR model.

In conclusion, this work has presented an initial, but seminal, experimental study in the allocation of point patterns through secondary data. The resulting improvements using spatial point pattern statistics and mixed models have been shown to improve the prediction from a global model, however the approaches to point allocation inside polygons remains an open and significant problem.

RFERENCES

Back, T., D. Fogel & Z. Michalewicz 1997, *Handbook of Evolutionary Computation*, Oxford University Press and Institute of Physics Publishing, Bristol, City, .

Brunsdon, C., A. S. Fotheringham & M. Charlton (1998) Spatial nonstationarity and autoregressive models. *Environment and Planning A,* 30: pp. 957-973.

Farber, S. & A. Paez (2007) A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. *J. Geograph Syst,* 9: pp. 371-396.

Jelinski, D. & W. Jianguo (1996) The modifiable areal unit problem and implications for landscape ecology. *Landscape Ecology,* 11:3, pp. 129-140.

Openshaw, S. & P. Taylor 1979, 'A million or so correlation coefficients: three experiments on the modifiable areal unit problem.' in *Statistical Applications in the Spatial Sciences*, Ed N. Wrigley, Pion, London..

Yao, X. 1999, *Evolutionary Computation Theory and Applications*, World Scientific Publishing Co. Pty. Ltd., City, pp 360.