A Preliminary Investigation of the Stability of Geographically-Weighted Regression

Peter Whigham[1] & Geoff Hay[2]

[1]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

[2]Injury Prevention Research Unit
University of Otago, Dunedin, New Zealand

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

This paper describes preliminary work analysing the stability of parameter coefficient estimates for Geographically-Weighted Regression (GWR). Based on a large dataset (35721 points) various random samplings of this data were performed and models built using GWR. An analysis of the coefficient values for the independent variables showed that these values could varying significantly both between runs and between sampling sizes. This suggests that the results from GWR must be carefully considered in terms of the form of data, assumed coefficient surface being modelled, and the confidence of the resulting parameter estimates.

***Keywords and phrases*:** GWR, weighted regression, sampling, coefficient estimates

1.0 INTRODUCTION

Geographically weighted regression (Brundsdon, Fotheringham & Charlton 1996; Fotheringham, Brunsdon & Charlton 2002) has become a standard tool for constructing spatial models and to investigate spatial non-stationarity in point-based datasets. Although many of the examples using GWR are based on datasets with a relatively small geographic extent and few sampling points, this situation is changing with the advent of national datasets and more powerful computers. Clearly a model over a small geographic extent is not likely to show significant departures in coefficient values from one location to the next. However, as the spatial extent increases the probability that the local model coefficients will vary increases, especially when the factors that drive these models vary over space.

This paper will consider the stability of GWR in terms of the coefficient estimates as the sampling over the spatial extent increases. Stability will be described in terms of the range of values for each coefficient both between random samplings of the same size, and between different sampling sizes. The main hypotheses to be tested are: (1) that as sampling size increases the stability of parameter estimates increase; and (2) that although coefficient values will vary they will not change sign (from positive to negative, or vice-versa) .

This study is important since it will begin to address the issue of how confident the user of GWR can be when interpreting the coefficient values of the resulting model. A theoretical approach to addressing this problem could be achieved by generating coefficient surfaces with a range of properties and sampling the resulting surfaces to build the data points. However, for this preliminary study a real data-set will be used to indicate whether stability may be an issue when sampling over large spatial extents.

2.0 METHODS

A detailed description of the dataset used in this study is described in the paper by Hay et al. in this proceedings (Hay, Whigham, Kypri et al. 2007). Briefly, the location of current bar liquor licenses (bars, pubs, taverns, night

clubs and other non-restaurant on-licenses) active on the 6th March 2001 (census night) for all of New Zealand were obtained from the Ministry of Justice and geocoded using GeoStan Map software (Version 2.1.1, Critchlow Limited, 2006). The meshblock census data for 2001 for all of New Zealand was then used to construct a roadway distance measure to the nearest bar from the population-weighted centroid (Martin 1989; Pearce, Witten & Bartie 2006) of each meshblock. Meshblocks that did not contain a bar were excluded from the analysis, resulting in 35521 datapoints for the complete dataset. Two independent variables from the census data were used to construct the predictive model for distance to bars: each meshblock also had an associated deprivation index (Salmond & Crampton 2002) which measured the social and material deprivation on a scale from 1 (least deprived) to 10 (most deprived); and each meshblock was classified as to its degree of *urban influence* using the Statistics New Zealand Urban/Rural Profile Classification (Bayley & Goodyear 2001). This measure ranged from 1 (urban) through to 7 (rural), with values between these extremes indicating rural characteristics of the meshblock with varying degrees of urban influence.

This dataset was used to build local linear models which predicted the distance to the nearest bar for sampled meshblocks based on the independent variables of deprivation and urban/rural classification. The following GWR methods were used to construct the models:

I. **Fixed bandwidth determined by cross-validation:** This model used a cross-validation procedure to find the best predicting model based on using a fixed sized (in terms of distance) bandwidth which was applied at each data-point.

II. **Fixed adaptive bandwidth:** This model was given a value between zero and one representing the proportion of total data-points to be used in each local model. For example, a bandwidth of 0.1 for a sample size of 100 indicated that the nearest 10 data-points were used to construct the kernel when building the linear model at each data-point. This approach is termed adaptive since the size of the kernel (i.e. the geographic extent of the kernel) changes depending on the density of points surrounding the point where the model is being evaluated. For this paper bandwidth sizes of 0.1, 0.2, 0.5 and 0.9 were used.

For each GWR method described above ten (10) random samplings from the total point dataset were performed to produce datasets of size 50, 100, 500, 1000 and 2000. For each sampling and model type the coefficients were analysed to address the previously stated hypotheses. Although the number of sampling does not give a strong statistical power the results should be adequate for this pilot study to consider whether further, more rigorous, analyses is justified. The "R" programming language (Ihaka & Gentleman 1996) and associated GWR libraries were used for all experiments.

3.0 RESULTS

The results will be presented as a series of figures, commencing with the fixed kernel experiments. For each panel the coefficient densities for each run, and the associated box-plots for the coefficients, are shown. Note that since there is some local spatial variation in the model structure is means that there is variation in the coefficient values for any particular run. However, the notion of stability refers to the variation observed between runs (for a given sample size), and therefore the main consideration is whether there is significant variation between coefficients across all sampling runs.

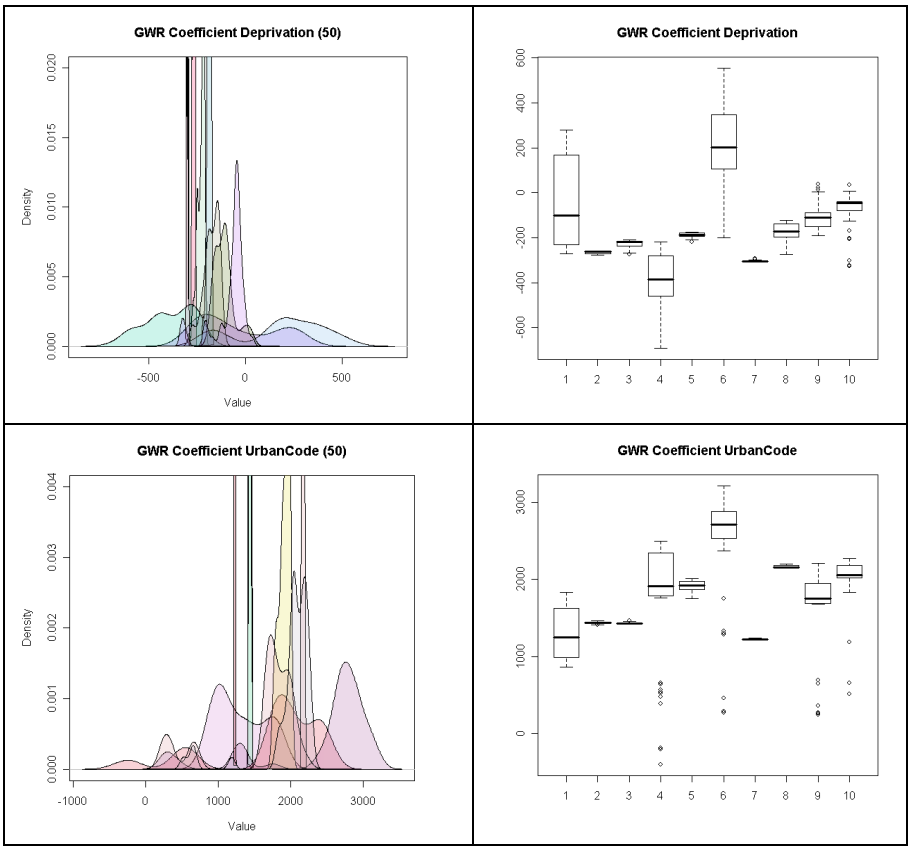

Figure 1. Coefficient ranges for sample size 50.

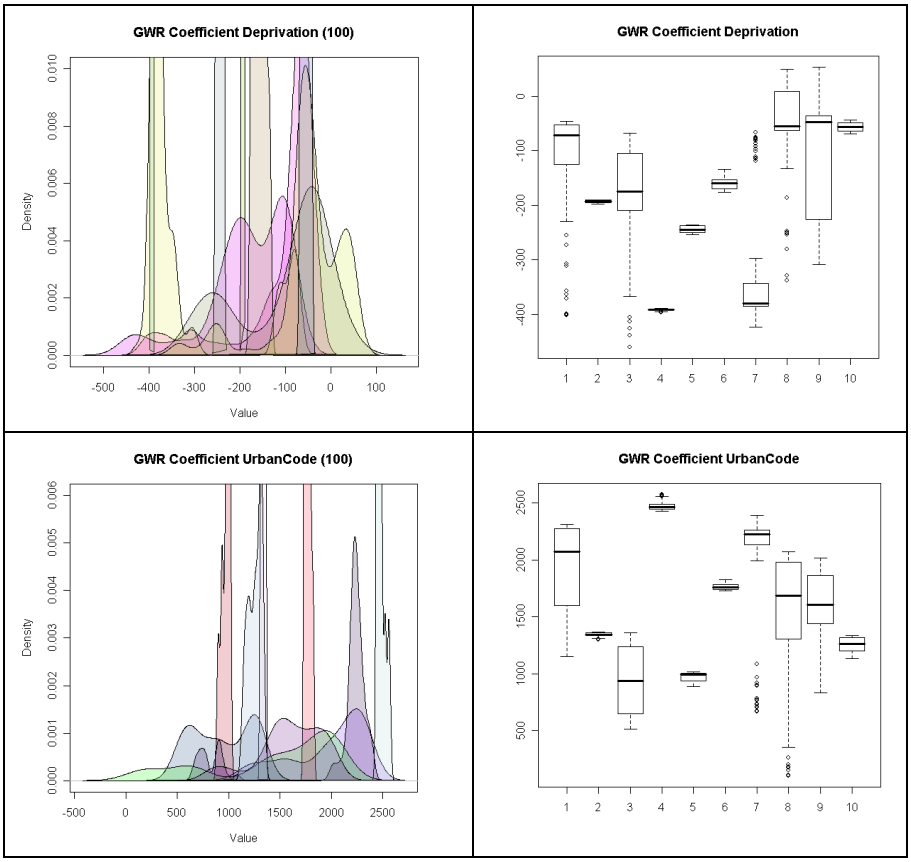

Figure 2. Coefficient ranges for sample size 100.

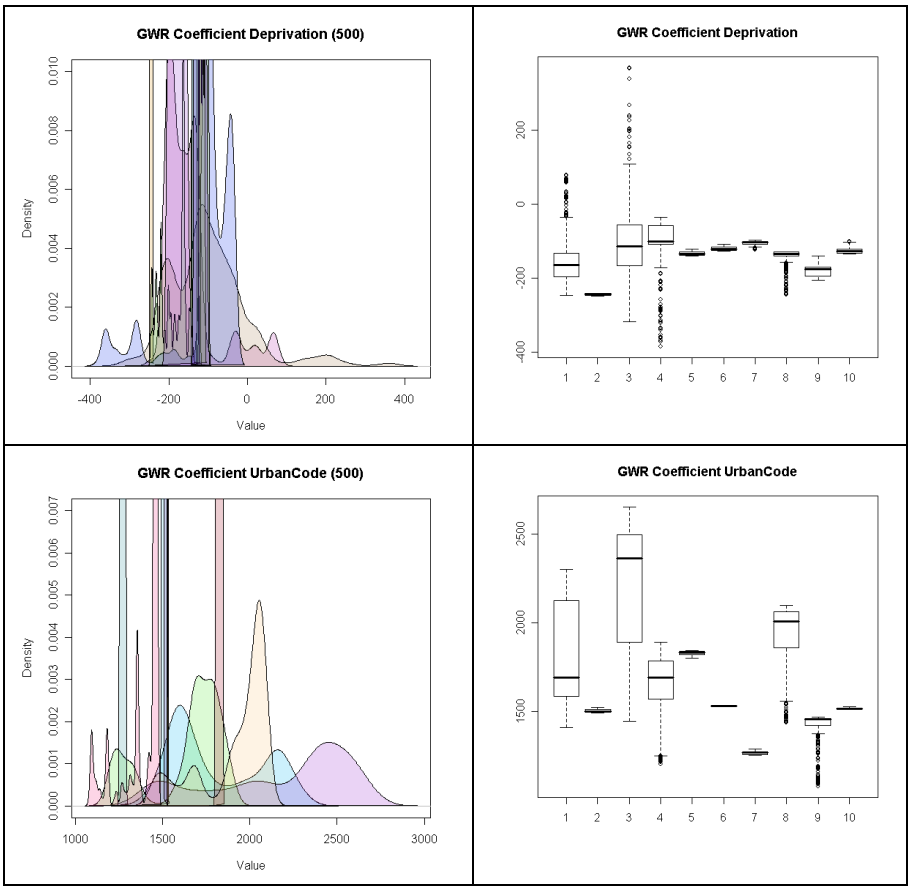

Figure 3. Coefficient ranges for sample size 500.

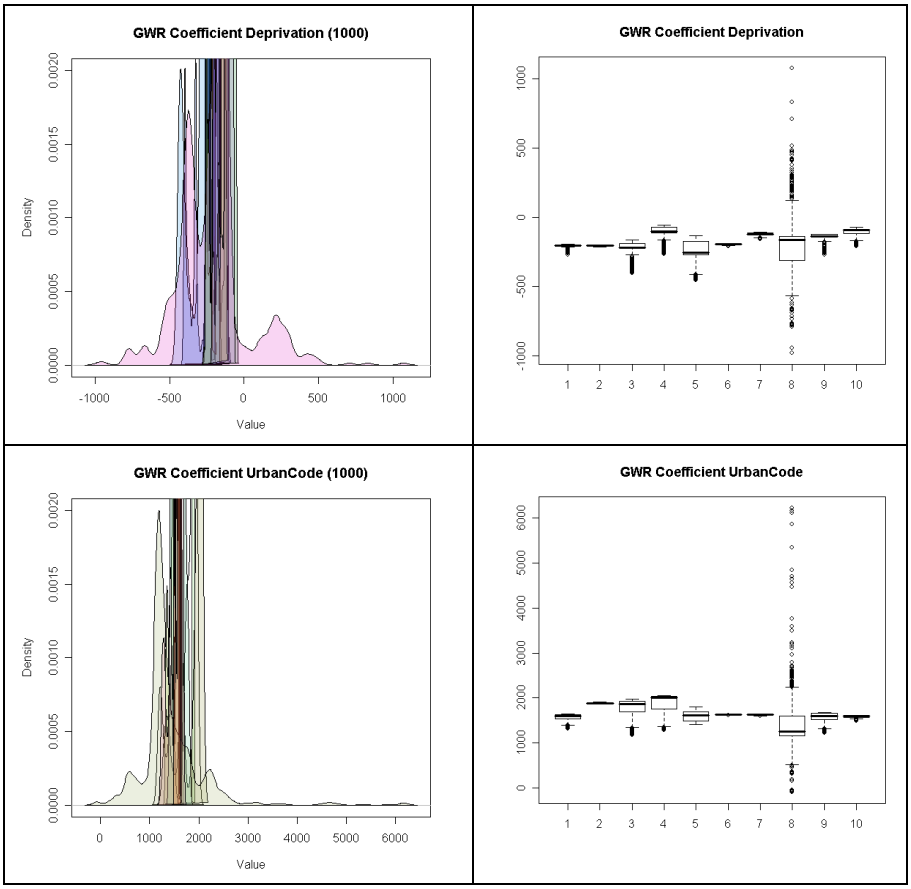

Figure 4. Coefficient ranges for sample size 1000.

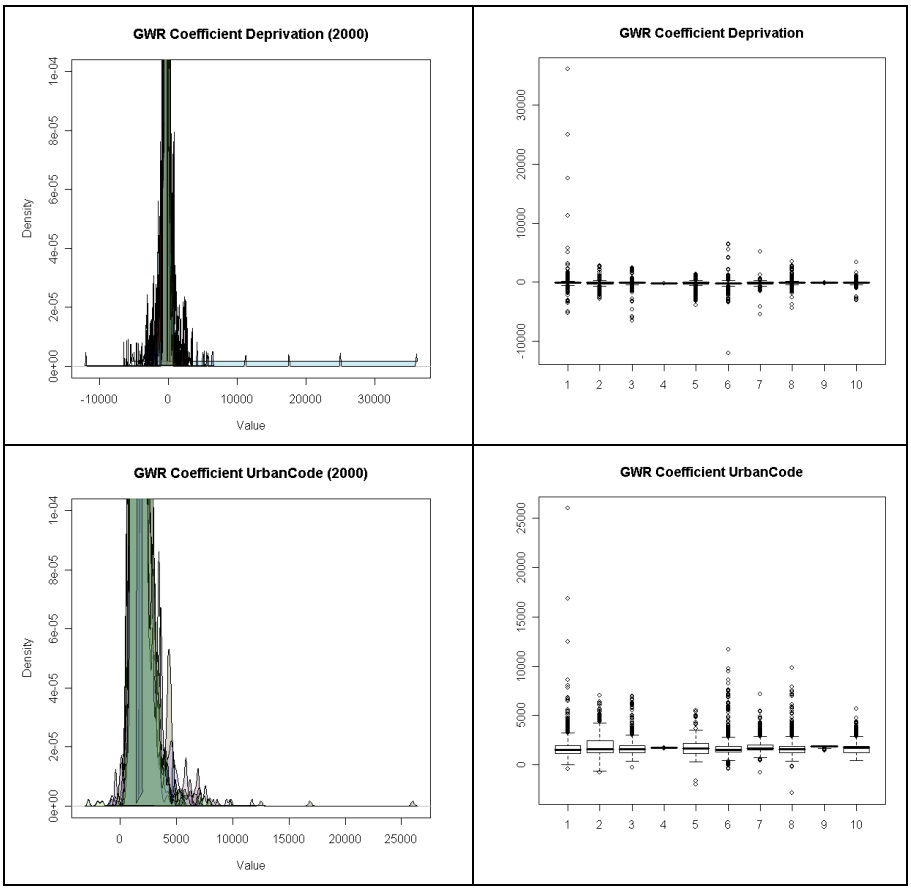

Figure 5. Coefficient ranges for sample size 2000.

3.1 Fixed Kernel Results

Figures 1-5 support the conjecture that increasing the sampling size will reduce the variance in the estimated coefficient values. Figure 5 shows the greatest stability, although run#1 in this figure suggests that there is greater spatial non-stationarity for the UrbanCode coefficient in the model then the other nine runs. Since the point locations were randomly sampled, this would suggest that a larger number of rural areas happen to have occurred in run#1, which has resulted in the increased distance response of this variable.

It is not surprising that for small sample sizes the variation in coefficient values is high. Note from Figures 1-5 that as the sample size increases the average value of each coefficient becomes closer for each run. However, even when the sampling size is 2000 there is notable variation: runs 1,6 & 8 show a greater variation in the UrbanCode coefficient which would be interpreted as greater spatial non-stationarity for this variable. Since typically the user is just given one sampling size for the set of data, this implies that it would be possible to interpret the corresponding GWR model as either stationary or non-stationary, even though the underlying true relationship could be either. Clearly the sample design is a significant contributor to the interpretation of results with GWR.

3.2 Adaptive Kernel Results

The results when using the adaptive kernel setting were similar to those of Figures 1-5, and therefore we will only show a few results to highlight unusual aspects of the results. Figures 6 ,7 and 8 show the coefficient values for a range of bandwidths (0.1,0.2,0.5,0.9) for sample sizes of 500, 1000 and 2000 respectively. There is clearly a pattern of increasing variation in coefficient estimates as the bandwidth increases. This shows that although a larger proportion of the points are used for each local regression, the sampling has produced a variety of quite dissimilar models. When bandwidth is small (for example, with a bandwidth of 0.1 the nearest 50 points are used to construct the model for a sample size of 500) the average parameter estimates are much closer to each other

for all runs. In addition for some runs with this small bandwidth spatial non-stationarity is more pronounced. Since the use of cross-validation or an AIC measure is often used to avoid overfitting of the local models, the bandwidth is typically of the order of 0.1-0.2. Hence this would imply that if the data is not being sampled at the appropriate rate for the underlying process that the results are likely to be highly variable and therefore not easily interpreted.

The same pattern of increased variation as bandwidth increases in shown in Figures 7 & 8 for the sample sizes of 1000 and 2000 respectively. This is most evident with a bandwidth of 0.9: here the deprivation coefficient could either be found to have almost no spatial variation (runs 6,9,10) or could have low large negative values with significant spatial variation (for example, run 7). Similar patterns can be observed for the UrbanCode coefficient estimates.

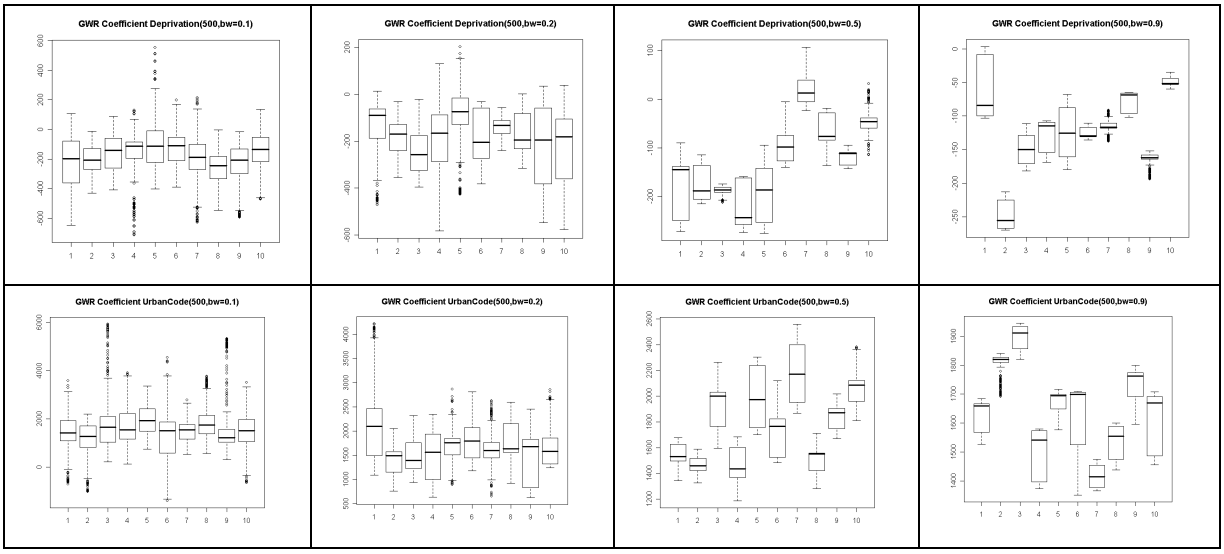

Figure 6. Deprivation and UrbanCode coefficients for increasing bandwidth (Sample size = 500)

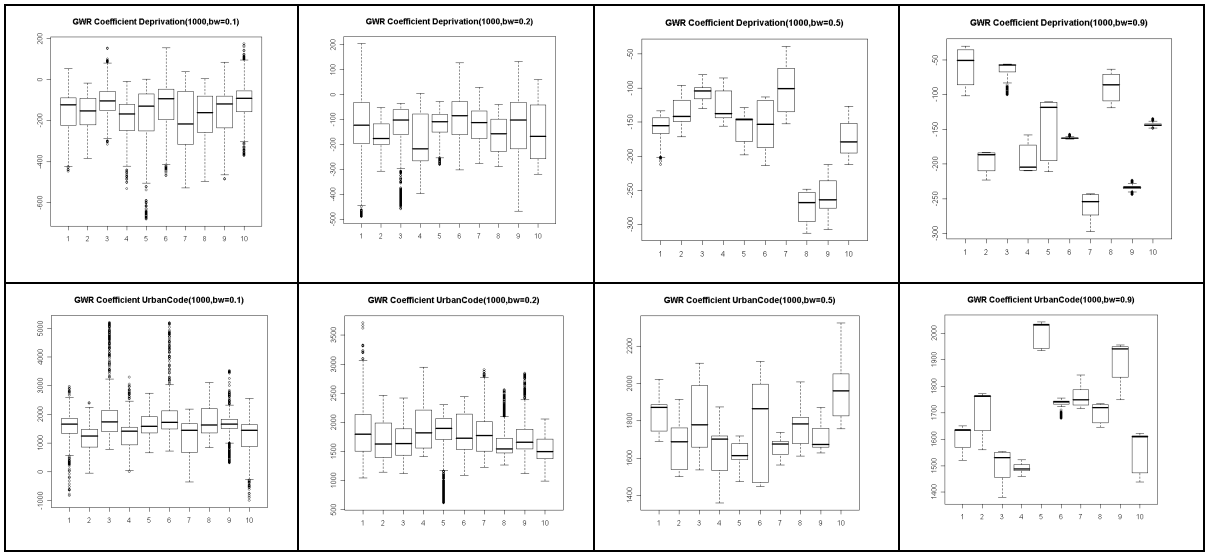

Figure 7. Deprivation and UrbanCode coefficients for increasing bandwidth (Sample size = 1000)

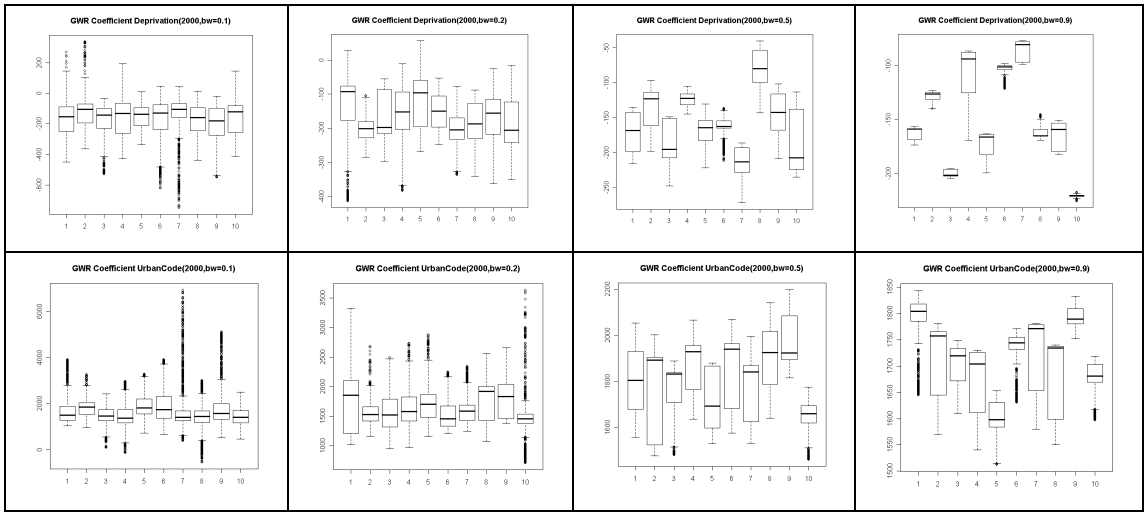

Figure 8. Deprivation and UrbanCode coefficients for increasing bandwidth (Sample size = 2000)

Since the observed variation in parameters may be due to the nature of the sampling, it would be expected that building global models based on the sampled data would also exhibit variation in the parameters. To ascertain whether this is the case the final figure shows the results for a global linear regression model and parameter estimates. For each sample size (50,100,500,1000,2000,5000,10000) there were 100 trials. The box plots show the variation for the intercept, deprivation and urbancode coefficient estimates.

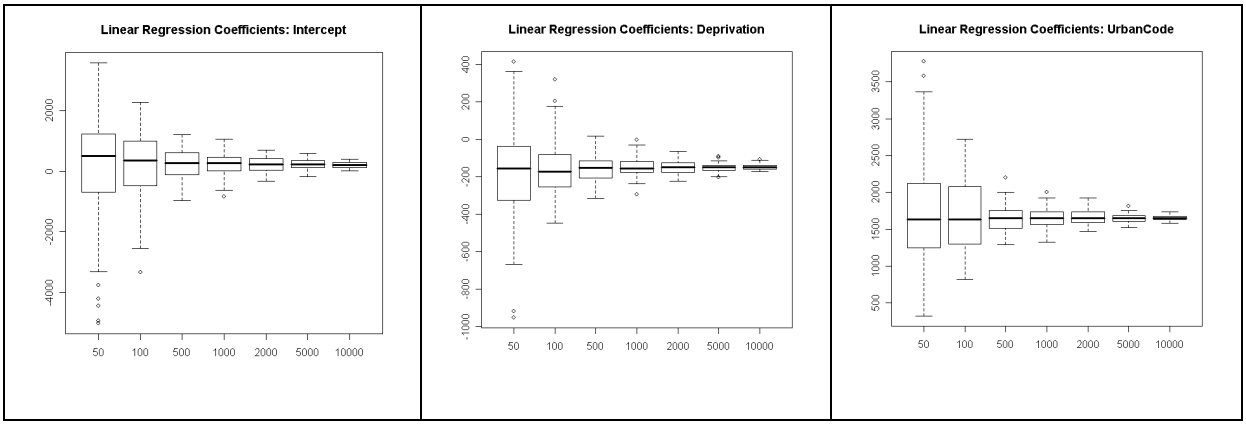

Figure 8. Global Linear Regression Coefficient Estimates for a range of sample sizes

Although there is variation for small sample sizes with the global regression, the average values of the coefficient estimates are relatively constant. This is in stark contrast to the variation in the average values for the fixed bandwidth runs for bandwidth sizes of 0.5 or greater. Clearly the local variation in sample points outweighs the global measure of these points. This suggests that the modeled response surface may show a large amount of variability locally, which is driving the values of the GWR coefficients. Since these estimates vary significantly between samplings it does raise concerns regarding how best to measure the effect of sampling of the resulting GWR model and its interpretation.

DISCUSSION

Previous research into the properties of GWR have considered the role and sensitivity of cross-validation (Farber & Paez 2007). In this work it was shown that a small number of outliers can often drive the cross-validation process and therefore produce results that are not consistent with the majority of the dataset. The work here has

shown some preliminary investigations of variation in coefficient estimates based on sampling rate for one dataset. Of course, a single dataset and few trials means that the conclusions of the work cannot be generalized, however it certainly supports the conjecture that GWR results must be carefully interpreted to avoid artifacts due to the manner in which sampling of the point data has occurred. In particular, the variation due to changes in bandwidth versus the stability of coefficients for global models, suggests that once we move from a global to local model the sampling regime becomes more critical.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Ministry of Justice for providing liquor license data, Jamie Pearce from Canterbury University for supplying the population-weighted centroid data, and the IPRU (Univ. Otago) for permission to use the outlet meshblock data in this study.

REFERENCES

Bayley, A. & R. Goodyear (2001) *An Urban/Rural Profile,* Statistics New Zealand, .

Brundsdon, C., A. S. Fotheringham & M. Charlton (1996) Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis,* 28: pp. 281-289.

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.

Fotheringham, A. S., C. Brunsdon & M. Charlton (2002) *Geographically Weighted Regression: the analysis of spatially varying relationships,* John Wiley & Sons, Ltd, .

Hay, G., P. Whigham, K. Kypri & J. Langley 2007, 'Spatial variation in the association between neighbourhood deprivation and access to alcohol outlets', in *Proceedings of the 19th Annual Colloquium of the Spatial Information Research Centre*, Ed P. Whigham, Otago University Print, Dunedin, New Zealand.

Ihaka, R. & R. Gentleman (1996) R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics,* 5:3, pp. 299-314.

Martin, D. (1989) Mapping Population Data from Zone Centroid Locations. *Transactions of the Institute of British Geographers,* 14:1, pp. 90-97.

Pearce, J., K. Witten & P. Bartie (2006) Neighbourhoods and health: a GIS approach to measuring community resource accessibility. *J Epidemiol Community Health,* 60: pp. 389-395.

Salmond, C. & P. Crampton 2002, *NZDep Index of Deprivation*, Department of Public Health, Wellington School of Medicine and Health Sciences, City, .