Testing geographically weighted multicollinearity diagnostics

Stamatis Kalogirou[1]

[1]Harokopio University, Department of Geography, El. Venizelou 70, Kallithea, 17676 Athens, Greece

Tel. +30 210 9549163 Fax +30 210 9514759

skalo@hua.gr, http://www.geo.hua.gr

KEYWORDS: spatial analysis, local correlation coefficient, geographically weighted regression, multiple hypotheses testing

1. Introduction and Background

The increasing application of local methods of explanatory spatial data analysis such as local regression methods motivates the need for local statistical inference diagnostics. The development of techniques that examine the existence of local multicollinearity is an open field for research in spatial analysis. Recent contributions by Brunsdon (2009) and Wheeler and Paez (2010) stimulated the testing of a simple local multicollinearity diagnostic; this is a local version of Pearson Correlation Coefficient (Kalogirou 2010a, 2011).

This paper tests the Geographically Weighted Pearson Correlation Coefficient (GWPCC). The definition of a geographically weighted correlation coefficient has already been described in Fotheringham et al. (2002). The GWPCC here has been calculated for three pairs of explanatory variables of mean recorded household income in Local Authorities in Greece in order to examine the existence of local multicollinearity in these variables. This is a necessary test for the corresponding geographically weighted regression model. The GWPCC is a simpler approach for checking multicollinearity among explanatory variables in local regression compared to the local diagnostics tools such as the VIF presented by Wheeler (Wheeler, 2006; 2007). A Geographically Weighted VIF could also be developed for the purpose of this paper.

However, the lack of off-the-self software for computing local multicollinearity diagnostics in local regression methods has been recognised by Wheeler and Tiefelsdorf (2005) as well as by Wheeler and Paez (2010). Thus, for the calculation of the GWPCC it was necessary to develop a computer program (LC Tools) an alpha version of which is available at http://gisc.gr/index.php/lctools or upon request to the author of this paper. It is necessary to note that the *Geographically weighted regression with penalties and diagnostic tools (gwcc)* developed by David Wheeler was recently made available as a package in R (http://cran.r-project.org/web/packages/gwrr/index.html).

The suggested multicollinearity diagnostics has been applied to test the potential multicollinearity of the determinants of mean recorded household income in Greece. The income model configuration is based on the findings of previous work (Kalogirou and Hatzichristos, 2007).

2. Data

The model defined for the application of the suggested diagnostics refers to the determinants of the mean recorded household income earned during the calendar year 2001. The latter refers to the average income recorded in tax forms for each postcode aggregated to the Local Authority level. The term “recorded” in this case refers to the household income declared in the tax form inflated according

to the wealth of the household, such as properties and equities (Kalogirou and Hatzichristos 2007). The source of income data is the General Secretariat for Information Systems (GSIS) of the Ministry of Finance of the Hellenic Republic. The postcode data have been initially aggregated to the 1033 Local Authority geography (Kalogirou 2010b) and then further aggregated according to the new geography of the 325 Municipalities in Greece. The latter is the result of a local government restructuring law called *The New Architecture of the Local Government and Decentralised Governance - Program Kallikratis*. A choropleth map of mean recorded household income is presented in Figure 1.

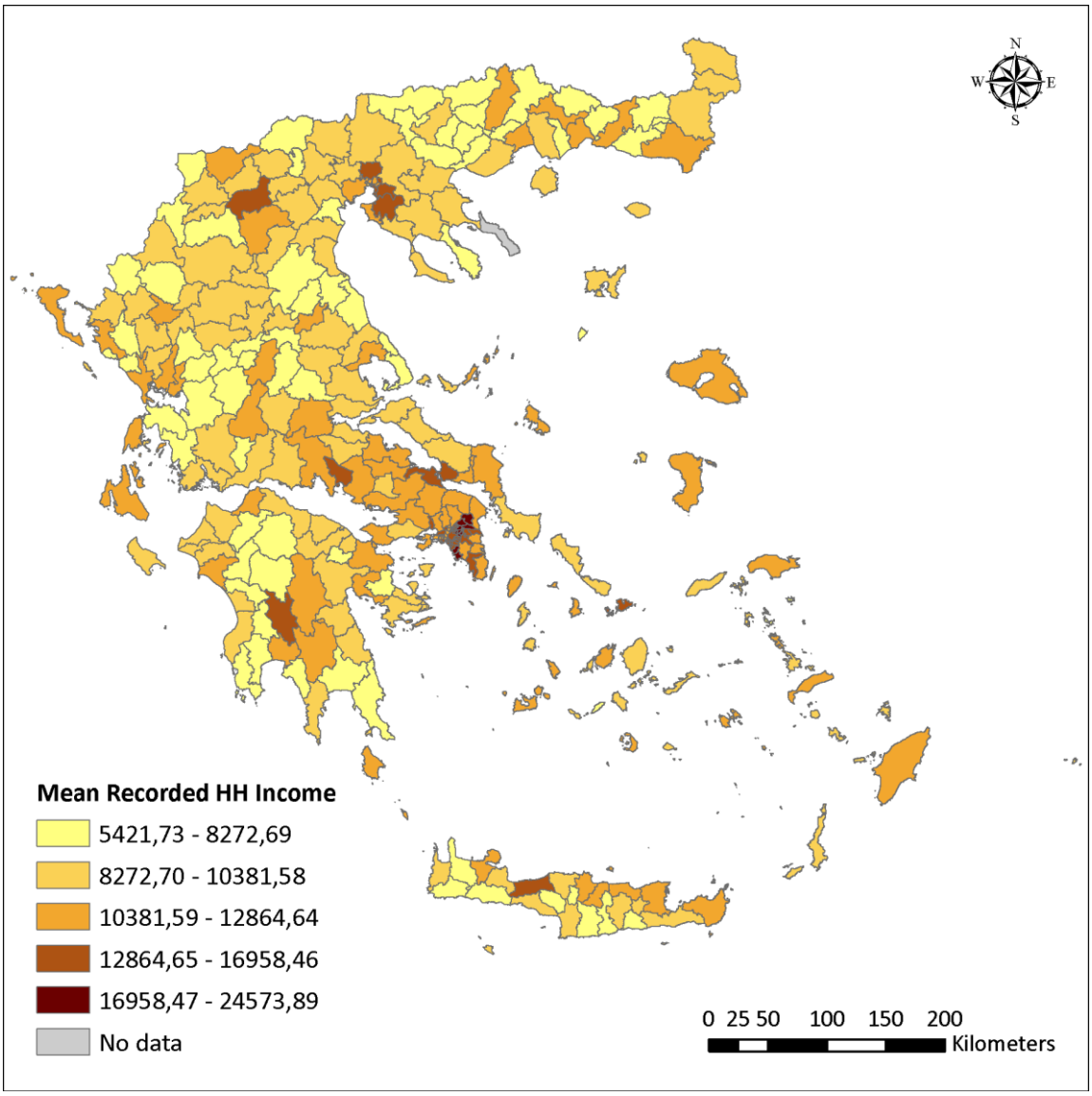

Figure 1. Map of mean recorded household income at the local authority level in Greece

The three determinants of income include: the proportion of people with high educational attainment (this includes college, university and postgraduate and PhD degree holders); the proportion of people working in agriculture and fishing related industries (NACE A and B); and the total unemployment rate. The data source of these determinants is the 2001 Census for Population in Greece.

3. Methodology

The methodology adopted here for modelling the mean recorded household income at a local level has been recently presented by Kalogirou and Hatzichristos (2007). A short description of the Local

Pearson Correlation Coefficient and the Geographically Weighted Pearson Correlation Coefficient follows.

3.1 Local Pearson Correlation

The Pearson Correlation Coefficient *r* is a standard statistic for checking for multicollinearity in the independent variables of a linear regression model calibrated using Ordinary Least Squares (OLS). The formula to calculate *r* in order to examine for the correlation between two variables *X* and *Y* that have a normal distribution, mean values of $\bar{x}$ and $\bar{y}$, and standard deviations of s_x and s_y, respectively, is:

$$r = \frac{\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)}{n-1} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{(n-1)}}\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{(n-1)}}} \quad (1)$$

where *n* is the number of observations. The equation can also be written as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (2)$$

The coefficient *r* is statistically significant at a given significance level *a* if the absolute value of *t* given by equation 3 is higher than the value of the two tailed t-student distribution for *n-2* degrees of freedom and significance level *a*. The formula of *t* is:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \Rightarrow t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (3)$$

where *n* is the number of observations (Kalogirou, 2010a).

The Local Pearson Correlation Coefficient (LPCC) for each observation point *i* is calculated as follows:

$$r_i = \frac{\sum_{j=1}^{k}(x_j-\bar{x}_i)(y_j-\bar{y}_i)}{\sqrt{\sum_{j=1}^{k}(x_j-\bar{x}_i)^2}\sqrt{\sum_{j=1}^{k}(y_j-\bar{y}_i)^2}} \quad (4)$$

where *k* is the number of nearest neighbours around observation point *i*, $\bar{x}_i$ and $\bar{y}_i$ are the mean values of *x*s and *y*s of the *k* nearest neighbours of *i*. For example, $\bar{x}_i = \sum_{j=1}^{k} x_j \big/ k$.

The number of nearest neighbours can be determined a priory or could be based on the optimal number resulted in after the application of Geographically Weighted Regression (GWR) analysis using an adaptive kernel (Fotheringham et al, 2002).

3.2 Geographically Weighted Pearson Correlation

The GWPCC is a geographically weighted moment-based statistic that adopts the idea of geographical weighting of the values around an observation for which local statistics are calculated (Fotheringham et al. 2002). The formula to calculate $gwpcc_i$ in order to examine for local correlation between two variables X and Y that have a normal distribution, geographically weighted mean values of $\bar{x}_i = \sum_{j=1}^{k} x_j w_{ij} \Big/ \sum_{j=1}^{k} w_{ij}$ and $\bar{y}_i = \sum_{j=1}^{k} y_j w_{ij} \Big/ \sum_{j=1}^{k} w_{ij}$, and geographically weighted standard deviations of s_{xi} and s_{yi}, respectively, is:

$$gwpcc_i = \frac{\sum_{j=1}^{k} w_{ij}(x_j - \bar{x}_i)(y_j - \bar{y}_i)}{\sqrt{\sum_{j=1}^{k} w_{ij}\left(x_j - \bar{x}_i\right)^2}\sqrt{\sum_{j=1}^{k} w_{ij}\left(y_j - \bar{y}_i\right)^2}} \quad (5)$$

The function to calculate the weights w_{ij} in equation 5 is a bi-square function usually adopted as an adaptive kernel weighting scheme in GWR (Fotheringham et al., 2002). The corresponding formula is:

$$w_{ij} = \begin{cases} [1-(d_{ij}/h_i)^2]^2 \text{ if } d_{ij} \le h_i \\ 0 \text{ otherwise} \end{cases} \quad (6)$$

where d_{ij} is the distance of each of the k nearest neighbours of i and h_i is the distance of the most distant neighbour.

Fotheringham et al. (2002) recognise that a geographically weighted correlation coefficient is a useful diagnostic tool for GWR calibrated models. It could be argued that the GWPCC is a more appropriate diagnostic for testing for multicollinearity among the explanatory variables in a GWR model than the LPCC.

4. Regression results and statistical inference

The results of the calibration of both global and local income estimation models are presented in Table 1. It is necessary to note that the model configuration is simple and does not account for all possible explanatory variables for which data is available. This is the case because the aim of this paper is to examine for the existence of local multicollinearity among income determinants and not to fit a well defined income estimation model. However, the three determinants of income alone explain 87% of its variation. The goodness of fit statistics by means of the Adjusted R^2 and the AIC_C indicate that the GWR model has a higher explanatory power than the global OLS model.

The proportion of people with high educational attainment has a strong positive effect on mean recorded household income at the local authority level. Both the proportion of those working in the primary sector of production and the total unemployment rate, have a significantly negative effect on mean recorded income.

Table 1. Global and local regression results

Variable	OLS Parameter Estimates	t-student test (Sig.)	GWR Parameter Estimates interval	Monte Carlo test
Constant	8671.367	27.497 (0. 000)	6952.163 - 11731.029	0.600
High education attainment	316.993	25.123 (0.000)	122.921 - 402.149	0.230

Working in Agriculture and Fisheries	-39.592	-8.948 (0.000)	-77.691 - -15.427	0.200
Unemployment rate	-95.774	-4.447 (0.000)	-436.050 - 30.541	0.000
R^2		0.872		0.908
Adjusted R^2		0.871		0.898
AIC_c		5454.028		5410.359
Observations /Nearest Neighbours		325		101

Table 2 presents the global Pearson Correlation Coefficients between the three explanatory variables and the corresponding significance p-values (in parenthesis). For the two out of three possible pairs of variables there is an apparent significant correlation. However, with the exception of the correlation between high educational attainment and the proportion of people working in the primary sector, the correlation coefficients are rather low. The VIF values are low for all three variables suggesting the lack of significant local multicollinearity. It could be argued that the independence criterion for the OLS regression is satisfied.

Table 2. Global Pearson Correlation Coefficients

	High education attainment	**Primary Sector**
Primary Sector	-0.683 (.000)**	
Unemployment rate	-0.053 (.340)	-0.266(.000)**

However, the LPCCs and the corresponding GWPCCs calculated for the location of each observation assuming 101 nearest neighbours (this number of nearest neighbours minimizes the AIC_c in the corresponding GWR analysis) show a different picture. The averaging nature of global statistical analysis of spatial data can be demonstrated by the local coefficients between high education attainment and unemployment rate. The global coefficient is -0.053 whereas the LPCCs range from -0.327 to 0.295 (Figure 2) and the GWPCCs range from -0.380 - 0.382 (Figure 3).

A summary of the LPCCs is presented in Table 3a and a summary of the GWPCCs is presented in Table 4a. In each cell, the range of the local coefficients is presented. If we assume that each t-student test is independent and the corresponding data do not exhibit a significant spatial autocorrelation, then a local coefficient is significant at the 95% confidence level for 99 degrees of freedom (two-tailed) if the calculated value of the local t-student test is equal or lower than -1.9842 or equal or higher than 1.9842 (in order to reject the null hypothesis that the correlation coefficient is 0).

However, there are several problems when spatial data and multiple hypothesis testing are concerned. In order to account for the latter, one could use a Bonferroni (1935), a Sidak (1967) or a Holm (1979) adjustment that requires a higher critical value for t. In the Bonferroni approach $p_{critical} = a'/n$; in the Sidak approach $p_{critical} = 1-(1-a)^{1/n}$; and in the Holm approach $p_{critical} = a/(n-i+1)$ where *n* is the number of tests and *i* refers to the *i*-th test. In this paper, in order for the family wise error for the 325 hypothesis tests to be 0.05, each null hypothesis is rejected if $p \leq 0.0001538$ ($t \geq 3.9365$) in the Bonferroni approach; $p \leq 0.0001578$ ($t \geq 3.9294$) in the Sidak approach; and $p_i = 0.05/(326-i)$ in the Holm approach, respectively (two-tailed).

Brunsdon and Charlton (2011) review the practice of the above multiple hypothesis testing along with the BH approach proposed by Benjamini and Hochberg (1995) and the BH2S (a BH two step approach) proposed by Benjamini et al (2006). The cases in which the LPCCs and the corresponding GWPCCs are significant are presented in Tables 3a and 4a, respectively.

Table 3a. Local Pearson Correlation Coefficients

	High education attainment	Primary Sector
Primary Sector	-0.776 - -0.604	
Unemployment rate	-0.327 - 0.295	-0.619 - -0.003

Table 3b. LPCCs significance

Times H_o is Rejected	High education attainment	Primary Sector
Primary Sector	Unadjusted = 325 Bonferroni = 325 Sidak = 325 Holm = 325 BH (FDR) = 325 BH2S = 325 (p_{max}=2.37E-11)	
Unemployment rate	Unadjusted = 156 Bonferroni = 0 Sidak = 0 Holm = 0 BH (FDR) = 143 ($p_{critical}$ = 0.0189) BH2S = 153 ($p_{critical}$ = 0.0386)	Unadjusted = 197 Bonferroni = 93 Sidak = 93 Holm = 143 BH (FDR) = 189 ($p_{critical}$=0.0188) BH2S = 199 ($p_{critical}$ = 0.0657)

Table 4a. Geographically Weighted Pearson Correlation Coefficients

	High education attainment	Primary Sector
Primary Sector	-0.795 - -0.557	
Unemployment rate	-0.380 - 0.382	-0.699 - 0.049

Table 4b. GWPCCs significance

Times H_o is Rejected	High education attainment	Primary Sector
Primary Sector	Unadjusted = 325 Bonferroni = 325 Sidak = 325 Holm = 325 BH (FDR) = 325 BH2S = 325 (p_{max}= 1.42E-09)	
Unemployment rate	Unadjusted = 161 Bonferroni = 54 Sidak = 58 Holm = 63 BH (FDR) = 129 ($p_{critical}$ = 0.0185) BH2S = 156 ($p_{critical}$ = 0.0376)	Unadjusted = 231 Bonferroni = 133 Sidak = 133 Holm = 143 BH (FDR) = 229 ($p_{critical}$ = 0.0346) BH2S = 238 ($p_{critical}$ = 0.0851)

It is interesting to map all pairs of variables for which the Local and the Geographically Weighted Pearson Correlation Coefficients vary spatially in terms of magnitude and significance. This would allow for the understanding of their spatial patterns. Figures 2 and 3 present a map of the LPCCs and a map of the GWPCCs as well as the corresponding local p-values for the local correlation between *high educational attainment* and *total unemployment rate*, respectively. The two maps show similar spatial patterns. However, there is a difference between the LPCC and the GWPCC for a given location *i*. In several cases the GWPCCs found to be significant and of higher magnitude than the LPCCs. There are several cases where the GWPCCs are significant after adjusting for multiple hypotheses testing, especially using the Bonferroni, Sidak and Holm approaches that are rather conservative. Figure 4 presents a scatter plot of the pairs of LPCCs and GWPCCs of the local correlation between *high educational attainment* and *total unemployment rate*. Theoretically, the GWPCC is a more appropriate diagnostic for local multicollinearity in a GWR model. It uses the weighting scheme for the data calibrated using GWR which accounts for spatial autocorrelation typically found in spatial data.

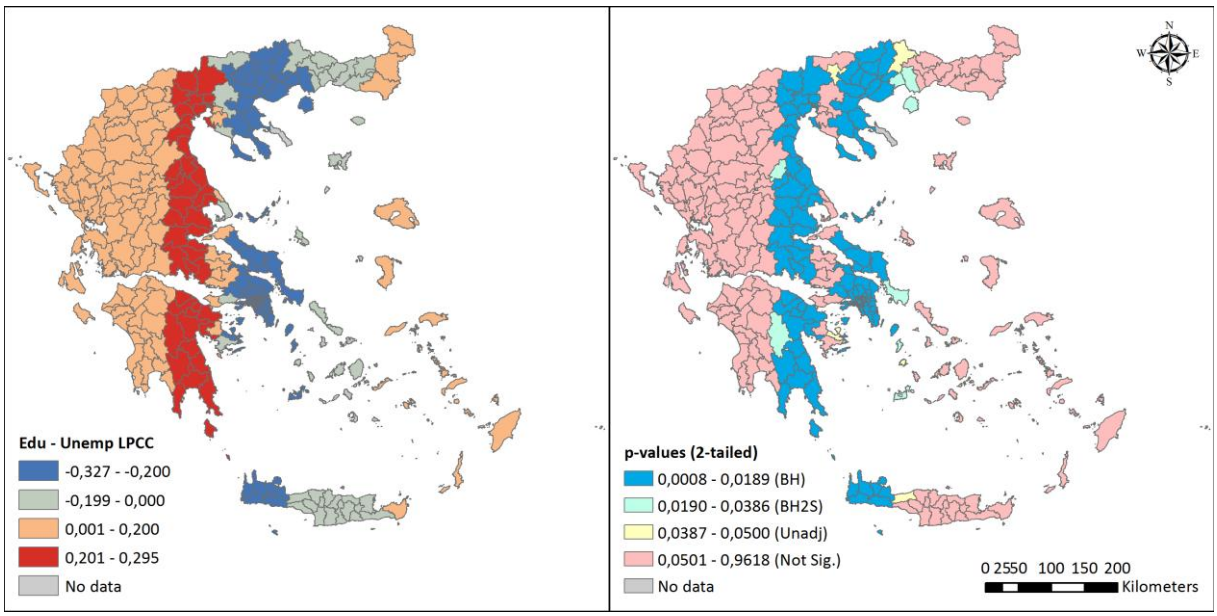

Figure 2. Maps of Local Pearson Correlation coefficients between high educational attainment and unemployment rate and the corresponding p-values

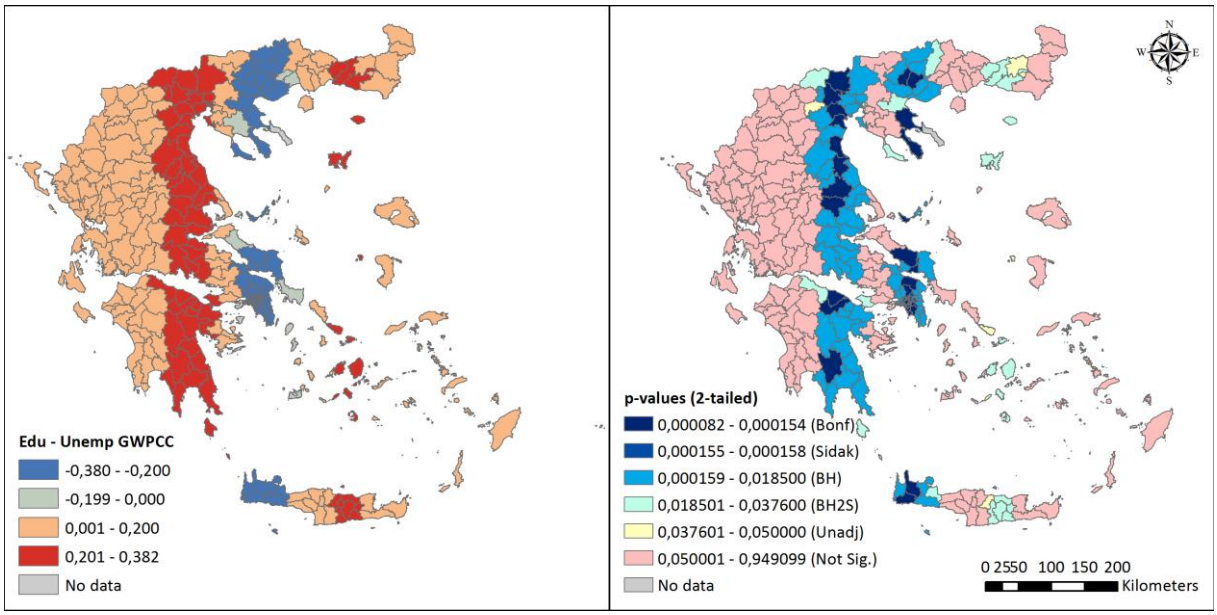

Figure 3. Maps of Geographically Weighted Pearson Correlation coefficients between high

educational attainment and unemployment rate and the corresponding p-values

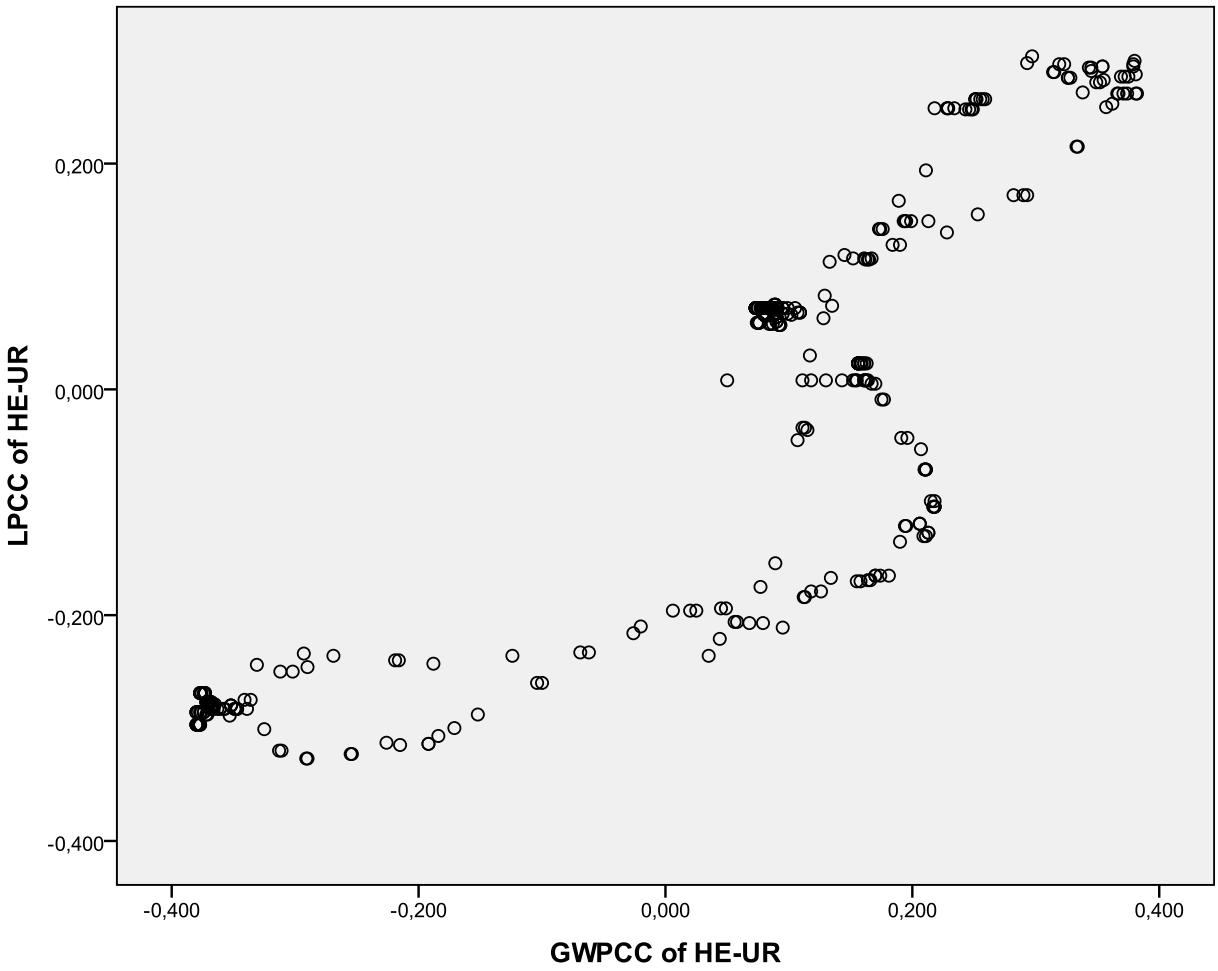

Figure 4. Map of the pairs of LPCCs and GWPCCs between high educational attainment and unemployment rate

5. Conclusions

The findings of this paper support the need for local diagnostic tools and software to calculate these easily when local regression methods are performed. The examination for the existence or lack of local multicollinearity in global regression is not appropriate when local modelling is performed. The results presented in this paper provide empirical evidence that a near zero and not significant global correlation coefficient between two variables can range widely around zero and be significant even after adjusting for multiple hypotheses testing when it is locally calculated using the kernel defined in the corresponding GWR model. Simple local and geographically weighted correlation coefficients show a similar spatial pattern but do differ in magnitude and significance levels. Although theoretically the GWPCC should be more important to test, as it accounts for the existence of spatial autocorrelation in the data, a more thorough study is necessary in order to provide strong evidence for this conclusion. Checking for local multicollinearity among the independent variables is necessary when local regression methods are applied.

6. Acknowledgements

I would like to thank Prof. Chris Brunsdon who keeps helping me understand statistical inference in spatial analysis and its importance since when I was a postgraduate student. I would also like to thank Prof. Stewart Fotheringham for encouraging me to test a Geographically Weighted Pearson Correlation Coefficient versus the local Pearson correlation coefficient. The use of GWR was made possible by the use the software package *GWR 3.0*. This software, of which the National University of

Ireland at Maynooth is the copyright holder (http://ncg.nuim.ie/ncg/GWR), was kindly provided for academic use only by Professor A. Stewart Fotheringham and Mr. Martin Charlton.

7. References

Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing *Journal of the Royal Statistical Society Series B* **57** pp 289–300

Benjamini Y, Kreiger A and Yekutieli D (2006) Adaptive linear step-up procedures that control the false discovery rate *Biometrika* **93** 491–507

Bonferroni C E (1935) Il calcolo delle assicurazioni su gruppi di teste", in: *Studi in Onore del Professore Salvatore Ortu Carboni*. Rome. pp 13–60

Brunsdon C (2009) Statistical Inference for Geographical Processes In Fotheringham AS and Rogerson P (eds) *The SAGE handbook of spatial analysis*. Sage Publications. London pp 207–224

Brunsdon C and Charlton M (2011) An assessment of the effectiveness of multiple hypothesis testing for geographical anomaly detection *Environment and Planning B: Planning and Design* **38** pp 216–230

Fotheringham AS, Brunsdon C and Charlton M (2002). *Geographically Weighted Regression: the analysis of spatially varying relationships*. John Wiley and Sons, Chichester.

Holm S (1979) A simple sequentially rejective multiple test procedure *Scandinavian Journal of Statistics* **6** pp 65–70

Kalogirou S (2010a) Testing local versions of correlation coefficients. In: *Proceedings of the 50th Anniversary European Congress of the Regional Science Association International (ERSA 2010)*. Jönköping, Sweden. paper no 529.

Kalogirou S (2010b) Spatial inequalities in income and post-graduate educational attainment in Greece *Journal of Maps* **6(1)** pp 393–400

Kalogirou S (2011) Testing local versions of correlation coefficients *Review of Regional Research - Jahrbuch für Regionalwissenschaft* **32(1)** pp 45 – 61

Kalogirou S and Hatzichristos T (2007) A spatial modelling framework for income estimation *Spatial Economic Analysis* **2(3)** pp 297–316

Sidak Z (1967) Rectangular confidence region for the means of multivariate normal distributions *Journal of the American Statistical Association* **62** 626–633

Wheeler DC (2006). *Diagnostic tools and remedial methods for collinearity in linear regression models with spatially varying coefficients*. PhD Thesis. Ohio State University, USA.

Wheeler DC (2007) Diagnostic tools and a remedial method for collinearity in geographically weighted regression *Environment and Planning A* **39(10)** pp 2464–2481

Wheeler DC and Paez A (2010). Geographically Weighted Regression. In: Fischer MM and Getis A (eds) *Handbook of Applied Spatial Analysis*. Springer-Verlag. Berlin. pp 461-486

Wheeler D and Tiefelsdorf M (2005) Multicollinearity and correlation among local regression coefficients in geographically weighted regression *Journal of Geographical Systems* **7(2)** pp161–187

Biography

Dr. Stamatis Kalogirou is a Lecturer in Applied Spatial Analysis, Department of Geography, Harokopio University of Athens, Greece and an Affiliate of the National Centre for Geocomputation in Ireland. His research interests include spatial analysis; inference in spatial statistics; internal migration modelling; population projections and ageing; spatial inequalities; and geocomputation applications.