

Local Spatial Analysis of Cardiovascular Diseases in Canadian Health Regions

Analisi Spaziale Locale delle malattie cardiovascolari nelle regioni sanitarie canadesi.

Eugenia Nissi and Annalina Sarra

Abstract Cardiovascular diseases are one of the major worldwide health concerns, responsible for premature death and disability all over the world. The present analysis has as its primary intention in assessing the extent to which the association between cardiovascular disease prevalence in the Canadian Health Regions and some well-established risk factors vary spatially. We adopt the GWR to explore if the effects of some cardiovascular disease predictors are heterogeneous across space and hence vary from place to place. Since GWR amplifies the negative effects of correlation in the weighted explanatory variables, in this study we also employ a GWR model where a ridge regression parameter has been incorporated to reduce model complications arising from collinearity.

Abstract

Le malattie cardiovascolari rappresentano uno dei principali rischi per la salute umana e sono ritenute responsabili di morte prematura e di gravi disabilità. La presente analisi ha come obiettivo principale quello di valutare la misura in cui l'associazione tra l'incidenza delle malattie cardiovascolari e alcuni fattori di rischio ad esse collegate variano spazialmente nei distretti sanitari canadesi. In questo lavoro, si adotta la GWR per esplorare l'eterogeneità spaziale di alcuni predittori di malattia cardiovascolare. Tenuto conto che i modelli di regressione locale tendono ad amplificare gli effetti negativi attribuibili alla collinearità tra le variabili esplicative, si suggerisce l'utilizzo di un modello GWR che contempli un parametro di ridge regression locale.

Key words Geographically Weighted Regression (GWR), Locally Compensated Ridge Regression, Canadian Health Regions.

1 Introduction

Cardiovascular diseases (heart diseases, stroke, hypertension) represent one of the major worldwide wellbeing concern. The implications arising from cardiovascular diseases have severe consequences for health and life expectancy. Cardiovascular diseases are the main cause of premature death and disability throughout the World. According to the latest cardiovascular diseases statistics around 7.3 million people die annually for coronary heart disease and 6.2 million for stroke. Most cardiovascular diseases can be avoided by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity, high blood pressure and diabetes.

This paper aims to present an exploratory spatial analysis for ascertaining Canadian regional variations in the relationships between hypertension disease prevalence and

¹

Eugenia Nissi; University G. d'Annunzio, Department of Economics, nissi@unich.it

Annalina Sarra, University G. d'Annunzio, Department of Economics, asarra@unich.it

some well-established risk factors. Since the geographic variation in risk factors for hypertension disease is too complex to be captured by a single set of regression coefficients, a local regression technique is employed. In particular, in this study, we make use of geographically weighted regression (GWR) models with a ridge regression parameter to condense model complications related to the occurrences of local collinearity in the weighted explanatory variables.

2. Statistical methods: Geographically Weighted Regression (GWR) modelling

The GWR is a local estimation procedure accounting for spatially changing relationships and it is deemed a complement tool to global modelling.

The main appeal of method is that GWR relaxes the assumption in traditional OLS models that the relationships (regression coefficients) between dependent and independent variables being modelled is constant across a study area, by allowing local rather than global parameters to be estimated. A typical model of GWR can be written as:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \quad (1)$$

where β_{ki} ($k=1,2,\dots,M$) are the regression coefficients for each location i and each variable k ; (u_i, v_i) denotes the coordinates of i -th point in space and ε_i are the error terms. Equation 1 creates a continuous surface of estimated parameters values. The local parameters $\beta_k(u_i, v_i)$ are estimated by weighted least-squares estimator, given by :

$$\hat{\beta}(i) = (x^T W(i) x)^{-1} x^T W(i) y \quad (2)$$

In (2) the weight matrix is no longer constant but varies according to the location of point i . An essential step in GWR modelling is the choice of spatial weighting function, which quantifies the spatial dependency between the observed variables. The building of the weight matrix involves the selection of the distance function. The distance can be calculated in various ways and does not have to be necessarily Euclidean. The adoption of a given kernel function and bandwidth are also worth noting. Several weighting functions can be considered and calibrated. For a comprehensive overview about GWR methodology see Fotheringham et al. (2002).

A growing body of literature has argued some problems with using GWR for statistical inference on regression relationships. Recent studies show that this technique is not able to estimate the regression coefficients accurately (see, among others, Wheeler and Calder 2007). One important issue is related to the occurrence of local collinearity in the weighted explanatory variables. The adverse effects of collinearity in the predictors of a linear model are more pronounced in the localised framework of GWR. In that context, the smaller samples used to calibrate each regression along with the potential spatially heterogeneous correlation structures of the data, exacerbate the detrimental influence of collinearity among the predictor variables. Different diagnostics to identify explanatory variable collinearity have

been proposed. Variance inflation factor (VIF) and Condition Number are traditionally employed to measure to what extent the estimates of a linear model are amplified by the collinearity in the predictors. One way to constrain and stabilize regression coefficients is to calibrate a ridge regression. The initial adoption of the ridge regression in the context of GWR is due to Wheeler (2007). Later Brundson et al. (2012) introduce a GWR with a locally compensated ridge term (GWR LCR). A distinguishing feature of that approach is basically the possibility to fit local ridge regressions where the ridge parameter is allowed to vary across space. Besides, to achieve an in-depth understanding of where the undesirable influences of local collinearity might occur, the GWR LCR calibrates the ridge regressions only at locations where the condition number is above a user's specified threshold. Accordingly, there is a condition number associated with every point in the study area at which GWR coefficients are estimated. Formally, the estimator for this locally compensated ridge regression model is:

$$\hat{\beta}(i) = (x^T W(i) x + \lambda(i))^{-1} x^T W(i) y \quad (3)$$

where $\lambda(i)$ is the locally compensated value of the ridge term λ at location i . As in the standard GWR regression, the cross-validation procedure can be adopted to estimate the optimal bandwidth.

3. Data and results

Our data come from the Canadian Community Health Survey (CCHS) which is a series of national cross-sectional surveys that have been carried out by Statistics Canada since 2001. This survey was designed to produce regular and timely cross-sectional estimates of health determinants, health status and health system utilization at provincial and sub-provincial levels (Health Region or combination of Health Region). Owing to the limited availability of data the subsequent analyses have been carried out only in 70 health regions. The base population for each Health Region provides the option of normalizing the data. Our independent variable of interest in all empirical models considered in this study, is the per cent of people aged 12 and over who report that they have been diagnosed by a health professional as having high blood pressure (HBP). Albeit the exact causes of high blood pressure are usually unknown, there are several factors that have been highly associated with this condition. In the current study is acknowledged that life style behaviours, specifically weight loss, regular increase in physical activity, avoiding tobacco smoking and a healthy diet, rich in the consume of fruits and vegetables per day, can effectively lower blood pressure. In addition, research from the WHO highlights the importance of raised diabetes as a risk factor for HBP. All employed explanatory variables are summarised in Table 1.

Table 1: Definitions of the explanatory variables

Variables Name	Variables Description	Codename
Dietary practices	Population aged 12 and over, by the average number of times per day that they consume fruit and vegetables.	DIETS
Weight (BMI)	Overweight (BMI of 25.0-29.9), by number of persons 18 and over, excluding pregnant women, both sexes	OVERBMI
Physical activity	Household population 12 and over, physically inactive	PHYS_IN
Smoking Status	Household population 12 and over, daily smokers	CSMOKE
Diabetes	Population aged 12 and over who reported that they have been diagnosed by a health professional as having diabetes	DIAB

We first carried out an OLS regression and the results are reported in Table 2.

Table 2. OLS parameter estimate summaries

	Coefficient	SE	t_statistic	pvalue	Sign.
Intercept	-1.06	6.24	-0.17	0.87	
DIAB	0.57	0.23	2.42	0.02	*
CSMOKE	-0.20	0.06	-3.52	0.00	***
OVERBMI	0.35	0.08	4.59	0.00	***
DIETS	0.07	0.05	1.29	0.20	
PHYS_IN	-0.04	0.08	-0.55	0.58	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The hypothesized relationship between the HBP and the explanatory variables are supported quite well by the data. Indeed, most of the predictor variables, except those related to dietary practices (Consume of Vegetables and Fruits) and Physical Activity, are statistically significant according to their t-values at the significance level of $\alpha = 0.05$. The global parameter estimates have negative signs for “current smoker”. The coefficient of determination R^2 is 0.49.

In order to run the GWR, we simplify the general format of the model by removing the independent variables which do not exhibit statistically significant effects in the OLS model. Hence, using the same data set, GWR was also tested. Table 3 provides a detailed description of goodness of fit of OLS and GWR models.

Table 3 Goodness-of-fit test for improvement in model fit of GWR over global model (OLS)

Source	SS	DF	MS	F
OLS residuals	480.2	5		
GWR improvement	107.7	9.74	11.0	
GWR residuals	372.5	54.2	6.86	1.61
Diagnostic information	OLS	GWR		
R^2	0.53			
Adjusted R^2	0.49	0.63		

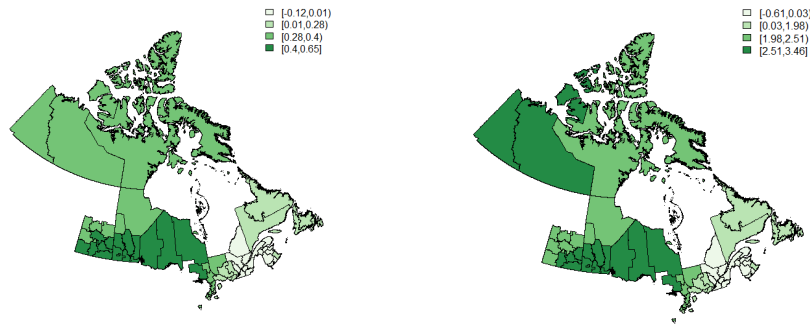
SS= sum of squares; DF=degree of freedom;

MS= residual mean square; F= F-statistic

The GWR estimation improves the overall explanatory power of the HBP model by raising the adjusted R^2 values from 0.49 to 0.63, indicating that are spatially circumstances influencing the per cent of people with HBP.

The Monte Carlo test results for spatially variability of parameters suggested that the association between HBP and almost all the predictors included in the GWR model (OVERBMI, SMOKE and DIAB) is nonstationary across space. The spatial variation of the OVERBMI explanatory variable can be better understood by looking at Figure 1(a), displaying its local parameter estimates. More pronounced and diverse spatial non stationarity is evident in the effect of DIAB on HBP. The relationship is positive and strongest in the South whereas becomes weaker in the North.

Besides, the GWR findings indicate that the largest coefficients for OVERBMI and SMOKE variables are also concentrated in the southern Health Regions. Further enlightening information is provided by maps of t-statistics, obtained dividing each local estimate of the regression coefficients by its corresponding local standard error. In reality, there are some theoretical difficulties in interpreting these t-statistics values. As pointed out by Waller et al. 2007, the pseudo t-values should be used in an exploratory fashion since they do not represent a formal statistical estimate, but they result from a relatively ad hoc inference.



1(a) GWR estimate for OVERBMI variable Adaptive Bandwidth

1 (b) GWR t-values for Estimated Parameters for OVERBMI

Figure 1. Spatial mapping of OVERBMI (GWR) modelling.

We also investigated the issue of collinearity in the analyzed data using the approach of locally compensated ridge GWR, described in Section 2. To this end, we compare the coefficient estimates for the un-adjusted basic GWR with those from a GWR LCR.

The local condition numbers for a basic GW regression are large everywhere, ranging from 50.51 to 105.79. Thus, we calibrate a GWR LCR, which involves a local compensation only at locations displaying condition numbers higher than a given threshold (in our case a threshold of 30 has been specified). As a result, for

this data set, a local ridge term will be found at all locations. The empirical findings suggest that the greatest adjustments were required in northern health districts. After performing different local compensations in relation to the varying levels of collinearity among the predictors, it would result that the larger coefficients for the basic GWR model are reduced in magnitude and the smaller coefficients are raised (see Fig. 2)

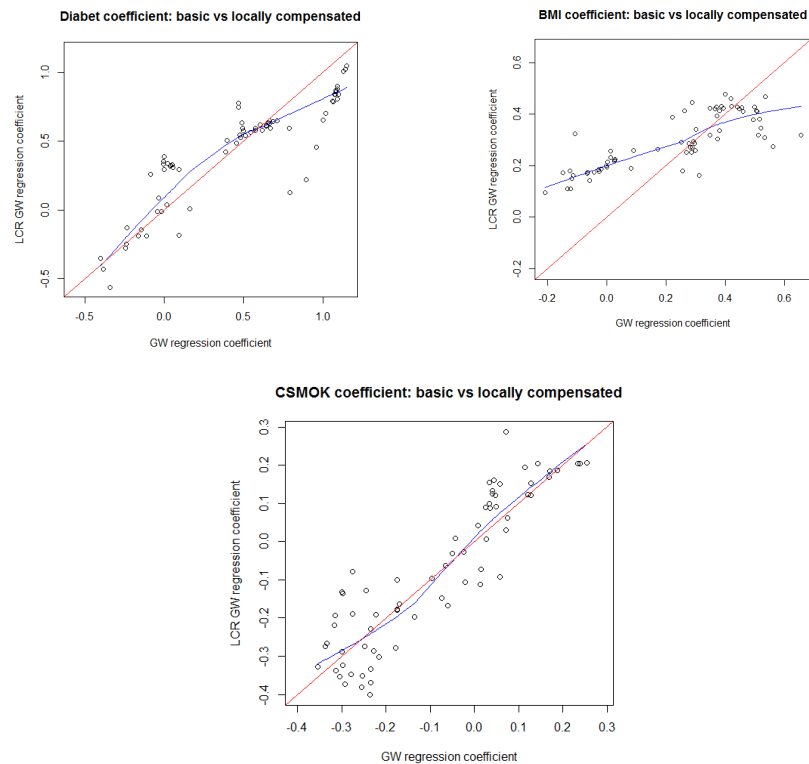


Figure 2 Comparison of coefficient estimates.

References

- Brunsdon, C.; Charlton, M.; Harris, P.,(2012), Living with collinearity in Local Regression Models. Spatial Accuracy Conference Brazil 2012.
- Fotheringham, A. S., Brunsdon, C., and Charlton, M. E. (2002). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships, Wiley, Chichester.
- Wheeler D C, 2007, "Diagnostic tools and a remedial method for collinearity in geographically weighted regression" *Environment and Planning A* **39**(10) 2464-2481.
- Wheeler DC, Calder CA, 2007, "An assessment of coefficient accuracy in linear regression models with spatially varying coefficients" *Journal of Geographical*.