Evaluating the Use of Alternative Distance Metrics in Spatial Regression Analysis of Health Data: A Spatio-temporal Comparison

Stefania Bertazzon and Scott Olson

Department of Geography, University of Calgary
2500 University Dr. NW, Calgary, AB, T2N 1N4, Canada
bertazzs@ucalgary.ca, smolson@ucalgary.ca

Abstract. A method is discussed to enhance the reliability of multivariate spatial regression analysis: alternative values of the Minkowski distance metric are used in the spatial weight matrix. The method is tested on an analysis of the association between heart disease incidence and a pool of socio-economic variables in Calgary over two consecutive census surveys. The method provides a reliable model, which can guide locational decisions to mitigate present and future disease incidence. The model is underpinned by a quantitative definition of neighbourhood connectivity throughout the city. Such connectivity, usually described by Euclidean distance, can be more effectively described by a specifically calibrated distance metric. The analytical results are meaningful, robust to neighbourhood size, and relatively constant over time. Owing to its effectiveness and simplicity, the procedure is generalizable to other health and socio-economic analysis. An automatic implementation is suggested, to assist in the definition of reliable spatial regression models.

Keywords: Spatial regression, distance metric, Minkowski, neighbourhood connectivity, reliability, health, socio-economic, GIS.

1 Introduction

The outbreaks of SARS (Severe Acute Respiratory Syndrome), West Nile virus, and most recently swine flu are but a few examples from recent headlines that point to the compelling need to develop effective analytical tools to model occurrence, transmission, and causes of disease. Many of the most urgent health concerns of today's society are fundamentally spatial in nature: effective accessibility to health care services; prompt and efficient response to epidemic outbreaks; detection and monitoring of environmental health hazards; and consequent urban planning. Spatial analytical methods can be useful management and policy tools to address these concerns, but their use rests on assumptions that are often violated by empirical processes, with the result that much current applied research fails to bring this toolset to its full potential. Presently, management decisions are often supported by quantitative models, specifically regression models; these are potentially desirable tools that can link, for

M.L. Gavrilova and C.J.K. Tan (Eds.): Trans. on Comput. Sci. VI, LNCS 5730, pp. 267–287, 2009.

example, disease incidence to residents' age, thus providing a realistic picture of where health care services will be most needed in the near future. Unfortunately, current models are often uncertain or unreliable. In the best cases, unreliable models provide decision makers with a realistic, but blurry picture of the factors they need to manage, potentially leading to ineffective decisions; in the worst cases the picture is so blurry that it may lead to management decisions that are not just ineffective, but harmful. The uncertainty stems from two properties of geographical phenomena: spatial non-stationarity (things tend to vary unevenly in space), and spatial dependence (near things tend to be more similar than distant things) [1]. Addressing the limitations of regression models is key to improving the reliability of much current quantitative analysis, if, as noted by Griffith and Amrhein [2], most of the multivariate techniques commonly used by geographers can be formulated or reformulated in terms of regression analysis.

The focus of this application is on spatial autoregressive modelling, a technique that specifically addresses the effect of spatial dependence on regression models [3]. In order to maximize the effectiveness of this technique, one crucial element is the correct specification of a spatial weight matrix, capable of providing an accurate representation of the spatial neighbourhood and hence the configuration of the observed spatial dependence. In turn, the spatial weight matrix is specified by a small number of parameters: an appropriate measurement of distance, a correct definition of the range of the observed spatial dependence, and a correct definition of the distance decay effect within the defined neighbourhood [3]. In this paper, the focus is on the distance measurement, discussing the use of a range of distance metrics known as Minkowski distance.

The effect of alternative distance metrics is evaluated on a case study that is relevant from many applied perspectives, including health care provision and urban management. The spatial regression model analyzes the association of demographic and socio-economic factors with the incidence of heart disease in Calgary, a large Canadian city. Population distribution within the city, clustering of age groups, and socio-economic pattern are the main factors considered; urban connectivity enters the model through the measurement of distance among spatial units. By optimizing the distance measurement and specifying an appropriate spatial weight metric, this study provides a method for enhancing the reliability of the estimated model parameters. As a consequence, the model represents an effective analytical support tool for policy and planning decisions. The same spatial regression analysis is estimated for two temporal intervals, hinging on two consecutive census surveys (2001 and 2006). The comparison of the analytical results constitutes a preliminary but important exploration of spatio-temporal dynamics within the city, its health and demographic characteristics, and its socio-economic structure. The analytical findings provide the foundation for more advanced computational developments.

All the statistical computations are conducted in Splus 7 and Splus Spatial Statistics 1.5, with the exception of the bivariate Pearson correlations that are computed in SPSS 15. Geographical data management and visualization are performed in ArcGIS 9.1.

Section 2 provides some background information and an introduction to the case study. Section 3 outlines the various aspects of the methodology: the specification of exploratory analysis and spatial autocorrelation analysis; the definition, selection, and estimation of spatial regression models; and the definition of the distance

metrics of interest. In Section 4 the results of the spatial dependence analyses and regression models are presented for various distance metrics, and in Section 5, the results of the two survey periods are compared and the critical aspects of the methodology are discussed. The final section offers some concluding remarks and future lines of enquiry.

2 Background and Case Study

Heart disease (myocardial infarction) is one of the leading causes of death in the developed world. In addition to the individual characteristics that correlate with the disease, there are a number of factors that are related to a complex variable usually referred to as "lifestyle". Individual characteristics, such as genetic background or simultaneous presence of other conditions, are known as non-modifiable risk factors, in contrast with modifiable risk factors, which include such factors as physical activity, smoking, and diet. These modifiable risk factors tend to correlate with demographic and socio-economic characteristics of individuals [4]. At most geographical scales, demographic and socio-economic characteristics tend to display a pattern, or spatial clustering; disease prevalence, likewise, presents a characteristic geographical distribution, or spatial pattern. For this reason a spatial regression model is an appropriate tool to analyze the spatial pattern of disease occurrence as a function of localized demographic and socio-economic characteristics. This is also the reason why this type of phenomena tends to manifest spatial dependence and non-stationarity. Spatial autoregressive modelling is therefore an effective analytical tool, capable of providing estimates of the coefficients linking disease prevalence to each demographic and socioeconomic factor. The reliable parameters obtained with this method can later be used for analysis and prediction, to ultimately design proactive policy solutions aimed at alleviating and mitigating disease prevalence over the study region.

The spatial regression models discussed in this paper make use of medical records from the APPROACH Project, an ongoing data collection initiative begun in 1995, containing information on all patients undergoing cardiac catheterization in Alberta, and census variables. The medical records represent disease prevalence, and the census variables represent demographic and socio-economic factors. Cardiac catheterization is an invasive procedure for patients experiencing cardiovascular symptoms, which provides important prognostic information for individuals affected by cardiovascular conditions [5]. For the present analysis, a subset of patient records were selected from the provincial database, obtaining a sample of patients undergoing the procedure in Calgary from 1998 to 2002[1], and from 2003 to 2007[2]. Patient address is released at the postal code level; postal code conversion files (PCCFs) from Census Canada were used to calculate the geographic coordinates in latitude and longitude, which were subsequently converted to easting and northing coordinates prior to performing distance computations.

[1] A sample of 11,345 cases, on a total population of 875,245 residents.
[2] A sample of 16,355 cases, on a total population of 988,193 residents.

Socio-economic and demographic variables were drawn from the 2001 and 2006 census surveys, respectively. These variables are available at the dissemination area[3] and census tract levels: for this analysis, the census tract aggregation level was used. At this spatial resolution the spatial dependence is more severe; hence there is a stronger need to implement efficient spatial regression models. In addition, these relatively large units are more meaningful in terms or urban planning and health policies; therefore, a model calibrated at this scale is more useful and applicable than one calibrated at the dissemination area level. The cardiac data were spatially aggregated to match the census tracts, resulting in 182 valid census tract records for the period around the 2001 census, and 186 records for the period around the 2006 census. Fig. 1 shows the Calgary census tracts and the distribution of catheterization cases over the two study periods: Fig. 1a for the 2001 census and Fig. 1b for the 2006 census. It may be worth noting that during the study period, the procedure was available only at the Foothills Hospital, located in the northwest sector of Calgary.

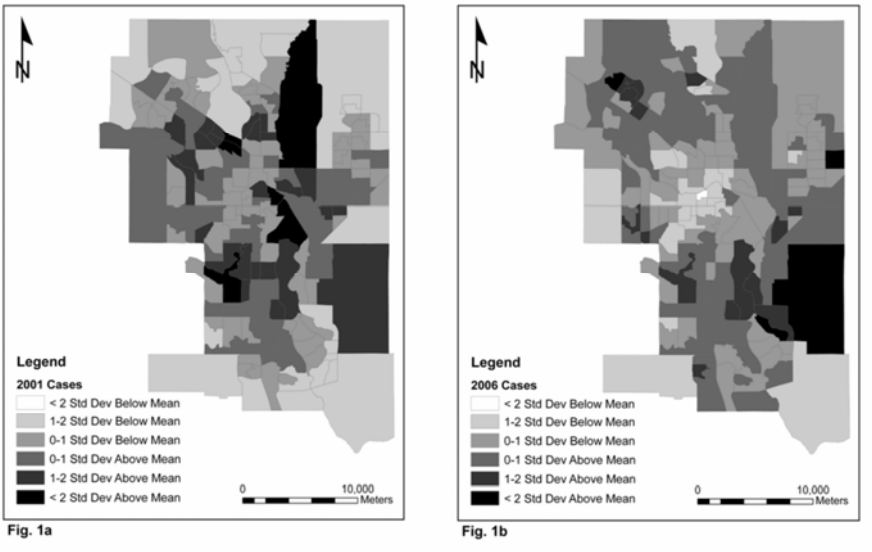

Fig. 1. Spatial distribution of cardiac catheterization cases in 2001 and 2006

Calgary's urban structure is a combination of numerous development episteme. Local patterns of connectivity vary according to local design. For instance, grid pattern road development of the inner city offers different travel options than the circular, cul-de-sac design of its outlying suburban counterparts. Furthermore, large variations in both physical size and shape of neighbourhood form are very apparent in the city. Thus, there is a need to capture how varying urban patterns affect neighbourhood connectivity.

[3] A small, relatively stable geographic area composed of one or more neighbouring blocks standardized through uniform population sizes targeted at 400 to 700 persons. These areas are usually delineated by physical features (roads, water, powerlines, etc.) and respect the boundaries of census subdivisions and census tracts (Statistics Canada, 2007).

3 Methodology

The methodology developed in this paper aims at reducing the inflated variance caused by spatial dependence in regression model estimates. The method will be discussed in detail in this section, and can be usefully summarized as follows. Step 1: calculation of distance using a standard metric; Step 2: calculation of spatial autocorrelation using the standard distance measurement; Step 3: estimation of a spatial regression model based on the standard distance measurement; Step 4: assessment of the model variance; Step 5: replication of Steps 1–4 for an array of candidate Minkowski metrics; Step 6: identification of the distance metric that minimizes the model variance.

3.1 Exploratory Analysis and Spatial Autocorrelation

Census variables and medical records consist of incident numbers. For this reason, a normalization of all the variables was conducted, preliminary to any analysis. In most cases the normalization involved the use of the total resident population as the standardizing variable (e.g., number of cardiac catheterizations), while in other cases it involved the use of a pertinent subset of residents (e.g., population over 20 was used to standardize education levels and population over 15 to standardize marital status). On the normalized variables, descriptive spatial and statistical analyses were conducted, to test their normality and other statistical properties. Cross-correlation analysis (Pearson's coefficient) was used to test the strength of the correlation between each explanatory variable and the dependent variable, as well as the cross-correlation among explanatory variables. The latter test was particularly important to ensure the statistical independence of the explanatory variables introduced in the regression model, in order to avoid multcollinearity in the multivariate regression models.

A traditional spatial autocorrelation test, Moran's I [6], is used to test all instances of spatial autocorrelation throughout the analysis presented in this paper, i.e., to assess the spatial dependence in variables as well as in model residuals. The values of Moran's I can vary between -1 and 1, where -1 indicates perfect negative spatial autocorrelation, 0 indicates absence of spatial autocorrelation, and + 1 indicates perfect positive spatial autocorrelation. The computation of this index requires the specification of a model of spatial dependence, defined by a spatial weight matrix. The matrix defined for the calculation of the spatial autocorrelation index will also be used for the estimation of the spatial autoregressive models (Section 3.2). In its simplest form, the matrix is a binary structure, but it is common to use more complex specifications, which include various types of weights to describe distance decay effects. There are several ways of specifying spatial contiguity [7, 8]: a common method is the definition of k orders of spatial neighbours; an alternative method is a threshold distance; a third method is based on shared borders (for areal units only). While some methods are heavily dependent on the topology of the spatial units, the computation of spatial neighbours is a very general method. In all cases, the extent of the spatial dependence must be defined, either via a maximum distance parameter, or via a number (k) of nearest neighbours.

The correct specification of this matrix is key to minimizing the variance of the spatial regression model. Prior to the specification of a regression model, the spatial autocorrelation is estimated for all the model variables. An intermediate objective of the presented methodology, before the final model can be estimated and its variance assessed, is the identification of the spatial weight matrix that produces the highest value of the spatial autocorrelation index for the variable of interest, i.e., the dependent variable of the regression model: this matrix contains the neighbourhood specification that best captures the spatial dependencies in that variable.

3.2 Spatial Regression Models

The number of spatial regression techniques discussed in the academic literature has grown considerably ([9], [10], [11], [12]) in response to the increasing availability of spatial data, easier access to specialized software, and increased awareness of the inadequacy of traditional analytical techniques in dealing with the unique properties of spatial data [12]. Perhaps the most critical of these properties is spatial dependence, which results in a redundancy of information that inflates the variance (uncertainty) associated with the parameter estimates. Large parameter variance also inflates classical inferential tests, resulting in a more frequent rejection of the null hypothesis. As a consequence, inefficient parameter estimates are not only unreliable, but potentially misleading [3].

Various types of spatial analytical methods have been developed to analyze spatial data, including Bayesian approaches [14] and multilevel models [15]. Spatial autoregressive methods include Generalized Least Squares (GLS) and Maximum Likelihood (ML) models; the covariance structure is typically expressed by a conditional autoregressive (CAR), simultaneous autoregressive (SAR), or moving average (MA) specification [10]. In all cases, a contiguity matrix (Section 3.1) determines which units are spatially dependent [7]. The effectiveness of the regression model depends largely upon the choice of the contiguity matrix and the underlying model of spatial dependence. However, defining contiguity remains difficult and subjective, often dependent on the spatial process under consideration [8].

The contiguity—or spatial weight—matrix is used in the computation of spatial autocorrelation indices (e.g., Moran's I) as well as in the spatial regression:

$$Y = X\beta + \rho WY + \varepsilon \tag{1}$$

where ρ (rho) is the autoregressive parameter and W is the contiguity matrix. Each element of the spatial weight matrix is a spatial weight w_{ij}, which defines the extent of the spatial dependence and the correlation between spatial units as a function of their distance. The spatial autoregressive coefficient, rho, varies between -1 and +1. For rho values approaching 0, the spatial regression model reduces to a standard regression.

The method presented in this paper for the specification of a spatial weight matrix (W) is developed around the nearest neighbour method, where the use of different distance metrics allows for the computation of distance in a way that approximates travel along the road network and actual physical connection, better representing the

actual neighbourhood connectivity. The use of alternative distance metrics produces alternative definitions of nearest neighbours (Section 4.1, Fig. 4). The neighbourhood configuration that best represents actual community structure is expected to better capture the spatial dependence, thereby enhancing the effectiveness of the autoregressive component of the model, expressed by the value and significance of the autoregressive coefficient rho (Equation 1). A model that can best capture spatial dependence via an effective autoregressive specification presents lower variance of the estimated parameters, which are therefore more reliable.

In order to identify the regression which can be considered the best model, the criterion applied here is the minimization of the model's variance, evaluated by means of a broad set of indicators. The value of the spatial autocorrelation index in the model's residuals is considered the most important indicator; this index, in fact, indicates whether or not the model was estimated in violation of the hypothesis of independence of the residuals, which determines the properties of all the model's estimates. In addition to the variance minimization criterion, the model's goodness of fit will be weighed heavily, as it indicates the model's effectiveness in explaining the variation of the dependent variable.

3.3 Alternative Distance Functions

Distance can be measured in many ways: travel time and travel cost [16] are very useful in some contexts, but lack fundamental geometric properties (triangle inequality); Mahalanobis distance has a way of accounting for spatial dependencies. Our work focuses on one category of distance metrics, known as Minkowski distance [17]. This array of metrics was chosen because of its flexibility. In fact, once this class of metrics is chosen, a range of parameters can be selected from within that class; therefore, a single yet flexible measurement method can be defined for the optimal estimation of spatial dependence in spatial autoregressive models. The most commonly used distance metric is the Euclidean or straight line distance:

$$d_{ij} = [(x_i - x_j)^2 + (y_i - y_j)^2]^{1/2} \tag{2}$$

Alternatively, Manhattan distance, also known as City Block Distance [18], is the distance between two points measured along the axes at right angles:

$$d_{ij} = |x_i - x_j| + |y_i - y_j| \tag{3}$$

The Minkowski distance is described by a general formula, of which Euclidean and Manhattan are special cases:

$$d_{ij} = [(x_i - x_j)^p + (y_i - y_j)^p]^{1/p} \tag{4}$$

As visually represented in Fig. 2, Minkowski distance can provide intermediate values between Euclidean and Manhattan distance, producing a more realistic overall

representation of travel in a city; for example, a road network is typically a mixture of straight-lines, curves, and grid-like patterns [19, 20, 21]. Unlike distances measured empirically along a given road network, the use of a specific distance metric provides a consistent model of distance throughout a city or a region, which provides the benefits of generalization but filters out local detail. These distance measurements have been usefully applied in practice for the effective provision and accessibility of health care services [22]. Preliminary analysis has suggested a value of p=1.54 as the best Minkowski value to represent travel along the Calgary road network [23]. Our purpose is not to mimic the city road network but to select a distance metric that best represents neighbourhood connectivity, which in turn is defined by the interplay of road network and urban design.

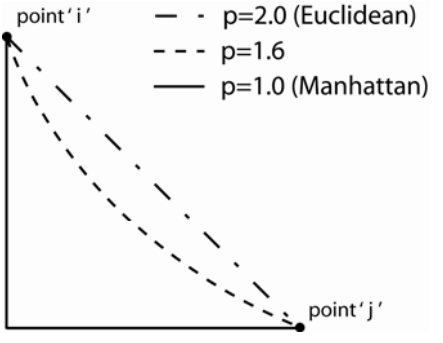

Fig. 2. Travel path for the two extreme and one intermediate Minkowski *p* values

The general Minkowski formula (Equation 4) is used in this application to experiment with a systematic sample of *p* values (i.e., $p = 1.1$, $p = 1.2$, etc.) in the interval $[1 \leq p \leq 2]$, to evaluate the performance of alternative spatial neighbourhood definitions based on the various *p* values for a given order of neighbourhood and distance decay functions. The criterion for choosing an optimal *p* value is the minimization of the variance of the estimates of the spatial regression (SAR) model (Section 3.1). The correlation is modeled in the spatial weight matrix not only by a distance metric but also by a distance decay function and a distance range, defined by the number of nearest neighbours. Even though these three parameters are not mutually independent, the method discussed in this paper focuses on the distance metric. Through a series of empirical experiments, the best value of the other parameters was also identified (number of nearest neighbours and distance decay function, respectively). Given these parameters, we determine the distance metric that minimizes the variance of the model estimates. This process leads to the specification of a set of alternative spatial regression models, based on an array of spatial weight matrices and distance metrics. For each survey period, an iterative process guides us to the selection of the metric that, all else being equal, leads to the lowest model variance. It is our intention to extend this line of work to encompass the definition of an algorithm for the selection of the optimal metric.

4 Results

A number of alternative spatial weight matrices were tested for the dependent variable (cardiac catheterization cases) and the independent variables (demographic and socio-economic variables) in both survey periods. Only the results relative to the dependent variable, "Catheterization Cases", will be discussed in detail in the following subsections. Experiments with increasing orders of neighbourhood (i.e., k = 2, 3,…., 10) have confirmed that, for all the variables, the spatial autocorrelation index is constantly higher for lower orders of neighbourhood, suggesting that the spatial dependence is more pronounced over short distances. Alternative distance decay functions were also tested: the function that best captures the distance decay effect for most variables is an inverse squared distance function weighted by the area of each spatial unit (census tract)[4]. This analysis of various distance decay functions confirms the indication emerging from the inverse relationship between neighbourhood order and spatial autocorrelation index: spatial dependency is more severe over short distances and decreases sharply as distance increases and more spatial units are considered. These results are consistent for both the temporal periods analyzed.

4.1 Spatial Autocorrelation Index, 2001

Based on the experiments summarized above, systematic analyses were conducted for one and two orders of neighbourhood.[5]

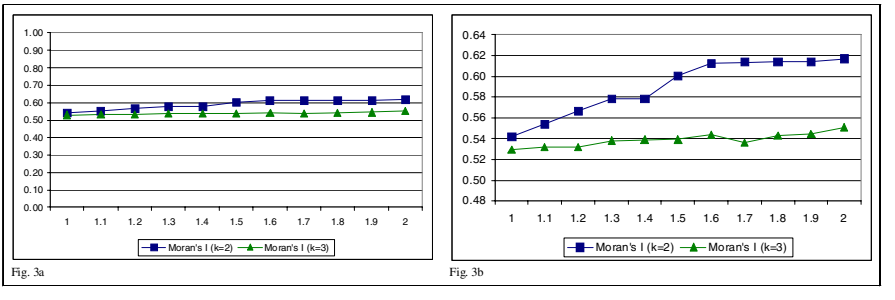

Fig. 3. Spatial autocorrelation in the interest variable for varying p values and distance ranges, 2001

Fig. 3 represents the variation of the spatial autocorrelation index as a function of the p value that defines the Minkowski distance metric. As evidenced in Fig.3a, the difference between one and two nearest neighbours is relatively minor, suggesting that the method is robust with respect to the choice of a neighbourhood order. Fig.3b highlights in greater detail the local features of the spatial autocorrelation function in

[4] Inner-city census tracts tend to have smaller areas than peripheral ones; therefore a pure distance weighted specification would tend to under-estimate the neighbourhood connectivity in the suburbs.

[5] k=2 and k=3, respectively, in Splus.

the interval [$1 \leq p \leq 2$]. For k=2 the function presents an overall increasing trend. A leap upwards is observed at p=1.4; at p=1.6 the line reaches a plateau that remains approximately constant until p=2.0. For k=3 the function presents a relatively stable trend. The line also displays an anomaly, or a peak, at p=1.6, it drops slightly at p=1.7, and then rises again constantly until p=2.0. From this initial analysis, the value p=1.6 emerges as the candidate metric that can best capture the spatial dependence in the variable "Catheterization Cases".

The effect of the distance metric on the selection of nearest neighbours operated by alternative metrics can be better appreciated visually. Fig. 4 shows the comparison between the neighbourhood selection for the extreme p values (p=1 and p=2) as well as for the value p=1.6 that was identified in the spatial autocorrelation analysis (Fig. 3). Fig. 4 presents the analysis conducted for two orders of neighbourhood (k=3), as the visualization results are most effective for this value.

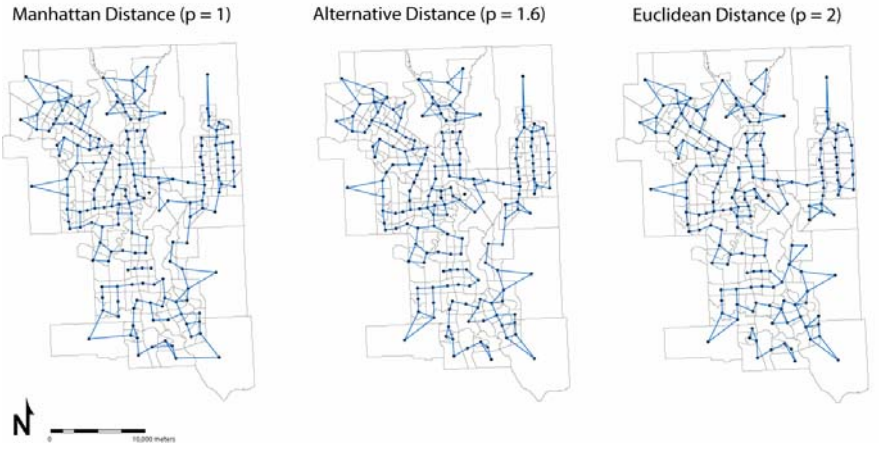

Fig. 4. Second order nearest neighbour connections according to varying distance metrics

A careful examination of the plots in Fig. 4 reveals that the selection of nearest neighbours varies in many parts of the city: individual points selected as nearest neighbours vary for each metric, and these differences become increasingly pronounced as the order of neighbourhood increases.

4.2 Spatial Autocorrelation Index, 2006

The spatial autocorrelation analysis for the second survey period reveals a noticeable change over the 2001 period. Fig. 5 summarizes the variation of the spatial autocorrelation index for the dependent variable as a function of the distance metric.

The main aspect emerging from the analysis summarized in Fig. 5 is the sharp decrease of the spatial autocorrelation value, from an average value[6] of approximately 0.6 in 2001 to a value of approximately 0.3 in 2006. Other analyses conducted on

[6] Depending on the number of nearest neighbours.

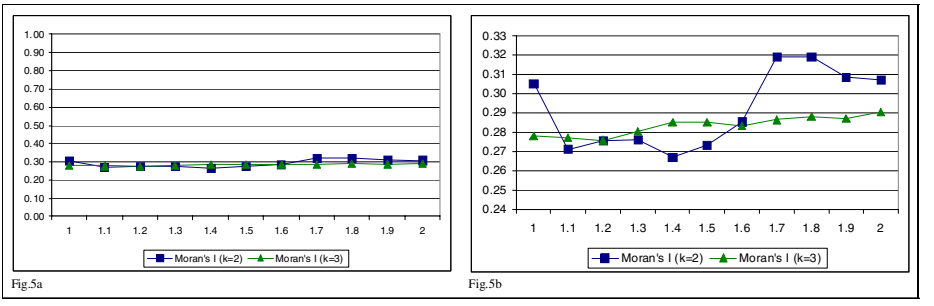

Fig. 5. Spatial autocorrelation in the interest variable for varying p values and distance ranges, 2006

these data have suggested a strong correlation between catheterization cases and population over 65 years for the 2001 survey period [25]. Those analyses indeed suggest that the spatial autocorrelation observed in the disease pattern may be driven by the spatial autocorrelation in the distribution of senior citizens. The pattern of cross-correlation observed for the 2006 period (Section 4.3) suggests, instead, a lower correlation between disease and senior citizens, accompanied by higher correlations between disease and younger age groups. These observations can explain the lower spatial autocorrelation observed in the disease pattern: younger age groups tend to be less clustered in Calgary, therefore less spatially autocorrelated. As a consequence also the disease pattern displays a lower spatial autocorrelation for the 2006 survey period.

As for the 2006 census period, Fig. 5a confirms that the number of neighbours has only a marginal effect on the overall variation. As evidenced by Fig. 5b, the variation of the spatial autocorrelation as a function of the distance metric has somewhat changed between the two survey periods. For two nearest neighbours, i.e., k=3, the trend is generally similar to the one observed in 2001, with an overall upwards trend in the interval $[1 \leq p \leq 2]$. The most significant increase occurs over the interval between p=1.4 and p=1.5. After a minor decline at p=1.6, a second and higher peak is reached at p=1.8; after another minor decline, the final, highest value is reached at p=2, suggesting the Euclidean distance as a better metric for this number of neighbours.

A more complex pattern is shown by the function for 1 nearest neighbour, i.e., k=2. The spatial autocorrelation index exhibits almost identical values at the two extremes of the interval, i.e., for the Manhattan and Euclidean metrics, respectively. Within the $[1 \leq p \leq 2]$ interval, however, the values remain lower between p=1.1 and p=1.6, while the highest peak is reached for the values p=1.7 to p=1.8. For these p values, the spatial autocorrelation index is a whole decimal degree higher than the values obtained for the Manhattan and Euclidean metrics. Consequently, the single p value that maximizes the spatial autocorrelation function for both k=2 and k=3 is p=1.8, for the 2006 period.

The shift from a best p value of p=1.6 in 2001 and of p=1.8 in 2006 is likely related to the shift in the spatial distribution and clustering of the "Catheterization Cases" variable. This may also be due to a shift in the population distribution, and consequently in the disease pattern. The economic boom experienced by the city during the

2003–2007 period caused a massive population increase, mainly constituted by young individuals, which in turn caused an increase in the number of residents, particularly in the suburbs [26]. The spatial autocorrelation index is a joint measure of distance and correlation: these results indicate that it is correlation that drives the change in the spatial autocorrelation index, and therefore the p value of the distance metric evolves, to best capture the new correlation pattern. Over the interval between the two census surveys, Calgary has undergone major social and demographic changes, but lesser physical changes affecting its neighbourhood connectivity.

4.3 Regression Model, 2001

A pool of candidate explanatory variables for the regression model was selected from the census data for each survey period. Table 1 summarizes descriptive spatial statistics and spatial autocorrelation for the dependent variable and the subset of variables used in the 2001 regression models.[7] The standard descriptive statistics evidence the normality of the data, while the spatial autocorrelation index shows that all the variables present significant and generally high spatial dependence.

Table 1. Descriptive statistics and spatial autocorrelation for selected regression variables, 2001

*** Summary Statistics for data in: Master.CT.Norm ***

	cases	males	a45.54	a55.64	a65pl	2p.wchld	gr13ls	non.uni	f.m.inc
Mean:	1.34	49.77	14.46	7.61	9.64	47.31	30.64	36.68	66.61
Median:	1.28	49.80	13.92	7.24	8.16	48.18	28.47	37.04	63.13
Variance:	0.22	2.76	10.00	6.53	32.66	181.27	122.98	29.90	330.68
Std Dev.:	0.47	1.66	3.16	2.55	5.71	13.46	11.09	5.47	18.18
SE Mean:	0.03	0.12	0.24	0.19	0.42	1.00	0.82	0.41	1.35
Skewness:	0.40	0.01	0.60	0.77	0.81	-0.18	0.64	-0.39	0.61
Kurtosis:	-0.29	2.44	0.40	0.32	0.11	-0.53	-0.34	0.11	-0.56

*** Spatial Correlations ***

Correlation	0.62	0.47	0.48	0.57	0.73	0.86	0.82	0.37	0.63
Variance	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Std. Error	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
Normal statistic	6.17	4.73	4.80	5.68	7.25	8.60	8.16	3.69	6.25
p-value (2-sided)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Cross-correlation analysis was conducted to assess the correlation between the dependent and each independent variable and the cross-correlation among the independent variables. In Table 2, census variables have been grouped into homogeneous categories and a sample of 2 representative variables for each category is presented.

[7] Unless otherwise specified, spatial statistics on any variable are conducted using Euclidean distance (p=2.0) and two orders of nearest neighbours (k=3), following the convention used in Splus: based on the cross-correlations among dependent and independent variables, we select the multiple regression that best expresses the relationship between lifestyle and heart disease incidence).

Table 2. Cross-correlation analysis for dependent and selected independent variables, 2001

		Demographics		Family		Housing		Education		Economics	
	cases	a55.64	a65pl	mar.claw	2p.wchld	owned	s.detach	gr13ls	non.uni	unemp	f.m.inc.k
cases	1.000	.569(**)	.794(**)	-.377(**)	-.495(**)	-.285(**)	-.273(**)	.181(*)	-.235(**)	.171(*)	-.229(**)
a55.64	.569(**)	1.000	.415(**)	0.047	-0.074	0.144	0.054	-0.026	-.292(**)	0.090	.195(**)
a65pl	.794(**)	.415(**)	1.000	-.416(**)	-.555(**)	-.359(**)	-.353(**)	-0.098	-.334(**)	0.045	-0.099
mar.claw	-.377(**)	0.047	-.416(**)	1.000	.819(**)	.909(**)	.871(**)	-.224(**)	0.127	-.367(**)	.665(**)
2p.wchld	-.495(**)	-0.074	-.555(**)	.819(**)	1.000	.818(**)	.803(**)	-0.087	0.044	-.153(*)	.572(**)
owned	-.285(**)	0.144	-.359(**)	.909(**)	.818(**)	1.000	.912(**)	-0.133	0.099	-.347(**)	.647(**)
s.detach	-.273(**)	0.054	-.353(**)	.871(**)	.803(**)	.912(**)	1.000	-0.120	0.093	-.322(**)	.593(**)
gr13ls	.181(*)	-0.026	-0.098	-.224(**)	-0.087	-0.133	-0.120	1.000	.251(**)	.336(**)	-.698(**)
non.uni	-.235(**)	-.292(**)	-.334(**)	0.127	0.044	0.099	0.093	.251(**)	1.000	-.160(*)	-.294(**)
unemp	.171(*)	0.090	0.045	-.367(**)	-.153(*)	-.347(**)	-.322(**)	.336(**)	-.160(*)	1.000	-.366(**)
f.m.inc.k	-.229(**)	.195(**)	-0.099	.665(**)	.572(**)	.647(**)	.593(**)	-.698(**)	-.294(**)	-.366(**)	1.000

The cross-correlations analysis provides an interesting portrait of the socio-economic structure of Calgary, as shown, for example, by the high correlation between "owning a house", "married or in common law", and "single detached home", which suggests a predominant traditional family model, and a widespread wealth. Several high cross-correlation values limited our choice of independent variables[8], but, at the same time, the variables included in the models are also representative of those that could not be directly entered in the regressions.

As evidenced in Table 2, the demographic variables, and particularly those indicating old age, tend to display a very high correlation with the dependent variable. Other multivariate analyses of these data [23, 25] have indicated that the correlation between the dependent and demographic variables is so high that the inclusion of these variables in any regression model results in the exclusion of all the socioeconomic variables. For this reason, in this paper, the demographic variables are excluded from the selection process. Alternative model specifications [25] have also confirmed that age and sex standardization of the disease variable has only a minor impact on the regression results.

After experimenting with various combinations of independent variables, a backwards selection process produced the following regression model:

$$CC = f(2p.w.chld, Uni, F.m.inc, Gr13ls) \tag{5}$$

where *CC* is the number of catheterization cases; *2p.w.chld* is the number of families of two parents with children at home; *Non.uni* is the number of persons with a post-secondary, non-university degree; *F.m.inc* is the family median income; and *Gr13ls* is the number of persons with grade 13 or lower education.

The model parameters and selected diagnostics are summarized in Table 3.

[8] Only variables correlated less than +/- 0.7 were included in each model.

Table 3. Multivariate spatial regression model coefficient and selected diagnostics, 2001

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	1.7981	0.3648	4.9294	0.0000
2p.w.chld	-0.0226	0.0032	6.9717	0.0000
Gr13ls	0.0195	0.0043	2.9257	0.0000
F.m.inc	-0.0163	0.0056	2.8138	0.0039
Non.uni	0.0089	0.0032	4.5262	0.0055

L.Likelihood	Pseudo-R^2	Rho	Sigma^2	Res.Std. Err	Res. Moran
-277.8252	0.3724	0.8504	0.1165	0.3413	-0.0293

The model describes the incidence of heart disease as a function of family structure, education, and income. This set of variables possesses a satisfactory explanatory power (pseudo-R^2 = .37).[9] The negative and highly significant coefficient of families with children (t value) suggests a negative correlation between disease and individuals in young families and appears to be related to fairly young individuals, at early to mid stages of their career, likely with relatively high education and moderately high income. The positive relationship between disease incidence and low education (grade 13 or less) suggests a relationship with old age and fringes of poverty and low social status. The negative coefficient linking disease incidence and areas dominated by post-secondary, non- university education identifies trade workers and professionals: a category with fairly high income levels. Overall, the education variables suggest that higher education levels are associated with greater income and lower disease incidence, suggesting that higher education levels may lead to healthier lifestyle and lower risk of disease. Finally, the positive relationship between disease incidence and income suggests the hypothesis of higher disease incidence in individuals with higher levels of stress and responsibility and appears to be related to more mature professionals, therefore possibly encompassing a latent age factor.

The regression summarized in Equation 5 and Table 3 contains a number of significant explanatory variables, which produce a satisfactory model, from a statistical as well as a conceptual point of view. This model is estimated from a spatial weight matrix based on the standard Euclidean metric for the distance measurement. The model will now be re-estimated, using an array of alternative spatial weight matrices, each based on a different value of the Minkowski metric in the $[1 \leq p \leq 2]$ interval.

The results of this experiment are presented and compared in Fig. 6, which depicts a selection of key regression indicators as a function of the distance metric. The first two values are indicators of the model's performance and goodness-of-fit, i.e., the logarithm of the likelihood and the pseudo-R^2; the rho value (Section 3.1), indicates the relative importance of the autoregressive coefficient; residual standard error and σ^2 (sigma^2) are indicators of the model's variance; finally Moran's I measures the spatial autocorrelation in the model's residuals. All the indicators have been scaled and plotted in one single graph so that they can be compared more easily.

[9] Following Anselin [26], the pseudo-R^2 is calculated as the square of the correlation between observations and regression fit.

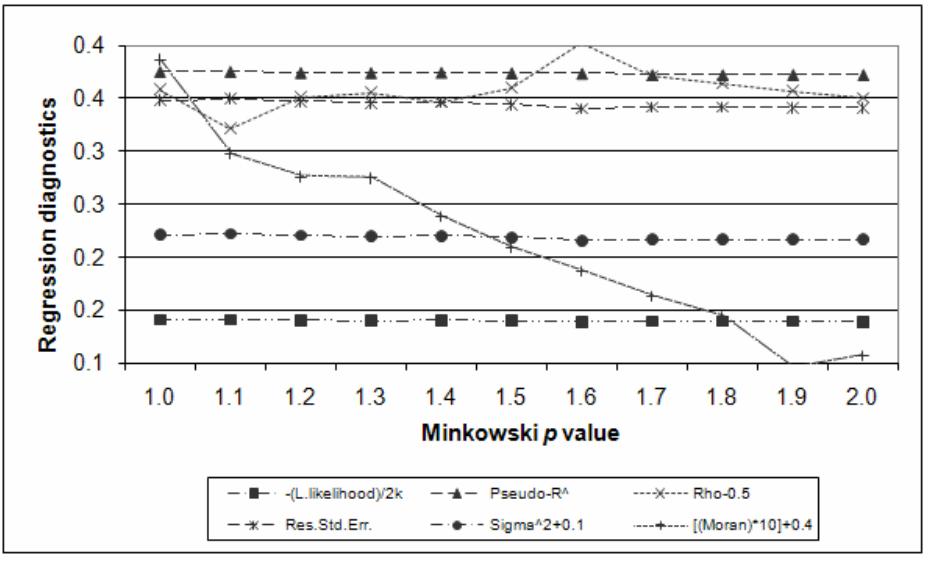

Fig. 6. Spatial regression model selected diagnostics for varying *p* values, 2001

The trends presented in Fig. 5 indicate that the variance indices of the model are affected, to varying degrees, by variations in the distance metric, or *p* value, through the spatial weight matrix. Conversely, the use of alternative distance metrics has only a negligible impact on the goodness-of-fit indicators. The indices that are most affected are the autoregressive coefficient, rho, and the spatial autocorrelation index on the model residuals, Moran's I. These results confirm that an appropriate choice of the *p* value can impact the neighbourhood definition and consequently the model's capacity to effectively capture spatial dependencies, thus ultimately enhancing the reliability of the estimates.

The results cannot be considered conclusive, but from the majority of our tests the value p=1.6 continues to emerge as the best candidate for the optimal distance metric. The rho value displays a distinct peak at p=1.6, suggesting that with that metric the autoregressive coefficient is most effective at capturing the spatial dependence in the model's residuals. For the same *p* value, a corresponding trough is displayed by residual standard error and sigma square. Visually, the latter features appear to be much less pronounced than the rho value, but range and scaling of these indices should be considered. Somewhat in contrast with the above values, the spatial autocorrelation index in the model's residuals (Moran's I) displays an overall decreasing trend, reaching its lowest value at p = 1.9. The interpretation of this trend presents some difficulties, in that all its values are negative, indicating that negative spatial autocorrelation increases for increasing *p* values. Moreover, the values of residual Moran's I over the $[1 \leq p \leq 2]$ interval are not only insignificant, but very low, as its maximum value is -0.001 for p = 1 and its minimum value is -0.030 for p = 1.9. For all these reasons, the trend of the rho coefficient and of the variance indicators will be weighed more

heavily than residual Moran's I in the identification of a suitable *p* value. A *p* value of 1.6 is therefore the most suitable value emerging from this analysis.

4.4 Regression Model, 2006

Descriptive statistics and spatial autocorrelation index for the variables used in the 2006 regression models are summarized in Table 4.

Overall the changes during the previous period are relatively minor. Most of the variables can still be considered normal, although the variables "Catheterization Cases" and "Male Residents" present very high values of kurtosis. This may be explained by the considerations presented in Section 4.2 on more recent socio-demographic trends. The spatial autocorrelation index presents overall lower values than for the 2001 survey period.

Table 4. Descriptive statistics and spatial autocorrelation for selected regression variables, 2006

*** Summary Statistics ***									
	cases	males	a45.54	a55.64	a65pl	single	gr13ls	non.uni	f.m.inc
Mean:	2.25	49.93	16.35	9.73	10.67	36.08	39.20	27.75	82.01
Median:	2.24	49.77	16.16	9.33	9.65	34.88	36.65	27.78	78.05
Variance:	0.77	4.68	9.81	7.79	30.65	73.30	134.62	25.83	576.84
Std.Dev	0.88	2.16	3.13	2.79	5.54	8.56	11.60	5.08	24.02
SE. Mean	0.06	0.16	0.23	0.20	0.41	0.63	0.85	0.37	1.76
Skewness:	3.07	3.01	0.20	0.76	0.76	0.82	0.61	-0.26	0.96
Kurtosis:	23.27	23.47	0.25	0.61	-0.05	0.38	-0.43	0.41	1.71
*** Spatial Correlation ***									
Correlation	0.29	0.37	0.30	0.35	0.53	0.70	0.80	0.42	0.47
Variance	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Std.Error	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
Normal statistic	4.60	5.84	4.80	5.57	8.32	11.02	12.51	6.56	7.46
p=value(2-sided)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

The correlation analysis for the 2006 survey period is summarized in Table 5.

Table 5. Cross-correlation analysis for dependent and selected independent variables, 2006

		Demographics		**Family**		**Housing**		**Education**		**Economics**	
	cases	a45.54	a55.64	single	2p.wchld	owned	sdetach	gr13ls	non.uni	unemp	f.m.inc.k
cases	1	.404**	.427**	-.449**	0.046	.363**	.392**	.213**	0.116	0.046	0.106
a45.54	.404**	1	.544**	-.320**	.333**	.417**	.451**	-0.138	0.034	-0.046	.494**
a55.64	.427**	.544**	1	-.268**	-0.048	.200**	.160*	-0.025	-0.067	0.056	.216**
single	-.449**	-.320**	-.268**	1	-.619**	-.870**	-.817**	0.138	-.167*	.218**	-.617**
2p.wchld	0.046	.333**	-0.048	-.619**	1	.718**	.691**	0.05	.160*	-.146*	.462**
owned	.363**	.417**	.200**	-.870**	.718**	1	.899**	-.159*	.232**	-.322**	.667**
sdetach	.392**	.451**	.160*	-.817**	.691**	.899**	1	-0.107	.250**	-.269**	.602**
gr13ls	.213**	-0.138	-0.025	0.138	0.05	-.159*	-0.107	1	.245**	.257**	-.655**
non.uni	0.116	0.034	-0.067	-.167*	.160*	.232**	.250**	.245**	1	-.193**	-.176*
unemp	0.046	-0.046	0.056	.218**	-.146*	-.322**	-.269**	.257**	-.193**	1	-.369**
f.m.inc.k	0.106	.494**	.216**	-.617**	.462**	.667**	.602**	-.655**	-.176*	-.369**	1

The 2006 cross-correlation analysis, compared with the 2001 analysis, indicates a different picture of the socio-economic structure of the city of Calgary. The correlation between disease and senior citizens ("age 65 plus") has significantly decreased, and has been surpassed by the correlation between disease incidence and younger age groups, i.e., "age 55–64" and "age 45–54". The variable most highly correlated with disease in 2006 is no longer "age 65 plus", but "single". Most likely this is due to the recent phenomenon of internal immigration (from less wealthy parts of the country) of young persons seeking work in the booming Alberta economy (see also Section 4.2). In addition, availability of new medications and changes in Alberta Health Care policies [28] are likely to have affected the frequency of cardiac catheterization in different age groups.

The model selection procedure results in the model described in Equation 6 and summarized in Table 6.

$$CC = f(Single, Gr13ls) \tag{6}$$

where *CC* is the number of catheterization cases; *Single* is the number of persons who were never married; and *Gr13ls* is the number of persons with grade 13 or lower education.

Table 6. Multivariate spatial regression model coefficient and selected diagnostics, 2006

	Value	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.9993	0.3248	9.2357	0.0000	
Single	-0.0501	0.0075	-6.6779	0.0000	
Gr13ls	0.0267	0.0057	4.6477	0.0000	
L.Likelihood	**Pseudo-R^2**	**Rho**	**Sigma^2**	**Res.Std. Err**	**Res. Moran**
-426.9000	0.2751	0.1203	0.5272	0.7261	0.0078

The model consists of only two variables, i.e., "single" (negative coefficient) and "grade 13 and less" (positive coefficient) suggesting that the disease incidence is inversely correlated with family status and directly correlated with lower education. Compared with the 2001 model, this is a simpler, but less informative model. The presence of the variable "grade 13 and less" confirms the importance of education attainment in explaining the disease incidence. The variable "single" may indicate a social evolution, in that the negative correlation between disease and of families with children has been replaced by a negative correlation between disease and singles. These results are likely related to the demographic and social transformations induced by the recent economic boom experienced by the city of Calgary.

The re-computation of this regression with varying spatial weight matrices based on alternative *p* values is summarized in Fig. 7.

Similar to with the 2001 regressions, the impact of alternative distance metrics is remarkable on the model variance indicators, but negligible on the goodness-of-fit indicators. Again for this census period, the effect of the distance metric is most pronounced on the rho parameter and the index of spatial autocorrelation on the

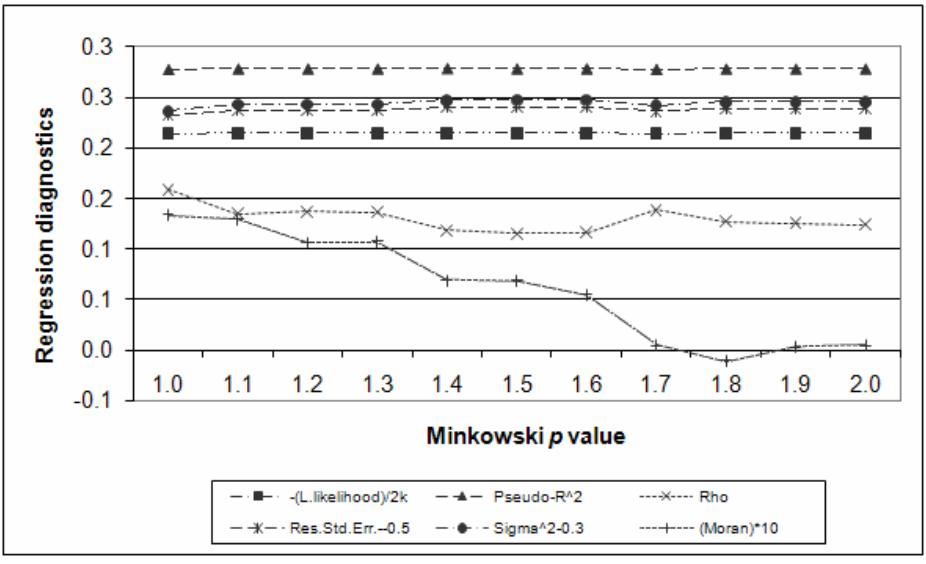

Fig. 7. Spatial regression model selected diagnostics for varying p values, 2006

residuals. For the rho value, the highest value corresponds to the p value of 1.7. For the spatial autocorrelation index, the lowest value is reached at p=1.8. The latter is a negative value, (-0,011), while in absolute value the lowest value is 0.003 corresponding to p=1.9.

Once again, the p value that maximizes the spatial autocorrelation index for the dependent variable emerges also as the best value for the regression model. These results provide important elements to evaluate the value and applicability of the methodology presented.

5 Discussion

The methodology presented in this paper hinges on the proposition that different ways of measuring distance can affect the definition of the spatial weight matrix, and, ultimately, the reliability of multivariate spatial regression models. The methodology is simple and, owing to its simplicity, presents a strong potential for implementation in an automated procedure. A few critical aspects require an in-depth discussion.

The analysis is limited to the consideration of the Minkowski distance, as this method allows for the identification of parameters that can provide a model of empirical distance, or travelling mode, on a given road network. This is a practical decision, presenting the advantage of limiting the number of possible metrics to consider and, at the same time, allowing for the choice of a specific parameter, i.e., p value, within a given interval. Following this method, it is possible to maintain a uniform framework, and simply modify the p value to reflect variations in the correlation pattern or the physical network connectivity, as shown by the shift in the best p value from p=1.6 to p=1.8 over the interval between the 2001 and the 2006 census surveys (Section 4.2).

Such simple adjustments of the p value can be rapidly implemented to maintain the highest reliability of the spatial regression model.

The subsequent steps of the process also follow a simple process: initially, a set of 11 spatial weight matrices are defined, based on regular increments of the Minkowski p value in the interval $[1 \leq p \leq 2]$. This series of spatial weight matrices is initially used to compute the spatial autocorrelation index for the variable of interest. This analysis identifies a potential best p value that maximizes the value of the spatial autocorrelation index for that variable. The consequent estimation of a series of spatial regression models has consistently confirmed that the p value preliminarily identified by the spatial autocorrelation analysis is also the best p value for the multivariate spatial regression models, as indicated most strongly by the variation of the spatial autoregressive coefficient (rho). This is a vital point, which suggests that one can simply identify an optimal p value from the spatial autocorrelation analysis, and then proceed to the definition and selection of the spatial regression model using the spatial weight matrix based on that p value.

Indeed, the analysis of the regression model diagnostics (Sections 4.3 and 4.4) confirmed that modification of the spatial weight matrix (i.e., through modifications of the p value) can have a significant impact on the model's variance, most noticeably evidenced by the values of the autoregressive coefficient (i.e., the rho parameter) and the index of spatial autocorrelation in the model's residuals. Conversely, the impact of such variation is constantly negligible on the indicators of goodness-of-fit. These results corroborate the indication that the spatial regression model can be specified and selected independently of the spatial weight matrix specification, thereby suggesting a simplified procedure that can pursue the two main processes, (1) spatial weight matrix definition and (2) model specification and selection, independently.

With reference to the case study presented, the p values identified for the two subsequent census surveys appear to be meaningful values when cross-checked from different perspectives: they are consistent with other findings [23]; they are robust to variation in the distance threshold (i.e., k value, or number of nearest neighbours) considered; and remain relatively consistent over time. Moreover the transition to a higher p value appears to be explained by the major social and demographic transformations experienced by the city over this period.

Overall, the methodology presented in this paper provides an effective tool to enhance the reliability of multivariate regression by altering the distance metric that is used in the spatial weight matrix. The method is simple and can be split into a small number of easy steps. For this reason, a semi-automatic procedure can be envisaged, to select the p value that, all else being equal, minimizes the model's variance. This procedure can accompany the more delicate tasks of model definition and selection, which are not substantially altered by the spatial weight matrix definition, but require judgment and subjective choices, and should therefore continue to be performed manually.

6 Conclusion

The analysis presented in this paper strongly supports the proposition that the choice of a distance metric affects the definition of the spatial weight matrix and can thus

lead to the specification of a more reliable spatial autoregressive model. The research presented provides one comprehensive approach to the solution of one of the most common and serious problems affecting the analysis of spatial data: spatial dependence. By estimating efficient regression parameters, the research provides effective support for explicitly spatial policy decisions.

The methodology has been tested on the same variables over two consecutive census surveys: not only have the tests produced consistent results, but the specific differences between the two periods are explained by the major social and demographic changes that have affected the city over that period. The results have also proven meaningful and robust to variations in other parameters of the spatial weight matrix, i.e., distance range. Further analysis of the spatial regression diagnostics have confirmed that variations in the spatial weight matrix can substantially affect the model's variance, but have a negligible effect on its goodness-of-fit.

The analysis confirms the value of the methodology presented. The desirable characteristics of the method, along with the simplicity and modularity of the procedure support the feasibility of its automated implementation to enhance the reliability of applied health and socio-economic spatial regression modelling.

Acknowledgements

We would like to acknowledge the GEOIDE network, and our partners and collaborators for supporting our research project "Multivariate Spatial Regression in the Social Sciences: Alternative Computational Approaches for Estimating Spatial Dependence". We would also like to thank the APPROACH initiative and researchers for providing us with data and support throughout our work. We also appreciate the contributions and suggestions of all the students who helped us with this project, particularly with data preparation and Splus scripting.

References

1. Cliff, D., Ord, J.K.: Spatial Processes. Models and Applications. Pion, London (1981)
2. Griffith, D.A., Amrhein, C.G.: Statistical Analysis for Geographers. Prentice Hall, Englewood Cliffs (1991)
3. Anselin, L.: Spatial Econometrics: Methods and Models. Kluwer Academic Publisher, New York (1988)
4. Kaplan, G.A., Keil, J.E.: Socioeconomic factors and cardiovascular disease: a review of the literature. Circulation 88(4), 1973–1998 (1993)
5. Ghali, W.A., Knudtson, M.L.: Overview of the Alberta Provincial Project for Outcome Assessment in Coronary Heart Disease. Canadian Journal of Cardiology 16(10), 1225–1230 (2000)
6. Getis, A.: A history of the concept of spatial autocorrelation: A geographer's perspective. Geographical Analysis 40(3), 297–309 (2008)
7. Getis, A., Aldstadt, J.: Constructing the Spatial Weights Matrix Using a Local Statistic. Geographical Analysis 36, 90–104 (2004)
8. Bertazzon, S.: A definition of contiguity for spatial regression analysis in GISc: Conceptual and computational aspects of spatial dependence. Rivista Geografica Italiana 2(CX), 247–280 (2003)

9. Anselin, L.: Under the hood. Issues in the specification and interpretation of spatial regression models. Agricultural Economics 27(3), 247–267 (2002)
10. Cressie, N.: Statistics for Spatial Data. Wiley, New York (1993)
11. Fotheringham, A.S., Brundson, C., Charlton, M.: Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, Chichester (2002)
12. Ward, M.D.: Spatial regression models. Sage, Los Angeles (2008)
13. Openshaw, S., Alvanides, S.: Applying geocomputation to the analysis of spatial distributions. In: Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W., et al. (eds.) Geographical Information Systems: Principles and Technical issues, vol. 1, pp. 267–282 (1999)
14. Besag, J., Green, P.: Spatial statistics and Bayesian computation. Journal of the Royal Statistical Society B 55, 25–37 (1993)
15. Duncan, C., Jones, K.: Using multilevel models to model heterogeneity: Potential and pitfalls. Geographical Analysis 32, 279–305 (2000)
16. Bailey, T., Gatrell, A.: Interactive Spatial Data Analysis. Wiley, New York (1995)
17. Haggett, P., Cliff, A.D., Frey, A.: Locational Analysis in Human Geography. Edward Arnold, London (1977)
18. Krause, E.F.: Taxicab geometry. Addison-Wesley, Menlo Park (1975)
19. Apparicio, P., Abdelmajid, M., Riva, M., Shearmur, R.: Comparing alternative approaches to measuring the geographical accessibility of urban health services: Distance types and aggregation-error issues. International Journal of Health Geographics 7, 7 (2008)
20. Laurini, R., Thompson, D.: Fundamentals of Spatial Information System. The A.P.I.C. Series, vol. 37. Academic Press, London (1992)
21. Phibbs, C.S., Luft, H.S.: Correlation of travel time on roads versus straight line distance. Medical Care Research and Review 52(4), 532–542 (1995)
22. Kohli, S., Sahlen, K., Sivertun, A., et al.: Distance from the primary health center: A GIS method to study geographical access to health care. Journal of Medical Systems 19(6), 425–436 (1995)
23. Shahid, R.: GWR in Health: An Application to Cardiac Catheterization in Calgary. In: Proceedings ESRI Health GIS Conference 2007 (2007)
24. Bertazzon, S.: Cardiovascular disease and socio-economic risk factors: an empirical spatial analysis of Calgary (Canada). Rivista Geografica Italiana 116(3) (forthcoming, 2009)
25. Bertazzon, S., Olson, S., Knudtson, M.: A spatial analysis of the demographic and socio-economic variables associated with cardiovascular disease in Calgary (Canada). Applied Spatial Analysis and Policy (2009), doi:10.1007/s12061-009-9027-7
26. Statistics Canada: Report on the Demographic Situation in Canada (2008)
27. Anselin, L.: SpaceStat tutorial. Regional Research Institute. West Virginia University, Morgantown, West Virginia (1993)
28. Alberta Health and Wellness, `http://www.health.alberta.ca/health-care-insurance-plan.html` (accessed, 13/05/2009)