Alternative Distance Metrics for Enhanced Reliability of Spatial Regression Analysis of Health Data

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. We present a spatial autoregressive model (SAR) to investigate the relationship between the incidence of heart disease and a pool of selected socio-economic factors in Calgary (Canada). Our goal is to provide decision makers with a reliable model, which can guide locational decisions to address current disease occurrence and mitigate its future occurrence and severity. To this end, the applied model rests on a quantitative definition of neighbourhood relationships in the city of Calgary. Our proposition is that such relationships, usually described by Euclidean distance, can be more effectively described by alternative distance metrics. The use of the most appropriate metric can improve the regression model by reducing the uncertainty of its estimates, ultimately providing a more reliable analytical tool for management and policy decision making.

1 Introduction

The outbreaks of SARS (Severe Acute Respiratory Syndrome), West Nile virus, and avian flu are but a few examples from recent headlines that point to the compelling need for the development of the most 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 process, so 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, which are potentially desirable tools that can link, for 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 vary unevenly in space), and spatial dependence (near things are

O. Gervasi et al. (Eds.): ICCSA 2008, Part I, LNCS 5072, pp. 361–374, 2008.

more similar than distant things) [1]. Addressing the limitations of regression models is the key to improve 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.

This paper presents an application of spatial regression analysis, aimed at reducing the uncertainty of the model by optimizing the specification of the spatial weight matrix. The core of our work is the evaluation of the performance of alternative distance metrics in capturing the spatial dependence of heart disease incidence and its related socio-economic factors. 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 obtained using ArcGIS 9.1.

In the following section we provide background information and introduce the case study; in section 3 we discuss the methodology; in section 4 we present and discuss the results; and finally we offer some conclusion and future lines of work.

2 Background and Case Study

Heart Disease (myocardial infarction) has become one of the leading causes of death in the developed world. "It is not obvious, however, what the relative importance is of such factors as stress, limited physical activity, smoking, high intake of calories and high proportion of saturated fats, or what the relation is between these characteristics and elevated blood pressure, serum cholesterol and triglycerides (blood fat)" [3] . All these factors are in turn related to a complex variable usually referred to as *lifestyle*, which is hard to characterize and measure. Demographic indicators (e.g. age, sex), socio-economic indicators (i.e. income, job type) and environmental indicators (e.g. recreation sport facilities, pollution), can provide an indication of lifestyle.

Our spatial regression model rests on medical records (APPROACH Project) and census variables, expressing *lifestyle* factors. The APPROACH Project is an ongoing data collection initiative, begun in 1995, containing information on all patients undergoing cardiac catheterization in Alberta; cardiac catheterization refers to an emergency procedure for patients experiencing myocardial infarction [4]. For this work we selected from the provincial database approximately 12,000 records of patients undergoing the procedure in the city of Calgary from 1998 to 2002; patient address is released at the postal code level. During the study period, the procedure was available only at the Foothills hospital, located in the Northwest of the City.

Socio-economic and demographic variables are drawn from the 2001 census data. These variables are available at the dissemination area[1] and census tract levels. Postal code conversion files (PCCF) from Census Canada provided the geographic coordinates in latitude and longitude, which were subsequently converted to easting and northing coordinates prior to performing distance computations. The cardiac data

[1] 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).

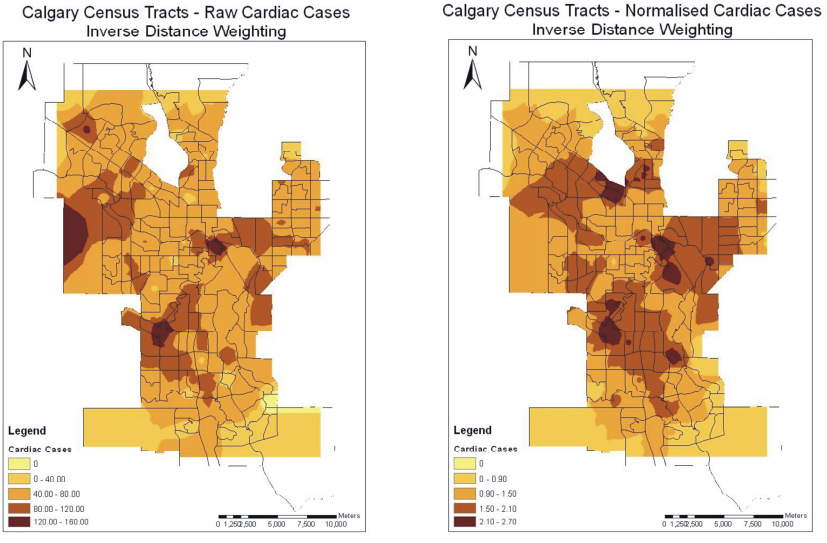


Fig. 1. Inverse weighting distance interpolation maps for raw and normalized number of cases of incidence of cardiac disease in Calgary

were spatially aggregated to match these two aggregation levels, resulting in approximately 1300 valid dissemination area records and 180 census tract records. Fig. 1 shows the distribution of catheterization cases over the entire study period and the Calgary census tracts.

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 Methodology

Here, a spatial regression model is calibrated to investigate the relationship between disease incidence and an array of demographic and socio-economic factors that can be used by decision makers as an effective and reliable policy and planning tool. The reliability of the model is enhanced by specifying a spatial weighting matrix that best captures the neighbourhood connectivity, hence the spatial dependence in the observed variables. The method we propose to achieve this goal involves altering the method used for calculating the distance metrics inherent to the foundation of the spatial weighting matrix in spatial autoregressive models. This alternative approach can reflect overall urban development and road network connectivity more accurately than the traditionally utilized distance metrics.

3.1 Spatial Regression Models

The number of spatial regression techniques discussed in the academic literature has grown considerably in recent years ([5], [6], [7]). Increased availability of spatial data and more accessible specialized software have certainly played a role, but the main reason for such developments relates to increasing awareness of the inadequacy of traditional analytical techniques in dealing with the unique properties of spatial data [8]. Perhaps the most critical of such properties is spatial dependence, which introduces 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.

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 [6]. In all cases, a constant covariance structure is assumed, and a contiguity matrix determines which units are spatially dependent [9]. 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 [10].

The spatial weighting, or contiguity matrix, is used in the computation of spatial autocorrelation indices (e.g., Moran's I) as well as in the spatial regression (equation 1).

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

where ρ (rho) is the autoregressive parameter and W is the contiguity matrix.

In its simplest form, W is a binary structure, while some more complex specifications include various types of weights that describe distance decay effects [9]. There are several ways of specifying spatial contiguity [10]: 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 [11].

Our proposed method is developed around the nearest neighbour method, and the use of different distance metrics allows the computation of distances in a way that resembles 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. The neighbourhood configuration that better represents actual community structure is expected to best capture the spatial dependence, thereby enhancing the effectiveness of the autoregressive component of the model, which is expressed by the value and significance of the autoregressive coefficient, rho. A model that can best capture spatial dependence via an effective autoregressive specification presents lower variance of the estimated parameters, which are thus ultimately more reliable.

3.2 Alternative Distance Functions

Distance can be measured in many ways: travel time and travel cost [12] are very useful in some contexts, but lack fundamental geometric properties (triangle inequality); Mahalanobis distance is an interesting method to consider spatial dependencies. Our work considers only one category of distance metrics, which can serve as a basis for the definition of a single criterion for the selection of an 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 [13] is the distance between two points measured along the axis 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. Unlike distances measured empirically along an empirical 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. 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.

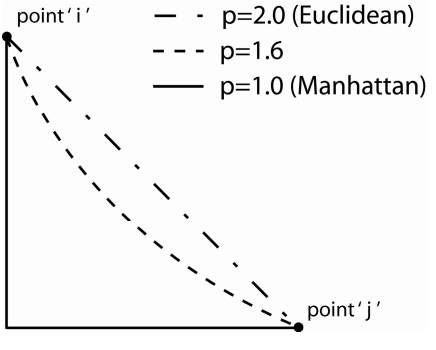


Fig. 2. The difference in travel path for varying p values as defined by Euclidean, Manhattan, and one possible intermediate metric of determining distance between two points in space

Using the general Minkowski formula (equation 4), we experiment with a systematic sample of p values (i.e., $p = 1.1$, $p = 1.2$, etc.) in the interval $[1 \leq p \leq 2]$, and examine alternative spatial autocorrelation functions based on various p values several orders of neighbourhood. The criterion for choosing an optimal model is based on the minimization of the variance of the estimates of the regression model. The main result is the identification of the p value that minimizes the variance in the spatial regression (SAR) model. Estimating this model requires the calibration of spatial weights, w_{ij}, that define the extent of the spatial dependence, and the correlation between spatial units as a function of their distance. The correlation is modeled by a distance metric and a distance decay function that is also chosen to minimize the variance of the model estimates. We first calibrate the distance decay function, for the two extreme distance metrics (Euclidean and Manhattan distance, respectively), and after this we proceed to select the most appropriate distance metric. This takes us to the specification of a set of alternative spatial regression models, based on an array of contiguity matrices and distance metrics. 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 define an algorithm for the selection of the optimal metric.

4 Results and Discussion

The discussed methodology was tested at the dissemination area level as well as at the census tract level. In this paper we chose to present mainly the analyses conducted at the census tract level: at this spatial resolution, we believe, the spatial dependence is more severe, hence there is a stronger need to implement efficient spatial regression models; in addition, relationships among variables can more easily be identified at this scale; most importantly, we believe that these relatively larger units are more meaningful in terms or urban planning and health policies, therefore a model calibrated at this scale is more useful and applicable.

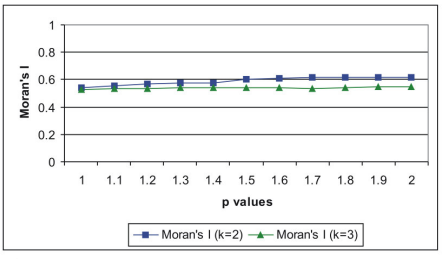
Initial exploratory analyses on the variables presented in section 2 reveals that generally variables are not normally distributed, and much of the multivariate relationships are driven by the magnitude of the population residing in each spatial unit. We therefore implemented a "normalization" of each variable, which in some cases involved the use of the total resident population as the standardizing variable (e.g., number of cardiac catheterizations), other cases 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). Descriptive spatial analyses (i.e., spatial autocorrelation indices) as well as multivariate aspatial[2] descriptive statistics (cross-correlations) and multivariate regressions produce more robust and meaningful results on the normalized variables. Unless otherwise indicated, all the results presented in this paper were obtained from normalized variables at the census tract level.

[2] Aspatial data refers to the association of data that is not spatially ascribed (age, gender) to spatial data (latitude and longitude). In spatial regression analysis, this data is used to investigate the relational processes contributing to the spatial distribution of the dependent variable based on the premise that similar attributes group together in space.

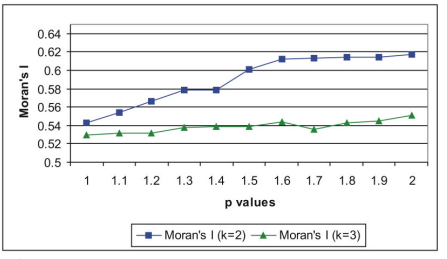
4.1 Spatial Autocorrelation Index

A spatial neighbourhood weighting matrix (section 3.1) is required for the computation of spatial autocorrelation indices, i.e., Moran's I. The correct specification of this matrix is the key of our proposed method and will be further discussed in the SAR model context. At the initial stage, we experimented with several orders of neighbourhood, different spatial weights, and alternative distance metrics. All else being equal, we pursued the weight matrix that produces the highest value of the spatial autocorrelation index, i.e., Moran's I, in the belief that this is the neighbourhood specification that best captures the spatial dependencies in the variables of interest.

Having chosen a nearest neighbourhood specification, we experimented with increasing orders of neighbourhood, and found that the spatial autocorrelation index is constantly higher for lower orders of neighbourhood, suggesting that the spatial dependence is more pronounced over short distances. We also tested different distance weights: we assigned no weights, used a standardizing variable[3], used various distance decay functions, and finally included the area of each census tract as the standardizing variable[4]: a squared inverse distance weight appears best capture the distance decay effect. This confirms the indication emerging from the inverse relationship between neighbourhood order and spatial autocorrelation index, suggesting that overall spatial dependencies are stronger over small areas, and decrease sharply as distance increases and more spatial units are considered. Based on these findings, we conducted further analyses for one and two orders of neighbourhood[5].



3a



3b

Fig. 3. 3a depicts the spatial autocorrelation index for varying distance metric values (p) according to two specifications of nearest neighbourhood range (k=2 and k=3). 3b depicts the same index at a scale that highlight the changes of the spatial autocorrelation over the varying p values.

Fig. 3 represents the variation of the spatial autocorrelation index as a function of the p value that defines the distance metric. Fig.3a evidences that the difference between one and two nearest neighbours is relatively minor, indicating that the method

[3] We tested "population aged 65 and older", for its high correlation with the disease incidence.

[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.

is robust with respect to the choice of a neighbourhood order. Fig.3b highlights the local features of the spatial autocorrelation function: for k=2 an interesting 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. While presenting a more stable trend, for k=3 the function also shows an anomaly, or a peak, at p=1.6, drops at p=1.7, and then rises again constantly until p=2.0. This initial analysis supports our hypothesis that the measured autocorrelation index is affected by the distance metric used in the definition of the spatial weight matrix. The value p=1.6 emerges as the candidate metric that can best capture the spatial dependence in the data.

The effect of the distance metric and specifically the different selection of nearest neighbours operated by alternative metrics can be better appreciated visually. In Fig. 4, we compare 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). The following figure presents two orders of neighbourhood (k=3), as the visualization results are most effective for this value.

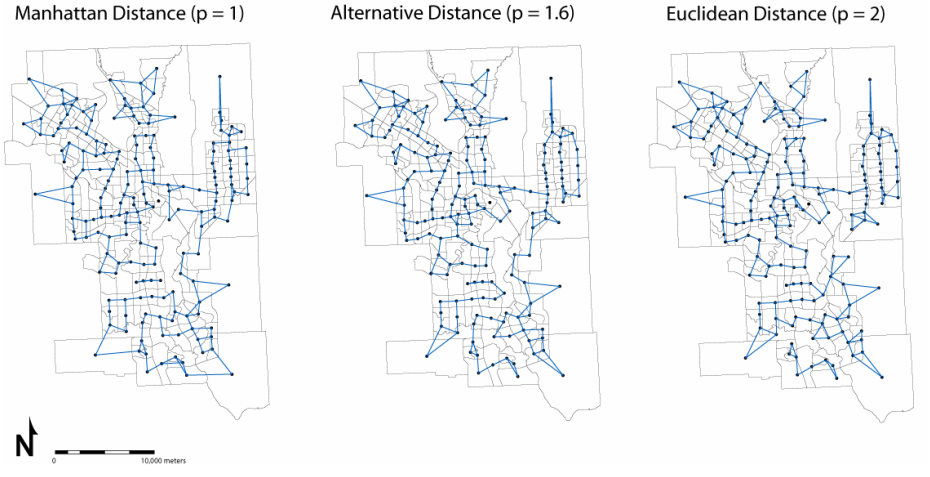


Fig. 4. Second order nearest neighbour connections for urban communities in Calgary, Alberta according to varying distance metrics and their respective p values

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 Regression Model Specification

A pool of 28 census variables was originally selected from the 2001 census: 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 1 summarizes descriptive spatial and non-spatial statistics on the

dependent variable and the subset of variables used in the regressions presented in this paper[6]. 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. Select descriptive statistics for the census data variables used in the analysis of cardiac catheterization cases

*** 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

An analysis of the cross-correlations among these variables provides an exceptionally informative portrait of the socio-economic structure of Calgary. Unfortunately, we were not able to effectively summarize the 28*28 correlation matrix in a suitable form for this paper. In Table 2 census variables have been grouped into homogeneous categories and only a sample of 2 representative variables for each category is presented.

This sample of socio-economic variables highlights the correlations across groups more so than within groups. As an example, the correlation between "owning a house" and "married or in common law" or "single detached home" suggests a predominant traditional family model, and a widespread wealth. The high correlations among most of the variables imposed serious constraints upon the choice of variables to be entered in the regression model. We have tested several models and considered some alternative approaches, including data reduction techniques, i.e., Factor Analysis, but presenting these results lie beyond the scope of the present discussion. In interpreting the following regressions, we would like to point out that while the cross-correlations limited our choice of independent variables, these same high cross-correlations allow the variables that were eventually included to be representative of those that could not be directly entered in the regressions. Therefore, the models are conceptually richer and more meaningful than it may appear at first sight.

[6] Unless otherwise specified, spatial statistics on any variable are conducting using Euclidean distance (p=2.0) and two orders of nearest neighbours (k=3, following the convention used in Splus).

Table 2. Cross-correlations among census data variables used in the analysis of cardiac catheterization cases

		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

After experimenting with an array of combinations of independent variables and after performing a backwards selection process, we chose the regression model described in equation(5): a simple model, which aims at maximizing the goodness-of-fit.

$$CC = f(a65pl, a55.64, males) \tag{5}$$

Where:

CC = number of catheterization cases;
a65pl = number of persons aged 65 years and older;
a55.64 = number of persons aged 55 to 64 years;
males = number of males.

Table 3. Spatial regression results for explaining cardiac catherdization cases with independent variables determined by a backwards selection process starting from 28 initial census variables

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.6055	0.6370	-4.0904	0.0001
n.a65pl	0.0637	0.0041	15.6500	0.0000
n.a55.64	0.0475	0.0081	5.8324	0.0000
n.males	0.0597	0.0126	4.7522	0.0000

L.likelihood	Pseudo R2	rho	sig^2	Res.Std. Err	Std. Err	Res. Moran
-212.9376	0.7339	0.2237	0.0487	0.2434	0.1010	-0.0087

Table 3 summarizes a selection of regression diagnostics. This regression presents some interesting aspects: it possesses a remarkably good fit (pseudo-R^2 = .73)[7], and it confirms the well known relationship between heart disease and

[7] Following Anselin [14], the pseudo- R^2 is calculated as the square of the correlation between observations and regression fit.

demographic factors such as age and gender. This is an important result, and its validity in a spatial model indicates the presence of fringes of population at higher risk with specific location in the urban setting. The model, however, remains in part unsatisfactory for statistical as well as conceptual reasons. From the statistical standpoint, the relatively low value of the rho parameter suggests that the autoregressive component of the model fails to capture most of the spatial dependence in the data. Considering the high cross-correlation between the dependent and some of the independent variables[8], we suspect that high spatial cross-correlations also exist among these variables, and the effect of these correlations is reflected in the rho value. We have not tested for spatial cross-correlations, as we were not prepared to address them in a comprehensive spatial model, but we believe that this is an important line of future enquiry. From a conceptual standpoint, one important goal of this work is to identify socio-economic variables, not simply demographic variables, which can help identify social and economic factors found in association with the disease incidence. A model that includes such variables would better assist in the definition of more effective social policies, the provision of appropriate health services, and most importantly would help identify localized pockets of risk beyond the known demographic factors. Using standard model selection procedures, the demographic variables remain the only significant variables in any multivariate specification. Therefore, in order to force other variables into the model, we deliberately omitted the demographic ones, obtaining a model that necessarily has a lower goodness-of-fit, but bears perhaps a greater value from a planning and policy making point of view. Our second model is detailed in equation (6).

$$CC = f(2p.w.chld, n.uni, f.m.inc, gr13ls) \tag{6}$$

Where:

CC = number of catheterization cases;
2p.w.chld = number of 2 parent families with children at home;
n.uni= number of persons with a post-secondary, non-university degree
f.m.inc= family median income
gr13ls= number of persons with grade 13 or lower education.

The new model diagnostics are summarized in Table 4. This model describes the incidence of heart disease as a function of family structure, education, and income. Even though we deliberately excluded the variables with the highest explanatory power, the model still explains a large portion of the observed variable (pseudo-R^2 = .37). The significance is relatively constant across variables (*t* value), unlike in the previous model, where retirement age is by far the most significant variable. The negative and highly significant coefficient of families with children 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, likely residing in the suburbs.

[8] Variables depicting age.

Table 4. Spatial regression results for explaining cardiac catherdization cases with independent census variables representing socio-economic indicators

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	1.7981	0.3648	4.9294	0.0000
n.2p.wchld	-0.0226	0.0032	-6.9717	0.0000
n.non.uni	-0.0163	0.0056	-2.9257	0.0039
f.m.inc.k	0.0089	0.0032	2.8138	0.0055
n.gr13ls	0.0195	0.0043	4.5262	0.0000

L.likelihood	Pseudo R2	rho	sig^2	Res.Std. Err	Std. Err	Res.Moran
-277.8000	0.3724	0.8504	0.1165	0.3413	0.1010	-0.0293

The positive relationship between disease incidence and low education (less than grade 13) may relate to represents old age and fringes of poverty and low social status. The negative coefficient linking disease incidence and individual with post-secondary, but non university education identifies trade workers and professionals: a category with fairly high income levels, possibly lower than those of individuals with university degrees. We would like to note that the variable "individuals with university degrees" correlates highly (and negatively) with "individuals with grade 13 or lower education" and therefore cannot be entered in the same regression; however, alternative model specifications present a significant and negative coefficient of "university degree". Overall, the education variables indicate that higher education levels are associated with greater income but lower disease incidence, suggesting higher education levels may lead to healthier lifestyle and lower risk of disease. Finally, the positive relationship between disease incidence and income suggests higher incidence in individuals with higher levels of stress and responsibility and appears related to more mature professionals, therefore suggesting a latent age factor.

Once the final regression has been identified and considered satisfactory from a statistical and a conceptual standpoint, we recomputed the spatial contiguity matrix for each p value in the $[1 \leq p \leq 2]$ interval and compared some key spatial indicators: the rho value, to assess the importance of the autoregressive coefficient; the pseudo-R^2, and some key indicators of the model's variance, including variance and standard errors. All these indicators have been scaled and plotted in one single graph, along with the Moran's I values already presented in Fig. 3. The trends presented in Fig 5 confirm that all the indices examined are affected by variation of the distance metric, or p value, supporting our proposition 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.

We cannot consider our results conclusive, but from the majority of our tests the value p=1.6 emerges as the best candidate for the optimal distance metric. Fig. 3 and Fig. 5 show a progressive increase in the spatial autocorrelation index for increasing p values, but the most noticeable growth occurs at p=1.6, and after this point it levels off. Even more intriguing is the peak displayed by the rho value, which suggests that with that metric the autoregressive coefficient is most effective at capturing the spatial dependence. This is confirmed by the corresponding trough in all the variance indicators examined. These improvements are far more meaningful than the slight decrease in the pseudo-R^2 value.

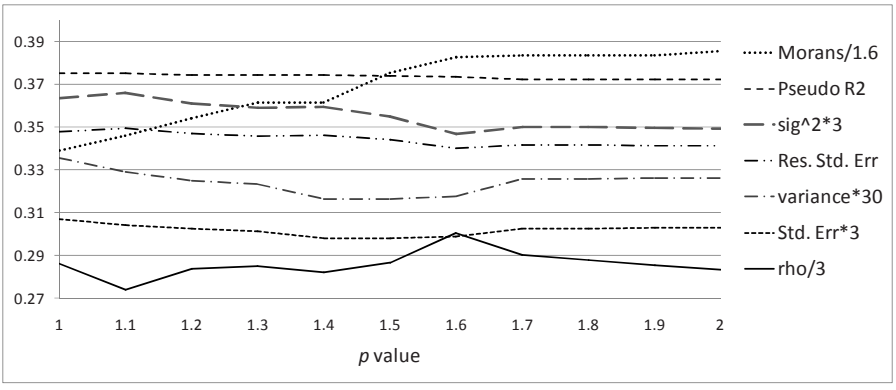


Fig. 4. SAR model indices for varying p values. Some indices are scaled for comparative purposes.

5 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 metric 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 for meaningful spatial units, our research provides effective support for spatial decisions by reliably identifying the parameters of spatial processes involving health, society, and the environment.

Further research is required to jointly address spatial dependencies in the dependent as well as the whole set of independent variables. A thorough investigation of urban design and neighbourhood connectivity should aid the interpretation of the optimal p value emerging from the statistical analysis; this line of research is likely to produce generalizeable conclusions, applicable to other urban patterns and geographical phenomena. We envisage the extension of this work to include local analyses, which can improve the reliability of the model by addressing non-stationarities in the observed relationships. Finally, a procedure to "semi automate" the selection of the optimal distance metric will enhance the usability of our proposed method.

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 APPROACH project researchers for providing us with data and support for our work. We also appreciate the contributions and suggestions of all the students who helped us with this project 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. Ahlbom, A., Norell, S.: Introduction to Modern Epidemiology. Epidemiology Resources Incorporated (1984)
4. 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)
5. Anselin, L.: Under the Hood. Issues in the Specification and Interpretation of Spatial Regression Models. Agricultural Economics, 27(3), 247–267 (2002)
6. Cressie, N.: Statistics for Spatial Data. Wiley, New York (1993)
7. Fotheringham, A.S., Brundson, C., Charlton, M.: Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, Chichester (2002)
8. 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. (eds.) Geographical Information Systems: Principles and Technical issues, vol. 1, pp. 267–282 (1999)
9. Getis, A., Aldstadt, J.: Constructing the Spatial Weights Matrix Using a Local Statistic. Geographical Analysis 36, 90–104 (2004)
10. 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)
11. Bailey, T., Gatrell, A.: Interactive Spatial Data Analysis. Wiley, New York (1995)
12. Haggett, P., Cliff, A.D., Frey, A.: Locational Analysis in Human Geography. Edward Arnold, London (1977)
13. Krause, E.F.: Taxicab geometry. Addison-Wesley, Menlo Park, California (1975)
14. Anselin, L.: SpaceStat tutorial. Regional Research Institute. West Virginia University. Morgantown, West Virginia (1993)