Appl. Spatial Analysis (2010) 3:1–23
DOI 10.1007/s12061-009-9027-7

A Spatial Analysis of the Demographic and Socio-economic Variables Associated with Cardiovascular Disease in Calgary (Canada)

Stefania Bertazzon · Scott Olson · Merril Knudtson

Received: 21 January 2009 / Accepted: 12 June 2009 /
Published online: 10 July 2009

Abstract The association between cardiovascular disease and a pool of demographic and socioeconomic variables is analyzed, for a large Canadian city, by means of multivariate spatial regression analysis. The analysis suggests that the spatial dependence observed in the disease prevalence is driven by the spatial distribution of senior citizens. A spatially autoregressive specification on a pool of solely socio-economic variables produces a model whose main predictors are family status, income, and educational attainments. This model can provide an effective analytical tool to support policy decisions, because it identifies a set of socioeconomic, not simply demographic predictors of disease. These socioeconomic variables can be targeted by social policies much more effectively than demographic variables. A further analytical step recombines the significant explanatory variables based on their spatial patterns. Thus the model is used to identify areas of social and economic concern, and to enable the initiation of specifically localized preventative health measures. Owing to its generality, the method can be applied to other conditions and to analyze multivariate relationships involving not only socioeconomic variables, but also environmental factors.

Keywords Spatial autoregressive · Cardiovascular disease · Socio-economic determinants of health · Health policy

S. Bertazzon (✉) · S. Olson
Department of Geography, University of Calgary, 2500 University Dr. NW, Calgary, AB T2N 1N4, Canada
e-mail: bertazzs@ucalgary.ca

M. Knudtson
Department of Medicine and Community Health Sciences, University of Calgary,
2500 University Dr. NW, Calgary, AB T2N 1N4, Canada

Introduction

Many primary health concerns of contemporary western societies are inherently 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 and regional planning. Management and planning decisions are often supported by quantitative models; indeed the use of quantitative and statistical methods in health research is well established, and the application of spatial analytical methods has gained acceptance over the last several years (Elliott et al. 2000; Waller and Gotway 2003; Elliott and Wartenberg 2004). It has been argued that the integration of analytical and visual methods can improve the effectiveness of spatial analysis as a decision support tool for policy and management (Guo 2007). Further, it has been argued that, though some difficulties still exist, Geographic Information Science has a potential role in improving public health (Rushton 2000).

Despite its potential role assisting management and planning decisions, the application of quantitative methods to spatial data remains prone to estimate uncertainty, which stems from two intrinsic properties of geographical phenomena: spatial dependence (i.e., near things are more related than distant things) and spatial non-stationarity (i.e., inconstant variability over space) (Cliff and Ord 1981). Violation of either assumption inflates the variance—and hence the uncertainty—of the regression estimates, resulting in less reliable models (Anselin 1988). Two broad families of spatial analytical techniques have been developed to increase the reliability of traditional statistical analysis when applied to spatial data: spatial autoregressive methods to address spatial dependence, and geographically weighted methods to address spatial non-stationarity (Fotheringham et al. 2002; Legendre et al. 2002; Fortin and Dale 2005). Other spatial analytical methods exist, for example, Bayesian approaches (Besag and Green 1993), and multilevel models (Duncan and Jones 2000), among others. However, a simplistic application of such analytical methods will not necessarily lead to the best results. Spatial analytical methods are applied to data, i.e., to a simplified representation of the phenomena of interest. If analytical models are to be truly meaningful, their mathematical rigour must be complemented with a rich understanding of the phenomena under consideration.

This paper presents the use of multivariate regression analysis to identify demographic and socio-economic variables significantly associated with cardiovascular disease prevalence in a large Canadian city, Calgary. In consideration of the available data, i.e., spatially aggregated records, spatial regression techniques are applied, in the presence of spatial dependence, to enhance the reliability of the model parameters. A comparative analysis of spatial dependence in the crucial variables and in alternative model specifications provides an indication of the likely source of the dependence, allowing for a deeper understanding of the processes involved. The spatial distribution of the explanatory variables, along with the sign and value of their regression coefficients, provide a reliable indication of areas of the city where targeted health and social policies should be implemented, in order to effectively reduce the disease incidence among high-risk population.

While the paper focuses on one application, it also discusses a general method that can be applied to other conditions and extended to analyze their spatial relationship not only with socioeconomic variables, but also with environmental factors. Subsequently, Section "Background and Case Study" provides some background information and introduces the case study, Section "Methodology" outlines the methods employed, Section "Results" presents the analytical results, Section "Discussion" provides a critical discussion of the analyses, and Section "Conclusion" offers some conclusions along with future lines of work.

Background and Case Study

Cardiovascular diseases remain one of the leading causes of death in the developed world (Kaplan and Keil 1993; Manuel et al. 2003; Canadian Heart Health Strategy and Action Plan 2009). Disease occurrence is related to personal characteristics, such as age, gender, genetic background, and the simultaneous presence of other conditions; all these are known as non-modifiable risk factors. In addition to these, disease prevalence has been found in association with a number of modifiable risk factors, including stress, limited physical activity, smoking, high intake of calories, and high proportion of saturated fats. These modifiable risk factors, in turn, tend to correlate with demographic and socio-economic characteristics of individuals, such as age, occupation, or income, which can be measured by census variables (Diez Roux et al. 2001; Chaix et al. 2007; Augustin et al. 2008; Canadian Cardiovascular Outcomes Research Team Atlas 2009). At most geographical scales (and particularly at the urban scale), demographic and socio-economic characteristics tend to display a pattern, or spatial clustering; disease prevalence, likewise, tends to present a characteristic geographical distribution, or spatial pattern. For this reason a multivariate regression model is an effective tool to analyze the spatial pattern of disease occurrence as a function of localized demographic and socio-economic characteristics and of specific patterns of land use. The utility of the model is its capacity to identify distinct spatial patterns of disease with a consideration for localized socioeconomic variables in order to apply defined preventative social policy.

In a city that is expanding in both space and populace, effective resource allocation (hospitals, specialists, instruments, etc.) requires a thorough understanding of disease distribution, but also of the target population segments and their distribution in space. In addition, identifying demographic variables (e.g., age and gender) and socio-economic variables (e.g., income and education levels) most highly associated with disease prevalence helps to identify areas where higher disease incidence may be expected in the near future, hence suggesting locational decisions that could promote accessible health care services to the population at risk.

Calgary is one of the largest Canadian cities, with a population over 1 million (2008 civic census), and the economic centre of the province of Alberta, based on the tertiary activities induced by the mining of resources (oil and precious minerals) in the northern part of the province. Over its historical development the city has annexed many existing towns and it is now politically administered under a unicity concept. In recent years, an economic boom, driven by geopolitical factors, has increased the city's prosperity, with a consequent population increase, largely driven by internal immigration, mainly from some of the Canadian eastern provinces. This has contributed to a more pronounced tendency toward urban sprawl and a residential pattern characterized largely by single-family housing. Industrial productive activities are predominantly located in the eastern side of the city, along with a north-easterly located airport. The socio-economic structure is characterized by a relatively young population, with high average income and educational attainment levels.[1]

Calgary's geography and recent history make it a dynamic and diverse city, and therefore a suitable case study for a pilot multivariate regression model application. The city covers a large geographic area[2] and a proportionally large residential area, which is diverse in terms of housing type and neighbourhood age, residential density, accessibility to urban services and proximity to noxious facilities. Overall, there is a dichotomous geographic split between the residents with high income located more to the west and low income residents nearer the industrial east. A relatively low average age, coupled with a population predominantly formed by young families implies the coexistence of many diverse age groups, ranging from young children to the elderly. Over the last decades, alternating economic prosperity and decline have been followed by an overall radial urban development pattern, locally confounded with gentrification phenomena. As a consequence, the current socio-economic landscape is characterized by a distinct spatial pattern driven mainly by age, income, and education levels.

For all these characteristics, the city of Calgary constitutes a rich test bed to analyze the spatial association between disease occurrence and demographic as well as socio-economic variables. It should finally be noted that "In accordance with the Canada Health Act, Alberta has a publicly administered and funded health care system that guarantees Albertans receive universal access to medically necessary hospital and health care services", and that the modest health premiums paid by Albertans have been eliminated as of January 1, 2009 (Alberta Health and Wellness).

The medical data were provided by the APPROACH Project (Ghali and Knudtson 2000), a clinical registry, begun in 1995, that records information on all patients undergoing cardiac catheterization in Alberta. Cardiac catheterization is an invasive procedure for patients experiencing cardiovascular symptoms and defines coronary anatomy, left ventricular and valvular function; it provides important prognostic information for individuals affected by cardiovascular conditions (Ghali and Knudtson 2000). The dependent variable is obtained by selecting from the provincial database the records of Calgary residents (approximately 12,000) undergoing the procedure at the Foothills Hospital between 1998 and 2002;[3] patient address is released at the postal code level.[4]

[1] Median age is approximately 35 years, and median family income slightly over 63,000 CDN $, for the 2001 census.

[2] Approximately 750 Km^2.

[3] This temporal interval was defined in order to best match the census data (particularly in a long-term perspective) so that the analysis can be repeated for corresponding intervals as new census and medical data become available.

[4] The data consisted of complete postal codes, wherein the last three characters of the code identifies the Local Delivery Unit (LDU), A LDU in an urban environment services an average of 19 households specifically defined to one side of a road segment (Statistics Canada 2007).

Socio-economic and demographic variables are drawn from the 2001 census of Canada data. The analysis is conducted at the census tract level. The cardiac data were spatially aggregated using Census Canada's postal code conversion files (PCCF) to match the census data, resulting in approximately 180 valid census tract records. Figure 1b summarizes the distribution of catheterization cases over the entire study period and the census tracts[5] for Calgary. Figure 1a provides a synthetic representation of the city's major features and predominant land use: this shall serve as reference for the interpretation of the spatial distribution of the variables discussed throughout the paper.

The authors are aware of the limitations of the clinical database used in this work: specifically, cardiac catheterization records are not necessarily the most accurate representation of patients affected by cardiovascular disease. However, other clinical records, e.g., hospitalizations for symptoms of acute coronary syndrome, are only available, at the postal code level, over a very short period (2007–2008), which is also a poor temporal match for any Census survey. For these reasons, the catheterization database was preferred for this analysis. The limitations of the data should not impact on the value of the analysis, which is presented in this paper as a methodological exemplar.

Methodology

The prevalence of heart disease within a city can be modelled as a spatial process (Cressie 1993), and so can all the demographic and socioeconomic variables associated with the disease prevalence. Each of these observed processes is likely to display spatial dependence and non-stationarity. With reference to the disease prevalence, individuals living in the same or nearby neighbourhoods tend to have similar age, income, and access to health care, and consequently similar rates of disease prevalence: the process therefore displays positive spatial dependence. The process is non-stationary, because the disease prevalence is inconstant over space: prevalence rates vary from young and wealthy neighbourhoods to retirement communities (inconstant mean), the variability within a young neighbourhood is greater than in an older one (inconstant variance), and the spatial extent of the spatial dependence varies across the city, from densely populated central areas to recent suburban communities (inconstant covariance).

A multivariate regression model for the above spatial process should contain specific procedures to minimize the model's variance in the simultaneous presence of spatial dependence and non-stationarities. As illustrated by the example, in this type of process the two properties do not simply occur simultaneously, but they are also mutually related. Despite the known effects of this relationship (Tiefelsdorf 2003), most existing advanced spatial methods address only one of the properties: spatial autoregressive methods (Anselin 1988) focus on spatial dependence, but typically disregard non-stationarity; and local, or geographically weighted, methods (Fotheringham et al. 2002) focus on non-stationarity, but typically disregard spatial

[5] Urban areas with a population greater than 50,000 are subdivided into census tracts which are spatial units with populations ranging from 2,000 to 8,000.

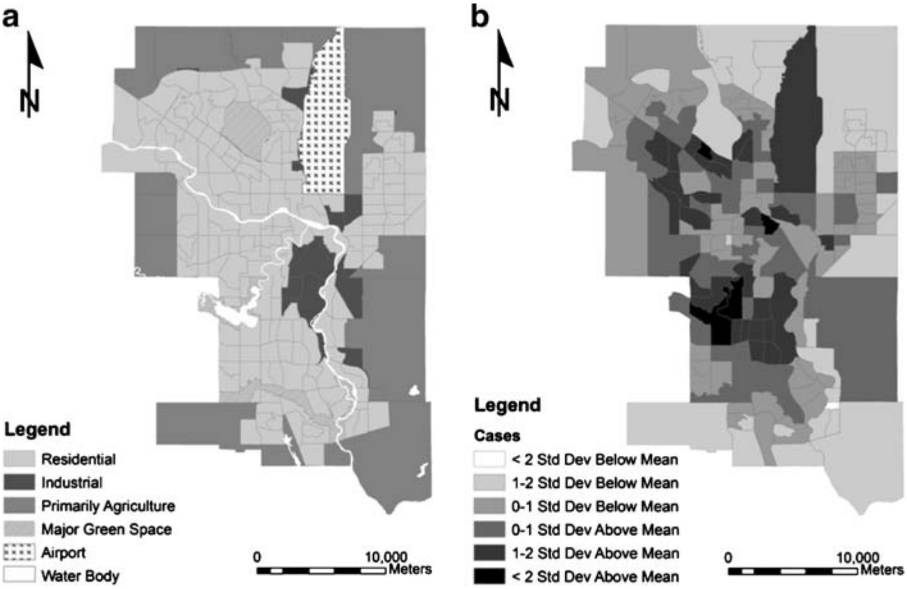

Fig. 1 The city of Calgary and cardiac catheterization cases, 1998–2002

dependence. Only recently have some methods been proposed to integrate local and autoregressive methods, for example, by Aldstadt and Getis (2006) and by Law et al. (2006).

The scope of this paper is limited to the application of spatial autoregressive procedures; the analytical implementation therefore begins with an examination of the spatial dependence in all the processes involved. Spatial autocorrelation measures based on the Moran's I (Getis 2008) are commonly used to test clustering tendency of medical data, including analysis in multivariate specifications (Lin and Zhang 2007). Likewise, throughout this paper the traditional spatial autocorrelation test, Moran's I (Getis 2008), will be used. While the authors are aware of the limitations of this index (Li et al. 2007), its results can be interpreted as a broad indication of the presence and magnitude of spatial dependence. The use of a single index for various analyses (i.e., exploratory analysis, individual variables, and model residuals) is important for the discussion of spatial dependence presented in this paper. The computation of this index requires the specification of a model of spatial dependence, defined by a contiguity, or spatial weight matrix, which can be a simple binary structure or a more complex specification, including various types of weights that describe distance decay effects. There are several ways of specifying spatial contiguity (Getis and Aldstadt 2004): 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 maximum number (*k*) of nearest neighbours.

Spatial autoregressive methods (Anselin 1988) 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. Generally, a constant covariance structure is assumed, and a spatial weight matrix determines which spatial units are spatially dependent (Cressie 1993). The model is expressed formally by Eq. 1:

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

where ρ (rho) is the autoregressive parameter and W is the spatial weight matrix. The autoregressive parameter represents a correlation coefficient, whose value can vary between −1 and +1. For the definition of the spatial weight matrix, the same considerations apply as for the calculation of the spatial autocorrelation index. For the application discussed in this paper, a simultaneous autoregressive (SAR) specification is used; the spatial weight matrix is the same one defined for the computation of the spatial autocorrelation index. A backwards model selection procedure is conducted for all the regressions. Following each regression specification, the spatial autocorrelation index is computed, in the described fashion, on the regression residuals.

Throughout the outlined methodology, a key role is played by the spatial weight matrix, which largely determines the value of the spatial autocorrelation index—in the variables as well as in the model residuals—and the efficacy of the spatial autoregressive models. However, defining a spatial weight matrix remains subjective, and rests on an estimate of the spatial dependence in the spatial processes involved. Previous work (Bertazzon and Olson 2008) has focused on the definition of an optimal spatial weight matrix: distance metric, number of nearest neighbours and distance decay function were considered. An array of spatial weight matrices was defined by altering the values of each of the three parameters. The specification that produced the highest value of Moran's I for the dependent variable was preferred, as this was considered the neighbourhood specification that best captures the spatial dependencies in the variable of interest. Using the same criteria, the spatial weight matrix used throughout this paper was chosen: the neighbourhood is defined by the two nearest neighbouring census tracts,[6] the distance is calculated following the Euclidean metric (Bertazzon and Olson 2008), and a squared inverse distance decay function is employed, with the census tract area as weight. The matrix thus defined is used to compute Moran's I for all the variables and all model residuals discussed.

For all the analyses presented, the variables are normalized, in most cases using the total resident population as denominators (e.g., "Number of cardiac catheterizations"), in other cases using specific denominators (e.g., "Population over 20" was used to standardize education levels and "Population over 15" to standardize marital status). Following this normalization, all the variables that originally were numbers become rates. The bivariate Pearson correlation is used to test the cross-correlation among variables. The dependent variable is standardized for age and sex, using inverse standardization. The standardization is conducted on the raw data, and

[6] k=3, following the convention used in S-PLUS.

the standardized variable is subsequently normalized like the other variables, using the total resident population as denominator.

Distributions are cartographically represented in standard deviations from the mean, with class width defined as a proportion of the standard deviation, and each variable subdivided into six classes (Evans 1977; Mennis 2006). This method was chosen because standard deviations allow for a direct comparison of population characteristic distributions of areas, whereas other methods such as Jenks or Quartiles do not maintain a consistent reference point between maps. Also, standard deviations identify extreme population characteristics rather than suppressing them in groups of equal interval: in this research the skewness of a distribution is important information for identifying the specific population pockets considered to be in dire need.

All the statistical computations are conducted in S-PLUS 7 and S-PLUS Spatial Statistics 1.5, with the exception of the bivariate Pearson correlations, which are computed in SPSS 15. Geographical data management and visualization are implemented in ArcGIS 9.1.

Results

Correlation Analysis

A pool of variables was selected from the 2001 census. Table 1 summarizes the name and definition of the variables used throughout all the analyses and reported in the following tables.

Table 1 Variables: categories, names, and definitions

Variable category	Variable name	Variable definition
Dependent variable	*cases*	Cardiac catheterization cases
Demographic variables	*males*	Male residents
	age 45–54	Residents aged between 45 years and 54 years
	age 55–64	Residents aged between 55 years and 64 years
	age over 65	Residents aged 65 years or older
Family variables	*families*	Families of two parents with children
	couples	Residents married or living in common law
	single parents	Single parent families
Housing variables	*owned*	Owned dwellings
	detached	Single family detached units
Education variables	*secondary*	Residents with grade 13 or lower education
	non-university	Residents with post-secondary, non university education
	university	Residents with university education
Economic variables	*income*	Family median income
	unemployment	Unemployment rate

Initial exploratory analyses on these variables have indicated that most of them do not follow a normal distribution, and in addition, preliminary regression analyses have suggested that apparent multivariate relationships may be driven by the magnitude of the variables recorded in each spatial unit. For both reasons, normalization is conducted on all pertinent variables.[7] In the interest of brevity, the following tables do not present statistical analyses on the entire pool of variables, but on the subset of variables that remain significant in the regressions presented throughout the paper. Table 2 summarizes the values of descriptive statistics computed on the dependent, standardized dependent, and selected explanatory variables.

Based on the descriptive statistics presented in Table 2, it can be concluded that the standardized variables are normally distributed. Table 3 summarizes the values of the spatial autocorrelation index for the dependent, standardized dependent, and selected explanatory variables. All the variables present significant and generally high spatial autocorrelation values; specifically, the value for the dependent variable ("*cases*") is 0.62, and 0.70 for the standardized dependent ("*std.cases*"); the greatest values (above 0.8) are displayed by the variables "*families*" and "*secondary*", while the lowest value is shown by the variable "*non university*" (Table 3). A cross-correlation analysis among independent variables is used to identify sets of uncorrelated variables on which regression models and model selection procedures are run (Table 4). Due to the difficulty of effectively conveying the information contained in the complete correlation matrix, in Table 5 the census variables are grouped into homogeneous categories (defined in Table 1) and only a sample of two representative variables for each category is presented.

The cross-correlations provide an exceptionally informative portrait of the socio-economic structure of the city of Calgary. As an example, the correlations between "*owned*", "*couples*", and "*detached*" suggests a predominant model of traditional family, a relative stability and wealth, and a characteristic urban pattern (Section "Background and Case Study"). Conversely, the demographic variables present low cross-correlations, which allow for the inclusion of a rich set of age groups in the regression model.

The high cross-correlations imposed severe limitations on the choice of independent variables for the regression models, but at the same time those cross-correlations imply that the variables that are eventually entered in the regressions are also representative of those that are not directly entered: therefore, the models are conceptually richer and more meaningful than the set of independent variables may suggest.

Regression Analysis

The first analytical test is a regression including socio-economic as well as demographic variables. In this model the dependent variable is normalized, but not standardized (Section "Methodology"). Given the high cross-correlations among most of the independent variables,[8] an array of combinations was tested and a

[7] Such variables as median family income or male/female ratio were not normalized.

[8] Alternative approaches were explored, for example the use of data reduction techniques, such as factor analysis. While these alternative analyses are beyond the scope of this paper, exploratory tests have indicated that their results are overall consistent with those discussed in this paper.

Table 2 Descriptive statistics

Summary statistics for data in Master.CT.2001

	Mean	Median	Variance	Std. dev.	SE mean	Skewness	Kurtosis
cases	1.34	1.28	0.22	0.47	0.03	0.40	−0.29
std. cases	1.34	1.31	0.18	0.42	0.03	0.38	−0.15
males	49.77	49.80	2.76	1.66	0.12	0.01	2.44
age 45–54	14.46	13.92	10.00	3.16	0.24	0.60	0.40
age 55–64	7.61	7.24	6.53	2.55	0.19	0.77	0.32
age over 65	9.64	8.16	32.66	5.71	0.42	0.81	0.11
families	47.31	48.18	181.27	13.46	1.00	−0.18	−0.53
secondary	30.64	28.47	122.98	11.09	0.82	0.64	−0.34
non university	36.68	37.04	29.90	5.47	0.41	−0.39	0.11
university	32.67	31.63	183.12	13.53	1.01	0.22	−0.58
income	66.61	63.13	330.68	18.18	1.35	0.61	−0.56

backwards selection procedure was applied on each of them. In all cases, the selection procedure led to the same subset of significant variables, converging to the model described in Eq. 2 (see Table 1 for a definition of the variables in the equation).

$$cases = [males, age45-54, age55-64, age-over-65, university]\beta + \varepsilon \qquad (2)$$

The main regression results and a selection of regression diagnostics are summarized in Table 6. The variables are listed in decreasing order of significance.

Table 3 Spatial autocorrelation on dependent and explanatory variables

Moran's I

	Correlation	Variance	Std. error	Normal statistic	Norm. p-val. 2-sided
cases	0.62	0.01	0.10	6.17	0.00
std. cases	0.70	0.01	0.10	6.96	0.00
males	0.47	0.01	0.10	4.73	0.00
age 45–54	0.48	0.01	0.10	4.80	0.00
age 55–64	0.57	0.01	0.10	5.68	0.00
age over 65	0.73	0.01	0.10	7.25	0.00
families	0.86	0.01	0.10	8.60	0.00
secondary	0.82	0.01	0.10	8.16	0.00
non university	0.37	0.01	0.10	3.69	0.00
university	0.70	0.01	0.10	6.97	0.00
income	0.63	0.01	0.10	6.25	0.00

Table 4 Bivariate correlations between dependent and explanatory variables

		Dependent variables	
		cases	std.cases
Demographic variables	*males*	−0.14	−0.34
	age 45–54	0.04	0.18
	age 55–64	0.57	0.69
	age over 65	0.79	0.93
Family variables	*families*	−0.50	−0.43
	couples	−0.38	−0.29
	single parents	0.32	0.10
Housing variables	*owned*	−0.29	−0.20
	detached	−0.27	−0.22
Education variables	*secondary*	0.18	−0.12
	non-university	−0.24	−0.41
	university	−0.05	0.26
Economic variables	*income*	−0.23	0.05
	unemployment	0.17	0.06

The results of this regression are important from a theoretical perspective as they confirm the well known results of medical research, i.e., age and gender (male) are key factors in cardiovascular disease prevalence (Gerber et al. 2006). The high value of the R^2 index (0.7896) suggests that the model (through this set of variables, which point mostly to non-modifiable risk factors) explains almost 80% of the variation of the disease prevalence. Among the regression diagnostics, the variance indicators (σ^2 and residual standard error) present relatively low values, and the spatial autocorrelation index calculated on the regression residuals (Residual Moran) presents a low and insignificant value of 0.0039 (normal statistic=0.1102; p-value=0.9122). This suggests that, even though the dependent variable, and most explanatory variables, present significant spatial autocorrelation values (Table 3), there is no evidence of spatial dependence in the regression residuals. This is also corroborated by the low values of individual variables' standard errors. Consequently, the assumptions of the regression model are not violated by the presence of spatial dependence, and parameter estimates can be considered reliable.

This set of explanatory variables points almost exclusively to non-modifiable risk factors, demographic in nature, all linked to the dependent by positive coefficients. The variable expressing retirement age, "*age over* 65", is by far the most significant (or *leading*) variable in the model. However, the regression also contains one variable which is not demographic: the number of residents with university education. This is the second variable by significance, with a negative coefficient, describing a negative relationship between education and disease prevalence. The spatial distribution of the dependent variable and the leading explanatory variable are shown in Fig. 3, and Fig. 4 presents the spatial distribution of the remaining (or *secondary*) explanatory variables.

Table 5 Cross-correlation analysis

	Demographic		Family		Housing		Education		Economic	
	Age 55–64	Age over65	Families	Couples	Owned	Detached	Secondary	Non-university	Unemployment	Income
age 55–64	1	0.42	0.05	0.00	0.14	0.05	−0.03	−0.29	0.09	0.20
age over65	0.42	1	−0.42	−0.16	−0.36	−0.35	−0.10	−0.33	0.04	−0.10
families	0.05	−0.42	1	0.24	0.91	0.87	−0.22	0.13	−0.37	0.67
couples	0.00	−0.16	0.24	1	0.19	0.15	−0.22	0.03	-0.06	0.27
owned	0.14	−0.36	0.91	0.19	1	0.91	−0.13	0.10	−0.35	0.65
detached	0.05	−0.35	0.87	0.15	0.91	1	−0.12	0.09	−0.32	0.59
secondary	−0.03	−0.10	−0.22	−0.22	−0.13	−0.12	1	0.25	0.34	−0.70
non-university	−0.29	−0.33	0.13	0.03	0.10	0.09	0.25	1	−0.16	−0.29
unemployment	0.09	0.04	−0.37	−0.06	−0.35	−0.32	0.34	−0.16	1	−0.37
income	0.20	−0.10	0.67	0.27	0.65	0.59	−0.70	−0.29	−0.37	1

Table 6 Standard regression model

	Value	Std. error	t value	Pr(>|t|)
(Intercept)	−2.0674	0.5917	−3.4940	0.0006
age over 65	0.0714	0.0040	17.6705	0.0000
university	−0.0096	0.0014	−6.8076	0.0000
males	0.0492	0.0113	4.3519	0.0000
age 55–64	0.0343	0.0092	3.7343	0.0003
age 45–54	0.0228	0.0074	3.0977	0.0023
	R^2	**Sigma^2**	**Res. Std. Err**	**Res. Moran**
	0.7896	0.0463	0.2189	0.0039

Perhaps the most interesting aspect of Fig. 2 is that the distribution of retirees “*age over 65*” (Fig. 2b) spatially overlays almost exactly the distribution of disease prevalence “*cases*” (Fig. 2a).

In Fig. 3a, the variable “*university*”, in consideration of its negative sign, is represented by a reverse colour scheme, with lighter shades representing higher values. Of the four variables, this is the one that presents the most interesting spatial distribution, with a constantly strong presence in the western part of the city, and consistently low values in the eastern part. It is worthwhile observing the clustering of this variable around the main post-secondary institutions, easily identified as the spatial units displaying the lightest colour.[9] The variable “*males*” (Fig. 3b) presents peaks in the downtown area, followed by a relatively strong presence in the northeast: this pattern is likely related to young, wealthy, single males, with high job positions in the downtown core, and blue-collar jobs in the north-east. This demographic pattern is likely related to recent internal immigration and temporary positions induced by the booming Alberta oil industry (Section “Background and Case Study”).

Based almost entirely on retirement age, marginally complemented by mostly other demographic variables, this model is unlikely to provide an effective tool for early detection of high risk population, based on socio-economic factors. As an alternative, a second regression is proposed, which was obtained by deliberately excluding all the demographic variables from the original pool of census variables. Experiments were conducted using the normalized dependent variable and its age- and sex-standardized version. In all cases, the use of either variable leads to the same final model, with minor differences in the values of the regression coefficients and the diagnostics. In all the regressions presented in the remainder of the paper the dependent variable is the age- and sex- standardized variable. Following the procedure outlined above, alternative regressions were tested and a backwards model selection procedure was implemented. Again, all the alternative specifications converged to the model defined by Eq. 3.

$$std.cases = [families, secondary, non - university, income]\beta + \varepsilon \tag{3}$$

[9] The University of Calgary and Mount Royal College, and the Southern Alberta Institute of Technology: the University of Calgary (in the Northwest) and Mount Royal College (with a more southern location) are most easily discernible in Fig. 3a.

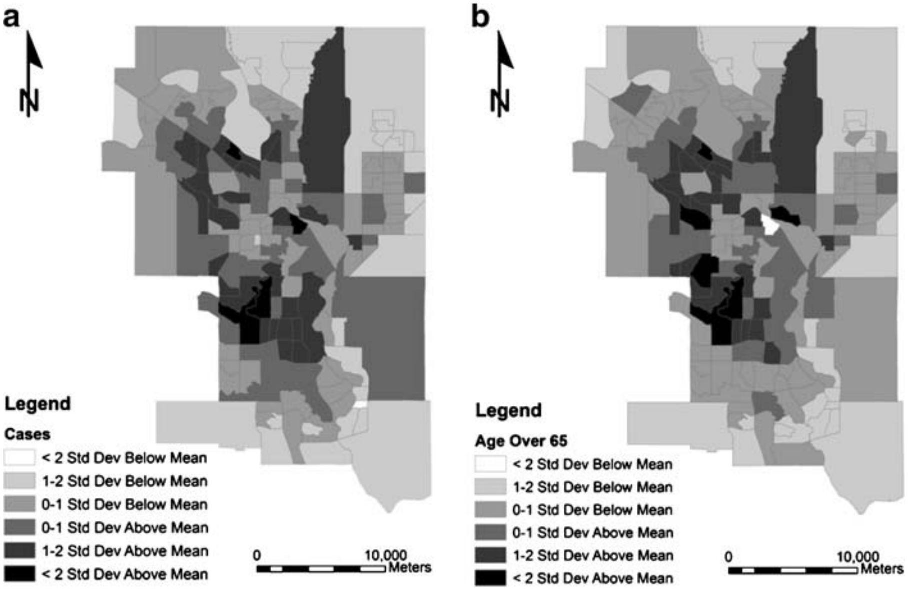

Fig. 2 Standard regression model: dependent and leading explanatory variable

The variables in this regression make it a more promising analytical tool; however, the spatial autocorrelation index calculated on the regression residuals shows a value of 0.6286, which is significant[10] and largely increased in comparison with the previous model. This suggests that the removal of the demographic variables has caused an increase in the spatial dependence in the model's residuals, with consequent loss of reliability of its estimates. The parameter estimates produced by this model are obtained in violation of one of the main model assumptions, and for this reason they cannot be trusted; therefore, the complete regression coefficients and diagnostics shall not be presented. Considering the potential practical value of this model, a spatial regression model will be alternatively specified on the restricted pool of variables.

The application of the usual cross-correlation and model selection procedures, but within a spatially autoregressive specification, produces the model described in Eq. 4.

$$std.cases = [families, secondary, non - university, income]\beta + \rho W[cases]\varepsilon \quad (4)$$

Again, alternative combinations of variables led consistently to this subset of significant variables, in addition to the autoregressive element. The main regression results and diagnostics are summarized in Table 7.

The most important aspect of this model is the value of the spatial autocorrelation in the regression residuals, which attests that the spatially autoregressive specification produced a drastic reduction of the index, from 0.6286 in the residuals of the standard regression model (Eq. 3), to -0.02681 (normal statistic=−0.2745; p-value=

[10] Normal statistic=8.193; p-value=2.556×10^{-4}.

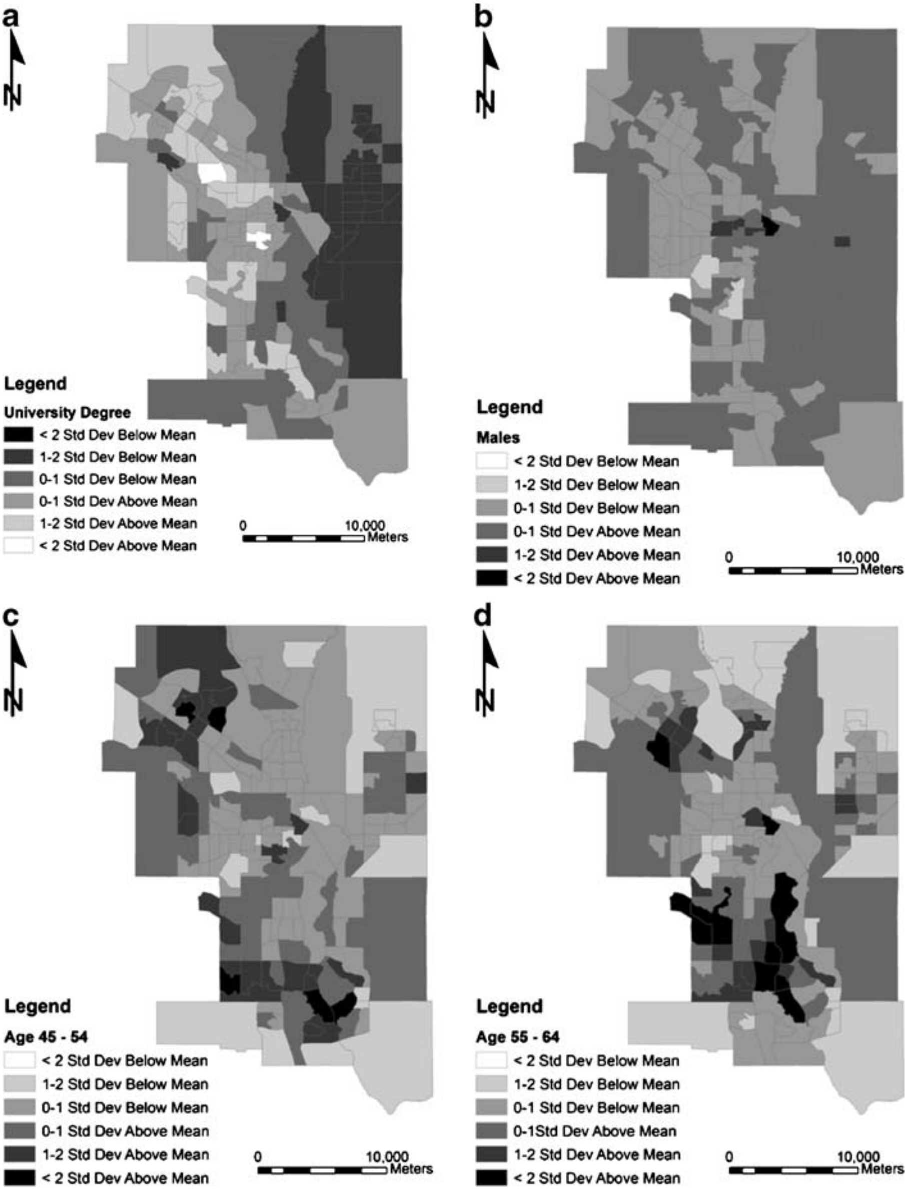

Fig. 3 Standard regression model: secondary explanatory variables

0.7837) in the residual of the spatial regression model (Eq. 4). The effectiveness of the spatial autoregressive specification is confirmed by the high value (0.94) of the autoregressive parameter, rho. The spatial autoregressive procedure has rid the model of the spatial dependence in the residuals, so that the regression hypotheses are now met and the reliability of the parameter estimates is restored.

An important observation emerges from the comparison of the spatial autocorrelation in the dependent variable and in the residuals of the various models. The

Table 7 Spatial regression model, excluding non-modifiable risk factors

	Value	Std. error	t value	Pr(>|t|)	
(Intercept)	1.9705	0.3114	6.3270	0.0000	
families	−0.0188	0.0027	−7.0132	0.0000	
non-university	−0.0245	0.0044	−5.5963	0.0000	
income	0.0106	0.0025	4.2718	0.0000	
secondary	0.0130	0.0036	3.6283	0.0004	
L.likelihood	**Pseudo-R^2**	**Rho**	**Sigma^2**	**Res. Std. Err**	**Res. Moran**
−229.0000	0.3198	0.9406	0.0599	0.2483	−0.0268

value of the spatial autocorrelation index is 0.62 for the dependent variable, and 0.70 for its standardized version; 0.004 for the residuals of the standard regression on the raw variable including demographic variables (Eq. 2); 0.63 for the residuals of the standard regression on the standardized variable[11] excluding demographic variables (Eq. 3); and −0.003 for the residuals of the spatial regression on the standardized variable excluding demographic variables (Eq. 4). This seems to suggest that the spatial dependence observed in the dependent variable is induced by the spatial dependence in the demographic variables (Table 3), which are also the variables most highly correlated with the dependent. The explicit inclusion in the model of these variables compensates for the spatial dependence in the dependent variable, resulting in a model (Eq. 2) that even in a standard specification presents insignificant spatial dependence in the residuals. The exclusion from the model of the variables causing the dependence results in significantly spatially dependent residuals (Eq. 3), that can only be corrected by a spatially autoregressive specification (Equation 4). This consideration may be important to understand the nature of the spatial dependence observed in the variable, and to identify the processes that may have caused it. For this reason alone, a model that is mostly based on demographic variables (Eq. 2) is unlikely to increase the spatial knowledge of the process of interest, as its explanatory variables are likely to present the same spatial clustering observed for the dependent variable (Fig. 2).

These considerations cast a new light on the measurement of spatial autocorrelation and the nature of the spatial dependence observed on the dependent variable. The criteria used to define the spatial weight matrix for these models are currently centered on the dependent variable (Section "Methodology"): in light of the above considerations, alternative criteria, focusing on the demographic variables instead, shall be explored. Likewise, these considerations impact on the meaning of the observed dependence and the conceptualization of space underlying these models, suggesting new lines of investigation.

The spatially autoregressive model (Eq. 4) describes the prevalence of cardiovascular disease as a function of family structure, education, and income. Even without the demographic variables, which have the greatest explanatory power, the model still

[11] The value is 0.5011 for the same regression on the raw variable.

explains a large portion of the variation of the disease prevalence (pseudo-R^2=0.32).[12] This result has important theoretical and practical implications, as it suggests that over 30% of the disease prevalence can be explained by socioeconomic variables only, even though demographic factors are implicitly accounted for by these variables. As shown in Table 7, the significance (t value) is relatively constant across variables, unlike in the standard regression model, where retirement age is by far the most significant of all variables. The spatial patterns of this set of variables are shown in Fig. 4. Again, the variables with positive coefficients are represented according to the traditional convention, where darker colours represent higher values, while for variables with negative sign, the colour scheme is reversed.

The negative and highly significant coefficient of the variable "*families*" (Fig. 4a) suggests that the disease prevalence is lowest in areas characterized by the predominance of families with children, likely formed by fairly young individuals, at early to middle stages of their career, with relatively high education and moderately high income, as these are the main traits of the spatial pattern of this variable. The negative coefficient linking disease prevalence and "*non-university*" (Fig. 4b) identifies areas of low disease prevalence in association with strong presence of residents with a post-secondary, technical education: trade workers and professionals, i.e, a category with fairly high income levels, not necessarily lower than those of individuals with university degrees.[13] The positive relationship between disease prevalence and lower education or "*secondary*" (Fig. 4d) seems to point to fringes of poverty and low social status, possibly related also to old age. Mirroring the variable "*university*" (Fig. 3a), discussed within the standard regression model, this variable displays a remarkably high presence in the eastern part of the city. Overall, the education variables suggest that higher education attainment levels are found in association with lower disease prevalence; from this it may be inferred that higher education levels tend to be negatively associated with modifiable risk factors, likely a consequence of a healthier lifestyle characterized, for example, by higher levels of physical activity, lower consumption of tobacco and alcohol, and healthier dietary choices. Finally, the positive relationship between disease prevalence and "*income*" (Fig. 4c) suggests higher disease prevalence in higher income areas, and may point to areas inhabited by mature professionals, possibly implying also a latent age factor. The relationship between disease prevalence and "*income*" also appears to be affected by extreme values of the latter, i.e., areas of very high and very low income. This may be explained by the consideration that the variable refers to median family income, which tends to present the highest values for one-person families. Consistently, the cross-correlation analysis has identified a link between disease and various categories of lonely persons, ranging from singles with very high income to single parents, and divorced, separated, or widowed persons. The high value of the variance of "*income*" (Table 2) is another indication of the weight exerted by extreme values of this variable.

[12] Following Anselin (1993), pseudo-R^2 is defined as the squared correlation between observed and estimated values of the dependent variable.

[13] It should be observed that the variable "*university*" correlates almost perfectly (and negatively) with "*secondary*". The choice of entering either variable in the various models was dictated by the correlations with other explanatory variables.

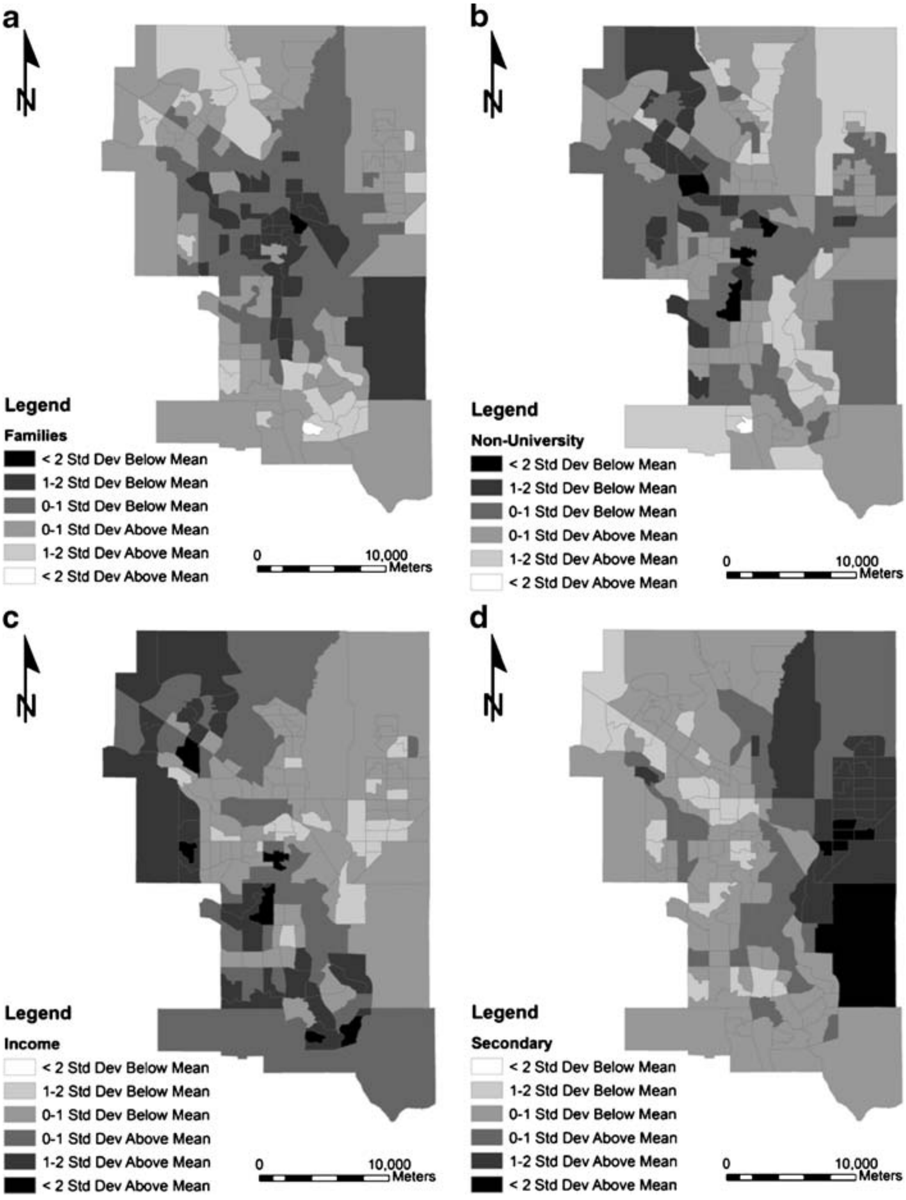

Fig. 4 Spatial regression model: explanatory variables

Discussion

The spatial regression model (Eq. 4) represents a potential analytical tool for the definition of social and health policies for the reduction of the prevalence of cardiovascular disease. Through its final set of explanatory variables, their sign and significance, the model casts a new light on the socio-economic pattern of the city.

By an in-depth analysis of the areas where the explanatory variables present their highest values (in the case of positive regression coefficients) or lowest values (in the case of negative regression coefficients), the model can aid in identifying localized pockets of high risk population, which may not be immediately apparent from the spatial pattern of the disease prevalence. To realize this potential of the model, it is imperative that the parameters linking each explanatory variable to the dependent variable are estimated reliably, and this condition was met by the transition from a standard to a spatial autoregressive specification.

In order to achieve the stated policy objective, an acceptable risk level should be defined, along with a risk threshold for each independent variable. The definition of a risk function is beyond the scope of this work; however, the results of this analysis can serve as an example of the model's potential. As discussed in Section "Regression Analysis", the standard model has little potential as a policy tool, dominated as it is by the variable "*age over 65*", which so closely mimics the spatial distribution and clustering of the dependent variable. Conversely, all the four variables in the spatial regression model provide meaningful contributions to the model, without any single one dominating the regression; in addition, the spatial distribution of all these variables is distinct from the spatial pattern of the dependent. Of the four variables, "*income*" and "*non-university*", both with positive sign, are indicators of relatively high economic if not social status; even though this is not a guarantee of wellbeing, areas dominated by high values of these variables present relatively low urgency for the definition of social and health policies. In contrast, areas dominated by low education levels and by scarcity of young families, i.e., high values of "*secondary*" and low values of "*families*", may point to situations of social and economic concern, where the need for social and health policies is more urgent. Interestingly, the two wealth indicators, "*income*" and "*non-university*" display similar spatial patterns, and their greatest (positive) values are markedly located in the west and northwest of the city (Fig. 4). On the contrary, a diffused high value of "*secondary*" is a prevalent characteristic of the east, whereas low values of "*families*" is more severe in the central areas, expanding to the east and south. The latter variable presents an almost radial pattern, likely reflecting the age of settlement of the various communities. The relative scarcity of families with children in the city center is not found in association with low education (unlike in the north-east), as these are areas mostly inhabited by single, young, and wealthy residents.

From the above considerations emerges a dichotomous spatial pattern, where wealth and high social status are consistently found in the west part of the city, whereas lower social and economic status characterizes the east. For demonstration purposes we propose the spatial overlay[14] of the two indicators of low social and economic status, which, identifying areas of simultaneous presence of the two variables, more narrowly define a set of potential risk pockets, shown in Fig. 5.

The pockets thus identified are mostly located in the eastern part of the city. Of little relevance are the large census tracts, mostly occupied by the airport, industrial installations, and farmland (Fig. 1a), while more important is the identification of the

[14] For each variable, the two highest and lowest classes, respectively, were selected and spatially overlaid, by means of a simple query. Alternative classifications were tested, but they all resulted in an almost identical spatial pattern.

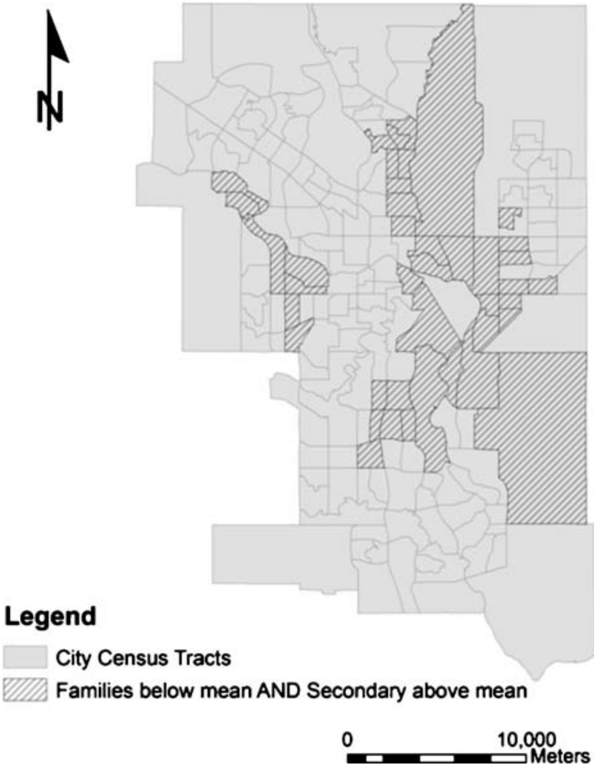

Fig. 5 Potential high risk areas identified by the spatial regression model

north-eastern residential fringes west and south of the airport, which have lower income levels and are located near the more noxious facilities and industries of the city. This corridor is becoming an area of growing concern to city officials: it has been confirmed to be home to greater than average numbers of elderly citizens and it is becoming an area of high psycho-social stress and crime (Calgary Police Service 2008).

Large portions of the northeast do not emerge as areas of concern, where the presence of low education is compensated by the abundance of young families. Interesting is also the northwest corridor, nearing a wealthy area, but characterized by situations of social isolation. This area consists of five census tracts, four of which have median income levels well below the city average[15] and have been reported to contain a higher percentage of population that are considered socially isolated according to the City's civic census.[16] Thus, although the northwest sector

[15] Lower income in this area is also confirmed through Fig. 4 above.

[16] According to the City Census, 9.1% of the population of Calgary is considered socially isolated. This indicator presents an average value of 11.08% (1.98 above the city) for the 5 communities identified, and rises to 12.075% (2.975 above the city), excluding the community of Wildwood. Likewise, the median income of Calgary is 57,879, in contrast to an average of 51,823 for the 5 communities, and 46,676, excluding Wildwood.

of Calgary is usually considered affluent by its residents, four of the five areas identified as high risk pockets in this area display variables that are more similar to the eastern part of Calgary. This could also be attributed to the absolute age of the census tracts. The census tracts in this corridor are not part of the peripheral suburban expansion that has become accustomed to Calgary's growth patterns; instead, this area consisted of existing towns that were annexed many years ago. This indicates more affordable housing for the lower income, less educated population. Furthermore, an older, physically smaller housing style in this area combined with contemporary ideals of necessary house size render these dwellings more appropriate for families without children.

Even though the identification of potential risk areas was performed mainly for demonstration purposes, it has identified areas that are not generally considered of greatest social concern, but at a closer inspection, they present alarming signals. Thus the analysis effectively demonstrates the potential use of the proposed model for the identification of explicit, localized targets for early (and thus more effective) health and social policies. In addition, the process that led to the final stage through the spatial regression model provides a rich analysis of the socioeconomic pattern of the city. The multiple cross-correlations among variables reflect the interplay of several factors that appear to contribute to lifestyle choices that may affect modifiable risk factors. One final remarkable aspect of the analysis is the strong and constant presence of education variables in the regression models. Even though each of these variables hints also to other variables, such as age and income, their significance is so prominent as to raise the question as to whether a direct, negative link may indeed exist between disease prevalence and education, and this in itself might constitute a valuable policy recommendation.

Conclusion

The association between cardiovascular disease and a pool of demographic and socioeconomic variables was analyzed, for the city of Calgary (Canada), over a 5-year interval around the 2001 census.

The analytical results suggest that the spatial dependence observed in the dependent variable is driven by the spatial dependence in the variable "*age over 65*", which is considered representative of retirement age, and is most closely correlated with the dependent. This finding may have important consequences not simply for the specification of a reliable statistical model, but, more importantly, for a deeper conceptual understanding of the roots of the spatial dependence observed in the disease prevalence.

A multivariate specification including demographic variables results in a model characterized by insignificant residual spatial autocorrelation, but dominated by the retirement age indicator (Driver et al. 2008). Conversely, a spatially autoregressive specification on a pool of solely socio-economic variables produces a model whose explanatory variables range from family status to income and education levels. This regression presents the greatest potential as an analytical tool to support policy decisions, because the disease prevalence is not associated simply with old age, but with a set of social and economic variables that can be targeted by effective social policies before the disease insurgence becomes inevitable.

The reliable identification of variables associated with the disease prevalence is followed by an analysis of the spatial distribution of each one of these variables; an additional analytical step recombines the significant variables based on their spatial patterns. Thus the analysis serves to identify localized areas of social and economic concern, characterized by a significant presence of the variables found in association with the disease prevalence.

Our future lines of investigation shall involve an analysis of the quantitative and conceptual implications of the model of spatial dependence, as a complex pattern of spatial dependence emerged from the comparison of different models. A separate line of research shall investigate the local variation of the multivariate relationship, in order to provide a more comprehensive analytical solution by the integration of local and global analytical methods. Complementary to the latter investigation, an analysis of the multivariate relationships emerging from this paper shall be conducted at different geographical scales, due to the imperfection of any spatial aggregation and the ever present modifiable areal unit problem. Also, further analyses shall specifically include air pollution and other relevant environmental data, to clarify their influence on the spatial pattern of disease prevalence and their interaction with social and economic factors. Finally, the policy potential of the model shall be enhanced by the construction of a spatial risk framework and the definition of risk thresholds for each independent variable.

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. Finally we would like to thank the editor of ASAP and the anonymous referees for their insightful comments and constructive criticism.

References

Aldstadt, J., & Getis, A. (2006). Using AMOEBA to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis*, *38*(4), 327–343.

Anselin, L. (1988). *Spatial econometrics: Methods and models*. New York: Kluwer.

Anselin, L. (1993). *SpaceStat tutorial*. Morgantown: Regional Research Institute, West Virginia University.

Augustin, T., Glass, T. A., James, B. D., & Schwartz, B. S. (2008). Neighborhood psychosocial hazards and cardiovascular disease: The Baltimore memory study. *American Journal of Public Health*, *98*(9), 1664–1670.

Bertazzon, S., & Olson, S. (2008). Alternative distance metrics for enhanced reliability of spatial regression analysis of health data. In O. Gervasi, B. Murgante, A. Laganà, D. Taniar, Y. Mun, & M. Gavrilova (Eds.), Proceedings of the International Conference on Computational Science and its Applications, Part I, volume 5072 of Lecture Notes in Computer Science, pp. 361–374.

Besag, J., & Green, P. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society B*, *55*, 25–37.

Calgary Police Service. (2008). Annual statistical report 2003–2007 (http://www.calgarypolice.ca/news/pdf/Annual_Statistical_Report_2007.pdf).

Canadian Cardiovascular Outcomes Research Team. (2009). (http://www.ccort.ca/CardiovascularAtlas/Atlasdescription/tabid/62/Default.aspx).

Canadian Heart Health Strategy and Action Plan. (2009). Building a heart healthy Canada (http://www.chhs-scsc.ca/web/wp-content/uploads/60408strategyeng.pdf).

Chaix, B., Rosvall, M., & Merlo, J. (2007). Neighborhood socioeconomic deprivation and residential instability: effects on incidence of ischemic heart disease and survival after myocardial infarction. *Epidemiology*, *18*(1), 104–111.

Cliff, D., & Ord, J. K. (1981). *Spatial processes. Models and applications*. London: Pion.

Cressie, N. A. C. (1993). *Statistics for spatial data*. New York: Wiley.

Diez Roux, A. V., Merkin, S., Arnett, D., Chambless, L., Massing, M., Nieto, F., et al. (2001). Neighborhood of residence and incidence of coronary heart disease. *New England Journal of Medicine*, *345*(2), 99–106.

Driver, J. A., Djoussé, L., Logroscino, G., Gaziano, J. M., & Kurth, T. (2008). Incidence of cardiovascular disease and cancer in advanced age: prospective cohort study. *British Medical Journal*, *337*(7683), 1400–1403.

Duncan, C., & Jones, K. (2000). Using multilevel models to model heterogeneity: potential and pitfalls. *Geographical Analysis*, *32*, 279–305.

Elliott, P., & Wartenberg, D. (2004). Spatial epidemiology: current approaches and future challenges. *Environmental Health Perspectives*, *12*(9), 998–1006.

Elliott, P., Wakefield, J. C., Best, N. G., & Briggs, D. J. (2000). *Spatial epidemiology: Methods and applications*. Oxford: Oxford University Press.

Evans, I. (1977). The selection of class intervals. *Transactions of the Institute of British Geographers, New Series*, *2*(1), 98–124.

Fortin, M. -J., & Dale, M. R. T. (2005). *Spatial analysis: A guide for ecologists*. Cambridge University Press.

Fotheringham, A. S., Brundson, C., & Charlton, M. (2002). *Geographically weighted regression: The analysis of spatially varying relationships*. Chichester: Wiley.

Ghali, W. A., & Knudtson, M. L. (2000). Overview of the Alberta provincial project for outcome assessment in coronary heart disease. *Canadian Journal of Cardiology*, *16*(10), 1225–1230.

Gerber, Y., Jacobsen, S., Frye, R., Weston, S., Killian, J., & Roger, V. (2006). Secular trends in deaths from cardiovascular diseases. A 25-year community study. *Circulation*, *113*(19), 2285–2292.

Getis, A. (2008). A history of the concept of spatial autocorrelation: a geographer's perspective. *Geographical Analysis*, *40*(3), 297–309.

Getis, A., & Aldstadt, J. (2004). Constructing the spatial weights matrix using a local statistic. *Geographical Analysis*, *36*, 90–104.

Guo, D. (2007). Visual analytics of spatial interaction patterns for pandemic decision support. *International Journal of Geographical Information Science*, *21*(8), 859–877.

Kaplan, G. A., & Keil, J. E. (1993). Socioeconomic factors and cardiovascular disease: a review of the literature. *Circulation*, *88*(4), 1973–1998.

Law, J., Haining, R., Maheswaran, R., & Pearson, T. (2006). Analyzing the relationship between smoking and coronary heart disease at the small area level: a Bayesian approach to spatial modeling. *Geographical Analysis*, *38*(2), 140–159.

Legendre, P., Dale, M. R. T., Fortin, M.-J., Gurevitch, J., Hohn, M., & Myers, D. (2002). The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography*, *25*, 601–615.

Li, H., Calder, C. A., & Cressie, N. (2007). Beyond Moran's *I*: testing for spatial dependence based on the spatial autoregressive model. *Geographical Analysis*, *39*, 357–375.

Lin, G., & Zhang, T. (2007). Loglinear residual tests of Moran's *I* autocorrelation and their applications to Kentucky breast cancer data. *Geographical Analysis*, *39*, 293–310.

Manuel, D., Leung, M., & Nguyen, K. (2003). Burden of cardiovascular disease in Canada. *Canadian Journal of Cardiology*, *19*(9), 997–1004.

Mennis, J. (2006). Mapping the results of geographically weighted regression. *The Cartographic Journal*, *43*(2), 171–179.

Rushton, G. (2000). GIS to improve public health. *Transactions in GIS*, *4*, 1–4. (reference to authors' own work: withheld for review purpose).

Statistics Canada. (2007). *More information on postal codes*. Retrieved April 5, 2009, from Statistics Canada 2006 Census Definitions: http://www12.statcan.ca/english/census06/reference/dictionary/geo035a.cfm. 06 11.

Tiefelsdorf, M. (2003). Misspecifications in interaction model distance-decay relationships: a spatial structure effect. *Journal of Geographical Systems*, *5*(1), 25–50.

Waller, L. A., & Gotway, C. A. (2003). *Applied spatial analysis of public health data*. New York: Wiley.