Localized spatial regression modelling of health data

S. Bertazzon, S. Olson, J. Norris, O. Elikan

Department of Geography, University of Calgary.
2500 University Dr. NW Calgary, AB, T2N 1N4 CANADA
Telephone: +1 (403) 220-7355
Fax: +1 (403) 282-6561
Email: bertazzs@ucalgary.ca

1. Introduction and context

Effective locational decision making is essential for properly addressing many health related concerns, such as detection and monitoring of environmental hazards, efficient response to epidemic outbreaks, and effective access to health care services. Presently, these decisions are supported by quantitative models, which are *potentially* powerful tools, but whose estimates are often affected by uncertainty, which reduces their reliability: unreliable models may lead to ineffective or even harmful management decisions (Elliott and Wartenberg, 2004). Uncertainty in the model parameters stems from two properties of geographical phenomena: spatial non-stationarity (i.e., inconstant variability over space), and spatial dependence (i.e., near things are more related than distant things). These two properties are mutually related, and most observed processes exhibit both, simultaneously (Cliff and Ord, 1981). Advanced spatial analytical methods exist to correct for the effects of each property; however, despite the recognized simultaneity of their occurrence, each advanced spatial method is designed to address only one property: spatial autoregressive methods (SAR) address spatial dependence (Anselin, 1988), but do not account for non-stationarity; geographically weighted regression (GWR) addresses non-stationarity (Fotheringham et al., 2002), but does not account for spatial dependence. A serious implication of this is that in trying to reduce the effects of one confounding property, the effects of the other property may be worsened, thereby potentially increasing the overall uncertainty (Tiefelsdorf, 2003).

In response to these limitations, a method is proposed to deal with both properties: a localized version of SAR. The objective of the research is to model geographical phenomena at small, yet meaningful spatial units, where reliable quantitative models can provide effective support tools for locational decisions.

This paper builds on recent work (Bertazzon et al., 2009), where SAR was used to identify socio-economic variables associated with cardiovascular disease prevalence in Calgary, a large Canadian city with a population over 1 million (2008 civic census); family status, income, and education are significantly associated with disease prevalence. Examining the spatial pattern of these variables, the analysis identifies pockets of low income and social isolation, which constitute disease risk areas, and emerge from a relatively homogenous background, otherwise dominated by medium-high income and traditional families.

The reliability of these results may be enhanced by appropriate local analysis; indeed, uncertainty may affect the global model for two reasons. First, tests indicate that not all variables can be considered stationary over the entire city; therefore, the parameters were estimated in violation of the non-stationarity hypothesis in some parts of the city. Second, the global model selection procedure leads to a single set of explanatory variables, which are not necessarily significant in every part of the city. The GWR method is appropriate to address the former, but not the latter issue.

In the following paragraphs, we discuss a method to enhance the model reliability in all parts of the city by simultaneously addressing non-stationarities and spatial dependence.

2. Methods

Given that global SAR methods are unsatisfactory for processes that exhibit spatial non-stationarity and, conversely, GWR methods are unsatisfactory in the presence of spatial dependence, an integration between local and global regressions is desirable. Our proposed solution is a localization of SAR: first, meaningful micro-region are defined, based on administrative and statistical criteria; subsequently, the local modelling principle is applied on each micro-region, via the implementation of a localized SAR model, thereby addressing jointly spatial dependencies and non-stationarities.

Statistically, the proposed method enhances model reliability more effectively than either SAR or GWR alone, because it jointly addresses both sources of uncertainty: spatial dependence via SAR methods, and non-stationarity via localized modelling. Additionally, the approach is conceptually well founded and has positive management implications: defining meaningful micro-region and estimating models at that scale yields more comprehensible analyses that are more effective decision support tools than global or local models.

Ultimately, the research discussed in this paper will include a statistical procedure for the identification of spatially stationary micro-areas. The statistical procedure will be integrated with qualitative knowledge to ensure the stationarity of meaningful management units. For the present discussion, a set of meaningful micro-regions is defined following criteria based on dominant road patterns in combination with barriers of urban area connectivity, such as water bodies, protected areas and major highways. Road patterns are chosen as a delineating criterion since similar patterns infer a predominance of certain demographics (Jacobs, 1993). Southworth (1997) established the road classification used to differentiate regions. Statistical constraints are added, such as minimum sample size, to preserve the robustness of the analytical results.

3. Results and Discussion

Employing the criteria defined in Section 2, six regions are defined for Calgary (Fig. 1).

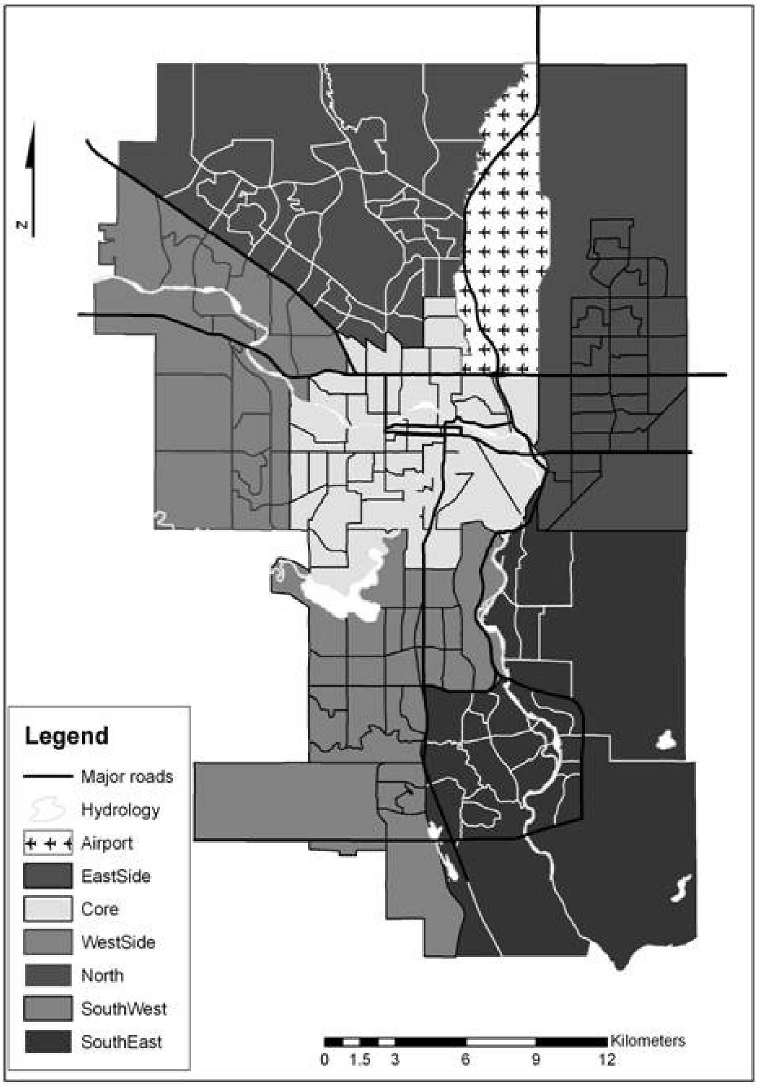

Fig. 1: Identification of six meaningful regions in Calgary

Spatial dependence (Getis, 2008) is estimated for each variable, in the whole city and in each region, by a spatial autocorrelation index (Table 1).

Table 1: Spatial autocorrelation for the regression variables

		Acute coronary syndrome	2 parents no chld. families	Widowed divorced separated	Post-sec. non univ. education	Grade 13 or lower education	Single detached units	Duplex units	Faminly Median Income
Global	I	0.06	0.76	0.79	0.42	0.32	0.59	0.52	0.49
	z	0.82	9.05	9.42	5.05	3.89	7.07	6.19	5.88
Core	I	0.42	0.54	0.26	0.12	0.50	0.49	0.29	0.04
	z	2.44	3.06	1.55	0.77	2.86	2.78	1.69	0.36
East Side	I	0.35	0.60	0.63	-0.17	0.54	0.35	0.38	-0.03
	z	2.04	3.40	3.58	-0.75	3.09	2.06	2.23	-0.01
North	I	0.66	0.74	0.56	0.53	0.90	0.28	0.46	0.40
	z	3.61	4.02	3.10	2.91	4.88	1.63	2.56	2.22
South East	I	0.01	0.32	0.14	-0.10	0.29	0.37	0.47	0.08
	z	0.20	1.52	0.74	-0.24	1.40	1.71	2.13	0.50
South West	I	0.46	0.63	0.44	-0.08	0.29	0.03	0.06	0.31
	z	2.22	2.97	2.14	-0.19	1.47	0.32	0.45	1.53
West Side	I	0.36	0.40	0.18	0.28	0.15	0.18	0.40	0.34
	z	1.91	2.10	1.06	1.53	0.90	1.05	2.09	1.82

The dependent variable exhibits insignificant spatial autocorrelation values, while all the independent variables exhibit significant values at the global level and lower values at the micro-region level. These results support the need for SAR in some parts of the city. As a measure of spatial stationarity, Getis-Ord Gi* tests (Fotheringham et al., 2002) are conducted on all variables at the global level and for each micro-region (Table 2).

Table 2: Global and local Getis–Ord Gi* stationarity tests

		Acute coronary syndrome	2 parents no chld. families	Widowed divorced separated	Post-sec. non univ. education	Grade 13 or lower education	Single detached units	Duplex units	Faminly Median Income
Global	Obs.	0.08	0.10	0.09	0.08	0.09	0.08	0.17	0.08
	z	-0.84	8.98	4.57	-3.60	2.69	-3.01	11.31	-3.65
Core	Obs.	0.11	0.13	0.12	0.12	0.12	0.11	0.11	0.13
	z	-1.60	2.19	-1.91	-2.40	-2.32	-1.26	-0.62	1.15
East Side	Obs.	0.21	0.20	0.21	0.18	0.20	0.19	0.25	0.20
	z	1.74	1.26	1.78	-2.64	0.85	-1.11	2.67	-1.83
North	Obs.	0.28	0.24	0.27	0.24	0.26	0.23	0.41	0.22
	z	3.54	1.08	3.26	1.14	3.34	-1.10	3.75	-3.65
South East	Obs.	0.19	0.32	0.28	0.35	0.32	0.37	0.42	0.35
	z	-1.59	-1.40	-1.97	0.37	-1.84	1.02	0.34	0.30
South West	Obs.	0.41	0.39	0.41	0.36	0.37	0.32	0.45	0.34
	z	2.85	3.49	3.35	0.60	1.85	-1.92	1.59	-1.83
West Side	Obs.	0.23	0.21	0.23	0.20	0.22	0.20	0.45	0.20
	z	2.06	1.74	2.53	-0.18	1.75	-0.35	3.50	-0.98

The dependent variable displays insignificant scores globally and in several regions. Conversely, all independent variables display significant scores globally, and insignificant scores in many micro-regions. Overall the test indicates, for most variables, global spatial non-stationarity, and stationarity at the micro-region level.

Cross-correlation analysis and model selection procedures (Bertazzon et al., 2009), are applied, individually, to each micro-region, producing six different models (Table 3).

Table 3:Global and local regression models

		2 par. no chld. Fam.	Widow. Divorc. Separ.	Post-sec. non-uni. ed.	Grade 13 or lower educ.	Single detach. units	Duplex units	Fam. Median Income	psuedo R^2	Rho	Res Std Error	Res Moran
Global	β	5.90	11.40	-5.10	-3.10	3.60	-2.40	-5.50	0.52	0.34	0.08	0.00
	t	6.13	5.31	-3.79	-3.01	8.94	-2.24	-3.23				
Core	β		14.50			3.20	-3.30	-5.80	0.57	0.30	0.08	0.02
	t		4.35			4.72	-2.43	-2.28				
East Side	β	12.10		-9.60	-5.80				0.58	0.09	0.06	0.02
	t	5.79		-2.03	-1.99							
North	β	6.70		-15.10		4.70		-8.20	0.78	-0.01	0.07	0.00
	t	3.24		-4.48		4.65		-2.53				
South East	β	11.10	19.60	-15.20	-4.70			-11.20	0.82	-0.23	0.07	-0.09
	t	8.07	2.45	-4.10	-3.51			-3.54				
South West	β	16.50		-11.00		3.50			0.71	-0.08	0.08	-0.04
	t	7.77		-3.73		0.65						
West Side	β	6.80							0.28	-0.08	0.10	-0.03
	t	3.05										

Each localized regression is a “custom-made” model, improved over both SAR and GWR models. Unlike GWR, the proposed approach leads to six different sets of variables, all significant in each micro-region, whereas GWR models are all based on an identical set of variables, determined by global model selection procedures. Consequently, not all these variables are necessarily significant in each region. As an example, Fig. 2 shows areas of significance ($\alpha = 0.05$) for one of the beta parameters.

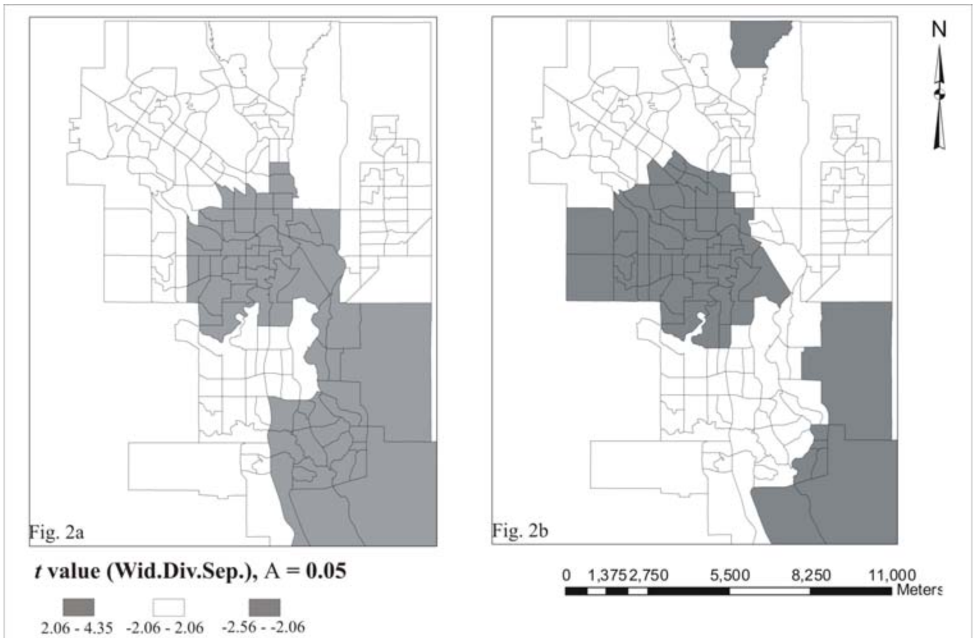

Fig. 2: Localized (a) vs. GWR (b) regressions: parameter significance

As shown in Table 3, this parameter is significant only in the “Core” and “SouthEast” regions; therefore, it is retained only in those two localized regressions (Fig. 2a): this ensures efficiency, robustness and meaningfulness of each model. The significance of the same beta parameter in GWR is mapped in Fig. 2b, which shows that the spatial distributions are almost identical for the two methods. What really does differentiate the two methods is that in GWR the parameters are retained in each local regression throughout the entire city, even though they are significant only in some areas, implying a corresponding loss of efficiency, robustness, and meaningfulness.

In addition, each localized model is spatially autoregressive, which minimizes the variance induced by spatial dependence. The reliability of model estimates has been therefore maximized, by addressing both sources of uncertainty in a localized approach integrating SAR and GWR methods at an intermediate spatial scale, most meaningful for decision support.

4. Conclusion

The research presented in this paper constitutes a successful experiment of localization of global SAR methods. The method proceeds through the identification of meaningful micro-regions, the assessment of the properties of spatial dependence and non-stationarity at the global and local levels, and the specification and selection of localized spatially autoregressive models for each micro-region. The method constitutes a valuable alternative to traditional GWR and SAR methods, possesses added flexibility in model selection, and provides effective decision support tools at meaningful scale.

5. Acknowledgements

We acknowledge the GEOIDE network for funding and the APPROACH initiative for data and support. Thanks to colleagues and students for their suggestions and help.

References

L. Anselin (1988). *Spatial Econometrics: Methods and Models*. New York: Kluwer Academic Publisher.

S. Bertazzon, S. Olson, and M. Knudtson (2009). A spatial analysis of the demographic and socio-economic variables associated with cardiovascular disease in Calgary (Canada). *Applied Spatial Analysis and Policy*. Forthcoming.

D. Cliff and J. Ord (1981). *Spatial Processes. Models and Applications*. London: Pion.

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

A.S. Fotheringham, C. Brundson, and M. Charlton (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester: Wiley.

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

A. Jacobs (1993). *Great Streets*. Cambridge, MA: MIT Press.

M. Southworth (1997). *Streets and the Shaping of Towns and Cities.* New York: McGraw-Hill.

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