Executive Summary

Multivariate Spatial Regression in Epidemiology: Feasibility Study of Alternative Computational Approaches for the Estimation of Spatial Dependence.

> "PEOPLE DIE each year because no one *BOTHERS to properly analyze DISEASE and DEATH* data for unusual localized concentrations"
>
> Stan Openshaw

Problem Statement:

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)" [Ahlbom & Norell 1984]. All these factors are in turn related to a complex variable usually referred to as *lifestyle*, which can be represented by demographic indicators (e.g. age, sex), socio-economic indicators (i.e. income, job type) and environmental indicators (e.g. recreation sport facilities, pollution).

The APPROACH Project is an ongoing data collection initiative, begun in 1995, containing information on all patients undergoing cardiac catheterization in Alberta [Ghali & Knudtson, 2000]. Preliminary analyses have shown evidence of a clear spatial pattern of the variable within the city of Calgary, indicating that the disease incidence should be analyzed as a spatial process, using appropriate analytical tools.

The scope of the project is to analyze the spatial variation of heart disease [*where* cardiac patients are located] and investigate the relationship between disease incidence and lifestyle indicators [*why* clusters of heart disease incidence occurs at those locations].

To address these questions a multivariate spatial regression model will be specified. The goal of the model is to analyze current patterns to simulate and predict future trends of disease incidence, based on the spatial variation of a set of lifestyle indicators. The model will represent an effective tool for policy, planning, service providing and facility location.

The novelty of the model is its spatial thrust. To guarantee its efficiency we propose an innovative approach to the measurement of distance based on the use of alternative metrics. This is an original approach that expands the scope of the project to fundamental questions in GIScience. The use of alternative metrics can provide a uniform criterion to estimate spatial autocorrelation, aiding in the definition of spatial contiguity and multivariate spatial correlation. This is also a methodological pilot study, that can be applied to the analysis other conditions.

Objectives:

The objective of the project is the development of a multivariate spatial regression model based on L_p metrics as an innovative analytical tool for prediction and simulation in spatial epidemiology. The main features of the model can be summarized as follows:

- The disease incidence is analyzed using demographic, socio-economic, and environmental factors: this provides a broader base for the disease prevention and mitigation, and enhances the predictive capacity of the model.
- Based on its ability to perform predictions and simulations, the model can be used to address "*what if …*" questions. This is a useful feature in planning, as it can address, for example, the question: "what will happen if a new clinic is opened in a neighborhood?"
- The model can be used to evaluate the response of the disease incidence to a change in each lifestyle factor (elasticity) independently or jointly, allowing for accurate and specific simulations.

- Due to the characteristics of the database, a dynamic model can be specified, for the detection of spatio-temporal trends. The model can serve for cross-temporal comparisons, and for advanced dynamic modeling: in this case, the analytical objective is to study the evolution of the functional parameters, their statistical significance, and their elasticity.
- The use of alternative distance metrics for an accurate measurement of the spatial autocorrelation in the data: this measurement ensures the *efficiency* of the model.
- The proposed metrics for the measurement of distance can be applied for route design and the provision of routine and emergency services, and optimal facility location using location-allocation models.
- The implementation of uni- and multi-variate (auto)correlation indices and the definition of a geometry based contiguity among different spatial objects.
- The investigation of spatial patterns will be performed using advanced computational methods, closely related to the methodology used to develop the spatial regression model.
- The implementation of adequate visualization tools for the representation of spatial data and the communication of analytical results. Visualization of spatial patterns and relationships among variables using topology-based models such as the TIN model and the generalized Delaunay triangulation, for the representation of diverse attributes (medical, social and environmental) of a spatial data structure.

Methodology:

The scope of regression analysis is to relate the variable of interest (dependent) to a set of explanatory variables by establishing a functional form of such relationship and estimating its functional parameters. For this reason, regression analysis is an ideal tool in epidemiology, where disease incidence typically results from the combination of several interacting factors. Even though regression analysis is an appealing tool, its simple application to spatial data would produce inefficient parameter estimates, due to the property of spatial dependence, that typically affects spatial data.

Spatial Regression Analysis.

The classical solution to spatial dependence requires the estimation of a correlation model, which is, in the case of spatial data, a spatial autocorrelation model. Spatial dependence can be informally defined using Tobler's (1979) so-called first of law of geography: "*everything is related to everything else, but near things are more related than distant things*". This double similarity (relatedness and nearness) implies that a SpAC model consists of two components: the attribute covariance, and the spatial component, usually represented by a spatial weight, which accounts for the effect of the spatial separation among units, an effect that tends to decrease as distance increases.

There are several commonly used methods, but not a single standard criterion to determine the value and spatial extension of the spatial weights. Usually the spatial weights are represented by a contiguity matrix, which is, in its simplest specification, a binary structure, defining which units are dependent, and which are not. There are many methods for determining contiguity[1]: some are heavily dependent on the topology of the spatial units; the most general method (i.e. threshold distance) involves an element of subjectivity [Bertazzon, 2002]. To the best of our knowledge there are, to date, only two commercial packages that can perform spatial regression analysis: Anselin's (1995) SpaceStat, and the Spatial Statistics extension of Splus [MathSoft, 2000], based on Cressie's (1993) work. Both packages use several methods, and both of them, for the threshold-distance method, require the specification of the maximum distance parameter, which represents the typical subjective choice.

Standard criterion for distance measurement in spatial regression models

We propose an innovative approach, which resolves this subjectivity by providing a uniform criterion, consisting of the use of L_p norms, to evaluate spatial autocorrelation. The method is essential to our project, because it allows for a uniform treatment of diverse data (medical, socio-economic, and environmental). Overcoming the present ambiguity, the proposed approach will enhance the reliability robustness of spatial regression estimates, particularly in the present quest for standardization and interoperability in GIS. The inclusion of the proposed approach in commercial software can be envisaged in the near future.

The proposed method is developed around the threshold-distance method, and makes use of L_p norms to provide a greater flexibility in the measurement of distance and, consequently, of spatial autcorrelation. Additionally, the distance-based method is intrinsically flexible[2], and is applicable in most spatial autocorrelation indices (i.e. Moran's and Geary's indices).

L_p norms will be used to develop a spatial autocorrelation function γ combining attribute values (i.e. the value of the observed variable) at different distance intervals. The γ function will provide a *range* value, or the spatial extension of the spatial autocorrelation: this is the threshold distance that will be used in defining the contiguity matrix. Along with the evaluation of alternative L_p metrics, we also propose different variations of the γ function, where the distance, as measured by the proposed metrics, is weighted in different ways by the attributes covariances. We regard the latter as a profoundly innovative method, in that, traditionally, the attribute covariance is considered as a fixed parameter, and the spatial structure only considered as a weight. The implementation of these two approaches will lead to the specification of a family, Γ, of spatial autocorrelation functions, among which the most appropriate will be selected for a specific spatial process of interest. To this purpose, we propose to explore several L_p norms [Okabe, 1999; Gavrilova, 2002] and suggest an algorithm to determine the *best* norm for the spatial process of interest. The optimizing algorithm will compare the model variance, selecting the norm that can provide the lowest variance.

Multidimensional spatial autocorrelation for multivariate spatial regression analysis

The methodology for the estimating spatial autocorrelation will be extended to bi- and multi-variate specifications, to obtain a model of spatial correlation among all the variables involved.

This is a crucial part of a multivariate regression model, particularly when the variables in the model have been sampled at different spatial units [Bertazzon, 2002]. This is typically the case in epidemiological studies, and particularly in the APPROACH database, where the disease data are released at the block level, and the lifestyle indicators are released at the census enumeration level. The result will be a multidimensional spatial (auto)correlation model, that can be intuitively visualized as a hyper-matrix that combines several attributes at varying distances.

Further application the L_p norm-based spatial autocorrelation function

The proposed γ function combines attribute and spatial weights. For this reason, we can envisage a direct use of this function in the estimation of the spatial regression parameters. This will overcome the need for a contiguity matrix, simplifying the entire process.

The calculation of an optimized γ function based on L_p norms can become a sub-routine of the spatial regression model. Beyond the conceptual merits of the method, this development could prove very appealing to the software industry, and promote the applications of spatial regression analysis. Such sub-routines can be used in other analytical procedures, first of all traditional (and dated) spatial autocorrelation indices, such as Moran's and Geary's indices.

Pattern analysis

Point pattern analysis is concerned with the location of events, and with answering questions about the distribution of those locations, specifically whether they are clustered, randomly or regularly distributed [Cressie,93]. The simplest way of exploring point pattern data is by examining a two-dimensional frequency distribution of counts within equal-area units imposed on the study area. Nearest-neighbor distances are also used to analyze intensity. It is proposed to investigate the spatial structure using partition-based point pattern analysis techniques, utilizing the topology-based computational geometry data structures, such as k-d trees, regular spatial partitioning and Delaunay tessellation [Okabe, 99]. Also, the new method for bitmap based pattern analysis, based on the worst-case time optimal Euclidean distance transform algorithm, will be utilized [Gavrilova and Alsuwaiyel, 2001]. The method performs the analysis of data (in the matrix form), with the purpose of discovering clusters.

Geometric visualization.

Geometric visualization provides an efficient tool for producing qualitative information through computational and graphical methods, allowing fast data interpretation [Padin, 2002]. In the framework of the proposed project, geometric visualization will assist in qualitative analysis of autocorrelation and multivariate spatial regression models. Visualization will be based on spatial data representation, which will include topological proximity, attribute values analysis, and multi-dimensional (i.e. 2- and 3- dimensional) graphical representation of sample points and their autocorrelation values. The underlying TIN and Delaunay triangulation models will be used. It is expected that visualization of the results will provide new insights on how distance, attribute correlation and spatial weights relate; and be useful in evaluating alternative metrics in medical applications.

Milestones and Deliverables:

Phase I (February 2003 – October 2003): Pattern analysis and visualization methods;

Phase II (October 2003 - March 2005): Implementation and testing of methodology

Phase II will include meeting with network partners, participation in GEOIDE workshops, organization of workshops with local mini-network, and initiation of the Spatial Epidemiology and Geomatic Research Group. Visits to project supporters (NCGIA, Hong Kong Polytechnic), student trips to GEOIDE Annual meetings and result dissemination through variety of venues is planned.

The main contributions and benefits of the proposed project can be summarized as follows:

- An efficient multivariate spatial regression model for the analysis and prediction of the spatial pattern of disease incidence as a function of localized socio-economic and environmental variables.
- Implementation of a dynamic spatial model, based on data collected systematically over several years, which will represent an effective tool for detecting spatial and temporal trends, and for monitoring zones of particular concern.
- The use of alternative metrics for the evaluation of distance potentially providing efficient computational methods suitable for route design and the provision of routine and emergency services, and optimal facility location using location-allocation models.
- Implementation of uni- and multi-variate (auto)correlation indices and definition of contiguity among different spatial objects. These will provide innovative and flexible solutions to the analysis of spatial autocorrelation in GIS.
- Implementation of methods for the representation and visualization of spatial data, including the use of attribute features as a weight for the locational feature in spatial autocorrelation models.
- Visualization of spatial patterns and relationships among variables using topological models such as the TIN model and Delaunay triangulation, for the representation of social and environmental attributes in a spatial data structure.
- Development of methods that will aid in the interpretation of disease incidence as a spatial process, also providing efficient tools for communication to the public and interest groups.
- The developed methodology will serve as a catalyst for further integration of computational science and geography disciplines, with the goal of finding an innovative solution to other pressing socio-economical and medical problems.

References

Ahlbom A, Norell S (1984) *Introduction to Modern Epidemiology*. Epidemiology Resources Incorporated

Anselin L (1995) *New Directions in Spatial Econometrics*, New York: Springer-Verlag.

Anselin L. (1995) *SpaceStat Tutorial*. University of Illinois, Urbana - Champaign.

Bertazzon S (2002) *Metaspace: From a Model of Spatial Contiguity to the Conceptualization of Space in Geo-Analyses*. Joint International Symposium on Geospatial Theory, Processes and Applications 2002, Ottawa, July 8-12.

Bertazzon S (2002) *A Definition of Contiguity for Spatial Regression Analysis in GISc: Conceptual and Computational Aspects of Spatial Dependence*. Accepted for publication in Rivista Geografica Italiana, in press.

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

Fotheringham A S, Brundson C, Charlton M, (2000), *Quantitative geography. Perspectives on Spatial Data Analysis*. London: Sage.

Gavrilova, M. (2002) On a Nearest-Neighbor Problem under Minkowski and Power Metrics for Large Data Sets, the Journal of Supercomputing, Special Issue on Computational Issues in Fluid Dynamics, Optimization and Simulation, Kluwer, 22 (1): 87-98.

Gavrilova, M., Alsuwaiyel, M. (2001) Two Algorithms for Computing the Euclidean Distance Transform, International Journal of Image & Graphics, World Scientific, 1(4): 635-646.

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, (October 2000).

Griffith D A, Layne L J (1999) *A Casebook for Spatial Statistical Data Analysis*, Oxford: University Press.

Mausner J S, Kramer S (1985) *Mausner and Bahn Epidemiology An Introductory Text*. Second Edition. W. B. Saunders Company.

MathSoft Inc (2000) *S+SpatialStats. User's Manual for Windows and Unix*. Mathsoft Inc., Seattle, Washington.

Okabe, A., Boots, B. and Sugihara, K (1999): Spatial Tessellations --- Concepts and Applications of Voronoi Diagrams, 2nd ed, John Wiley and Sons, Chichester.

Padin, M. Deltrieux, P. Soto H. (2002) Experimental study of population dynamics ICCS'02 2: 125-131.

Tobler W R (1979) "Cellular Geography" in: Gale S., Olsson G. (eds.), *Philosophy in Geography*, Dordecht: Reidel, 379-386.

[1] A common method is the definition of k orders of spatial neighbors; an alternative method is a threshold distance; a third method is based on shared borders (for areal units only).

[2] Unlike the shared-border method, it s applicable to all types of spatial units. The nearest-neighbor method also possesses a relative flexibility and contains an element of subjectivity, but we believe that its conceptual issues are encompassed within our proposed solution to the threshold-distance method.