CO-280

GEOSTATISTICAL MAPPING FOR ENVIRONMENTAL STUDIES

GOVOROV M.(1), GIENKO G.(2)

(1) Vancouver Island University, NANAIMO, CANADA ; (2) University of Alaska Anchorage, ANCHORAGE, UNITED STATES

ABSTRACT

Modern GIS packages are well equipped with statistical and geostatistical tools which allow researchers to explore variety of environmental data in spatial context. This paper illustrates the use of several geostatistical methods to analyze associations and trends of spatially distributed data using different techniques, available in modern GIS packages. The paper outlines spatial analyses workflow in mapping various environmental characteristics and uses concentration of fine particulates (PM2.5) in Canadian cities as an example.

KEY WORDS

geostatistics, data exploration, geovisualization

INTRODUCTION

A common research task is the investigation of the spatial structures of natural or social phenomena using point observations and quantitative analysis. Modern GIS packages are well equipped with statistical tools, which allow researchers to explore their data in spatial context, employing variety of geostatistical tools and approaches. This paper illustrates the use of several methods for mapping environmental phenomena using various geostatistical methods and techniques for analysis of spatially distributed data. Exploratory data analysis (or testing of data for normality, linearity and spatial patterns) is discussed in the first part of the paper, the second part describes linear regression (to explore correlation and bivariate and multivariate regression, including geographically weighted regression), and the last part of the paper is dedicated to analysis of data trends using different geostatistical techniques including linear kriging and co-kriging.

While working with geospatial data, spread over large geographical areas, one should pay attention (and exercise certain care) on parameters of coordinate system (including datum and projection) used for data analyses. Geostatistical techniques heavily rely on distances and directions between observation points. While GIS packages have different ways to calculate and represent distances, most geostatistical tools assume that the Euclidean distance is used as a measure of distance. This distance is calculated based on locations of two points, and will depend on properties of a particular coordinate system (and projection) used. In kriging interpolation, size and shape of lag bins, which are used to group the pairs of locations, based on their distance and directions from one another for estimation of empirical semivariogram, will change depending on used coordinated system. In this paper we also looked at the sensitivity of spatial modeling to different coordinate systems – we used two types of projections with different types of distortions - equal area (Albers Equal Area Conic projection) and equal-distance (Lambert Conformal Conic projection), for sensitivity analysis.

METHODOLOGY AND DATA

A typical spatial analysis of data, having spatial component, may be based on the following workflow. Exploratory spatial data analysis (ESDA) is used for initial data analysis, such as check for statistical distribution, linearity and presence (or absence) of a pattern (both in spatial and non-spatial domains). Based on results of ESDA and initial hypotheses, further exploration can be developed by testing data for auto-correlation, correlation, and building some regression model. Trend analysis, an essential part of data exploration, can supplement kriging and co-kriging techniques, designed for analysis of spatially distributed data. There is excessive literature describing various theoretical aspects of spatial statists and geostatistical analyses, some references are provided in the Reference section.

In this paper, we used environmental geospatial data from the national database service, maintained by the Environment Canada (www.ec.gc.ca). The Environment Canada provides specific environmental information across Canada from its website assessable by general public. In this study statistical data reflecting air quality in Canada (best available data from 1997-2008) were used. Original dataset from the Environment Canada web site was modified to suite requirements of this study. This dataset includes geospatial data for air quality, greenhouse gas emissions and water quality indicators for different spatial units.

The primary dataset, which is used in this study, includes points with measurements of fine particulates (PM2.5) by Canadian cities. PM2.5 is key measure of air quality such as smog. The PM2.5 indicator is based on the 24-hr daily average concentrations recorded at monitoring stations across Canada during the warm season (April 1 to September 30). Fine particulates are minute solid particles or tiny liquid droplets in the air. When inhaled deeply into the lungs, even small amounts can cause serious health problems (www.ec.gc.ca). According to the data description, 71% of fine particulates (PM2.5) came from two sources: industry (35%) and home firewood burning (36%). To explore the factors behind observed spatial pattern of fine particulates across Canada cities, additional social-economical characteristics (or explanatory variables) at monitoring stations were prepared. Thus, the attributes of fine particulates dataset also include population and labor indicators. These social-economical data were derived from the 2006 Profile of Census Subdivisions (CSD) and Census Tract (CT) of Statistics Canadian (www.statcan.gc.ca). This set of socio-economical variables is used explore distribution of PM2.5 concentration across Canadian territories (see Fig. 1).

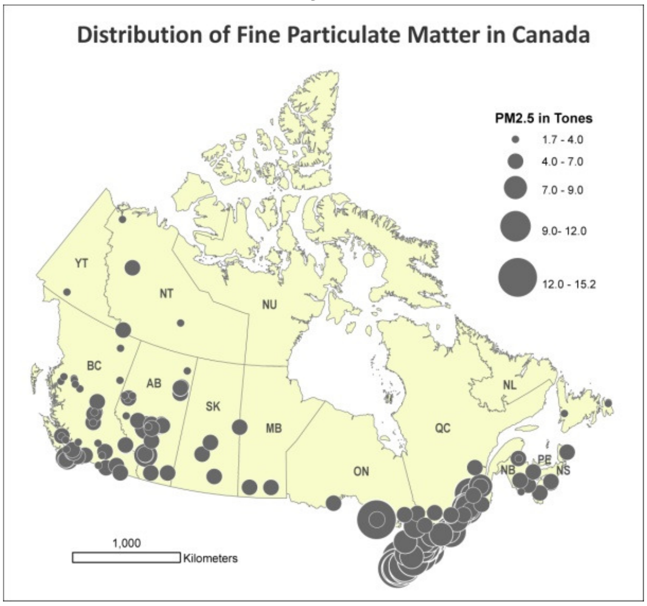

Fig.1. Distribution of Fine Particulate Matter (PM2.5) emissions in tones in Canadian cities (data source: www.ec.gc.ca)

Table 1 provides basic statistics of the source data.

Table 1. Basic statistics of the source data

	$PM_{2.5}$		$PM_{2.5}$		$PM_{2.5}$
Count	179	Range	13.5600	Variance	6.3246
Minimum	1.6840	Mean	6.2298	Coefficient of Variation	0.4036
Median	5.6226	Standard Deviation	2.5148	Coefficient of Skewness	0.7601
Maximum	15.2400			Coefficient of Kurtosis	3.2686

EXPLORATORY DATA ANALYSIS

Different statistical analyses techniques are based on certain hypotheses and assumptions, therefore, in order to get statistically correct results the source data should comply with particular requirements. The following data requirements were taken into consideration for regression and kriging analysis in our study.

Data requirements for regression. Spatial regression has to consider that explanatory variables in the model have a consistent relationship to the dependent variable both in geographic space and in data space. Therefore, the following requirements and issues of linear multiple regression have to be addressed:

• Variables should be measured on the interval or ration scale.

• Variables have a linear association.

• It is important not to miss significant explanatory variables (misspecification has to be avoided).

• Explanatory variables should not to be redundant (multicollinearity has to be avoided).

• For every value of dependent variable, the distribution of its residuals should be normal, and the mean of the residuals should equal zero.

• Residuals have a constant variance across all the values of the statistical variables (nonstationarity or spatial heterogeneity has to be considered).

• The value of each residual is independent of all other residual values (spatial autocorrelation has to be considered).

Data requirements for linear kriging. The kriging method predicts the best linear unbiased estimates of a surface at specified locations, based on the assumptions that the surface is stationary and the correct form of the semivariogram has been chosen. Therefore, the following requirements and issues of linear kriging have to be addressed:

• Variables are measured on the interval or ratio scale.

• Kriging methods are optimal when data:

• Come from normal distributions. Ordinary kriging accommodates non-normally distributed data as long as spatial autocorrelation structure is not masked by extreme-valued outliers.

• The underlying assumptions of second-order stationarity are met, i.e., at a minimum the mean and variance of the sample data remain invariant in space.

• The homogeneity, i.e., the data represents one single homogenous domain.

To explore the source data set to comply with the above requirements, we tested the data for normality, linearity and spatial patterns.

Test for Normality

For some methods of kriging, data should conform to the normal distribution. If the data does not exhibit a normal distribution (which can be checked graphically by plotting histogram or QQPlot), it may be necessary to transform the data to make it conform to a normal distribution before using certain interpolation techniques. According to analysis of the histogram and Normal QQPlot, as illustrated in Figure 2, distribution of PM2.5 values does not satisfy criteria of normality. However, applied normalization in form of logarithmic transformation brings the dataset closer to the normal.

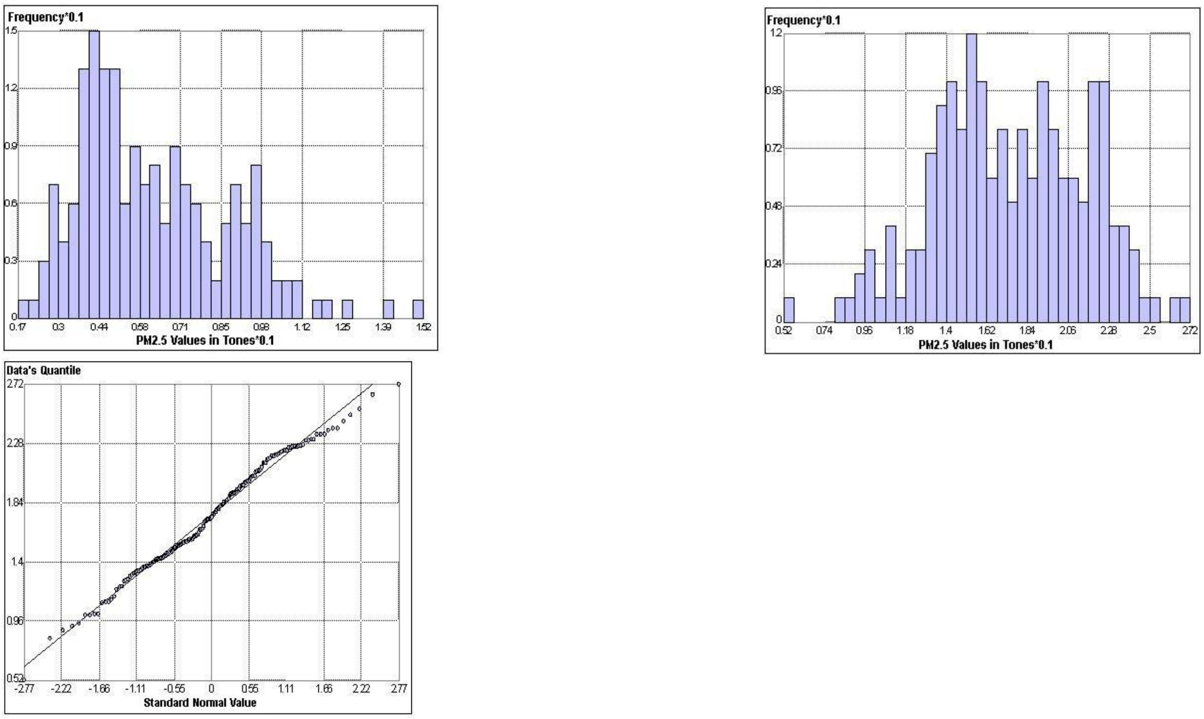

Fig.2. Histograms of the source and normalized PM2.5 values in tones and Normal QQPlot (after applying logarithmic transformation).

Test for Linearity

Often the first step in both correlation and regression analyses is to plot dependent and independent variables on a scatterplot graph. In this paper, we used linear regression methods. If the relationship between any of the explanatory variables and the dependent variable is nonlinear, the resultant regression model may perform poorly. In our case, the relationship between the most explanatory variables and the dependent variable is nonlinear, as illustrated in Figure 3.

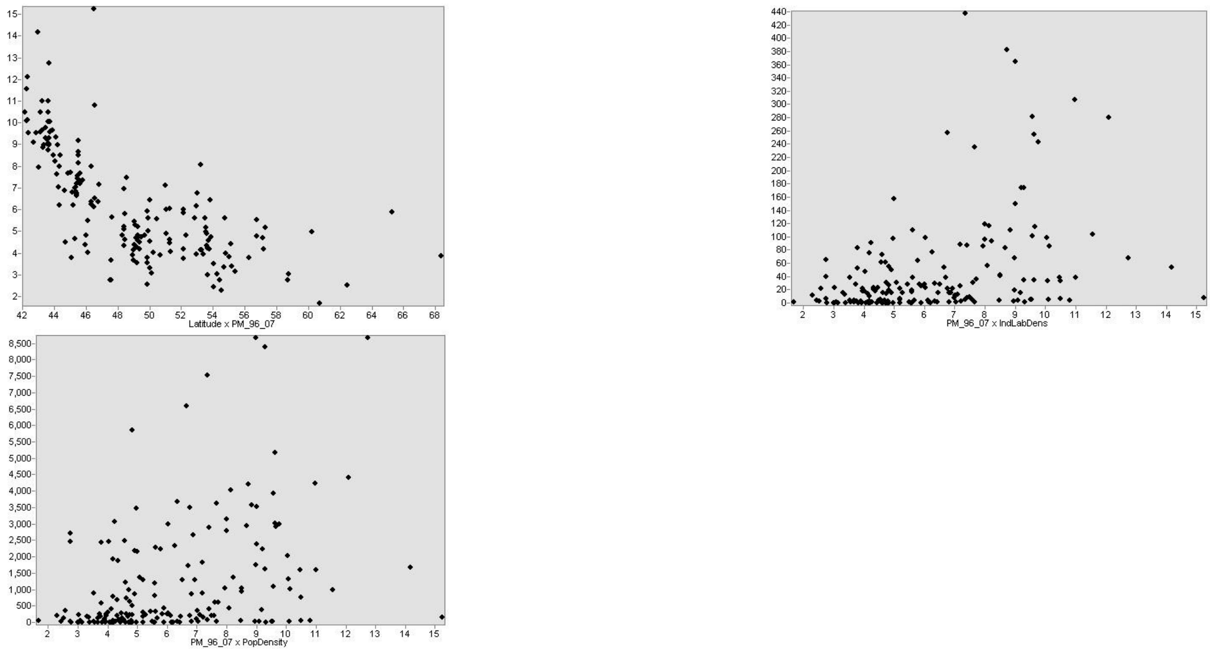

Fig.3. Scatter plots of PM2.5 values with Latitude, Density of all occupations of industrial labor and Population density 2006, respectively.

Another issue of multiple regression analysis is multicollinearity. This leads to an overcounting type of bias and an unstable/unreliable regression model. Multicollinearity occurs if one or a combination of explanatory variables used in regression is redundant. This usually happen when independent variables are correlated, and therefore they are not independent from each other. Strongly correlated variables should not be used together in one regression model. According to analysis of scatter plots, some independent variables in our dataset exhibits strong positive relationship between each other.

Pattern Analysis

We used spatial autocorrelation methods to identify patterns in PM2.5 spatial measurements. Most GIS packages utilize autocorrelation methods such as Moran’s I, G-Statistic, Cluster Anselin Local Moran’s I (Anselin, 1995) or/and Hot Spot Local G-Statistic (Getis, 1992) analysis. All these methods consider both locations of a measurement point and variation of attribute’s values at the locations. Spatial autocorrelation on the modeled variable indicates if the variable is spatially random, clustered, or dispersed. Results of spatial autocorrelation are useful in analysis of spatial regression.

According to Moran’s I and Getis-Ord analysis, the pattern of fine particulates PM2.5 can be described as highly clustered with statistical significance. Moran's I index is 0.9873 (p-value = 0.0) and Observed General G = 0.000005 (p-value = 0.0) for data in equal-distance projection. For data in equal-area projection results are as follows: Moran's I = 0.9957 (p-value = 0.0) and Observed General G = 0.000005 (p-value = 0.0). As seen from these numbers, Moran's I (global spatial autocorrelation) reveals some sensitivity to properties of used projections, while local pattern analysis is not so sensitive.

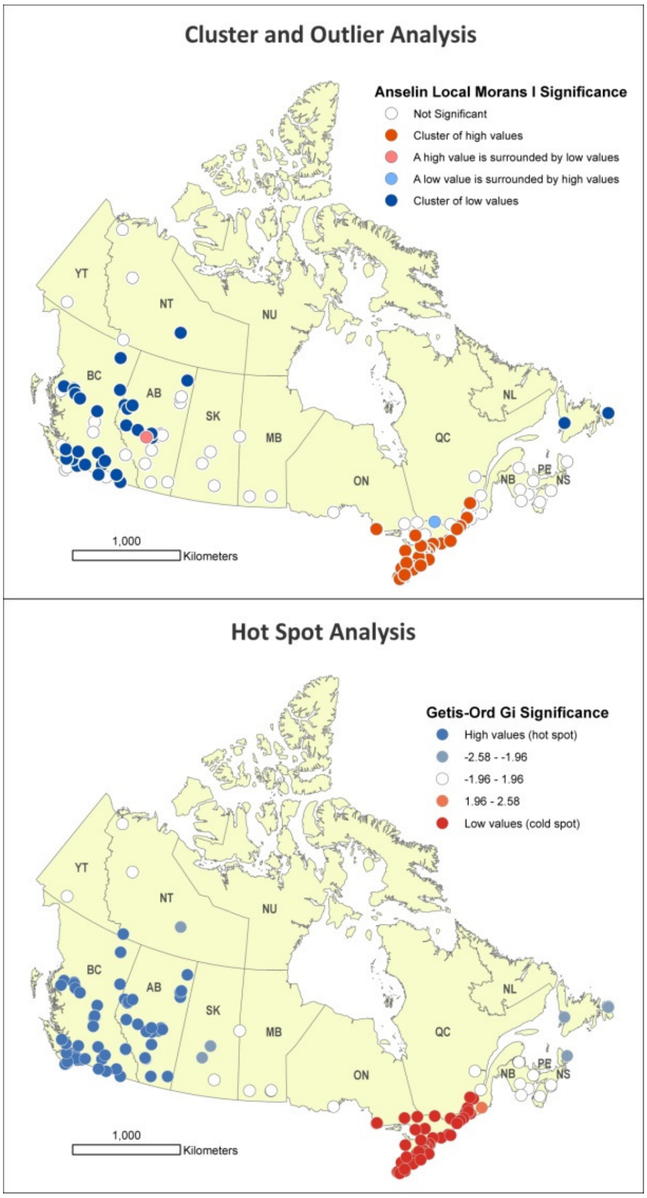

Fig.4. Cluster and outlier analysis (Anselin Local Moran's I) and Hot Spot Analysis (Getis-Ord Gi) of PM2.5 observations.

The maps in Figure 4 show high level of clustering with high values of PM2.5 in the south-eastern part of Canada and high level of clustering with low values of PM2.5 in the two western provinces of Canada. Thus, spatial clustering of PM2.5 observations should be taken into account in a regression and geostatistical models to avoid overcounting type of model bias.

REGRESSION ANALYSIS

Bivariate Linear Ordinary Least Squares Regression

Quantitative measure of correlation is used to confirm or reject relationship hypothesis by using correlation and regression analyses. Correlation analysis in our case study shows very weak association between the dependant PM2.5 variable and independent variables for population and labor indicators. In addition, the residual values are highly clustered and not normally distributed.

Population and labor indicators, used in this study, were collected by Census Canada for polygonal units, Census Subdivisions (CSD) and Census Tract (CT), varying by area. At the same time, fine particulates (PM2.5) were measured at monitoring stations (points). The measurements reflect condition of atmosphere in surrounding areas. To address this issue, population and labor variables were normalized by land area of respective observation polygonal units to be further used in regression analysis.

Normalized density variables have certainly improved the model. We examined few variables, which significantly contribute into concentration of fine particulates. Table 2 shows results of bivariate regression between dependant PM2.5 variable and several independent variables.

Table 2. Results of bivariate regression

	Population density, 2006 **PopDensity**	Density of all occupations of industrial labor **IndLabDens**	Density of employed labor force **EmplDens**	Density of all occupations of labor **AllLabDen**	Density of total population by labor force activity **TPLabDen**	Density of total population in the labor force **TPLInDen**	**Latitude**	**Longitude**
β_1 coefficient	0.000564	0.013272	0.000882	004737	0.000564	0.000832	-0.358161	0.064789
Probability of β_1	0.000000	0.000000	0.000000	0.000002	0.000000	0.000001	0.000000	0.000000
Robust probability of β_1	0.000000	0.000000	0.000001	0.000001	0.000000	0.000000	0.000000	0.000000
Determination coefficient R^2	0.146067	0.154383	0.129887	0.121647	0.146067	0.133101	0.480926	0.335308
Correlation coefficient r	0.382	0.393	0.360	0.349	0.382	0.365	0.693	0.579
Adjusted determination coefficient R_{adj}^2	0.141242	0.149605	0.124972	0.116685	0.141242	0.128203	0.477993	0.331553
P-value of F-statistic for R^2	0.000000	0.000000	0.000001	0.000002	0.000000	0.000001	0.000000	0.000000
P-value of chi-squared for ***Wald***-statistic	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
P-value of chi-squared for ***BP***-statistic	0.889440	0.966295	0.859884	0.863996	0.889440	0.868857	0.828194	0.000001
P-value of chi-squared for ***JB***-statistic	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

The results show that explanatory variables justify considerable percent of fine particulates concentration in the air and correlation coefficients are statistically significant (F-statistic for R2, p-values < 0.05). P-values of chi-squared for Koenker K (BP) statistic, which considers spatial effect within data, are not statistically significant. Therefore, effect of non-stationarity can be neglected, except for association between PM2.5 and Longitude. However, according to Jarque-Bera (JB) statistics, regression residuals are not normally distributed (p-value for this test is less than 0.05 for 95% confidence level), which indicate possible model misspecification. In addition, Moran's I indices show that the residual values are highly clustered or have spatial autocorrelation (see discussion above), and it may indicate that the regression models are missing one or more key explanatory variables.

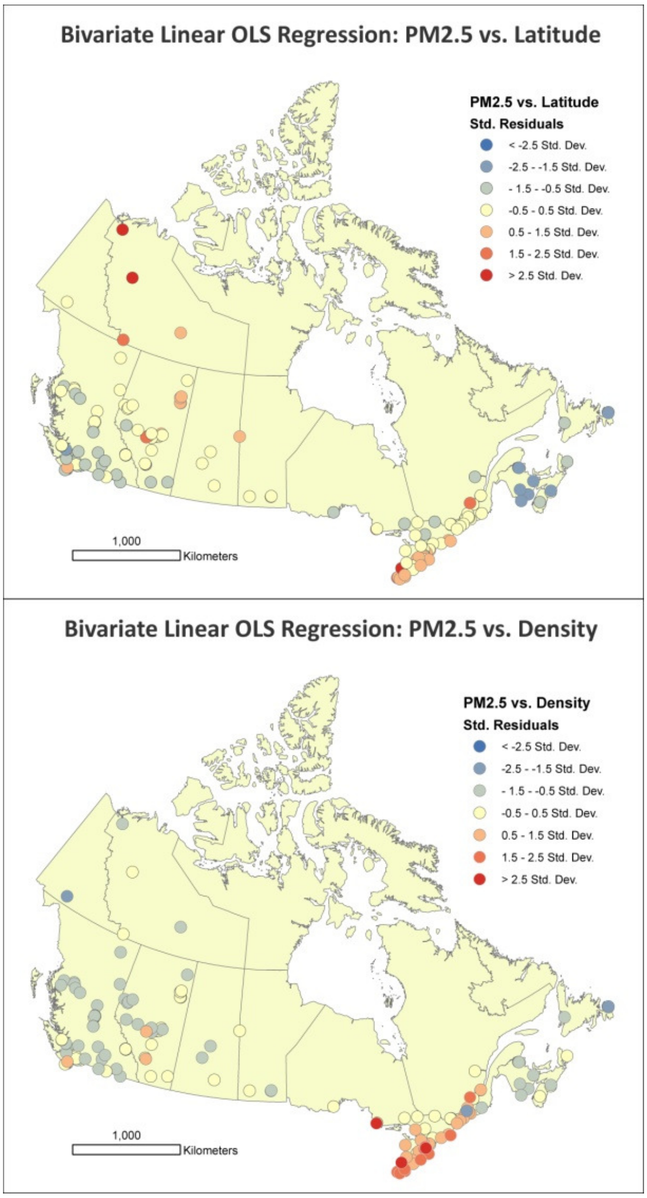

Fig.5. Bivariate linear ordinary least squares regression: a) PM2.5 vs. Latitude (PM2.5 = 23.7248-0.3581 x Latitude), correlation coefficient r = 0.693, goodness of fit SE = 1.822; and b) PM2.5 vs. Density of all occupations of industrial labor (PM2.5 = 5.6278 + 0.0133 x IndLabDens), correlation coefficient r = 0.393, goodness of fit SE = 2.326.

Multivariate Linear Ordinary Least Squares Regression (OLS)

In the next step, relationships between dependent variable and several explanatory variables were modeled by using multivariate regression analysis. We used several iterations to build the best multivariate regression model by including a different number and different variation of explanatory variables and comparing these multivariate regression models among each other by using RMSE criteria until we choose the best one.

We have chosen Latitude, Density of all occupations of labor (AllLabDen), and Density of all occupations of industrial labor (IndLabDens) as key explanatory variables for PM2.5 multivariate regression model. This combination shows acceptable results with all correlation coefficient being statistically significant. The model explains observed dependent variable values with 51.6% of R2 and 50.7% and Radj2. These are highest values from all combinations of variables. One of the measures used to evaluate regression is goodness of fit. Standard error SE, as one of the goodness of fit measures, is the estimated standard deviation for the residuals. In this case, SE has the lowest value equal to 1.77.

The Akaike Information Criterion (AIC) is a measure of model performance and is helpful for comparing different regression models and it equals to 716.37, which is smallest value comparing to other models. The p-value of F-statistic and Wald-statistic are statistically significant. According with K(BP)-statistic, non-stationarity is not statistically significant, however the distribution of PM2.5 residuals is not normal. In addition, the Variance Inflation Factor (VIF) values of two regression coefficients are larger than 7.5, which indicate strong correlation between AllLabDen and IndLabDens variables. While this is an indicator of not the best conditions for a regression, the other regression parameters and diagnostics still suggest using these two variables in the model. In addition, Moran's I indices indicate presence of spatial pattern in PM2.5 residuals (the model is clustered or spatially correlated). Figure 6 illustrates results of the multivariate regression.

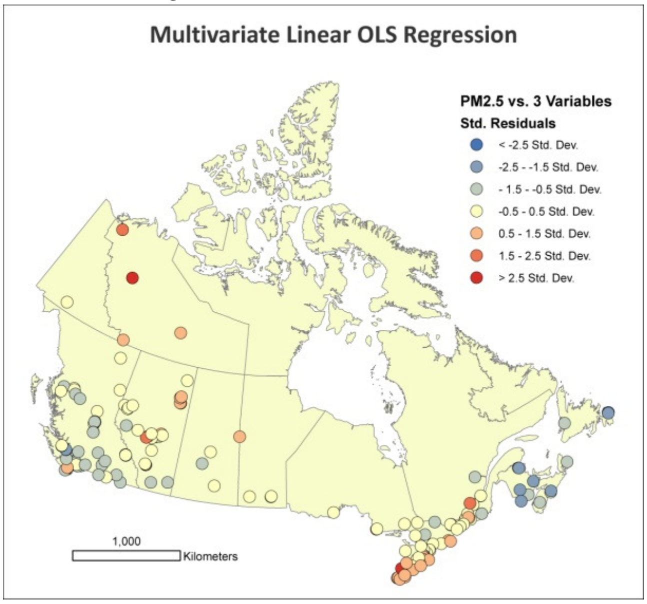

Fig.6. Multivariate linear ordinary least squares regression: PM2.5 = 22.0715 - 0.3275xLatitude + 0.0177x IndLabDens - 0.0051xAllLabDen. R2= 0.5157, goodness of fit SE = 1.77. Results of the regression are similar for both types of projections.

The model has 52% of coefficient of multiple determination R2; in other words, it explains 52% of PM2.5 variations. Thus, the AllLabDen variable can somehow explain the home firewood burning component of fine particulates pollution. The IndLabDens variable can explain industry component of the fine particulates pollution. The model is still missing 48% of PM2.5 variations, which can be result of non-linear relationship between the dependent variable and the independent variables; there are evidences of spatial correlation effect or clustering, and some important key variables are missing.

Local Geographically Weighted Regression

There is a possibility to improve model results by applying local Geographically Weighted Regression (GWR), which takes into account effect of spatial correlation (Fotheringham, 2002). Comparison of R2, SE and AIC values from GWR and Bivariate Linear Ordinary Least Squares Regression (OLS) can advocate for using local regression model (GWR) instead of global model (OLS). However, the GWR PM2.5 local model has not dramatically improved the OLS PM2.5 global model with the Latitude, AllLabDen and IndLabDens explanatory variables. The AIC values are still comparable (716.46 in GWR vs. 716.37 in OLS), however the R2 value in GWR is higher (0.522) comparing to OLS (0.516). The SE is also slightly improved from 1.77 (GWR) to 1.759 (OLS).

The GWR local model with the key explanatory variables Latitude, AllLabDen and IndLabDens explains 52.2% of fine particulates (PM2.5) values in the selected Canadian cites. The PM2.5 residuals of the model are still highly correlated according the Moran's I test. This indicates that the model is still not properly specified and some important key variables are missing, which can explain 48% of PM2.5 residuals.

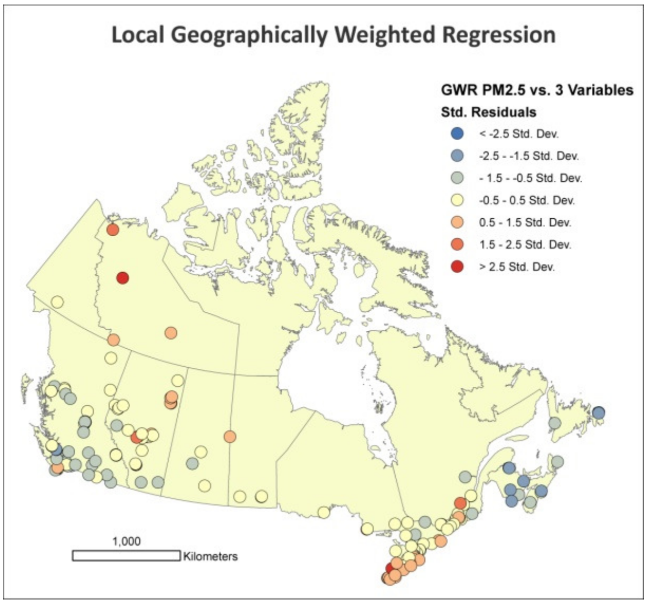

Fig.7. Local Geographically Weighted Regression: SE = 1.7589; AICc = 716.4563; R2 = 0.5224. Results of the regression are similar for both types of projections.

Spatial interaction (spatial autocorrelation) can be explored using spatial econometrics technique (Anselin, 2001). Usually, autocorrelation creates an over-count type of bias for traditional (non-spatial) regression methods; however, this discussion is beyond the scope of this paper.

GEOSTATISTICS

Trend Analysis: Global Polynomial Interpolation

If an experiment contains a quantitative independent variable, then shape of a function relating the levels of this quantitative independent variable to the dependent variable is often of interest. In this case study, we used several basic techniques to explore trends in PM2.5 observations. Figure 8 illustrates trends in PM2.5 observations as the results of applying polynomial models of the first, second and fourth orders.

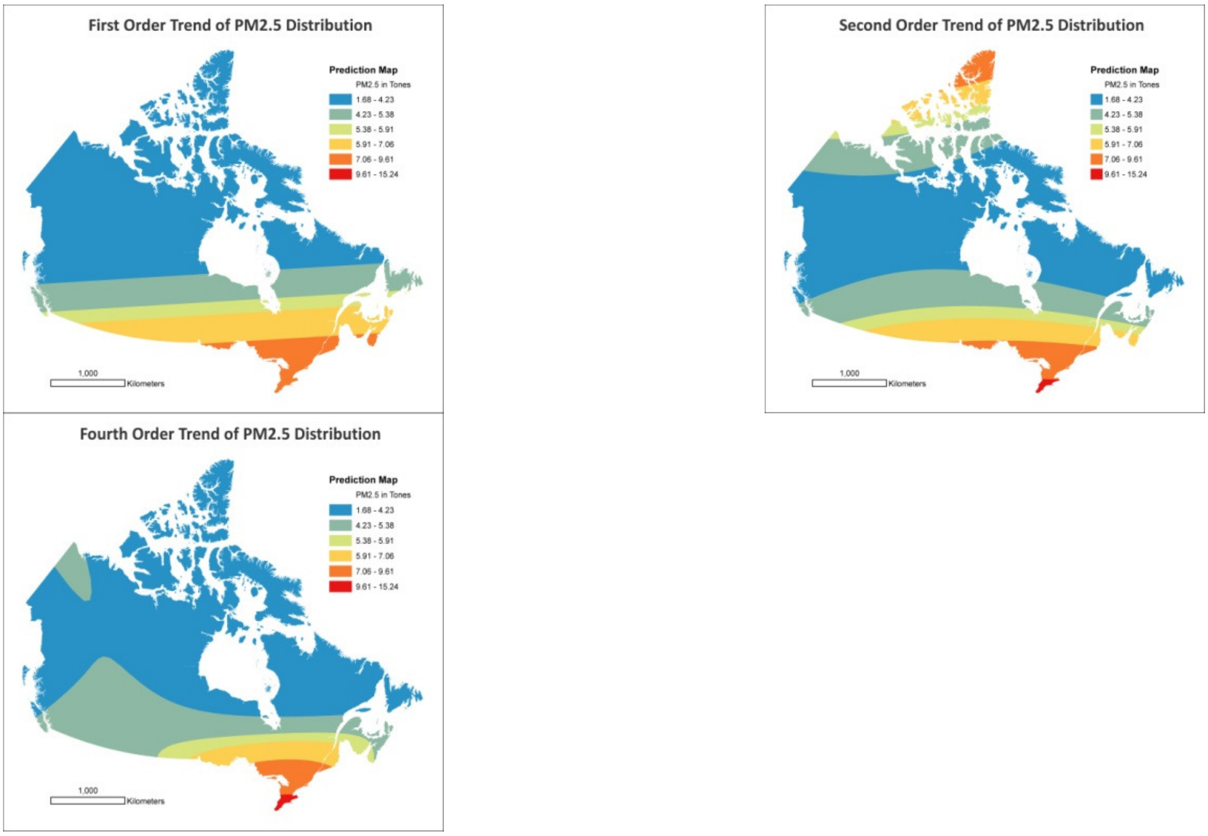

Fig.8. Trend analysis using polynomial models: a) first order (RMSE=1.6190 for Conformal projection, RMSE=1.6351 for Equal Area projection); b) second order, RMSE=1.3295; and c) fourth order,

RMSE=1.553. Apart from first order polynomials, type of projection (Conformal or Equal Area projection) does not significantly affect results of modeling.

The existing trend of PM2.5 data shows the long-range variation of data values. The dominant direction of data changes is from north-west to south-east. It has approximately 140-degree azimuth (based on min RMSE value). Figure 9 illustrates longitudinal and latitudinal projections of PM2.5 data and their trends.

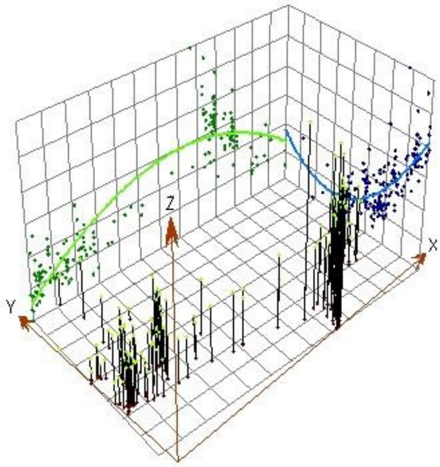

Fig.9. Visual presentation of PM2.5 values in longitudinal and latitudinal projection and their trends.

Linear Kriging and Co-kriging

In this part of research, we used geostatistical approach (kriging) to predict values of PM2.5 observations. Several geostatistical kriging methods were tested to get the best result based on RMSE of PM2.5 values.

Ordinary kriging assumes the model Z(s) = m + e(s), where m is an unknown constant (trend) and e(s) is a fluctuation part. One of the main issues concerning ordinary kriging is whether the assumption of a constant mean is reasonable. Sometimes there are good scientific reasons to reject this assumption. However, as a simple prediction method, kriging has remarkable flexibility. A universal kriging assumes the model, Z(s) = m(s) + e(s), where m(s) is some deterministic function or an n-order polynomial (describing the trend), and e(s) is a fluctuation part (the errors, which are assumed to be random with the mean of all e(s)= 0). Conceptually, autocorrelation is modeled from the random errors e(s) (Journel, 1978). Co-kriging uses supplementary information on co-regionalized variables (co-variables) to improve prediction of a target variable. We used Density of all occupations of industrial labor (IndLabDens) and Latitude as co-variables to model PM2.5 values distribution.

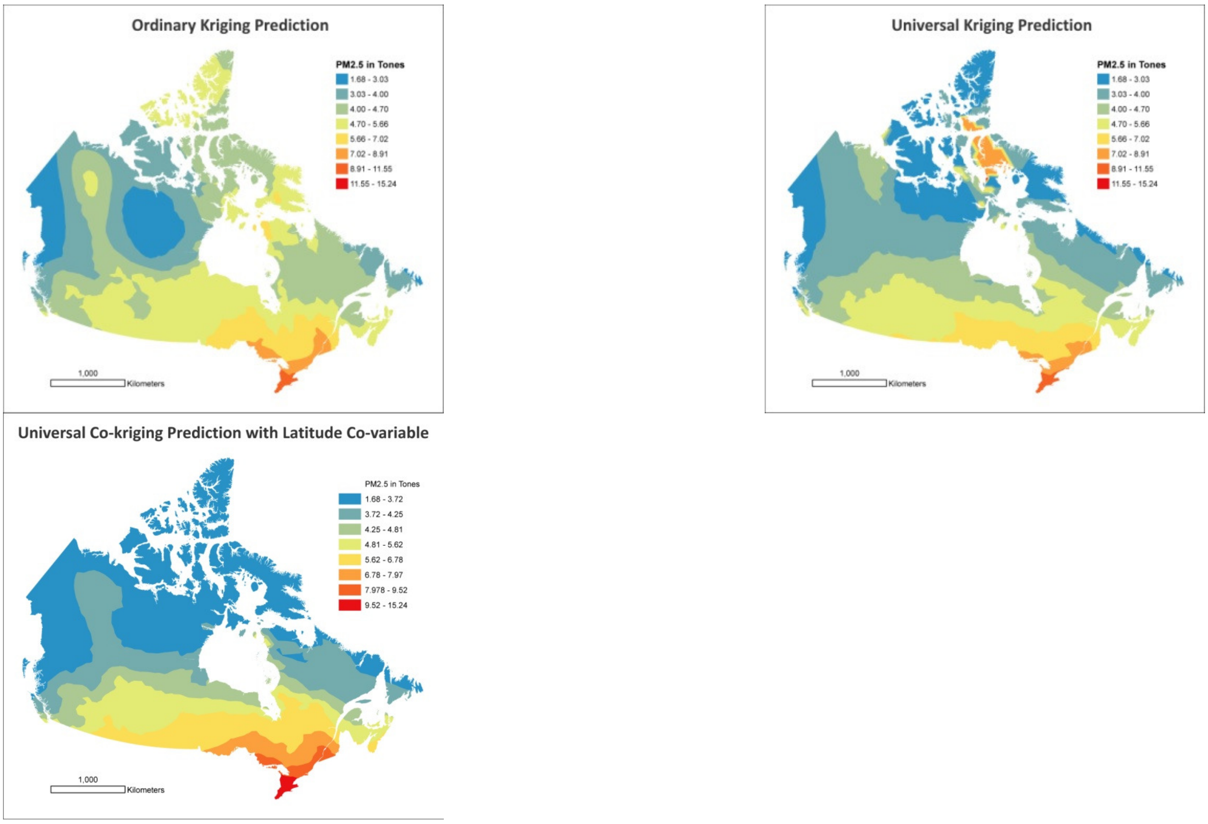

Fig.10. Predictions of PM2.5 observation values based on kriging and co-kriging models: a) Ordinary kriging (RMSE=1.2914); b) Universal kriging: RMSE=1.2749 (conformal projection) and RMSE=1.3002 (equal area projection); c) Linear co-kriging with Latitude co-variable: RMSE=1.2851 (conformal projection) and RMSE=1.2860 (equal area projection). All kriging methods are with 140 deg anisotropy and use normalized source data as a result of logarithmic transformation.

Adding the trend to interpolation in kriging improves data prediction, especially using Conformal projection - the RMSE has slightly decreased from 1.2914 to 1.2768 for ordinary kriging and universal kriging respectively. Adding co-variables IndLabDens and Latitude to interpolation does not improve prediction (co-kriging results in RMSE=1.2851 for conformal projection, and RMSE=1.2860 for equal area projection).

MODEL ANALYSIS AND DISCUSSION

Regression analysis and kriging optimal interpolation are used for prediction PM2.5 values in known locations. According to results of our case study, kriging-based predictions give lower RMSE in situations when a regression model is not properly specified, i.e. when some important key variables are missing. Figure 11 illustrates best regression models (GWR and kriging) for PM2.5 observations, based on analysis of available data. However, if the regression model is specified reasonable, regression and kriging may supplement each other and used for further validation of models.

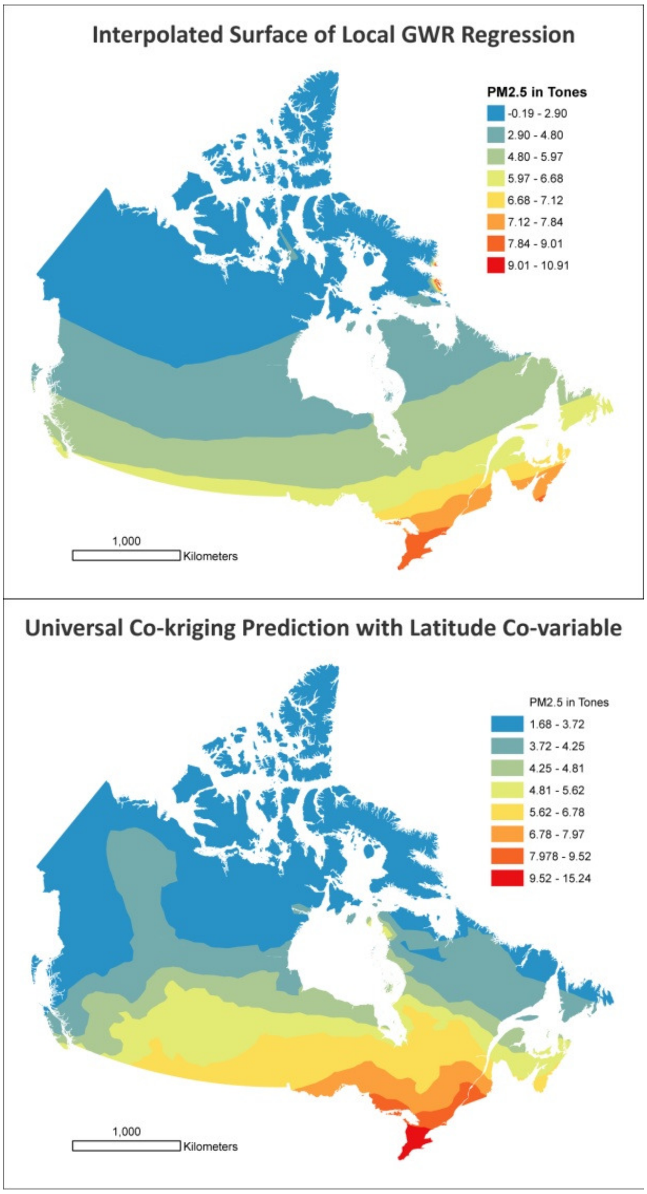

Fig.11. Interpolated surface of a) local GWR Regression, RMSE=1.738; and b) Universal co-kriging (Latitude as co-variable), RMSE=1.286.

Combination of two methods can be beneficial in several situations. Regression and kriging modeling can be used for prediction at all locations where predictions are required. However, for regression analysis, values of the independent variables have to be available in predicted locations. Kriging modeling can be used to estimate the independent variables in predicted locations with the following use in regression. In addition, if one of independent variables of regression is a coordinate, this variable can be defined exactly in required prediction locations, and thus it can improve prediction based on the regression model. In addition to prediction, regression model can be used to examine and explore spatial relationships between dependent variable and independent (explanatory) variables. Key exploratory variables from regression modeling can be used as co-variables in co-kriging modeling. Moreover, it can be revealed even from the respective maps, that the results and natures of regression and global polynomial interpolations are very similar. Therefore, it can be an option to use the regression surface to obtain residuals for kriging modeling.

Explored geostatistical methods also shown high level of robustness to parameters of projections used to represent source dataset, with some excerption of global spatial autocorrelation (Moran's I index) and universal kriging.

CONCLUSION

Geostatistical mapping of environmental characteristics is a rapidly evolving area, which based on combined power of well developed methods of spatial statistics and geovisualization capabilities of modern GIS packages. Geostatistical mapping, partially based on classical statistics and mapping, has a number of unique methods, specifically designed to address spatiality of geographical data. While the real world provides very large range of applications for geostatistical mapping, some major steps of environmental spatial data analysis, described in this paper, are common and relevant to most studies.

REFERENCES

Anselin, L. 2001, Spatial econometrics, A Companion to Theoretical, Econometrics, in Baltagi B. (ed.), Blackwell, Oxford, 310–330.

Anselin, L. 1995, Local Indicators of Spatial Association - LISA, Geographical Analysis 27(2), 93–115.

Getis, A., and Ord, J. K. 1992, The Analysis of Spatial Association by Use of Distance Statistics, Geographical Analysis 24, no. 3.

Fotheringham, S. A., Brunsdon, C., and Charlton, M. 2002, Geographically Weighted Regression: the analysis of spatially varying relationships, John Wiley & Sons.

Journel, A.G. and Huijbregts, C.J. 1978, Mining Geostatistics, Academic Press London.