SPATIAL MODELLING OF GEOCODED CRIME DATA

Krista Collins

ABSTRACT

How are different types of crime distributed across cities? What factors are associated with high neighbourhood crime rates? These are but a few of the questions Statistics Canada is addressing in the analysis of crime rates from city neighbourhoods, as part of the project on geocoding crime data. To ensure the analysis is accurate the effects of spatial autocorrelation must be accounted for in the model. This paper will give an overview of the spatial autoregressive models used to analyse the crime rate in several Canadian cities and the potential influence these models have for crime reduction strategies in Canada.

KEY WORDS: Simultaneous autoregressive model, Spatial autocorrelation, Spatial error, Spatial lag model

RÉSUMÉ


MOTS CLÉS : Autocorrélation spatiale; modèle à autocorrélation spatiale des erreurs; modèle à variable endogène décalée; modèle autorégressif simultané; modèle à variable endogène décalée;

1. INTRODUCTION

There are often many reasons to expect that data which are located close to each other in space will be more related than data that are further apart; this is a basic law of geography (Cressie 1993). By analysing the spatial structure of the data, more information can be learnt about the nature of the spatial process. For example, the distribution of crime within a city is known to be affected by the properties of the different city regions and by the social interactions of the people in those areas (Puech 2004). By mapping the distribution of crime, regions with a high concentration of crime, known as hot-spots can be detected. The crime data can then be related to social and demographic characteristics to discover the factors associated with high and low crime neighbourhoods. The distribution of neighbourhood crime can also be compared across major Canadian cities. This information can then be used by police services and community planners to identify potentially high-risk areas and to make informed decisions on resource allocation and crime reduction strategies. This is precisely the objective of the geocoding crime analysis project at Statistics Canada (Savoie et al., 2006; Savoie, 2008; Statistics Canada, 2008).

Geocoding is the process of assigning an exact latitude and longitude coordinate to a street address (Savoie et al., 2006). In this case the street address corresponds to the location where a criminal incident took place. Geocoded crime data can then be integrated into a geographic information system to relate the incidents to the characteristics of the neighbourhoods in which they reside. The relationship between crime and neighbourhood characteristics can then be modelled. If the spatial structure in the data is not completely accounted for in the regression model then the residuals will be spatially autocorrelated. Spatial autocorrelation, simply defined as the non-zero covariance between observations that are related in space, violates the fundamental assumption of independence of the observations in a standard regression model. Spatial
autocorrelation can cause inefficient estimation of the parameters in a standard regression model and can lead to inaccurate estimation of the sample variance and significance tests.

This paper will give an overview of the geocoding crime analysis project at Statistics Canada. The focus will be on the spatial models used to relate the neighbourhood crime rate to various socio-economic and demographic variables from the neighbourhood. Interested persons are referred to the original analysis papers for more information about the other aspects of the analysis covered in the geocoding project (Savoie et al., 2006; Savoie, 2008; Statistics Canada, 2008). The paper will start with an overview of the spatial models used in the analysis of neighbourhood crime. In section 3, the importance of modeling the spatial structure of the data is illustrated with an analysis of violent crime in Halifax and property crime in Montréal. The paper concludes with some comments about the use of spatial models.

2. SPATIAL AUTOREGRESSIVE MODELS

Spatial data can be classified into one of three types: regional, geo-statistical and point patterns (Cressie, 1993). The geocoded incidents of crime are classified as point patterns which can be mapped to illustrate the distribution of crime in a given area (Figure 3B and 4B). However, to create a model relating crime to the characteristics of the neighbourhood in which it occurs, the crime data is aggregated to regions rather than modelling the exact location of all the criminal incidents. This is classified as regional data. By definition the regions span the entire study area and the focus of the analysis is on modelling the regional data as it is defined. There is hence no possibility of interpolation between two adjacent regions. The observed data points are assumed to be representative of the entire region in which they reside.

Although there are two main realms for the spatial analysis of regional data, only simultaneous autoregressive (SAR) models, which are commonly used for the analysis of crime in the literature, were used for the geocoding project. SAR models are hence the only form spatial models considered in this paper.

2.1 Taxonomy of Simultaneous Autoregressive Models

Prior to writing a formula for the SAR models, a definition of how the regions relate to one another in the model is needed. In the simplest form of the model, regions influence one another only if they are considered neighbours. Neighbouring locations can be defined by a contiguity structure, which is based on the geographical arrangement of the neighbourhoods; by a distance band, in which all regions within a certain specified distance are classified as neighbours; or by k-nearest neighbours, in which the k-specified closest regions are considered neighbours (Durbin, 1998). These different possibilities are illustrated in Figure 1.

Though there are two main realms for the spatial analysis of regional data, only simultaneous autoregressive (SAR) models, which are commonly used for the analysis of crime in the literature, were used for the geocoding project. SAR models are hence the only form spatial models considered in this paper.

2.1 Taxonomy of Simultaneous Autoregressive Models

Prior to writing a formula for the SAR models, a definition of how the regions relate to one another in the model is needed. In the simplest form of the model, regions influence one another only if they are considered neighbours. Neighbouring locations can be defined by a contiguity structure, which is based on the geographical arrangement of the neighbourhoods; by a distance band, in which all regions within a certain specified distance are classified as neighbours; or by k-nearest neighbours, in which the k-specified closest regions are considered neighbours (Durbin, 1998). These different possibilities are illustrated in Figure 1.

![Figure 1: Different methods for defining neighbouring regions for the selected region in yellow. A) Rook contiguity – includes locations with a common border, B) Queen contiguity – includes regions with a common border or vertex, and C) Distance – includes all regions whose centroid is within the specified distance band.](image)

The desired neighbourhood structure is represented in the model by a spatial weights matrix, \( W \). The spatial weights matrix is a binary \( N \times N \) matrix, where the off-diagonal elements, \( W_{ij} \) for \( i \neq j \), equal one if location \( j \) is a neighbour of location \( i \) and zero otherwise. Thus the spatial weights matrix allows for the potential interaction of neighbouring locations in the model and rules out spatial dependence for non-neighbouring regions. By convention, zeros are placed on the diagonal elements of the matrix, \( W_{ii} \), indicating that a location cannot be a neighbour of itself. The elements of the matrix are then row standardised so that the sum of each row is equal to one.

The generic form of the SAR model can then be written as:

\[
y = X\beta + \rho W_y + \varepsilon = \lambda W_2\varepsilon + u
\]  

(1)
where $y$ is the $N \times 1$ vector of the dependent variable, $X$ is the $N \times k$ matrix of independent variables, $\beta$ is the $k \times 1$ vector of regression coefficients, $\rho$ and $\lambda$ are scalar spatial autoregressive parameters, $W_1$ and $W_2$ are spatial weights matrices and $\varepsilon$ and $u$ are the $N \times 1$ vectors of error terms (Anselin, 1988). For this paper the residuals, $u$, are assumed to be normal with a homoskedastic variance, although it is possible to introduce heterogeneity in the model by relaxing this assumption.

Equation 1 shows that there are two different ways of adding spatial effects into a standard regression model; either by adding a spatial autoregressive term $\rho W_1 y$ or by modeling spatial dependence in the error terms, $\lambda W_2 \varepsilon$. The use of the subscripts 1 and 2 on the spatial weights matrices is simply to denote the possibility of modelling a different spatial structure for the different spatial terms. Since the weights matrices are row-standardized, the terms $W_1 y$ and $W_2 \varepsilon$ represent the average value of the dependent variable and the residuals from neighbouring locations, respectively.

By setting either or both of the spatial parameters, $\rho$ and $\lambda$ in equation 1 equal to zero, different forms of SAR models can be obtained. If both $\rho$ and $\lambda$ are set to zero, then the SAR model reduces to a standard linear regression model ($y = X\beta + \varepsilon$). By removing the spatial dependence in the error terms ($\lambda = 0$) the SAR model reduces to the spatial lag model:

$$y = X\beta + \rho W_1 y + \varepsilon$$

where the term $\rho W_1 y$ is called the spatial lag (Anselin, 1988). If the spatial dependence is only in the error terms, obtained by setting $\rho = 0$, then the resulting SAR model is called the spatial error model:

$$y = X\beta + \varepsilon, \quad \varepsilon = \lambda W_2 \varepsilon + u$$

The full SAR model, equation 1, contains both the spatial lag and spatial error terms. This model is rarely used in practice and is not described in this paper. The spatial lag model (2) and the spatial error model (3) are described in detail in the next two sub-sections and they are illustrated in the analysis of property crime in Montréal in sections 3.3 and 3.4.

### 2.2 Spatial Lag Model

In the spatial lag model (2), the inclusion of the spatial lag as a covariate in the model implies there is a theoretical assumption that the dependent value at one location has a direct influence on the value at neighbouring locations. In the case of crime data this direct effect of neighbouring regions could be caused by spatial diffusion as crime spreads from one location to another or by other social interactions. The magnitude of the spatial coefficient, $\rho$, could then be interpreted as a measure of the effect of neighbouring locations given all the other variables in the model; however this is not necessarily the case. The spatial lag coefficient only represents the effect of neighbouring locations if the spatial units that are used to aggregate the data match the underlying spatial structure in the data. In other words, the variable of interest needs to be homogeneously distributed in each neighbourhood, as shown in Figure 2A. Otherwise there is a mismatch between the spatial scale used to measure the phenomenon and the spatial scale at which it occurs (Figure 2B).

The spatial lag coefficient then indicates the strength of the mismatch between the scales. This mismatch between the spatial units is common when predefined administrative units are used for the analysis. In this case the spatial lag term has no real meaning in the interpretation of the final model. In practice for most social phenomena the spatial lag coefficient represents in part the direct effect of neighbouring locations and the effect of the mismatched scale. The spatial coefficient is still an important part of the model as it accounts for the spatial structure and allows the other variables in the model to be interpreted in the same way as coefficients in a standard regression model.

![Figure 2](image)

Figure 2: Interpretation of the spatial lag coefficient: A) illustrates the spatial units matching and B) illustrates a different spatial scale of measurement and occurrence.

The spatial lag model can also be written as:

$$y = (I - \rho W_1)^{-1} X\beta + (I - \rho W_1)^{-1} \varepsilon$$

In this form it is apparent that the value of $y$ at a given location is dependent on not only the independent variables from that location but it is also dependent on the independent variables from all other locations by means of the spatial
multiplier, \((I - \rho W_i)^{-1}\) (Anselin, 2002). The dependent variable is also correlated with the error term at all locations. These correlations between the variables that are induced by the model make least squares an inefficient method of estimating the model parameters. The parameters must instead be estimated by maximum likelihood, a two-stage least squares approach or by instrumental variables. In the maximum likelihood approach, an estimate of \(\rho\) is obtained first by the non-linear optimization of the following log-likelihood equation:

\[
L = -N \ln \left( \frac{(e_o - \rho e_i)'}{(e_o - \rho e_i)} \right) + \sum \ln(1 - \rho w_i),
\]

where the variables \(w_i\) represent the eigenvalues of the spatial weights matrix, \(e_o\) represents the residuals from a regression of \(y\) on \(X\), and \(e_i\) represents the residuals from the regression of \(W_i y\) on \(X\) (Anselin and Bera, 1998). Maximum likelihood can then be used to estimate the other coefficients and the standard deviation by using the estimated value of \(\hat{\rho}\).

The model estimates have the asymptotic properties of consistency, normality and efficiency (Anselin and Bera, 1998). The asymptotic variance, used for hypothesis testing, can be estimated as:

\[
\text{AsyVar}[\hat{\rho}, \hat{\beta}, \hat{\sigma}] = \begin{bmatrix}
\text{tr}(W_A) + \text{tr}(W_i W_A) + \frac{W_i X \hat{\beta}}{\hat{\sigma}^2} W_i X \hat{\beta}' & \frac{(X' W_i X \hat{\beta})'}{\hat{\sigma}^2} \frac{1}{N} W_i X \\
\frac{1}{\hat{\sigma}^2} & 0 \\
\frac{1}{\hat{\sigma}^2} & 0 \\
\end{bmatrix}
\]

where \(W_A = W_i (I - \rho W_i)^{-1}\). As shown in the matrix, the covariance between the error term and the regression coefficient is zero, but there is a non-zero covariance between the spatial coefficient and the error and between the spatial lag and regression coefficients.

Contrary to a standard regression model there are two types of observed error terms for the spatial lag model, as the prediction error and the residuals are not equal (Anselin, 2005). As usual, the prediction errors are defined as the difference between the observed and fitted values, \(p.e. = y - \hat{y}\), where \(\hat{y} = (I - \hat{\rho} W_i)^{-1} X \hat{\beta}\). By definition the prediction errors are spatially autocorrelated. The residuals are defined as \(e = (I - \hat{\rho} W_i) y - X \hat{\beta}\), and they will be uncorrelated if the model adequately accounts for the spatial autocorrelation in the data. Therefore all tests for the model assumptions and for any remaining unexplained spatial autocorrelation in the data are done with the residuals and not the prediction errors.

### 2.3 Spatial Error Model

The spatial error model (3) can also be written as:

\[
y = X\beta + (I - \lambda W_i)^{-1} u
\]

which illustrates that the spatial disturbance is completely modelled in the residuals. The spatial error model is used when spatial autocorrelation is found in the residuals from a standard regression model and the dependence needs to be accounted for in the model to obtain unbiased and efficient estimates of the regression parameters. This spatial dependence could be the effect of unmeasured variables or measurement errors which have a systematic spatial pattern. There is thus no direct interpretation of the spatial error coefficient, \(\lambda\), so it is treated as a nuisance parameter.

The spatial error model can be estimated using generalised least squares conditional on the value of \(\lambda\), since the spatial dependence is completely contained in the model error terms. The value of \(\lambda\) is estimated first by the non-linear optimization of the following log-likelihood equation:

\[
L = -\frac{N}{2} \ln \left( \frac{u' u}{N} \right) + \sum \ln(1 - \lambda w_i)
\]

where \(u' u = y_L' y_L - y_L' X_L [X_L' X_L]^{-1} X_L' y_L\), and \(y_L\) and \(X_L\) are spatially filtered variables: \(y_L = y - \lambda W_i y\) and \(X_L = X - \lambda W_i X\), and the variables \(w_i\) represent the eigenvalues from the spatial weights matrix. The resulting estimate of \(\hat{\lambda}\) is then used to estimate the other regression coefficients and the variance by generalised least squares. The asymptotic variance matrix of the parameter estimates is block diagonal between \(\hat{\beta}\) and the other two parameters, \(\hat{\sigma}^2\) and \(\hat{\lambda}\). It is estimated by
AsyVar[\hat{\beta}] = \hat{\sigma}^2 [X'X]^{-1} \tag{9}

and

AsyVar[\hat{\sigma}^2, \hat{\lambda}] = \begin{bmatrix}
N / 2\hat{\sigma}^4 & tr(W_b) \\
tr(W_b) & tr(W_b)^2 + tr(W_b'W_b)
\end{bmatrix}^{-1} \tag{10}

where \( W_b = W_x (I - \hat{\lambda}W_x)^{-1} \) (Anselin and Bera, 1998).

Different structures for the spatial dependence in the error terms can also be used to create alternate forms of the spatial error model instead of the autoregressive process that is illustrated here. For example, a spatial moving average process, in which the error term is dependent on a random uncorrelated error term and the average neighbouring values of the uncorrelated error, could also be used with this model structure (Anselin and Bera, 1998).

3. Analysis of Neighbourhood Crime

3.1 Description of Variables

3.1.1 Crime Data

The crime data used in all of the spatial models is obtained from police services through the Uniform Crime Reporting Survey (UCR). The UCR, administered by the Canadian Centre for Justice Statistics (CCJS) at Statistics Canada, collects administrative data from all of the police services across Canada. Thus any analysis of UCR data is based only on the crimes known to the police, which gives a specific picture of the nature and the extent of crime in Canada (Savoie et al., 2006).

For most of the analysis done at CCJS the UCR data is grouped into three broad categories of crime: violent, property and other offences. Violent offences are defined as crimes against the person and include: homicide, attempted murder, sexual assault, assault and robbery. Property offences are defined as crimes against property and include: arson, breaking and entering, theft, possession of stolen goods and mischief. The category for other offences includes all the offences that are not classified as violent or property, such as bail violations, drug possession or trafficking, and counterfeiting currency. Many of these ‘other’ crimes cannot be modelled spatially as there is no physical location where the offence takes place. For example, violations against the administration of justice, such as bail violations and failure to appear in court are usually given a default location of the court house or police station. For this reason only violent and property crimes are analysed with the spatial models.

Since the volume of crime in any given region is directly related to the size of the population in the region, crime statistics are usually displayed as a rate of the number of incidents per population. This provides a good representation of the crime in large geography areas, such as cities or provinces, where the population on a daily basis is relatively constant. However a crime rate based solely on the residential population is not sufficient for representing the rate within a city neighbourhood, since people move between neighbourhood during the day and have the potential to be a victim of crime in any of the neighbourhoods that they visit. Instead, for the purpose of this analysis, the crime rate is calculated as the number of criminal incidents per population at risk, where the population at risk is defined as the combined population of residents and people who work in the area. Using the population at risk helps to stabilise the crime rate which would otherwise be inflated for downtown and commercial areas where there is a small residential population but a high concentration of people working in the area or engaging in other activities (Savoie et al., 2006). In all of the models the crime rate is log transformed to obtain an approximately normal distribution for the dependent variable.

3.1.2 Explanatory Variables

The crime data was related to various neighbourhood characteristics that describe the demographics and socio-economic status of the resident population, the condition of the dwellings in the neighbourhood and the area is zoned for the different types of land use. The demographic and dwelling characteristics were obtained from the Census of Population conducted on May 15, 2001. The demographic characteristics include the percentage of males aged 15 to 24 in the
resident population, the proportion of the population who are aged 15 and over and have never been married, the percentage of the population who are members of a visible minority, the proportion of single mother families among economic families living in private households, and the proportion of residents who live alone. The socio-economic characteristics included variables about the median household income, the proportion of residents over 20 with a bachelor’s degree and the proportion of the population living in a private household with a low income in 2000. Variables describing the conditions of the dwellings in the neighbourhood, such as the proportion of dwellings in need of major repair and the proportion of household that are occupied by the owner, were also included in the preliminary models.

The land-use variables describe the proportion of the neighbourhood that was zoned for commercial land use and for single and multiple family dwellings. The land-use data was obtained directly from each of the cities analysed in this project. Additional variables describing other factors thought pertinent to the city were also added to the analysis. For example, this includes a variable representing the density of bars and other drinking establishments over the geographic area in Montréal, and the distance from the downtown core in Thunder Bay.

Prior to the analysis the explanatory variables were transformed to have a mean of zero and a standard variance. For further information on the variables used in this analysis and information on their potential influence on the crime rate refer to Savoie et al. (2006) and Savoie (2008).

3.2 Analysis of Crime in Halifax

The study area of Halifax included the portion of the Halifax Regional Municipality that is primarily serviced by the Halifax Regional Police. The Halifax Regional Municipality consists of four separate municipalities: Halifax, Dartmouth, Bedford and Halifax County, which were amalgamated in 1996 to be governed by a single city council. The study area includes the region south-west of the harbour, the former municipality of Halifax and the region north-east of the harbour, the former municipality of Dartmouth (Figure 3A). The remainder of the Halifax Regional Municipality, including the region north of the harbour that connects Halifax and Dartmouth, known as Bedford, are serviced by the Halifax Country Rural RCMP and were not included in the analysis. The study area is thus divided into two separate regions by the Halifax harbour. This is the first indication that there is a location effect in the data that needs to be considered in the analysis.

![Figure 3: A) The Halifax study area with neighbourhoods defined by census tracts. B) The distribution of violent crime.](image)

The study area spans 160 square kilometres and had a resident population of 191,514 people in 2001. The neighbourhood boundaries used in this analysis are defined by the 51 census tracts in the areas. Census tracts are small relatively stable geographic areas with a population of 2,500 to 8,000 residents, defined by Statistics Canada in conjunction with local specialists. For the purpose of the spatial analysis, the census tracts are considered neighbours of each other based on a queen contiguity structure, which includes regions that either have a border or vertex in common.
The distribution of violent crime in Halifax is concentrated in the downtown area, which is located on either side of the Halifax harbour (Figure 3B). The log-transformed violent crime rate was modelled by the various neighbourhood characteristics using a standard linear regression model in SAS. Since the study area is divided into two separate regions by the harbour, a binary location variable indicating if the neighbourhood was located south-west of the harbour (in Halifax) or north-east of the harbour (in Dartmouth) was included as a variable in the analysis. A stepwise procedure was then used to determine the optimal set of variables to include in the model and variance inflation factors (VIFs) were used to test for multicollinearity amongst the explanatory variables.

The location variable was marginally significant in the model of the entire study area (Table 1) so separate models were also fit to the regions north-east and south-west of the harbour. A comparison of the regional models indicates that different variables are more closely related to violent crime in the different regions of the city, which is not apparent from a model of the entire study area with the binary location variable. For example, neighbourhoods where a higher percentage of the resident population had a bachelor’s degree tended to have less violent crime in the north-east regions, all else being equal. This was not true in the regions south-west of the city which is where the three local universities are located. This difference may be because residents with a bachelor’s degree are more evenly spread out in the south-west regions around the three universities which are all in neighbourhoods with a varying level of violent crime. Regardless of the reason for the difference, accounting for the spatial nature of the city allows for a better fit of the model to the data. The adjusted R-squared value for the model of the entire study area was 0.48, while the value for the model north-east of the harbour was 0.81 and for the regions in the south-west the value was 0.60.

Table 1: The standard linear regression models for violent crime in the entire Halifax study area and the neighbourhoods north-east and south-west of the Halifax harbour.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Entire Study Area</th>
<th></th>
<th>North-East</th>
<th></th>
<th>South-West</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff</td>
<td>p-value</td>
<td>Coeff</td>
<td>p-value</td>
<td>Coeff</td>
<td>p-value</td>
</tr>
<tr>
<td>Intercept</td>
<td>3.15</td>
<td>&lt;0.0001</td>
<td>2.62</td>
<td>&lt;0.0001</td>
<td>2.62</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>% Lone mother family</td>
<td>0.27</td>
<td>0.120</td>
<td>0.42</td>
<td>0.0080</td>
<td>0.3</td>
<td>0.0077</td>
</tr>
<tr>
<td>% Living alone</td>
<td>0.46</td>
<td>&lt;0.0001</td>
<td>..</td>
<td>..</td>
<td>0.82</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>% Major repairs</td>
<td>0.26</td>
<td>0.0102</td>
<td>..</td>
<td>..</td>
<td>0.27</td>
<td>0.033</td>
</tr>
<tr>
<td>% Commercial zone</td>
<td>..</td>
<td>..</td>
<td>0.26</td>
<td>0.011</td>
<td>..</td>
<td>..</td>
</tr>
<tr>
<td>% Bachelor degree</td>
<td>..</td>
<td>..</td>
<td>-0.45</td>
<td>0.0022</td>
<td>..</td>
<td>..</td>
</tr>
<tr>
<td>Binary Location Variable</td>
<td>-0.44</td>
<td>0.0408</td>
<td>..</td>
<td>..</td>
<td>..</td>
<td>..</td>
</tr>
</tbody>
</table>

The residuals from the least squares models were then tested for the presence of spatial autocorrelation using Moran’s I statistic, a global measure of the linear association between the residuals and their average value at neighbouring locations (Anselin and Bera 1998). The Moran’s I value for the model of the entire region was 0.129 (p=0.0853), the value was -0.077 (p=0.2997) in the north-east and the value for the south-west model was -0.065 (p=0.2773), which indicates that the spatial dependence shown in the distribution of violent crime in the city is accounted for by the explanatory variables in the model and a spatial model is not necessary.

3.3 Analysis of Crime in Montréal

The study area for the Montréal analysis was the Island of Montréal which is part of the larger Census Metropolitan Area (Figure 4A). The island of Montréal has a residential population of approximately 1.8 million and spans an area of approximately 500 square kilometres. The neighbourhood boundaries for the analysis were defined by the 521 census tracts on the island. Fifteen of the census tracts were removed from the analysis because the resident population was less than 250 inhabitants. Statistics Canada suppresses information on regions with a resident population that is this small due to issues of confidentiality and data quality. As with the Halifax analysis, neighbouring regions were defined by the queen contiguity structure.

Property crime on the island of Montréal is concentrated in one hot-spot located near the downtown core of the island (Figure 4B). The log-transformed property crime rate was first modelled with the neighbourhood characteristics using a standard linear regression model. The stepwise procedure in SAS Proc REG was used to select an initial set of variables to include in the model (Table 2). Since many of the explanatory variables are highly correlated with each other, VIFs were used to ensure there was no problem of multicollinearity in the model. All of the VIFs were less than 3 indicating there
was no multicollinearity among the predictor variables. The residuals of the standard regression model were then tested for the presence of spatial autocorrelation using a Moran’s I statistic. The value of the statistic was 0.23 for property crime \( (p<0.0001) \), which indicated there was a significant amount of spatial autocorrelation in the residuals from the least squares model making the estimates inaccurate.

Figure 4: A) The island of Montréal with neighbourhoods defined by census tracts. B) The distribution of violent crime.

Lagrange Multiplier (LM) tests were then used to determine the form of the spatial dependence in the data (Anselin et al., 1996). The standard versions of the LM tests are powerful but not robust to local misspecification of the model, so the LM test for spatial error dependence can be significant even if the form of the spatial dependence more closely resembles a spatial lag structure or vice versa. The standard LM tests for spatial lag and spatial error dependence were both significant, 105.90 \( (p<0.001) \) and 74.88 \( (p<0.001) \), respectively, so the robust versions of the LM test must be used instead. The robust LM tests account for one form of spatial dependence while testing for the other form by adjusting the asymptotic mean and variance of the test. The robust tests are less powerful than the standard tests but are more robust to model misspecification. The robust LM tests for spatial lag dependence was significant \( (LM=32.47, p<0.001) \), while the robust LM test for spatial error dependence was not \( (LM=1.36, p=0.243) \). Thus a spatial lag model was fit to the property crime data in Montréal.

The spatial lag model was estimated using programs written in the SAS procedure PROC IML. A manual stepwise procedure was then used to determine the best model for the data (Table 2). A comparison between the standard linear regression model and the spatial lag model for the property crime rate shows that for most of the variables there are only minor differences in the estimated regression coefficients. However it is more important to note the differences between the level of significance of some of the variables. The percentage of dwellings in need of major repair is a significant variable in the standard regression model at a 5% level of significance, but this variable is clearly not significant in the spatial lag model. There is also a minor change in the p-values for the variables representing the percentage of residents who are members of a visible minority and the number of bars per square kilometre. The reason for the differences in the level of significance was caused by the positive spatial autocorrelation that was present in the residuals from the standard regression model (Legendre, 1993). Due to the positive spatial autocorrelation, the sample variance in the least squares model was underestimated making the variables appear more significant than they were in actuality. The significance of the ‘major repairs’ variable may have also been influenced by the proportionately large change in the estimated coefficient which was -0.039 in the standard linear model and -0.013 in the spatial model. The variable ‘major repairs’ was removed from the final model. All of the other variables are retained in the model and they have the same interpretation as a standard least squares regression model, except for the spatial lag coefficient, which represents in part the mismatch between the spatial units and the direct effect of neighbouring locations.

Other model diagnostics were used to confirm that the spatial lag model was more appropriate for the property crime data and that the addition of the spatial lag term did not results in a reduction of the model fit. The Akaike Information Criterion (AIC) was 220.5 for the spatial model compared to 304.1 for the standard regression model. Thus despite adding an additional term to the model there was still a significant increase in the log-likelihood of the model. The square correlation between the observed and predicted values was also slightly higher for the spatial model \( (0.62) \) than the standard regression model \( (0.59) \). Although this was only an approximate measure of the goodness of fit of the spatial model, it did indicate the spatial lag model performs at least as well as the standard regression model.
To ensure the spatial autocorrelation in the data was completely accounted for by the final model, the residuals were tested for remaining spatial autocorrelation using another LM test, in which the residuals from the spatial lag model are tested for spatial error dependence (Anselin et al., 1996). The results of the LM test was 1.72 (p=0.190) indicating that the spatial structure of the data was adequately explained by the spatial lag model.

Table 2: The standard linear regression and spatial lag models fit to the rate of property crime on the island of Montréal.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Standard Reg</th>
<th>Spatial Lag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.47</td>
<td>1.98</td>
</tr>
<tr>
<td>Coeff</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Low Inc. Private</td>
<td>0.17</td>
<td>0.11</td>
</tr>
<tr>
<td>Coeff</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Visible Minority</td>
<td>-0.08</td>
<td>-0.05</td>
</tr>
<tr>
<td>Coeff</td>
<td>&lt;.0001</td>
<td>0.0040</td>
</tr>
<tr>
<td>Single</td>
<td>0.19</td>
<td>0.11</td>
</tr>
<tr>
<td>Coeff</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Commercial Area</td>
<td>0.14</td>
<td>0.12</td>
</tr>
<tr>
<td>Coeff</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Bar Density</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>Coeff</td>
<td>&lt;.0001</td>
<td>0.0011</td>
</tr>
<tr>
<td>Major Repairs</td>
<td>-0.04</td>
<td>-0.01</td>
</tr>
<tr>
<td>Coeff</td>
<td>0.0107</td>
<td>0.3405</td>
</tr>
<tr>
<td>Spatial Lag</td>
<td></td>
<td>0.43</td>
</tr>
<tr>
<td>Coeff</td>
<td></td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

3.3 Effect of spatial model misspecification

As noted previously the spatial model used for any given dataset should reflect the underlying spatial pattern in the data. To illustrate the effect of misspecifying the structure of the spatial dependence, a spatial error model was fit to the property crime rate in Montréal (Table 3). The estimated regression coefficients for the two models were very similar. In both models the variable major repairs was not significant indicating it should be removed from the model. The most noteworthy difference between the two models was the significance of the proportion of neighbourhood residents who were members of a visible minority. This variable was significant in the spatial lag model but was not significant at the 5% level in the spatial error model. The reason for this discrepancy was that the spatial error model did not completely account for the spatial structure of the data. This was confirmed by a test for remaining spatial autocorrelation in the form of spatial lag dependence in the model residuals from the spatial error model. The value of the LM test was 5.59 (p=0.018) indicating that the spatial error model did not completely account for the spatial dependence in the data.

Table 3: The spatial lag and a spatial error model fit to the rate of property crime on the island of Montréal.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Spatial Lag</th>
<th>Spatial Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.98</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Coeff</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>Low Inc. Private</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>Coeff</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Visible Minority</td>
<td>-0.05</td>
<td>-0.04</td>
</tr>
<tr>
<td>Coeff</td>
<td>0.0040</td>
<td>0.0878</td>
</tr>
<tr>
<td>Single</td>
<td>0.11</td>
<td>0.19</td>
</tr>
<tr>
<td>Coeff</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Commercial Area</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>Coeff</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Bar Density</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Coeff</td>
<td>0.0011</td>
<td>0.0009</td>
</tr>
<tr>
<td>Major Repairs</td>
<td>-0.01</td>
<td>-0.02</td>
</tr>
<tr>
<td>Coeff</td>
<td>0.3405</td>
<td>0.2700</td>
</tr>
<tr>
<td>Spatial Lag</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>Coeff</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>Spatial Error</td>
<td></td>
<td>0.50</td>
</tr>
<tr>
<td>Coeff</td>
<td></td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

A comparison of the model diagnostics for the spatial lag and spatial error models further confirms that the spatial lag model was preferred for the property crime rate in Montréal. The AIC for the spatial lag model was 220.5 compared to 241.2 in the spatial error model. The spatial error model did perform better than the standard regression model, as indicated by the larger log-likelihood, but the remaining spatial autocorrelation in the residuals confirms that it was not appropriate for the Montréal property crime data.
4. CONCLUSION

Whenever the dataset being analysed is representative of a geographic location it is important to consider the spatial structure of the data in the analysis. As shown with the analysis of violent crime in Halifax, a different model was obtained for the two separate areas in the study. In both regions a standard linear regression model did account for the spatial structure of the data, but by allowing a different model in the two regions of the city, the models provided a better fit to the data and provided more information about the different factors influencing crime in the area. In the analysis of property crime on the island of Montréal, a spatial model was necessary to allow for accurate inference of the other model parameters. If a standard linear model had been used for this data it may have erroneously resulted in the recommendation of a policy to repair housing in at-risk neighbourhoods as a possible means of reducing crime. Although that may be advantageous to the people living in those regions, the spatial model did not show that neighbourhoods with a high proportion of houses in need of ‘major repair’ was one of the most important variables in predicting the rate of property crime. The spatial lag model instead illustrated that neighbourhoods with a high property crime rate were associated with neighbourhoods with a high percentage of low-income private households, people living alone, commercial zoning and number of bars per square kilometre and a low percentage of residents who are members of a visible minority. Finally, the difference between the spatial lag and spatial error models for the Montréal property crime data illustrated the importance of ensuring the form of the spatial structure in the data is accurately represented in the model. This can be easily detecting using LM tests applied to the residuals of a standard linear model.

REFERENCES


