The role of spatial representation in the development of a LUR model for Ottawa, Canada

Marie-Pierre Parenteau · Michael Charles Sawada

Abstract A land use regression (LUR) model for the mapping of NO_2 concentrations in Ottawa, Canada was created based on data from 29 passive air quality samplers from the City of Ottawa's National Capital Air Quality Mapping Project and two permanent stations. Model sensitivity was assessed against three spatial representations of population: population at the dissemination area level, population at the dissemination block level and a dasymetrically derived population representation. A spatial database with land use, roads, population, zoning, greenspaces and elevation was created. Polycategorical zoning data were used in dasymetric mapping to spatially focus population data derived from the dissemination blocks to a sub-block level for comparison purposes. Dasymetric population mapping provided no significant LUR model improvement in explained variance when compared to block level population; however, both the former were significantly better than the dissemination area level population representations. However, where block level population is not available or too costly to acquire, our method using polycategorical zoning data provides a viable alternative in LUR modelling endeavours.

Keywords GIS, LUR, Dasymetric mapping, Scale LUR, Land use regression · NO_2, Nitrogen dioxide · DA, Dissemination area · DISB, Dissemination block · CMA, Census Metropolitan Area · GIS, Geographical Information System · PDF, Population density fraction · AR, Area ratio ·

TF, Total fraction · RMSE, Root-mean-square error · VIF, Variation inflation factor · CI, Condition index · LOOCV, Leave-one-out cross-validation · MAE, Mean absolute error

M.-P. Parenteau · M. C. Sawada (✉)
Department of Geography,
Laboratory for Applied Geomatics and GIS Science (LAGGISS),
University of Ottawa,
Simard Hall, 60 University Pvt., Room 047,
Ottawa, Ontario K1N 6N5, Canada
e-mail: msawada@uottawa.ca

Introduction

Modelling chronic air pollution exposure to constituents like nitrogen dioxide (NO_2) at an intra-urban scale is fundamental for health planning and intervention within cities. The land use regression (LUR) model was first introduced in 1997 by a team of European researchers (Briggs et al. 1997), but it was not until 2005 that a first attempt at using this methodology in North America was published (Gilbert et al. 2005). Since then, LUR models have been developed for only a limited number of large centres in Canada (Gilbert et al. 2005; Henderson et al. 2007; Jerrett et al. 2007; Marshall et al. 2008; Su et al. 2008; Wheeler et al. 2008; Sahsuvaroglu et al. 2009; Poplawski et al. 2009). The modelling of air quality based on LUR requires accurate data on a number of human and environmental factors such as land use, street networks, location of greenspace and population distribution. Each of these variables can, and has been, integrated within LUR models using a wide variety of spatial representations and spatial scales. As the number of articles published that employ land use regression models has been increasing, a research agenda that focuses on the role of spatial representation and scale in the LUR model performance is warranted.

In order to improve LUR model development and choice, our principal objective is to examine the role of spatial representation of the LUR-independent variables used to model atmospheric NO_2 concentrations. A second

objective of this research aims at developing a reasonably accurate LUR model for Ottawa, Canada, which has not been attempted before. Ottawa is often considered a unique city because of its small manufacturing base and large government and technology sector activities. So, developing a LUR for this city is challenging considering the size of Ottawa and the low industrial activity found within its boundaries (Wentz et al. 2002; Jerrett et al. 2007). The use of a population independent predictor is the key element that many published European and North American LUR models have in common. There are a numerous ways in which population has been represented in LUR modelling efforts and these commonly include the number of dwellings per unit area, the population count per unit area and population density (Henderson et al. 2007; Ryan et al. 2008; Beelen et al. 2009). For the most part, operationalising the population variable is achieved by using available census data at different geographic levels. The use of multiple population representations found in the learned literature begs the question: how robust are LUR results to the use of different population representations at different spatial scales? Since the question of spatial representation is a fundamental consideration, the modifiable areal unit problem (Openshaw 1984) in population representation is a concern (Andresen and Brantingham 2008). With regard to our present LUR undertaking, we have tried to limit our scope by looking more specifically at the spatial representation of the population variable and its ensuing effects on LUR model output performance. We thus address the specific question: How robust are LUR models to different population representations as independent variables?

To our knowledge, no research has yet studied the role of spatial representation in the development of a LUR model, more specifically for the population variable. With this research, we propose to address this issue for the first time by developing regression models based on three different representations of population from the Canadian Census of Population: the population count at the dissemination area (DA) level, the population count at the dissemination block (DISB) level and the population count at a sub-dissemination block level using dasymetric mapping (DASYM).

The need for data integration arises when one wants to use data collected under a different spatial division (e.g. non-census tract or non-dissemination area level but a finer or custom geographic boundary set) than the one used by the census (Fisher and Langford 1996) or when wanting to understand or intervene at a scale that is finer than that collected by the census. This would be the case when the goal is to examine natural socioeconomic processes that are indifferent to the imposed non-physical boundaries. As such, spatial units are often incompatible with respect to the required or intended needs of the researcher and so areal

interpolation techniques are required (Langford 2007). Solving this problem of incompatibility requires the assignment of one aggregated dataset to another incompatible dataset using various available spatial algorithms (Sadahiro 1999; Mennis 2003; Reibel and Bufalino 2005; Reibel and Agrawal 2007; Langford 2007). The approaches developed to solve the problem of incompatible spatial units have the capability of generating a more precise map of population distribution or many other census derived variables. Dasymetric mapping, which can also be pycnophylactic (Tobler 1979) in nature, is the spatial interpolation method used in this research. It is a method that is based on the integration of ancillary spatial data. Ancillary datasets like roads, greenspace, water, land use and cadastral data can help to define both where people could live as well as where they cannot live within a predefined area. As such, a dasymetric approach provides a method by which the original dataset representing, for example, population counts in census tracts can be disaggregated and redistributed to a finer spatial scale. The use of this approach also corresponds to the last goal of this research, which is to work toward the development of a standardised methodology for dasymetric mapping (Langford and Higgs 2006).

We use dasymetric mapping in the context of LUR, but it could also have numerous other applications where population data at fine spatial resolutions are required. For example, the availability of an accurate representation of the population distribution for governance (*the governance of oneself and of others*) is very important for the task of administrating services (Crampton 2004). It can be argued that an accurate map of human population is essential to municipal planning; even more so for public health planning and healthcare provisions (Hay et al. 2005). Global disaster management for the developing world has given rise to projects like LandScan (Bhaduri et al. 2007) and others that are producing dasymetric gridded global population estimates at fine spatial resolutions that compare well with known population distributions in the developed world (Sutton et al. 2003; Sabesan et al. 2007; Patterson et al. 2009). It is clear that accurate data on the spatial distribution of population is fundamental to a number of endeavours (Liu et al. 2008), but few studies have focused on the question of spatial representation in terms of population distribution and its impact on policies. One of the few examples of research on the subject is the work of Langford and Higgs (2006) who investigated the influence of alternative spatial population representations to the measure of potential access to primary healthcare services. The authors found that the modelling method for population impacted the results. The authors concluded that the use of dasymetric mapping consistently provides lower estimates of accessibility to healthcare, which in terms of policy and planning could have a significant impact. Research in the field of environmental

justice has also started addressing the question of using alternative population representations (Most et al. 2004; Brindley et al. 2005; Mohai and Saha 2006, 2007). Hence, this research will also contribute to the advancement of these other fields of studies.

In Europe and North America, LUR modelling efforts to map exposure to NO_2 have performed generally well with R^2 values varying from approximately 0.5 to 0.9. In general, Canadian research has yielded acceptable results, with R^2 values between 0.54 and 0.77. Even though the predictor variables have been generally the same for most studies, their specifications have been significantly different (Jerrett et al. 2007; Ryan and LeMasters 2007). Hence, not only it is very important to understand the sensitivity of the models to spatial representation in order to obtain consistency in the results but it is also very important as these exposure models are in most cases one of the first steps in the study of the relationship between exposure and health. Exposure models with an improved spatial resolution have been found to produce more robust associations (Sahsuvaroglu et al. 2009) when compared to health conditions. We expect dasymetric mapping will aid in obtaining better LUR models. Hence, better LUR models will contribute to the literature on the relationship between exposure and health. Still, the main objective of this research is to give a first approximation to the role of spatial representation in the development of LUR models and contribute to the research agenda on the role of spatial representation more generally.

Land use regression model

Land use regression models were developed for application in European cities, and research on the subject was first published by Briggs et al. (1997). Those authors were interested in exposure models that would allow the study of the relationship between health and air pollution at a local scale. Work by these authors also corresponds to the period when academia started to be interested in the spatial distribution of pollutants, not only in the context of inter-urban studies but also for intra-urban studies (Jerrett et al. 2005).

The general approach to modelling aims at predicting concentration (NO_2 or other pollutants) at an arbitrary location using observed concentrations at selected sampling sites together with a number of spatial explanatory variables characterising the environment in the proximity of those air quality sampling sites (Jerrett et al. 2005; Ryan and LeMasters 2007; Hoek et al. 2008). LUR is, hence, based on two important principals: "1) environmental conditions for the variable of interest can be estimated from a small number of readily measurable predictor variables 2) the relationship between the target variable and these predictors

can be reliably assessed on the basis of a small sample survey or training area" (Briggs et al. 2000). Among the main advantages of this approach are the facts that it can easily be adapted to the environment in which the study is taking place and that the development of such a model is less costly than increasing the observed air quality sampling network (Jerrett et al. 2005).

Over the last few years, a number of North American studies have been published (Kanaroglou et al. 2005; Gilbert et al. 2005; Jerrett et al. 2007; Wheeler et al. 2008). The results of these studies confirm the potential of LUR (Jerrett et al. 2007). An element that makes the development of a LUR model for Ottawa, Canada challenging is the reality that most published studies took place in large urban centres (Hoek et al. 2008). The City of Ottawa, part of Ottawa-Gatieau census metropolitan area (CMA; Ontario part) may be one of the largest cities in Canada with a population count of 812,129 in 2006 (Statistics Canada 2007), but it cannot be compared to large European and American cities in terms of population, and particularly, as mentioned, Ottawa lacks significant industrial activities.

In a literature review looking at the published LUR models, Hoek et al. (2008) found that a number of explanatory variables were more frequently used then others. Variables for population, traffic, land use, altitude, meteorology and location tended to be included in the LUR models developed. In addition, Hoek et al. (2008) found that problems frequently associated with spatial data, such as completeness and precision, had not been addressed in an appropriate manner within LUR methodologies. In particular, the role of spatial representation of the explanatory spatial variables in LUR modelling requires more attention.

Areal interpolation

When data are compiled according to different geographic boundaries, one is faced with a problem of geographic mismatch. Areal interpolation methods have been developed as a solution to this problem. Different methods of areal interpolation are available, the majority are based on a spatial overlay algorithm and each method is based on a number of different assumptions. Areal interpolation methods are generally classified into two broad categories: (1) techniques not using ancillary data; (2) techniques using ancillary data. In this article, dasymetric mapping, which falls in the category of techniques that use ancillary data, is considered. Dasymetric mapping can be implemented without a lot of additional data (Hay et al. 2005) and has the ability to provide results that are closer to the underlying spatial processes (Langford et al. 2008).

Dasymetric mapping

Dasymetric mapping is an areal interpolation method that takes advantage of ancillary data and so considers the heterogeneous distribution of a phenomenon through space. Poulsen and Kennedy (2004) describe dasymetric mapping as "a technique that involves estimating the distribution of aggregated data within the units of analysis, by adding additional information that provides insights on how these data are potentially distributed". Through the use of ancillary data, dasymetric mapping is capable of creating subzones with a higher level of homogeneity than the initial source zones and so more accurately represents the underlying geographic patterns of population when compared to the finest available enumeration areas (Fisher and Langford 1996; Mennis 2003; Holt et al. 2004; Langford and Higgs 2006; Fiedler et al. 2006; Mennis and Hultgren 2006). To date, this approach has relied in most cases on classified multi-spectral satellite imagery as a source of ancillary data for the identification of non-residential discontinuities such as parks and cemeteries in the urban environment (Reibel and Bufalino 2005; Langford and Higgs 2006; Langford 2007; Langford et al. 2008). The general reliance on remotely sensed imagery may have been an obstacle to the implementation of this areal interpolation method as it requires expertise in the field of remote sensing (Langford 2007) and the cost of image acquisition. An alternative to the use of satellite imagery, and raster data more generally, has been proposed. A non-image-based technique uses street network data (Reibel and Bufalino 2005) and has been successfully implemented in a few studies. In its simplest form, dasymetric mapping distinguishes between inhabited and uninhabited areas, which is called binary dasymetric mapping (Langford and Higgs 2006; Mennis and Hultgren 2006). It is the traditional approach to dasymetric mapping and has proven to perform better than area-weighted interpolation or other methods of areal interpolation that do not exploit ancillary data. The second type of dasymetric mapping is a polycategorical approach where three or more population density (land use) classes or categories are defined and is considered a more advanced and complex method to implement (Langford 2007; Eicher and Brewer 2001).

For the modelling of air quality, dasymetric mapping has been used only in research by Beelen et al. (2009). That study is quite different from the one presented here as Beelen et al. (2009) were interested in the possibility of modelling air quality over a large geographical area, namely the European Union using the LUR. They used dasymetric mapping to refine population distribution from NUTS-5 level using CORINE land cover and light emissions data. Conceptually, our work resembles the research of Beelen et al. (2009), but our goal is not to

determine how well the LUR can perform at a large scale using data at a coarse spatial resolution but rather how using different spatial representations, all at a fairly high resolution, will have an impact on the performance of the LUR.

Sources of ancillary data

As mentioned, most implementation cases of dasymetric mapping have been based on the use of satellite imagery as a source of ancillary data. A study by Fisher and Langford (1996) showed that population estimates by dasymetric mapping based on Landsat imagery are robust. Given the wide availability of free Landsat imagery, their approach is feasible if the modeller has the remote sensing expertise and the available imagery is sufficient. However, in their research, the authors found that dasymetric mapping still outperforms other areal interpolation techniques in those instances where low image classification accuracy is involved. In fact, they found that it takes classification accuracy as low as 60% for other areal interpolation methods to perform better. Where there is a lack of ancillary data, remote sensing approach is necessary and will provide the best estimate.

An alternative to the use of raster data for dasymetric mapping is the use of street network data (Xie 1995), making this approach more accessible to analysts who may not have the raster GIS skills (Reibel and Bufalino 2005). This use of the vector data is based on the assumption that population distribution is closely linked to the configuration of roads across the landscape (Hawley and Moellering 2005). Results of research show that the use of the street network data, which is a simple solution to the problem in comparison to the use of satellite imagery, provides better results than areal weighted interpolation (Xie 1995; Reibel and Bufalino 2005). Since areal weighting redistributes population based solely on the intersection of the target zone and the source zone, a homogeneous distribution of the population is assumed. The methodology implemented in this research takes into account both this aspect of the problem but adds another component that allows us to account for the differences in population density of the target zones.

Methods

Study design

We developed a LUR model for the mapping of NO_2 concentrations in Ottawa, Canada. Our focus is on the role of different spatial representations on model performance. Air pollution data were obtained from the City of Ottawa

through the National Capital Air Quality Mapping Project. In order to develop the LUR model, a spatial database that contained information on land use, roads, population, zoning, greenspaces and elevation was created. This database was also used for the mapping of population using the dasymetric mapping approach and will be discussed in the following sections.

Data

The National Capital Air Quality Mapping Project was launched in the fall of 2007 by the City of Ottawa with the help of Environment Canada and Health Canada. Through this project, 30 Ogawa passive samplers were installed throughout the City of Ottawa on three different occasions, each time for a period of 2 weeks (Fig. 1). Unfortunately, due to logistical problems, only the data collected from 29 samplers over the period of 29 September 2007–13 October 2007 could be used in the development of the LUR model. The monitor removed from the analysis (AQMP-09) was located in Gatineau, Quebec. The information collected at this location would have had a minor impact on the variation of the NO_2 concentrations (output of LUR model) in Ottawa. The monitor was located in a quiet residential area, very similar to the environmental setting of AQMP-17 and AQMP-26. As a consequence, removing AQMP-09 did not translate into the loss of key information.

The samplers measured the concentration in NO_2, O_3 and SO_2 at each location. Data collected at two permanent stations were also added to the dataset. Even though data on O_3 and SO_2 were available, only the NO_2 modelling is part of this ongoing research as it is an essential marker for traffic-related air pollution and strongly correlated with other common air-quality indices. This pollutant has been used as a proxy to evaluate exposure to traffic emissions (Jerrett et al. 2007).

Spatial data from Statistics Canada are made available to Canadian universities through the Data Liberation Initiative (Statistics Canada 2009). For this particular reason, the 2008 street network file from Statistics Canada was used as it represents the 2007 streets in Ottawa. The 2006 dissemination block spatial boundary file from Statistics Canada was also used as it corresponds to the most recent Census of Population in Canada. The dissemination block is a spatial unit defined on all sides by roads and/or boundaries and is the smallest geographic area for which population and dwelling counts are publicly reported (Statistics Canada 2007). Another important dataset used in the current research is the zoning data obtained from the City of Ottawa. The original dataset contained 39 zoning classes or categories which were reclassified into 23 classes according to the permitted uses for each of the main zoning types as described in the City of Ottawa Zoning By-law 2008-250 Consolidation (Table 1).

The zoning data eliminates the need to use remote sensing data, which in many cases is an obstacle to the implementation of dasymetric mapping (Langford 2007; Reibel and Agrawal 2007). Zoning data provides information that allows the

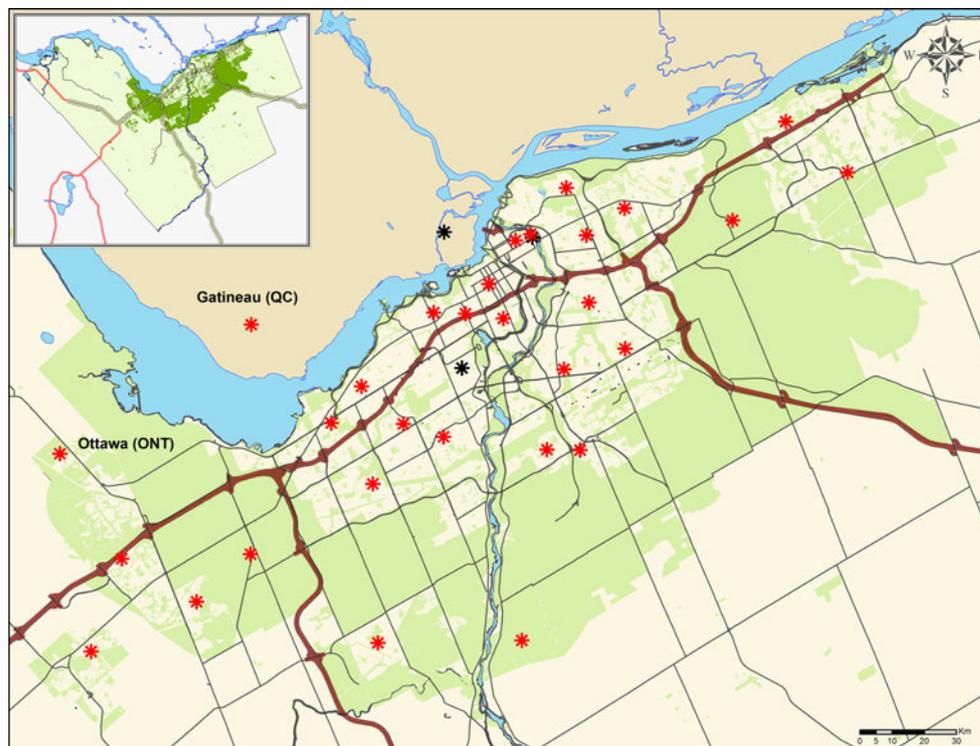Fig. 1 Distribution of the passive samplers (red) and permanent stations (black) in Ottawa

Table 1 Aggregation of the zoning classes according to the City of Ottawa Zoning By-law 2008-250 Consolidation

Original zoning type	New zoning type	New definition
Agricultural zone	AG	Agricultural zone
Arterial mainstreet zone	AM	Arterial mainstreet zone
Development reserve zone	DR	Development reserve zone
Environment protection zone	EP	Environment protection zone
General mixed use zone	GM	General mixed use zone
Minor institutional zone	NoPop	No population
Major institutional zone	NoPop	No population
General industrial zone	IG	General industrial zone
Heavy industrial zone	NoPop	No population
Light industrial zone	IL	Light industrial zone
Business park industrial Zone	IP	Business park industrial zone
Community leisure facility zone	NoPop	No population
Major leisure facility zone	NoPop	No population
Central experimental farm zone	NoPop	No population
Local commercial zone	R3	Local commercial zone
Mixed use centre zone	MC	Mixed use centre zone
Mixed use downtown zone	MD	Mixed use downtown zone
Mineral extraction zone	ME	Mineral extraction zone/mineral aggregate reserve zone
Mineral aggregate reserve zone	ME	Mineral extraction zone/mineral aggregate reserve zone
Parks and open space zone	O1	Parks and open space zone
Residential first density zone	R1	Residential first density zone
Residential second density zone	R2	Residential second density zone
Residential third density zone	R3	Residential third density zone
Residential fourth density zone	R4	Residential fourth density zone
Residential fifth density zone	R5	Residential fifth density zone
Rural commercial zone	NoPop	No population
Rural general industrial zone	NoPop	No population
Rural heavy industrial zone	NoPop	No population
Rural institutional zone	NoPop	No population
Mobile home park zone	RM	Mobile home park zone
Rural residential zone	RR	Rural residential zone

Table 1 (continued)

Original zoning type	New zoning type	New definition
Rural countryside zone	RU	Rural countryside zone
Air transportation facility zone	NoPop	No population
Ground transportation facility zone	NoPop	No population
Traditional main street zone	TM	Traditional main street zone
Village mixed use zone	VillRes	Village residential
Village residential first density zone	VillRes	Village residential
Village residential second density zone	VillRes	Village residential
Village residential third density zone	VillRes	Village residential

identification of uninhabited areas as well as detailed information on the different levels of population density throughout the region. We attributed a population count value of zero to zoning types where residential habitation use is not allowed and the other zoning types were set aside for the evaluation of the population density fraction (PDF). The calculation of the PDF is one of the main steps of dasymetric mapping.

Dasymetric mapping

Dasymetric mapping was undertaken at three scales, first at the scale of the dissemination block, second using the zoning data and third, the fusion of both datasets provided us with the necessary information to further refine the spatial resolution of the Census population data.

Because the zoning data were reclassified into 23 classes of land use, the dasymetric approach used in this research is polycategorical. This approach should represent an improvement in comparison to the traditional binary approach where information on the different levels of population densities within a census reporting zone are not accounted for (Langford and Higgs 2006). Our methodology also facilitates the implementation of dasymetric mapping, requiring no data or knowledge of remote sensing. This approach also has the advantage of overcoming the misclassification problem that can occur between residential apartment buildings and commercial facilities (Langford 2007).

The methodology used in this research is based on the work of Mennis (2003). Mennis' approach has four main steps: (1) calculation of the PDF; (2) calculation of the area ratio (AR); (3) calculation of the TF, and (4) the mapping of population (or other variable). The main difficulty with the

implementation of the polycategorical approach is the identification/calculation of the different relative ratios of the PDF values that are assigned to each population density class/category (land use class). Our approach to overcoming that issue is accomplished using selective sampling within the dataset (for more details, see Mennis 2003). The Mennis' method was selected for its simplicity, as one of our goals is to define a methodology that can be implemented by local governments without the need for considerable expertise. The downside of this approach is that it requires that some source zones are totally nested into one single population density class in order to be able to calculate the relative ratios of population density values (Mennis 2003; Langford and Higgs 2006; Reibel and Agrawal 2007). For example, in this study, the source zones were the dissemination blocks. Hence, it was necessary to have at least several blocks that were completely contained within one single land use class for the calculation of the PDF. This approach was first implemented by Mennis (2003); alternative approaches are found in Eicher and Brewer (2001) and Yuan et al. (1997). These may not have the problem associated with selective sampling but present other issues such as vulnerability to outliers and incorrect sampling in the case of the centroid method (Mennis and Hultgren 2006).

The use of dasymetric mapping allowed us to refine the census geography, defining a smaller geographic scale than the dissemination block. Considering that population data are not disseminated at a lower geographic level than the dissemination block, our methodology represents a solution for the production of a population map at a sub-census unit level. For example, with the use of dasymetric mapping, we were able to split a dissemination block into a number of parts (Fig. 2). Some of these parts, for example, do not have residential as a permitted land use and so they were attributed a population value of zero (see Table 1) and the total population count was reassigned within the larger

census zone to the new sub-units according to the PDF and the AR.

LUR

The measured concentrations of NO_2 at the sampling locations (dependent variable) were used in a multiple least-squares regression model (Ryan and LeMasters 2007). Multiple explanatory variables were identified based on the literature, with an emphasis on the representation of the population variable. As such, population variables were created based on the population data at the DA, the DISB and the refined dissemination block using DASYM. Buffers ranging from 50 to 2,000 m at an interval of 50 m were created for each of the sampling sites. At each interval, each explanatory variable was compared with NO_2 concentrations (40 comparisons per covariate). The use of the series of buffers provided a means by which to determine an optimised distance for each spatial covariate (Su et al. 2008). As NO_2 concentrations are known to fluctuate over distances as small as 50 m, the use of multiple buffers allowed us to identify the distance at which the measured correlation between the pollutant (mean NO_2 concentrations for the 2-week period) and the different covariates was the strongest. This preliminary analysis used Pearson's r coefficient of correlation and bivariate regressions to look at the absolute strength of the variables and possible deviations from linear relationships with NO_2 by examination of residual plots (Poplawski et al. 2009). This analysis provided us with initial information on the relationship between the variables and measured NO_2 concentrations as well as a way to confirm the anticipated direction of the effect (positive or negative) (Henderson et al. 2007). For each variable, the distance of maximum correlation was used to define the explanatory variable used in the model building exercise. Because we were undertaking multiple tests ($n=40$), for each variable, we interpreted statistical significance at a Bonferroni corrected level of $p=0.05/40$ (Cooper 1968).
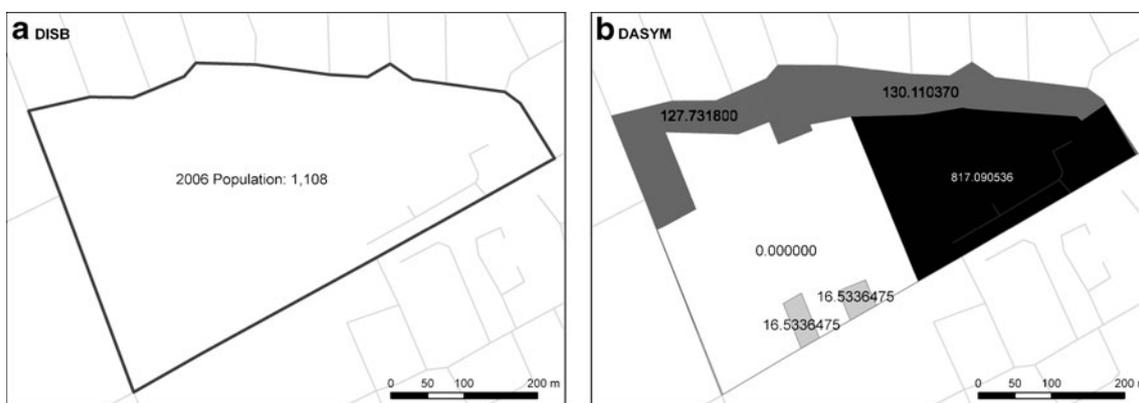

Fig. 2 Example of dasymetric mapping using zoning data. The population is redistributed (b) within census zone (DISB) shown in (a) as some of the zoning types within the selected DISB does not allow residential use

All three representations of the population variable had the strongest correlation with NO$_2$ at a distance of 1,750 m, showing coherence. Pearson's r for the correlation between the measured NO$_2$ concentration and population at the DA level only has a value of 0.28719 and is not statistically significant. The population at the DISB level and DASYM has similar Pearson's r values for the same relationship, with the dasymetric mapping having a slightly higher level (0.78968 versus 0.78893). However, NO$_2$ and DISB and DASYM are statistically significant as indicated in Table 2 but not significantly different from each other.

Since this research is concerned with the role of spatial representation, three different LUR models were developed. The first model included a population variable derived from DASYM, the second model's population variable was derived by the use of the data at the geographic level of the DISB and the third model had population derived from the DA level. Stepwise selection was the method used to develop the different regression models (Jerrett et al. 2007; Su et al. 2008). Variables in the model had to be statistically significant at $p=0.05$ and models were evaluated based on the R^2 values, RMSE, VIF eigenvalues and the CI. The methodology used for the development of all the regression models was the same. For example, to develop the model using population at the DA level, DISB population and DASYM population representations were excluded from the list of covariates. All models were tested for spatial autocorrelation in the residuals using Moran's I statistic (Bailey and Gatrell 1995), and no significant spatial autocorrelation was present.

The model containing the DASYM population variable (model 1) provided an R^2 of 0.8055 (Table 3) while the model using the DISB variable provided a very similar value of 0.8038 (model 2). Using the population data at the geographic level of the DA, model 3 did perform poorly, yielding a result of R^2 of 0.6962, much lower than models 1 and 2. Study of the correlation between the variables and NO$_2$ concentration as well as the development and evaluation of the regression models was achieved in the environment of SAS Enterprise (version 3.0).

The two models using the finest levels of aggregation as spatial representations of the population variables are very similar in terms of the variables included in the models, parameter estimates and statistics to evaluate the perfor-

Table 2 Correlation between NO$_2$ concentration and the different population representations

Spatial representation	Pearson's r	Prob>|r|
DA	0.28719	0.1141
DISB	0.78893	<0.0001
DASYM	0.78968	<0.0001

Table 3 Definition of the LUR models using different population representations

Variable	Parameter estimate	Pr>t	VIF
Model 1 (R^2=0.8055)			
Intercept	3.66507	0.0027	0
Maj_Rds_250m	0.00041824	0.0116	1.47569
PopDasym_1750m	0.00010586	0.001	2.24357
Greenspace_700m	−0.00000216	0.0229	2.39553
Industrial_250m	0.00005783	0.0398	1.28638
Model 2 (R^2=0.8038)			
Intercept	3.49419	0.0058	0
Maj_Rds_250m	0.00041876	0.0119	1.47878
PopDISB_1750m	0.00010492	0.0011	2.35903
Greenspace_700m	−0.00000206	0.03046	2.53099
Industrial_250m	0.00002917	0.037	1.29209
Model 3 (R^2=0.6962)			
Intercept	7.18141	<0.0001	0
Maj_Rds_250m	0.00065638	0.0009	1.2866
PopDA_1750m	0.00001609	0.2665	1.26329
Greenspace_700m	−0.00000488	<0.0001	1.02323

Probability and variation inflation factor values associated with each of the variables

Pr Probability, VIF variation inflation factor, t Student's t distribution

mance of the models. The model making use of spatial representation based on the data at the DA level had a different composition in terms of covariates, with the Industrial_250m variable being excluded from the model. Most importantly, the population variable included in the DA model, PopDA_1750m, was not statistically significant at 95% level.

As mentioned earlier, all three model residuals were tested for spatial autocorrelation, and none showed significant spatial dependency. As such, we are not concerned that spatial autocorrelation in the residuals will affect the size and significance of risk estimates from this analysis as others have found (Jerrett and Finkelstein 2005). Graphical representation of the residuals does not indicate presence of trends in any of the three models (Fig. 3). Once again, plots of the residuals of models 1 and 2 (Fig. 3a and 3b respectively) are very similar, but the residual plot of model 3 (Fig. 3c) is clearly different.

Model validation

Validation of the LUR model was achieved using leave-one-out cross-validation (LOOCV) (Hoek et al. 2008; Mukerjee et al. 2009). Using this methodology, the regression model was re-estimated 31 times, each time leaving out one observation ($n-1$). The root-mean-square error and the

Fig. 3 Regression (*left*) and residuals (*right*) of the three different models: **a** PopDASYM **b** PopDISB **c** PopDA. Predicted values of NO$_2$ level in ppb vs. measured NO$_2$ level in ppb (*left*). Predicted NO$_2$ level in ppb vs the residual in ppb (*right*)
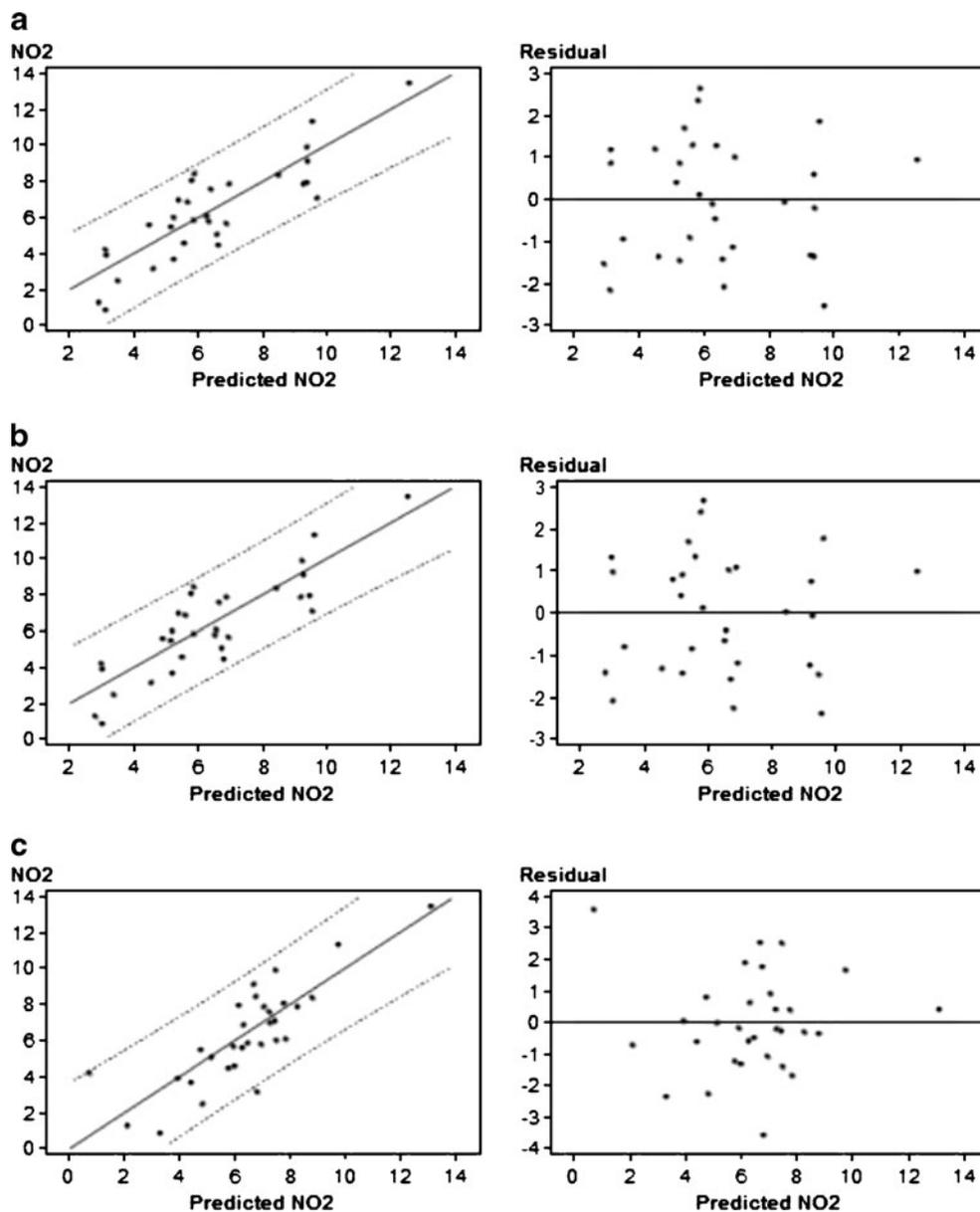
mean absolute error were computed in order to assess the performance of the model. The results (RMSE=1.05 ppb and MAE=0.86 ppb) indicate that the NO$_2$ values predicted by the model are reliable.

Discussion

Our results indicate that the accuracy of the land use regression model is affected by the use of different spatial representations of the population variable. The mapping of population at a sub-block geographic level using the dasymetric approach showed no significant differences over using population at the dissemination block. On the other hand, both of these models outperformed the model

based on dissemination area population data. As such, LUR model performance increases with mapping population distribution at a lower geographic level than the dissemination area. In addition, with the more detailed population, estimates of other socioeconomic and demographic covariates can potentially be better focused spatially to where the population resides.

With an R^2 of 0.8055, 0.8038 and 0.6962, all three LUR models have explained variances within the range found in the literature (Hoek et al. 2008). These values also demonstrate that it is possible to use the LUR model in a smaller city where the air quality is often considered generally good. This finding for Ottawa is important as it enlarges the applicability of the model to a number of additional centres, especially in Canada—where the number

of large urban centres is limited. The NO$_2$ surface created through the implementation of the LUR model corroborates a priori knowledge of the air quality in Ottawa with pockets of higher NO$_2$ concentration located in areas known to have air quality problems. The variables included in the LUR model for Ottawa (length of major roads, population count, area of greenspace and industrial land use) are generally the same as those included in similar models developed for other Canadian cities. The size of the buffers is also within the same range as those found within the literature on Canadian LUR models (Henderson et al. 2007; Jerrett et al. 2007; Marshall et al. 2008; Wheeler et al. 2008; Sahsuvaroglu et al. 2009). The main difference we found when building the LUR model for Ottawa was the importance of the variable "Greenspace". The only other model with a similar variable was published by Sahsuvaroglu et al. (2009) where a LUR model for Hamilton, Ontario is developed using a variable "Open land use". The presence of greenspaces is an important feature of the landscape in Ottawa, explaining why it is not found in most other LUR models developed for Canadian cities.

The City of Ottawa has a unique landscape created by the existence of a large corridor of greenspace that represents the boundary between the urban and the suburban part of the city (Fig. 4). Included in the original planning of the city in 1950 by Jacques Greber, a French planner, the purpose of the 'Greenbelt' was to prevent urban sprawl and protect the rural land surrounding the city (NCC 2009). Due to population

growth, residential and commercial developments were constructed beyond the greenbelt. The location of the greenbelt can easily be identified in the LUR NO$_2$ concentration surface map; it corresponds to the large linear area where the concentration of NO$_2$ is considerably lower than the surroundings. Hence, the existence of the greenbelt potentially has a strong effect on residents' health.

We believe the combination of a number of factors is behind the fact that the use of dasymetric mapping did not significantly improve the results of NO$_2$ modelling. One first factor is potentially found in the level of heterogeneity of the DISBs within Ottawa. In general, population distribution within a particular dissemination block is highly homogeneous. As a consequence, refinement of population distribution is not significant enough to considerably improve the results of the modelling.

Also, the scale at which NO$_2$ concentration varies is of importance. Refinement of the map of population distribution at a geographic level lower than the dissemination block may be much finer a scale than at which we can observe NO$_2$ variations. For this reason, spatial representation does not appear to be playing an important role when comparing population counts at the block level and dasymetric population. On the contrary, the role of spatial representation is significant when comparing LUR with block or sub-block level data and DA level data.

Lastly, the poor contribution of dasymetric mapping to the LUR may be due to the fact that population count (all

Fig. 4 LUR for NO$_2$ concentrations (ppb) in Ottawa
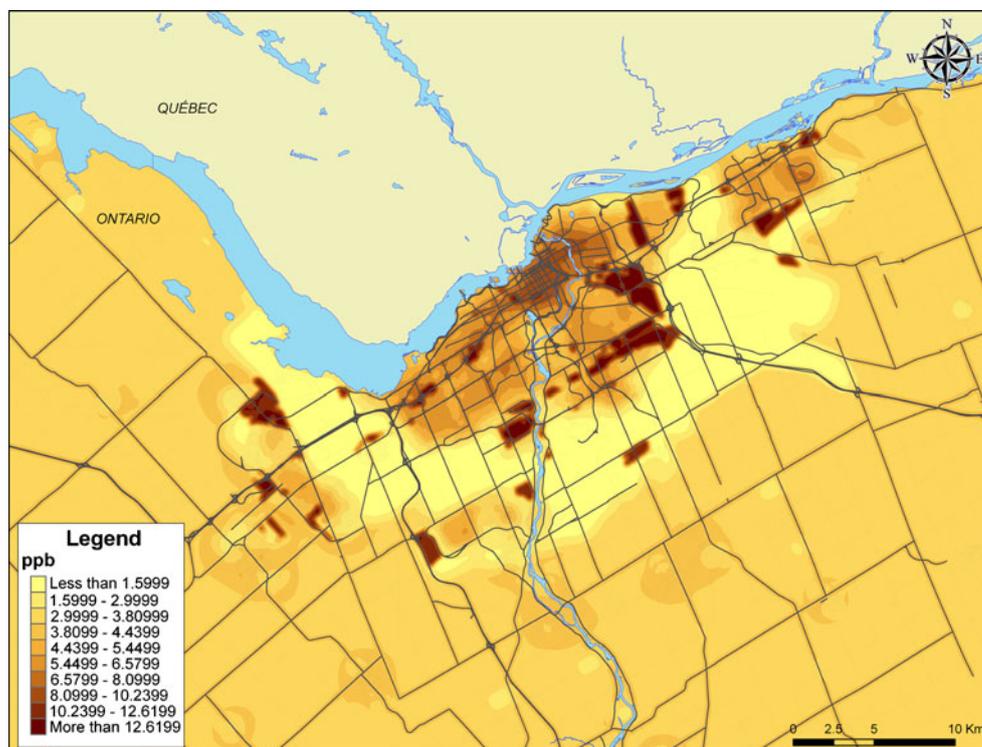
representations) is highly correlated to NO_2 concentration at a distance of 1,750 m from the sampling sites. In a study on the evaluation of population at risk using dasymetric mapping, Higgs and Langford (2009) conclude that "as [the] buffer size increases with respect to census tracts, partial intersections proportionately decrease and correspondingly the results [of different population representations] tend to converge". This suggests that if the strongest correlation of population with NO_2 would have been measured at a smaller distance, potentially the use of dasymetric mapping could have considerably improved the results. Such a hypothesis begs testing in this urban area by using data in different seasons for example.

Our results indicate that the use of dasymetric mapping to refine the population data at the sub-dissemination block level is possible. In terms of model improvement, the result is not significantly different from using block level population representations, and we could not recommend the use of this approach when population data are available at a fine geographic level (as the dissemination block). On the other hand, we perceive this methodological approach to be of great value not only to redistribute population aggregated at a higher geographic level but also for a number of other applications. In addition, our approach would be beneficial for local level governments that have access to local zoning data but who may not have access or who do not want the added expense of utilising block level population in Canada and elsewhere in urban areas where block level population may not be collected.

The main limitation of this study is the City of Ottawa's geography. The City of Ottawa is located on the Ontario side of the Ottawa River, representing only one side of the Ottawa-Gatineau CMA. The other side of the Ottawa River is the City of Gatineau, located in the nation of Québec. The data used in the modelling of the NO_2 with the LUR model were acquired from the City of Ottawa and do not cover the City of Gatineau. An airshed does not follow political boundaries, and it is realistic to assume that the integration of the Québec portion would have resulted in a more accurate model.

Another limitation of this research is the fact that the NO_2 data used were collected through a limited air sampler deployment. Considering that Ottawa is characterised by large variability in temperatures measured at different times of the year, model improvements could certainly be made with NO_2 measurements taken during different seasons. As stated earlier, seasonal variations in the spatial correlation of NO_2 with covariates might improve model predictions. However, Wheeler et al. (2008), in a study of the correlation between NO_2 measurements taken throughout different Canadian seasons with yearly averages, found that all seasons were highly correlated with the annual average. The problem would then be one of overestimation or

underestimation, depending on the season at which measurements were taken and would not affect the general spatial distribution of NO_2 concentrations.

Also, the location of the samplers did not include rural sites. As a large proportion of the territory of the City of Ottawa is rural, having sampling sites in these areas would have potentially improved the models beyond the urban core and in the suburbs. The location of the samplers may be more important than the number or samplers deployed (Ryan and LeMasters 2007). However, such was out of our control but should be considered in the future deployments by the city. Lastly, the importance of the temporal period of sampler deployment and resultant effects on NO_2 modelling in regard to the selection of covariates needs attention and were beyond the scope of this study.

Conclusion

As the land use regression model grows in popularity as a simple modelling approach, we believe that it is time that the research community focuses on the question of spatial representation and the impact on the mapping of pollutants. The secondary goal of this research was to develop a LUR model for Ottawa. Our results indicate that it is possible to develop such a model for a smaller urban centre where air quality is considered to be good.

Our results point to the fact that the output of LUR modelling for NO_2 mapping is affected by the use of different spatial representations of population. Where we had expected the LUR model using results of dasymetric mapping to outperform the one using block level population counts, we did find essentially identical results. This confirms the work of Higgs and Langford (2009) who found that the best results with dasymetric mapping arise at small buffer sizes. Consequently, we recommend the use of dasymetric mapping for LUR models if the strongest correlation between NO_2 and population is found in close proximity of the air quality sampling sites. Moreover, the availability of zoning datasets by municipalities can easily replace the need for block level population in cases where such data are not available or too costly to acquire. In general, the use of zoning data obtained in a vector format did allow the refinement of population distribution with the dasymetric mapping methodology. This also has the advantage of easy adoption by local governments for the creation of population maps at a sub-block level without the need to use remote sensing data.

References

Andresen MA, Brantingham PL (2008) Visualizing ambient population data within census boundaries: a dasymetric mapping procedure. Cartographica 43:267–275

Bailey TC, Gatrell AC (1995) Interactive spatial data analysis. Wiley, New York

Beelen R, Hoek G, Pebesma E, Vienneau D, de Hoogh K, Briggs DJ (2009) Mapping of background air pollution at a fine spatial scale across the European Union. Sci Total Environ 407:1852–1867

Bhaduri B, Bright E, Coleman P, Urban ML (2007) LandScan USA: a high-resolution geospatial and temporal modelling approach for population distribution and dynamics. GeoJournal 69(1–2):103–117

Briggs DJ, Collins S, Elliott P, Fisher P, Kingham S, Lebret E, Pryl K, Van Reeuwijk H, Smallbone K, Van Der Veen A (1997) Mapping urban air pollution using GIS: a regression-based approach. Int J geogr Inf Sci 11(7):699–718

Briggs DJ, de Hoogh C, Gulliver J, Willis J, Elliott P, Kingham S, Smallbone K (2000) A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. Sci Total Environ 253:151–167

Brindley P, Wise SM, Maheswaran R, Haining RP (2005) The effect of alternative representations of population location on the areal interpolation of air pollution exposure. Comput Environ Urban Syst 29:455–469

Cooper DW (1968) The significance level in multiple tests made simultaneously. Heredity 23:614–617

Crampton JW (2004) GIS and geographic governance: reconstructing the choropleth map. Cartographica 39(1):42–53

Eicher CL, Brewer CA (2001) Dasymetric mapping and areal interpolation: implementation and evaluation. Cartogr Geogr Inf Syst 28(2):125–138

Fiedler R, Schuurman N, Hyndman J (2006) Improving census-based socioeconomic GIS for public policy: recent immigrants, spatially concentrated poverty and housing need in Vancouver. ACME 4(1):145–171

Fisher PF, Langford M (1996) Modelling sensitivity to accuracy in classified imagery: a study of areal interpolation by dasymetric mapping. Prof Geogr 48(3):299–309

Gilbert NL, Goldberg MS, Beckerman B, Brook JR, Jerrett M (2005) Assessing spatial variability of ambient nitrogen dioxide in Montreal, Canada, with a land-use regression model. J Air Waste Manage Assoc 55:1059–1063

Hawley K, Moellering H (2005) A comparative analysis of areal interpolation method. Cartogr Geogr Inf Sci 32(4):411–423

Hay SI, Noor AM, Nelson A, Tatem AJ (2005) The accuracy of human population maps for public health application. Trop Med Int Health 10(1):1073–1086

Henderson SB, Beckerman B, Jerrett M, Brauer M (2007) Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. Environ Sci Technol 41(7):2422–2428

Higgs G, Langford M (2009) GIScience, environmental justice, & estimating populations at risk: the case of landfills in Wales. Appl Geogr 29:63–76

Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fisher P, Briggs D (2008) A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos Environ 42:7561–7578

Holt JB, Lo CP, Hodler TW (2004) Dasymetric estimation of population density and areal interpolation of census data. Cartogr Geogr Inf Sci 31(2):103–121

Jerrett M, Finkelstein M (2005) Geographies of risk in studies linking chronic air pollution exposure to health outcomes. J Toxicol Environ Health A 68:1207–1242

Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, Morrison J, Giovis C (2005) A review and evaluation of intraurban air pollution exposure models. J Expo Anal Environ Epidemiol 15:185–204

Jerrett M, Arain MA, Kanaroglou P, Beckerman B, Crouse D, Gilbert NL, Brook JR, Finkelstein N, Finkelstein MM (2007) Modelling the intraurban variability of ambient traffic pollution in Toronto, Canada. J Toxicol Environ Health Part 1(70):200–212

Kanaroglou PS, Jerrett M, Morrison J, Beckerman B, Arain MA, Gilbert NL, Brook JR (2005) Establishing an air pollution monitoring network for intra-urban population exposure assessment: a location-allocation approach. Atmos Environ 39:2399–2409

Langford M (2007) Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps. Comput Environ Urban Syst 31:19–32

Langford M, Higgs G (2006) Measuring potential access to primary healthcare services: the influence of alternative spatial representations of population. Prof Geogr 58(3):294–306

Langford M, Higgs G, Radcliff J, White S (2008) Urban population distribution models and service accessibility estimation. Comput Environ Urban Syst 32(1):66–80

Liu XH, Kyriadis PC, Goodchild MF (2008) Population-density estimation using regression and area-to-point residual kriging. Int J Geogr Inf Sci 22(4):431–447

Marshall JD, Nethery E, Brauer M (2008) Within-urban variability in ambient air pollution: comparison of estimation methods. Atmos Environ 42:1359–1369

Mennis J (2003) Generating surface models of population using dasymetric mapping. Prof Geogr 55(1):31–42

Mennis J, Hultgren T (2006) Intelligent dasymetric mapping and its application to areal interpolation. Cartogr Geogr Inf Sci 33 (3):179–194

Mohai P, Saha R (2006) Reassessing racial and socioeconomic disparities in environmental justice research. Demography 43 (2):383–399

Mohai P, Saha R (2007) Racial inequality in the distribution of hazardous water: a national-level reassessment. Soc Probl 54 (3):343–370

Most MT, Sengupta R, Burgener MA (2004) Spatial scale and population assignment choices in environmental justice analyses. Prof Geogr 56(4):574–586

Mukerjee S, Smith LA, Johnson MM, Neas LM, Stallings CA (2009) Spatial analysis and land use regression of VOCs and NO2 from school-based urban air monitoring in Detroit/Dearborn, USA. Sci Total Environ 407:4642–4651

National Capital Commission (2009) The National Capital Greenbelt; History and Culture http://www.canadascapital.gc.ca/bins/ncc_web_content_page.asp?cid=16297-16299-9735-113846-9742&lang= Accessed 20 January 2010

Openshaw S (1984) The modifiable areal unit problem. Concepts and techniques in modern geography 38, Geobook, Norwich

Patterson L, Urban M, Myers A, Bhaduri B, Bright E, Coleman P (2009) The effects of quality control on decreasing error propagation in the LandScan USA population distribution model: a case study of Philadelphia county. Trans GIS 13(2):215–228

Poplawski K, Gold T, Setton E, Allen R, Su J, Larson T, Henderson S, Brauer M, Hystad P, Lightowlers C, Keller P, Cohen M, Silva C, Buzzelli M (2009) Intercity transferability of land use regression models for estimating ambient concentrations of nitrogen dioxide. Jounal Expo Environ Epidemiol 19:107–117

Poulsen E, Kennedy LW (2004) Using dasymetric mapping for spatially aggregated crime data. J Quant Criminol 20(3):243–262

Reibel M, Agrawal A (2007) Areal interpolation of population counts using pre-classified land cover data. Popul Res Policy Rev 26:619–633

Reibel M, Bufalino ME (2005) Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. Environ Plann A 37:127–139

Ryan PH, LeMasters GK (2007) A review of land-use regression models for characterizing intraurban air pollution exposure. Inhal Toxicol 19(Suppl 1):127–133

Ryan PH, LeMasters GK, Levin L, Burkle J, Biswas P, Hu S, Grishpun S, Reponen T (2008) A Land-use regression model for estimating microenvironmental diesel exposure given multiple addresses from birth through childhood. Sci Total Environ 404:139–147

Sabesan A, Abercrombie K, Ganguly AR, Bhaduri B, Bright EA, Coleman PR (2007) Metrics for the comparative analysis of geospatial datasets with applications to high-resolution grid-based population data. GeoJournal 69(1–2):81–91

Sadahiro Y (1999) Accuracy of areal interpolation: a comparison of alternative methods. J Geogr Syst 1:323–346

Sahsuvaroglu T, Jerrett M, Sears MR, McConnell R, Finkelstein N, Arain A, Newbold B, Burnett R (2009) Spatial analysis of air pollution and childhood asthma in Hamilton, Canada: comparing exposure methods in sensitive subgroups. Environ Health 8(14)

Statistics Canada (2010) Community Profiles Ottawa. http://www12.statcan.gc.ca/census-recensement/2006/dp-pd/prof/92-591/details/page.cfm?Lang=E&Geo1=CD&Code1=3506&Geo2=PR&Code2=35&Data=Count&SearchText=ottawa&SearchType=Begins&SearchPR=01&B1=All&CustomREFCODE=10&LANG=E&GEO_LEVEL=102&TYPE=L, Accessed November 24th 2009

Statistics Canada (2009) Data Liberation Initiative (DLI). http://www.statcan.gc.ca/dli-ild/dli-idd-eng.htm, Accessed August 4th 2009

Su JG, Brauer M, Buzzelli M (2008) Estimating urban morphometry at the neighborhood scale for improvement in modelling long-term average air pollution concentrations. Atmos Environ 42:7884–7893

Sutton PC, Elvidge C, Obremski T (2003) Building and evaluating models to estimate ambient population density. Photogramm Eng Remote Sens 69(5):545–553

Tobler WR (1979) Smooth pycnophylactic interpolation for geographical regions. J Am Stat Assoc 74:519–530

Wentz E, Gober P, Balling RC, Day TA (2002) Spatial patterns and determinants of winter atmospheric carbon dioxide concentrations in an urban environment. Ann Assoc Am Geogr 92:15–28

Wheeler AJ, Smith-Doiron M, Xu X, Gilbert NL, Brook JR (2008) Intra-urban variability of air pollution in Windsor, Ontario—Measurement and modelling for human exposure assessment. Environ Res 106:7–16

Xie Y (1995) The overlaid network algorithms for areal interpolation problem. Comput Environ Urban Syst 19(4):287–306

Yuan Y, Smith RM, Limp WF (1997) Remodelling census population with spatial information from Landsat TM imagery. Comput Environ Urban Syst 21(3–4):245–258