Modeling spatial patterns of traffic emissions across 5570 municipal districts in Brazil

Weeberb J. Requia*, Henrique L. Roig, Petros Koutrakis, Matthew D. Adams

McMaster University, School of Geography and Earth Sciences, Hamilton, Ontario, Canada

A R T I C L E   I N F O

Article history:
Received 12 September 2016
Received in revised form
30 January 2017
Accepted 2 February 2017
Available online 8 February 2017

Keywords:
Air pollution
Traffic emissions
Spatial analysis
Spatial patterns

A B S T R A C T

Traffic emissions pose a major environmental threat to human health in many countries. Most studies of traffic emissions have applied spatial models to achieve more accurate exposure assessments. Spatial analysis is a technique that takes into account various geographic phenomena, providing data that can be communicated to potentially affected people and government agencies considering policies to reduce exposure and adverse health effects. In Brazil, only a few studies have evaluated air pollution and no studies have conducted a spatial analysis of traffic emissions across the entire country. Brazil is a large continental region with 200 million inhabitants, where traffic emissions present a critical health challenge. In this study, we used three geostatistical approaches (Getis-Ord Gi*, K means, and spatial regression) to assess the spatial patterns of traffic emissions for all 5570 municipal districts in Brazil. We identified five groups of municipal districts (spatial clusters) with distinct patterns of traffic emissions based on six variables: population, gross domestic product, urbanization rate, length of highways, human development index, and distance from the state capital. One group represents municipal districts in the Northeast and Northwest regions, which have lower income, traffic, and emissions. In contrast, the others groups represent regions with high income, traffic, and emissions. Finally, we found a significant association between emissions inventories and the six variables used to evaluate spatial clusters. Our results can help inform the design of targeted, cost effective air pollution control strategies to reduce the adverse health effects of traffic emissions in developing countries.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Air pollution poses a major environmental risk to human health in many countries, including the United States (Lamsal et al., 2013; Lee et al., 2012; Zou et al., 2014), China (Cao et al., 2011; Chan and Yao, 2008; Chowdhury et al., 2013), Mexico (Mugica-Álvarez et al., 2012), England (Beever et al., 2013; Lim et al., 2012), and Brazil (de Oliveira et al., 2012; Miraglia et al., 2013; Requía Júnior et al., 2015b; Silva et al., 2012). Several studies have shown that, among air pollution sources, traffic emissions are strongly associated with poor health outcomes in these countries (Bind et al., 2015; Chandran et al., 2013; Masri et al., 2015; Requía Júnior et al., 2015a; Song et al., 2012). For example, Bruegg et al. (2013) show that there is an increased risk for people who live near highways due to the concentration gradient profile. Karagulan et al. (2015) show that 25% of PM$_{2.5}$ in urban areas are emitted by traffic globally. Furthermore, of all emissions worldwide, transportation is responsible for 30% of NO$_x$, 14% CO$_2$, 54% CO, and 47% non-methane hydrocarbon - NMHC (Sokhi, 2011). Traffic emissions are significantly linked with premature mortality (Lelieveld et al., 2015), lung cancer (Fajersztajn et al., 2013), cardiorespiratory diseases (Paula Santos et al., 2005), and asthma (Amancio and Nascimento, 2012). According to Jacobson (2007), traffic emissions are anticipated to cause at least 10,000 premature deaths in the United States in 2020. Though a large number of studies on air pollution levels have been conducted in developed countries, between 1983 and 2013, only 5% of air pollution studies published worldwide were conducted in developing countries, such as Brazil (Fajersztajn et al., 2013). No studies have presented a spatial analysis of air pollution from traffic emissions across Brazil. Brazil is a large region ($8,515,767$ km$^2$) with 200 million inhabitants, where the traffic emissions pose a critical challenge to environmental health. The Brazilian fleet represents a non-negligible fraction of the global fleet; from 2000 to 2010, Brazil had the second largest fleet volume...
increase after China (WHO, 2012). The Brazilian Energy Agency projected that in 2050 the vehicle fleet in Brazil will increase more than 300%, totaling 130 million vehicles registered (EPE, 2014). Some local-level studies (at the city level) have shown that traffic emissions are associated with an increase in cardiorespiratory diseases in Brazil (Marcilio and Gouveia, 2007; Moura et al., 2012; Réqua Júnior and Roig, 2015; Ribeiro and Cardoso, 2003). Over the last 8 years (2008–2015), 1,239,665 people died due to cardiorespiratory diseases in Brazil and 18,091,947 were hospitalized (Datosus, 2015).

Many studies have measured traffic emissions in order to achieve more accurate assessments of exposure and its associated adverse health effects. Most of these studies have applied spatial models, including land use regression (Basagaña et al., 2012), interpolation (Deligiorgi and Philippopoulos, 2011), and cluster analysis (Zibert and Praznikar, 2012). Researchers have shown that spatial analysis allows a better understanding of geographic phenomena that impact emissions levels (Kurland and Gorr, 2012; Mitchell, 1999). Spatial analysis methods involve monitoring of conditions on the ground to calculate temporal changes, compare populations, and, especially, communicate actionable data to potentially affected people as well as governmental agencies and policymakers (Bateman et al., 2013). In this study, we assessed the spatial patterns of traffic emissions for all 5570 municipal districts in Brazil.

2. Materials and methods

2.1. Study design

The spatial domain of our study included all 5570 municipal districts in Brazil, which are organized into 26 states, plus a Federal District where the capital is located (Fig. 1).

This study was performed in three stages: i) consolidation of the traffic emissions data, ii) evaluation of spatial clusters of traffic emission patterns, and; iii) analysis of the spatial relationship between traffic emissions and socioeconomic variables.

2.2. Traffic emissions data

Traffic emissions were estimated by our research group in a previous analysis (Réqua et al., 2016). We estimated vehicle emissions inventories (tons per year) for each of the 5570 municipal districts in Brazil using a bottom-up approach. We predicted emissions for the period 2001–2012. We calculated emissions for carbon monoxide (CO), non-methane hydrocarbon (NMHC), methane (CH₄), nitrogen oxides (NOₓ), particulate matter (PM), and carbon dioxide (CO₂) from six vehicle categories: light vehicles (passenger cars), utility vehicles (for transport of passengers or goods), motorcycles, trucks (light, middle, and heavy duty), urban buses, and interstate buses. In Appendix, we present the characteristics for each vehicle type. We used the following equation to estimate emissions:

\[ E_{x, i, z} = \left( \frac{V_{f,x,i,y} \times a_y \times D_{t,y} \times E_{f,i,y}}{10^6} \right) \]  

(1)

where: \( E \) are the annual emissions in metric tons; \( V_f \) is the number of vehicles; \( a \) is fraction of fleet in use; \( D_t \) is the average distance travelled in km/year; \( E_f \) is the emission factor in grams of pollutant per unit distance – g/km; \( i \) is the municipal district (5,570); \( y \) is the vehicle type. Finally, to estimate the total emissions of a pollutant \( z \) for year \( i \), and for municipal district

\[ x \], we used the following equation:

\[ TAE_{x, z} = E_{L_v, x, i, z} + E_{U_v, x, i, z} + E_{M_o, x, i, z} + E_{T_r, x, i, z} + E_{U_b, x, i, z} + E_{I_b, x, i, z} \]  

(2)

where \( TAE \) are the emissions, in tons, for light vehicles, \( L_v \), utility vehicles \( U_v \), motorcycles, \( M_o \), trucks, \( T_r \); urban buses, \( U_b \), and interstate buses, \( I_b \).

2.3. Cluster mapping

We applied two geostatistical methods to identify spatial clusters of estimated emissions across the municipal districts of Brazil: Getis-Ord Gi* and K means. Studies on environmental health have used these methods (Austin et al., 2013; Réqua et al., 2016; Tsai et al., 2009).

Getis-Ord Gi* (Equation (3)) identifies spatial clusters of high values (hot spots, municipal districts clustered by high emissions) and low values (cold spots, municipal districts clustered by low emissions) within the context of neighboring features (Getis and Ord, 1992).

\[ G_{i}^{*} = \frac{\sum_{j=1}^{n} W_{ij} X_{j} - \overline{X} \sum_{j=1}^{n} W_{ij}}{\sqrt{\sum_{j=1}^{n} W_{ij}^{2} - \left( \overline{X} \sum_{j=1}^{n} W_{ij} \right)^{2}}} \]  

(3)

where \( G_{i}^{*} \) is the Getis-Ord value, which is a z-score. The larger the
A significant positive z-score is, the more intense the clustering for a hot spot. The smaller the significant negative z-score is, the more intense the clustering for a cold spot. \(X_j\) is the attribute value (emission inventories) for feature \(j\) (for a specific municipal district); \(W_{ij}\) is the spatial weight between feature \(i\) and \(j\), which each feature is analyzed within the context of neighboring features. Neighboring features inside the distance band (we used the method Euclidean Distance) receive a weight of one, while neighboring features outside the distance band receive a weight of zero. Finally, \(n\) is the total number of features; and \(\bar{x}\) and \(S\) are described by Equations (4) and (5), respectively.

\[
\bar{x} = \frac{\sum_{j=1}^{n} X_j}{n} \quad (4)
\]

\[
S = \sqrt{\frac{\sum_{j=1}^{n} X_j^2}{n} - \left(\bar{x}\right)^2} \quad (5)
\]

The K means method identifies groups where all feature values within each group are as similar as possible, and all groups themselves are as different as possible. In other words, the K means algorithm searches for a best solution that maximizes both intra-

Fig. 2. Hot and cold spots for 2012, l, confidence interval.
group similarity and inter-group difference (Hartigan and Wong, 1979).

In order to create each group, the K means algorithm first identifies a seed feature randomly, then assigns all features to the closest seed feature. Finally, it calculates a mean data center for each group of features, and reassigns each feature to the closest center. We configured the K mean algorithm to group features using the K nearest neighbors approach. This approach determined that a municipal district will be included in a group only if at least one other municipal district in the same group is a natural neighbor.

The number of seed features selected randomly matched the number of groups. To estimate the optimal number of groups, we applied the Calinski-Harabasz pseudo F-statistic, as presented by the following equations:

$$CH = \frac{\frac{T_1}{nc-1}}{\frac{T_2}{n-nc}}$$  \hspace{1cm} (6)

where $CH$ is the Calinski-Harabasz pseudo F-statistic; $nc$ is the number of classes (groups); $n$ is the number of features (municipal districts, which is equal to 5570); and $T_2$ is defined as follows:

![Fig. 3. K means analysis. Descriptive statistics for each group for emission inventory (tons), population, HDI, and urbanization rate (%). Std. Dev., standard deviation; Min, minimum; Max, maximum.](image-url)
where \( SST \) is a reflection of between-group differences (described by Equation (8)) and \( SSE \) is a reflection within-group similarity (described by Equation (9)).

\[
SST = \sum_{m=1}^{nc} \sum_{i=1}^{ni} \sum_{k=1}^{nv} \left( V_{mj}^k - \bar{V}_k \right)^2 \tag{8}
\]

\[
SSE = \sum_{m=1}^{nc} \sum_{i=1}^{ni} \sum_{k=1}^{nv} \left( V_{mj}^k - \bar{V}_g \right)^2 \tag{9}
\]

where \( ni \) is the number of features in group \( m \); \( nc \) is the number of classes (groups); \( nv \) is the number of variables used to group features; \( V_{mj}^k \) is the value of the \( k \)th variable of the \( j \)th feature in the \( m \)th group; \( \bar{V}_k \) is the mean value of the \( k \)th variable; and \( \bar{V}_g \) is the mean value of the \( k \)th variable in group \( m \).

We first performed simulations in order to estimate the optimal number of groups. Then, based on the simulations, we established the number of clusters and applied the K means.

We used a number of variables to group the municipal districts’ features (\( nv \)), such as emission inventories, population, gross domestic product (GDP), urbanization rate (% of urbanized area), length of highways (km), Human Development Index (HDI), and distance of the municipal district from its state capital (km). We calculated length of highways and distance from the capital using GIS techniques. We obtained population, GDP, urbanization rate, and HDI information from the Brazilian Institute of Geography and Statistics (IBGE, 2012). As these data were collected only for 2012, we performed the K mean analysis only for the emission inventories from 2012. Finally, we used the software ArcGIS to perform all geostatistical analysis.

### 2.4. Spatial relationship

We assessed spatial patterns of emission inventories (only for 2012) using the covariates population, GDP, urbanization rate, length of highways, HDI, and distance of the municipal district from its state capital. Our outcome was the air pollutants. We used the Ordinary Least Squares (OLS) regression to perform this analysis. For each regression model the number of observation was 5570.
3. Results and discussion

3.1. Spatial cluster analysis using Getis-Ord Gi*  

We performed Getis-Ord Gi* to identify hot and cold spots and found similar results for each of the study years (2001–2012). Fig. 2 shows the results for year 2012 and Appendices 1–6 display the results for the other years. The hot and cold spot analysis identified the municipal districts with relatively higher and lower traffic emissions. The statistical significance of this result is based on the p-value, z-score and confidence interval (Gi Bin). The regions identified as hot spots were located in the southeastern region and were similar for all pollutants. For CO, we identified 665 (12%) municipal districts grouped as hot spots; for NMHC, 679 (12%) municipal districts; for CH4, 682 (12%) municipal districts; for NOx, 812 (15%) municipal districts; for PM, 761 (14%) municipal districts; and for CO2, 732 (13%) municipal districts (Fig. 2). São Paulo, the largest Brazilian city, is located in a hot spot municipal district and also contains the highest number of vehicles, approximately 9% of the national fleet. Similarly, Rio de Janeiro represents 3% of the national fleet and is located in a municipal district identified as a hot spot in our analysis.

Cold spots were concentrated mostly in the northeast and differed by pollutant. For CO, we identified 741 (13%) municipal districts grouped as cold spots; for NMHC, 902 (16%) municipal districts; for CH4, 835 (15%) municipal districts; for NOx, 1300 (23%) municipal districts; for PM, 1201 (22%) municipal districts; and for CO2, 1090 (20%) municipal districts (Fig. 2). Although the Northeast has had a strong fleet growth rate since 2001, it still contains a relatively low number of vehicles. The observed differences among the six pollutant emissions in the cold spots are probably related to the spatial distribution of the diesel fleet in these municipal districts, especially trucks. Greater emissions across the cold spot municipal districts were observed for NOx, PM, and CO2, which may be due to higher truck emissions in these regions.

Fig. 5. Grouping analysis. Spatial distribution of groups and variables. Emission inventory variable is CO. For the variables CO, population, HDI, urbanization rate, GDP, distance from the capital, and length of highways, we considered the average value of the group.
3.2. Spatial cluster analysis using K means

We next performed a K means analysis to identify groups of municipal districts with similar values of emissions. The K means grouping analysis resulted in five groups of municipal districts, which represent the best solution that maximizes both intra-group similarity and inter-group dissimilarity. We determined the number of groups using a pseudo F-statistic test (p value < 0.01). We performed this test for 15 simulations (simulating 2 groups, 3 groups, 4 groups, up to 15 groups). The highest pseudo F-statistic value was found for five groups (value equal to 611.43; Appendix 7).

We initially performed the K means analysis using data on emissions inventories (total emissions for all vehicles, 2012 only) and six additional variables: population, GDP, urbanization rate, length of highways, HDI, and distance from the state capital. We arrived at similar special cluster analysis results when considering the different types of pollutants in the emissions inventory data. Therefore, we present here only data for CO as the emission inventory variable.

The K mean analysis calculates an R² value for each variable. This value was found for the different parameters are as follows: 0.68 for emission inventory; 0.57 for population; 0.35 for HDI; 0.31 for urbanization rate; 0.12 for GDP; 0.11 for distance from the capital, and; 0.006 for length of highways. Considering R² values from emission inventory, for example, it reflects how much of the variation in the original emissions data was retained after the grouping process.

Other studies have shown that environmental and socioeconomic variables are important in distinguishing spatial clusters. Austin et al. (2013), for example, identified particle composition clusters in the United States based on geographic locations, emissions profiles, population, and proximity to major emissions sources. Zou et al. (2014) examined spatial clusters of air pollution exposure across the United States using population, educational level, and income as factors. Also, Perc and Szolnoki (2010) show that spatial structure and environment factors (e.g. population, income, traffic) are significant variables to guide public strategies that focus on resolving social dilemmas.

Figs. 3 and 4 summarize the statistics for each variable in each group. On average, group 1 (blue) and group 2 (red) reflect of municipal districts more effectively. The larger the R² value, the better the discrimination among the municipal districts. The R² values for the different parameters are as follows: 0.68 for emission inventory; 0.57 for population; 0.35 for HDI; 0.31 for urbanization rate; 0.12 for GDP; 0.11 for distance from the capital, and; 0.006 for length of highways. Considering R² values from emission inventory, for example, it reflects how much of the variation in the original emissions data was retained after the grouping process.

Other studies have shown that environmental and socioeconomic variables are important in distinguishing spatial clusters. Austin et al. (2013), for example, identified particle composition clusters in the United States based on geographic locations, emissions profiles, population, and proximity to major emissions sources. Zou et al. (2014) examined spatial clusters of air pollution exposure across the United States using population, educational level, and income as factors. Also, Perc and Szolnoki (2010) show that spatial structure and environment factors (e.g. population, income, traffic) are significant variables to guide public strategies that focus on resolving social dilemmas.

Figs. 3 and 4 summarize the statistics for each variable in each group. On average, group 1 (blue) and group 2 (red) reflect

### Table 1
Spatial regression results (number of observation for each regression = 5570).

<table>
<thead>
<tr>
<th>Po</th>
<th>Var</th>
<th>Coef</th>
<th>SE</th>
<th>t</th>
<th>Prob</th>
<th>VIF</th>
<th>R²</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO</td>
<td>Int</td>
<td>−274.332</td>
<td>37.501</td>
<td>−7.315</td>
<td>0.000</td>
<td>−</td>
<td>0.944</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Pop</td>
<td>0.009</td>
<td>0.000</td>
<td>297.756</td>
<td>0.000</td>
<td>1.061</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GDP</td>
<td>0.002</td>
<td>0.000</td>
<td>3.336</td>
<td>0.000</td>
<td>1.088</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Urb</td>
<td>0.271</td>
<td>0.374</td>
<td>0.722</td>
<td>0.003</td>
<td>1.323</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LH</td>
<td>0.263</td>
<td>0.088</td>
<td>2.983</td>
<td>0.002</td>
<td>1.044</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dist</td>
<td>0.000</td>
<td>0.000</td>
<td>3.818</td>
<td>0.000</td>
<td>1.038</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** Po, Outcome - pollutant; Var, Predictor variable; Coef, Coefficient; SE, Standard error; t, t-statistic; Prob, Probability of the coefficients; VIF, Variance inflation factor; R², adjusted R-squared; Prob*, Overall model significance by Joint Wald Statistic; CO, carbon monoxide; Int, Intercept; Pop, Population; GDP, Gross domestic product; Urb, Urbanization rate; LH, Length of highways; Dist, Distance from the capital; HDI, Human Development Index; NMHC, non-methane hydrocarbon; CH₄, methane; NOₓ, nitrogen oxides; PM, particulate matter; CO₂, carbon dioxide.

* p < 0.01, statistically significant.
municipal districts with lower emissions and population (<global lower quantile >) and Interquartile range values - IQR (between the global lower quantile and global upper quantile) for the other variables. The difference between these two groups is the IQR values. For example, group 1 considers the upper quantile value of the variable GDP, while group 2 considers the lower quantile values.

Group 3 (green) reflects municipal districts with very high values (>global upper whisker) for emissions, population, and length of highways; high values (>global upper quantile) for HDI, GDP, and urbanization rate; and lower values for distance from the capital (Figs. 3 and 4).

Group 4 (orange) represents municipal districts with lower values of emissions and population and lower IQR values for HDI, urbanization rate, GDP, distance from the capital and length of highways (Figs. 3 and 4).

Finally, group 5 (purple) reflects municipal districts with lower values for emissions and population, lower IQR values for distance from the capital and length of highways, and high values for HDI, urbanization rate and GDP (Figs. 3 and 4).

Fig. 5 shows the spatial distribution of both the municipal district groups and the variables. Group 2 has the highest number of municipal districts (2,517), while group 3 has the lowest number (1). Note that Figs. 3, 4, and 5 use the same color key to represent each group of municipal districts.

Our findings suggest a similarity between the spatial distribution of emissions and location of the groups defined by K means analysis (Fig. 5). Group 2 represents municipal districts in the Northeast and Northwest regions, which have lower income, traffic, and emissions. In contrast, groups 1, 3, 4, and 5 represent regions with high income, traffic, and emissions. Specifically, group 3 includes São Paulo, with large emissions, length of highways, and population.

3.3. Spatial regression

The factors contributing to emission inventories were estimated using regression analysis, based on data from 2012 (Table 1). We found statistically significant coefficients (p < 0.01) in all models (six models, one for each pollutant). All predictors were positively associated with emissions for the six models. HDI had the highest coefficient. In addition, there is no multicollinearity (Variance inflation factor < 7.5) between the predictor variables and our models had high performance (R² > 0.80; Table 1).

In Appendix 8, we present the residuals map obtained from each model. Overall, residuals for the Northern region were < -0.5, while those for the South were > 0.5. Appendices 9 to 14 show the histogram of standardized residuals, indicating no bias in our models. We used a simple regression model to assess the association between emissions and covariates, which can lead to a limitation of the specific contributing variables in order to reduce exposure and mitigate adverse health outcomes. Our simple regression models could be an effective tool for developing cost effective air pollution control policies that take into account socio-economic factors as important covariates for the implementation of further complex models.

Acknowledgements

This publication was made possible by USEPA grant RD-834798 and RD-835872. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclepro.2017.02.010.

References


4. Conclusions

To our knowledge, this is the first study to model spatial patterns of traffic emissions across continental Brazil at the municipal district level. Brazil has a rapidly growing fleet and its emissions contribute significantly to South America and even at a global level. This is an essential step to guide and inform policy decisions. Our approach can be an effective tool to develop mitigation strategies tailored to specific regions in Brazil. Our analysis identified groups of municipal districts with varying levels of traffic emissions, suggesting that public policies need to be tailored to address the specific contributing variables in order to reduce exposure and mitigate adverse health outcomes. Our simple regression models could be an effective tool for developing cost effective air pollution control policies that take into account socio-economic factors as important covariates for the implementation of further complex models.


