Environmental Science and Pollution Research (2018) 25:25942–25954
https://doi.org/10.1007/s11356-018-2614-x

RESEARCH ARTICLE

Modelling local uncertainty in relations between birth weight and air quality within an urban area: combining geographically weighted regression with geostatistical simulation

Manuel Castro Ribeiro[1] · Maria João Pereira[2]

Received: 19 January 2018 / Accepted: 18 June 2018 / Published online: 1 July 2018

Abstract

In this study, we combine known methods to present a new approach to assess local distributions of estimated parameters measuring associations between air quality and birth weight in the urban area of Sines (Portugal). To model exposure and capture short-distance variations in air quality, we use a Regression Kriging estimator combining air quality point data with land use auxiliary data. To assess uncertainty of exposure, the Kriging estimator is incorporated in a sequential Gaussian simulation algorithm (sGs) providing a set of simulated exposure maps with similar spatial structural dependence and statistical properties of observed data. Following the completion of the simulation runs, we fit a geographically weighted generalized linear model (GWGLM) for each mother's place of residence, using observed health data and simulated exposure data, and repeat this procedure for each simulated map. Once the fit of GWGLM with all exposure maps is finished, we take the distribution of local estimated parameters measuring associations between exposure and birth weight, thus providing a measure of uncertainty in the local estimates. Results reveal that the distribution of local parameters did not vary substantially. Combining both methods (GWGLM and sGs), however, we are able to incorporate local uncertainty on the estimated associations providing an additional tool for analysis of the impacts of place in health.

Keywords Environmental epidemiology · Exposure · GWR · Regression Kriging · Uncertainty

Introduction

In most studies correlating health outcomes with air pollution, spatial epidemiology plays an important role, combining methods from epidemiology and spatial modelling to describe and analyse the impact of pollution in health. Usually, personal exposure assignments are based on misaligned data collected at air-quality monitoring stations not coinciding with health data locations (Gryparis et al. 2009), so spatial interpolators are needed to predict air pollution measurements in unsampled locations and to assign exposures at health data locations (Jerrett et al. 2005). Methodological developments in geostatistical interpolators over the last two decades may play an important role in the future of spatial epidemiology, especially in situations where location data (e.g. addresses) are available (Lawson et al. 2016). Besides improvements in methods to map the risk of disease (Kyriakidis 2004; Goovaerts et al. 2005; Goovaerts 2009; Hampton et al. 2011), geostatistics also contribute to provide clearer pictures in geographic correlation studies by mapping air pollution exposures to link with health data (Lee et al. 2012; Kalkbrenner et al. 2015; Fei et al. 2017). In addition, geostatistics provide suitable methods to assess their spatial uncertainty, based on simulation algorithms (Young et al. 2008; Goldman et al. 2012). This is important because measures of exposure can be misleading if they do not take into account the uncertainty of predictions, since the extent of uncertainty varies throughout the spatial domain. Waller and Gotway (2004) first explored spatial exposure uncertainty of particulate matter to measure associations with very low birth weights using stochastic simulation and drawn geostatistical uncertainty from generalized linear model (GLM) analysis,

Responsible editor: Philippe Garrigues

✉ Manuel Castro Ribeiro
manuel.ribeiro@tecnico.ulisboa.pt

[1] CERENA, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisbon, Portugal

[2] CERENA, DECivil, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisbon, Portugal

with adjustment for well-known potential confounders. More recently, Ribeiro et al. (2016) applied similar methods in an urban area. In both studies, however, no appropriate local adjustments for variations in the estimated associations were taken into account. An alternative regression model developed in the recent years (Brunsdon et al. 1998; Fotheringham et al. 1998) addresses this issue, with adjustments for variations in local associations carried out directly in the model parameters. This class of models, generically known as geographically weighted regression (GWR), allows the parameters to vary smoothly as a function of spatial neighbourhoods and the correlations to have spatially varied random effects (Brunsdon et al. 1998). The combined use of GWR and geostatistical models for spatial prediction has been evaluated (Harris et al. 2010), and a recent application combining GWR and geostatistical methods was proposed for mapping high spatial resolution soil moisture data (Jin et al. 2018).

Applications of GWR in the spatial epidemiology field include, for example, analysis of association between the exposure to alcohol and violence outcomes (Waller et al. 2007) or ozone concentrations and myocardial infarctions (Young et al. 2008). Besides conventional Gaussian models, GWR has been recently adapted to model data with other error distributions belonging to the exponential family (Nelder and Wedderburn 1972), like negative binomial distribution (da Silva and Rodrigues 2013) or Poisson distribution (Nakaya et al. 2005), also known as geographically weighted generalized linear models (Brunsdon and Singleton 2015).

In this work, within a small urban area, we describe a new methodology based on geostatistical simulation and geographically weighted generalized linear models (GWGLM) to estimate local variations in the associations between air quality levels and birth weight variations, while incorporating geostatistical uncertainty of exposures. Because variations in air quality tend to vary at short distances within city limits due to variations in land use and in land use intensity, predicted air pollution exposures at unsampled locations are assessed with Regression Kriging (Hengl et al. 2007). With this spatial interpolation technique, air quality at unsampled locations can be predicted, incorporating auxiliary land use data, known to be well correlated with air quality, while modelling the residual as a stationary random spatial function (Fortin et al. 2012). Regression Kriging is then incorporated within a geostatistical simulation algorithm to assess spatial uncertainty of predicted exposures. In the final part of the analysis, simulated exposures are used to derive local distribution of associations between air quality and birth weight drawn from an analysis performed with GWGLM.

Data

In this study, we used air quality data and personal health data collected in the city of Sines (Fig. 1) under the Gestão Integrada Saúde e Ambiente (GISA) project, a health and environment project developed in the coastal Alentejo region (Portugal) during 2007–2010.

Health data

The health dataset, collected in the period 2008–2010, included information on 227 mothers enrolled in the GISA project and living in the city of Sines during pregnancy. Mothers' places of residence were geocoded using a georeferenced street map provided by Correios de Portugal (CTT Correios 2010). The following risk factors were included in analysis: maternal body mass index (BMI), maternal age, maternal weight gain during pregnancy, gestational age (weeks of gestation) pregnancy surveillance (first antenatal visit and number of visits during pregnancy), maternal smoking during pregnancy (active smoking) and exposure to environmental tobacco smoke (or passive smoking), education and occupation. All these risk factors are known to be associated with variations in birth weight (Kramer 2003).

Birth weight is highly related to duration of gestation and to rate of foetal growth (Kramer 2003). Air pollutants may be involved with birth weight directly through effects on the rate of foetal growth or indirectly by impairing maternal health (Glinianaia et al. 2004). Therefore, instead of birth weight, the outcome analysed in this study was birth weight percentile by sex and gestational age. Birth weight percentiles, hereafter called birth weight, provided a clearer analysis of relations between air pollution and rate of foetal growth by removing the effect of duration of gestation (and sex) from the analysis.

Air quality data (AQIv)

Lichens are bioindicators of air pollution because they are very sensitive to variations in atmospheric pollution. In polluted areas, their frequency and diversity tend to decline (Rose and Hawksworth 1981). They are affordable bioindicators of air quality, as they can be found on a wide range of places on the planet, including ground, water, rocks or human-made structures and tree bark. Lichens as bioindicators of air quality have been applied in several studies (Garty 1993; Conti and Cecchetti 2001; Wolterbeek et al. 2003; Pinho et al. 2008b; Canha et al. 2014; Munzi et al. 2014; Pinho et al. 2014). A standard protocol (Asta et al. 2002) provided guidelines to use lichens as bioindicators of air quality, which have been applied in several studies (Loppi et al. 2002; Pinho et al. 2004; Ribeiro et al. 2012; Llop et al. 2012; Paoli et al. 2015).

Due to scarcity of existing air-quality monitoring stations in the city (only one), this study considered the use of bioindicators as a surrogate of air pollution, as they exist in several trees within the city, providing a relatively dense network of 83 sample sites. The protocol was applied to all available trees and a value of air quality, Air Quality Index value

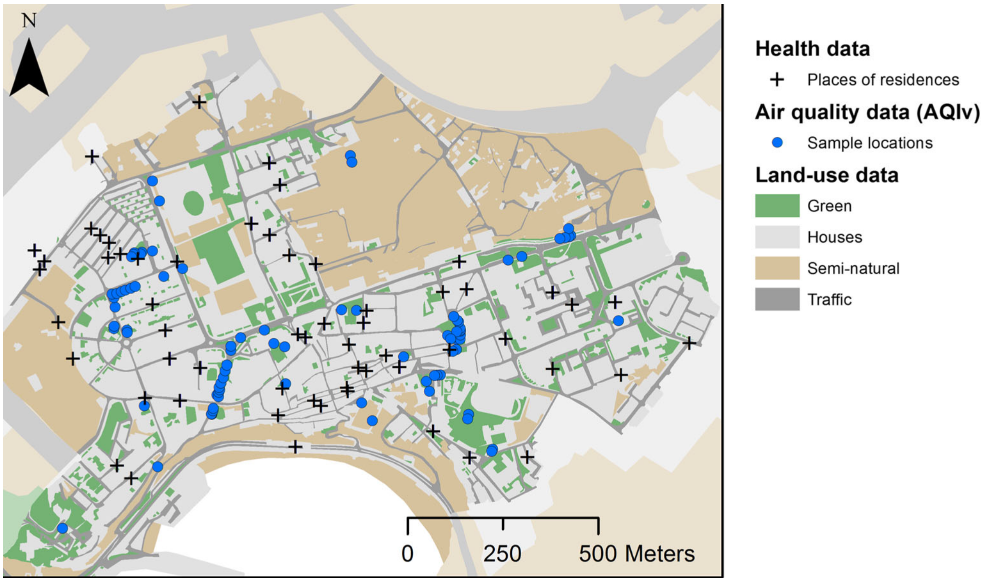


Fig. 1 Map illustrating land use categories, mothers' places of residence and air quality sample locations in the city of Sines (from Ribeiro et al. 2016)

(AQIv) was assigned to each sampling site. A higher value of the AQIv indicates better air quality.

Combined with the lichen biomonitoring network, land use data was also used to model air pollution spatial distribution, because they are well correlated (Pinho et al. 2008a; Llop et al. 2017). The land use dataset was provided by Llop and colleagues (Llop et al. 2012). Four types of land use were considered for analysis: green areas, corresponding to public garden areas within the city; semi-natural areas, corresponding to areas with abandoned agriculture and semi-natural vegetation areas; houses, corresponding to residential areas with low traffic; and traffic areas, corresponding to the vicinity of main roads.

Methods

To describe and analyse observed data on air quality and health variables, we conducted descriptive and inferential statistical analysis. Then, we modelled spatial covariance of air quality data to interpolate an exposure map with Kriging estimator. The exposure map was further used to select a proper radial buffer distance for personal exposure assignment. That was achieved by comparing the goodness-of-fit from generalized linear models (GLM) with exposures estimated at different radial buffer distances centred at mothers' places of residence.

The spatial covariance of air quality data was further used within sequential Gaussian simulation (sGs) to generate multiple exposure maps with similar spatial patterns but with extreme exposure scenarios. The goal was to assess uncertainty of exposures. For each simulated map, we assigned mean personal exposures using the buffer distance selected before and used geographically weighted generalized linear models (GWGLM) to quantify local relationships with birth weight, while controlling the effect of health covariates. After fitting GWGLM for all simulations, we have calculated the distribution of exposure parameters estimated at each mother's place of residence, reflecting the local uncertainty of estimated associations with birth weight.

In Fig. 2, we illustrate each one of these steps and the remainder of this section describes in detail the methods used. All statistical and geostatistical modelling was performed in R (R Core Team 2014). Packages raster (Hijmans and van Etten 2012) and gstat (Pebesma 2004) were used to model air quality and generate geostatistical simulations. Arc GIS software (Environmental Systems Research Institute 2006) was used to integrate all spatial data in one work environment and to produce the maps.

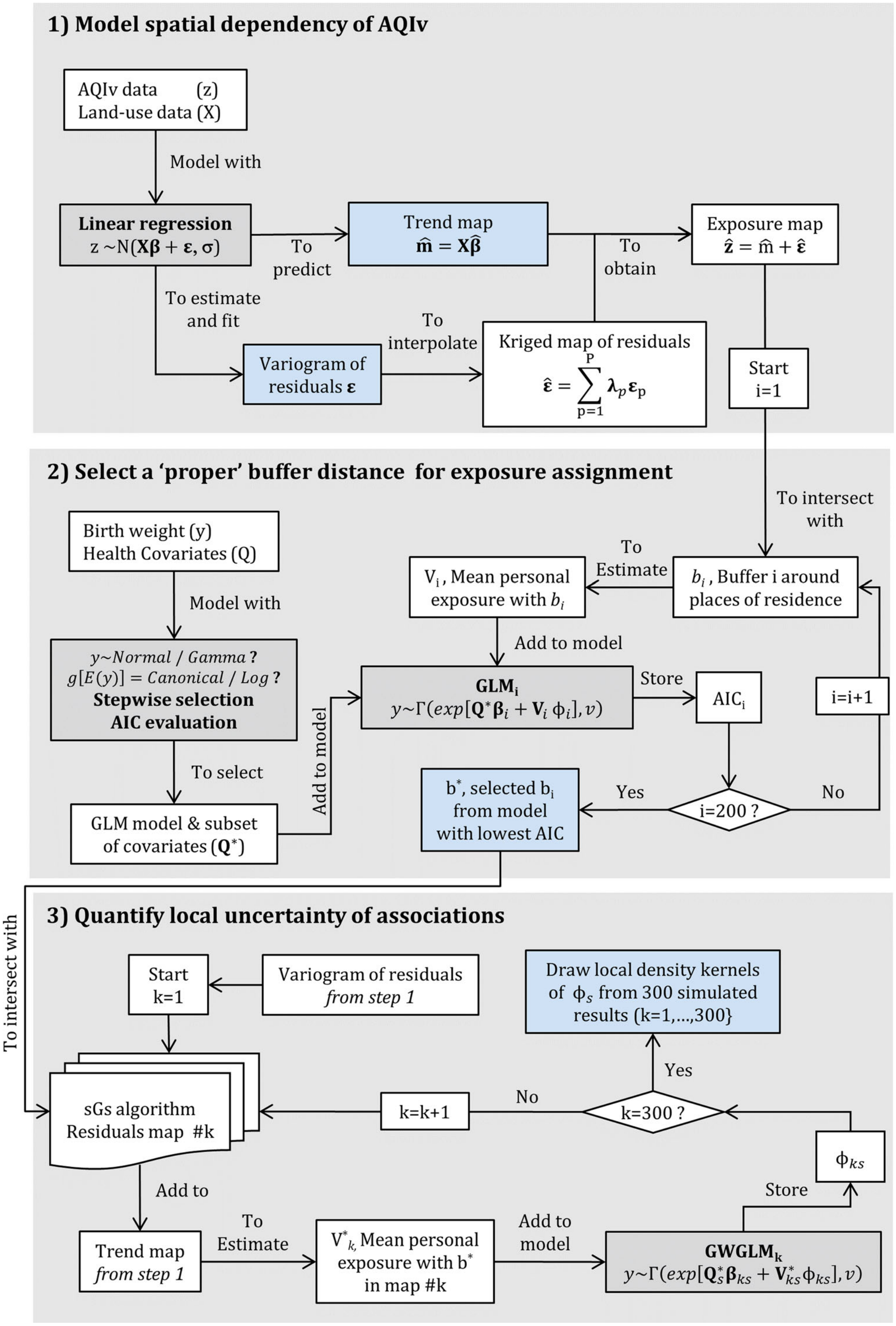
Descriptive and inferential statistical analysis

To find if mean AQIv is different between land use categories, we used the analysis of variance (ANOVA). We assessed AQIv differences between land use categories, using paired *t* tests. In these analysis, however, we emphasize that our primary goal was not to test hypothesis but to find any indications of dependence and trends between AQIv and land use categories, because these are relevant for posterior geostatistical modelling.

Model spatial dependency of AQIv

Air quality tends to vary at short distances within urban areas due to variations in land use (e.g. green spaces, roadways, residential areas) and in land use intensity (e.g. intensity of traffic-related pollution is different in major roads and secondary roads). To capture such short-distance variations in spatial predictions, we used Regression Kriging (Hengl et al. 2007) also known as Kriging with External Drift (Goovaerts 1997)

Fig. 2 Schematic diagram describing the proposed methodology in three steps. Text boxes represent variables, algorithms, control flows and functions and the arrow lines represent relations between them as described in text along with the lines. Text boxes with blue background represent final outputs of each step. Text boxes with grey background represent the regression models applied



or Residual Kriging (Neuman and Jacobson 1984). Regression Kriging interpolates a principal variable sampled at some locations (AQIv data in this study), by modelling the relationship with spatially correlated auxiliary variables, available everywhere in spatial domain (land use data in this study).

Typically, this approach combines multiple linear regression to model the relationship between principal (dependent) and auxiliary (explanatory) variables, with Ordinary Kriging or Simple Kriging estimator to predict the residuals. Once the trend is modelled, the residuals can be interpolated with a Kriging estimator and added to the trend (Hengl et al. 2007), as shown in Eq. (1). In Eq. (1), the predicted value of the principal variable at any unsampled location, $\hat{z}_0$, is obtained by summing a trend, $\hat{m}_0$ (derived from the relation between principal and auxiliary variable), with the sum of neighbour residuals ε_p (with $p = 1, \ldots, P$), derived from P neighbour observations, which are weighted by the Kriging optimal weights, λ_p.

$$\hat{z}_0 = \hat{m}_0 + \sum_{p=1}^{p} \lambda_p \varepsilon_p \tag{1}$$

To estimate the parameters of the regression and its residuals, the following steps are taken:

1- Fit a regression model to predict value, z, as a linear function of auxiliary data, X. Because auxiliary data in this study is a categorical variable, matrix X is represented with dummy indicator variables. The vector of estimated parameters associated with auxiliary variables, $\hat{\boldsymbol{\beta}}_{ols}$, is derived by ordinary least squares (OLS). The residual of the model fitted for a known observation located at p, ε_p (with $p = 1, \ldots, P$), can then be computed:

$$\varepsilon_p = z_p - X_p \hat{\boldsymbol{\beta}}_{ols}$$

2- Estimate the semi-variogram of ε and fit a variogram model, $\hat{\gamma}\left(h; \hat{\boldsymbol{\theta}}\right)$, with vector $\hat{\boldsymbol{\theta}} = (\hat{c}_0, \hat{c}_e, \hat{a}_e)$ representing the estimates of $\hat{c}_0$ (nugget effect), $\hat{c}_e$ (partial sill) and $\hat{a}_e$ (range), and h represents some distance between pairs of observations. The variogram model parameters are estimated iteratively using a weighted least-squares estimator (Pebesma 2004). The initial values specified for parameter estimation of c_0, c_e and a_e are based on the visual inspection of the semi-variogram estimate and close to the expected estimates.

3- Estimate the (spatial) covariance matrix of residuals, $\hat{\mathbf{C}}$, based on the relation between the spatial covariance and the corresponding variogram: $C(h) = \sigma^2 - \gamma(h)$, where σ^2 represents the variance (or spatial covariance at zero distance).

4- Re-fit a linear regression model to predict response value as a function of auxiliary data. But now, instead of OLS, the parameters are estimated by generalized least squares (GLS),

$$\hat{\boldsymbol{\beta}}_{gls} = \left(\mathbf{X}^{\mathbf{T}} \hat{\mathbf{C}}^{-\mathbf{1}} \mathbf{X}\right)^{-1} \mathbf{X}^{\mathbf{T}} \hat{\mathbf{C}}^{-\mathbf{1}} \mathbf{z}$$

$\hat{\boldsymbol{\beta}}_{gls}$ represents the vector of parameters estimated with generalized least squares, $\hat{\mathbf{C}}$ represents the covariance matrix between estimated residuals, $\mathbf{X}$ is the matrix of auxiliary data and $\mathbf{z}$ is the vector of observed response data.

5- The residuals of the re-fitted model at observation located at p, ε_p^*, can then be computed as follows:

$$\varepsilon_p^* = z_p - X_p \hat{\boldsymbol{\beta}}_{gls}$$

6- Steps 2–5 are repeated until no significant change occurs in parameter estimates.

Hengl et al. (2007) refers the work of Kitanidis (1993) to stress that in most applications, no relevant gain occurs in estimating spatial covariance after the first iteration with OLS. Some studies (Odeh et al. 1994; Minasny and McBratney 2007; Hengl et al. 2007) applied this “short version” of the method with satisfactory results. In this study, we follow the latter version, confining the estimate of spatial covariance to the residuals derived from OLS estimation (steps 1–2), as it captures the essence of the Regression Kriging approach.

If the principal variable is correlated with auxiliary variable, the regression model will explain some part of the variability. First, we predict the value of the principal variable at every unsampled locations of the spatial domain, using the regression model and the auxiliary data available everywhere (trend part). Then, we predict the part of variability not explained by the model (residual part) with the Ordinary Kriging linear estimator. The predicted value of residual at any unsampled location, $\hat{\varepsilon}_0$, is obtained by summing neighbour residuals ε_p (with $p = 1, \ldots, P$), derived from P neighbour observations, which are weighted by the Kriging optimal weights, λ_p:

$$\hat{\varepsilon}_0 = \sum_{\mathrm{p}=1}^{\mathrm{P}} \lambda_\mathrm{p} \varepsilon_\mathrm{p}$$

We finally obtain a map of principal variable, by summing both trend and residual maps as presented in Eq. (1).

Select a “proper” buffer distance for exposure assignment

To assign a personal exposure, we computed a mean AQIv using a radial buffer distance around the place of residence. However, the selection of that distance was not straightforward, since a too small radial buffer distance could provide unreliable personal average exposure estimates, whereas a too large buffer distance would add bias to the estimates. Therefore, to select an “appropriate” radial buffer distance, we performed a two-step modelling approach.

In the first step, we selected a regression model with a column vector for birth weight responses, all health covariates (data matrix **Q**) and no AQIv data included. This way we gained insight into the contributions of several well-established health risk factors to explain birth weight variations. We used stepwise procedures, evaluated goodness-of-fit (Akaike Information Criteria) and performed residual diagnostics for the normal and gamma GLM models (as both belong to the exponential family) with canonical and log-link functions ($g[E(y)]$ in Fig. 2). The gamma model with log-link function specified with a subset of health covariates (data matrix $\mathbf{Q}^*$) and associated vector of parameters, β, presented better results among all candidates and was selected for further statistical analysis. With this model (we designate this as the "health model"), presented in model (2), we assumed that birth weight is an independent variable following a gamma distribution with variance proportional to its mean square (constant shape v) and that health covariates have a linear relationship with mean birth weight (in terms of the link function).

$$y \sim \Gamma\left[exp\left(\mathbf{Q}^*\beta\right), v\right] \tag{2}$$

In the second step, we used the map of AQIv interpolated before (see "Model spatial dependency of air quality (AQIv)") to compute the mean AQIv by mother's place of residence and tested 200 different radial buffer distances (5–1000 m, with 5 m steps) centred at each place of residence. For each buffer distance, b_i (with $i = 1, \ldots, 200$), we estimated the column vector $\mathbf{V}_\mathrm{i}$ of personal average exposures and refitted the health model [2], but now with an additional parameter, ϕ_i, associated to the new variable $\mathbf{V}_\mathrm{i}$ as shown in model (3). At each $\mathbf{V}_i$, the vector of parameters associated to the subset of health covariates, $\boldsymbol{\beta}_\mathrm{i}$, is obviously estimated.

$$y \sim \Gamma\left[exp\left(\mathbf{Q}^*\beta_\mathrm{i} + \mathbf{V}_\mathrm{i}\phi_\mathrm{i}\right), v\right] \tag{3}$$

Our goal was to find the lowest Akaike Information Criteria from the different models quantifying the relationship between the exposure and birth weight, while controlling the effect of health covariates. We extracted the Akaike Information Criterion (AIC) (Akaike 1973) score associated with each model. Finally, we selected a proper buffer distance for exposure assignment from the model with the lowest AIC score among all the models considered.

Quantify local uncertainty of associations

Sequential Gaussian Simulation (sGs)

To map spatial uncertainty of AQIv, we generated conditional simulations of AQIv residuals data and added the simulated residuals to the trend map to provide several maps that match the statistical properties of observed data. We considered the use of sequential Gaussian simulation algorithm in this study, but we could have considered any other sequential simulation algorithm suited to continuous data. In a nutshell, the workflow of the algorithm can be described in the following manner:

1. The algorithm randomly defines a path over the entire study area (defined as a regular grid) passing through all grid nodes to be simulated.
2. For the first node, the Regression Kriging interpolator estimates a local average and local variance of residuals conditioned to the cumulative distribution function of residuals. A simulated value is obtained using the inverse Gaussian distribution (Hengl 2009), which is then added to the conditioning dataset.
3. The same procedure (2.) is followed for the next node, looping until all nodes of the grid have been visited and simulated.
4. Add the simulated residuals to the trend map.
5. Using this approach, we generated 300 possible exposure scenarios, from which spatial uncertainty could be retrieved.

Geographically weighted generalized linear model (GWGLM)

Generalized linear models (GLM) are widely used to model health events, because they are flexible and generally suited for analysing correlations that are generally poorly represented by Gaussian distributions. When using such models, it is assumed that a single global model describes the relationships between variables which may fail to capture spatial variations in the relationships. To cope with spatially varying associations, Brunsdon et al. (1998) proposed a geographically weighted regression (GWR), which uses distance-weighted subsets of neighbour observations to estimate the regression parameters at each sampling site. However, this model was implemented to predict continuous variables with Gaussian error distribution. More recently, the use of GWR has been extended to the exponential family of distributions (Nakaya et al. 2005; Chen and Yang 2012; da Silva and Rodrigues 2013; Li et al. 2013), also known as geographically weighted generalized linear model (GWGLM). We explored the use of GWGLM with gamma distribution and log-link function presented in model (4).

$$y_s \sim \Gamma\left[exp\left(\mathbf{Q}_s^*\beta_{ks} + \mathbf{V}_{k\mathbf{s}}^*\phi_{k\mathbf{s}}\right), v\right] \tag{4}$$

In model (4), $y_\mathbf{s}$ represents birth weight variable at mother's place of residence with vector of coordinates $\mathbf{s} = (x_\mathbf{s}, y_\mathbf{s})$, following a gamma distribution with scale parameter $exp\left(\mathbf{Q}_\mathbf{s}{}^*\beta_{ks} + \mathbf{V}_k^*\phi_{ks}\right)$ and with constant shape v, $\mathbf{Q}_\mathbf{s}{}^*$ represents the local data matrix of health covariates at location $\mathbf{s}$, $\boldsymbol{\beta}_{k\mathbf{s}}$ represents the vector of local parameters associated with $\mathbf{Q}_\mathbf{s}{}^*$ in simulation k at location $\mathbf{s}$, $\mathbf{V}_{k\mathbf{s}}^*$ is a column vector of local personal average exposures at simulation k at location $\mathbf{s}$ and $\phi_{k\mathbf{s}}$ is the local parameter associated with $\mathbf{V}_{k\mathbf{s}}^*$ at simulation k at location $\mathbf{s}$. Local parameters at location $\mathbf{s}$ are estimated by maximizing the log-likelihood.

In this maximization process, a suitable local bandwidth around each place of residence is selected to build a geographical weighting matrix for neighbour observations. A natural solution is to use some distance-decaying function between location $\mathbf{s}$ and each of neighbours' locations within bandwidth limits. The classical weighting functions used in GWR, also denoted as kernels, are fixed or adaptive type. In fixed type, the weight value of a neighbour location $\mathbf{s}^*$ to estimate the local parameters at mother's place of residence located at $\mathbf{s}$, $r_{\mathbf{ss}^*}$, can be obtained with the Gaussian kernel function,

$$r_{\mathbf{ss}^*} = e^{\left(-d_{\mathbf{ss}^*}/w\right)^2}$$

where $d_{\mathbf{ss}^*}$ is the distance between locations $\mathbf{s}$ and $\mathbf{s}^*$, and w is the bandwidth selected. In this type of kernel, w is constant and sets the rate at which the weights gradually decay. In this study, however, we considered an adaptive kernel type, better suited when geographic density of observations varies substantially. The function used, known as Gaussian adaptive kernel, calculates $r_{\mathbf{ss}^*}$ in the following manner,

$$r_{\mathbf{ss}^*} = e^{\left(-d_{\mathbf{ss}^*}/w_u\right)^2}$$

where $d_{\mathbf{ss}^*}$ is the distance between locations $\mathbf{s}$ and $\mathbf{s}^*$, and w_u is an adaptive bandwidth size defined as the distance to the uth nearest neighbour. In this type of kernel, the bandwidth size may vary, as it adapts to the variations in the density of neighbours: larger bandwidths are selected if the number of neighbour locations is smaller, and smaller bandwidths are selected if the number of neighbour locations is bigger.

To select the "best" GWGLM model, we used the corrected Akaike Information Criterion (AIC$_\text{c}$) presented in Eq. (5).

$$AIC_c(w) = d(w) + 2m(w) + 2\left\{\frac{m(w)m(w)+1}{n-m(w)-1}\right\} \tag{5}$$

In Eq. (5), AIC$_c$(w) is the corrected AIC of model with bandwidth w, $d(w)$ is its deviance statistic, $m(w)$ is the number of effective parameters and n the number of observations. The optimal model is achieved with the bandwidth w that provides the smallest AIC$_c$ value, among all fitted models.

Local uncertainty of AQIv parameter

We assessed the local means and confidence intervals for exposure parameters numerically, using the maps produced with the sGs algorithm. For each simulation, we fitted a GWGLM that provided a vector of parameter estimates at each mother's place of residence. From the set of simulations, we could draw the local distribution of each parameter estimated at each place of residence, reflecting the local uncertainty of estimated associations with the birth weight variable.

Results

Health data places of residence, land use data and air quality data sample locations are mapped in Fig. 1. The average birth weight percentile among the 227 observations is 40, and the distribution of the variable is right skewed (Fig. 3a) (skewness = 0.58).

We measured an Air Quality Index value (AQIv) in 83 different locations within the Sines city. The overall mean is 6.0 (variance = 8.63), and the distribution of values is slightly right skewed (Fig. 3b) (skewness = 0.45). By land use category, the mean AQIv is 8.5 for green, 7.6 for semi-natural, 5.3 for houses and 4.9 for traffic. Figure 4 presents the boxplots for AQIv by land use category. Despite being presented elsewhere in the literature (Ribeiro et al. 2016), we include this figure here for the sake of completeness.

Geostatistical model of exposure

We specified a linear regression to model the response of AQIv as a function of land use to capture the part of variation in AQIv explained by land use categories (trend component). The fitted regression model explained 21% of AQIv variability (adjusted $R^2 = 0.207$). The estimated parameters are shown in regression model (6).

$$\hat{m} = 8.50 - 0.93\text{SNA} + 3.15\text{HOU} - 3.60\text{TRA} \tag{6}$$

When compared with green category (the reference category in the model), AQIv decreases as we move from land use areas with lower air pollution emissions (SNA, semi-natural) to areas with higher emissions (TRA, traffic). The t statistics, associated with the land use categories (SNA, $t = -0.78$, p-value = 0.436; HOU, $t = -3.87$, p-value < 0.001 and; TRA, $t = -4.46$, p-value < 0.001), suggest that land use variable has a statistically significant linear relationship (at 0.1%) with AQIv. We used this model to predict the trend part of AQIv at every locations of the spatial domain obtaining therefore the trend map.

With regard to the residuals of model (6), we analysed their spatial dependency. The experimental semi-variogram of the residuals showed spatial correlation. Initial parameters for variogram model were drawn from visual inspection of experimental semi-variogram. We considered an exponential model with no nugget effect, sill equal to the total variance of residuals (variance = 6.59) and range asymptotically reached at 180 m. The fitting of the exponential model converged, explaining well the spatial structure up to the first two lags (Fig. 5).

With estimation of covariance structure of the residuals, we interpolated the map of residuals using the Ordinary Kriging linear estimator. Summing both trend and residual maps, we obtained a smoothed exposure map (not shown), which was useful to find an adequate buffer for personal exposure assignment.

Assessing the radial buffer distance for exposure assignment

To assign a personal exposure, we first needed to find a proper (or optimal) radial buffer distance around mothers' places of residence.

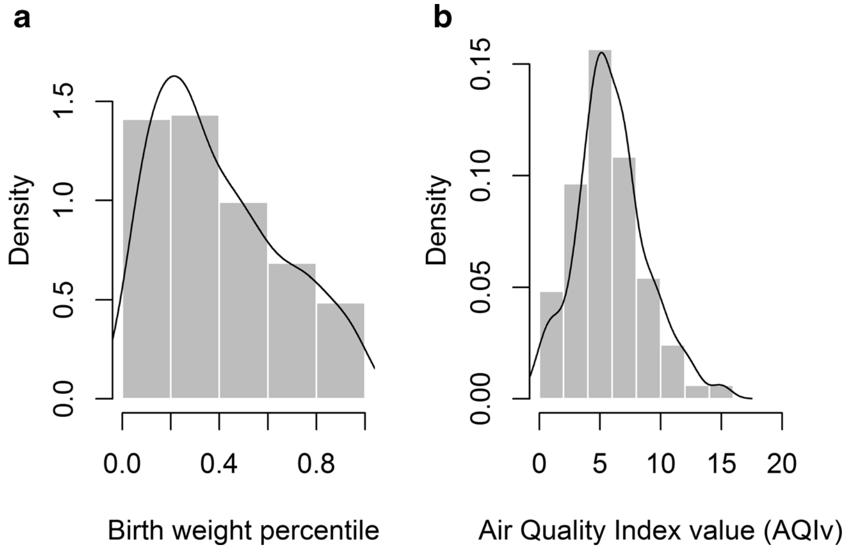


Fig. 3 Histogram (grey bars) and density function estimate (black curve) from birth weight percentile (**a**) and Air Quality Index value (**b**). Both variables showed positive skewness, more exuberant in the birth weight case

Firstly, we selected a generalized linear model using birth weight and all health covariates but did not include the exposure variable. This way we gained insight into the contributions of several well-established health risk factors to explain birth weight variations. We evaluated generalized linear models with normal and gamma distributions and canonical and log-link functions, applied a stepwise procedure for variable selection, evaluated goodness-of-fit (Akaike Information Criteria) and performed residual diagnostics (results not shown). The gamma model with log-link function presented better results among all candidates and was selected for further statistical analysis. The final model resulted in a health model retaining the variables active smoking during pregnancy (SMO, *t* value = − 1.511, *p*-value = 0.1322), maternal body mass index (BMI, *t* value = 2.612, *p*-value = 0.0096) and gestational age (GES, *t* value = 2.509, *p*-value = 0.0128). The selected health model and parameter estimates are shown below in Eq. (7).

$$\hat{y} = exp(-3.963 - 0.163SMO + 0.032BMI + 0.059GES) \quad (7)$$

We used the map of AQIv interpolated before to compute the mean AQIv by mother's place of residence and tested 200 different radial buffer distances (5–1000 m, with 5 m steps) centred at each place of residence. For each buffer distance, we estimated a mean exposure by mother's place of residence and refitted model (7) but now with an additional term associated to the new exposure variable. After fitting the regression models with all the 200 buffer sizes, we stored the goodness-of-fit Akaike Information Criterion (AIC) score associated with each one and selected the buffer of 265 m, since it was the one associated to the model with the lowest AIC score among the models considered (Fig. 6).

Therefore, for further statistical analysis, we assigned personal exposure during pregnancy, based on mean AQIv

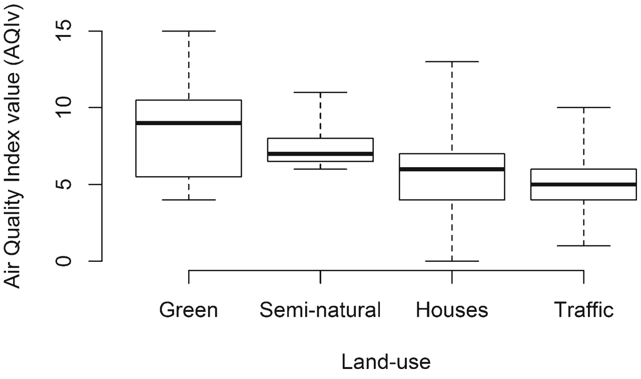


Fig. 4 Boxplot graphs for AQIv in different land use categories (from Ribeiro et al. 2016). The limits of the boxes represent the interquartile range, and the midline represents the median. The boxplot for green category is higher than the boxplots for houses (paired *t* test, *p* -value < 0.001) and traffic categories (paired *t* test, *p* -value < 0.001)

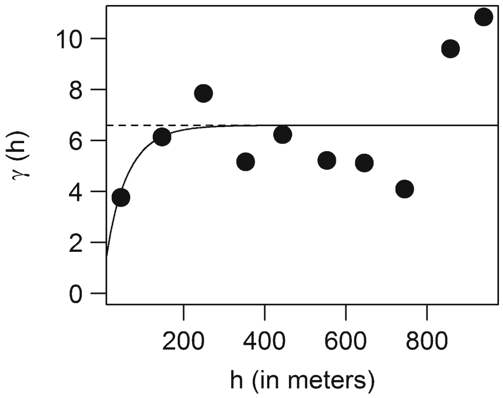


Fig. 5 Experimental semi-variogram of residuals from OLS regression model (black dots) and the exponential variogram model fitted (continuous line) up to the sill (dashed line)

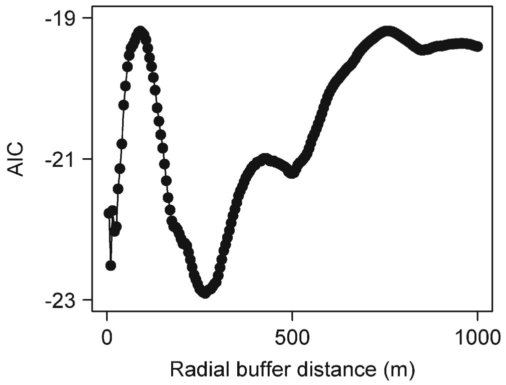


Fig. 6 Akaike Information Criteria (AIC) scores associated to 200 models fitted with different radial buffer distances (5–1000 m). The model with the lowest AIC score (AIC = − 22.91) was found with radial buffer distance set at 265 m

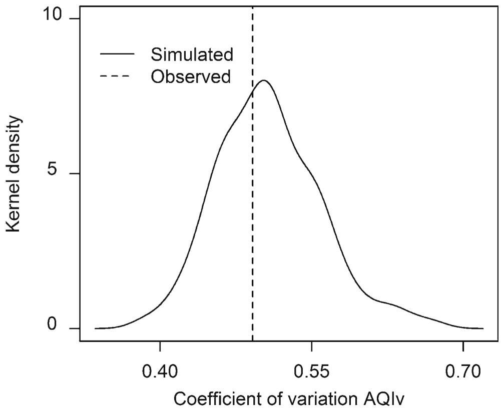


Fig. 7 Kernel density distribution of the coefficient of variation computed from 300 simulated exposure maps (solid line) and the coefficient of variation estimated from observed exposure data (dashed line)

computed within 265 m buffer size around each mother's place of residence.

Geostatistical simulations with sGs

From previous geostatistical analysis, we have estimated parameters and selected the form of variogram to model exposure residuals (see "Select a "proper" buffer distance for exposure assignment"). Now, to assess spatial uncertainty of exposures, we incorporated the output from that analysis to run geostatistical simulations. The spatial location of the set of 83 AQIv samples sites (Fig. 1) was taken into account for the simulation step. We ran 300 simulations of AQIv residuals on a 570 × 719 grid (409,830 nodes) with 4.25 m pixel resolution. We used the sequential Gaussian simulation (sGs) algorithm incorporating the Ordinary Kriging estimator to draw AQIv residuals at the visiting nodes. Each exposure map was then obtained by summing trend, obtained with Eq. (6), and the simulated residuals map. The averages AQIv calculated from simulations and observed data are 6.47 and 5.97, with standard deviations of 3.27 and 2.93, respectively.

We performed an assessment to evaluate whether the statistical properties of simulated maps reproduced the statistical properties of observed data. We computed the coefficient of variation (standard deviation/mean) for the 300 simulations and found that the distribution of coefficients of variation from simulations and the observed coefficient variation agree. The agreement can be observed in Fig. 7, where the distribution of coefficients of variation obtained among simulations is centred at 0.50 and the coefficient of variation of observed data is 0.49.

The simulations obtained through conditional sGs incorporate a deterministic part that captured the spatial trend of AQIv, provided by land use spatial distribution. In Fig. 8 are illustrated four possible scenarios of air quality obtained with the sequential algorithm. Most of major patterns observed in land use map are successfully captured, in particular some areas of the road network, while the differences between them are provided by the simulated residuals.

Local uncertainty of association with GWGLM

The GWGLM model presented in Eq. (8) was combined with geostatistical simulations for analysis of spatial uncertainty in associations between AQIv and birth weight. For the generalized linear modelling processes, we estimated the parameters of the models assuming that birth weight variable has gamma distribution with log-link function.

$$\hat{y}_s = exp\left(\hat{\beta}_{0ks} + \hat{\beta}_{1ks}SMO_s + \hat{\beta}_{2ks}BMI_s + \hat{\beta}_{3ks}GES_s + \hat{\phi}_{ks}AQIv_{ks}\right) \quad (8)$$

With each simulation (with $k = 1, \ldots, 300$), we computed personal mean AQIv around each mother's place of residence, fitted the GWGLM models and obtained local parameter estimates for each variable. After repeating this procedure for the 300 simulations, we were able to assess a histogram of the AQIv parameters that reflect the geostatistical uncertainty of associations with birth weight. In Fig. 9, we illustrate these results by computing smooth Kernel density estimates that enable to visualize the underlying distributions of AQIv parameters. Exposure uncertainty at place of residence B is marginally smaller when compared with the one located at A, as the local distribution of AQIv parameter showed a marginally narrower 95% confidence interval (CI95) (CI95 − 0.183; 0.019) when compared with place of residence A (CI95 − 0.187; 0.020). In terms of associations, these results do not show any significant association between birth weight and air quality.

Discussion

The main purpose of this study was to present and illustrate a new approach to assess local distributions of estimated

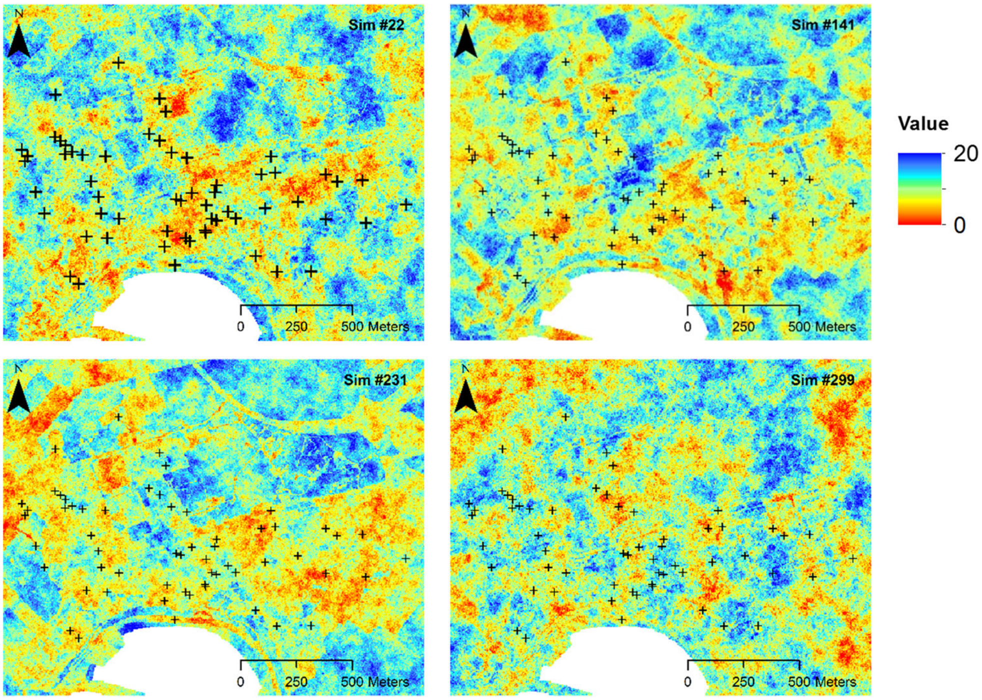


Fig. 8 Examples of four geostatistical simulations of AQIv obtained with Sequential Gaussian Simulation. These exposure maps captured the major patterns of land use and the patchy patterns provided by simulated residuals

parameters measuring associations between air quality and birth weight. Combining GWGLM techniques and geostatistical simulation algorithms, we provided clues on the spatial processes underlying spatial variations in associations between air quality and birth weight. Moreover, we incorporated spatial uncertainty of air quality predictions in the estimation of local parameters. Thus, we consider the use of this approach an additional tool to health analysts in assessment of the local impacts of environmental risk factors in health. This is critical for successful development of the spatial epidemiology methods, since estimates are interpreted in the context of exposure uncertainty varying throughout the spatial domain.

Kernel curves computed from GWGLM models (Fig. 9) overlapped in the large majority of cases and, according to the principle of parsimony, would suggest no gains in the analysis of local associations or, in other words, a global model with no local parameters would be sufficient for the analysis. The most plausible explanation for this result is that the spatial process we modelled is essentially stationary. Nevertheless, we found marginal differences in estimations, illustrated in Fig. 9 with the results for addresses A and B, showing the potential of geostatistical simulation to adjust the spatial uncertainty of personal exposures in GWGLM models.

The map produced with the Regression Kriging technique was not enough to quantify the spatial variability of exposure patterns, since interpolators only create a smooth map revealing the major spatial patterns of the data. So, to model uncertainty of exposures, we turned to geostatistical simulation algorithms (in particular the sequential Gaussian simulation algorithm) which provided a set of maps with spatial patterns of high- and low-exposure values. In the simulation process, we have tried to run 200, 300, 500 and 1000 simulations. We found that after 300 simulations, changes in results were negligible, so we considered 300, a sufficient number of simulations to show the benefits of the proposed method.

Simulated maps presented in Fig. 8 air quality exhibited, more often, lower values near traffic areas. That was not surprising to find, since small-scale variations of air pollution in urban areas are typically from road traffic (de Hoogh et al. 2014) and associations between air pollution and traffic in Sines city have been previously reported (Llop et al. 2017). However, we were able to explicitly separate the spatial variations of air quality in two components: a trend component, modelled with a regression to fit the variations explained by the auxiliary land use variable, and a spatially random component, modelled with a Kriging estimator to fit the unexplained variation or residuals. The contribution of the random component in spatial variation of air quality could be related with the dispersion of gaseous pollutants emitted by traffic or the dispersion of particles (mostly dust) emitted by semi-natural areas.

Selecting a radial buffer distance to derive personal average exposure was not trivial. Selecting too small radial distances would provide more variability in personal average exposures, but these estimates would be unreliable since they would be based on few exposure values. On the other hand, too large distances would add bias to the estimates (towards 0) since personal average exposure at different addresses would tend to get quite similar. Our choice to overcome this problem was to

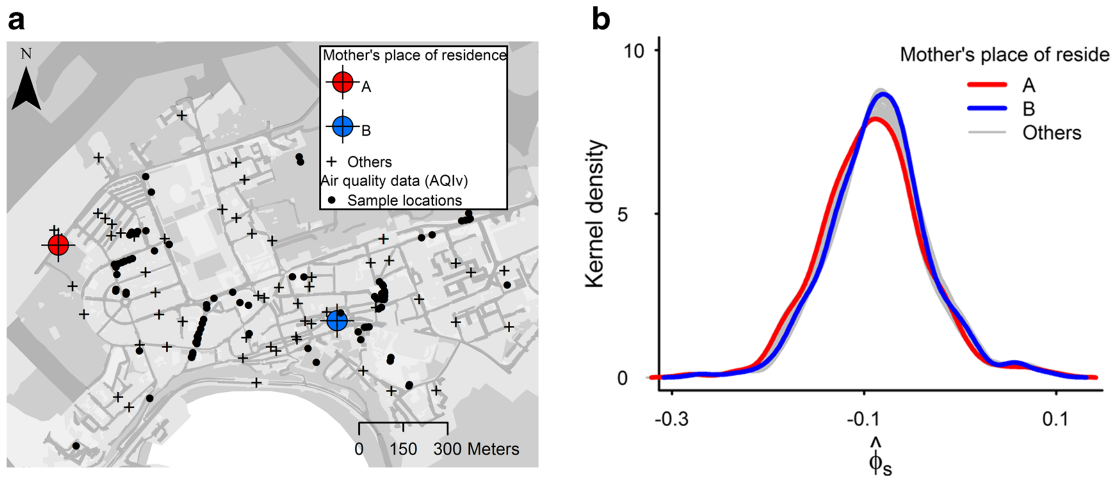


Fig. 9 **a** The city of Sines map showing mothers' places of residence and AQIv sample locations. Circled and highlighted crosses (red and blue) correspond to locations of two mother's place of residence, A and B. **b** The kernel density local distributions of AQIv-estimated parameters. Curves reflect the results of geostatistical uncertainty of exposures. Highlighted (red and blue) are the two curves representing the local distributions of exposure parameters estimated for mother's place of residence, A and B

use a criteria based on a measure of goodness-of-fit. We tested 200 different radial buffer distances up to 1 km to find the lowest AIC as the indicator of the optimal radius distance. Results showed that the optimal distance was reached at 265 m, which is in line with other buffer distances selected for estimation of local variability of air pollutants in urban environments (Kanaroglou et al. 2005; Zandbergen 2007; Pasquier and André 2017).

The correlations found between AQIv and land use combined with Regression Kriging method provided a way to provide exposure maps while tackling changes of air quality at short distances. Nevertheless, the pathway with regard to the estimation process was not perfectly smooth. Obtaining the exposure maps was required in the first place to remove the trend captured by the regression model fitted with the land use predictor (Eq. [6]). However, the percentage of variability in AQIv explained by land use was only moderate (21%). One possible way to improve it could be by adding more auxiliary variables known to be associated with variations in urban air quality, such as traffic volume (Faria et al. 2017) or distance to major roads (Hoffmann et al. 2009). In this case, those variables were not available.

We are also aware that the sampling design of AQIv affected the estimation of experimental semi-variogram of residuals and the specification of the theoretical variogram model. The spatial representativeness of AQIv samples was limited to the samples available, located only in some areas of the city, mostly traffic and residential areas. We would prefer to have a sampling design with locations evenly distributed within the urban area and spatially stratified by land use category. This would provide the possibility of getting an adequate number of samples at various intervals with small lag-distances to estimate the experimental semi-variogram. In our study, we could not consider such small lag-distances, because there would be too few points in some intervals, leading to less reliable variogram estimates and a more complicated spatial model. Therefore, we had to consider a sufficiently large lag-distance (100 m) that allowed us to fit a suitable theoretical model capturing the major spatial patterns of AQIv residuals within the city.

A key issue that may have masked variations in local distributions concerns the accuracy of mothers' places of residence. Due to confidentiality constraints, geocoded place of residence information is available in aggregate form and is assigned to street centrelines. While these geocoded information are reasonably accurate proxies for mothers' places of residence, misplacements of addresses might have introduce bias and error in personal exposure assignments, as underlined by Kirby and co-authors in their recent work (Kirby et al. 2017). For example, different places of residence located in the same street were assigned the same centreline coordinates, meaning that the same exposure values were assigned to those collocated addresses, which may have resulted in a misevaluation of spatial associations.

Conclusions

The new approach presented in this study combines known spatial statistical methods to measure local associations between air quality and health outcomes in urban areas while incorporating their local uncertainty, providing an additional tool to health analysts in assessment of the impacts of air pollution in health. Personal exposure assignments are based on interpolated predictions with Regression Kriging, and their spatial uncertainty is tackled with the sequential Gaussian simulation algorithm. Since the extent of exposure uncertainty varies from place to place, it is essential to take it into account; otherwise, exposure predictions can be misleading. The associations between birth weight and the resulting set of simulated exposures are measured with geographically weighted generalized linear models, in order to derive the uncertainty in the local parameters.

The new approach presented here can be applied in other urban areas where health data are available; the geographical density of air quality data is high and complemented with additional auxiliary variables like land use data or air pollution emissions data.

Funding information This study was funded by the CERENA (strategic project FCT-UID/ECI/04028/2013).

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) Second International Symposium on Information Theory, pp 267–281

Asta J, Erhardt W, Ferretti M, Fornasier F (2002) European guideline for mapping lichen diversity as an indicator of environmental stress. Br Lichen 1–20

Brunsdon C, Singleton AD (2015) Geocomputation: a pratical primer, 1st edn. SAGE Publications Ltd, London

Brunsdon C, Fotheringham S, Charlton M (1998) Geographically weighted regression—modelling spatial non-stationarity. J R Stat Soc Ser D 47:431–443

Canha N, Almeida SM, Freitas MC, Wolterbeek HT (2014) Indoor and outdoor biomonitoring using lichens at urban and rural primary schools. J Toxicol Environ Health A 77:900–915. https://doi.org/10.1080/15287394.2014.911130

Chen VY-J, Yang T-C (2012) SAS macro programs for geographically weighted generalized linear modeling with spatial point data: applications to health research. Comput Methods Prog Biomed 107:262–273. https://doi.org/10.1016/j.cmpb.2011.10.006

Conti ME, Cecchetti G (2001) Biological monitoring: lichens as bioindicators of air pollution assessment—a review. Environ Pollut 114:471–492

CTT Correios (2010) Geoindex standard. http://geoindex.ctt.pt/. Accessed 2 May 2011

da Silva AR, Rodrigues TCV (2013) Geographically weighted negative binomial regression—incorporating overdispersion. Stat Comput 24:769–783. https://doi.org/10.1007/s11222-013-9401-9

de Hoogh K, Korek M, Vienneau D, Keuken M, Kukkonen J, Nieuwenhuijsen MJ, Badaloni C, Beelen R, Bolignano A, Cesaroni G, Pradas MC, Cyrys J, Douros J, Eeftens M, Forastiere F, Forsberg B, Fuks K, Gehring U, Gryparis A, Gulliver J, Hansell AL, Hoffmann B, Johansson C, Jonkers S, Kangas L, Katsouyanni K, Künzli N, Lanki T, Memmesheimer M, Moussiopoulos N, Modig L, Pershagen G, Probst-Hensch N, Schindler C, Schikowski T, Sugiri D, Teixidó O, Tsai MY, Yli-Tuomi T, Brunekreef B, Hoek G, Bellander T (2014) Comparing land use regression and dispersion modelling to assess residential exposure to ambient air pollution for epidemiological studies. Environ Int 73:382–392. https://doi.org/10.1016/j.envint.2014.08.011

Environmental Systems Research Institute (2006) ArcGIS. ESRI Inc 2011

Faria MV, Duarte GO, Baptista PC, Farias TL (2017) Scenario-based analysis of traffic-related PM2.5 concentration: Lisbon case study. Environ Sci Pollut Res 24:12026–12037. https://doi.org/10.1007/s11356-015-5556-6

Fei J-C, Min X-B, Wang Z-X, Pang ZH, Liang YJ, Ke Y (2017) Health and ecological risk assessment of heavy metals pollution in an antimony mining region: a case study from South China. Environ Sci Pollut Res 24:27573–27586. https://doi.org/10.1007/s11356-017-0310-x

Fortin M-J, James PMA, MacKenzie A et al (2012) Spatial statistics, spatial regression, and graph theory in ecology. Spatial Statistics 1: 100–109. https://doi.org/10.1016/j.spasta.2012.02.004

Fotheringham AS, Charlton ME, Brunsdon C (1998) Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. Environ Plan A 30:1905–1927. https://doi.org/10.1068/a301905

Garty J (1993) Lichens as biomonitors of heavy metal pollution. In: Weinheim MB (ed) Plants as biomonitors: indicators for heavy metals in the terrestrial environment. pp 193–257

Glinianaia SV, Rankin J, Bell R, Pless-Mulloli T, Howel D (2004) Particulate air pollution and fetal health: a systematic review of the epidemiologic evidence. Epidemiology 15:36–45. https://doi.org/10.1097/01.ede.0000101023.41844.ac

Goldman GT, Mulholland JA, Russell AG, Gass K, Strickland MJ, Tolbert PE (2012) Characterization of ambient air pollution measurement error in a time-series health study using a geostatistical simulation approach. Atmos Environ 57:101–108. https://doi.org/10.1016/j.atmosenv.2012.04.045

Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press

Goovaerts P (2009) Medical geography: a promising field of application for geostatistics. Math Geosci 41:243–264

Goovaerts P, Jacquez GM, Greiling D (2005) Exploring scale-dependent correlations between cancer mortality rates using factorial kriging and population-weighted semivariograms. Geogr Anal 37:152–182. https://doi.org/10.1111/j.1538-4632.2005.00634.x

Gryparis A, Paciorek CJ, Zeka A, Schwartz J, Coull BA (2009) Measurement error caused by spatial misalignment in environmental epidemiology. Biostatistics 10:258–274. https://doi.org/10.1093/biostatistics/kxn033

Hampton KH, Serre ML, Gesink DC, Pilcher CD, Miller WC (2011) Adjusting for sampling variability in sparse data: geostatistical approaches to disease mapping. Int J Health Geogr 10:54. https://doi.org/10.1186/1476-072X-10-54

Harris P, Fotheringham AS, Crespo R, Charlton M (2010) The use of geographically weighted regression for spatial prediction: an evaluation of models using simulated data sets. Math Geosci 42:657–680. https://doi.org/10.1007/s11004-010-9284-7

Hengl T (2009) A practical guide to geostatistical mapping, 2nd edn. Office for Official Publications of the European Communities, Luxembourg

Hengl T, Heuvelink GBM, Rossiter DG (2007) About regression-kriging: from equations to case studies. Comput Geosci 33:1301–1315. https://doi.org/10.1016/j.cageo.2007.05.001

Hijmans RJ, van Etten J (2012) Raster: geographic analysis and modeling with raster data. R package version 2.4.20

Hoffmann B, Moebus S, Dragano N, Stang A, Möhlenkamp S, Schmermund A, Memmesheimer M, Bröcker-Preuss M, Mann K, Erbel R, Jöckel KH (2009) Chronic residential exposure to particulate matter air pollution and systemic inflammatory markers. Environ Health Perspect 117:1302–1308. https://doi.org/10.1289/ehp.0800362

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. https://doi.org/10.1038/sj.jea.7500388

Jin Y, Ge Y, Wang J, Chen Y, Heuvelink GBM, Atkinson PM (2018) Downscaling AMSR-2 soil moisture data with geographically weighted area-to-area regression kriging. IEEE Trans Geosci

Remote Sens 56:2362–2376. https://doi.org/10.1109/TGRS.2017.2778420

Kalkbrenner AE, Windham GC, Serre ML, Akita Y, Wang X, Hoffman K, Thayer BP, Daniels JL (2015) Particulate matter exposure, prenatal and postnatal windows of susceptibility, and autism spectrum disorders. Epidemiology 26:30–42. https://doi.org/10.1097/EDE.0000000000000173

Kanaroglou P, Jerrett M, Morrison J et al (2005) Establishing an air pollution monitoring network for intra-urban population exposure assessment: a location-allocation approach. Atmos Environ 39: 2399–2409. https://doi.org/10.1016/j.atmosenv.2004.06.049

Kirby RS, Delmelle E, Eberth JM (2017) Advances in spatial epidemiology and geographic information systems. Ann Epidemiol 27:1–9. https://doi.org/10.1016/j.annepidem.2016.12.001

Kitanidis PK (1993) Generalized covariance functions in estimation. Math Geol 25:525–540. https://doi.org/10.1007/BF00890244

Kramer MS (2003) The epidemiology of adverse pregnancy outcomes: an overview. J Nutr 133:1592S–1596S

Kyriakidis P (2004) A geostatistical framework for area to point spatial interpolation. Geogr Anal 36:259–289

Lawson A, Banerjee S, Haining R, Ugarte L (2016) Handbook of spatial epidemiology. CRC Press-Taylor & Francis Group

Lee S, Serre ML, Van DA et al (2012) Comparison of geostatistical interpolation and remote sensing techniques for estimating long-term exposure to ambient PM2.5 concentrations across the continental United States. Environ Health Perspect 120:1727–1733

Li Z, Wang W, Liu P, Bigham JM, Ragland DR (2013) Using geographically weighted Poisson regression for county-level crash modeling in California. Saf Sci 58:89–97. https://doi.org/10.1016/j.ssci2013.04.005

Llop E, Pinho P, Matos P, Pereira MJ, Branquinho C (2012) The use of lichen functional groups as indicators of air quality in a Mediterranean urban environment. Ecol Indic 13:215–221. https://doi.org/10.1016/j.ecolind.2011.06.005

Llop E, Pinho P, Ribeiro MC, Pereira MJ, Branquinho C (2017) Traffic represents the main source of pollution in small Mediterranean urban areas as seen by lichen functional groups. Environ Sci Pollut Res 24:1–10. https://doi.org/10.1007/s11356-017-8598-0

Loppi S, Ivanov D, Boccardi R (2002) Biodiversity of epiphytic lichens and air pollution in the town of Siena (Central Italy). Environ Pollut 116:123–128

Minasny B, McBratney AB (2007) Spatial prediction of soil properties using EBLUP with the Matérn covariance function. Geoderma 140: 324–336. https://doi.org/10.1016/j.geoderma.2007.04.028

Munzi S, Correia O, Silva P, Lopes N, Freitas C, Branquinho C, Pinho P (2014) Lichens as ecological indicators in urban areas: beyond the effects of pollutants. J Appl Ecol 51:1750–1757. https://doi.org/10.1111/1365-2664.12304

Nakaya T, Fotheringham S, Brunsdon C, Charlton M (2005) Geographically weighted Poisson regression for disease association mapping. Stat Med 24:2695–2717. https://doi.org/10.1002/sim.2129

Nelder JA, Wedderburn RW (1972) Generalized linear models. J R Stat Soc 135:370–384

Neuman SP, Jacobson EA (1984) Analysis of nonintrinsic spatial variability by residual kriging with application to regional groundwater levels. J Int Assoc Math Geol 16:499–521. https://doi.org/10.1007/BF01886329

Odeh IOA, Mcbratney AB, Chittleborough DJ (1994) Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63:197–214

Paoli L, Munzi S, Guttová A, et al (2015) Lichens as suitable indicators of the biological effects of atmospheric pollutants around a municipal solid waste incinerator (S Italy). Ecol Indic 52:362–370. https://doi.org/10.1016/j.ecolind.2014.12.018

Pasquier A, André M (2017) Considering criteria related to spatial variabilities for the assessment of air pollution from traffic. Transportation Research Procedia 25:3354–3369. https://doi.org/10.1016/j.trpro.2017.05.210

Pebesma EJ (2004) Multivariable geostatistics in S: the gstat package. Comput Geosci 30:683–691

Pinho P, Augusto S, Branquinho C, Bio A, Pereira MJ, Soares A, Catarino F (2004) Mapping lichen diversity as a first step for air quality assessment. J Atmos Chem 49:377–389. https://doi.org/10.1007/s10874-004-1253-4

Pinho P, Augusto S, Máguas C, Pereira MJ, Soares A, Branquinho C (2008a) Impact of neighbourhood land-cover in epiphytic lichen diversity: analysis of multiple factors working at different spatial scales. Environ Pollut 151:414–422. https://doi.org/10.1016/j.envpol.2007.06.015

Pinho P, Augusto S, Martins-Loução M et al (2008b) Causes of change in nitrophytic and oligotrophic lichen species in a Mediterranean climate: impact of land cover and atmospheric pollutants. Environ Pollut 154:380–389. https://doi.org/10.1016/j.envpol.2007.11.028

Pinho P, Llop E, Ribeiro MC, Cruz C, Soares A, Pereira MJ, Branquinho C (2014) Tools for determining critical levels of atmospheric ammonia under the influence of multiple disturbances. Environ Pollut 188:88–93. https://doi.org/10.1016/j.envpol.2014.01.024

R Core Team (2014) R: a language and environment for statistical computing.

Ribeiro MC, Llop E, Branquinho C et al (2012) A retrospective cohort study to assess the association between outdoor air quality and low birth weight. Arch Dis Child 97:A283. https://doi.org/10.1136/archdischild-2012-302724.0990

Ribeiro MC, Pinho P, Llop E et al (2016) Geostatistical uncertainty of assessing air quality using high-spatial-resolution lichen data: a health study in the urban area of Sines, Portugal. Sci Total Environ 562:740–750

Rose CI, Hawksworth DL (1981) Lichen recolonization in London's cleaner air. Nature 289:289–292

Waller LA, Gotway CA (2004) Applied spatial statistics for public health data. Wiley, Hoboken

Waller LA, Zhu L, Gotway CA, Gorman DM, Gruenewald PJ (2007) Quantifying geographic variations in associations between alcohol distribution and violence: a comparison of geographically weighted regression and spatially varying coefficient models. Stoch Environ Res Risk A 21:573–588. https://doi.org/10.1007/s00477-007-0139-9

Wolterbeek HT, Garty J, Reis MA, Freitas MC (2003) Chapter 11 Biomonitors in use: lichens and metal air pollution. In: Markert BA, Breure AM, HGZBT-TM and other C in the E (eds) Bioindicators & biomonitors principles, concepts and applications. Elsevier, pp 377–419

Young LJ, Gotway CA, Yang J, Kearney G, DuClos C (2008) Assessing the association between environmental impacts and health outcomes: a case study from Florida. Stat Med 27:3998–4015. https://doi.org/10.1002/sim

Zandbergen PA (2007) Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. BMC Public Health 7:1–13. https://doi.org/10.1186/1471-2458-7-37