Environmental Modelling & Software 105 (2018) 132–152

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A polynomial approximation of the traffic contributions for kriging-based interpolation of urban air quality model

Maxime Beauchamp [a, b, *], Laure Malherbe [b], Chantal de Fouquet [a], Laurent Létinois [b], Frédéric Tognet [b]

[a] Mines ParisTech, Géosciences, Equipe géostatistique 35 rue Saint Honoré, 77305, Fontainebleau, France

[b] Institut National de l'Environnement Industriel et des Risques (INERIS) Direction des risques chroniques, Parc Technologique Alata, 60550, Verneuil-en-Halatte, France

ARTICLE INFO

Article history:
Received 1 September 2017
Received in revised form
26 March 2018
Accepted 30 March 2018

Keywords:
Geostatistics
Kriging
Air pollution
Urban scale
Series expansion drift

ABSTRACT

The European directives for ambient air quality require to assess areas where air pollutant concentrations exceed a regulatory threshold. As the spatial distribution of the pollutant is not exactly known, deterministic atmospheric dispersion models are commonly used to supplement the observational network. To reduce the computational time, the simulations are made on irregular grids, especially in urban areas where the grid is refined close to the roads. An interpolation method is then necessary to map the dispersion model at any location. We propose a new geostatistical approach based on kriging with external drift to distinguish the information along and across the roads. An exponential function is introduced to describe the decrease of the concentrations across the roads. Its series expansion is used to build a set of polynomial auxiliary predictors with unknown coefficients. This framework leads to a drift that is more generic and flexible in the kriging system.

1. Introduction

The European legislation (Directive, 2008/50/EC, 2008) on ambient air quality defines some environmental objectives, expressed as annual/daily/hourly averaged concentrations or maximum numbers of days/hours in exceedance depending on the pollutant and the time resolution.

When one or several thresholds are exceeded, the member states have to delineate the spatial extent and the population exposed to these exceedances. A lot of studies estimate the exposure to exceedances by crossing a map of concentrations with a static map of the population, i.e. counted at their place of residence, which is also the norm in the European regulation. For this reason, this work falls within this framework and is focused on the improvement of the concentration maps, without worrying about the issue of the cross combination with the population. However, a growing umber of studies have recently estimated the same exposure, but taking into account the dynamic aspect of the population that moves and sometimes works away from the residential areas. These works are mostly built on activity-based exposure models (Beckx et al., 2009; Hatzopoulou and Miller, 2010; Jantunen et al., 1999; WHO, 2005) or simulations of the daily movement of the population, using for instance mobile phone tracking (Liu et al., 2013; Gariazzo et al., 2016).

When occurring in urban areas - and it is mostly the case for NO_2 and PM_{10} - the exceedances are usually assessed by using urban scale dispersion models. Let us note $\mathscr{M}$ this type of model. To limit the computational costs, a widespread practice is to calculate the concentrations in a two-step procedure: first, the concentrations $Z(\mathbf{x}_\alpha)$ are simulated on the irregular grid $\{\mathbf{x}_\alpha\}_{\alpha=1,\cdots,p}$ with a coarse regular resolution in background areas and a higher adaptative resolution close to the roads (see e.g. Leelőssy et al., 2014); next, they are interpolated on a regular grid with high resolution over the whole domain of simulation $\mathscr{D}$. Regarding the observational data $T(\mathbf{x}_{\alpha'})$ (fixed monitoring sites, passive sampling measurements) available at locations $\{\mathbf{x}_{\alpha'}\}_{\alpha'=1,\cdots,d}$ $(d<p)$, they are seldom introduced in a data assimilation framework (Tilloy et al., 2013) to reduce the errors made by the dispersion model $\mathscr{M}$. A kriging-based combination of simulated data and passive sampling measurements can also be performed, see e.g. ASPA (2014) technical report. In that case, the frequency and the extent of the sampling

* Corresponding author. Mines ParisTech, Géosciences, Equipe géostatistique 35 rue Saint Honoré, 77305, Fontainebleau, France.
E-mail address: maxime.beauchamp@mines-paristech.fr (M. Beauchamp).

https://doi.org/10.1016/j.envsoft.2018.03.033
1364-8152/

campaigns play an important part in the quality of the results.

As a consequence, the final mapping of $Z(\mathbf{x})$ over $\mathscr{D}$ mostly depends on three criteria: the number p of $\mathbf{x}_\alpha$ in the simulation (also denoted as the receptor points), their locations and the precision of the interpolation technique. In air quality, the simulated concentrations are usually interpolated by linear methods implemented in post-processing tools, such as Climate Data Operators (CDO), Netcdf Operators (NCO), see Zender (2016). In the urban configuration, there is an additional difficulty because the concentrations come from both traffic-related and background sources of pollution. Thus, according to the spatial distribution of the receptor points, the usual interpolation methods may lead to some artifacts.

In this paper, an original adaptation of kriging-based interpolation is proposed for the simulations made by an urban-scale dispersion model. First, a review of the standard existing approaches is presented. Then, an external drift modelling is introduced to assess the behaviour of the concentrations across the roads. The concentration $Z(\mathbf{x})$ simulated by the model is seen as a stochastic process explicitly decomposed into a deterministic part $m_Z(\mathbf{x})$ and a zero mean second-order stationary random field $W(\mathbf{x})$. The series expansion of the exponential function is used to linearize the expression of the deterministic part, and thus stick to the usual kriging with external drift framework (Chiles and Delfiner, 2012).

The three next sections are dedicated to an application of the methodology on the French city of Orléans in 2010 (source LIG'AIR): the NO_2 annual mean simulated by the ADMS-Urban air quality model is considered. In Section 3, the dataset for both $Z(\mathbf{x}_\alpha)$ and $T(\mathbf{x}_{\alpha'})$ are presented, as well as the traffic emissions used to build the predictors of $m(\mathbf{x})$. Section 4 compares the results of a selection of standard interpolators and the exponential drift framework in terms of mapping and cross-validation scores. A detailed study concerning the advantages of the series expansion is also provided with additional guidelines regarding the truncation of the series and the minimal number of receptor points to use. Last, a discussion is given to comment some key points of the methodology, and in particular the consequences of the polynomial approximation on the coefficients of the drift estimated by kriging. Section 5 is focused on a kriging where the interpolation $Z^{\mathrm{K}}(\mathbf{x})$ of the simulation dataset is used as an external drift to improve the estimation $T^{\mathrm{K}}(\mathbf{x})$ of the data collected from a passive sampling campaign. Finally, Section 6 is dedicated to software availability, just before the conclusion of the paper.

2. Materials

2.1. Overview of commonly used approaches

The linear interpolation is widely used to provide air quality model outputs anywhere in space, whatever the scale the model deals with. It is considered as valid when the variations between the $\mathbf{x}_\alpha$'s are low or when the distances between the $\mathbf{x}_\alpha$'s are small: a few meters in traffic-related configuration but up to a few kilometers for low background concentrations. More recently, Fortin et al. (2012) developed a method based on the Delaunay triangulation; see also Lixin et al. (2011) for further applications of the Delaunay triangulation to air quality interpolation. Assuming that the pollutant concentration along the triangle edges varies linearly, polygons linking all the positions where the concentration is equal to a given value are defined. Last, the (ordinary) kriging (Chiles and Delfiner, 2012) is a linear combination of the data with optimal weights satisfying unbiasedness constraint $\mathrm{E}[Z^{\mathrm{K}}(\mathbf{x}) - Z(\mathbf{x})] = 0$ and optimality for the variance of the error $\mathrm{Var}[Z^{\mathrm{K}}(\mathbf{x}) - Z(\mathbf{x})]$.

The statistical, geostatistical and GIS software make it possible to implement linear or Delaunay interpolation and ordinary kriging with limited effort. In addition, an evaluation study (Beauchamp et al., 2016) using 5 different French urban datasets (Bourges, Nantes, Niort, Orléans, Reims and Tours) has shown the relative efficiency of the linear and Delaunay interpolation compared to ordinary kriging. This can be explained by the non-stationarity related to the multi-source origin of the concentrations (either influenced by the traffic or only reflecting the background pollution). Ordinary kriging is not able to account for it when dealing with few data. Jeannee and Lemarchand (2012) handled this issue by considering locally varying anisotropies, but other kriging options are available, in particular universal kriging based approaches (see Section 2.2), in which the non stationarities are taken into account through the modelling of the deterministic part $m(\mathbf{x})$, also called the drift, of the random process. Other interpolators (inverse distance, nearest neighbour, Akima's interpolator) will not be considered in this paper because they are too simplistic and only relevant for smooth concentrations fields.

More sophisticated estimation methods are not addressed in this paper. They could be useful if hourly/daily simulation outputs are considered instead of the annual-averaged values. Among them, the Bayesian Maximum Entropy methodology (BME) that is a superset of classical geostatistics interpolators has already been used in air pollution and atmospheric studies (Christakos and Li, 1998; Christakos et al., 2004; Yu et al., 2016, 2011). The growing popular estimation approach through the SPDE (Stochastic Partial Differential Equation) framework (Lindgren et al., 2011) must also be mentioned. An interesting application to space-time PM_{10} pollution data is made in Cameletti et al. (2012) using the INLA (Integrated Nested Laplace Approximation) computational approach. A basic introduction to these methods is given in Appendix A.

2.2. Traffic-related external drift modelling

In urban areas, the pollution at a location $\mathbf{x}$ can be considered as the sum of a background component and a traffic-related contribution due to the emissions of the roads in the close neighbourhood of $\mathbf{x}$. To simplify, these two components will be supposed spatially independent, see for instance Font et al. (2014). Let denote $Z(\mathbf{x})$ the concentration simulated by the model $\mathscr{M}$ at $\mathbf{x}$, not to be mistaken with the true value of the concentration, denoted as $T(\mathbf{x})$ and introduced later in Section 2.5. $Y(\mathbf{x})$ is the random function related to the background feature and $S(\mathbf{x})$ is the random function that deals with the pollution increment related to traffic emissions:

$$Z(\mathbf{x}) = Y(\mathbf{x}) + S(\mathbf{x}) \tag{1}$$

The independence between $Y(\mathbf{x})$ and $S(\mathbf{x})$ can be checked out by averaging the values $S(\mathbf{x}_\alpha)$ at locations $\mathbf{x}_\alpha$'s considered as traffic-influenced per urban background concentration classes $(y_1, \cdots, y_P)$. $Y(\mathbf{x}_\alpha)$ is not known but may be obtained either by roughly removing the traffic-related sources in the air quality simulation or by kriging the background data, i.e the receptor points located far enough from the road network. In the literature, more sophisticated models can also be used to estimate this quantity, see e.g. Pournazeri et al. (2014). The appropriate definition of "far enough" depends on the corresponding environments, so that the model in Eq. (1) is correct. According to the literature (see for example Baldwin et al., 2015; Zou et al., 2006; Gilbert et al., 2003; Roorde-Knape et al., 1999) and repeated results from regular campaigns in France, the direct impact of road traffic can be considered insignificant 400 m away from the road: thus, it possibly holds as true for similar environments elsewhere. As a consequence, when $\mathbf{x}$ is located more than 400 m away from the road, $S(\mathbf{x})$ is neglected.