Spatial Statistics 22 (2017) 18–46

Contents lists available at ScienceDirect

Spatial Statistics

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

Dealing with non-stationarity through explanatory variables in kriging-based air quality maps

Maxime Beauchamp [a,b,*], Chantal de Fouquet [a], Laure Malherbe [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 23 September 2016
Accepted 12 August 2017
Available online 24 August 2017

Keywords:
Geostatistics
Kriging
Air pollution
Non-stationarity
Auxiliary variables
Composite space

ABSTRACT

Kriging-based estimation is used to interpolate air quality data on large scale regular grids from observations at monitoring sites. However, the sources of pollution and therefore the data typology are highly variable (rural background, urban and suburban background, traffic-related pollution), invalidating the assumption of stationarity. On a large scale (regional, national, continental), the site typology can be described by an auxiliary variable, typically the pollutant emission in a fixed radius around the site. In particular the quantity of NO_x emissions,introduced in the kriging as external drift, is a useful variable to account for human activity and related pollution in urban areas. In this paper a new formulation of this question was examined by introducing a new space 2D×V where $\nu \in V$ is the explanatory variable of the concentration. The way of computing the variogram is addressed with two different options. First, after rescaling the third dimension to $\mathbb{R}^2$, the classic Euclidean distance is used in the newly built space through a metric variogram. Secondly, both the Euclidean distance in the 2D space and the absolute deviation in ν are used within a product-sum model to deal with the underlying complex anisotropy. The results obtained by ordinary kriging with metric or product-sum variogram in the composite space are compared to ordinary kriging and kriging with ν as external drift using a purely spatial variogram. The use of Chemistry-Transport Models (CTM) is also considered

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

http://dx.doi.org/10.1016/j.spasta.2017.08.003

as a potential improvement is also approached and the way of introducing this type of data in the kriging is discussed. Finally, in the light of the results provided by this set of estimators applied to NO_2 datasets, kriging with external drift in the 2D space remains the most convenient way on average to deal with non-stationarity and to include all the explanatory informations in the estimator. Nevertheless, the product-sum variogram model used in the composite space 2D×V sometimes exhibits a better performance for some data configurations, suggesting that there is no single best method to account for non-stationarity in air quality processes.

1. Introduction

Air quality data have been used for many years to produce pollution maps (Wackernagel et al., 2004; Horálek. et al., 2013). However, the monitoring network is not sufficient to produce reliable maps since the pollution level at measurement sites can be explained by different sources such as traffic-related pollution, domestic heating, industry, etc. Geostatistics makes it possible to combine different types of data and take the lack of stationarity into account through the use of appropriate explanatory variables. Then, thanks to the development of environmental databases in the last decade, mapping techniques now usually aim at mixing pollution levels observed at monitoring sites with high resolution variables (emissions, land use, population, etc.). An external drift modelling, see e.g. Wackernagel (1995), is a suitable way to do so. A common approach is to aggregate in a fixed radius the variable targeted as potentially explanatory around the monitoring sites to build a variable well correlated with the pollutant concentration (Malherbe et al., 2008).

In the external drift modelling, a linear deterministic link with unknown coefficients is introduced between the variable to estimate and the auxiliary variable. As the coefficients of the (linear) drift are not known, this relation can be locally adjusted using a moving neighbourhood. But if there are too few monitoring sites in the neighbourhood, the relation can be badly estimated. Moreover, if the drift takes very low values, some negative estimations may occur. The aim of this paper is to examine different ways of considering an explanatory variable without making the single assumption of a linear relation between the data and the explanatory variables. The idea is to use the explanatory variable $\nu \in V$ as a new dimension of a composite space 2D× V in which the sample covariance function is computed either using a 3D Euclidean distance with the appropriate rescaling or addressing the question of anisotropy by separately dealing with spatial distance and absolute deviation in ν.

The first approach is theoretically less flexible because of its naive conception: the Euclidean distance is computed in the composite space 2D× V by considering the three dimensions in the same way. It could be improved with a zonal anisotropy but a pure metric model is rather restricted since both marginal variograms will be of the same type and have the same sill (if the variogram model is bounded).

The second approach may offer more possibilities. In a similar way to what is done in space–time applications, the covariance is non separable. A product-sum model (De Cesare et al., 1997; De Iaco et al., 2001) is chosen to fit an authorized covariance function. A preliminary fit of the marginal variograms can be used to build the complete model.

Chemistry-transport models (CTMs) are increasingly used to complete the information on air quality delivered by monitoring sites with concentration fields computed on a regional scale. Their capacity to reproduce the spatial and temporal variations of concentrations is regularly assessed through evaluation exercises, see e.g. Bessagnet et al. (2014). It is possible to build a linear model of coregionalization (LMC) between the observations and the CTM to cokrige the concentration of the pollutant but the most popular solution is to interpolate CTMs output on the grid and use it as an external drift (Malherbe et al., 2011). However, because of computational issues, the resolution of CTMs (commonly 5–25 km) is usually not high enough to finely reproduce concentration gradients due to local sources and specific dispersion conditions. The idea is therefore to both use information