Science of the Total Environment 598 (2017) 780–788

Contents lists available at ScienceDirect

Science of the Total Environment

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

Improving modeled air pollution concentration maps by residual interpolation

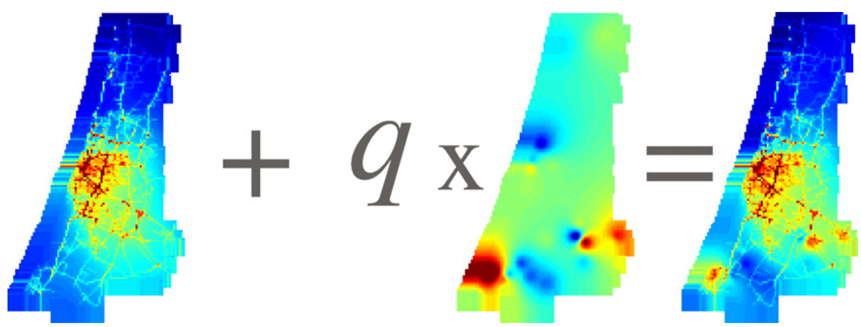
Yuval [a], Ilan Levy [b], David M. Broday [a,*]

[a] Faculty of Civil and Environmental Engineering, Technion, Israel Institute of Technology, Haifa, Israel
[b] Division of Air Quality and Climate Change, Ministry of Environmental Protection, 125 Menachem Begin road, Tel Aviv 61071, Israel

HIGHLIGHTS

- Interpolation of model residuals can improve the accuracy of air pollution maps.
- The residual map is useful for detecting regional bias.
- Using a cross-validation process prevents fitting random observation errors.
- Correction should not be applied if the residuals are randomly distributed in space.
- The methodology can accommodate optimization of different performance measures.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 15 December 2016
Received in revised form 17 March 2017
Accepted 15 April 2017
Available online 28 April 2017

Editor: D. Barcelo

Keywords:
Air pollution concentration maps
Cross-validation
Exposure
Residuals
Spatial autocorrelation

ABSTRACT

Models that are used to map air pollutant concentrations are not free of errors. A possible approach for improving the final concentration map is to interpolate the residuals of the initial model concentration estimates. Due to the possible spatial autocorrelation of the residuals of the initial model estimates, Bayesian inference schemes were suggested for this task, since they can correctly adjust the level of fitting of the residuals to the random measurement errors. However, the complexity of Bayesian methods often discourages their use. Here, we present an alternative and simpler approach, using a leave-one-out cross-validation to determine the optimal level of fitting of the residual correction. We show that the optimal correction level is related to the extent of the spatial autocorrelation of the cross-validated residuals. Namely, when the residuals are not autocorrelated residual correction is unnecessary, and if employed may actually degrade the quality of the final concentration map. Moreover, our approach enables to optimize the residual correction based on different target performance measures, with a possibly different optimal correction depending on the performance measure used. Hence, different target performance measures can be chosen to fit best the specific application of interest. The method is demonstrated using output of three different models used for estimating NO_x and NO_2 concentrations over Israel. We show that our approach can be used as an exploratory step, for assessing the potential benefit of residual correction, and as a simple alternative to Bayesian schemes.

1. Introduction

Air pollution estimates are used for regulatory purposes, air resources management, as well as for assessing associations between them and different health outcomes. The power of the statistical

* Corresponding author at: Faculty of Civil and Environmental Engineering, Technion I.I.T., Haifa 32000, Israel.
E-mail address: dbroday@tx.technion.ac.il (D.M. Broday).

http://dx.doi.org/10.1016/j.scitotenv.2017.04.117

analysis, the strength of the obtained associations and the width of the confidence intervals depend, among other things, on the accuracy of the exposure estimates and therefore on the underlying concentration maps (Thomas et al., 1993; Armstrong, 1998; Zeger et al., 2000; Basagana et al., 2013). For example, exposure errors of the classical type tend to push the recovered association towards the null (Thomas et al., 1993), and inadequate accuracy of the exposure estimates was suggested to be the key reason for the relatively weak associations found between air pollution and different health outcomes (Zeger et al., 2000).

One approach to deal with exposure errors is to account for their effect within the epidemiological model that looks for an association between the exposure and the health outcome. For example, Gryparis et al. (2009) accounted for the effect of spatial misalignment on risk estimates using generalized linear model framework. Molitor et al. (2006) and Eitan et al. (2010) accounted for random exposure errors using hierarchical Bayesian models, and Molitor et al. (2007) extended it for cases with spatially autocorrelated exposure errors. Madsen et al. (2008) developed a parametric bootstrap approach that jointly fits the exposure and health models, and Szpiro et al. (2011) improved the computational efficiency of the method and accounted for the effect of spatially correlated errors. However, in general, adjusting epidemiological models to exposure errors is difficult and computationally challenging, while not always assuring full account for their effects (Szpiro et al., 2011; Basagana et al., 2013).

A different approach for improving the exposure estimates is by minimizing the errors in the pollutant concentration estimates. Refining the pollutant estimates, from which exposure is derived, is undoubtedly the most desired route. Yet, until models that predict pollutant dispersion and fate in the atmosphere are improved, which often takes some time, an alternative approach is to use the information that is contained in the residuals of the pollutant concentration maps in the locations where observations exist. Overall, residuals can result from two general sources: measurement errors due to instrument failure and/or calibration errors, and inaccuracy/bias in the estimation process, i.e. modelling errors. While modelling errors are expected to result in a coherent autocorrelated residual pattern, the distribution of measurement errors is oftentimes random. Our goal is thus to correct the coherent spatial pattern of the residuals but to avoid fitting the random measurement errors. This challenge, sometimes referred to as the bias/variance dilemma (Geman et al., 1992), can be addressed using either Bayesian methods or cross-validation techniques.

Best fitting the spatial features of the coherent errors can be obtained using geostatistical interpolations. Indeed, since the early 2000s researchers use corrections based on geospatial interpolations of the residuals for improving the estimated pollutant concentrations (Blond et al., 2003; Wackernagel et al., 2004; van de Kassteele et al., 2009). For example, Hogrefe et al. (2009) corrected concentration maps produced by the photochemical Community Multiscale Air Quality Model (CMAQ; Byun and Schere, 2006) using a one stage Inverse Distance Weighting (IDW) (Isaaks and Srivastava, 1989) to map the ratio of the observed-to-modeled concentrations. Bergen et al. (2013) and Mercer et al. (2011) used a two-stage approach for improving estimates of exposure to $PM_{2.5}$ and NO_x. In the primary stage they used a Land Use Regression (LUR) model, based on an optimally-sized set of covariates, while the second stage involved refining of the exposure estimates by a kriging interpolation of the concentration residuals (i.e. presumably with the nugget set to fit the residuals).

Kriging interpolation provides the best linear unbiased prediction based on the spatial covariance structure of the residuals (Cressie, 1993). The latter is described by a variogram, whose mathematical form relies on rather subjective decisions regarding the number of distance bins (lag sizes between paired observations) and the function that describes and smooths the histogram of the observations (by choosing its functional shape and parameters: nugget, sill, and range). Yet, it is very challenging to find the optimal variogram parameters, especially when tens of thousands of maps are required and while allowing for possible temporal variation of the variogram model parameters. In particular, although applying cross-validation is conceptually simple, finding the optimal parameters of a kriging interpolation by a cross-validation process is computationally prohibitive. Hence, implementation of geostatistical interpolations for residual correction has been carried out using Bayesian inference schemes (van de Kassteele et al., 2009; Beckerman et al., 2013; Akita et al., 2014). Bayesian Maximum Entropy (BME) interpolation (Christakos et al., 2001) can combine different information sources, including concentration residuals, and improve the original concentration estimations. For example, Beckerman et al. (2013) used the BME to correct LUR model monthly $PM_{2.5}$ residuals across the contiguous USA, and Akita et al. (2014) used BME to integrate monitoring data and results of both a LUR and a chemical transport models for estimating NO_2 concentrations in Catalunya, SP. However, Bayesian schemes are rather complex and may be computationally intensive (Blangiardo et al., 2013). Moreover, poor selection of the Bayesian prior probability distributions may result in an inferior residual correction.

In this work, we demonstrate application of a leave-one-out cross-validation (LOOCV) procedure to optimally correct pollution concentration maps, using simple interpolation of the concentration residuals. We demonstrate that the residuals can serve for examining the error structure of the original model, and that the LOOCV approach can be used as an exploratory step for assessing the potential benefits of performing a more comprehensive residual correction, as well as a simple alternative for complex schemes like Bayesian methods. We demonstrate this approach by correcting estimated ambient NO_x and NO_2 concentrations obtained using three different models that are currently in use for air quality management and research in Israel.

2. Methods

2.1. Study area

The study domains of the three models used for demonstrating the residual correction method are not identical, although they overlap over the central coastal strip of Israel (Fig. 1). Most of the population in this area, as well as the local emission sources, reside in a narrow strip of width < 35 km along the Mediterranean coastline. Ambient NO_x concentrations in this area are mostly attributed to traffic, with minor contributions from heavy industry and power plants. Israel's littoral geography results in a vigorous land-sea breeze cycle that maintains good ventilation of the lower atmosphere. Since the daily mixing layer is very rarely thinner than 500 m and due to usually strong daytime winds, atmospheric stratification plays a relatively minor role in determining air pollutant concentrations across the study area, and on concentrations build-up during the high traffic hours. The common 75–150 m thickness of the night-time boundary layer is normally associated with low traffic emissions and with easterly (land-sea) winds that disperse the pollutants offshore to the Mediterranean.

2.2. Monitoring data

Air quality is observed in Israel by a network of air quality monitoring stations that are maintained by the Ministry of Environmental Protection. Most of the stations are of an ambient monitoring type and comply with the EU Council Directive 1999/30/EC for protection of human health. Few stations are deployed close to major roads and are designated traffic monitoring stations. The air quality monitoring data used in this study were obtained from the Technion Center of Excellence in Exposure Science and Environmental Health's air pollution monitoring database (TAPMD). The database collates all the air quality monitoring data observed in Israel since 1997 and, subsequently, quality assurance/quality control procedures are applied to enhance them. In this study we used the NO_x and NO_2 records observed by ambient