Environmental Pollution 178 (2013) 411–418

Contents lists available at SciVerse ScienceDirect

Environmental Pollution

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

Application of validation data for assessing spatial interpolation methods for 8-h ozone or other sparsely monitored constituents

John Joseph [a,*], Hatim O. Sharif [a], Thankam Sunil [b], Hasanat Alamgir [c,1]

[a] The University of Texas at San Antonio, Department of Civil and Environmental Engineering, BSE 1.202, One UTSA Circle, San Antonio, TX 78249, USA
[b] The University of Texas at San Antonio, Department of Sociology, MS 4.02.66, One UTSA Circle, San Antonio, TX 78249, USA
[c] One Technology Center, 7411 John Smith Drive, Suite 1100, San Antonio, TX 78229, USA

ARTICLE INFO

Article history:
Received 29 November 2012
Received in revised form
11 March 2013
Accepted 16 March 2013

Keywords:
Air quality
Spatial interpolation
Ozone
Overfitting
Inverse distance weighting
Nearest neighbor
Kriging

ABSTRACT

The adverse health effects of high concentrations of ground-level ozone are well-known, but estimating exposure is difficult due to the sparseness of urban monitoring networks. This sparseness discourages the reservation of a portion of the monitoring stations for validation of interpolation techniques precisely when the risk of overfitting is greatest. In this study, we test a variety of simple spatial interpolation techniques for 8-h ozone with thousands of randomly selected subsets of data from two urban areas with monitoring stations sufficiently numerous to allow for true validation. Results indicate that ordinary kriging with only the range parameter calibrated in an exponential variogram is the generally superior method, and yields reliable confidence intervals. Sparse data sets may contain sufficient information for calibration of the range parameter even if the Moran I *p*-value is close to unity. R script is made available to apply the methodology to other sparsely monitored constituents.

1. Introduction

Adverse health impacts of ground-level ozone are well-documented. Several epidemiological studies have indicated the short and long term adverse effects of tropospheric ozone on health (Kinney et al., 1988; Romieu et al., 1996; Gryparis et al., 2004). The short term effects of exposure to ozone include increasing hospital admissions and emergency department visits and chronic respiratory conditions (Bell et al., 2004; Lippmann, 1993). Elevated concentration of ozone also results in daily mortality for respiratory as well as cardiovascular diseases (Stafoggia et al., 2010; Bell et al., 2004; Gryparis et al., 2004; Ito et al., 2005). Adverse health effects have led to standards such as the 8-h maximum average by the World Health Organization (2006).

Estimating exposure levels in urban areas is indispensable in the formulation of responses. Cost constraints typically limit monitoring ozone to a small number of stations in an urban area of concern. Interpolation methods may be heavily relied upon to estimate concentrations throughout such an area. Mulholland et al. (1998) apply universal kriging to interpolate 1-h and 8-h data from 10 stations in the area of Atlanta, Georgia, USA. Rojas-Avellaneda (2007) compares inverse distance weighting and other interpolation methods for peak-hour ozone data from 16 stations in Mexico City, Mexico. Sanchez et al. (2009) apply a kriging method to interpolate data from 8 stations in the Guadalajara urban area of Mexico. Son et al. (2010) apply of variety of interpolation techniques for 8-h ozone concentration data from 13 stations in the urban area of Ulsan, Korea. Other studies, such as that of Temiyasathis et al. (2009), who use 8-h ozone data from 14 stations in the Dallas-Fort Worth area of Texas, rely on sophisticated procedures that incorporate meteorological data or models of atmospheric dynamics in conjunction with an interpolation method such as kriging. In a study for Madrid, Spain, Montero et al. (2010) apply ordinary kriging to annualized ozone data from 27 continuous monitoring stations, an unusually high number for a single urban area. We the authors of this paper need to estimate 8-h ozone exposure in the area of San Antonio, Texas, USA, which has at most 11 active ozone monitoring stations. This present study is motivated by our need to clarify for ourselves which interpolation method would be most suitable, given that we are unable to sacrifice a portion of so few stations for validation.

Spatial interpolation methods typically involve the calibration of parameters so that the values predicted most closely match the

* Corresponding author.
E-mail addresses: john.joseph@utsa.edu (J. Joseph), hatim.sharif@utsa.edu (H.O. Sharif).
[1] Tel.: +1 210 562 5516; fax: +1 210 562 5528.

0269-7491/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envpol.2013.03.035

values measured at the monitoring stations. Sound statistical practice requires that observed data be separated into two subsets, a calibration or "training" subset, and a validation or "unseen" subset. However, in the case of 8-h ozone concentrations in urban areas, the number of monitoring stations is typically too low to allow for sacrificing a portion for validation. Yet when the number of monitoring stations is low, the risk of overfitting is great, and validation is most needed.

One method often used to help compensate for the unavailability of true validation data is a cross-validation process in which one observed data point at a time is excluded on a rotating basis (e.g., Son et al., 2010). This process, however, presents complications especially when the interpolation model contains parameters which are to be calibrated. In such cases, the model is initially calibrated to minimize the error function for the set of included points, and then the resulting residual (or error) is determined at the excluded point. This process is repeated until each point has had a turn at being excluded, and the distribution of residuals that occur at the excluded points might then be assumed to represent the distribution of errors in predicting concentrations at points where there are no monitoring stations. However, each set of included points yields a different set of parameter estimates. A single set of parameter estimates must be used for predicting concentrations throughout the entire area of concern. Therefore, the parameters must be re-calibrated to minimize the residual function at all the excluded points simultaneously, as is done in conventional (as opposed to one-at-a-time cross-validation) calibration. Now, which residuals are to be used along with the conventionally calibrated parameters to represent the distribution of residuals expected to occur at the non-monitored points? If the residuals generated by the conventional calibration are used, the entire one-at-a-time cross-validation has little value, as nothing is used from it. If the residuals associated with the one-at-a-time cross-validation are used, they do not correspond to the actual parameter values used in the model for predicting concentrations at non-monitored points, and justification for their usage, while not impossible, becomes complicated.

As the set of points used in the one-at-a-time cross-validation process becomes large, we may feel more confident that the residuals generated are representative of the residuals that would be found at non-monitored points. This is because the possibility of overfitting, i.e., the adjustment of parameter values to random effects rather than to actual phenomena, becomes less as the set of points becomes large relative to the number of parameters to be calibrated. Yet the question remains as to how large that set needs to be. This question needs to be answered by testing against truly unseen (validation) data. In our literature review, we did not find any study which utilizes a validation subset to truly validate any interpolation method for 8-h ozone concentrations in urban areas.

Presently, there is no reliable guideline or "rule of thumb" that would allow us to be reasonably confident *a priori* that overfitting is not occurring in any particular interpolation method applied to an 8-h ozone data set in an urban area. The likelihood of overfitting is not easily discerned because it depends on a variety of interacting factors, including the ratio of the number of parameters to the number of data points, constraints assigned to possible parameter values, and how well the structure of the model represents the underlying phenomena (e.g., Whittaker et al., 2010). However, if particular models and parameters are applied to various data sets representing the same basic underlying phenomena repeatedly (in our case, 8-h ozone in urban regions), and checked against validation data, one would expect a rule of thumb to emerge regarding which models and parameters are most appropriate, and how large the data sets must be to avoid overfitting. Then one could proceed with reasonable confidence in applying the tested interpolation methods and parameters where the sparseness of data makes validation impractical. This study is an effort toward developing such a rule of thumb.

More sophisticated methods may be used for estimating 8-h ozone concentrations between monitoring stations than are presented in this study. These methods may include models that utilize land use classifications, ozone source locations, meteorological conditions, dynamics of dispersion and atmospheric chemistry, and other sophisticated measures (e.g., Xing et al., 2011; Carslaw and Ropkins, 2012). Such efforts require more resources. This paper deliberately excludes such additional information for the sake of developing screening tools that may be quickly and easily used. Simple methods such as those reviewed in this study are to be utilized first. If they yield confidence intervals adequate for decision-making, then resources need not be wasted on more sophisticated efforts.

2. Data and methods

2.1. Data

We selected two urban areas with exceptionally large and dense monitoring networks so that a portion of data may be reserved for validation – the Los Angeles/Riverside, California, USA urban area (herein referred to as the "Los Angeles area"), which has up to 27 active stations, and the Houston/Galveston, Texas, USA urban area (herein referred to as the "Houston area") which has up to 42 active stations. A shapefile of the urban populated areas as of the year 2010 was obtained from the United States Census Bureau at http://www2.census.gov/geo/tiger/TIGER2010/UA/2010/. ArcGIS 10 was used to develop Fig. 1.

For each of the years 2009, 2010, and 2011, the dates having the maximum 8-h ozone concentration for the Los Angeles area and the Houston area were selected. For the Los Angeles area, all of these dates fell on a weekend, and so, to help ensure a better representation of the variety of spatial distributions, the date with the second highest 8-hr average was chosen for 2011, as this fell on a weekday. Hourly ozone concentrations for Houston area were obtained through the Texas Commission on Environmental Quality (TCEQ) at http://www.tceq.texas.gov/cgi-bin/compliance/monops/daily_summary.pl. The data is from stations forming TCEQ's Region 12. Hourly data for the Los Angeles area were obtained from the California Environmental Protection Agency Air Resources Board (ARB) at http://www.arb.ca.gov/adam/hourly/hourly1.php, and are of the ARB's Region 61 data. Ozone analyzers and their calibration are to meet the requirements of Title 40 of the United States Code of Federal Regulations, Part 53. Geographic coordinates were obtained through links at these TCEQ and ARB websites, and then projected using the GEOmap (Lees, 2012) package of R statistical software version 15.1 (R Core Team, 2012).

For each of the dates at least one of the monitoring stations was inactive due to malfunctioning or maintenance, so that the exact number of data points varied. The dates and numbers of stations having available 8-h ozone data are shown in Table 1 for each urban area. Also shown is the average area covered per station. As is discussed below, calibration sets of size 10 and 20 would be randomly selected from the full data sets. The last column of Table 1 displays the approximate area covered per station for the calibrations sets of size 10.

2.2. Creation of calibration and validation sets

A comparison of interpolation methods cannot be achieved without separating each of the six sets of data into calibration and validation sets. It is not unusual for some urban areas to be limited to approximately 10 ozone monitoring stations, while it is unusual for them to exceed 20 stations. We therefore chose calibration set sizes of 10 and 20. An exception is the October 22, 2011 dataset for the Los Angeles area, for which the calibration set sizes were 10 and 14 due to the desire to have the validation set size to be no fewer than 7.

Any particular randomly chosen calibration subset may unfairly favor one interpolation over another due to effects that are merely random. The number of all possible sets is extremely large. We limited the number of calibration sets randomly selected from each data set to approximately 5,000.

The expected number of times that each data point would be selected was the same for all data points, and the urban area was subdivided such that each set displayed a realistic spread.

2.3. Selection of interpolation methods

Data was explored for autocorrelation and trends. In Fig. 1 each 8-h ozone measurement is represented by a circle of size proportional to its value. These measured values range from 31.0 parts per billion by volume (ppb_v) to 112.5 ppb_v for the Houston area data sets, and from 10.8 ppb_v to 117.1 ppb_v for the Los Angeles area