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

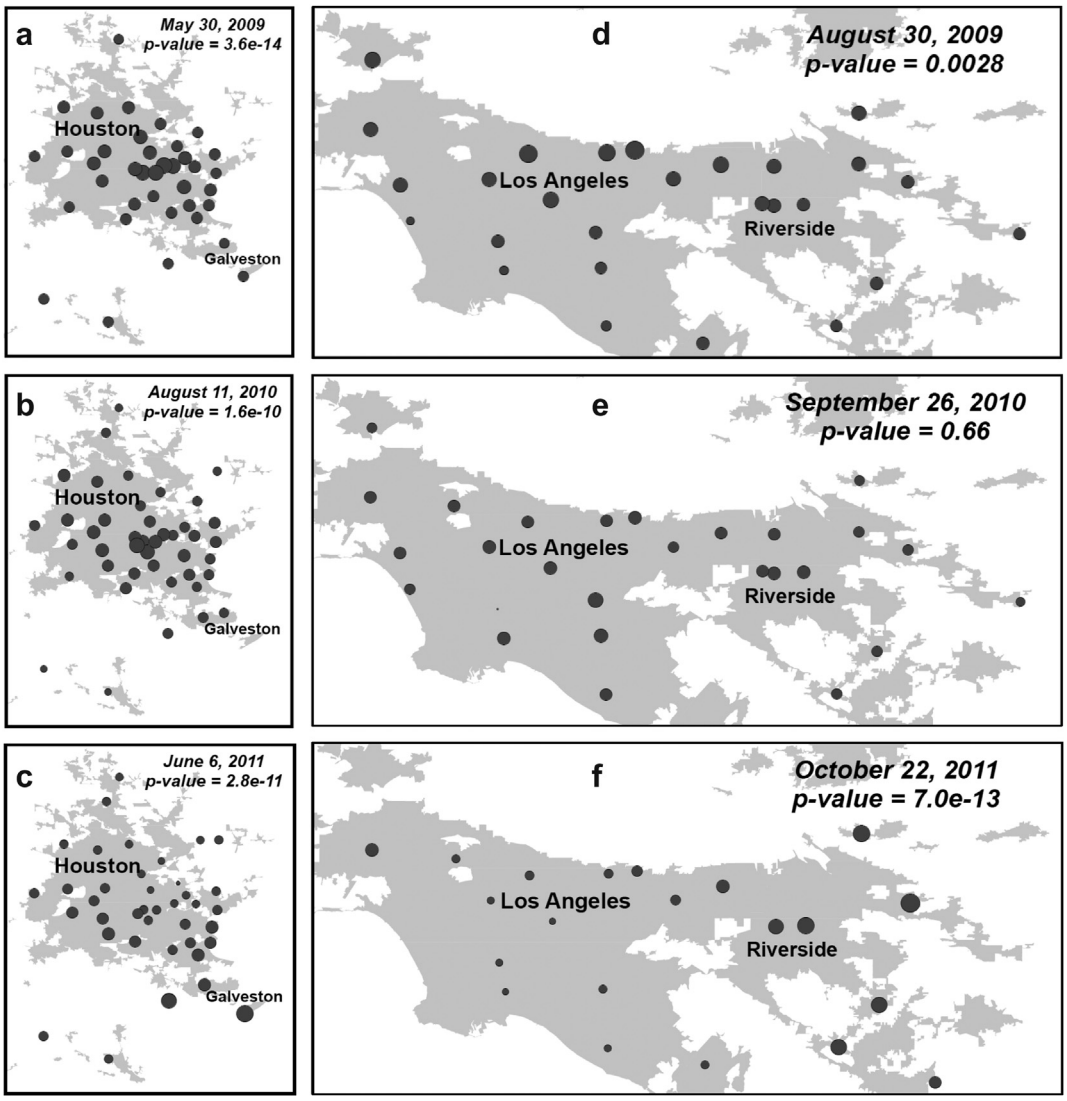

Fig. 1. a–f. Locations of stations providing observed data, with circle sizes proportional to measured 8-h ozone values. The Houston area is the left column (a–c), and the Los Angeles area is the right column (d–f), ordered from top to bottom. The gray patches represent the urban areas identified by the 2010 United States census. The Moran I *p*-value is in the upper right-hand corner for each data set.

data sets. The *p*-value for the Moran I test was calculated using the R package "ape" (Paradis et al., 2004). These *p*-values tend to be far less than typical levels of significance, such as 0.10 or 0.05, meaning that we would reject the hypothesis that the observed values are not autocorrelated. Therefore, interpolation techniques in which near values are weighted more than far values are worth considering. For the September 26, 2010 Los Angeles area data set, the *p*-value is quite high, and no trend seems evident, suggesting that simple averaging also be considered. Trends may also be suggested in some of the datasets displayed in Fig. 1. Based on this data exploration and frequent appearance in the literature, we selected simple averaging, nearest neighbor, inverse distance weighting (IDW), ordinary kriging, and universal kriging as the simple interpolation methods to be tested and compared. Autocorrelation and trends are expected to be less evident in the randomly selected calibration sets, which are only a fraction of the full sets displayed in Fig. 1.

2.4. Common basis of comparison and one-at-a-time cross-validation as a calibration process

To compare interpolation methods, a common objective function must be maximized (or minimized) for calibration, and then the value of that function, with parameters calibrated, is to be calculated for the validation data. The objective function value for the validation data may be a criterion to identify the best interpolation method.

An advantage of a likelihood function as an objective function is its flexibility in assigning appropriate weights to residuals. However, a calibration set size of 10 provides insufficient information for identification of the density function and its statistical parameters needed for the likelihood function. We therefore chose the sum of the square of the residuals (SSR) as the objective function. Minimizing this function is the same as minimizing a variety of commonly used measures for any fixed sample size, including the mean square error, the root-mean-square error (RMSE), the sample variance, and the sample standard deviation (SD). For the validation data, SSR was converted to RMSE to provide the reader with a better sense of the magnitude of the errors relative to the observed values. For each of these statistics, the extension ".cal" and ".val" indicate that the statistic pertains to the calibration set or the validation set, respectively.

For each of the interpolation methods, one member of the calibration set was excluded at a time, and the predicted value at that point was compared with the observed value to determine a residual. The number of residuals for each calibration set was thus equal to the size of the calibration set. For models containing parameters to be calibrated, the parameters were calibrated until the SSR from the one-at-a-time process was minimized, and thus the parameter estimates and their corresponding residuals are what would be yielded through conventional calibration. A more thorough discussion of the calibration process is detailed below, where the interpolation models are described. For some of the interpolation methods – simple averaging, nearest neighbor, and several IDW models – there was no parameter to adjust and hence no calibration.

Another criterion, in addition to the objective function value for the validations data, is the narrowness of the 95% confidence intervals and the percentage of validation points that fall within those intervals. These confidence intervals were calculated by assuming the residuals of the calibration set to be normally distributed, i.e., as having the width (−1.96*SD.cal, +1.96*SD.cal). Actual confidence

Table 1
Numbers of stations providing ozone data for each of the dates and urban areas of this study, and the average area covered per station. The "subset of size 10" represents 10 stations randomly selected from the full network. Average areas for subsets of size 20 are not shown, but are simply half those shown for subsets of size 10.

Urban area	Date	Number of stations	Average area per station (km^2/station)	Average area per station in subset of size 10 (km^2/station)
Houston, Texas	Saturday, May 30, 2009	37	309	1140
	Wednesday, August 11, 2010	42	262	1100
	Monday, June 6, 2011	38	321	1220
Los Angeles, California	Sunday, August 30, 2009	27	440	1190
	Sunday, September 26, 2010	27	423	1140
	Wednesday, October 22, 2011	21	473	990

intervals will tend to fluctuate spatially, tending to be wider where observed values are more sparse, and can be calculated for kriging, but we apply this simplifying assumption to facilitate comparisons with IDW, simple averaging, and nearest neighbor.

To present the tens of thousands of results in a comprehensible form, averages over the 5000 trials for each of the six data sets were calculated for SD, RMSE, and the percentage of validation data values falling within the 95% confidence intervals.

2.5. Descriptions of interpolation methods

R script for each of the interpolation methods is available from the author, and may be readily adapted for other constituents or other sites.

2.5.1. Simple averaging, nearest neighbor, and inverse distance weighting

In simple averaging (SA), the predicted ozone concentration at any point in the area equals the simple average of the observed ozone values, and thus does not vary spatially.

In the nearest neighbor approach (NN), the predicted concentration at any point equals the value observed at the nearest station.

In inverse distance weighting (IDW), the predicted ozone concentration at any point depends on values at observed points, with observed values from closer stations given more weight than those from farther stations. In the Shepard Method (Shepard, 1968), the simplest form of IDW and that which is used in this paper, the weighting is given according to the following formula:

$$C_{j,\text{predicted}} = \frac{\sum_{i=1}^{N}\left(d_{ij}^{-p}\cdot C_{i,\text{observed}}\right)}{\sum_{i=1}^{N} d_{ij}^{-p}}$$

where $C_{j,\text{predicted}}$ is the predicted concentration at any point j, $C_{i,\text{observed}}$ is the observed concentration at station i, and d_{ij} is the distance from point i to point j. The power parameter, p, is a positive real number.

As p approaches 0, the distance weights, d_{ij}, all approach 1, and the IDW predicted values approach that of SA. As p approaches infinity, the weight for the shortest distance becomes very large compared to the weights for all other distances, and the IDW predicted values approach those of NN.

We initially assigned values of 1, 2, 3 and 4 to p in our comparison of methods to adequately represent the range of possibilities between SA and NN. However, results for $p = 3$ were always very close to the results for either $p = 2$ or $p = 4$, and thus we excluded $p = 3$. We also treated p as a calibrating parameter, i.e., for each calibration set, we found the p parameter value that would yield the lowest SSR in the one-at-a-time cross-validation process.

2.5.2. Ordinary kriging and universal kriging

An explanation of kriging is provided below only insofar as needed to discuss its implementation for this study. For more detailed discussions, the reader is referred to Journel (1989) or other geostatistics primers.

In ordinary kriging, a variogram is used to quantify the tendency for the differences between values at points to increase with distance. We would expect ozone concentrations at points very near to each other to be close in value, but would not be surprised to find points separated by great distance to differ greatly. A variety of variogram models may be selected. For sufficiently large data sets, the model may be selected and parameters estimated based on plots of the data. However, a plot of even our largest data set (that of August 11, 2010, in the Houston area) with the R package "geoR" (Diggle and Ribeiro, 2007; Ribeiro and Diggle, 2001) did not yield a sufficiently clear pattern to suggest an appropriate variogram model. We therefore chose the commonly used exponential variogram model. Its representation of exponential decay is simple and reasonable for tropospheric ozone concentrations. Furthermore, in Rojas-Avellaneda (2007) it appears to be the overall best performing of several variogram models for kriging one-hour ozone concentrations in Mexico City, Mexico. In the matrix operation of minimizing SSR for a given r (the range, or the distance at which the semi-variance between ozone concentrations no longer increases with distance), a Lagrange parameter is utilized (e.g., Journel, 1989) and the sill parameter σ^2 (the semi-variance between ozone concentrations at distances greater than the range) is factored out so that predicted values are dependent only on the range r. A third parameter, the nugget, τ^2, which would represent measurement errors in the ozone measurement equipment, was assumed to be negligible relative to the prediction errors associated with distance. Having only one parameter, the range r, reduces the tendency toward overfitting.

To find the optimal r value for each calibration set, r values were selected sequentially as 1%, 2%, 5%, 10%, 15%,…, 95%, and 100% of the distance between the two points farthest from each other in the calibration set. The optimal r value achieves the lowest SSR in the one-at-a-time cross-validation process.

In universal kriging, additional parameters were utilized to allow for representing a linear trend in both the north–south and the east–west directions. Three Lagrange parameters were therefore utilized (e.g., Journel, 1989). As the optimal linear trend parameters are unique for a given r value, we selected various values only for r, and selected them sequentially as we did for ordinary kriging, until finding the one corresponding to the lowest SSR.cal in the one-at-a-time cross-validation process.

2.6. SSR.cal and Moran I p-value as potential indicators of best method

The practitioner utilizing an urban air quality monitoring network typically will not have the luxury of sacrificing a substantial portion of data for validation of the spatial interpolation method. We therefore tested whether the interpolation method with the lowest SSR.cal might also consistently be the interpolation method with the lowest RMSE.val and have close to 95% of validation data falling within the 95% confidence interval. This would allow the practitioner without validation data to select the interpolation method yielding the lowest SSR in one-at-a-time cross validation.

Also, in theory, spatially correlated data is better represented by interpolation methods that give more weight to nearer values than it is by simple averaging, unless overfitting occurs. The authors do not doubt that actual 8-h ozone concentrations are autocorrelated, but an important question remains: If the available data is so sparsely distributed that it does not provide evidence (by, e.g., the Moran I test) of the underlying autocorrelation, might the data set still contain enough information to guide the selection of the interpolation method. We therefore used the R package "ape" (Paradis et al., 2004) to calculate the Moran I p-value for each calibration set, to assess whether this p-value might be helpful in selecting an interpolation method when validation data is unavailable.

3. Results and discussion

Tables 2 and 3 display the results for each of the calibration subset sizes for each of the three dates for the Houston and Los Angeles areas. The three statistics reported in Tables 2 and 3 are averages for the 5000 or more randomly chosen calibration sets. Best results for each date and calibration set size are highlighted.

In the "Average SD of residuals, calibration set" column, the values appearing for calibrated models are for the residuals generated by the parameter estimates that minimize the SSR. SD values in this column are similar in magnitude to the interpolation residuals for 8-h ozone in urban areas reported by Son et al. in Ulsan, South Korea (2010) and Rojas-Avenellada in Mexico City (2007). Houston more closely matches the more densely networked Ulsan, while Los Angeles more closely matches the more sparsely networked Mexico City, though a thorough investigation would likely reveal that significant differences cannot be explained merely by network density or numbers of monitoring stations.

The results for the three dates in the Houston area (Table 2) show that the 95% confidence intervals are generally reliably in capturing the validation data points. The percent captured is never far below the anticipated 95%, and in some cases exceeds 95%.

Table 2 also indicates that the ordinary kriging method is generally superior to the other techniques for these dates in the Houston area. In each of the six combinations of date and

Table 2
Results for Houston area, with calibration subsets of size $n = 10$ and $n = 20$. The gray shading highlights the overall best option in the first column, and best values in the other columns.

	Average SD of residuals, calibration set		% of validation points in 95% CI		Average RMSE of validation points	
	$n = 10$	$n = 20$	$n = 10$	$n = 20$	$n = 10$	$n = 20$
May 30, 2009						
Simple average	12.70	11.42	92.91	95.93	11.30	10.77
Nearest neighbor	11.58	8.76	96.05	94.18	9.54	8.41
IDW, $p = 1$	11.03	9.77	97.58	96.70	9.36	9.13
IDW, $p = 2$	10.23	8.29	97.83	95.11	8.67	8.08
IDW, $p = 4$	10.20	7.58	96.74	93.73	8.53	7.56
IDW, optimal p	9.72	7.46	95.39	93.24	8.97	7.63
Ordinary Kriging, optimal r	8.74	6.16	97.54	93.59	7.01	5.84
Universal Kriging, linear	10.37	6.61	95.59	92.63	8.15	6.36
August 11, 2010						
Simple average	12.55	11.92	91.77	91.34	11.83	11.79
Nearest neighbor	12.61	10.29	95.03	95.32	11.07	9.51
IDW, $p = 1$	11.81	10.83	92.94	91.51	10.55	10.25
IDW, $p = 2$	11.42	9.98	93.51	93.04	10.17	9.47
IDW, $p = 4$	11.65	9.64	94.15	94.91	10.22	9.07
IDW, optimal p	11.07	9.34	92.30	92.90	10.51	9.35
Ordinary Kriging, optimal r	10.54	8.29	93.87	95.66	9.08	7.44
Universal Kriging, linear	14.04	9.56	95.47	95.73	10.01	7.90
June 6, 2011						
Simple average	16.97	16.06	93.25	92.11	15.35	15.07
Nearest neighbor	14.49	12.41	93.88	93.86	12.74	11.24
IDW, $p = 1$	15.38	13.94	93.65	92.22	13.12	12.72
IDW, $p = 2$	13.92	12.04	93.73	92.37	11.65	10.86
IDW, $p = 4$	13.18	13.15	94.24	94.14	10.88	10.90
IDW, optimal p	12.77	10.45	93.19	91.97	11.46	10.15
Ordinary Kriging, optimal r	12.82	10.04	94.03	92.50	10.39	9.59
Universal Kriging, linear	16.83	11.89	92.75	92.85	13.29	10.42

calibration set size, it yields the lowest average RMSE.val. While it does not maintain the highest percentage of validation points within the 95% confidence interval, it does maintain a percentage close to and occasionally higher than 95%, despite typically having the narrowest intervals due to having the lowest SD.cal.

In general, the technique that yields the lowest *average* SD.cal also yields the lowest *average* RMSE.val, and maintains close to 95% of the validation points within the 95% confidence interval.

However, in the one-at-a-time cross validation process, the subjection of parameters to calibration may limit the reliability of

Table 3
Results for Los Angeles area, with calibration subsets of size $n = 10$ and $n = 20$. The gray shading highlights the overall best option in the first column, and best values in the other columns.

	Average SD of residuals, calibration set		% of validation points in 95% CI		RMSE of validation points	
	$n = 10$	$n = 20$[a]	$n = 10$	$n = 20$[a]	$n = 10$	$n = 20$[a]
August 30, 2009						
Simple average	16.96	15.92	93.61	92.78	15.16	14.80
Nearest neighbor	18.14	15.96	93.09	92.75	15.52	15.04
IDW, $p = 1$	16.57	14.80	93.99	93.08	13.86	13.32
IDW, $p = 2$	16.33	13.90	94.17	93.47	13.34	12.45
IDW, $p = 4$	16.31	14.22	93.48	92.14	13.64	13.24
IDW, optimal p	15.12	13.67	92.69	91.07	14.28	13.18
Ordinary Kriging, optimal r	13.95	10.86	93.44	95.13	12.21	10.20
Universal Kriging, linear	13.58	10.89	91.94	93.15	12.43	10.85
September 26, 2010						
Simple average	14.79	15.43	91.39	94.51	15.27	12.96
Nearest neighbor	18.66	19.48	91.25	91.57	21.01	19.06
IDW, $p = 1$	15.09	15.58	92.26	94.19	16.03	13.59
IDW, $p = 2$	15.71	16.55	91.95	91.12	17.70	15.45
IDW, $p = 4$	16.40	18.51	91.18	91.90	19.30	17.84
IDW, optimal p	14.29	14.99	90.87	92.76	16.24	16.37
Ordinary Kriging, optimal r	13.90	14.84	91.17	92.41	16.79	19.22
Universal Kriging, linear	17.53	16.14	93.22	92.63	18.96	19.70
June 22, 2011						
Simple average	27.64	26.79	98.29	99.03	25.13	24.35
Nearest neighbor	16.07	13.98	93.61	92.51	13.71	11.54
IDW, $p = 1$	20.29	18.90	97.85	97.76	17.60	16.54
IDW, $p = 2$	15.81	14.19	95.78	93.87	13.64	12.25
IDW, $p = 4$	13.92	12.76	92.06	91.58	12.55	11.04
IDW, optimal p	13.48	12.49	91.09	90.31	12.87	11.49
Ordinary Kriging, optimal r	13.78	11.67	95.77	94.59	11.43	10.22
Universal Kriging, linear	14.78	13.59	90.71	92.14	13.44	11.61

[a] The larger calibration set size for June 22, 2011 is actually 14 rather than 20.

using the lowest SD for selecting the most accurate of the techniques. Among the five techniques in which no parameter was subjected to calibration (simple average, nearest neighbor, and IDW with $p = 1, 2, 4$), the one with the lowest average SD is also always the one with the lowest average RMSE.val in Table 2. Also, in most cases, the ranking is exactly maintained from first through fifth. Among the three techniques in which at least one parameter is subjected to calibration (the last three techniques listed under each date) such is not the case, and dramatic reversals of order are possible. For example, on August 11, 2010 "Universal Kriging, linear" has an average SD.cal of 14.04 for the calibration sets of size 10, while "IDW, optimal *p*" has a substantially lower value of 11.07. Yet the average RMSE.val is less for "Universal Kriging, linear" than for "IDW, optimal *p*". Such reversals may be due to overfitting.

For the small calibration set size of 10, the overfitting in "Universal Kriging, linear" may be severe. The one point excluded while the parameters are fit to the other 9 typically has a very large error associated with it due to this overfitting, and collectively the 10 points have a high SD.cal. Selecting the range parameter that minimizes the square of such errors in one-at-a-time cross validation reduces the severity of this overfitting, yielding an average RMSE.val that is in many cases competitive with those yielded by other methods. For the larger calibration set size of 20, the impact of overfitting is substantially reduced relative to that of size 10, as shown by pronounced reductions in the average SD.cal. Yet, even for calibration sets of size 20, vestiges of overfitting remain. "Universal Kriging, linear" provides a consistently inferior average RMSE.val to that of "Ordinary Kriging, optimal *r*".

A comparison of the results for "IDW, optimal *p*", "Ordinary Kriging, optimal *r*" and the three IDW techniques with fixed *p* clarifies that it is not the sheer number of calibrating parameters that causes overfitting. Even having only one calibrating parameter may contribute to overfitting if that parameter does not adequately represent the underlying phenomena. For calibration sets of size 10 and 20, "IDW, optimal *p*" consistently achieves a lower average SD.cal than do the IDW techniques with fixed *p*. Yet, for calibration sets of size 10, its average RMSE.val is always higher than that of at least one of the IDW techniques with fixed *p*, suggesting that the parameter may be overfit for calibration sets of size 10. Meanwhile, "Ordinary Kriging, optimal *r*" in which the range is the only calibrating parameter, consistently delivers the lowest average RMSE.val, even for calibration sets of size 10. The exponential model with parameter *r* better represents the underlying reality than does IDW with parameter *p*.

Much of what is observed of the Houston area (Table 2) also applies to August 30, 2009 and June 22, 2011 of the Los Angeles area (Table 3). The 95% confidence intervals are generally reliable. "Ordinary kriging, optimal *r*" emerges as the superior method for these two dates, though "Universal Kriging, linear" becomes more competitive. This relative improvement in "Universal Kriging, linear" is not surprising, as Fig. 1d and f suggest a general decline in observed concentrations from northeast to southwest. A model more sophisticated than a linear model might provide an even greater improvement, but would best be supported by data collection efforts that would go beyond the basic screening and assessment which our methodology is intended to support.

The results for the Los Angeles area on September 26, 2010 in Table 3 are much different from those of the other two dates, and from those of any of the dates in Houston. Simple averaging yields the lowest average RMSE.val for both calibration set sizes, and does so by a substantial margin. The *p*-value for the Moran I test is 0.66. The *p*-value for each of the other five combinations of urban region and date, where "Ordinary Kriging, optimal *r*" yields the lowest average RMSE.val, is securely below typical test significance levels, as shown in Fig. 1. Theoretically, if no autocorrelation exists, then simple averaging is the most favorable interpolation method. The results thus appear consistent with theory.

However, Fig. 1e shows a single ozone concentration of 10.8 ppb_v that is extremely small relative to that date's other 26 values, which range from 60.8 to 96.6 ppb_v. (This small value is the barely visible speck in the lower left quadrant of Fig. 1e.) If the value of 10.8 ppb_v on September 26, 2010 is treated as an anomaly and removed, the Moran I *p*-value for the remaining 26 values drops to 0.008. Also, as will be discussed below when Table 4 is presented, "Ordinary Kriging, optimal *r*" then becomes the method that most frequently yields the lowest RMSE.val. Once again we see what we would expect from theory: low *p*-values reflecting autocorrelation favor techniques such as "Ordinary Kriging, optimal *r*" over simple averaging.

The data set *available* to a practitioner for an urban area will often be of stations sparser or fewer than those of this study. For example, the authors are interested in the area of San Antonio, Texas, U.S.A., which has at most 11 active monitoring stations at a density of 1 per 470 km^2. For such networks, the observed values may fail to provide evidence of autocorrelation when autocorrelation is actually present. The use of the Moran I *p*-value to select between simple averaging and other techniques might therefore be impractical. As shown in Table 4, many of the approximately 5,000 calibration sets of size 10 (a size that may be typically available to the practitioner) have *p*-values greater than the test significance level of 0.1. For these, the hypothesis of no autocorrelation cannot be rejected, though the *p*-values for the corresponding full data sets are typically multiple orders of magnitude smaller than 0.1. The percent of such cases in which "Ordinary Kriging, optimal *r*" ("% OK best" column) yields the lowest RMSE.val ranges from 43% to 89% (if we exclude the September 26, 2010 date). For simple averaging, the percent never exceeds 0.5% ("% SA best" column). A high *p*-value for data set sizes typically available is thus unlikely to be a sufficient reason for selecting simple averaging as an interpolation method. Furthermore, even for *p*-values approaching 1 the superiority of "Ordinary Kriging, optimal *r*" may be maintained. For example, of the 230 calibration sets of size 10 having a *p*-value that is greater than 0.9 (not shown) for the August 30, 2009 Los Angeles area data set, "Ordinary Kriging, optimal *r*" yields the lowest RMSE.val 73% of the time. Simple averaging yields the lowest RMSE.val for none of these 230 sets.

A statistic that offers more promise than does the Moran I test *p*-value for selecting the best interpolation method is SSR.cal. If we consider only simple averaging and "Ordinary Kriging, optimal *r*", two techniques embodying strongly contrasting assumptions, we find that the method which has the lower SSR.cal also has the

Table 4
General superiority of "Ordinary Kriging, optimal *r*" regardless of Moran I *p*-value for calibration sets of size 10. "% OK best" and "% SA best" are the percentages of cases in which "Ordinary Kriging, optimal *r*" and "simple averaging", respectively, are the interpolation methods yielding the lowest RMSE.val.

	Moran I *p*-value <0.1			Moran I *p*-value >0.1		
	Number of cases	% OK best	% SA best	Number of cases	% OK best	% SA best
Houston area						
May 30, 2009	2104	95%	0.0%	2911	89%	0.0%
August 11, 2010	1153	92%	0.0%	3900	81%	0.1%
June 6, 2011	2769	59%	0.0%	2247	64%	0.0%
Los Angeles area						
August 30, 2009	195	88%	0.0%	4805	71%	0.5%
September 26, 2010	598	18%	72%	4402	21%	50%
September 26, 2010[a]	747[a]	43%[a]	7.2%[a]	4253[a]	35%[a]	1.0%[a]
June 22, 2011	4913	64%	0.0%	87	43%	0.0%

[a] Measured 8-h concentration of 10.8 ppb_v treated as an anomaly and excluded.

lowest RMSE.val in 88% of the approximately 60,000 (5000 or slightly more for each of the 12 combinations of date, site, and calibration set size) cases. (This and other percentages listed in this paragraph are not shown in Table 4.) If we exclude the September 26, 2010 Los Angeles data, this overall percentage rises to approximately 99%, and the calibration set size of 10 holds this predictive power well relative to that of size 20. For the sets of size 10, the percentage is 98.6%, while for the sets of size 20 it is 98.9%. Simple averaging yields the smaller SSR.cal in only about 3% of the cases, almost all of which occur for the set size of 10 rather than 20, and of this 3%, simple averaging achieves the lowest RMSE.val 81% of the time. (If the September 26, 2010 Los Angeles data is included, simple averaging achieves the lowest SSR.cal approximately 5% of the time, and of this 5%, simple averaging achieves the lowest RMSE.val 69% of the time.) In a small number cases the optimal r was small and "Ordinary Kriging, optimal r" yielded essentially the same results as simple averaging.

When the other methods are included along with simple averaging and "Ordinary Kriging, optimal r", the success rate of using SSR.cal to predict the method that would yield the lowest RMSE.val drops to less than 50%, even with the September 26, 2010 Los Angeles data excluded. This is primarily because these other techniques apparently are not as capable of truly representing the underlying phenomena, and achieve the lowest SSR.cal primarily by chance. "IDW, optimal p" provides the lowest SSR.cal the most frequently among these other methods, but only 5%–15% (depending on the site and date) of these yield the lowest RMSE.val, perhaps due to an overfitting tendency. For "IDW, $p = 1$", "IDW, $p = 2$", "IDW, $p = 4$", and "nearest neighbor" the lowest SSR.cal leads to a corresponding lowest RMSE.val in less than 1% of the approximately 60,000 calibration sets.

4. Conclusion

The 95% confidence intervals based on the residuals from the one-at-a-time cross-validation and assumed to be normally distributed were shown to be generally reliable for the data sets of this study. Of the methods compared, the narrowest intervals and the best fit with validation data (as measured by the RMSE) are generally that of the particular single-parameter ordinary kriging model applied.

Simple averaging, as opposed to interpolation methods which give nearer values more weight, would be expected to perform better than such interpolation methods if ozone concentrations are not autocorrelated. However, a high Moran I p-value, even if close to 1 for data set sizes typically available to the practitioner, is unlikely to be an indicator that simple averaging would outperform kriging or other interpolation methods. A more helpful indicator of the best interpolation method is that which yields the lowest sum of the square of the residuals obtained through one-at-a-time cross-validation, and this statistic is a most reliable selecting tool when the alternatives being considered are substantially different from each other, i.e., when they represent the underlying distribution in fundamentally contrasting ways.

This study is not a comprehensive comparison of interpolation techniques, but only of simple ones which have low requirements of time and expertise, and can be employed without collecting and inputting additional information such as land use and meteorological conditions. Further study to include other variogram models, such as Gaussian and spherical models, may be worthwhile, as would including a nugget effect and applying the general methodology to other large cities having relatively high numbers of monitoring stations. For now, the emerging rule of thumb for 8-h ozone concentrations in urban areas appears to be that "Ordinary Kriging, range r" is generally the best of the simple methods compared, unless the sum of the square of the residuals in one-at-a-time cross validation indicates simple averaging, and that the range parameter r can be calibrated without overfitting if the dataset consists of 10 or more active stations each of which covers an average area no greater than approximately 1,200 km^2, as suggested by the last column of Table 1. Further study with additional dates and urban areas may help in strengthening this rule of thumb and defining its limitations. Further study may also help reveal whether this or a similar rule of thumb applies to ozone where damage to vegetation is predicted by criteria other than 8-h concentrations or in rural areas where monitoring networks are sparser than in urban areas (Hunová et al., 2012; De Marco et al., 2010) and to other sparsely monitored constituents in general.

References

Bell, M.L., McDermott, A., Zeger, S.L., Samet, J.M., Dominici, F., 2004. Ozone and short-term mortality in 95 US urban communities, 1987–2000. JAMA 292, 2372–2378.

Carslaw, D.C., Ropkins, K., 2012. Open air – an R package for air quality data analysis. Environmental Modelling and Software 27–28, 52–61.

De Marco, A., Screpanti, A., Paoletti, E., 2010. Geostatistics as a validation tool for setting ozone standards for durum wheat. Environmental Pollution 158, 536–542.

Diggle, P.J., Ribeiro Jr., P.J., 2007. Model Based Geostatistics. Springer, New York.

Gryparis, A., Forsberg, B., Katsouyanni, K., Analitis, A., touloumi, G., Schwartz, J., et al., 2004. Acute effects of ozone on mortality from the "air pollution and health: a European approach", project. American Journal of Respiratory and Critical Care Medicine 170, 1080–1087.

Hunová, I., Horálek, J., Schreiberová, M., Zapletal, M., 2012. Ambient ozone exposure in Czech forests: a GIS-based approach to spatial distribution assessment. The Scientific World Journal 2012, 10. http://dx.doi.org/10.1100/2012/123760. Article ID 123760.

Ito, K., De Leon, S.F., Lippmann, M., 2005. Associations between ozone and daily mortality: analysis and meta-analysis. Epidemiology 16, 446–457.

Journel, A.G., 1989. Fundamentals of Geostatistics in Five Lessons. American Geophysical Union, Washington, D.C.

Kinney, P.L., Ware, J.H., Spengler, J.D., 1988. A critical evaluation of acute ozone epidemiology results. Archives of Environmental Health 43, 168–173.

Lees, J.M., 2012. GEOmap: Topographic and Geologic Mapping. R package version 1.6-06. http://CRAN.R-project.org/package=GEOmap.

Lippmann, M., 1993. Health effects of tropospheric ozone: review of recent research findings and their implications to ambient air quality standards. Journal of Exposure Analysis and Environmental Epidemiology 3, 103–129.

Montero, J.-M., Chasco, C., Larraz, B., 2010. Building an environmental quality index for a big city: a spatial interpolation approach combined with a distance indicator. Journal of Geographical System 12, 435–459.

Mulholland, J.A., Butler, A.J., Wilkinson, J.G., Russell, A.G., 1998. Temporal and spatial distribution of ozone in Atlanta: regulatory and epidemiologic implications. Journal of Air & Waste Management Association 48, 418–426.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290.

R Core Team, 2012. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. URL. http://www.R-project.org/.

Ribeiro Jr., P.J., Diggle, P.J., 2001. geoR: a package for geostatistical analysis. R-NEWS 1 (2), 15–18.

Rojas-Avellaneda, D., 2007. Spatial interpolation techniques for estimating levels of pollutant concentrations in the atmosphere. Revista Mexicana de Fisica 53 (6), 447–454.

Romieu, I., Meneses, F., Ruiz, S., Sienra, J.J., Huerta, J., White, M.C., Etzel, R.A., 1996. Effects of air pollution on the respiratory health of asthmatic children living in Mexico City. American Journal of Respiratory and Critical Care Medicine 154, 300–307.

Sanchez, H.U.R., Garcia, M.D.A., Bejaran, R., Guadalupe, M.E.G., Vazquez, A.W., Toledano, A.C.P., Villasenor, O.D., 2009. The spatia–temporal distribution of the atmospheric polluting agents during the period 2000–2005 in the urban area of Guadalajara, Jalisco, Mexico. Journal of Hazardous Materials 165, 1128–1141.

Shepard, D., 1968. A Two-dimensional Interpolation Function for Irregularly Spaced Data, pp. 517–524. Proceedings of the 1968 ACM National Conference.

Son, J.-Y., Bell, M.L., Lee, J.-T., 2010. Individual exposure to air pollution and lung function in Korea: spatial analysis using multiple exposure approaches. Environmental Research 110, 739–749.

Stafoggia, M., Forastiere, F., Faustini, A., Biggeri, A., Bisanti, L., Cadum, E., et al., 2010. Susceptibility factors to ozone-related mortality: a population based case-crossover analysis. American Journal of Respiratory and Critical Care Medicine 182, 376–384.

Temiyasathis, C., Kim, S.B., Park, S.-K., 2009. Spatial prediction of ozone concentration profiles. Computational Statistics and Data Analysis 53, 3892–3906.

Whittaker, G., Confesor, R., Di Luzio, M., Arnold, J.G., 2010. Detection of over-parameterization and overfitting in an automatic calibration of SWAT. Transactions of the ASABE 43 (5), 1487–1499.
World Health Organization, 2006. WHO Air Quality Guidelines for Particulate Matter, Ozone, Nitrogen Dioxide and Sulfur Dioxide. WHO Press, Geneva, Switzerland.
Xing, J., Wang, S.X., Jang, C., Zhu, Y., Hao, J.M., 2011. Nonlinear response of ozone to precursor emission changes in China: a modeling study using response surface methodology. Atmospheric Chemistry and Physics 11, 5027–5044.

Web references

California Environmental Protection Agency Air Resources Board at http://www.arb.ca.gov/adam/hourly/hourly1.php, (accessed 1.10.12.).
Texas Commission on Environmental Quality, http://www.tceq.texas.gov/cgi518bin/compliance/monops/daily_summary.pl, (accessed 1.10.12.).
United States Census Bureau, http://www2.census.gov/geo/tiger/TIGER2010/UA/2010/, (accessed 1.10.12.).