ELSEVIER

Environmental Pollution 119 (2002) 99–117

ENVIRONMENTAL POLLUTION

www.elsevier.com/locate/envpol

Predictive mapping of air pollution involving sparse spatial observations

Jeremy E. Diem[a,*], Andrew C. Comrie[b]

[a]*Department of Anthropology and Geography, Georgia State University, Atlanta, GA 30303, USA*
[b]*Department of Geography and Regional Development, The University of Arizona, Tucson, AZ 85721, USA*

Received 3 July 2001; accepted 16 October 2001

"Capsule": *Multiple linear regression can be used to map air pollution levels in metropolitan areas given the availability of multi-temporal air pollution measurements as well as spatially and temporally resolved inventories of atmospheric pollutant emissions.*

Abstract

A limited number of sample points greatly reduces the availability of appropriate spatial interpolation methods. This is a common problem when one attempts to accurately predict air pollution levels across a metropolitan area. Using ground-level ozone concentrations in the Tucson, Arizona, region as an example, this paper discusses the above problem and its solution, which involves the use of linear regression. A large range of temporal variability is used to compensate for sparse spatial observations (i.e. few ozone monitors). Gridded estimates of emissions of ozone precursor chemicals, which are developed, stored, and manipulated within a geographic information system, are the core predictor variables in multiple linear regression models. Cross-validation of the pooled models reveals an overall R^2 of 0.90 and approximately 7% error. Composite ozone maps predict that the highest ozone concentrations occur in a monitor-less area on the eastern edge of Tucson. The maps also reveal the need for ozone monitors in industrialized areas and in rural, forested areas.

Keywords: Ozone; Air pollution; Linear regression; Mapping; GIS

1. Introduction

Spatial interpolation is common practice in a variety of studies, especially those involving environmental variables. Essentially, the goal of interpolation is to discern the spatial patterns of a phenomenon by estimating/predicting values at unsampled locations based on measurements at sample points. When concerning air pollution, the result is a surface (map) of air pollution concentrations. These maps are typically produced to visualize the spatial distribution of air pollution levels and to estimate human and vegetation exposure to pollution. However, a problem arises when one attempts to use a spatial interpolation method to produce an accurate, high-resolution surface based on a small number of spatial observations (i.e. small sample size). This problem leads to the following questions:

1. Is a suitable mapping method available?
2. Can accurate surfaces be created with the above method?

One possible solution to these questions is represented by predictive mapping rather than spatial interpolation. This solution involves the integration of geospatial databases, multi-temporal data, and multi-variate statistical techniques.

1.1. Aims and questions

This paper's main objective is to outline an approach that others can use to map air pollution concentrations for areas with a limited number of spatial observations, but which have an abundance of temporal observations and sufficient ancillary geospatial data. We present an example of a linear regression-based mapping method that uses extensive, multi-temporal, ground-level ozone data at limited spatial locations and extensive, spatially and temporally resolved atmospheric pollutant

* Corresponding author. Tel.: +1-404-651-2504; fax: +1-404-651-3235.
E-mail address: gegjed@langate.gsu.edu (J.E. Diem).

0269-7491/02/$ - see front matter
PII: S0269-7491(01)00308-6

emissions data to predict the spatial patterns of ozone in the Tucson, Arizona region. The period of April–September of 1995–1998 is selected because it is a recent time period and those particular months represent the ozone season (Diem and Comrie, 2001a).

This study is driven by several key research questions, which are as follows:

1. Can linear regression-based mapping be used successfully in a metropolitan area?
2. What are appropriate spatial predictors for ozone?
3. What are the most important spatial processes affecting ozone?
4. What are the spatial patterns of ozone concentrations?
5. What are the policy implications of the resulting maps?

Even though these questions are answered based on results from a single case study, the answers should be moderately applicable to other study regions and should thus help guide future air quality mapping in those regions.

1.2. Sample points and spatial interpolation methods

Many metropolitan areas in the Unites States have relatively low sampling densities of air quality monitors (e.g. a monitor located every 10–20 km) due to the high costs of monitor acquisition and operation. Therefore, even large metropolitan areas with millions of inhabitants have less than 20 monitors for a given pollutant. Monitors are typically placed at perceived pollutant "hot spots" or they are placed in rural areas to obtain estimates of assumed background pollutant concentrations (i.e. outside the urban plume).

Many air pollution studies have employed spatial interpolation methods to produce maps of air pollution concentrations. These studies have mostly employed distance-weighting methods (Greenland and Yorty, 1985; De Leeuw and Van Zantvoort, 1997; Phillips et al., 1997) and kriging (Lefohn et al., 1987, 1988; Casado et al., 1994; Loibl et al., 1994; Westenbarger and Frisvold, 1994; Liu and Rossini, 1996; Godzik, 1997; Phillips et al., 1997; Mulholland et al., 1998; Holland et al., 2000; Tayanc, 2000). Kriging is more complex and has been used more widely than distance-weighting methods, and some comparison studies (Leenaers et al., 1990; Phillips et al., 1997) have found kriging to outperform distance-weighting methods. For a comprehensive review of most available spatial interpolation methods refer to Lam (1983). To predict values at unsampled locations, distance-weighting typically assigns more weight to nearby points than to distant points (Myers, 1991), thus inverse distance-weighting (IDW) is a popular form. Kriging is a regression-based technique that estimates values at unsampled locations using weights reflecting the correlation between data at two sample locations or between a sample location and the location to be estimated (Myers, 1991). Kriging has the advantage of providing unbiased estimates of values at unsampled locations with minimum estimated variance (Leenaers et al., 1990). Both distance-weighting and kriging directly use coordinate information of sample points to perform interpolation, and kriging's performance especially is dependent on the presence of spatial autocorrelation (i.e. values at nearby points are more similar than are values at distant points). The degree of spatial autocorrelation (spatial dependency) latent in a geo-referenced data set to some degree determines how successful spatial interpolation will be (Griffith and Lane, 1999).

Distance-weighting and kriging typically are not suitable for mapping air pollution in most metropolitan areas. The major obstacles to using these interpolation methods are the relative paucity of air quality monitors (i.e. small sample size) and the likely poor spatial distribution of those monitors (i.e. inappropriate sampling scheme). Distance-weighting requires a dense network of uniformly spaced, spatially-autocorrelated observations (Myers, 1994); however, this type of network is usually not available for the measurement of air pollution and many other environmental phenomena. Clustering of the sample points is troublesome, since an equal weight is assigned to each of the points even if it is in a cluster (Lam, 1983). Kriging explicitly requires spatial autocorrelation (Myers, 1991) as well as an abundance of sample points to be an accurate spatial interpolation method (Myers, 1991; Daly et al., 1994; Lesch et al., 1995; Liu and Rossini, 1996). If the degree of spatial autocorrelation is minimal the resulting surface is just "gibberish" (Berry, 1996). Cressie (1991) recommends a minimum of 30 pairs per distance class (i.e. points within a specified distance of each other) while the central limit theorem suggests as many as 100 pairs per class (Griffith and Lane, 1999). The accuracy is also questionable if the sample data do not represent the actual spatial variability of the predicted variable (Daly et al., 1994). As mentioned previously, most air pollution monitoring networks typically do not resolve the actual spatial variability of air pollution levels due to monitor placement bias. Therefore, when used on small geographic scales (e.g. metropolitan areas) that have an insufficient number of monitors, kriging is problematic. In addition to the above problems, if there are a small number of spatial observations, kriging can oversimplify (smooth) the spatial pattern of pollution levels by not capturing the operational scale (i.e. spatial complexity or scale of effect) of the pollutant.

1.3. Predictive mapping with linear regression

From both a theoretical and applied perspective, linear regression appears to be an appealing technique for

mapping air pollution in most metropolitan areas. Unlike distance-weighting and kriging, linear regression models do not require spatially autocorrelated observations to produce an accurate surface. For spatially autocorrelated data, measured observations behave as partially repeated measures of a single observation rather than as single observations (Griffith, 1992). When employing linear regression, one assumes all observations to be independent. A linear regression analysis produces an equation derived from statistical relationships between a dependent variable and one or more independent variables (i.e. predictor variables). The equation yields predicted values of the dependent variable. The general equation is as follows:

$$\hat{Y} = a + b_1 X_1 + \ldots b_n X_n,$$

where, in the case of air pollution, $\hat{Y}$ is the predicted air pollution concentration, a is a constant, and b_1 to b_n are the coefficients for the associated predictor variables X_1 to X_n. Linear regression assumes a significant relationship between air pollution concentrations and other variables (e.g. estimated emissions of pollutants) at the monitors; consequently, the values of the predictor variables are used to predict air pollution concentrations at all locations, not just at the monitors. Predictive mapping with linear regression depends on the availability of spatially continuous predictor variables (i.e. variables with values at every spatial location), which can be stored as geospatial databases within a GIS. Only a few air pollution studies have employed linear regression as a mapping method (Briggs et al., 1997, 2000). The heavy dependence on geospatial databases might be a major reason for linear regression's limited use.

Even though within a metropolitan area there are usually a small number of air quality monitors, linear regression can still be used to predictively map air pollution concentrations. This is enabled by multi-temporal measurements of air pollution concentrations at each monitor and a temporally dynamic pollutant emissions landscape (surface). The monitors are exposed to different emissions over space and time and the measured air pollution concentrations are a consequence of this truly dynamic emissions landscape. Thus, the changing emissions landscape enables an abundance of temporal observations to compensate for a lack of spatial observations.

The usefulness of the linear regression model, which is essentially its ability to produce a reliable surface of air pollution concentrations, depends on the placement of monitors in a wide range of emissions environments and the availability of a reasonably accurate spatially and temporally resolved emissions inventory. In addition, a GIS or similar software is needed to calculate emissions estimates within different neighborhoods of a specific location to account for the range of spatial processes responsible for the pollutant's spatial pattern, as a pollutant's concentration is a function of both local and distant spatial processes (e.g. emissions).

A linear regression model can be developed for both predictive and explanatory purposes. However, a linear regression equation that yields the "best" predictions may not provide the "best" explanations, especially if more than one of linear regression's principal assumptions is violated (Mark and Peucker, 1978; Griffith and Lane, 1999). See Anselin (1988) and Crown (1998) for a detailed treatment of the assumptions of linear regression, since a discussion of the assumptions is beyond the scope of this paper. Fortunately, when only prediction is the objective of the analysis, which is the case in this paper, linear regression is a robust procedure that is influenced marginally by departures from assumptions (Mark, 1984).

In the context of air pollution mapping, linear regression has several substantial benefits. Overall, a theory-based spatial model can be created that accounts for spatial processes and thus predicts spatial patterns. To capture the spatial processes, linear regression analysis utilizes GIS capabilities and extracts the maximum amount of information from the different data sets (Briggs et al., 1997). On a more technical level, linear regression can also produce reliable estimates beyond the sample area (i.e. the polygon anchored by the sample points) as well as beyond the range of the measured values if the relationships between the predictor variables and the predictand (i.e. dependent variable) remain constant. Spatial interpolation methods tend to produce inaccurate spatial predictions outside the sample area, but the spatial extrapolation capability of a linear regression model may be better if the values of the predictor variables have a large range and are distributed well. Linear regression can produce predictions that are both higher and lower than sample point values, thus global minimum and maximum levels can be identified more accurately.

Linear regression's major disadvantages in the current context are its dependence on a large amount of geospatial data, its assumption that the predictor variables (i.e. independent variables) are free of measurement error, and its potentially limited versatility. To optimize the model's specification, many spatially continuous predictor variables might be needed initially to ensure that important predictor variables are not excluded. For gridded data that have been subjected to aggregation and resampling, the violation of the "x is free of error" assumption is inevitable. Most spatial data stored in a GIS contain errors from a wide variety of sources, and these errors may have a significant impact on the validity of applying linear regression equations in a GIS environment (refer to Elston et al., 1997). Finally, since linear regression models are empirical, their versatility is dependent on the input data. Hence, there is no

guarantee as to the reliability of the model once it is extrapolated beyond the range of the input data used to construct it (Chock et al., 1975). Therefore, if the sample points do not cover a representative range of values of the available predictor variables, many locations will have unreliable predictions.

2. Case study: ozone in the Tucson region

2.1. Overview of ground-level ozone

Ground-level ozone is an environmental concern because of its adverse impacts on human health, crops, and forest ecosystems (Sillman, 1999). Ozone can be formed by the oxidation of the volatile organic compounds (VOCs) in the presence of nitrogen oxides (NO_x) and sunlight (Chameides et al., 1992). The accumulation of ozone is critically dependent upon the physical parameters that characterize the planetary boundary layer, such as temperature, wind speed, wind direction, and mixing height (Cardelino and Chameides, 1995). Generally, hot, sunny, and calm conditions are conducive to elevated ambient ozone concentrations. Consequently, concentrations tend to peak during the early afternoon. The above conditions not only increase the photochemical production of ozone but can also increase the atmospheric emissions and concentrations of ozone precursor chemicals (i.e. VOCs and NO_x), which then usually lead to higher ozone concentrations.

2.2. Ozone monitoring in the Tucson region

The Tucson region (centered at $\sim 32^\circ$ N latitude and $\sim 111^\circ$ W longitude) is located in southern Arizona, USA and contains urban areas as well as surrounding desert, agricultural, mining, and mountainous areas. Elevation ranges from ~ 600 to over 2800 m above sea level (a.s.l.), with peaks in the Santa Catalina, Rincon, and Santa Rita Mountains to the north, east, and south of the city, respectively (Fig. 1).

Tucson's ozone monitoring network is typical of those in many urbanized areas. For example, the sampling density, which is approximately one monitor per 240 km^2, is equivalent to that of Atlanta, Georgia, a heavily studied area with respect to ozone pollution. The monitors exist in a range of environments and are scattered across the Tucson metropolitan area. Since the monitors are expensive, they have been placed in locations where measurement redundancy over space is minimized. Therefore, linear regression-hindering spatial autocorrelation should not exist. Unfortunately, monitors do not exist in rural areas, such as desert and forest areas, or in industrial areas, which limits linear regression's spatial extrapolation capabilities. Between 1995 and 1998, the ozone monitoring network included an upwind, semi-rural monitor [Tangerine (TANG)], two upwind, suburban monitors [Pomona (POM) and River (RIV)], a downtown monitor (DT), a downwind, suburban monitor [22nd and Craycroft (22&C)], two downwind, urban-fringe monitors [Hidden Valley (HV) and Saguaro National Park East (SNP)], and a downwind, semi-rural monitor [Fairgrounds (FG)]. POM and RIV were not in operation simultaneously.

The highest ozone concentrations occur between April and September with a persistent mountain-valley circulation responsible for transporting pollutants eastward throughout the afternoon, thereby causing SNP typically to have the highest ozone concentrations of all the monitors. The lowest concentrations occur at DT due to the removal of ozone by freshly emitted NO (i.e. NO-scavenging; Diem and Comrie, 2001a). The US National Ambient Air Quality Standard (NAAQS) for ozone has never been violated in Tucson; however, some monitor-less areas may have ozone concentrations that exceed the NAAQS.

3. Data

The modeling and associated methods required six types of data: ozone concentrations, VOC and NO_x emissions estimates, meteorological values, land use information, population information, and a digital elevation model (DEM). Hourly ozone concentrations from April to September of 1995–1998 were obtained from the United States Environmental Protection Agency's (EPA) Aerometric Information Retrieval System (AIRS). Data for the complete time period were available for 22&C, DT, FG, SNP, and TANG. Ozone data for HV, POM, and RIV existed from April 1995 to August 1996, from August 1997 to September 1998, and from April 1995 to June 1996, respectively. Gridded (500 m), multi-temporal estimates of atmospheric VOC and NO_x emissions were developed separately and are described in Diem and Comrie (2000, 2001b). These inventories contain estimates of average daily emissions for each month and type of day (i.e. weekday and weekend) within each month. Daily maximum temperature, daily minimum relative humidity, and average daily atmospheric pressure data collected by the National Weather Service (NWS) at Tucson International Airport (TIA) from 1995 to 1998 were acquired from the National Climatic Data Center (NCDC). Hourly wind speed, wind direction, and insolation data measured at The University of Arizona's Campus Agricultural Center from 1995 to 1998 were acquired from the Arizona Meteorological Network (AZMET). Hourly wind speed and wind direction data collected at several air quality monitoring sites scattered throughout the metropolitan area from 1995 to 1998 were obtained from AIRS. Spatially

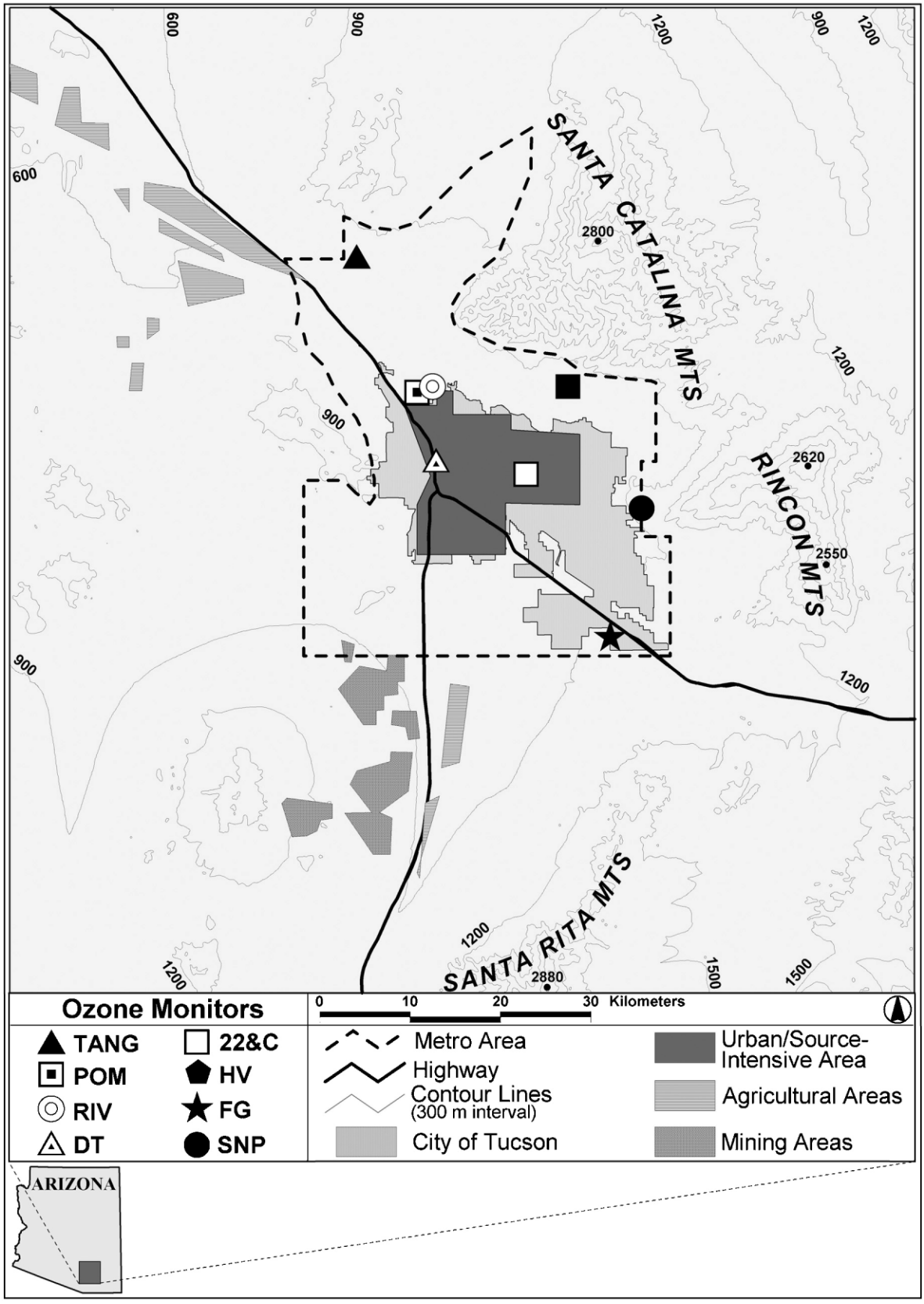

Fig. 1. Map of the Tucson region showing the City of Tucson, the urban/source-intensive area, the metropolitan area, topography, agricultural areas, mining areas, and the eight ozone monitoring sites. The numbers correspond to the elevations (m a.s.l.) of the contour lines and mountain peaks.

resolved 1995 land use data for the region were provided by The University of Arizona's School of Renewable Natural Resources. A spatially resolved 1997 database of the region's street network was provided by the Pima County Department of Transportation. Spatially resolved 1998 population estimates for the region were provided by the Pima Association of Governments. Finally, a high-resolution (30 m) DEM was acquired from the Arizona Regional Image Archive (ARIA).

4. Methods

The methods presented in this paper included (1) variable creation, (2) temporal clustering, (3) variable screening, (4) multiple linear regression modeling, (5) model evaluation, and (6) the creation of ozone design value maps (Fig. 2). The methods consisted primarily of data reduction techniques (i.e. cluster analysis and PCA) so that parsimonious regression models could eventually be developed. The models were based on the theory that ozone concentrations are a function of local emissions of VOCs and NO_x as well as transported VOCs, NO_x, and ozone from upwind areas.

4.1. Model development

4.1.1. Variable creation

Gridded emissions inventories (500 m resolution) were manipulated within a GIS to produce potential predictor variables of the same spatial resolution. Based on the success of Briggs et al. (1997) with using neighborhood totals of emission proxy values (i.e. traffic volume and land cover) as predictor variables in a regression-based nitrogen dioxide model, similar variables were created for this study. The spatial variables used in this study consisted of actual emissions variables, variables that were a proxy for ozone transport, and an exposure variable. The variables represented various physical and chemical processes affecting daily maximum ozone concentrations, which typically occurred during the early afternoon (i.e. 13:00–14:00).

Since motor vehicles are the most important VOC and NO_x source in the Tucson region (Diem and Comrie, 2001b), predictor variables included motor vehicle VOC and NO_x (MVOC and MNO_x) emissions estimates that were specific to mid-mornings and entire days. Hourly motor vehicle emissions between 08:00 and 10:00 were estimated to account for weekday vs. weekend differences in NO-scavenging during the morning rush-hour period (i.e. weekdays have more traffic and more scavenging). These values were calculated by applying estimated hourly coefficients to the daily emissions estimates. Daily emissions were assigned to the 11:00 to 13:00 interval (i.e. mid-day) since weekday vs. weekend differences in daily emissions were similar to differences in mid-day emissions.

Emissions from other anthropogenic and biogenic sources were only determined on a daily basis; therefore, daily emission totals were assigned to the mid-day period while emissions specific to the morning period were not estimated. The diversity of anthropogenic sources made it difficult to generalize their diurnal emissions profiles and thus estimate generalized morning and mid-day emissions factors. It was reasonable to associate daily biogenic VOC (BVOC) emissions totals with the mid-day period since it is likely that, similar to Atlanta, over 75% of daily BVOC emissions occur between 10:00 and 18:00 (Geron et al., 1995).

Neighborhood emissions estimates were calculated at each grid cell for many temporal situations (e.g. weekday morning in May). Local and upwind MVOC and MNO_x, other anthropogenic VOC and NO_x (OVOC and ONO_x), and BVOC emissions at each ozone monitor were included in the model development process (Table 1). At each cell, cumulative emissions were calculated for short radial distances (2 and 5 km) and longer wedge-like distances (10 and 20 km). Six, 60° wedges (centered at 45, 105, 165, 225, 285, and 345°, respectively) extended towards the direction from which the wind was coming during a certain time period (e.g. 11:00 to 13:00).

To account for the transport of ozone, cumulative totals of built-up land (BU; i.e. commercial land, industrial land, malls, mines, schools, major roadways, railway yards, and airports) and road length (RL) were calculated for long wedge-like distances (20, 30, and 40 km). Theoretically, as cumulative totals of built-up land and road length upwind of a cell increase so should the amount of ozone transported to that cell. This is evident in Tucson, for SNP is downwind of the urban area and has the highest ozone concentrations resulting from pollutant transport (Diem and Comrie, 2001a). These proxy variables were intended to explain variance in the ozone data that was not accounted for by the emissions variables.

A final variable was included to make a distinction between cells that were exposed fully to the major air pollution plumes and those that were relatively isolated from the plumes. This variable, EXP, was developed within a GIS by conducting a visibility analysis on a high-resolution DEM. Approximately 40 observation points within the metropolitan area were used in the visibility analysis to determine which cells were exposed directly to pollution within the metropolitan Tucson area.

The emissions variables were multiplied by the meteorological variables to increase the temporal variation among the emissions variables and to possibly make the emissions more representative of actual emissions on any given day. The meteorological variables

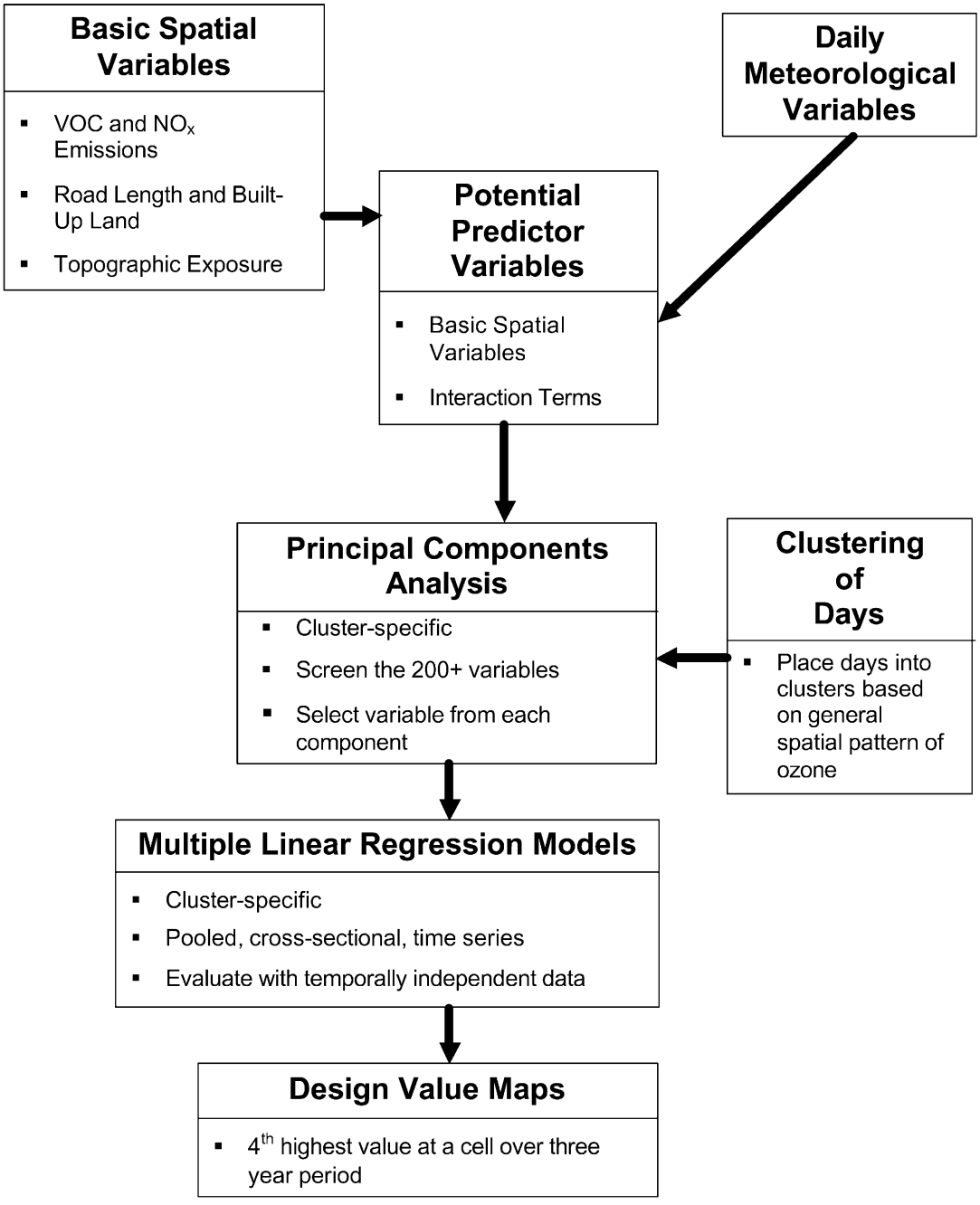

Fig. 2. Flow chart of the methods used to develop the ozone maps.

were as follows: maximum daily temperature (MAXT), minimum relative humidity (MINRH), inverse of average daily wind speed (IWS_D), inverse of 08:00–10:00 wind speed (IWS_{8to10}), inverse of 11:00–13:00 wind speed (IWS_{11to1}), average daily atmospheric pressure (PRESS), and total daily insolation (INSOL). Meteorology influences pollutant emissions, and the impacts of meteorology vary depending on the emissions source. For example, ONO_x emissions, which are dominated in the summer months by emissions from a local electric power generating facility, should increase with increases in temperature and relative humidity; hot and humid conditions lead to increased air conditioner use and thus increased NO_x emissions from the power plant.

4.1.2. Temporal clustering

Many different emissions/meteorological situations existed among the days considered in this study. For instance, local ONO_x emissions might have been the most important predictors on some days, while on other days long-distance MVOC emissions might have dominated. Therefore, a single model for all of the days probably would have had difficulty identifying important predictive relationships among the variables. By

Table 1
Attributes of basic spatial variables

Spatial variable	Description	Wind direction times	Single cell	Radius of circle (km)		Radius of wedge (km)[a]				Wedge differences (km)[a]		
				2	5	10	20	30	40	(40−30)	(40−20)	(30−20)
MVOC	Motor vehicle VOCs	08:00–10:00; 11:00–13:00		X	X	X	X					
MNO_X	Motor vehicle NO_X	08:00–10:00; 11:00–13:00		X	X	X	X					
OVOC	Other anthropogenic VOCs	11:00–13:00		X	X	X	X					
ONO_X	Other anthropogenic NO_X	11:00–13:00		X	X	X	X					
BVOC	Biogenic VOCs	11:00–13:00		X	X	X	X					
BU	Built-up land	11:00–13:00					X	X	X	X	X	X
RL	Road length	11:00–13:00					X	X	X	X	X	X
EXP	Exposed	NA[b]	X									

[a] Wedges are 60° and extend towards the direction from which the wind is coming during the various wind direction times.
[b] NA, not applicable.

clustering the days, the unique cluster-specific dependence of ozone on emissions/meteorological interactions could be captured. Clustering leads to a better specification of significant predictor variables in the regression models, thereby resulting in smaller errors and larger coefficients of determination (R^2) than one would achieve with a non-clustering approach (Davis et al., 1998).

Days were placed into clusters based on monitor-specific deviations from the average daily maximum ozone concentration so that the clusters represented several dominant ozone patterns. The average daily maximum concentration was the average of the daily maximum values at the five core monitors (TANG, DT, 22&C, SNP, and FG). Days with missing values at one or more of the monitors were excluded from the analysis. Consequently, approximately 10% of the original 732 days were not clustered.

The remaining days were clustered with K-means clustering. K-means clustering is a non-hierarchical iterative clustering method that requires the number of clusters and cluster "seeds" to be specified initially. It allows for the reclassification of days after they have been placed into an initial cluster (Davis et al., 1998). In this study, iterations ended when the locations of cluster centers remained stable in five-dimensional (i.e. five monitors) ozone deviation space.

Solutions were found for a range of cluster numbers. The optimal number of clusters was based on the following criteria: (1) each cluster had to have at least 40 days (~240 observations) to ensure a robust sample (at least 80 observations) for independent analysis; and (2) there should have been a relatively small change in an R^2 value, which was derived from mean monitor-specific ozone deviations per cluster regressed against actual ozone deviations, when moving from the chosen number of clusters to a larger number of clusters.

Moran's I tests were used to test for spatial autocorrelation among ozone concentrations (measured at six to seven monitors) for each cluster. Significant spatial autocorrelation was deemed present when the Z-value at either a 10, 15, 20, or 25 km lag was significant at the $\alpha = 0.05$ level. However, the distribution of the Moran's I test statistic is asymptotically normal, thus the distribution may not be normal for small samples (i.e. less than 50) and the use of the normal distribution could lead to mistaken inferences (Odland, 1988). Therefore, examinations of plots of distance between monitors vs. absolute difference in ozone concentrations between monitors were also used to detect spatial autocorrelation among ozone concentrations. These examinations involved 26 cases for each cluster. A significant ($\alpha = 0.05$) correlation between distance and difference was considered an indicator of spatial autocorrelation.

4.1.3. Variable screening

Over 200 potential predictor variables were created, so principal components analysis (PCA) was used to reduce the list of variables to a much smaller number of reasonably uncorrelated variables that represented the various emissions categories. A standardized PCA with rotation (Varimax) of all appropriate variables was performed for each cluster. Only the most important components (i.e. eigenvalue greater than one) were extracted. From each of the important components, a single variable was selected that had both a relatively high loading and a relatively high correlation with the dependent variable (i.e. ozone deviation). The selected variables therefore did not exhibit strong multicollinearity and were used as predictor variables in multiple linear regression analyses.

4.1.4. Multiple linear regression modeling

Multiple linear regression models were developed for each cluster using a random selection of two-thirds of the available days. The remaining one-third of the days were used as independent data during model evaluations. The dependent variable in the multiple linear

regression models was the deviation from the area-wide mean ozone concentration. The predicted residual surface was then added to the mean ozone concentration to yield predicted daily maximum ozone concentrations. Each model was initially developed under the assumption that every variable (from the standardized PCA) belonged in the final model. Backward variable elimination removed the variables from the model one at a time based on removal criteria. In this study, a variable was removed if its F-value is greater than 0.05. This criterion prevented the inclusion of confounding variables in the models.

4.2. Model evaluation

Even though the multiple linear regression models were developed almost entirely for predictive purposes, linear regression assumptions were still obeyed to a considerable degree. Even so, since this paper does not deal explicitly with the inferential aspects of multiple linear regression, diagnostic tests were not used to determine whether or not assumptions had been violated. Model evaluation consisted primarily of analyses of residuals to assess model bias and misspecification.

Inaccurate predictions should have been the result of (1) biased model coefficients, (2) errors associated with the spatial variation in the independent variables, and (3) model misspecification (Miron, 1984; Heuvelink et al., 1989). The first two causes of error were difficult to remove, for they stemmed from errors in the gridded emissions inventories, which had been made as error-free as possible (Diem and Comrie, 2000, 2001b). However, model misspecification, such as the exclusion of important spatial and temporal predictor variables, could be detected and fixed. Thus, model-misspecification tests were conducted that consisted of Moran's I tests of errors and cluster-specific examinations of plots of distance between monitors vs. absolute difference in residuals (i.e. predicted minus observed values) between monitors. For the Moran's I tests, significant spatial autocorrelation was deemed present when the Z-value at either a 15, 20, or 25 km lag was significant at the $\alpha = 0.05$ level, while a significant ($\alpha = 0.05$) correlation between distance and residual difference also signified spatial autocorrelation. In addition, year-, month-, and day-specific (i.e. weekday and weekend) errors were examined to determine if large errors tended to occur during a certain year, month, or day. The presence of spatial autocorrelation and temporally varying errors denoted the exclusion of important spatial variables and temporal variables, respectively.

Model accuracy was evaluated with temporally independent data for a pooled model (pooling of predictions from the models) and predictions on high ozone days (HODs). HODs were defined as days that had region-wide daily maximum ozone concentrations (i.e. average of daily maximum ozone concentrations at 22&C, DT, FG, SNP, and TANG) that were in the top 10% of all summer (April–September) values from 1995 to 1998. Knowing the accuracy of extreme values was important when evaluating the validity of ozone design value maps, which are composites of extreme ozone values. Fully independent data, which were both spatially and temporally independent, could not be used for the evaluation. The monitors represented unique locations on the ozone precursor emission landscape; therefore, each monitor had critically important values for many of the predictor variables. Consequently, evaluating the models with "jack-knifed" data (i.e. withholding one of the monitor's data from the modeling process) would have resulted in unreliable ozone predictions at most of the monitors.

Multiple evaluation statistics were used to determine modeling accuracy. The coefficient of determination (R^2), root mean squared error (RMSE), mean biased error (MBE), percent error, index of agreement (D_1), and the proportion of systematic error (PSE) were calculated. D_1 is a dimensionless measure of the degree to which model predictions are error-free (Willmott, 1981). It ranges from 0.0 (complete disagreement between predicted and observed values) to 1.0 (perfect agreement between predicted and observed values). The proportion of systematic error (PSE) was calculated by dividing the mean-squared error that was systematic by the total mean-squared error (MSE_S/MSE). Low PSE values are optimal, for systematic error is model-derived while unsystematic error represents the natural variability of the data that cannot be reduced by a model (Wilmott, 1981; Comrie, 1997; Comrie and Diem, 1999).

4.3. Creation of predicted design value maps

An air planning area is in violation of the NAAQS for ozone if, at a particular monitor, the fourth highest daily maximum 1-h average ozone concentration over the past 3 years exceeds 125 ppb (parts per billion). The linear regression models were employed to produce many ozone maps that were then stacked and the fourth highest value at each cell was extracted. Maps of the fourth highest values (i.e. design value maps) were created and subsequently examined to determine possible ozone standard exceedance areas and to estimate the percentage of the region's population that might have been exposed to air pollution that exceeded the NAAQS.

5. Results and discussion

5.1. Clustering of days

The cluster analysis results in five clusters based on spatial ozone patterns among the monitors. The clusters

differ with respect to month of occurrence, atmospheric conditions, day of occurrence, and magnitude and spatial variation of ozone levels (Table 2).

Clusters 1 and 2 occur most frequently in August, cluster 3 occurs mostly in July and August, cluster 4 occurs most frequently in April, and cluster 5 occurs mostly in May and June. Clusters 1, 2, and 3 have higher temperatures and more westerly mid-day winds than do clusters 4 and 5. Clusters 4 and 5 are associated with windy mornings (or days) as well as calm, and, most importantly, relatively cool conditions at mid-day. Clusters 1 and 4 occur most frequently on weekends, cluster 5 occurs mostly on weekdays, and clusters 2 and 3 are neither weekday- nor weekend-biased.

The clusters can also be differentiated in terms of their average ozone concentrations and the spatial variation in ozone concentrations. Clusters 1, 2, and 3, the warm clusters, have high ozone concentrations while clusters 4 and 5, the cool clusters, have low concentrations. Cluster 1 has relatively higher ozone concentrations in the urban/source-intensive area (i.e. DT and 22&C) and lower concentrations in the semi-rural areas (i.e. FG and TANG). The highest levels and lowest levels, respectively, in cluster 2 and 3 occur east/southeast (downwind) and west/northwest (upwind) of the urban/source-intensive area. Ozone transport to areas east of the urban area prevails among these clusters. Ozone concentrations in cluster 4 exhibit little spatial variation. Cluster 5 has substantially depressed concentrations in downtown Tucson while also having relatively homogeneous concentrations throughout the rest of the metropolitan area.

Moran's I tests verify the presumed absence of significant ($\alpha = 0.05$) spatial autocorrelation among ozone concentrations for all of the clusters. Plots of distance between monitors vs. absolute difference in ozone concentrations between monitors also reveal no significant ($\alpha = 0.05$) spatial autocorrelation among ozone concentrations (Fig. 3). The ozone measurements at the monitors are therefore spatially independent of one another. For space reasons, detailed results from the Moran's I tests are not presented.

5.2. Principal components analysis

The cluster-specific standardized PCAs results in 17, 15, 17, 17, and 18 components, respectively, for clusters 1, 2, 3, 4, and 5. The components explain at least 96% of the variance in the predictor variable data sets. The number of components (which are orthogonal) denotes the maximum number of predictor variables to be included in the linear regression models. For each component, a single variable is selected that has both a relatively high loading and a relatively high correlation with the dependent variable. Every possible type of spatial variable (refer to Table 1) is represented by at least one of the cluster-specific components.

5.3. Multiple linear regression modeling

The models predict deviations from average daily maximum ozone concentrations thus the models' predicted concentrations are added to the average concentrations to produce final predictions. The frequency distributions of the deviations were approximately normally distributed. The cluster-specific models are developed with an initial group of 15–18 predictor variables. Backward stepwise regression (F-remove = 0.05) results in models containing between 5 and 10 variables. Final variables are summarized in Table 3, which include the letter codes used to detail the individual models in Table 4. All of the spatial variable categories (i.e. proxy variables and short and long distance VOC and NO_x emissions variables) and all of the meteorological variables are present within one or more of the predictor variables.

Table 2
Characteristics of the five clusters based on spatial ozone patterns[a]

Cluster	Total days	Mode month	Day type	Mean															
				22&C	DT	FG	HV	POM	RIV	SNP	TANG	Ozone	MAXT	MINRH	INSOL	MWS	NWS	MWD	NWD
1	105	August	WE	73	64	58	58	71	73	67	60	64	36.8	20.3	25.99	1.8	1.8	134	268
2	45	August	Both	79	59	66	62	72	68	82	55	68	37.0	25.7	26.04	1.3	1.7	144	283
3	109	July August	Both	67	54	67	54	63	61	71	55	63	35.9	20.3	26.51	1.6	2.2	145	282
4	231	April	WE	59	50	55	51	59	61	57	58	56	33.1	15.7	26.06	2.1	2.8	139	231
5	164	May June	WD	59	44	55	53	58	60	63	56	55	33.5	18.0	26.00	2.1	2.8	144	244

[a] Values for 22&C, DT, FG, HV, POM, RIV, SNP, and TANG are average ozone concentrations (ppb) at those ozone monitors while Mean Ozone is the average of concentrations at 22&C, DT, FG, SNP, and TANG. Concerning the type of day on which the clusters occur, WD refers to weekdays, WE refers to weekends, and both refers to both weekdays and weekends. The remaining variables pertain to meteorology and are as follows: MAXT, daily maximum temperature (°C); MINRH, minimum relative humidity (%); INSOL, total daily insolation (MJ m^{-2}); MWS, wind speed (m s^{-1}) between 08:00 and 10:00; NWS, wind speed (m s^{-1}) between 11:00 and 13:00; MWD, wind direction (°) between 08:00 and 10:00; and NWD, wind direction (°) between 11:00 and 13:00.

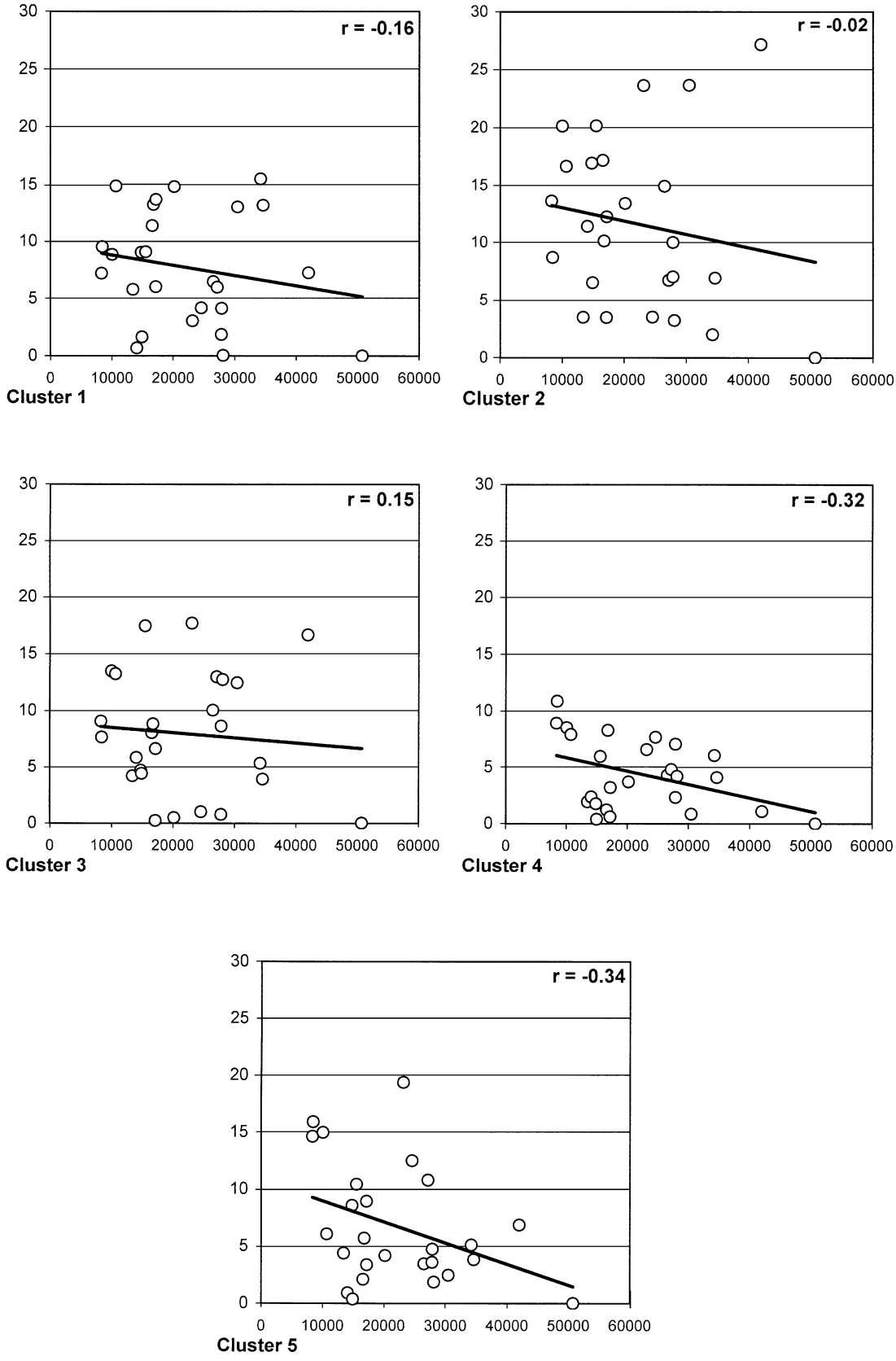

Fig. 3. Scatter plots of distance between monitors vs. absolute difference in ozone concentrations between monitors for each of the five clusters. Absolute difference in ozone concentrations (ppb) is on the y-axis while distance between monitors (m) is on the x-axis. The linear trend line and correlation values (r) are included on the plots. None of the trend lines have slopes that are significantly ($\alpha = 0.05$) different than zero.

Table 3
Cluster-specific linear regression models including predictor variables and associated standardized coefficients (β) and unstandardized coefficients

	β	Unstandardized coefficients
Cluster 1 variables		
Q	0.7265	4.1268×10^{-7}
I	−0.2743	-1.7272×10^{-8}
T	−0.2730	-2.2207×10^{-5}
H	0.2588	1.6161×10^{-6}
AB	0.1978	0.0071
E	0.1514	9.9718×10^{-9}
M	−0.1180	-4.7609×10^{-9}
(Constant)	–	−0.0010
Cluster 2 variables		
X	0.5356	4.8446×10^{-12}
G	0.3475	2.4632×10^{-9}
AB	0.2389	0.01550
P	−0.1525	-2.3344×10^{-9}
N	−0.1286	-1.0765×10^{-8}
(Constant)	–	−0.0259
Cluster 3 variables		
M	−0.4465	-1.98762×10^{-8}
AB	0.3684	-3.4253×10^{-9}
P	−0.3315	0.0128
G	0.2713	1.3448×10^{-9}
L	0.2157	2.6470×10^{-10}
V	0.1907	2.4804×10^{-9}
C	−0.1289	-1.0480×10^{-7}
(Constant)	–	−0.0131
Cluster 4 variables		
J	−0.7770	-4.7059×10^{-9}
S	0.6516	3.4325×10^{-7}
AB	0.4489	0.0078
U	−0.3973	-1.4694×10^{-8}
A	0.2028	2.9557×10^{-8}
F	0.1437	4.9911×10^{-7}
Y	0.1386	1.8557×10^{-7}
Z	−0.1283	-1.1370×10^{-5}
O	0.0904	2.5468×10^{-8}
(Constant)	–	−0.0063
Cluster 5 variables		
J	−1.3736	-1.1063×10^{-8}
R	0.6667	1.4947×10^{-8}
D	0.3540	8.6981×10^{-6}
AB	0.3340	0.0127
K	0.2448	4.9931×10^{-10}
U	−0.2374	-1.2554×10^{-8}
AA	0.1784	7.8537×10^{-9}
L	−0.1311	-1.6520×10^{-10}
W	−0.1126	-1.3318×10^{-9}
B	−0.0883	-1.5076×10^{-8}
(Constant)	–	−0.0085

As noted previously, the models are predictive rather than explanatory, however, we have confidence that most of the linear regression assumptions have been obeyed and that none of the assumptions have been violated severely. Therefore, denoting the relative importance of each predictor variable in each model, which is an explanatory procedure, is executed to merely highlight the differences between the models. The value of the standardized coefficient (β) is used as a proxy for importance.

Table 4
Letter codes and descriptions of predictor variables in the 1-h regression models[a]

Letter code	Description
A	MVOC (20 km)$_{8\ AM}\times$MAXT
B	MNO_X (20 km)$_{8\ AM}\times$MAXT
C	MVOC (10 km)$_{8\ AM}\times$MINRH
D	MNO_X (5 km)$_{8\ AM}\times IWS_{Daily}$
E	MVOC (20 km)$_{11\ AM}\times$MINRH
F	MVOC (20 km)$_{11\ AM}\times IWS_{Daily}$
G	MVOC (10 km)$_{11\ AM}\times$PRESS
H	MNO_X (10 km)$_{11\ AM}\times IWS_{11\ to\ 1}$
I	MVOC (5 km)$_{11\ AM}\times$MAXT
J	MNO_X (2 km)$_{11\ AM}\times$PRESS
K	OVOC (20 km)×PRESS
L	ONO_X (20 km)×PRESS
M	OVOC (10 km)×MAXT
N	ONO_X (10 km)×INSOL
O	ONO_X (5 km)×MINRH
P	ONO_X (5 km)×PRESS
Q	OVOC (2 km)×MAXT
R	OVOC (2 km)×PRESS
S	OVOC (2 km)×INSOL
T	ONO_X (2 km)$\times IWS_{11\ to\ 1}$
U	ONO_X (2 km)×PRESS
V	BU (40 km)×PRESS
W	RL (40 km)$\times IWS_{11\ to\ 1}$
X	RL (40 km)×PRESS
Y	BU (40−30 km)×MINRH
Z	BU (40−30 km)$\times IWS_{Daily}$
AA	BU (40−30 km)×PRESS
AB	EXP

[a] Refer to Table 1 for descriptions of emissions variables. Meteorological variables are as follows: MAXT, daily maximum temperature; PRESS, average daily pressure; $IWS_{11\ to\ 1}$, inverse of 11:00–13:00 wind speed; IWS_{Daily}, inverse of average daily wind speed; INSOL, total daily insolation; and MINRH, daily minimum relative humidity.

Most of the models contain a member of each broad, non-BVOC, spatial emissions category (i.e. local and upwind motor vehicle and other anthropogenic emissions). Variables related to BVOC emissions are absent presumably due to the relatively smooth variations in BVOC emissions across the metropolitan area. Consequently, BVOCs are not strong spatial predictors. VOC and NO_x variables and the exposure (EXP) variable are always present. Specifically, the cluster 1 model does not contain long distance, other anthropogenic emissions variables and proxy emissions variables (i.e. road length and built-up land). Local, other anthropogenic VOC emissions are the most important predictors. The cluster 2 and 3 models do not contain local, motor vehicle variables. The most important predictors in these models are road length and long distance, other anthropogenic VOC emissions, respectively. Cluster 2

represents summer situations where ozone production in Tucson is high, and the transport of ozone and its precursors eastward across the metropolitan area is an extremely important process. The cluster 4 model does not contain long distance, other anthropogenic variables while the cluster 5 model contains all of the non-BVOC variables. The most important predictors in these models are local, motor vehicle NO_x emissions, hence the scavenging of ozone by nitric oxide (NO) is a major controlling factor of ozone on days in clusters 4 and 5. Hence the overall low ozone concentrations and the extremely low concentrations at the DT monitor, especially on cluster 5 days. Due to increased motor vehicle traffic on cluster 5 days, which are mostly weekdays, these days have higher NO_x emissions and consequently more scavenging than do cluster 4 days, which are mostly weekend days.

5.4. Examination of cluster-specific errors

It is reasonably safe to assume that none of the models have excluded important predictor variables. Results from the Moran's I tests indicate that there is no significant ($\alpha = 0.05$) spatial autocorrelation among the errors for any of the clusters, while examinations of plots of distance between monitors vs. absolute difference in residuals between monitors reveal significant ($\alpha = 0.05$) spatial autocorrelation among residuals only for cluster 3 (Fig. 4). The possible spatial autocorrelation present in the cluster 3 model is caused by relatively large underpredictions at TANG compared to the rest of the monitors. However, the errors at all the monitors, including TANG, are relatively small compared to errors for the other models. For space reasons, results from the Moran's I tests and temporal examinations are not presented. In addition, large errors are not concentrated during certain years, months, weekdays, or weekends, thus important temporal variables were not excluded from the models. Finally, none of the monitors have anomalously large errors (Table 5). Large errors are a flag for a misspecified model. For example, if the EXP variable were not included in the models, predicted ozone concentrations at HV would be considerably more erroneous than those at the rest of the monitors.

5.5. Evaluation of pooled predictions

The models are evaluated through the examination of cross-validated error statistics (Table 6). Predictions of daily maximum 1-h average ozone concentrations are typically within 4 ppb (RMSE) of the observed values, and are neither positively nor negatively biased (MBE). In addition, the pooled predictions have just 7% error as well as an R^2 value of 0.90 and a D_1 value of 0.97. The PSE value of 0.08 indicates that 8% of the error is model-derived, and that the remaining 92% is natural variability of the data that cannot be reduced. Presumably, much of this error is derived both from inaccurate estimates of pollutant emissions and from spatial aggregation. The urban/source-intensive monitors, 22&C and DT, have the most accurate predictions while HV, POM, and RIV, the three monitors with the least historical data in the model development process, have the least accurate predictions.

Table 5
Root mean squared error values in ppb at the ozone monitors for each of the five clusters

Site	1	2	3	4	5
22&C	4.2	4.7	3.5	2.9	2.8
DT	4.5	4.5	3.4	3.1	3.1
FG	4.3	6.0	4.1	3.9	4.1
HV	5.9	5.4	4.0	4.2	3.5
POM	6.5	5.8	3.6	4.1	4.7
RIV	6.6	5.5	4.7	4.0	4.2
SNP	5.9	5.7	4.2	3.2	4.4
TANG	4.2	5.1	4.5	3.6	4.1

5.6. Evaluation of HOD predictions

Typically, a linear regression model performs poorly on the tails of the dependent variable's distribution, overestimating the lowest observed values and underestimating the highest observed values (Clark and Karl, 1982). However, developing separate models for each cluster decreases the errors associated with extreme values. The cluster-specific models account for the spatio-temporal processes responsible for spatio-temporal ozone patterns. HOD predictions are almost as accurate as the complete set of pooled predictions (Table 6). HOD predictions have slightly lower R^2 and D_1 values, but have equal or smaller percent error values. However, a considerable increase in PSE values from 0.08 to 0.32 indicates that a greater proportion of model-derived error is present in the HOD predictions. Nevertheless, the HOD predictions are still valid and useful for determining spatial variations in elevated ozone concentrations.

5.7. Maps of predicted ozone design values

Maps of ozone design values for 1997 and 1998 indicate a strong likelihood that an exceedance did not occur anywhere in the Tucson region in either of those years (Figs. 5 and 6). After adjusting the highest design values by adding the average residuals from the 22&C and SNP monitors, the maximum design value (102 ppb) is still less than 82% of the NAAQS (> 125 ppb; Tables 7 and 8). All the cells with design values in the top 1% of the 1997 values (i.e. the black cells) are located either in or downwind of the urban/source-intensive

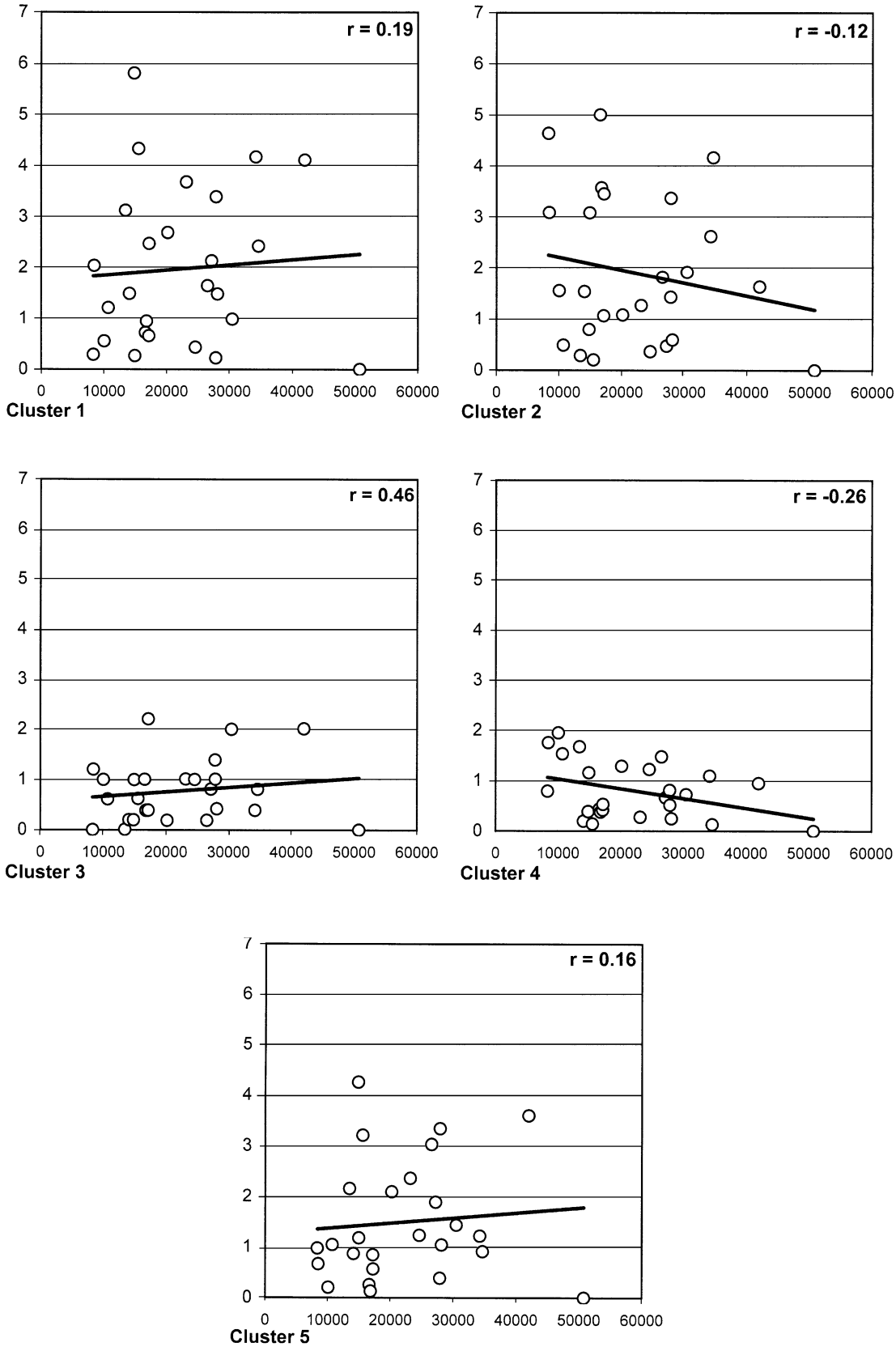

Fig. 4. Scatter plots of distance between monitors vs. absolute difference in residuals between monitors for each of the five clusters. Absolute difference in ozone concentrations (ppb) is on the y-axis while distance between monitors (m) is on the x-axis. The linear trend line and correlation values (r) are included on the plots. Only the trend line for cluster 3 has a slope that is significantly ($\alpha = 0.05$) different than zero.

Table 6
Evaluation statistics for pooled, daily maximum 1-h average ozone predictions for each monitor, all monitors, and HODs[a]

Site/days	n	O_{AVG}	P_{AVG}	O_{SD}	P_{SD}	MBE	RMSE	MAE	D_1	PSE	R^2	% Error
22&C	218	63.9	63.2	12.9	12.4	−0.7	3.3	2.5	0.98	0.12	0.94	5
DT	218	52.0	52.6	12.7	12.7	0.6	3.5	2.9	0.98	0.04	0.93	7
FG	218	58.2	59.3	11.5	11.0	1.2	4.2	3.3	0.96	0.16	0.88	7
HV	59	53.2	54.8	9.5	10.4	1.7	4.4	3.5	0.95	0.15	0.85	8
POM	84	62.5	61.9	12.9	11.7	−0.6	5.4	4.0	0.95	0.19	0.82	9
RIV	69	64.2	63.4	10.3	9.5	−0.8	5.0	3.9	0.93	0.17	0.76	8
SNP	218	64.3	62.9	12.5	12.3	−1.4	4.4	3.4	0.97	0.15	0.89	7
TANG	218	57.2	57.7	10.3	10.3	0.6	3.7	2.9	0.97	0.06	0.88	6
ALL	1302	59.3	59.4	12.7	12.2	0.0	4.0	3.1	0.97	0.08	0.90	7
HODs	149	76.8	76.9	10.8	8.9	0.1	5.5	4.2	0.92	0.32	0.75	7

[a] Descriptions of statistics are as follows: n, number of cases; O_{AVG}, average observed value (in ppb); P_{AVG}, average predicted value (in ppb); O_{SD}, standard deviation of observed values (in ppb); P_{SD}, standard deviation of predicted values (in ppb); MBE, mean biased error (in ppb); RMSE, root mean squared error (in ppb); MAE, mean absolute error (in ppb); D_1, index of agreement; PSE, proportion of systematic error; R^2, coefficient of determination; % Error, O_{AVG}/RMSE.

portion of the region. In 1997, the largest design values occur mostly between the 22&C and SNP monitors (i.e. downwind, suburban area). In 1998, the largest value occurs in the highly industrialized portion of the city while the downwind areas have slightly smaller values. With spatial interpolation methods, such as distance-weighting and kriging, the above potentially high ozone areas would not have been identified.

Both design value maps illustrate the importance of local VOC emissions, local NO_x emissions, especially from motor vehicles, and the predominant transport of VOCs, NO_x, and ozone from west to east across the Tucson metropolitan area. Upwind areas tend to have small VOC and NO_x emissions and receive negligible amounts of transported ozone. Within the source-intensive area, ozone levels are reduced by large NO_x emissions. The highest levels occur either in areas that have both large VOC and NO_x emissions or areas that are slightly downwind of the source-intensive area. More importantly, the downwind areas have more reliable predictions of ozone levels.

Most importantly, the maps show that the City of Tucson, which is the most populated part of the Tucson region, is almost completely covered with reliable predictions of design values. Approximately 90% of the region's population are associated with a reliable design value (Table 8). The remaining 10% are located mostly in the industrial (i.e. NO DATA) areas towards the center of the city (i.e. urban/source-intensive area). Large VOC and NO_x emissions, which are out of the range of the models, make ozone predictions unreliable in those industrial areas.

The maps also indicate that additional ozone monitors are needed in the Tucson region. Based on model results, ozone monitors should be placed in the following areas for the following reasons: (1) near the city's industrial areas to protect public health and improve modeling; (2) in a downwind, semi-rural location to protect public health; and (3) at a downwind, forested, high elevation site (e.g. peaks in the Santa Catalina and Rincon Mountains) to protect forest health and improve modeling. All these areas are substantially populated by sensitive receptors, either humans or trees, and have potentially unhealthy ozone levels.

Table 7
Monitor specific observed and predicted design values and residuals for 1997 and 1998

	1997	1997		1998	1998	
Monitor	Predicted	Observed	Residual	Predicted	Observed	Residual
22&C	96	99	−3.1	93	94	−1.0
DT	82	84	−2.5	81	82	−1.3
FG	84	89	−5.3	81	84	−2.7
SNP	98	101	−3.5	89	94	−5.2
TANG	80	83	−2.7	80	82	−2.6

Model improvement involves increasing the model's accuracy as well as increasing its spatial interpolation and extrapolation capabilities. This can be achieved by placing monitors in industrial and rural areas, which have relatively unreliable design values. These areas include heavy anthropogenic emissions areas and downwind areas (i.e. high elevation, forested areas) with negligible anthropogenic emissions but heavy biogenic emissions. With the dominance of BVOC emissions and the absence of local anthropogenic emissions in the Tucson region's forested areas, a monitor in a high elevation, forested environment would increase the presence and importance of BVOC emissions in the regression models. This monitor would also cause a substantial increase in reliable, predicted ozone levels

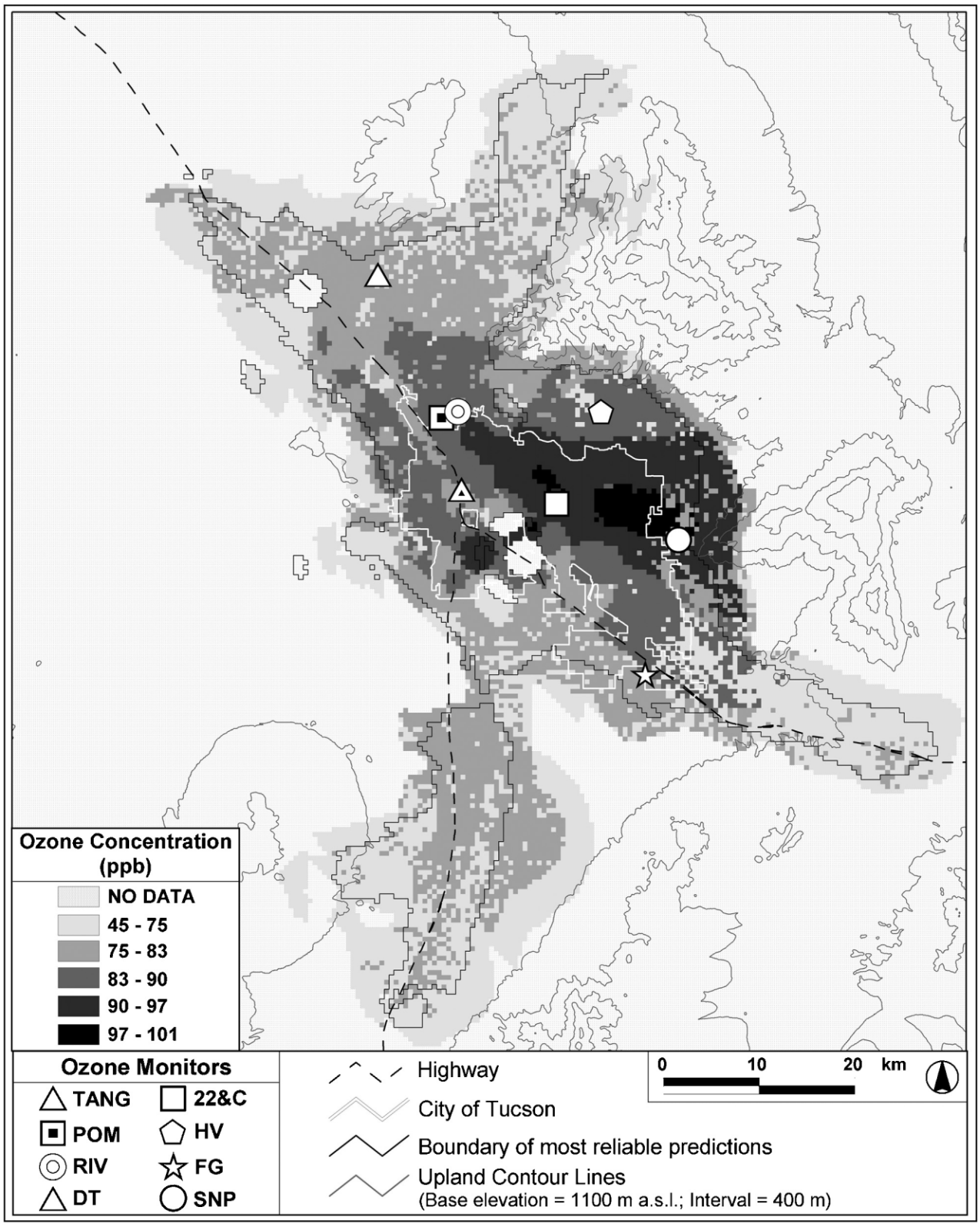

Fig. 5. Predicted 1997 design values of daily maximum 1-h average ozone concentrations in the Tucson region. A design value is the fourth highest value at a cell from 1995 to 1997.

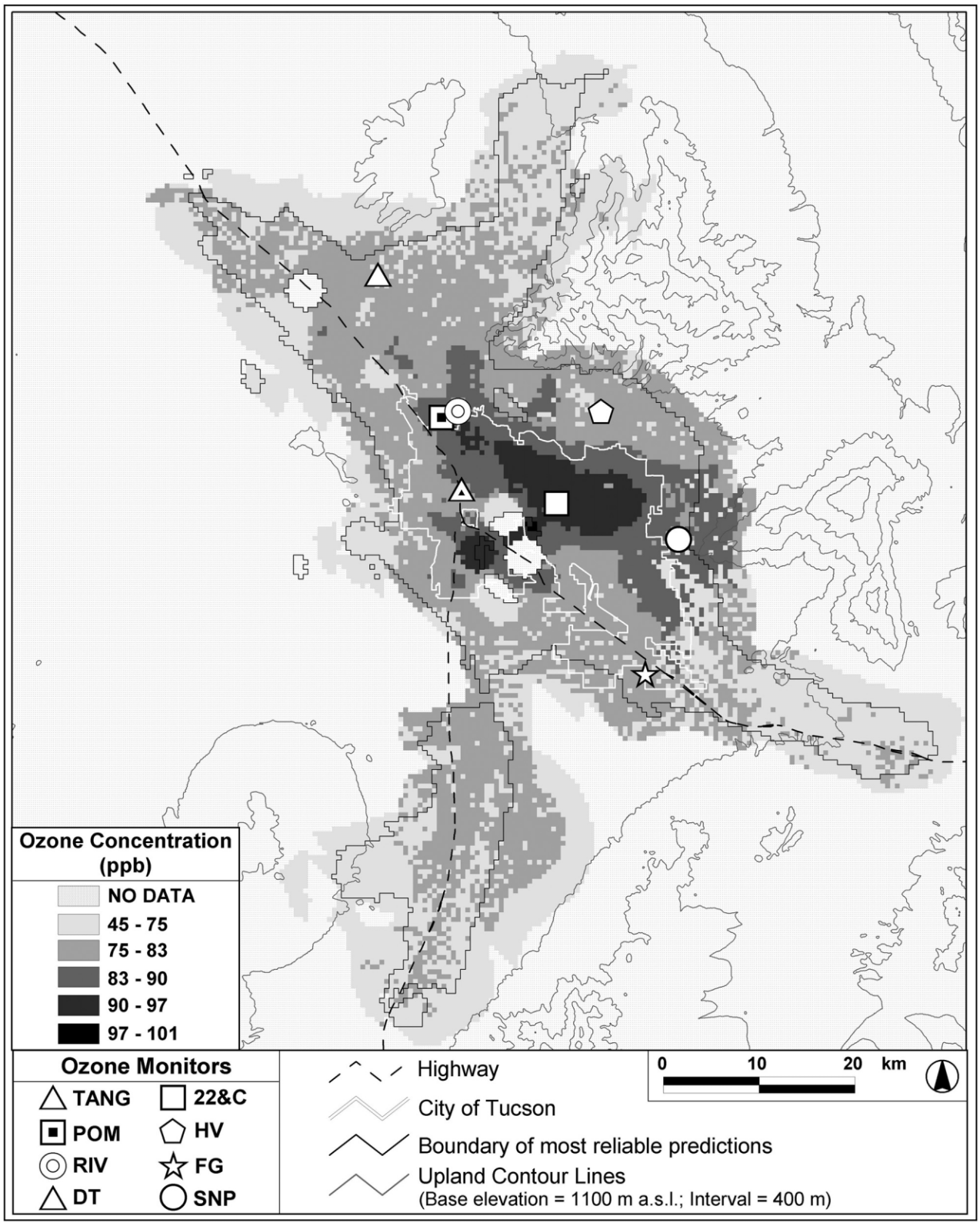

Fig. 6. Predicted 1998 design values of daily maximum 1-h average ozone concentrations in the Tucson region. A design value is the fourth highest value at a cell from 1996 to 1998.

Table 8
Maximum observed and predicted design values for 1997 and 1998

Year	Design value (monitors)	Design value (models)	Design value (adjusted)	% Change	% of Pop.	% of Pop. In Nonattainment[a]
1997	101	99	102	+1	90	0
1998	94	96	99	+5	90	0

[a] % Pop. and % Pop. in Nonattainment refer to the percent of population associated with a valid, predicted design value and percent of population associated with design values above the federal standard, respectively.

throughout the region's rural areas by increasing the range of BVOC, AVOC, and ANO_x emissions used during modeling. Consequently, ozone impacts on humans, crops, and forests could be better assessed.

6. Summary and conclusions

This paper illustrates the potential for overcoming the obstacle of sparse spatial observations in the context of air pollution mapping. A small number of air quality monitors greatly reduces the availability of appropriate mapping methods. Nevertheless, this paper presents a linear regression-based solution that involves the harvesting of multi-temporal measurements at monitors and multi-temporal, spatially continuous predictor variables to compensate for the relative lack of monitors across a region. Without the multi-temporal component, the predictive mapping of air pollution concentrations with few air quality monitors could not be performed adequately via linear regression. For the example presented in this paper, it is possible to use the multi-temporal component because the emissions environment affecting ozone concentrations at each of the monitors varies over time; these fluctuating environments provide the range of information to predict ozone levels over space. Using various predictor variables created partially within a GIS, different spatial processes responsible for the spatial patterns of ozone pollution are represented. The employment of cluster analysis, PCA, and a stepwise regression procedure reduces a large list of potential predictor variables to a reasonable number of variables that, when combined within a linear regression model, explain nearly all the variability in the ozone concentrations.

The regression-based mapping methodology produces accurate maps of ozone levels. Estimates of ozone precursor chemical emissions and proxy variables (e.g. road length) are suitable predictors. The maps illustrate the importance of local emissions of ozone precursor chemicals and the predominant transport of those chemicals and ozone from west to east across the Tucson metropolitan area. Composite maps are extremely important from an air pollution policy perspective, for they not only show where potential exceedances of the NAAQS might be occurring but the maps also provide some insight on suitable locations of future ozone monitors. The accurate mapping of ozone levels provides just one example of the fruitfulness of employing multi-temporal data and multi-variate statistical techniques in a mapping methodology. Although this paper employs linear regression, other predictive methods such as non-linear regression and artificial neural networks are definitely worth exploring in the context of air pollution mapping. Finally, this study's general methodology could certainly be extended to other atmospheric pollutants and to other environmental variables.

Acknowledgements

This research was funded by the Pima Association of Governments. We would like to sincerely thank the anonymous reviewers whose enlightening comments improved the quality of this manuscript considerably.

References

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer, Dordrecht.

Berry, J., 1996. Analyzing spatial dependency within a map. GIS World 9, 28–29.

Briggs, D.J., Wills, J., Elliott, P., Kingham, S., Smallbone, K., De Hoogh, C., Gulliver, J., 2000. A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. Science of the Total Environment 253 (1-3), 151–167.

Briggs, D.J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebret, E., Pryl, K., van Reeuwijk, H., Smallbone, K., van Der Veen, A., 1997. Mapping urban air pollution using GIS: a regression-based approach. International Journal of Geographical Information Science 11 (7), 699–718.

Cardelino, C.A., Chameides, W.L., 1995. An observation-based model for analyzing ozone precursor relationships in the atmosphere. Journal of the Air and Waste Management Association 45, 161–180.

Casado, L.S., Rouhani, S., Cardelino, C.A., Ferrier, A.J., 1994. Geostatistical analysis and visualization of hourly ozone data. Atmospheric Environment 28 (12), 2105–2118.

Chameides, W.L., Fehsenfeld, F., Rodgers, M.O., Vardelino, C., Martinez, J., Parrish, D., Lonneman, W., Lawson, D.R., Rasmussen, R.A., Zimmerman, P., Greenberg, J., Middleton, P., Wang, T., 1992. Ozone precursor relationships in the ambient atmosphere. Journal of Geophysical Research 97 (D5), 6037–6055.

Chock, P.D., Terrell, T.R., Levitt, S.B., 1975. Time series analysis of Riverside, California air quality data. Atmospheric Environment 9, 978–989.

Clark, T.L., Karl, T.R., 1982. Application of prognostic meteorological variables to forecasts of daily maximum one-hour ozone concentrations in the northeastern United States. Journal of Applied Meteorology 21 (11), 1662–1671.

Comrie, A.C., 1997. Comparing neural networks and regression models for ozone forecasting. Journal of the Air and Waste Management Association 47 (6), 653–663.

Comrie, A.C., Diem, J.E., 1999. Climatology and forecast modeling of ambient carbon monoxide in Phoenix, Arizona. Atmospheric Environment 33 (30), 5023–5036.

Cressie, N., 1991. Statistics for Spatial Data. Wiley, New York.

Crown, W.H., 1998. Statistical Models for the Social and Behavioral Sciences: Multiple Regression and Limited-Dependent Variable Models. Prager, Westport, CT.

Daly, C., Neilson, R.P., Phillips, D.L., 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. Journal of Applied Meteorology 33 (2), 140–158.

Davis, J.M., Eder, B.K., Nychka, D., Yang, Q., 1998. Modeling the effects of meteorology on ozone in Houston using cluster analysis and generalized additive models. Atmospheric Environment 32 (14), 2505–2520.

De Leeuw, F.A.A.M., Van Zantvoort, E.G., 1997. Mapping of exceedances of ozone critical levels for crops and forest trees in the Netherlands: preliminary results. Environmental Pollution 96 (1), 89–98.

Diem, J.E., Comrie, A.C., 2000. Integrating remote sensing and local vegetation information for a high resolution, biogenic emissions inventory—application to an urbanized, semi-arid region. Journal of the Air and Waste Management Association 50 (11), 1968–1979.

Diem, J.E., Comrie, A.C., 2001a. Air pollution, climate, and policy: a case study of ozone pollution in Tucson, Arizona. The Professional Geographer (in press).

Diem, J.E., Comrie, A.C., 2001b. Allocating anthropogenic pollutant emissions over space: application to ozone pollution management. Journal of Environmental Management (in press).

Elston, D.A., Jayansinghe, G., Buckland, S.T., Macmillan, D.C., Aspinall, R.J., 1997. Adapting regression equations to minimize the mean squared error of predictions using covariate data from a GIS. International Journal of Geographical Information Systems 11 (3), 265–280.

Geron, C.D., Pierce, T.E., Guenther, A.B., 1995. Reassessment of biogenic volatile organic compound emissions in the Atlanta area. Atmospheric Environment 29 (13), 1569–1578.

Godzik, B., 1997. Ground level ozone concentrations in the Krakow region, southern Poland. Environmental Pollution 98 (3), 273–280.

Greenland, D.A., Yorty, R.A., 1985. The spatial distribution of particulate concentrations in the Denver metropolitan area. Annals of the Association of American Geographers 75 (1), 69–82.

Griffith, D.A., 1992. What is spatial autocorrelation? Reflections on the past 25 years of spatial statistics. L'Espace geographique 3, 265–280.

Griffith, D.A., Layne, L.J., 1999. A Casebook for Spatial Statistical Data Analysis. Oxford University Press, New York.

Heuvelink, G.B.M., Burrough, P.A., Stein, A., 1989. Propagation of errors in spatial modeling with GIS. International Journal of Geographical Information Systems 3 (4), 303–322.

Holland, D.M., De Oliveira, V., Cox, L.H., Smith, R.L., 2000. Estimation of regional trends in sulfur dioxide over the eastern United States. Environmetrics 11 (4), 373–393.

Lam, N.S., 1983. Spatial interpolation methods: a review. The American Cartographer 10 (2), 129–149.

Leenaers, H., Okx, J.P., Burrough, P.A., 1990. Comparison of spatial prediction methods for mapping floodplain soil pollution. Catena 17 (6), 535–550.

Lefohn, A.S., Knudsen, H.P., McEvoy Jr, L.R., 1988. The use of kriging to estimate monthly ozone exposure parameters for the southeastern United States. Environmental Pollution 53 (1-4), 37–42.

Lefohn, A.S., Knudsen, H.P., Logan, J.A., Simpson, J., Bhumralkar, C., 1987. An evaluation of the kriging method to predict 7-h seasonal mean ozone concentrations for estimating crop losses. Journal of the Air Pollution Control Association 37, 595–602.

Lesch, S.M., Strauss, D.J., Rhoades, J.D., 1995. Spatial prediction of soil salinity using electromagnetic induction techniques, 1. Statistical prediction models: a comparison of multiple linear regressiona and cokriging. Water Resources Research 31 (2), 373–386.

Liu, L.J.S., Rossini, A.J., 1996. Use of kriging models to predict 12-hour mean ozone concentrations in metropolitan Toronto—a pilot study. Environment International 22 (6), 677–692.

Loibl, W., Winiwarter, W., Kopsca, A., Zueger, J., 1994. Estimating the spatial distribution of ozone concentrations in complex terrain. Atmospheric Environment 28 (16), 2557–2566.

Mark, D.M., 1984. Some problems with the use of regression analysis in geography. In: Gaile, G.L., Willmott, C.J. (Eds.), Spatial Statistics and Models. D. Reidel, Dordrecht, pp. 191–199.

Mark, D.M., Peucker, T.K., 1978. Regression analysis and geographic models. Canadian Geographer 22 (1), 51–65.

Miron, J., 1984. Spatial autocorrelation in regression analysis: a beginner's guide. In: Gaile, G.L., Willmott, C.J. (Eds.), Spatial Statistics and Models. D. Reidel, Dordrecht, pp. 201–202.

Mulholland, J.A., Butler, A.J., Wilkinson, J.G., Russell, A.G., 1998. Temporal and spatial distributions of ozone in Atlanta: regulatory and epidemiologic implications. Journal of the Air and Waste Management Association 48 (5), 418–426.

Myers, D.E., 1991. Interpolation and estimation with spatially located data. Chemometrics and Intelligent Laboratory Systems 11, 209–228.

Myers, D.E., 1994. Spatial interpolation: an overview. Geoderma 62 (1-3), 17–28.

Odland, J., 1988. Spatial Autocorrelation. Sage, Newbury Park, CA.

Phillips, D.L., Tingey, D.T., Lee, E.H., Herstrom, A.A., Hogsett, W.E., 1997. Use of auxiliary data for spatal interpolation of ozone exposure in southeastern forests. Environmetrics 8 (1), 43–61.

Sillman, S., 1999. The relation between ozone, NO_x, and hydrocarbons in urban and polluted rural environments. Atmospheric Environment 33 (12), 1821–1845.

Tayanc, M., 2000. An assessment of spatial and temporal variation of sulfur dioxide levels over Istanbul, Turkey. Environmental Pollution 107 (1), 61–69.

Westenbarger, D.A., Frisvold, G.B., 1994. Agricultural exposure to ozone and acid precipitation. Atmospheric Environment 28 (18), 2895–2907.

Willmott, C.J., 1981. On the validation of models. Physical Geography 2 (2), 184–194.