Published in final edited form as:

Sci Total Environ. 2012 May 15; 425: 27–34. doi:10.1016/j.scitotenv.2012.02.068.

Principal Component Analysis Optimization of a PM_{2.5} Land Use Regression Model with Small Monitoring Network

Hector A. Olvera¹, Mario Garcia¹, Wen-Whai Li¹, Hongling Yang², Maria A. Amaya⁴, Orrin Myers⁵, Scott W. Burchiel⁶, Marianne Berwick⁵, and Nicholas E. Pingitore Jr.^{3,4}

Hector A. Olvera: holvera@utep.edu; Mario Garcia: mario_garcia@urscorp.com; Wen-Whai Li: wli@utep.edu; Hongling Yang: hyang@utep.edu; Maria A. Amaya: mamaya@utep.edu; Orrin Myers: omyers@salud.unm.edu; Scott W. Burchiel: sburchiel@salud.unm.edu; Marianne Berwick: mberwick@salud.unm.edu; Nicholas E. Pingitore: nick@geo.utep.edu

¹Department of Civil Engineering, University of Texas at El Paso, 500 University Ave. El Paso TX, 79968, U.S.A

²Department of Mathematical Sciences, University of Texas at El Paso, 500 University Ave. El Paso TX, 79968, U.S.A

³Department of Geological Sciences, University of Texas at El Paso, 500 University Ave. El Paso TX, 79968, U.S.A

⁴School of Nursing, University of Texas at El Paso, 500 University Ave. El Paso TX, 79968, U.S.A

⁵Department of Internal Medicine, The University of New Mexico, Albuquerque NM, 87131, U.S.A

⁶Center for Environmental Health Sciences, The University of New Mexico, Albuquerque NM, 87131, U.S.A

Abstract

The use of land-use regression (LUR) techniques for modeling small-scale variations of intraurban air pollution has been increasing in the last decade. The most appealing feature of LUR techniques is the economical monitoring requirements. In this study, principal component analysis (PCA) was employed to optimize an LUR model for PM_{2.5}. The PM_{2.5} monitoring network consisted of 13 sites, which constrained the regression model to a maximum of one independent variable. An optimized surrogate of vehicle emissions was produced by PCA and employed as the predictor variable in the model. The vehicle emissions surrogate consisted of a linear combination of several traffic variables (e.g., vehicle miles traveled, speed, traffic demand, road length, and time) obtained from a road network used for traffic modeling. The vehicle-emissions surrogate produced by the PCA had a predictive capacity greater ($R^2 = .458$) than the traffic variable, Traffic Demand summarized for a 1 km buffer, with best predictive capacity ($R^2 = .341$). The PCA-based method

© 2012 Elsevier B.V. All rights reserved.

Correspondence to: Hector A. Olvera, holvera@utep.edu.

Hector A. Olvera, Phone: (915) 747-6518, Fax: (915) 747-8037

Mario Garcia, Phone: (915) 747-6518, Fax: (915) 747-8037

Wen-Whai Li, Phone: (915) 747-8755, Fax: (915) 747-8037

Hongling Yang, Phone: (915) 747-7001, Fax: (915) 747-8037

Maria A. Amaya, Phone: (915) 747-8583, Fax: (915) 747-8037

Orrin Myers, Phone: (505) 272-0920, Fax: (505) 272-6749

Scott W. Burchiel, Phone: (505) 272-0920, Fax: (505) 272-6749

Marianne Berwick, Phone: (505) 272-0920, Fax: (505) 272-6749

Nicholas E. Pingitore Jr., Phone: (915) 747-5754, Fax: (915) 747-8037

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

employed in this study was effective at increasing the fit of an ordinary LUR model by optimizing the utilization of a PM_{2.5} dataset from small-n monitoring network. In general, the method used can contribute to LUR techniques in two major ways: 1) by improving the predictive power of the input variable, by substituting a principal component for a single variable and 2) by creating an orthogonal set of predictor variables, and thus fulfilling the no colinearity assumption of the linear regression methods. The proposed PCA method, should be universally applicable to LUR methods and will expand their economical attractiveness.

Keywords

Air Pollution; Mapping; Ambient Monitoring; Location-Allocation; Exposure Assessment

1. Introduction

Numerous epidemiological studies demonstrate that significant adverse health effects are associated with exposure to urban air pollution (Adar and Kaufman, 2007; Brook et al., 2004; Krewski et al., 2005; Lewtas, 2007; O'Neill et al., 2006; Pope and Dockery, 2006). In particular, chronic exposure to particulate matter (PM) has been associated with increased mortality, morbidity, and risks of harmful effects to the pulmonary and cardiovascular systems (Dominici et al., 2005; Englert, 2004; Kappos et al., 2004; Pope and Dockery, 2006). Risks associated with certain PM sources, such as traffic-related emissions, and their impacts on populations with specific respiratory diseases are yet to be fully understood. In urban environments PM levels vary drastically at small scales driven by the spatial patterns of traffic and other sources (de Kok et al., 2006) (Hoek et al., 2008; Miller et al., 2007). In this regard, epidemiological study could greatly benefit from accurate characterizations of intra-urban PM pollution. Direct ambient PM measurements are costly and thus economical characterization methodologies are necessary to enhance the quality of epidemiological study, reduce the uncertainties associated with health risk estimates, and improve the effectiveness of environmental health policies.

The land-use regression (LUR) modeling technique successfully assessed small-scale variations of intra-urban air pollution for cities and metropolitan regions in Europe (Beelen et al., 2007; Brauer et al., 2003; Briggs, 2005; Briggs et al., 1997; Briggs et al., 2000; Carr et al., 2002; Morgenstern et al., 2007; Stedman et al., 1997) and North America (Gonzales et al., 2005; Henderson et al., 2007; Jerrett et al., 2007; Moore et al., 2007; Ross et al., 2007; Ryan et al., 2007; Smith et al., 2006). The majority of LUR studies focused on the variation of NO₂, due to the availability of an economical monitoring method and its strong association with traffic-related emissions (Hoek et al., 2008). However, health risks associated with chronic exposure to PM pollution might be misrepresented by assuming the same distribution for PM and NO₂ in an urban environment. The physicochemical transformations that traffic PM undergoes after emission might be drastic and the dispersion patterns from a gaseous pollutant such as NO₂ might be very different from that for an air pollutant with significant density, such as PM (Wilson et al., 2002). Therefore, the advancement of LUR techniques specific for intra-urban PM characterization is necessary.

A comprehensive description of the LUR methodology is provided elsewhere (Jerrett et al., 2007). In short, the LUR methodology seeks to predict pollution concentrations at a given location based on surrounding land use, traffic, and physical characteristics. Specifically, the method employs a linear regression model with pollution concentrations, measured at a limited number of sites, defined as the response variable and emission surrogates, summarized for circular areas around the monitoring locations (buffers), as predictors. The most appealing feature of LUR techniques is the economical monitoring requirements.

However, the approach to defining the number and locations of monitoring sites has varied considerably in the literature (Hoek et al., 2008; Ryan and Lemasters, 2007). Recently a location-allocation approach has shown promise in optimizing the design of monitoring networks for LUR modeling purposes (Kanakoglou et al., 2005). A shortcoming of this method is that it requires existing pollutant data to generate the demand surface. Data required for the location-allocation approach would be obtained from existing or purpose-deployed monitoring networks. Existing monitoring networks, mostly deployed for air quality regulatory compliance, are usually composed of a limited number of monitoring sites. The deployment of extensive preliminary monitoring networks would undermine the attractiveness of both the LUR and the location-allocation methods. Therefore approaches to optimize the use of data from monitoring networks with a limited number of sampling sites are necessary to conserve the economical characteristics of LUR techniques and advance the employment of the location-allocation tools.

In this study a method that employs principal component analysis (PCA) to optimize the input of land use data into an LUR model derived from monitoring networks with a limited number of sites was developed and evaluated. The proposed approach to optimize data from small monitoring networks will strongly complement both the location-allocation and LUR techniques. The PCA-based approach employed in this study is illustrated in Figure 1. The results of this study are part of an ongoing epidemiological study of air pollution and asthmatic Hispanic children in El Paso, Texas, USA.

2. Methods

2.1 Study area

El Paso, Texas, USA is located in southwest Texas bordering Juarez, Chihuahua, Mexico to the south and Sunland Park, New Mexico, USA to the west. The bi-national region has an urbanized area of approximately 903.8 km² and a population of more than 2.7 million (INEGI, 2011; USCB, 2011). Besides traffic-related emissions, other major PM sources include a petroleum refining facility, scrap-metal foundries, an electric power facility, and quarries in El Paso, and brick kilns, assembly factories, and numerous unpaved roads in Juarez (Currey and Pumfrey, 2006). For El Paso, Texas, LUR models for nitrogen dioxide (NO₂) and volatile organic compounds (VOC) have been successfully developed without showing strong interference from emissions generated in Juarez (Gonzales et al., 2005; Smith et al., 2006). In these studies, traffic-related parameters (e.g., traffic intensity and distance to highway) were found to be good predictors of NO₂ in El Paso. Also in both studies, distance to the port of entry (Bridge of the Americas) was used as a predictive variable, and might have partially captured the impact of emissions in Juarez (Gonzales et al., 2005; Smith et al., 2006). Due to political and practical constraints, this study also focused exclusively on El Paso, Texas. Figure 2 shows a map of the region.

2.2 Monitoring network

An ambient air monitoring program was conducted between August 2006 and August 2009 as part of a regional respiratory health study (hereinafter referred as the ARCH study). The PM_{2.5} monitoring network was not originally designed for purposes of LUR modeling. The number of monitoring sites was constrained by the availability of analytical and human resources and the locations of monitoring sites were selected to maximize the spatial coverage of the ARCH cohort based on population density. Twelve of the monitoring sites were located within 300 m of roadways. The Montana Vista site was located within 500 m from a major road. The sites were also located within residential areas where air pollutant exposures were of interest to the ARCH study. Initially, air samplers were deployed at eight sites. After the first year of monitoring, five sites were moved to new locations to increase

the spatial coverage of the network. Two sites were located in residential areas in the outskirts of the study area where PM_{2.5} levels were closest to background levels. All sites were active for at least one year. Weekly PM_{2.5} concentrations were averaged to generate long-term averages per site. As stated above, PM_{2.5} measurements were not always performed concurrently at all sites during the study. The potential of bias in annual averages across sites due to temporal variation was evaluated using data from a continuous measurement site following the approach presented by (Hoek et al., 2002). The Socorro site was used to adjust for temporal variations as the site was considered the least influenced by local sources, which might change over time (e.g. roads or major industrial emission sources). The total network coverage area was 649 km², resulting in a monitoring density of one site for every 50 km². Although the network comprised a small number of sites, its spatial coverage was comparable to other PM LUR studies reported in the literature (Brauer et al., 2003; Henderson et al., 2007; Moore et al., 2007; Morgenstern et al., 2007; Ross et al., 2007). Figure 2 shows the locations of the monitoring sites.

Two sites were co-located with continuous ambient monitoring sites (CAMS) operated by the Texas Commission on Environmental Quality. Data from CAMS sites collected with Tapered Element Oscillating Microbalances (TEOM®; Series 1400A, Thermo Scientific Co., Waltham, MA) was used for quality assurance purposes. Additionally, duplicate samples were collected at all sites in a rotating fashion during the monitoring campaign. Duplicates represented 10 % of the total number of samples collected. Agreement between duplicate measurements was assessed by the coefficient of variation (CV) calculated as the square root of the differences between duplicates divided by the overall mean (Parenteau and Sawada, 2010).

2.3 Sample collection

Weekly (7-day average) ambient PM samples were collected with Thermo Electron Corporation Series 241 dichotomous samplers (Thermo Electron Corp., Waltham, MA) on Teflon® filters, and mass concentrations were determined by gravimetric analysis. The sampling methodology was based on the U.S. EPA's Compendium Method IO-2.2 (USEPA, 1999). Detailed laboratory operating procedures are available elsewhere (Orquiz, 2001). Only data from PM_{2.5} samples was employed in this study. The dichotomous samplers also collect PM₁₀. However the evaluation of a PM₁₀ model was not feasible at this time. The study region is located in the heart of the Chihuahuan desert where PM₁₀ sources are ubiquitous and the influence of weather on ambient PM₁₀ levels is stronger. Here the spatial variability of PM₁₀ is less localized than in typical urban regions. Unfortunately the current network was unfit to capture the PM₁₀ spatial variability in the study region. A network that captures the upwind PM₁₀ levels to the west, the background levels to the northeast, and the impact from the numerous unpaved roads in north and southeast Juarez would be ideal for such application.

2.4 Modeling approach

Recent discussions on linear regression analysis strongly suggest a minimum sample size to predictive variable ratio (n/k) between 10–15 to avoid over fitting a model (Babiyak, 2004; Green, 1991; Mulaik, 2001). Based on these guidelines, the number of monitoring sites ($n = 13$) available for this study constrained the LUR model to a maximum of one predictor variable. In a recent review of land use regression publications presented by (Hoek et al., 2008), the typical LUR model included at least three predictive variables. Hoek et al., (2008) also discussed the typical number of sites used in studies at that time. It was suggested that 40–80 sites may be reasonable but that population and city size should be taken into account, and mentions specific studies that have explored this topic. For models with three predictive variables, the recommendation of at least 40 monitoring sites coincides with

suggestions of 10 to 15 samples per predictive variable for linear regression models (Babyak, 2004; Green, 1991; Mulaik, 2001). Although, the relationship between sample size and number of predictive variables in LUR models has not been formally discussed in the literature, the guidelines given for linear regression analysis in general should hold for all applications of the methods including LUR modeling. Consequently, the goal of this study was to optimize the predictive capacity of a model with a single predictor variable. The evaluation of a model with two predictor variables was avoided as such model might over fit the data.

Vehicle emissions was chosen a priori as the probable best single predictor of $PM_{2.5}$. A surrogate of vehicle emissions has been consistently identified as a strong predictor of ambient $PM_{2.5}$ in other LUR studies (Hoek et al., 2008; Moore et al., 2007; Morgenstern et al., 2007; Ross et al., 2007; Ryan and Lemasters, 2007). Still the types of surrogate variables used in LUR studies varied widely and an ideal vehicle-emissions measure or surrogate remains to be identified (Hoek et al., 2008; Moore et al., 2007; Ross et al., 2007; Ryan and Lemasters, 2007). In this study, a set of linear combinations of traffic-related variables was evaluated against single traffic-related variables to explore the possibility of an optimized vehicle-emissions surrogate for the El Paso, Texas urban region. The traffic dataset consisted of several parameters produced for transportation modeling purposes (El Paso MPO, 2011). The traffic-related parameters were geographically processed and summarized for buffers of varying sizes centered on 2240 grid points distributed orthogonally at a 500-m isotropic resolution over the El Paso urban area. The dimensionality of the gridded traffic data was reduced by means of principal component analysis. The PCA produced a set of new variables (linear combinations of the traffic variables), which were independently evaluated for their modeling performance of $PM_{2.5}$ by means of linear regression.

2.5 Traffic-related parameters

The following parameters were extracted from a road network used to model traffic in 2007: vehicle miles traveled, number of lanes, road capacity (vehicles per hour per lane; *vphl*), traffic flow (*vphl*), average speed (miles per hr), and road segment length (mile) (El Paso MPO, 2011). Vehicle miles traveled (VMT) was the total number of miles driven on a road segment by all vehicles over one year. Road capacity was the maximum volume of vehicles per hour that can pass through a lane of the road segment under design conditions. Traffic flow is the actual volume of vehicles per hour passing through a lane of the road segment. The *road length* variable was the sum of all road segments in a buffer. The *speed* variable represented the weighted average over all road segments with road length as weight. For road segments partially within a buffer the proportional VMT magnitude was used. *Number of lanes* was employed to compute new variables as explained below but was not summarized by buffer. *Road capacity* was divided by *total flow* and multiplied by the *number of lanes* to produce a variable that represented *traffic demand* per road segment. The *traffic demand* for all roads segments within a buffer was computed as the sum of *traffic demand* multiplied by *road length* and divided by total *road length*. The *road capacity* and *traffic flow* variables were not employed in the PCA only *traffic demand*. A crude estimate of average time spent by a vehicle within the buffer was computed by dividing the diameter of the buffer by the average speed for all segments within the buffer. This variable was labeled as *time*.

2.6 Principal Component Analysis

Principal component analysis was employed to reduce a dataset of 5 traffic-related variables, summarized for 6 different buffer sizes, onto a smaller set of variables composed of linear combinations of the traffic-related variables. Principal components (PCs) with eigenvalues greater than 1 were selected. The *varimax* rotation was employed to ensure that each

variable was maximally correlated with one PC and had minimal associations with the other components (Kaiser, 1958). The factor loadings and the linear combinations of the traffic-variables were checked to ensure the PCs were physically congruent as surrogates of vehicle emissions and that no double counting occurred. Factor score vectors were determined for each component at each of the 13 monitoring sites. Factor scores provide the variability of the principal components (linear combinations of traffic-variables); therefore, they were considered viable surrogates of vehicle-emission and evaluated as predictor variables of $PM_{2.5}$.

2.7 Linear Regression Analysis

The factor scores and the single traffic-related parameters calculated for each monitoring site were used as predictor variables of $PM_{2.5}$ in linear regression models and compared based on modeling performance (R^2). Each predictor variable was independently introduced into a simple linear regression model with $PM_{2.5}$ as the dependent variable. Normal probability plots of the standardized residuals were used to evaluate the normality of the residuals. Scatter plots of the standardized residuals against the predictor variable and the fitted values were used for linearity and homogeneity evaluations.

2.8 Model validation

The leave-one-out-cross-validation (LOOCV) technique was used to validate the performance of the regression models. The LOOCV has been widely used in LUR studies and has been shown to be adequate and robust for such purposes (Isakov et al., 2012; Johnson et al., 2010; Parenteau and Sawada, 2010; Rose et al., 2010). The LOOCV approach consisted of developing the model, originally developed with n observations, with $n - 1$ observations and using the $PM_{2.5}$ value at the location omitted from the modeling process to evaluate the performance. The process was repeated n times such that each site was left out once. The measure of performance in the LOOCV procedure was the R^2 parameter estimated for the fit between the observed and predicted values of ambient $PM_{2.5}$ averaged concentrations.

All geographical data was processed within a Geographical Information Systems (GIS) platform (ArcInfo 9.2, ESRI). Principal component analysis and linear regression models were performed with commercially available statistical software (SPSS 19, IBM).

3. Results

3.1 Data quality

A total of 96 duplicate $PM_{2.5}$ samples were collected during the monitoring campaign. The CV calculated from the duplicate sample was of 7.4%. Co-located $PM_{2.5}$ measured by the integrated monitors agreed reasonably well with those measured by continuous monitors at CAMS sites (CV 16.4 – 17.7%). The regional all-site averaged $PM_{2.5}/PM_{10}$ ratio was 0.27 with a standard deviation of 0.12. The observed $PM_{2.5}/PM_{10}$ ratio is consistent with values previously observed in the region (Li et al., 2001). To evaluate the degree of bias introduced by non-concurrent sampling, the long-term PM averages were adjusted as described in section 2.2. The Pearson correlation coefficient between the unadjusted and adjusted averages was 0.85. The adjusted values were used for modeling purposes.

3.2 Ambient PM Concentrations

The 941 $PM_{2.5}$ samples used in the analysis had an arithmetic mean of $7.2 \mu g m^{-3}$, with values ranging from 5.4 to $9.0 \mu g m^{-3}$ (SD = 4.1). A slight trend with highest values to the south was observed. The observed trend in El Paso is consistent with previous observations

(Gonzales et al., 2005; Noble et al., 2003) of vehicle-related air pollution increasing southward towards the U.S.-Mexico border.

3.3 Vehicle-emissions surrogates

A total of 67,200 traffic-related data points (scores) were produced from summarizing 5 traffic variables for 6 buffer sizes for 2240 grid points distributed over the study area. The traffic dataset was of adequate size for PCA purposes (Alii, 2010). Five components displayed eigenvalues greater than 1 and were retained for rotation. Combined, the selected components accounted for 66% of the total variance (see Table 1). The analysis produced a Keiser-Meyer-Olkin measure of sampling adequacy of .940 and the Bartlett's test of sphericity was significant ($p < .000$). Traffic variables and the corresponding factor loadings for the rotated components are presented in Table 2. In interpreting the rotated factor pattern, an item was said to load on a given component if the factor loading was .50 or greater for that component. Using these criteria, traffic variables clearly loaded on components 2 to 5 as function of buffer size (i.e., 300 m – 1500 m). Traffic variables from buffer sizes equal or greater than 1500 m loaded on component 1.

3.4 Land Use Regression model

The R^2 values for the linear regression models are summarized in Table 3. The traffic variable that best predicted ambient $PM_{2.5}$ levels was *Traffic Demand* for the 1 km buffer, intercept (C) = 6.007, slope (b) = 1.546, $t(11) = 10.611$, $p < .000$. *Traffic Demand* (1 km) also explained a significant proportion of variance in ambient $PM_{2.5}$ levels, $R^2 = .341$, $F(1,11) = 5.705$, $p < .05$. Component 4, composed of a linear combination of traffic variables at a 1 km buffer, also significantly predicted ambient $PM_{2.5}$ levels, $C = 7.133$, $b = .903$, $t(11) = 24.921$, $p < .000$, and explained a significant proportion of variance in ambient $PM_{2.5}$ levels, $R^2 = .377$, $F(1,11) = 6.663$, $p < .05$.

To explore the possibility of further optimizing the linear weights of component 4, the traffic variables for the 1 km buffer were subjected to PCA independently, while excluding all other variables. The PCA applied only to the variables summarized for the 1 km-buffer produced two components with eigenvalues greater than 1, which explained 72% of the variability in that dataset. The *post hoc* PCA analysis produced a Keiser-Meyer-Olkin measure of sampling adequacy of .747 and the Bartlett's test of sphericity was significant ($p < .000$). The two new components are referred as C4.1 and C4.2 to avoid confusion with the previous set of components. The *Traffic Demand*, *Speed*, *Time*, and VMT variables loaded on C4.1, whereas *Road Length* loaded in C4.2, as shown in Table 4. Factor scores for both components were computed for the site locations as before and regressed against ambient $PM_{2.5}$. The R^2 values for the linear regression models with components C4.1 and C4.2 as predictor variables were .458 and .006, respectively (see Table 3). Component C4.1 was selected as the optimized surrogate of vehicle emissions as it significantly predicted ambient $PM_{2.5}$ levels, $C = 6.524$, $b = 1.107$, $t(11) = 3.005$, $p < .05$, and explained a significant proportion of variance in ambient $PM_{2.5}$ levels, $R^2 = .458$, $F(1,11) = 9.029$, $p < .05$.

3.5 Leave-one-out model validation

Using the LOOCV technique described in section 2.8 the R^2 statistic was calculated for the models with *Traffic Demand* (1 km), component 4, and C4.1 as predictor variables. The C4.1 model produced the highest R^2 (.469), followed by the C4 model (.301) and the *Traffic Demand* (1 km) model (.298). The associated scatter plots are shown in Figure 3. The performance of the models, given the small sample size, was considered acceptable.

4. Discussion

Among all traffic variables, *Traffic Demand* (1 km) was the best predictor of ambient PM_{2.5} with an R^2 equal to .341. Principal component analysis was successfully applied to create a new set of variables, out of which a linear combination of the *Road Length*, *Traffic Demand*, *Speed*, *Time*, and VMT variables at a 1 km buffer (component 4), had a modeling performance ($R^2 = .377$) greater than that of *Traffic Demand* (1 km) independently. Furthermore, an optimized version of component 4 was obtained by subjecting only the 1 km-buffer variables to the PCA. The *Road Length* (1 km) variable did not load on the optimized component (C4.1), which had the highest R^2 value of .458 of all analyzed variables. Note that both components 4 and C4.1 include the *Traffic Demand* (1 km) variable. Therefore, the PCA increased the performance of *Traffic Demand* (1 km) by combining it with the other traffic variables summarized at the same buffer size. Also, the PCA results suggested that for this study area, the current arrangement (e.g., 500 m grid resolution), and this specific set of variables, a buffer of approximately 1000 m was best suitable for capturing the PM_{2.5} spatial variability.

The performance of all the significant models produced in this study was modest, with the highest significant R^2 equal to .458. Considering the small number of sites, the performance of the models was considered adequate. However, the objective of this study was to employ PCA to enhance the performance of LUR models, specifically those with small number of monitoring sites. In this regard, the improvements in model performance achieved by components 4 and C4.1 over the *Traffic Demand* (1 km) variable were significant and notable.

For the smallest buffer sizes (e.g., 300 – 1500 m) the traffic variables loaded on different components grouped by buffer size. For larger buffer sizes (e.g., 1500 – 5000 m) the variables loaded on the same component (C1). For a grid resolution of 500 meters buffers with diameters greater than 500 m overlap. As the overlap increases, the summarized variables capture a greater fraction of shared variation. Since the PCA is based on the correlation matrix of the variable dataset, variables with similar variation tend to load on the same component. Variable double counting occurs if the same variable, summarized for different buffer sizes, loads on the same component. An effective way to avoid double counting, without changing the resolution of the grid, is to perform the PCA for each buffer size as it was performed here *post hoc* for the 1km buffer size. However, the *post hoc* PCA approach employed here to determine components C4.1 and C4.2 uses additional degrees of freedom and thus introduces additional uncertainty. Still, the cost in degrees of freedom of using the PCA on a 2800-point dataset obtained from the grid is minimal compared to the cost of including an additional predictor variable to linear regression applied to 13 observations. Furthermore, degrees of freedom might be fully economized if the PCA approach is applied to the surrogate variables grouped by buffer size directly.

The method employed in this study was successful at optimizing the use of traffic-related variables by producing a vehicle-emission surrogate with better modeling performance, from a linear combination of the traffic-variables. The capacity to utilize PM_{2.5} datasets from small- n monitoring networks and produced acceptable LUR models, strongly complements the economical attractiveness of the LUR techniques. The use of data from a limited number of sites, either from existing or purposed deployed monitoring networks, is particularly important during study design or pilot studies. The PCA-LUR approach applied to data from small- n networks is ideal for the generation of preliminary pollutant surfaces and applications such as the location-allocation method, which optimizes the site selection process during monitoring network design.

The capacity of PCA to improve the performance of an LUR model with single predictor variable has applications beyond producing preliminary surfaces and network design from data collected from small- n monitoring networks. The PCA method presented here could also be applied to enhance LUR models with multiple predictor variables produced from extensive monitoring data. For instance, through additional monitoring extra variables could be introduced to the PCA model and further improve its explanatory capacity. Note that PCA in this study was employed exclusively on vehicle emissions surrogates, thus the introduction of surrogates for each of the other processes commonly included in LUR models (e.g., land-use, traffic, demographic characteristics) could further increase the certainty of models while economizing degrees of freedom. Furthermore, it could be argued that PCA could be employed to optimize all common LUR surrogates and subsequently construct an enhanced multiple regression model only from optimized components representing each of those surrogates. Follow-up studies should focus on evaluating such possibility.

A shortcoming of the use of PCA is that principal components might have complex physical meanings, and the physical interpretation of the resulting models might be challenged. In this regard, the biggest shortcoming has to do with the transferability of the model. Still the interpretation of components and the resulting models can be enhanced if only principal components with logical physical interpretation are extracted and by employing PCA separately on variables with similar physical meanings and for buffers of the same size, as it was done here. In this study PCA was employed exclusively on traffic variables, and the selected component represented variables extracted only from the 1 km buffer. In general, the transferability of the PCA approach is universal and should be applicable to any region where ordinary LUR could be applied. Furthermore, the specific model generated in this study might be transferable to similar arid regions (i.e. dryness, broad daily temperature changes, desert surroundings).

The PM_{2.5} surface generated in this study will be employed as input to a location-allocation protocol to design a monitoring network of 36 monitors within the study area for purposes of produce a two-variable land use regression model. Finally, although the PCA was effective at increasing the fit (R^2) of an LUR model by optimizing the use of a PM_{2.5} dataset, it remains to be seen whether the same applies to other pollutants.

5. Concluding remarks

Principal component analysis was successfully employed to optimize the use of long-term ambient PM_{2.5} averaged concentrations collected at 13 monitoring sites throughout El Paso, Texas, and to develop a Land Use Regression model with a single predictor variable, a vehicle-emissions surrogate. The vehicle-emissions surrogate was composed of a linear combination of the variables *Traffic Demand*, *Speed*, *Time*, and VMT within a buffer of 1 km. The PCA approach helped produce a model with a performance ($R^2 = .458$, $p < 0.05$) greater than that of the single variable, *Traffic Demand* at 1 km, with best performance ($R^2 = .341$, $p < 0.05$).

The PCA-based method presented in this study can help optimize the utilization of pollutant datasets from both small and large monitoring networks for LUR modeling purposes. When applied to small monitoring networks the PCA approach is ideal for producing preliminary pollutant surfaces necessary for network design applications, such as that of the Location-Allocation approach. For larger datasets, PCA can be used to produce multiple surrogates with better modeling capacity. In general, the PCA method can contribute to LUR techniques in two major ways: 1) by economizing degrees of freedom through the optimized use of independent variables, and 2) by creating an orthogonal set of predictor variables, and

thus fulfilling the no-colinearity requisite of the linear regression methods. The proposed PCA method, for combining predictor-variables, will advance the use LUR techniques by expanding their economical attractiveness and facilitating the application of the location-allocation monitoring network design approach.

Acknowledgments

This publication was made possible by grant number S11 ES013339 from the National Institute of Environmental Health Sciences (NIEHS), NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIEHS or NIH. The authors want to thank the El Paso Water Utilities, specifically Fernando Rico, and the Texas Commission on Environmental Quality for their invaluable support of this study.

The authors would also like to acknowledge the valuable feedback provided by the anonymous reviewers. Their comments enriched the analysis and considerably advanced the quality of this manuscript.

References

- Adar SD, Kaufman JD. Cardiovascular disease and air pollutants: evaluating and improving epidemiological data implicating traffic exposure. *Inhalation Toxicology*. 2007; 19 (Suppl 1):135–149. [PubMed: 17886061]
- Alii LK. Sample Size and Subject to Item Ratio in Principal Components Analysis and Exploratory Factor Analysis. *Journal of Biometrics & Biostatistics*. 2010:01.
- Babak M. What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*. 2004; 66:411–421. [PubMed: 15184705]
- Beelen R, Hoek G, Fischer P, Brandt PAVd, Brunekreef B. Estimated long-term outdoor air pollution concentrations in a cohort study. *Atmospheric Environment*. 2007; 41:1343–1358.
- Brauer M, Hoek G, van Vliet P, Meliefste K, Fischer P, Gehring U, et al. Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems. *Epidemiology (Cambridge, Mass)*. 2003; 14:228–239.
- Briggs D. The role of GIS: coping with space (and time) in air pollution exposure assessment. *Journal of Toxicology and Environmental Health, Part A*. 2005; 68:1243–1261. [PubMed: 16024500]
- Briggs D, Collins S, Elliott P, Fischer P, Kingham S, Lebre E, et al. Mapping urban air pollution using GIS: A regression-based approach. *International Journal of Geographical Information Science*. 1997; 11:699–718.
- Briggs DJ, de Hoogh C, Gulliver J, Wills J, Elliott P, Kingham S, et al. A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. *The Science of the Total Environment*. 2000; 253:151–167. [PubMed: 10843339]
- Brook RD, Franklin B, Cascio W, Hong Y, Howard G, Lipsett M, et al. Air pollution and cardiovascular disease: a statement for healthcare professionals from the Expert Panel on Population and Prevention Science of the American Heart Association. *Circulation*. 2004; 109:2655–2671. [PubMed: 15173049]
- Carr D, von Ehrenstein O, Weiland S, Wagner C, Wellie O, Nicolai T, et al. Modeling annual benzene, toluene, NO₂, and soot concentrations on the basis of road traffic characteristics. *Environmental Research*. 2002; 90:111–118. [PubMed: 12483801]
- Currey, R.; Pumfrey, R. Improving Air Quality in Paso del Norte. In: Pumfrey, R., editor. *The US-Mexican border environment: Binational air quality management*. Vol. 14. 2006. p. 109-122. SCERP Monograph Series
- de Kok TMCM, Driece HAL, Hogervorst JGF, Briedé JJ. Toxicological assessment of ambient and traffic-related particulate matter: a review of recent studies. *Mutation Research*. 2006; 613:103–122. [PubMed: 16949858]
- Dominici F, McDermott A, Daniels M, Zeger SL, Samet JM. Revised analyses of the National Morbidity, Mortality, and Air Pollution Study: mortality among residents of 90 cities. *Journal of Toxicology and Environmental Health, Part A*. 2005; 68:1071–1092. [PubMed: 16024489]

- El Paso, MPO. Office EPMP. Transportation Improvement Program. El Paso; Texas: 2011. Transportation Conformity Report for the Mission 2035 Metropolitan Transportation Plan-Mission 2011–2014.
- Englert N. Fine particles and human health--a review of epidemiological studies. *Toxicology letters*. 2004; 149:235–242. [PubMed: 15093269]
- Gonzales M, Qualls C, Hudgens E, Neas L. Characterization of a spatial gradient of nitrogen dioxide across a United States-Mexico border city during winter. *The Science of the Total Environment*. 2005; 337:163–173. [PubMed: 15626387]
- Green S. How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*. 1991; 26:499–510.
- Henderson SB, Beckerman B, Jerrett M, Brauer M. Application of Land Use Regression to Estimate Long-Term Concentrations of Traffic-Related Nitrogen Oxides and Fine Particulate Matter. *Environmental Science & Technology*. 2007; 41:2422–2428. [PubMed: 17438795]
- Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fischer P, et al. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*. 2008; 42:7561–7578.
- Hoek G, Meliefste K, Cyrus J, Lewne M, Bellander T, Brauer M, et al. Spatial variability of fine particle concentrations in three European areas. *Atmospheric Environment*. 2002; 36:4077–4088.
- INEGI. Instituto Nacional de Estadística y Geografía. Mexico: 2011.
- Isakov, V.; Johnson, M.; Touma, J.; Özkaynak, Hk. Development and Evaluation of Land-Use Regression Models Using Modeled Air Quality Concentrations, Air Pollution Modeling, and its Application XXI. Steyn, DGG.; Trini Castelli, ST., editors. Vol. 4. Springer; Netherlands: 2012. p. 717-722.
- Jerrett M, Arain MA, Kanaroglou P, Beckerman B, Crouse D, Gilbert NL, et al. Modeling the intraurban variability of ambient traffic pollution in Toronto, Canada. *Journal of Toxicology and Environmental Health, Part A*. 2007; 70:200–212. [PubMed: 17365582]
- Johnson M, Isakov V, Touma JS, Mukerjee S, Özkaynak H. Evaluation of land-use regression models used to predict air quality concentrations in an urban area. *Atmospheric Environment*. 2010; 44:3660–3668.
- Kaiser H. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*. 1958; 23:187–200.
- Kanaroglou P, Jerrett M, Morrison J, Beckerman B, Arain M, Gilbert N, et al. Establishing an air pollution monitoring network for intra-urban population exposure assessment: A location-allocation approach. *Atmospheric Environment*. 2005; 39:2399–2409.
- Kappos AD, Bruckmann P, Eikmann T, Englert N, Heinrich U, Höppe P, et al. Health effects of particles in ambient air. *International Journal of Hygiene and Environmental Health*. 2004; 207:399–407. [PubMed: 15471105]
- Krewski D, Burnett R, Jerrett M, Pope CA, Rainham D, Calle E, et al. Mortality and long-term exposure to ambient air pollution: ongoing analyses based on the American Cancer Society cohort. *Journal of Toxicology and Environmental Health, Part A*. 2005; 68:1093–1109. [PubMed: 16024490]
- Lewtas J. Air pollution combustion emissions: characterization of causative agents and mechanisms associated with cancer, reproductive, and cardiovascular effects. *Mutation Research*. 2007; 636:95–133. [PubMed: 17951105]
- Li WW, Orquiz R, Garcia JH, Espino TT, Pingitore NE, Gardea-Torresdey J, et al. Analysis of temporal and spatial dichotomous PM air samples in the El Paso-Cd. Juarez air quality basin. *Journal of the Air and Waste Management Association*. 2001; 51:1551–1560.
- Miller K, Siscovick D, Sheppard L, Shepherd K, Sullivan J, Anderson G, et al. Long-Term Exposure to Air Pollution and Incidence of Cardiovascular Events in Women. *New England Journal of Medicine*. 2007; 356:447. [PubMed: 17267905]
- Moore DK, Jerrett M, Mack WJ, Künzli N. A land use regression model for predicting ambient fine particulate matter across Los Angeles, CA. *J Environ Monit*. 2007; 9:246–252. [PubMed: 17344950]

- Morgenstern V, Zutavern A, Cyrys J, Brockow I, Gehring U, Koletzko S, et al. Respiratory health and individual estimated exposure to traffic-related air pollutants in a cohort of young children. *Occupational and Environmental Medicine*. 2007; 64:8–16. [PubMed: 16912084]
- Mulaik SA. The Curve-Fitting Problem: An Objectivist View. *Philosophy of Science*. 2001; 68:218–241.
- Noble CA, Mukerjee S, Gonzales M, Rodes CE, Lawless PA, Natarajan S, et al. Continuous measurement of fine and ultrafine particulate matter, criteria pollutants and meteorological conditions in urban El Paso, Texas. *Atmospheric Environment*. 2003; 37:827–840.
- O’neill MS, Veves A, Sarnat JA, Zanobetti A, Gold DR, Economides PA, et al. Air pollution and inflammation in type 2 diabetes: a mechanism for susceptibility. *Occupational and Environmental Medicine*. 2006; 64:373–379. [PubMed: 17182639]
- Orquiz, R. Juarez air basin Civil Engineering Masters of Science. University of Texas; El Paso, El Paso, Texas: 2001. Gravimetric and elemental analysis of particulate matter in the El Paso-Cd; p. 143
- Parenteau M-P, Sawada M. The role of spatial representation in the development of a LUR model for Ottawa, Canada. *Air Quality, Atmosphere & Health*. 2010:1–13.
- Pope CA, Dockery DW. Health effects of fine particulate air pollution: lines that connect. *Journal of the Air & Waste Management Association*. 2006; 56:709–742. [PubMed: 16805397]
- Rose N, Cowie C, Gillett R, Marks GB. Validation of a Spatiotemporal Land Use Regression Model Incorporating Fixed Site Monitors. *Environmental Science & Technology*. 2010; 45:294–299. [PubMed: 21133418]
- Ross Z, Jerrett M, Ito K, Tempalski B, Thurston G. A land use regression for predicting fine particulate matter concentrations in the New York City region. *Atmospheric Environment*. 2007; 41:2255–2269.
- Ryan PH, Lemasters GK. A Review of Land-use Regression Models for Characterizing Intraurban Air Pollution Exposure. *Inhalation Toxicology*. 2007; 19:127–133. [PubMed: 17886060]
- Ryan PH, Lemasters GK, Biswas P, Levin L, Hu S, Lindsey M, et al. A comparison of proximity and land use regression traffic exposure models and wheezing in infants. *Environmental Health Perspectives*. 2007; 115:278–284. [PubMed: 17384778]
- Smith L, Mukerjee S, Gonzales M, Stallings C, Neas L, Norris G, et al. Use of GIS and ancillary variables to predict volatile organic compound and nitrogen dioxide levels at unmonitored locations. *Atmospheric Environment*. 2006; 40:3773–3787.
- Stedman J, Vincent K, Campbell G, Goodwin J, Downing C. New high resolution maps of estimated background ambient NO_x and NO₂ concentrations in the UK. *Atmospheric Environment*. 1997; 31:3591–3602.
- USCB. El Paso, Texas - Fact Finder. Vol. 2011. U.S. Census Bureau; 2011.
- USEPA. Compendium of Methods for the Determination of Inorganic Compounds in Ambient Air. 1999.
- Wilson W, Chow J, Claiborn C, Fusheng W, Engelbrecht J, Watson J. Monitoring of particulate matter outdoors. *Chemosphere*. 2002; 49:1009–1043. [PubMed: 12492163]

Highlights

1. PCA was employed to produce a vehicle PM_{2.5} emissions surrogate for LUR modeling.
2. The surrogate had a predictive capacity 34% greater than any traffic-variable.
3. The PCA approach optimized data use from a small network (N = 13).
4. The PCA approach advances the use of LUR and location-allocation methods.

Figure 1.
PCA optimization of LUR for small-n monitoring network.

Figure 2.
Map of El Paso, Texas showing monitoring site locations and road network.

Figure 3.
Scatter plot of LOOCV results.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Loadings of Traffic Variables on Extracted and Rotated Components

	Communality		Component				
	Extraction	1	2	3	4	5	
Road Length (300 m)	.833			.899			
Traffic Demand (300 m)	.391			.591			
Speed (300 m)	.377			.594			
Time (300 m)	.700			.501			
VMT (300 m)	.344			.564			
Road Length (500 m)	.859		.908				
Traffic Demand (500 m)	.668		.804				
Speed (500 m)	.546		.506				
Time (500 m)	.376		.580				
VMT (500 m)	.644		.789				
Road Length (1 km)	.640				.773		
Traffic Demand 1 km)	.690				.821		
Speed (1 km)	.671				.766		
Time (1 km)	.399				.614		
VMT (1 km)	.532				.696		
Road Length (1.5 km)	.817	.590				.676	
Traffic Demand (1.5 km)	.811	.621				.645	
Speed (1.5 km)	.806	.615				.643	
Time (1.5 km)	.615	.458				.633	
VMT (1.5 km)	.737	.581				.618	
Road Length (2.5 km)	.824	.846					
Traffic Demand (2.5 km)	.842	.865					
Speed (2.5 km)	.818	.847					
Time (2.5 km)	.552	.725					
VMT (2.5 km)	.750	.815					
Road Length (5 km)	.876	.932					
Traffic Demand (5 km)	.890	.941					
Speed (5 km)	.881	.932					

	Communality		Component				
	Extraction		1	2	3	4	5
Time (5 km)	.724		.849				
VMT (5 km)	.843		.910				

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
^aRotation converged in 6 iterations.

Table 3

R² values for single linear model with PM_{2.5} as the dependent variable.

Predictor	-	Buffer Size				
		300 m	500 m	1 km	1.5 km	2.5 km
Road Length		0.277	0.211	0.294	0.288	0.312
Traffic Demand		0.101	0.150	0.341	0.340	0.320
Speed		0.151	0.210	0.283	0.255	0.180
Time		0.262	0.210	0.180	0.120	0.065
VMT		0.287	0.283	0.304	0.290	0.289
Component 1	0.181					
Component 2	0.181					
Component 3	0.227					
Component 4	0.377					
Component 5	0.049					
C4.1	0.458					
C4.2	0.006					

Bold font represents models significant at the 0.05 level (2-tailed).

Table 4

Rotated Components for PCA of traffic variables (1 km)

	Communalities	Components	
		1	2
Road Length (1 km)	0.790		0.877
Traffic Demand (1 km)	0.763	0.866	
Speed (1 km)	0.696	0.733	
Time (1 km)	0.598	0.706	
VMT (1 km)	0.742	0.777	

Extraction method: principal component analysis

Rotation method: varimax with Kaiser normalization.

^aRotation converged in 3 iterations