MULTI-PERSPECTIVE ANALYSIS AND SPATIOTEMPORAL MAPPING OF AIR POLLUTION MONITORING DATA

Alexander Kolovos,[1*] André Skupin,[2] Michael Jerrett,[3] and George Christakos[2]

1. SAS Institute, Inc., 100 SAS Campus Dr. S3042, Cary, NC 27513, USA.
2. Department of Geography, SDSU, San Diego, CA 92182-4493, USA.
3. School of Public Health, UC Berkeley, CA 94720-7360, USA.

Abstract

Space-time data analysis and assimilation techniques in atmospheric sciences typically consider input from monitoring measurements. The input is often processed in a manner that acknowledges characteristics of the measurements (e.g., underlying patterns, fluctuation features) under conditions of uncertainty, leads to the derivation of secondary information that serves study-oriented goals, and provides input to space-time prediction techniques. We present a novel approach that blends a rigorous space-time prediction model (Bayesian Maximum Entropy, BME) with a cognitively informed visualization of high-dimensional data (spatialization). The combined BME and spatialization approach (BME-S) is used to study monthly-averaged NO_2 and mean annual SO_4 measurements in California over the 15-year period 1988-2002. Using the original scattered measurements of these two pollutants BME generates spatiotemporal predictions on a regular grid across the state. Subsequently, the prediction network undergoes the spatialization transformation into a lower-dimensional geometric representation, aimed at revealing patterns and relationships that exist within the input data. The proposed BME-S provides a powerful spatiotemporal framework to study a variety of air pollution data sources.

Key words: *BME, spatiotemporal, prediction, mapping, spatialization, dimensional reduction*.

* To whom correspondence should be addressed: kolovos@email.unc.edu (phone: 919-531-2165)

1. Introduction

Nitrogen dioxide (NO_2) and sulfate (SO_4) are two air pollutants that typically occur as by-products of fossil fuel and biomass combustion. SO_4 mostly has the effect of increasing the air acidity that can lead to acid rain, whereas the role of NO_2 is more complicated as it contributes both to tropospheric ozone pollution and may have adverse effects on the cardio-respiratory system (*1*, *2*). A variety of statistical techniques can be used to study separately the spatial variation and temporal evolution of air pollutants and the associated risk assessment (e.g., *3*-*7*). The present work demonstrates a novel approach to study important characteristics of the SO_4 and NO_2 distributions in a *composite* space-time continuum. Such characteristics can be as evident and measurable as concentration values, or can be of a more subtle nature such as behavioral similarities across time for individual spatial locations.

More specifically, the proposed approach is the synthesis of a rigorous space-time stochastic analysis (BME, Bayesian Maximum Entropy, (*8*)) with a set of computational transformations for visualizing high-dimensional data (spatialization, (*9*)). BME is a well-established spatiotemporal statistics and geostatistics methodology (*8*) for spatiotemporal prediction of environmental attributes. With its advanced features (e.g., lack of restrictive assumptions, assimilating input from monitor measurements, a variety of certain and uncertain observations, and physical laws), BME has been employed in many atmospheric studies and has provided accurate prediction of air pollutants across space and time (e.g., *10*-*15*). Meanwhile, spatialization is an approach to making complex high-dimensional data accessible to the human perceptual and cognitive system through computational and visual means (*9*). This is grounded in the use of spatial metaphors for non-spatial data and typically involves processes of dimensionality reduction and spatial layout, followed by symbolization. We are employing the self-organizing map (SOM) method for clustering and dimensionality reduction (*16*, *17*). The synthetic BME-spatialization approach (BME-S) generates an array of visual output that offers insight, perspective, and facilitates the understanding of underlying mechanisms that govern the spatiotemporal distribution of the NO_2 and SO_4 pollutants.

In general, the monitors that generate pollutant measurements are characterized by sparseness and heterogeneity across space-time under conditions of uncertainty (*18*).

BME analysis provides a stochastic description of the combined uncertainty-heterogeneity that characterizes these pollutants and generates pollutant distributions in space-time (predicted distribution mean, median, etc.) and the associated prediction uncertainty (*10-14*). The BME analysis typically produces results on a composite space-time grid (involving multiple locations in space and numerous temporal instances). As such, for mapping purposes, the BME output could be considered as a higher-dimensional manifold, in the sense of (*16*).

Spatialization, with its series of conceptual, computational, and visual transformations, can lead to such high-dimensional data being seen in new ways (*9*). In this study, the gridded BME output enables the spatialization technology to search for informative space-time patterns of NO_2 and SO_4 concentrations. The particular spatialization approach introduced here involves reinterpreting the same data set in multiple ways, based on transformation of BME-generated values into different space-time-attribute configurations. That enables the combined BME-S approach to offer a valuable multi-perspective framework in the study of air pollutants in a composite space-time context.

To facilitate illustration of the BME-S approach, we selected on purpose the two pollutants NO_2 and SO_4 because their spatiotemporal distributions typically feature some distinctively different characteristics. Specifically, NO_2, a secondary pollution formed shortly after emission, is considered to be linked to traffic exhaust and is known to vary over small areas. The choice of SO_4 was motivated by the facts that sulfate is largely a regional secondary pollutant and it is the most spatially homogeneous component of ambient particles (*19*). To accent more on these differences, the geostatistical study of these two pollutants focuses on different time scales; namely, NO_2 is studied on a monthly temporal scale, whereas SO_4 concentration is predicted on the basis of annual-averaged values. In essence, the spatialization processing is then tested to investigate for space, time and attribute patterns between these two pollutants under these specific conditions. In the following, we exhibit that the proposed BME-S approach can be used to reveal connections between attributes that might be otherwise not apparent, as in the present study where we employ relatively unrelated attributes. In that sense, the BME-S approach is illustrated here in terms of the NO_2 and SO_4 datasets as a proof-of-concept.

The usefulness of BME-S extends beyond the scope of the current study; BME-S can be applied in a variety of space-time atmospheric systems involving large geographic areas, long time periods, and even multiple attributes captured by diverse monitors.

2. Nitrogen Dioxide and Sulfate Data

The study datasets were provided by the California Air Resources Board (CARB). The datasets consist of monthly-averaged NO_2 measurements from a total of 137 monitoring locations, and annual averages of SO_4 measurements from 166 monitoring locations over the 15-year period between 1988 and 2002. SO_4 was measured as a mass constituent of PM_{10} (particulate matter of 10 microns or less in aerodynamic diameter, although previous studies indicate that much of the SO_4 will have particle sizes below 2.5 microns in aerodynamic diameter; see, e.g., (*20*)). Figure 1 shows all the measurement sites for each pollutant during the study period. The geographic map units are meters and the data geographic coordinate system (GCS) is the 1983 North American GCS. About 70-100 measurements were available at every time instance for purposes of the BME analysis.

The prediction stage considers a moderately dense spatial rectangular grid of 21×24 nodes that encloses the state of California. The subsequent stage of spatialization analysis considers all the grid nodes within the state borders. The temporal grid resolution in NO_2 prediction is one month; it starts in January 1988 (month $t_M = 1$) and finishes in December 2002 (month $t_M = 180$). The corresponding temporal resolution for the SO_4 analysis is one year, where the grid starts in 1988 (year $t_Y = 1$) and finishes in 2002 (year $t_Y = 15$).

The analysis leads to maps of the pollutant concentration predictions in California that reveal the state-wide behavior of the pollutants. Some more information about the study data sets is given in section S1, Supporting Information.

3. Methodologies

3.1. *Conceptual and theoretical foundation of BME Spatiotemporal Prediction*

The study of the pollutants concentrations in space and time involves natural uncertainty, and for this reason we describe each pollutant in a stochastic manner. The spatiotemporal

random field theory (S/TRF, *21*) is a powerful tool to study attributes that vary across space-time under conditions of uncertainty.

In the synthetic BME-S approach, we consider the concentrations of each one of the pollutants as a S/TRF. The distribution of pollutant concentrations across space-time is represented by the S/TRF $X_{\boldsymbol{p}}$ at each space-time point $\boldsymbol{p} = (\boldsymbol{s},t)$ in the continuum of spatial coordinates $\boldsymbol{s} = (s_1,s_2)$ and time t. The BME technique for stochastic S/TRF analysis is applied to each one of the NO_2 and SO_4 datasets separately to predict the pollutant distributions across the specified grids during the study period.

The SEKS-GUI software library (*22*) is used to implement the space-time BME analysis. SEKS-GUI is an acronym for "Spatiotemporal Epistemic Knowledge Synthesis-Graphical User Interface". The software processes the fundamental BME equations of spatiotemporal dependence analysis and mapping that are presented in the Supporting Information section S2. These equations represent in a concise and coherent way the various elements of real world space-time analysis and prediction, including knowledge bases (KB), space-time geometry, and the probabilistic description of the phenomenon. The BME equations are mathematically very general, in the sense that they make no restrictive assumptions about the underlying probability distributions (non-Gaussian laws are automatically incorporated) and the shape of the space-time predictor (non-linear predictors are allowed). Hence, the BME framework is free of the restrictive assumptions commonly made in classical geostatistics and statistical regression modeling, and can thus operate on a broader scope of KB types and uncertain data (*23*).

The fundamental BME equations integrate general KB (*G*-KB) and site-specific KB (or *S*-KB) and provide a complete stochastic assessment of each pollutant at a set of predefined space-time nodes. The *G*-KB may include physical laws, theoretical models of space-time dependence (covariance, semivariogram, etc.), empirical relations, and logic-based assertions that are related to the pollutant $X_{\boldsymbol{p}}$. The site-specific information *S*-KB usually consists of observed hard and soft (probabilistic, intervals, etc.) data. In the present study, the *G*-KB consists of theoretical covariance models. The *S*-KB includes pollutant data. All observations have been reported as hard values, i.e., the study measurements do not include significant observation uncertainty. Hence, in the absence

of soft data the BME methodology mathematically reduces to BME kriging (*8*, Chapter 12).

SEKS-GUI represents the prediction grid by the space-time vectors $\boldsymbol{p}_k$, in which case the fundamental BME equations compute the complete prediction pdf f_K at each $\boldsymbol{p}_k$. Given f_K, different $X_{\boldsymbol{p}}$ predictions can be derived at each spatiotemporal node $\boldsymbol{p}_k$ of the mapping grid depending on the objectives of the study. In view of the *G*- and the *S*-KB used in the present study, the f_K mean (or *BMEmean*) is chosen as the pollutant predictor.

BME thus generates informative pollutant maps that completely cover the spatial and temporal continua within their respective extents. This enables asking new types of questions about complex spatiotemporal phenomena, such as the following: Are there pronounced regional patterns in how pollution at various cells has *changed* over time? Have some cells begun to diverge from each other, i.e., they used to have similar annual patterns, but behaved quite differently in later years? Similarly, has the behavior of particular pollutants – relative to other pollutants – changed, i.e., did they used to rise/fall in unison but not any more? Those are precisely the types of questions that the spatialization component of this project aims to address.

3.2. *Conceptualizing spatialization*

The use of spatialization in processing BME outputs is fundamentally rooted in a conceptualization of multi-temporal pollution measurements and interpolations as existing in various high-dimensional spaces. One can postulate different such spaces even for a single dataset, depending on the specific research questions one wants to pursue. Transformations between different spaces then become the basis for multi-perspective visualizations.

More specifically, in this study the BME-generated pollutant predictions in a composite space-time domain are interpreted within a *tri-space* that is formed by geographic space, time, and attributes (*24*). When dealing with cells at *S* geographic locations having been given values for *A* attributes at *T* different times, we are faced with a *single* set of $S \times A \times T$ observations (an *SAT* dataset). One can systematically explore this *tri-space* of *SAT* values by constructing a series of matrices in which rows and

column are constructed from combinations of tri-space elements. Figure S2 in Supporting Information gives a schematic representation of this concept.

For example, an individual row is identified with a composite identifier pointing to a particular geographic location $\boldsymbol{s}_i$ and a particular time slice t_k, whereas a column corresponds to particular attribute a_j. Hence, we would be looking at $S \times T$ space-time loci, existing in an A-dimensional space (the present study has a two-dimensional attribute space for pollutants NO_2 and SO_4). The questions one could then ask would deal with relationships among space-time loci. In addition, considering that T is part of the composite row identifier and that there tends to be known topology among different t_k –namely a particular, "natural" sequence– we can conceptualize an individual cell $\boldsymbol{s}_i$ as following a trajectory through the A-dimensional space. This would allow asking questions about multi-temporal similarity of multiple $\boldsymbol{s}_i$, including invoking notions of parallelism, divergence, and convergence, as demonstrated in (*25*) with multi-temporal Census data for Texas counties.

Alternatively, the same set of $S \times A \times T$ measurements could be transformed into a different matrix in which rows correspond to combinations of $\boldsymbol{s}_i$ and a_j, whereas columns correspond to different t_k. Altogether, six different matrices could be derived from one *SAT* source data set. Each of these six matrices offers a different perspective on the source pollutant data. In this case, different questions can be asked about the similarities of location-attribute composites. One could ask, e.g., whether NO_2 concentrations at one location have over time behaved similar to SO_4 concentrations at *another* location.

3.3. *Dimensionality reduction*

The dimensionality of row vectors is typically too large to be directly plotted. Some type of dimensionality reduction is then necessary for a visual depiction. For matrix-type data, the more popular options include multidimensional scaling (MDS), principal components analysis (PCA), and self-organizing maps (SOM). Of these, the SOM method is particularly suited for dealing with very voluminous and high-dimensional datasets (*16*). Applications of the SOM method abound, including for geographically referenced environmental data (*17*). There are two main products when using SOM: (i) a

low-dimensional model of the high-dimensional input space, the SOM itself; and (ii) a low-dimensional representation of high-dimensional vectors after they are mapped onto the trained SOM (*17*, pp. 1-20). The former is often useful in explaining patterns observed in the latter, once both are given visual form, as demonstrated in section 4.2.

3.4. *Visualization*

Once high-dimensional structures and relationships are expressed in a low-dimensional geometric form, they can theoretically be made visual. However, effective visualization depends on being able to further transform geometric structures in imaginative ways. Geographic information systems (GIS) are already uniquely suited to this goal, regardless whether the low-dimensional space in question is geographically referenced.

4. Implementation and results

4.1. *BME prediction*

First, the NO_2 and SO_4 data are processed to filter out potential surface (mean) trends. Mean trends represent larger scale variations that can obscure the underlying space-time dependence structure in the study scale. The SEKS-GUI software removes estimates of mean trends (see section S3, Supporting Information) and restores the trend component after the prediction computations.

For each one of the NO_2 and SO_4 pollutants, the detrended residuals are used to compute the empirical correlation among the pollutant values in space-time. This is used to fit a theoretical covariance model, which is key input to the prediction computations. Section S3 in Supporting Information contains the technical details regarding the correlation analysis in our study.

Figure 2 shows the predicted *BMEmean* values for NO_2 and SO_4 obtained by SEKS-GUI across the California grid for a few selected time instances. The examples shown in Figure 2 are characteristic of the pollutants' behavior, in that throughout the study period the highest pollutant concentrations are consistently found in the Los Angeles area. Specifically, there is an evident seasonality in the predicted NO_2 monthly averages across the state. These averages tend to peak during the winter months; an absolute maximum monthly mean of about 76 ppb over the 15-year period study was

predicted at $t_M = 95$ (November 1995). Overall, the BME analysis showed the annual mean of the highest predicted monthly values in California to range between about 33 to 49 ppb; these values are above the 30 ppb average annual limit set by the state, but they all fall below the corresponding current federal limit of 53 ppb. The NO_2 standard prediction error has been consistently estimated to be below 9.2 ppb at any location and exhibits typical values around 5.5 ppb. The SO_4 annual concentrations display no noticeable trend or significant fluctuations during those years. The highest annual average is predicted at nearly 7 $\mu g/m^3$ at $t_Y = 4$ (1991) and the maximum standard error value in the predictions was 0.82 $\mu g/m^3$.

Based on the original temporal grid specifications, there is a series of predictions at 150 monthly time instants for NO_2 and 15 annual instants for SO_4 between 1988 and 2002. These are the BME analysis results of the two pollutants that are forwarded to the spatialization segment of the BME-S approach.

4.2. *Spatialization*

This paper discusses two visualizations derived from NO_2 and/or SO_4 BME predictions, with the goal of illustrating the types of novel investigations that can be developed using a spatialization methodology. Considering the geostatistical generation of air pollution predictions in our selected spatiotemporal lattice of cells, we explore how spatialization might be used to *holistically* explore this data set. Only cells with predictions for both variables and all time periods (all 180 months for NO_2 and all 15 years for SO_4) were kept, and this explains the usefulness of the BME methodology in the proposed BME-S approach. Spatial filtering further eliminated all cells whose Voronoi region did not intersect California. The resulting set of geographic locations S consists of 204 cells.

The first experiment reported here focuses on just the NO_2 variable. Thus, though this is not a multi-attribute data set, with 204 cells and 180 time slices it still involves 36,720 NO_2 values. The first experiment arranges the monthly data in sequence to form a matrix of 204 rows representing geographic loci (S) and 180 columns representing temporal loci (A). One is thus able to ask questions about the temporal similarity of NO_2 observations across the 204 cells.

Before proceeding with dimensionality reduction (from 180 to 2 dimensions), NO_2 values were normalized in three alternative ways, each meant to illuminate different relationships among the cells. In each case, normalization is based on scaling values to a 0-1 range, proportional to the minimum and maximum NO_2 values.

(a) Global normalization

All 36,720 NO_2 values are here normalized to a 0-1 range in a single step, based on the smallest and largest monthly mean ever observed for any cell and any time period. Similarities among cells observed in the visualization (as expressed through similar colors) will thus be largely reflective of differences in the magnitude of NO_2 concentrations.

(b) Time normalization

Normalization occurs here in isolation for each time slice, based on minimum and maximum values for the respective slice. Assuming that the geographic distribution of relative NO_2 concentrations at different times is relatively constant –in that the relative ranking of cells does not change despite changes in absolute magnitude– one would expect patterns of cell similarity to be close to what is produced by global normalization.

(c) Cell normalization

Another form of normalization occurs within cells, i.e. within rows of the input matrix. With the smallest and largest value ever observed for a particular cell driving the normalization, this leads to more direct comparison of temporal NO_2 signatures. For example, it allows temporal alignment of local maxima and minima of different cells to be recognized despite differences in magnitude. One would expect that broad regional patterns effecting NO_2 concentration might come to the fore in this approach, since regional causes may drive concentrations up or down in similar patterns, irrespective of the magnitude of pollutant concentrations.

Three different input files were prepared according to the above three normalization approaches. From each one of these files, a SOM consisting of 16 neurons (4x4) was derived and the best-matching neuron was determined for each of the 204

cells. While this choice of granularity may seem arbitrary, it was informed by the desire to project the resulting grouping of the 204 cells from neuron space into geographic space. At 16 cells one can already represent significant variation. One advantage of the SOM method is that topological relationships among neuron-based clusters are explicitly represented, which can be considered when making color choices. In the representation of the SOM itself (i.e., the 4x4 neuron lattice) in Figure 3, complementary colors were chosen for opposite ends (i.e., red-green and orange-blue), and other neurons' colors correspond to transitional mixtures.

In the geographic map, each neuron receives the color of its best-matching neuron. In addition, the high-dimensional distance of the geographic cell from the neuron centroid is expressed as transparency, allowing visualization of within-cluster variation. For example, within the "red" cluster, fully saturated red corresponds to cells that are near the cluster centroid.

Albeit the three solutions shown in Figure 3 are similar, they are completely independent with no coordination of color schemes. For example, cells with generally high values of NO_2 concentration are found in the "red" cluster for the globally normalized solution in Figure 3a, but in the "green" cluster for the time normalized solution in Figure 3b. The main purpose of these visualizations is to observe multi-temporal regionalization of NO_2 concentrations across geographic space.

To explain the observed patterns, one would have to examine the content of neurons vectors. With every neuron being associated with a 180-dimensional vector, that would be a difficult task to achieve. To illustrate the principle, bar charts have been selectively generated and shown in Figure 3 from each neuron's trained value for the January component of all 15 years, running from 1988 to 2002 (left to right). An individual bar expresses how high the NO_2 values were for that particular January, according to the specific normalization approach. Again, one would have to either visually or computationally examine neuron values to understand precisely how the 180-dimensional space is structured by each 16-neuron SOM.

As expected, the globally normalized and time normalized SOM show very similar patterns. In the case of the extreme North Coast and the Central Coast and Coastal Range, one can observe similarity despite geographic separation. These areas

consistently exhibit the lowest NO_2 concentrations. In some cases, one can clearly observe transitions, such as when one moves from a corridor around Interstate 10 (I-10) northward. In the globally normalized solution, one moves from the red cluster gradually through the pink, then purple, then blue clusters. This corresponds to a predicted gradual decline of NO_2 values. However, notice also a relatively transitionless cross from the blue to the yellow cluster. This corresponds to a move from the orange to the blue/light-blue regions in the time-normalized solution. It is apparent that the Eastern Sierra Nevada and Owens Valley are quite different from the desert region north of the I-10 highway. Notice also the linear arrangement of yellowish-green cluster cells (Figure 3a) and purple cluster cells (Figure 3b) in the Central Valley along Interstate 5 (I-5) and how it is in both solutions separated from the neighboring Coastal Range by a full cluster. Los Angeles, with its known high NO_2 concentrations is placed within a contiguous cluster running from the coast along I-10 to the Arizona border. However, it is extreme enough to be located relatively far from the neuron's center, thus appearing less saturated than the rest of this cluster.

The cell-normalized solution is quite different (Figure 3c). As mentioned before, its goal is to reduce the effects of NO_2 magnitude to more clearly focus on similarities in temporal NO_2 regimes. In the bar charts there tends to be a wider range of bar lengths, indicating that January NO_2 concentrations varied significantly during the 1988-2002 time span. One can now observe that the generally low-NO_2 areas in the Coastal Range did show a very different temporal regime than its low-NO_2 counterparts at the North Coast. In fact, these two areas now appear in opposite clusters, with the red clusters showing the least amount of January NO_2 variation, while the North Coast saw a dramatic drop of NO_2 from its 1989 maximum. The Central Coast is now grouped with the adjacent portions of the Central Valley, indicative of underlying regional factors affecting NO_2 concentrations, despite the observed differences in NO_2 magnitude. In the cell normalized solution, the Mojave Desert region has become separated from the Los Angeles basin, due to its large rise of predicted January NO_2 concentrations following the 1997/98 minimum, with values stabilizing at a high level by 2001. Again, these are patterns that would have been obscured in a magnitude-focused analysis.

In a different analysis, an integration of NO_2 and SO_4 data was performed, leading to the visualization shown in Figure 4. One major advantage of the geostatistical generation of cell-level predictions is that NO_2 and SO_4 can be integrated despite differences in the spatial distribution of monitors. Spatial integration is thus feasible. Temporally, with SO_4 available as annual values, NO_2 had to be aggregated to the annual level. In this experiment, simple averaging of mean monthly values was used. One can then proceed to construct an *SAT* data set that actually has multiple attributes. The question pursued at this stage of the study was whether there are differences in the temporal regimes of NO_2 and SO_4. The 204 cells were conceptualized as existing in a 15-dimensional space, but with two instances of each cell, one for each attribute (see also Figure S2 in Supporting Information). The input data set thus generated consists of 408 vectors or rows. These were normalized to a 0-1 range at the level of the individual vector, with the goal of enabling comparison across multiple vectors *and* attributes, such that geographic patterns as well as attribute-level distinctions become apparent. A SOM of 2,500 neurons (50×50) was generated, and the 408 cell-attribute vectors were mapped onto it (Figure 4). If NO_2 and SO_4 concentrations were evolving in lock-step, subject to the same forces in the context of a particular cell, then we would expect to see *no* organized patterns in the distribution of NO_2 and SO_4 across the spatialization. However, a very strong organization can be observed instead, with most of the two pollutants being separated into contiguous regions in the SOM. It is apparent that the dominant pattern for NO_2 is the rapid decline of predicted annual concentrations after peaking around 1990, with a slight increase after a 1997 minimum. Meanwhile, SO_4 showed a similar pattern up to the 1997 minimum, but many cells then seem to have experienced a rapid rise in SO_4 concentrations after that. Exceptions from these broad patterns include the continued decline of SO_4 concentrations in the extreme north of the state and the rapid rise of predicted NO_2 concentrations in the Mojave Desert north of the I-10 highway. The SOM also registers graduated temporal regimes, such as on when one crosses the imaginary boundary between NO_2 and SO_4 regions in the center of the SOM. However, the contiguity of these regions even there points to a need to investigate differences in NO_2 / SO_4 regimes further.

With respect to Figures 3 and 4, this study did not investigate the relationships between observed patterns and the original monitor measurements. The space-filling prediction was accepted at face value, though some of the more consistent patterns occur relatively far from monitor locations, such as in the extreme north and east of the state.

Some additional technical details about the SOM implementation and one more visualization example are presented in the Supporting Information sections S4 and S5, respectively.

5. Discussion

The present study explored a new framework for a multi-perspective analysis and visualization of monitor data in space-time. The BME-S framework was introduced, which is based on the methodological contributions of the BME method for geostatistical prediction, and spatialization for visualization of high-dimensional data.

The BME component takes over the initial part of the analysis, where observations from monitors and a variety of other general and case-specific information sources can be gathered together to predict air pollution attributes in space and time by using an elaborate epistemic methodology. Spatialization depends on this output to perform its more introspective role, where the multiple dimensions of different attributes, locations and time instants are transformed in a variety of possible ways. These features offer additional paths to view and interpret the observed data and enhance the analysis of the attributes in a space-time context.

Our work exhibits for the first time this fruitful linking between the rigorous spatiotemporal prediction features of BME and the powerful dimensional reduction and visualization methodology of spatialization. The joint approach enables investigating and exploiting the characteristics of one or more sampled attributes in both space and time. Two pollutants (NO_2 and SO_4) were examined in this context, and a series of such investigations were illustrated with respect to the pollutants' spatiotemporal distributions and individual/joint regional pattern changes in space-time.

Usually, air pollution studies with more than one attribute consider the additional attributes as covariates that are suspected or known to have some effect on the main attribute. For example, a study that involves NO_2 might consider O_3 as a covariate

attribute, because these two gases often exhibit association and interaction in their spatiotemporal behavior. In our study, the two pollutants were selected to be relatively unrelated and non-interactive attributes, so that the BME-S approach can test potential connections between NO_2 and SO_4 on the basis of their spatiotemporal patterns and pattern changes alone. As illustrated in section 4.2, spatialization in BME-S provides a useful platform to explore potential links in similar scenarios. The findings suggest that the proposed approach can be a very attractive and helpful tool to enhance the spatiotemporal study of air pollution attributes.

Acknowledgements: Support for this work was provided by a grant from the California Air Resources Board, USA (Grant No. 55245A). We extend our appreciation and credit Wyson Pang for helping with some of the data arrangements and computer runs.

Supporting Information Available: Some additional figures and more technical details about the methodologies used in this manuscript are presented separately, as indicated in the manuscript text. This information is available free of charge via the Internet at http://pubs.acs.org .

References

(1) Jerrett, M.; Burnett, R.T.; Pope, C.A. III; Ito, K.; Thurston, G.; Krewski, D.; Shi, Y.; Calle, E.; Thun, M. Long-Term Ozone exposure and mortality. *The New England Journal of Medicine*. **2009**, 360, 1085-1095.

(2) Haining, R.; Law, J.; Maheswaran, R.; Pearson, T.; Brindley, P. Bayesian modelling of environmental risk: example using a small area ecological study of coronary heart disease mortality in relation to modelled outdoor nitrogen oxide levels. *Stochastic Environmental Research and Risk Assessment*. **2007**, *21*(5), 501-509.

(3) Bayraktar, H.; Sezer, F.; Turalioglu F. S. A Kriging-based approach for locating a sampling site—in the assessment of air quality. *Stochastic Environmental Research and Risk Assessment.* **2005**, *19*(4), 301-305.

(4) Kaya, I.; Cengiz Kahraman, C. Fuzzy robust process capability indices for risk assessment of air pollution. *Stochastic Environmental Research and Risk Assessment* **2009**, *23*(4), 529-541.

(5) Kumar, U.; Jain, V. K. ARIMA forecasting of ambient air pollutants (O_3, NO, NO_2 and CO). *Stochastic Environmental Research and Risk Assessment.* **2010**, *24*(5), 751-760.

(6) Li, H. L.; Huang, G. H.; Zou, Y. An integrated fuzzy-stochastic modeling approach for assessing health-impact risk from air pollution. *Stochastic Environmental Research and Risk Assessment.* **2008**, *22*(6), 789-803.

(7) Maxwell, R. M.; Kastenberg, W. E. A model for assessing and managing the risks of environmental lead emissions. *Stochastic Environmental Research and Risk Assessment.* **2008**, *13*(4), 231-250.

(8) Christakos, G. *Modern Spatiotemporal Geostatistics*; Oxford University Press: New York, NY, 2000.

(9) Skupin, A.; Fabrikant, S. I. Spatialization. In *The Handbook of Geographic Information Science*; Wilson, J. P., Fotheringham A. S., Eds.; Blackwell Publishing: 2008; pp 61-79.

(10) Christakos, G.; Kolovos, A. A study of the spatiotemporal health impacts of the ozone exposure. *J. Exposure Anal. and Env. Epidemiol*. **1999**, *9*, 322-335.

(11) Christakos, G.; Kolovos, A.; Serre, M. L.; Vukovich, F. Total ozone mapping by integrating data bases from remote sensing instruments and empirical models. *IEEE Trans. on Geosc. and Rem. Sensing*. **2004**, *42*(5), 991-1008.

(12) Bogaert, P.; Christakos, G.; Jerrett, M.; Yu, H-L. Spatiotemporal modelling of ozone distribution in the State of California. *Atm. Envir*. **2009**, *43*, 2471–2480.

(13) De Nazelle, A.; Serre, M. L. Ozone exposure assessment in North Carolina using Bayesian Maximum Entropy data integration of space time observations and air quality model prediction. *Epidemiology*. **2006**, *17*(6), S189.

(14) doi:10.1021/es100228w De Nazelle A.; Arunachalam, S.; Serre M.L. Bayesian Maximum Entropy integration of ozone observations and model predictions: an application for attainment demonstration in North Carolina. *Environmental Science & Technology*. **2010**, In press.

(15) doi:10.1016/j.atmosenv.2010.04.030 Yu, H-L; Wang, C-H. Retrospective prediction of intraurban spatiotemporal distribution of PM2.5 in Taipei. *Atmospheric Environment*. **2010**, In press.

(16) Kohonen, T. *Self-Organizing Maps*; Springer-Verlag: Berlin, 2001.

(17) Agarwal, P.; Skupin, A., (Eds.) *Self-Organising Maps: Applications in Geographic Information Science*; John Wiley & Sons: Chichester, England, 2008.

(18) Jerrett, M.; Newbold, K. B.; Burnett, R. T.; Thurston, G.; Lall, R.; Pope III, C. A.; Ma, R.; De Luca, P.; Thun, M.; Calle, J.; Krewski, D. Geographies of uncertainty in the health benefits of air quality improvements. *Stochastic Environmental Research and Risk Assessment.* **2007**, *21*(5), 511-522.

(19) Gilliland, F.; Avol, E.; Kinney, P.; Jerrett, M.; Dvonch, T.; Lurmann, F.; Buckley, T.; Breysse, P.; Keeler, G.; DeVilliers, T.; McConnell, R. Air pollution exposure assessment for epidemiologic studies of pregnant women and children: Lessons learned from the centers for children's environmental health and disease prevention research. *Env. Health Persp*. **2005**, *113*(10), 1447-1454.

(20) Zhuang, H.; Chan, C. K.; Fang, M.; Wexler, A. S. Size distributions of particulate sulfate, nitrate, and ammonium at a coastal site in Hong Kong. *Atmospheric Environment*. **1999**, *33*, 843-853.

(21) Christakos, G. *Random Field Models in Earth Sciences*; Academic Press, Inc.: San Diego, CA, 1992.

(22) Kolovos, A.; Yu, H-L.; Christakos, G. *SEKS-GUI v.0.6 User Manual*. Dept. of Geography, San Diego State University, San Diego, CA, 2006.

(23) Christakos, G.; Hristopulos, D. T. *Spatiotemporal Environmental Health Modelling*; Kluwer Academic Publ.: Boston, MA, 1998.

(24) Openshaw, S. Two exploratory space-time attribute pattern analysers relevant to GIS. In *GIS and Spatial Analysis*; Fotheringham, S., Rogerson, P., Eds.; Taylor & Francis: London, 1994; pp 83-104.

(25) Skupin, A.; Hagelman, R. Visualizing Demographic Trajectories with Self-Organizing Maps. *GeoInformatica*. **2005**, *9*(2), 159-179.

List of Figures

Figure 1

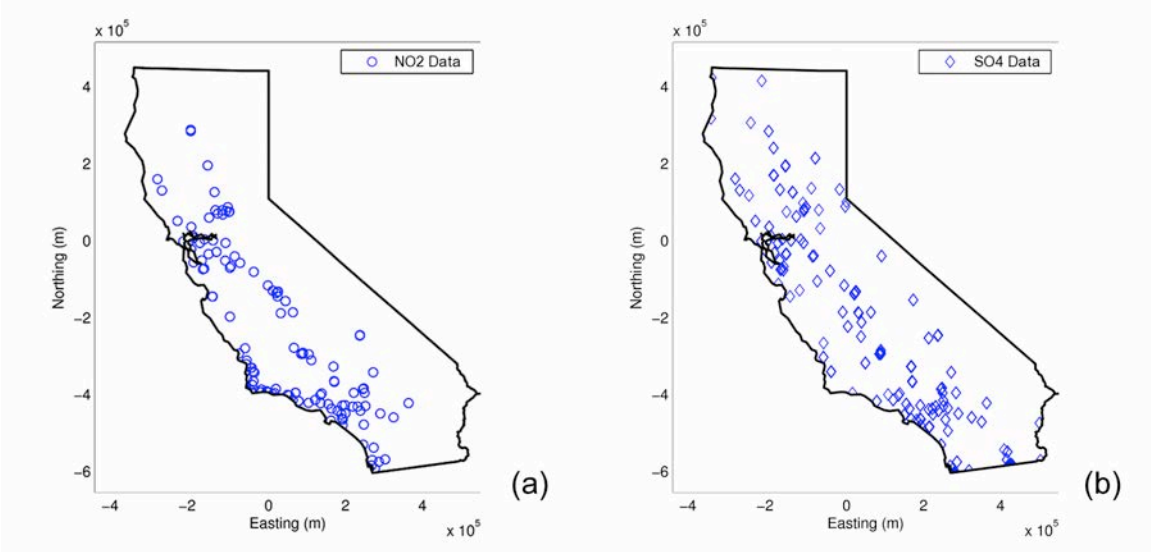

Figure 1: Sites of measurement monitors for the (a) NO_2 observations, shown as circles, and (b) SO_4 observations, shown as diamonds, during 1988-2002.

Figure 2

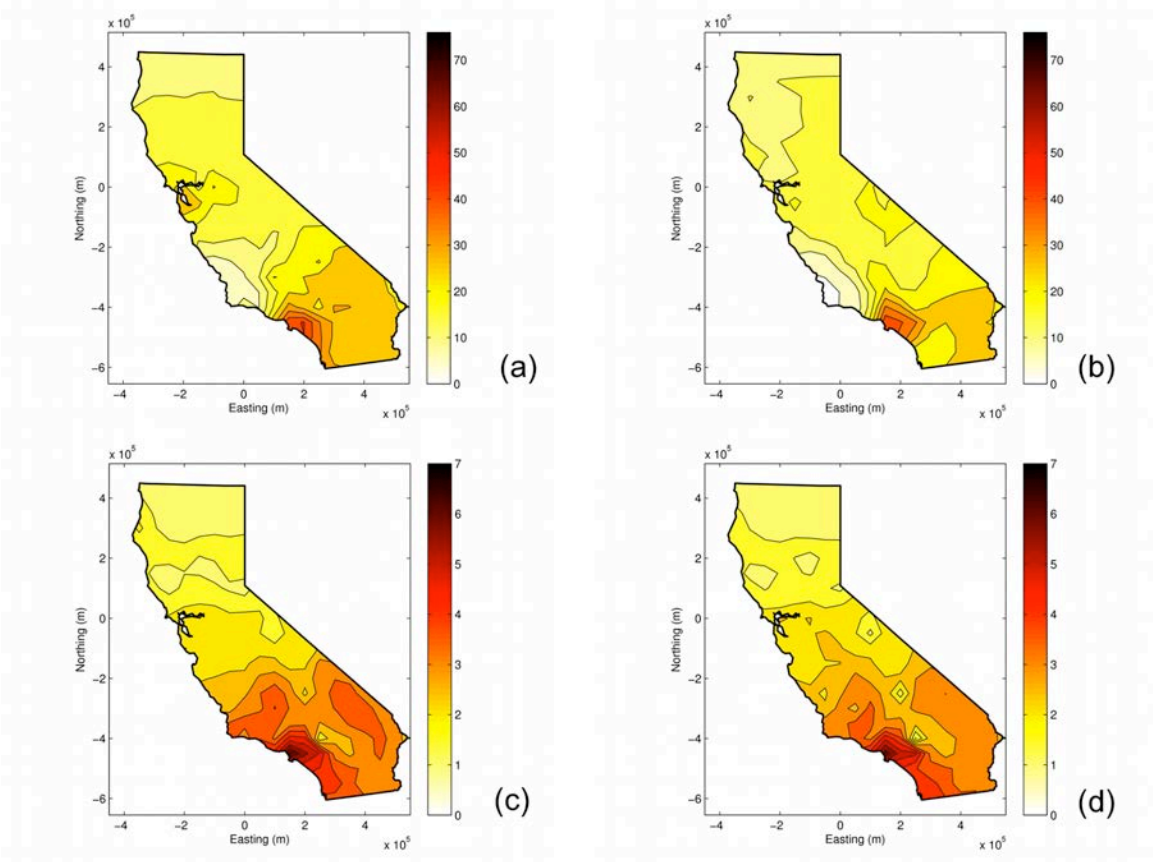

Figure 2: *BMEmean* maps for monthly-averaged NO_2 (in ppb units) and annual-averaged SO_4 (in $\mu g/m^3$ units) concentrations across California at selected time instances. The plot shows the NO_2 predicted distributions means at months (a) $t_M = 1$ (January 1988) and (b) $t_M = 7$ (July 1988), and the SO_4 predicted distributions means at years (c) $t_Y = 1$ (1988) and (d) $t_Y = 2$ (1989).

Figure 3

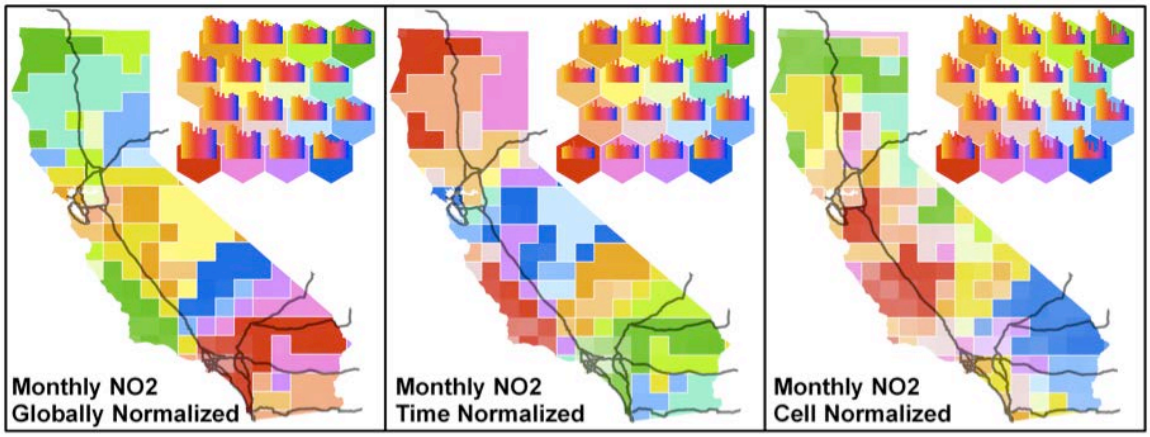

Figure 3: Monthly NO_2 values for 1988-2002 modeled as 180-dimensional vectors that are used to train a 16-neuron SOM. Geographic cells are classified based on matching SOM neurons. Normalization of cell values to a 0-1 range prior to SOM training based on (a) values of all cells across all time slices, (b) cell values within time slices, (c) values within same cell across all time slices. Transparency of cells in geographic map is dependent their n-dimensional distance from respective class center. Bar charts show annual sequence of January NO_2 values according to neuron vectors. Note that neurons actually have values for all 180 time slices, which one would need to investigate in order to completely understand how the SOM space is structured. In these plots, the gray lines designate Interstate highways crossing California. Interstate 5 (I-5) is on the north-south axis. Interstate 10 (I-10) is on the east-west axis in the southern part of the state.

Figure 4

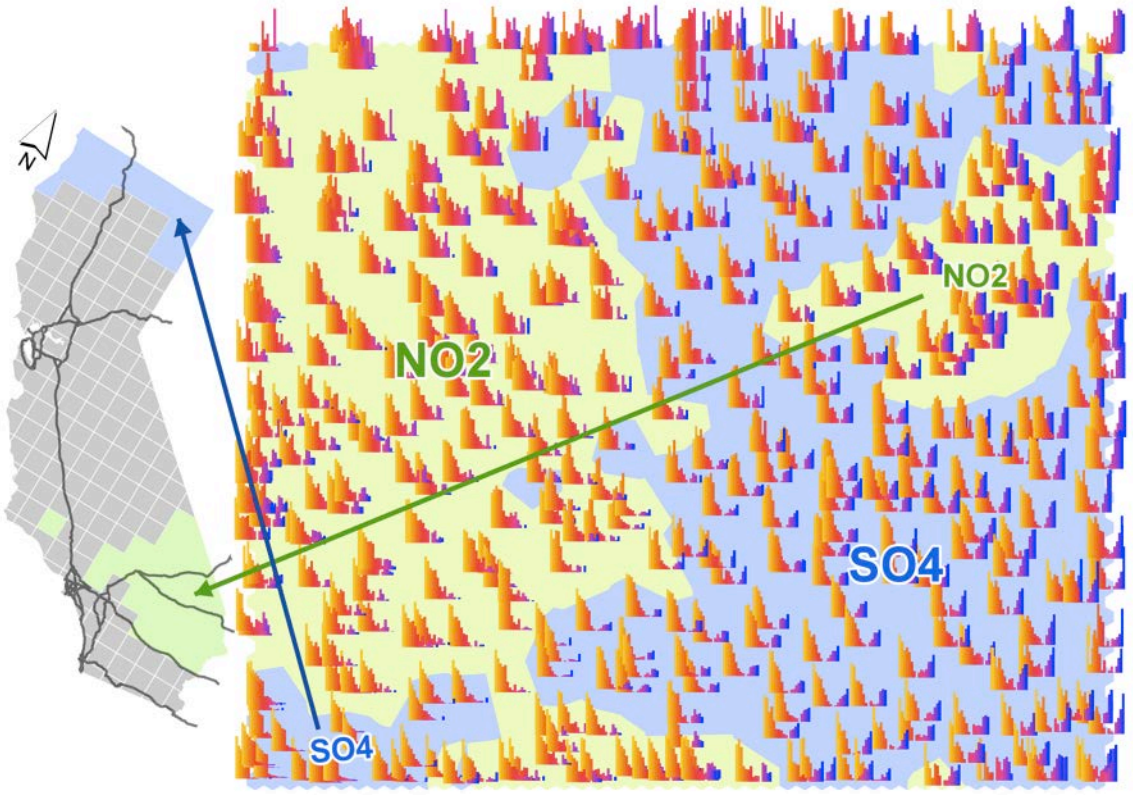

Figure 4: Annual NO_2 and SO_4, each modeled as 15-dimensional vectors, with two vectors per geographic cell, separated according to the two attributes. Input vectors are normalized to 0-1 range within the same cell across all 15 annual time slices. Vectors are used to train a 2500-neuron SOM and are then mapped onto it. Bar charts show sequence of annual values between 1988 and 2002. Notice the relative contiguity of regions formed by the two attributes. Notable exceptions are highlighted in geographic space.

Brief

A methodology is introduced to visualize high-dimensional data, which combines the Bayesian Maximum Entropy spatiotemporal prediction technique and the cognitively informed approach of Spatialization.

MULTI-PERSPECTIVE ANALYSIS AND SPATIOTEMPORAL MAPPING OF AIR POLLUTION MONITORING DATA

Alexander Kolovos,[1*] André Skupin,[2] Michael Jerrett,[3] and George Christakos[2]

1. SAS Institute, Inc., 100 SAS Campus Dr. S3042, Cary, NC 27513, USA.
2. Department of Geography, SDSU, San Diego, CA 92182-4493, USA.
3. School of Public Health, UC Berkeley, CA 94720-7360, USA.

Supporting Information For Publication

Number of Pages: 9

Number of Figures: 4

Number of Tables: 1

[*] To whom correspondence should be addressed: kolovos@email.unc.edu (phone: +1-919-531-2165)

S1. Nitrogen Dioxide and Sulfate Data Information from Monitored Measurements

The pollutant predictions complement the collected information from the individual monitoring stations. The information from the monitor measurements provides some basic characteristics about the pollutants. For example, plots of the monthly NO_2 concentrations at individual monitoring stations suggest a clear seasonal behavior of NO_2 with concentration peaks around the winter months. Also, similar plots of the annual-averaged SO_4 concentration values reveal a smoother behavior that is associated with longer-term patterns in the concentration. You can examine samples of pollutant concentration plots at selected individual stations in Figure S1.

S2. The Fundamental Bayesian Maximum Entropy (BME) Equations

The foundation of the BME methodology can be summarized in the following fundamental BME equations of spatiotemporal dependence analysis and mapping:

$$\left.\begin{aligned} &\int d\boldsymbol{\chi}(\boldsymbol{g}-\bar{\boldsymbol{g}})\, e^{\boldsymbol{\mu}^T \boldsymbol{g}} = 0 \\ &\int d\boldsymbol{\chi}\,\xi_S\, e^{\boldsymbol{\mu}^T \boldsymbol{g}} - A\, f_K(\boldsymbol{p}) = 0 \end{aligned}\right\}. \tag{S1}$$

Eqs. S1 integrate the available knowledge bases (KB), general KB (*G*-KB) and site-specific KB (*S*-KB) and provide a complete stochastic assessment of each pollutant at a set of predefined space-time nodes.

Specifically in eqs. S1, $\boldsymbol{g}$ is a vector of g_α-functions ($\alpha = 1, 2, ...$) that represents the *G*-KB, and the bar denotes stochastic expectation. Also, μ is a vector of μ_α-coefficients that depends on the space-time coordinates and is associated with $\boldsymbol{g}$ (i.e., the μ_α express the relative significance of each g_α-function in the composite solution sought), χ stands for the vector of possible $X_{\boldsymbol{p}}$ values (realization) for the spatiotemporal random field $X_{\boldsymbol{p}}$, the ξ_S is an operator that represents the *S*-KB, A is a normalization parameter, and f_K is the attribute probability density function (pdf) at each spatiotemporal point (the subscript K indicates the prediction pdf and means that f_K is based on the blending of *G*- and *S*-KB).

The inputs in eqs. S1 are $\boldsymbol{g}$ and ξ_S, whereas the unknowns are the μ and f_K across space-time locations $\boldsymbol{p}$. Once f_K is known at $\boldsymbol{p}$, different $X_{\boldsymbol{p}}$ predictions can be derived at that location, such as the most probable value of the distribution, the value that minimizes the prediction error, etc.

S3. Mean Trend Estimation and Covariance Analysis in the SEKS-GUI Software

The SEKS-GUI BME analysis features a stage where surface mean trends in the observed data are estimated and removed, before one continues with the empirical estimation of correlation in the data within the study scale. The software currently performs mean trend removal by means of an exponential kernel moving window across the area of interest. The user specifies the spatial and temporal ranges that define the extent of the moving window in the composite space-time domain. For our study, we selected a spatial range of 300 kilometers and a range of 5 temporal instances to detrend each one of the air pollutants.

In the detrending process, the kernel acts upon the observed values in the spatiotemporal window by smoothing them, and values at locations in-between are then interpolated. The resulting set of smoothed values across the domain (at the observations and the grid locations) represents the mean trend estimate. Since there usually exists no definite mean trend, there can be more than one trend estimates. It is possible to obtain different estimates by using different spatial and temporal range parameters that regulate the level of smoothing in the moving window. SEKS-GUI assumes that the resulting residuals are uncorrelated.

The detrended residuals for each one of the pollutants are used to estimate the empirical covariance of the corresponding attribute. A theoretical model is then fitted to these estimates to provide the input for the prediction stage. SEKS-GUI provides a variety of theoretical covariance models to fit. Table S1 lists the forms and coefficients that are used for each pollutant model in our study. In particular, a space-time separable covariance model that is exponential in space and exponential in time represents the spatiotemporal NO_2 correlation. For SO_4, a nested, non-separable spatiotemporal model with two structures in each one of the space and time components is selected. The second model structure employs a much longer temporal range compared to the first

structure. This suggests the presence of relatively elaborate mechanisms in the temporal evolution of the SO_4 concentration. The difference in the temporal ranges of the SO_4 nested model indicates that these mechanisms act in tandem, but the second structure has a much longer lasting influence than the first structure and implies a strong memory effect in the behavior of SO_4. Figure S3 illustrates the fitted theoretical space-time covariances for the two pollutants.

S4. Construction and Visualization of Self-Organizing Maps (SOM)

The freely available software package SOM_PAK (*1*) was used for training of the SOM neural networks discussed in the paper. In each case, a hexagonal neuron arrangement was used and each neuron's *n*-dimensional vector was initialized with random values and trained in the standard two-stage manner, with global structures formed in the first training stage and local structures formed during the second stage (*1*).

In this study, SOMs of different granularity were created in order to illustrate a broader range of applications then has been seen in the past. With a low number of neurons relative to the number of input vectors, the SOM acts mainly as a clustering mechanism, behaving similar to k-means clustering, but with explicit representation of topological relationships between clusters. In the case of the monthly NO_2 values for 1988-2002, the neural network consists of 16 neurons that are trained using a data set of 204 geographic cells. That invites displaying SOM neurons as legend boxes with topologically coordinated color design (Figure 3 in the article). Meanwhile, when there are a relatively large number of neurons, then the SOM acts primarily as a spatial layout mechanism and a detailed geometric layout of the input vectors can be derived. That is the case for the second SOM, in which 2,500 neurons are trained with 408 input vectors, two per geographic cell (Figure 4 in the article).

After SOM training, the codebook file representing each SOM and generated by SOM_PAK is converted to an ESRI Shape file. All further geometric transformation, visualization, and generation of the final figures is performed in ESRI ArcGIS software.

S5. An Additional Visualization Derived from NO_2 BME Predictions

It is possible one might want to discover finer distinctions among cells, specifically enabling the construction of cell trajectories that connect the sequence of annual vectors for each cell. Figure S4 is based on a conceptualization of monthly NO_2 predictions as being part of a 12-month vector for each geographic cell, with each cell thus contributing 15 such vectors, that is, one vector for each one of the 15 years used in our study. The resulting 3,060 input vectors were used to train a high-resolution SOM of 10,000 neurons (100×100).

References

(1) Kohonen, T. *Self-Organizing Maps*; Springer-Verlag: Berlin, 2001.

Table S1

Pollutant	Form (Spatial/Temporal)	Sill	Spatial range	Temporal range
NO_2	Exponential / Exponential	1	100000 m	13 months
SO_4	Exponential / Spherical	0.25	90000 m	2.2 years
	Exponential / Exponential	0.42	90000 m	300 years

Table S1: Spatiotemporal covariance models used in the geostatistical study.

Figure S1

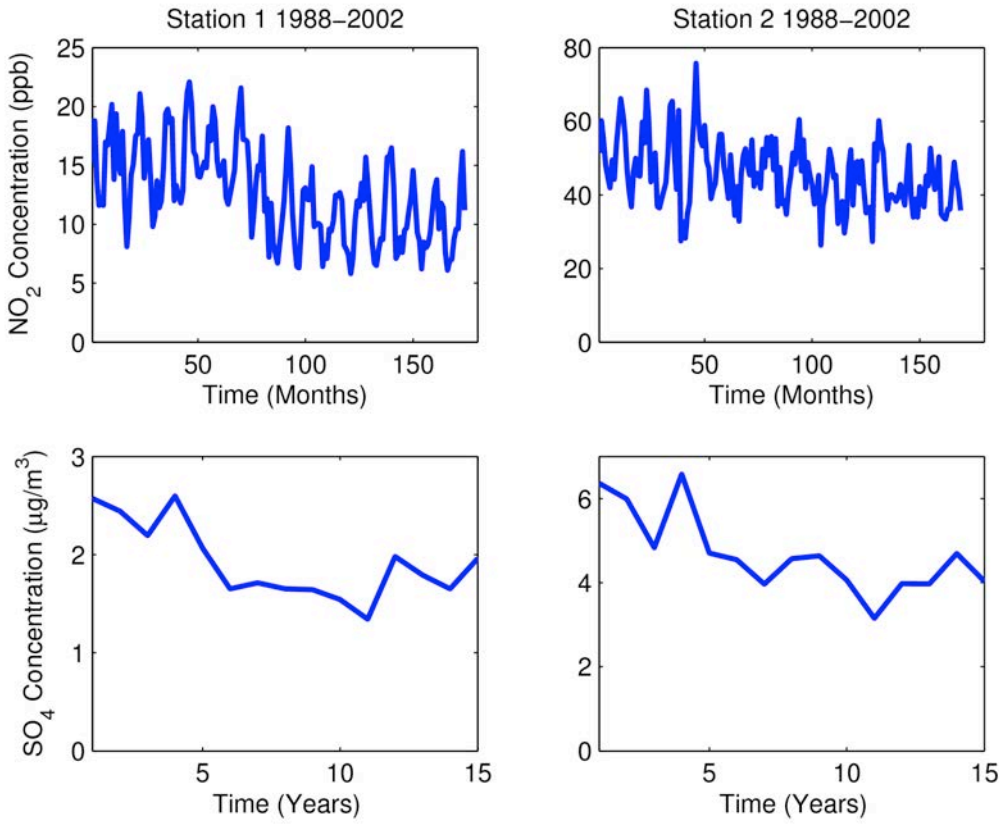

Figure S1: Time series of monthly-averaged NO_2 data in ppb units (upper row) and annual-averaged SO_4 data in $\mu g/m^3$ units (lower row) for two selected stations in the period 1988-2002. Station 1 (left column plots) is at coordinates (-143926.4, 167.9) near Sacramento, and Station 2 (right column plots) is located at (155134.3, -425284.2) in Los Angeles (coordinates are in meters). The x-axis count starts at month 1 (January 1988) for the NO_2 data, and year 1 (1988) for the SO_4 data.

Figure S2

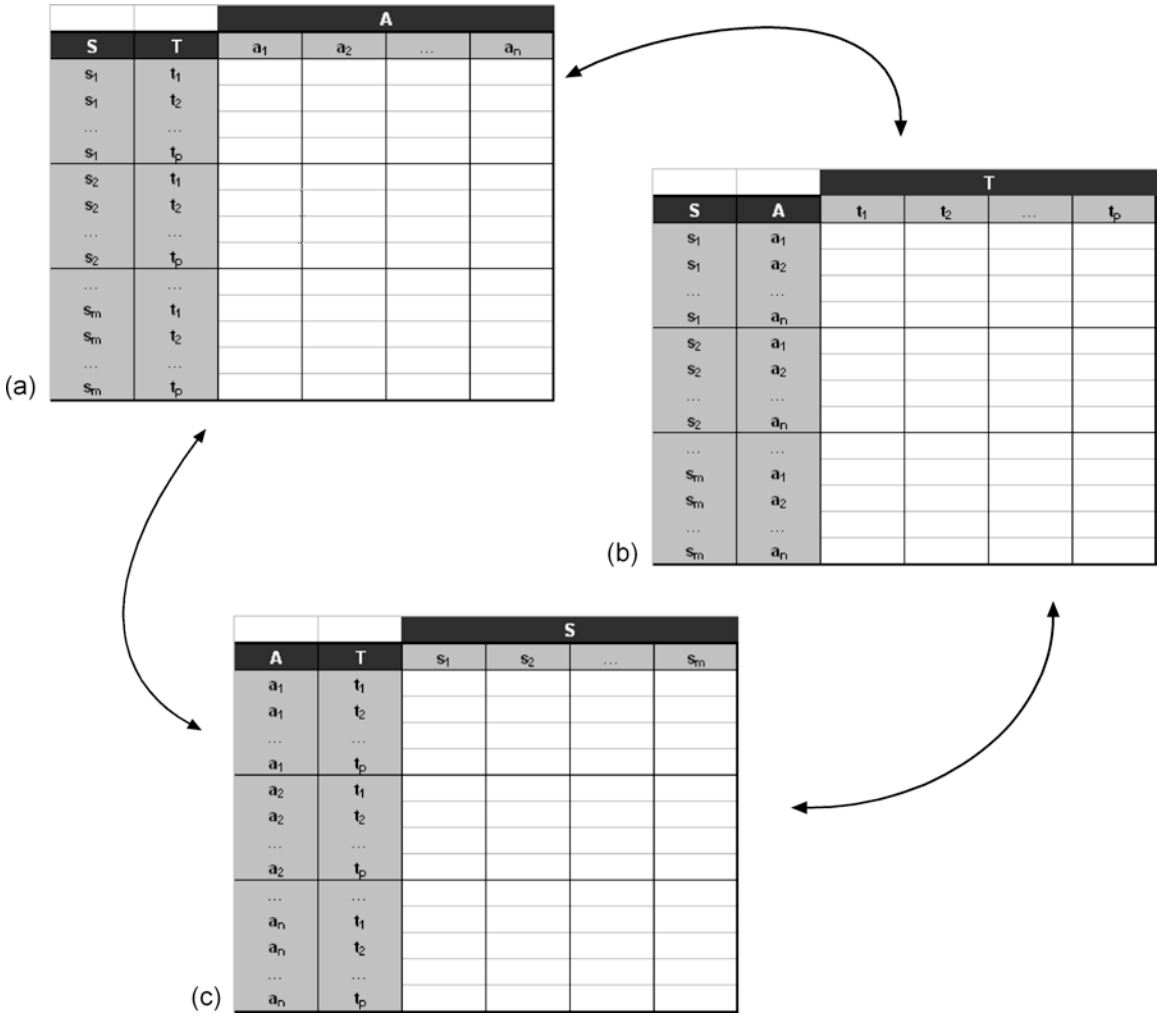

Figure S2: Three different representations that can be derived from the same space-time-attribute data set and could be morphed into each other. Rows are identified as combinations of two of the three tri-space components, leaving the remaining component to identify the columns. Transposing would yield three additional representations.

Figure S3

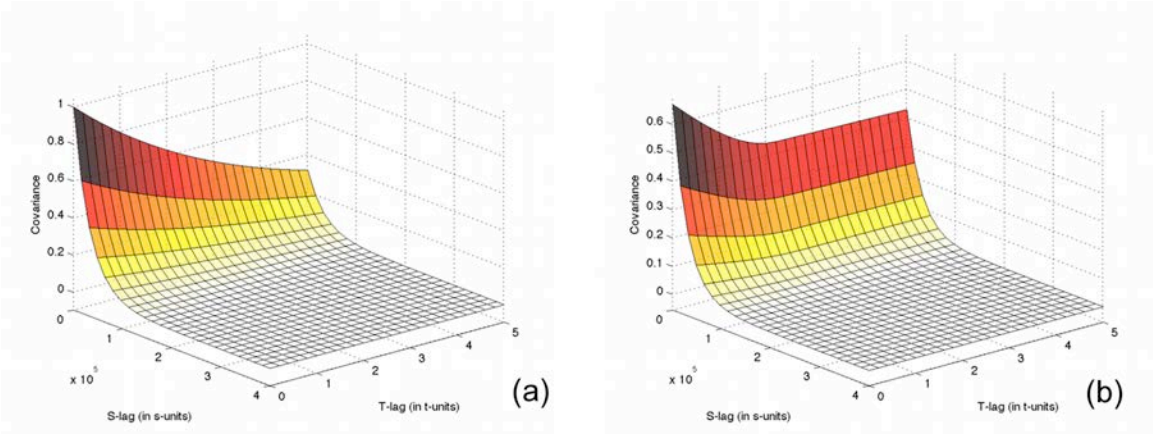

Figure S3: Theoretical spatiotemporal covariances for the (a) NO_2 (s-units are m, t-units are months), and (b) SO_4 (s-units are m, t-units are years) concentration random fields that illustrate the functions in Table S1. The covariance functions are depicted as surfaces that spread across the spatial and temporal distance axes.

Figure S4

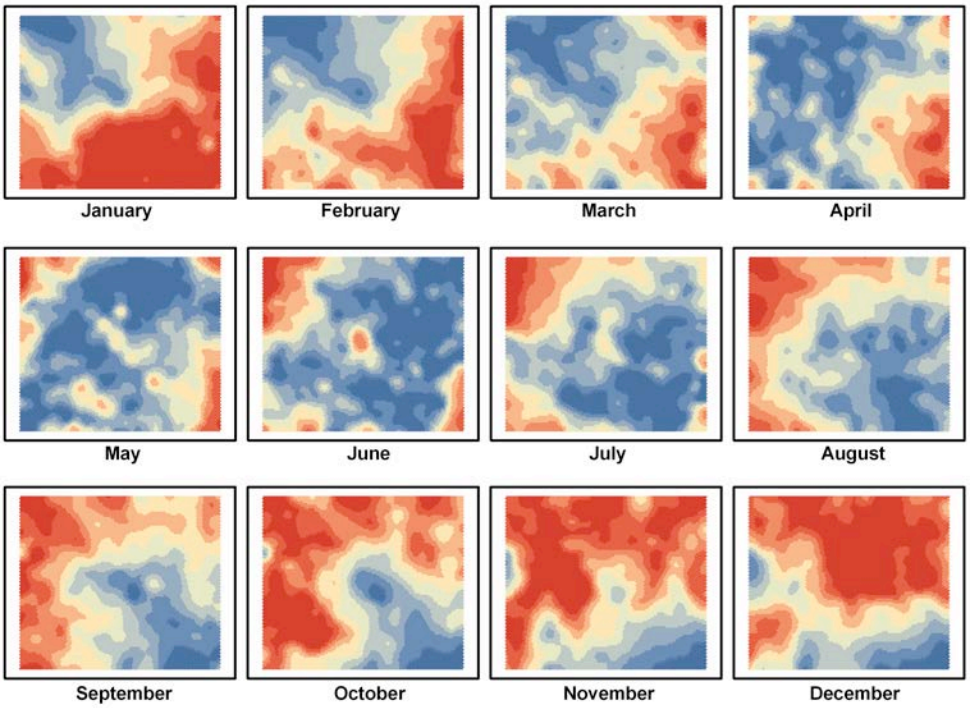

Figure S4: Monthly predicted NO_2 values for 1988-2002 modeled as 12-dimensional vectors, with each cell contributing 15 separate vectors. The 12 component planes are shown, with low-high values indicated as a blue-red color range. These planes clearly illustrate the dominant pattern of November through January as corresponding to the highest NO_2 concentrations, regardless of the absolute values.