Integrating PM_{25} observations, model estimates and satellite signals for the eastern United States by projection onto latent structures

P.S. Porter[1], J. Szykman[2], S.T. Rao[2], E. Gégo[3], C. Hogrefe[4,5] and V. Garcia[2]

[1]University of Idaho, Idaho Falls, ID, USA
[2]AMAD/NERL, U.S.E.P.A., Research Triangle Park, NC, USA
[3]Gego and Associates, Idaho Falls, ID, USA
[4]New York State Department of Environmental Conservation, Albany, NY, USA
[5]Atmospheric Sciences Research Center, University at Albany, Albany, NY, USA

Abstract Detailed, time-varying spatial fields of air contaminant concentrations are valuable to public health professionals seeking to identify relationships between human health and ambient air quality, and policy makers interested in assessing compliance with air quality regulations. In this paper PM_{25} fields are created from a linear model that predicts PM_{25} at unmonitored grid points from observed PM_{25} concentrations, CMAQ model outputs, and satellite estimates of aerosol optical density. The dimensionality of the input data set is first reduced using projection onto latent structures. Parameters of the linear model are mapped to the CMAQ model domain, permitting estimation of PM_{25} at unmonitored sites.

Introduction

Air quality observations are available in the US on a temporally dense but spatially sparse basis. To achieve additional spatial density, observations have been integrated (fused) with outputs of numerical air quality models. A more recent development is the integration of satellite data, used as proxies for ambient concentrations, into fused maps.

Viable integration techniques address the bias of model outputs and efficiently make use of the vast amount of available information. Here fused maps were created with *Projection onto Latent Structures*, also known as partial least-squares regression (PLSR). PLSR is a multivariate regression tool in which a response matrix (Y) is predicted from a matrix of predictors (X). With PLSR, the predictors (X) may be numerous (much greater than the dimension of Y), and correlated with each other (Wold etal, 2001, Trygg and Wold, 2002, Smoliak et al, 2010). The purpose of this paper is to present a technique for producing maps of ambient concentrations from these three sources of information.

Methods

The data used for this study were extracted from the *Remote Sensing Information Gateway* (RSIG, USEPA, 2010)). The goal of RSIG, an interactive web browser-based application hosted by the United States Environmental Protection Agency (USEPA), is to support researcher and analyst data gathering needs (sharing, visualization, and analysis), to extend air quality research and management to the larger air quality community. In addition to satellite data, RSIG hosts air quality observations and air quality model outputs. Multiple datasets can be visualized at the same time and downloaded. RSIG provides users the ability to integrate various data sets across different time and space scales. A valuable RSIG option is the display of different datasets on the CMAQ model grid.

We created fused maps of daily averaged PM_{25} concentrations for the period 1 June 2006 to 31 December 2006 for the eastern U.S. at a resolution of 12 km. PM_{25} concentrations at AIRS sites with continuous PM_{25} monitors were downloaded from RSIG. Sites 80% complete for the period of interest were retained (103 sites). Hourly community Multiscale Air Quality (CMAQ) model estimates of PM_{25} concentrations for the surface layer, and GOES EAST Aerosol/Smoke Product (GASP) Aerosol Optical Depth (AOD) remotely sensed (satellite) signals were also extracted from the RSIG site. Daily averages were formed from the hourly data. Data details can be found at the RSIG web site (USEPA, 2010).

The AIRS PM_{25} sites were randomly divided into 18 'predictor' and 85 'response' sets. The AIRS 'predictor' set, together with CMAQ and GASP, form X. The AIRS 'response' set forms Y. The dimension of X was reduced using PLSR (Matlab PLS function). PLSR transforms the numerous correlated predictors into a limited set of orthogonal (uncorrelated) latent structures defined to both capture X variability and best explain the Y response. PLSR is similar to principal component analysis (PCA) in that the dimension of a set of variables is reduced to a smaller orthogonal set. The goal of PLSR, prediction of Y, differs from the goal of PCA, which is to predict X without consideration of any response variable Y.

Following the detailed description of PLSR found in Wold etal (2001), the set of predictor variables is expressed as the product of 'scores' and 'loads':

$$\mathbf{X = X_S \bullet X_L + E,\quad X = predictor\ matrix} \qquad \textbf{(1)}$$

X dimension = N_t (number of time steps) x N_X (number of predictors)
X_S = X Scores [dimension = N_t x N_{Lv} (number of latent variables)]
X_L = X Loads [dimension = N_{Lv} x N_X]
E = portion of X not explained by $X_S \bullet X_L$

A few X scores explain X ($N_{Lv} << N_x$) and can be expressed as a linear combination of X:

$$\mathbf{X_S = X \bullet W,\quad W = set\ of\ constants} \qquad \textbf{(2)}$$

The response variable is predicted by X_S and the Y loads (Y_L):

$$\mathbf{Y = X_S \bullet Y_L + F, \quad F = Y\ variability\ not\ explained\ by\ the\ model} \qquad \mathbf{(3)}$$

and Y_L is found from:

$$\mathbf{Y = X \bullet W \bullet Y_L\ , \quad Y = X \bullet B\ , \quad B = W \bullet Y} \qquad \mathbf{(4)}$$

Fused maps were created as follows:

1. estimate the B coefficients at monitored grid points
2. map the B coefficients over the entire domain with a 2D cubic spline
3. estimate Y at unmonitored grids using the mapped B coefficients:

$$\mathbf{\hat{Y}\ \ (unmonitored\ \ grid)\ \ =\ \ X\ \ \bullet\ \ (mapping)}$$
(5)

where X is the collection of CMAQ, GASP, and AIRS time series described above. For our application, the model can also be written:

$$\mathbf{Y = Bo\ +\ B_1 \bullet X_1\ +\ B_2 \bullet X_2\ +\ B_3 \bullet X_3} \qquad \mathbf{(6)}$$

Where Y is one of the 85 AIRS sites in the response set (time series of 214 daily mean PM_{25}), X_1 consists of the other 18 AIRS sites (prediction set), X_2 = GASP, 3x3 neighborhood of the monitored grid (9 time series), and X_3 are time series in a CMAQ 3x3 neighborhood of the monitored (Y) grid (9 time series).

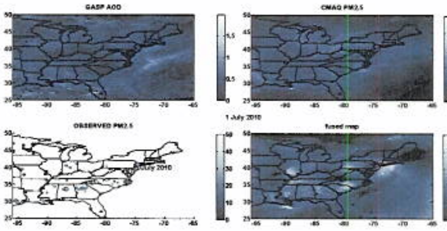

Figure 1. Fused map (36 predictors/ 5 latent variables)

The map is created from:

$$\mathbf{(unmonitored\ \ grid)\ \ =\ \ X\ \ \bullet\ \ (mapping)}$$
(7)

Results and Discussion

Figure 1 shows X (GASP, CMAQ and AIRS) and the fused map for 1 July 2006. 'Leave one out' cross-validation examples in Figures 2 and 3 show good agreement and poor agreement, respectively. Overall, cross-validation results summarized in Table 1 are not terribly impressive but are to be expected given the sparse network of observations that form the basis of the map.

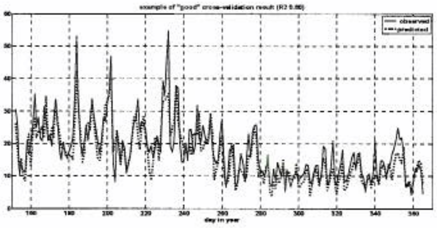

Figure 2. Example of 'good' cross-validation result (R^2 0.80)

Summary

PLSR can be used to reduce the dimension of the very large correlated set of explanatory variables represented by observations, CMAQ, and remotely sensed information. Fused maps produced by PLSR have relief derived from spatially dense CMAQ and satellite information. PLSR prediction performance in this study is limited by the sparse PM_{25} network utilized. Improvements might accrue from expansion of the set of predictor variables to other CMAQ layers, other remotely sensed parameters, and derived variables.

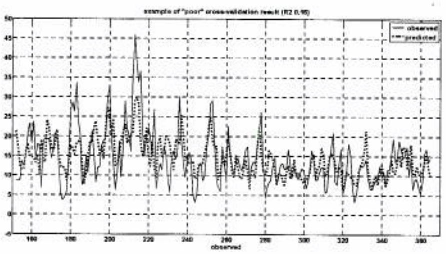

Figure 3. Example of 'poor' cross-validation result (R^2 0.16)

Table 1. Cross-validation results (5 latent vectors)

predictors	number of predictor grids	relative RMSE	relative mean bias	R^2
GASP	9 (3x3)	0.63	0.48	0.03
CMAQ	9 (3x3)	0.48	0.48	0.02
CMAQ	36 (3x3)	0.58	0.44	0.21
GASP	(3x3)			
AIRS	(18 sites)			

References

Smoliak, B.V., Wallace, J.M., Stoelinga, M.T., T.P. Mitchell. 2010. Application of partial least squares regression to the diagnosis of year-to-year variations in Pacific Northwest snowpack and Atlantic hurricanes. Geophysical Research Letters 37: L03801, doi:10.1029/2009GL041478

Trygg, J. and S. Wold. 2002. Orthogonal Projections to Latent Structures (O-PLS). Journal of Chemometrics 16: 119-128

Wold, S., Sjostrom, M., and L. Eriksson. 2001. PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems 58: 109–130

USEPA, 2011. Remote Sensing Information Gateway (RSIG), http://badger.epa.gov/rsig/ETRO final report: REanalysis of the TROpospheric chemical composition over the past 40 years – A long-term global modeling study of tropospheric chemistry, Reports on Earth System Science, edited by: Schulz, M. G., http://retro.enes.org/reports/RETRO_Final_Report.pdf, last access: 30 August 2010, Max Planck Institute for Meteorology, Hamburg, report no. 48/2007, ISSN 1614-1199, August 2007.

Question and Answer

Akula Venkatram: Does your method for combining model results with observations improve upon a purely statistical technique such as Kriging of observations.
P. Steven Porter: Kriged maps are too smooth for our purposes.

Jeremy Silver: How do you think the maps would look if the satellite data did not have areas missing due to cloudiness?
P. Steven Porter: Missing data are an important issue with GASP data. A significant effort goes into preprocessing, including filling-in missing values. The maps would undoubtedly look different with complete satellite information.