Application of statistical downscaling for ozone air quality in the US

Abstract

Tropospheric ozone is an air pollutant that can impact both human health and agriculture. Health impact studies typically require ozone concentration fields at high spatial resolution (km scale) but global numerical models of air quality are generally run at much coarser resolution (hundreds of km scale) due to computational cost. Statistical downscaling is a method of overcoming this computational constraint. It involves developing statistical relationships between coarse resolution predictor variables and high resolution predictand variables. Here we apply a regularized multitask regression method for surface level ozone in the United States using model output from a coarse resolution chemical transport model and fine-scale observed ozone at EPA monitoring sites.

1. Introduction

Tropospheric ozone is an air pollutant that can impact both human health and agriculture. Its production is controlled by chemical reactions between hydrogen oxide radicals (HO_x), nitrogen oxide radicals (NO_x), and volatile organic compounds (VOCs) (Seinfeld & Pandis, 2006). NO_x and VOCs have both natural and anthropogenic sources in the environment. Sources of NO_x include automobile exhuast, lightning, and soils. Sources of VOCs include forests and industrial emissions. Ozone is regulated under the national ambient air quality standards (NAAQS) as a criteria pollutant. The current regulation sets a standard of 75 parts per billion (ppb) maximum daily 8-hour average (MDA8) ozone (EPA), proposed to decrease to 70 ppb.

Although large suites of data are available for ozone, models are necessary because they are able to provide global spatial fields that are continuous in time. Global numerical models of air quality, called chemical transport models (CTMs), typically have spatial resolutions of hundreds of kilometers due to computational resource constraints as well as availability of high-resolution inputs. However, many applications of CTMs, such as health impact studies, typically require ozone concentration fields at high spatial resolution (km scale). In the air quality community, the common approach is dynamical downscaling, which involves using output from coarse-resolution CTMs as boundary conditions for running high-resolution CTMs over a limited region.

Statistical downscaling is a method of modeling fine-scale ozone without the computational constraints of numerical models. It involves developing statistical relationships between coarse resolution predictor variables and high resolution predictand variables. In the atmospheric sciences community, it was first applied to output from general circulation models (GCMs) for the purposes of studying climate-related variables (Wilby et al., 1998; Maurer & Hidalgo, 2008). Although many methods have been attempted, it is unclear whether any of them yield reliable results for operational use. Burger et al. (2011) compared five different statistical methods for climate models and found that all methods resulted in ”moderate” reliability for predicting precipitation events. There was no evidence that more complex neural network based methods produced better results than simple linear regression methods.

Statistical downscaling is less commonly used for air quality purposes. Alkuwari et al. (2013) developed a statistical method for ozone air quality using fitted empirical orthogonal functions as the predictor variables. Berrocal et al. (2014) use extreme value theory to derive a model for the distribution of fourth-highest MDA8 ozone based on coarse-resolution CTM output and observed ozone at AQS monitoring sites. None of these methods have been incorporated into operational modeling tools used by the EPA for air quality assessments.

Here we attempt a novel method of statistical downscaling of ozone using multi-task regression. With multi-task regression, rather than treating the regression at each monitoring site as an independent task, it seeks to improve performance by treating the tasks as related. Multi-task learning methods based on group Lasso are described in Roth & Fischer (2008); Kim & Xing (2010); Xu et al. (2010); Gong et al. (2012), while a regularized method similar to ridge regression is described in Evgeniou & Pontil (2004).

2. Methods

2.1. Data

The model is trained and validated with data from the EPA Air Quality System (AQS). AQS contains ozone data collected from 1173 monitors across the United States, along with meteorological information at each of the sites. Data is collected hourly and aggregated aggregated into hourly, daily, and monthly data sets. These data sets are publicly available for download at https://aqs.epa.gov/aqsweb/documents/data_mart_welcome.html. We remove data points flagged with possible measurement issues or associated with exceptional events, as well as any days with incomplete measurements.

Figure 1 shows the MDA8 ozone observed at each monitoring station on August 31, 2013. AQS monitors represent a variety of conditions, ranging from rural to urban, offering a fairly representative sample of surface air quality in the US. We see that coastal sites typically have lower ozone, while parts of the south and California has higher ozone.

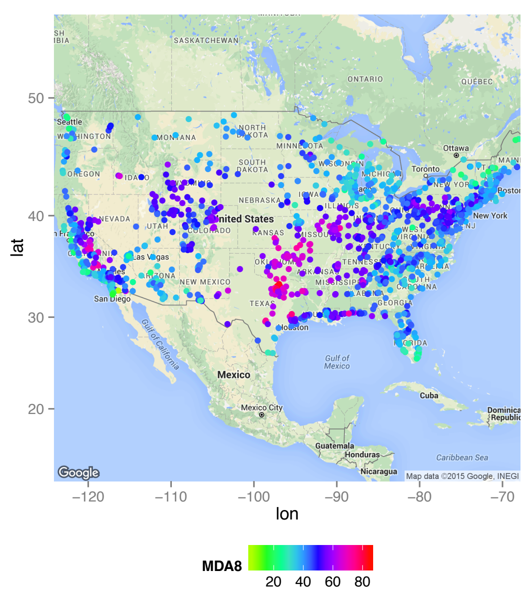

Figure 1. Observed MDA8 ozone at AQS monitors on August 31, 2013.

We use 9 months of data (Jan 2013 - Sep 2013) and randomly assign 90% of the data at each site to the training set and 10% to the validation set. This results in approximately 200 data points per monitoring site in the training set and approximatley 20 data points per monitoring site in the test set. Mean MDA8 ozone over this time period is 43 ppb, with a standard deviation of 12 ppb and values ranging from 2 ppb - 12 ppb.

2.2. Numerical model

The numerical model we use for providing coarse-scale ozone is the GEOS-Chem chemical transport model (CTM) (Bey et al., 2001), version 9-02, (http://wiki.seas.harvard.edu/geos-chem/index.php/GEOS-Chem_v9-02). GEOS-Chem has previously been used for a wide range of purposes ranging from studying regional air quality (Chen et al., 2009) to transcontinent pollution (Zhang et al., 2009). It is typically run at horizontal resolutions of 4^ox5^o ($\approx$ 400 km x 400 km) or 2^ox2.5^o ($\approx$ 200 km x 200 km) with 47 vertical levels. GEOS-Chem is driven by assimilated meteorology from the NASA GEOS-5 system (http://gmao.gsfc.nasa.gov/GEOS/), which is produced at a native resolution of 0.25^ox0.3125^o ($\approx$ 28 km x 28 km) and 72 vertical levels. GEOS-Chem models the chemical reactions between 196 chemical species (Parrella et al., 2012), using operator splitting to separately solve the equations of transport and chemistry. The timestep for the 4x5 simulation is 30 min for transport and 60 min for chemistry.

Figure 2 shows a snapshot of the modeled ozone from GEOS-Chem on August 31, 2013. As in the observations, there is regionally high ozone over the Southeast US, while coastal regions have lower ozone. However, unlike the observational sites, the coarse resolution model is unable to resolve the sometimes sharp gradients from one location to the next.

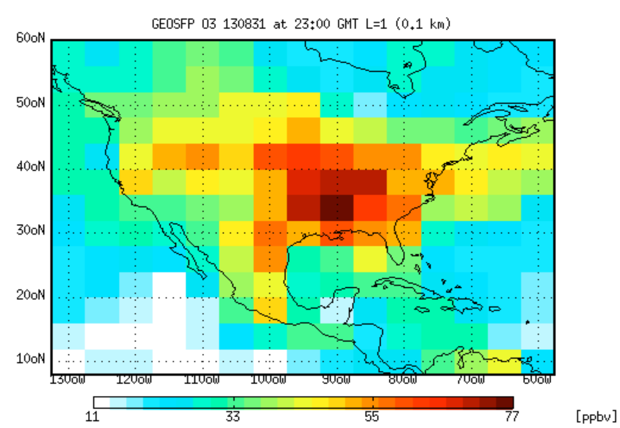

Figure 2. GEOS-Chem modeled ozone over North America.

In order to use GEOS-Chem output as predictor variables for regression, we sample the model hourly at all locations where there is an AQS monitoring site. Because the grid cells are so large, each grid cell may be sampled multiple times for different monitoring sites. This allows us to then compute MDA8 ozone from the model, which is directly comparable to the observed MDA8 ozone. The mean MDA8 ozone sampled at all AQS sites from GEOS-Chem is 43 ppb, with a standard deviation of 9 ppb and values ranging from 10 ppb - 80 ppb. From these aggregate statistics, we can see that GEOS-Chem produces less variability than the observations, but captures the region-wide mean fairly well.

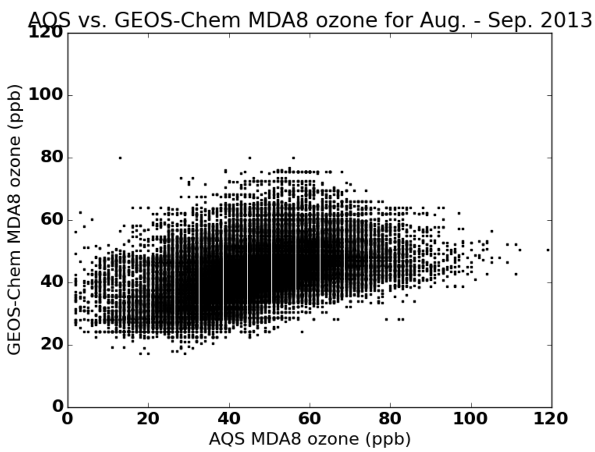

Figure 3. Scatter plot showing observed MDA8 ozone (x-axis) and coarse-scale modeled MDA8 ozone (y-axis). Only two months of data are plotted to reduce clutter.

Figure 3 shows a scatterplot of the observed and modeled MDA8 ozone for August and September 2013. As noted before, GEOS-Chem has much less variability compared to the observations. There is a large amount of scatter, indicating large errors at each individual site, despite the good agreement in the mean. We see that the model tends to underestimate the high end of the observations and overestimate the low end. The average RMSE of the GEOS-Chem model compared against the AQS measurements is 9.3 ppb averaged across all monitoring sites over the 9 month period. The largest RMSE is 24 ppb, while the lowest is 0.2 ppb.

2.3. Regularized multi-task regression

We apply the regularized multi-task learning method described in Evgeniou & Pontil (2004) as it is the most analogous to the ridge regression we use in the baseline. Here we describe this method in the context of our problem (Evgeniou & Pontil (2004) give the example of a classification problem while we are interested in regression), though we attempt to stick to the notation of (Evgeniou & Pontil, 2004) as close as possible.

For a particular measurement station t, the regression problem we attempt to solve can be described as the following set of linear equations

$$\mathbf{y}_t = \mathbf{w}_t \cdot \mathbf{X}_t + \epsilon$$

where $\mathbf{y}_t$ is the observed MDA8 ozone at the site, $\mathbf{X}_t$ is the matrix of predictor variables. In its simplest form, this corresponds to the coarse-scale ozone and a bias variable. $\mathbf{w_t}$ is the vector of regression weights that station. ϵ is a Gaussian random variable with mean zero and standard deviation σ.

Typically, for a set of T sites, each site is treated as independent of the others, with a separate $\mathbf{w_t}$ for each site. In multi-task regression, we can think of the regression weight vector for each site as

$$\mathbf{w}_t = \mathbf{w}_0 + \mathbf{v}_t$$

where the weight for each site, $\mathbf{w}_t$, is the sum of a common weight for all sites, $\mathbf{w}_0$, and $\mathbf{v}_t$, which measures how different each task is from the common weight. This amounts to solving the following minimization problem

$$\min_{\mathbf{w}_0, \mathbf{v}_t} \{J(\mathbf{w}_0, \mathbf{v}_t) = \tag{1}$$

$$\sum_{t=1}^{T} (\mathbf{y}_t - (\mathbf{w}_0 + \mathbf{v}_t)\mathbf{X}_t)^T \cdot (\mathbf{y}_t - (\mathbf{w}_0 + \mathbf{v}_t)\mathbf{X}_t) \tag{2}$$

$$+ \frac{\lambda_1}{T} \sum_{t=1}^{T} ||\mathbf{v}_t||^2 + \lambda_2 ||\mathbf{w}_0||^2\} \tag{3}$$

The last two terms are regularization terms for the individual and common regression weights, respectively, with regularization parameters λ_1 and λ_2. A large $\frac{\lambda_1}{\lambda_2}$ ratio will make the differet sites unrelated while a small ratio gives the same regression weights to all the sites.

The multi-task can be written in a more computationally expedient form as

$$\min_{\mathbf{w}_t} \{\sum_{t=1}^{T} (\mathbf{y}_t - \mathbf{w}_t \mathbf{X}_t)^T \cdot (\mathbf{y}_t - \mathbf{w}_t \mathbf{X}_t) + \tag{4}$$

$$\rho_1 \sum_{t=1}^{T} ||\mathbf{w}_t||^2 + \rho_2 \sum_{t=1}^{T} ||\mathbf{w}_t - \frac{1}{T} \sum_{s=1}^{T} \mathbf{w}_s||^2\} \tag{5}$$

where

$$\rho_1 = \frac{1}{T} \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2}$$

and

$$\rho_2 = \frac{1}{T} \frac{\lambda_1^2}{\lambda_1 + \lambda_2}$$

This method can be extended to non-linear regression using Reproducing Kernel Hilbert Spaces.

3. Results

3.1. Baseline

For the baseline, we performed linear regression using the coarse-scale GEOS-Chem ozone as the predictor variable along with a bias variable. We use ridge regression with regularization parameter $\lambda = 0.5$. This was chosen by testing a few different λ values and choosing the one that produced regression weights with the loweset RMSEs. Each site was treated individually. Figure 5 shows the RMSE between the down-scaled ozone and the observed ozone at each site for the test set.

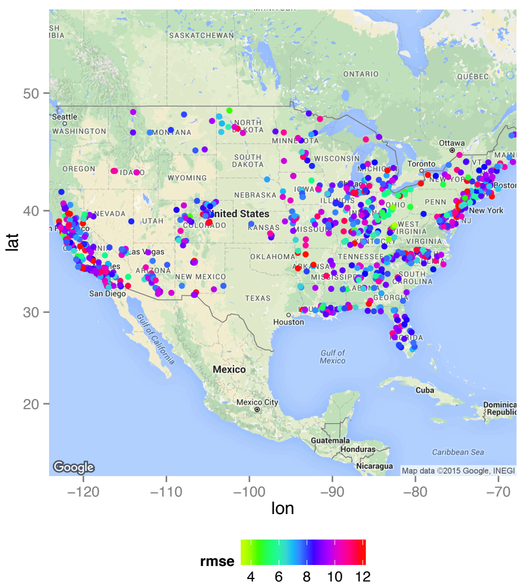

Figure 4. RMSE of downscaled ozone at each site for the test set.

Average RMSE on the test set is 8.8 ppb, which is only a minor improvement from not performing regression at all. The RMSE for all sites range from 3.4 ppb to 18 ppb. In terms of spatial patterns, there is no region that stands out, indicating smal-scale variabilities are more important than large-scale ones in this case.

3.2. Multi-task regression

We perform multi-task regression as described in section 2. We use parameters $\rho_1 = \rho_2 = 0.0002$ to stay consistent with the ridge regression regularization parameter of the baseline. We choose L-BFGS-B optimization from the scipy.optimize.minimize package.

The following figure shows the RMSE at each monitoring site in the test set.

The average RMSE on the test set is 8.7 ppb, which is only

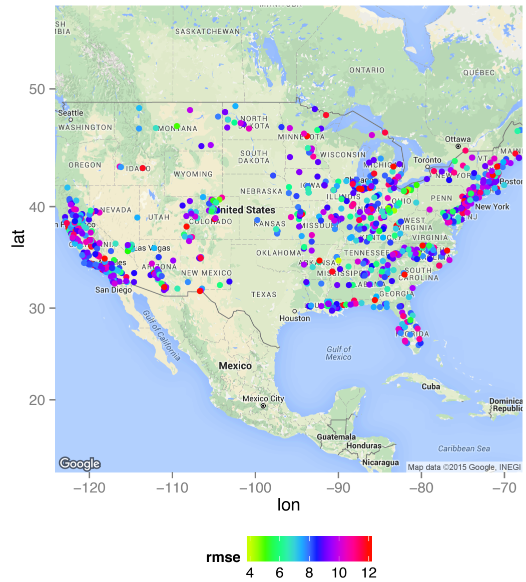

Figure 5. RMSE of downscaled ozone at each site for the test set using multitask regression.

0.1 ppb better than the baseline. Compared to the baseline, we see that there is more improvement in the inland regions of the US while the coastal regions do not visibly improve from the baseline.

4. Discussion

Multi-task regression failed to significantly improve upon the results of the baseline, which itself is not much better than directly using the results of the numerical model. There are several possible reasons for this:

- We do not include a temporal component in the regression, so we are essentially assume that the model gets the seasonal trend of ozone correct. If, for example, the model does a good job of simulating ozone in January but has high biases in August, then a regression weight that does not take this temporal component into account will lead to a correction in January that worsens the model output.

- The output of the numerical model is not a good predictor of observed ozone at monitoring sites. Perhaps the premise of using coarse-resolution output from a numerical output to predict fine-scale ozone is incorrect. Given the amount of variability between monitoring sites that are within close proximity to each other, the average ozone over a 400 km x 400 km grid box actually offers very little information about what the ozone value should be at a particular site. This is especially true if a coarse grid cell includes both urban and rural settings. In fact, simply predicting ozone using observed meteorlogical variables yields lower RMSEs than using the modeled coarse-scale ozone.

- AQS monitoring stations are too heavily biased towards urban and coastal sites, which are known to be difficult for the numerical model to capture.

- It was incorrect to treat each task as a monitoring site. Rather, we could first cluster sites then treat each cluster as a task.

Additionally, using optimization to find the regression weights was much slower than ridge regression. The alogrithm I chose to code up was relatively simple, and other algorithms could very well be more computationally efficient. Given these disadvantages, it does not seem to be a good method for this application.

Even though this project did not succeed in developing a robust method for downscaling ozone, it was useful to learn about statistical methods used in the field and now I have a better understanding for why they aren't applied more frequently.

5. Extensions

We test the hypothesis that temporal variations are not correctly captured in GEOS-Chem by adding an additional predictor variable for the month the data point was taken. I modified the code to do this, and attempted to run the optimization for the regression again, but did not have enough time to reach convergence. However, even by the 30th iteration, the we reached about the same RMSE as the regular multi-task regression by the time it converged, so it is reasonable to believe that this method would at least slightly improve the results of the regression.

6. Future work

Given the disappointing results of the previous sections, there is clearly much room for improvement. Here we outline several ideas that may produce improved regression results.

- Extend the multi-task linear regression method to multi-task non-linear regression through the use of Reproducing Kernel Hilbert Spaces (RKHS). This idea was briefly discussed in Evgeniou & Pontil (2004).

- Use PCA to identify patterns that may be useful in feature selection. Alkuwari et al. (2013) used PCA to generate the predictor variables that went into their regression.

- Inclusion of additional features, such as meteorological variables like wind and temperature. Ozone is known to be correlated with temperature, so including temperature will likely improve the accuracy of the prediction. Including additional chemical species, such as NO_x or VOCs, which are ozone precursors, may also improve the regression, but these variables would come from the coarse-resolution model and may not necessarily help.

- Attempt a more efficient multi-task regression algorithm. Kim & Xing (2010) describes a tree-guided group lasso method for multi-task regression that may potentially speed up the computation.

If I had more time to work on this project, I would spend more time on exploratory data analysis to determine what the appropriate predictor variables should be. Due to the time constraints, I implemented a machine learning algorithm that I read about without fully considering what the best method for this particular problem would be.

7. Code

The code is available at: https://github.com/kyu0110/CS281

References

Alkuwari, F. A., Guillas, S., and Wang, Y. Statistical downscaling of an air quality model using fitted empirical orthogonal functions. *Atmos. Env.*, 81:1–10, 2013.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. Assessing exceedance of ozone standards: a space-time downscaler for fourth highest ozone concentrations. *Environmetrics*, 25:279–291, 2014.

Bey, I., Jacob, D. J., Yantosca, R. M., Logan, J. A., Field, B., Fiore, A. M., Li, Q., Liu, H., Mickley, L. J., and Schultz, M. Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. *J. Geophys. Res.*, 106:23073–23096, 2001.

Burger, G., Murdock, T. Q., Werner, A. T., and Sobie, S. R. Downscaling extremes – an interomparison of multiple statistical methods for present climate. *Journal of Climate*, 25:4367–4385, 2011.

Chen, D., Wang, Y. X., McElroy, M. B., He, K., Yantosca, R. M., and Sager, P. Le. Regional co pollution in china simulated by the high-resolution nested-grid geos-chem model. *Atmos. Chem. Phys.*, 11:3825 – 3839, 2009.

EPA. National ambient air quality standards (NAAQS). URL http://www3.epa.gov/ttn/naaqs/criteria.html#3.

Evgeniou, T. and Pontil, M. Regularized multi-task learning. In *Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2004)*, pp. 109–117, Seattle, WA, 2004.

Gong, P., Ye, J., and Zhang, C. Robust multi-task feature learning. In *Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2012)*, pp. 895–903, Beijing, China, 2012.

Kim, S. and Xing, E. P. Tree-guided group lasson for multi-task regression with structured sparsity. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, Haifa, Israel, 2010.

Maurer, E. P. and Hidalgo, H. G. Utility of daily vs. monthly large-scale climate data: an intercomparison of two statistical downscaling methods. *Hydrol. Earth Syst. Sci.*, 12:551–563, 2008.

Parrella, J. P., Jacob, D. J., Liang, Q., Zhang, Y., Mickley, L. J., Miller, B., Evans, M. J., Yang, X., Pyle, J. A., Theys, N., and Roozendael, M. Van. Tropospheric bromine chemistry: implications for present and pre-industrial ozone and mercury. *Atmos. Chem. Phys.*, 12: 1823–1832, 2012.

Roth, V. and Fischer, B. The group-lasso for generalized linear models: Uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, pp. 848–855, Helsinki, Finland, 2008.

Seinfeld, J. H. and Pandis, S. N. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. John Wiley and Sons, Inc., Hoboken, NJ, 2006.

Wilby, R. L., Wigley, T. M. L., Conway, D., Jones, P. D., Hewitson, B. C., Main, J., and Wilks, D. S. Statistical downscaling of general circulation model output: A comparison of methods. *Water Resources Research*, 34 (11):2995–3008, 1998.

Xu, Z., Jin, R., Yang, H., King, I., and Lyu, M. R. Simple and efficient multiple kernel learning by group lasso. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, Haifa, Israel, 2010.

Zhang, L., Jacob, D. J., Kopacz, M., Henze, D. K., and Jaffe, D. A. Intercontinental source attribution of ozone pollution at western us sites using an adjoint method. *Geophys. Res. Lett.*, 36:L11810, 2009.