Geostatistical Analysis and Mapping of Ozone in California, August 2015

Bryan Cole

April 2016

Contents

1 ABSTRACT

Ground-level ozone is a harmful air pollutant with negative consequences to human health, sensitive vegetation, and various ecosystems. This research aims to understand and predict the distribution of ozone levels in California during August 2015 using spatial statistics. One of the key components of the research is integrating elevation data, population density data, and meteorology data in hopes of enhancing the analysis. Upon completion of data integration and data pre-processing, exploratory spatial data analysis is conducted through the use of descriptive statistics, empirical variograms, etc. A geostatistical model is presented, and the parameters of the model are estimated using maximum-likelihood estimation for various mean functions and covariance functions. Cross-validation is used to compare competing models, and Kriging is performed with the model that has the lowest RMSE to make predictions and evaluate their uncertainties. Lastly, a discussion of the results and recommendations to enhance the research going forward are expressed.

2 INTRODUCTION

2.1 Motivation and Problem Statement

We live in a fast-paced and heavily-industrialized world. As a result, it's no secret that we have introduced several problems to the environment. One of these is air pollution. While there are many pollutant groups that together make up pollution, this research deals with ground-level ozone (O_3). Ground-level ozone (which I will refer to simply as "ozone") is created by chemical reactions between oxides of nitrogen (NO_x) and volatile organic compounds (VOC) in the presence of sunlight. The Environmental Protection Agency (EPA) suggests that emissions from industrial facilities and electric utilities, motor vehicle exhaust, gasoline vapors, and chemical solvents are the major sources of NO_x and VOC. Ozone can instigate a large number of health problems such as respiratory symptoms, decrements in lung function, and inflammation of airways. As a result, ozone is associated with an increase of hospital admissions, asthma attacks, and daily mortality. What's more is that ozone can even negatively impact vegetation and various ecosystems.

Understanding the spatial distribution of ozone and monitoring it with geostatistical approaches is challenging but relevant to many people. In particular, it allows for the prediction of ozone concentration over some area which can be useful for many applications. This prediction can be further improved by integrating additional data sources such as elevation, population density, and meteorology. Therefore, the goal of this analysis is to explore the spatial patterns of ozone concentration integrated with these additional data sources.

2.2 Research Questions

1. What is the spatial distribution of ozone over the study area?
2. What is the spatial relationship between ozone, elevation, population density, and several meteorology covariates?

3. What are the predicted values of ozone concentration at unobserved locations in the study area, and what are the uncertainties of these predictions?

4. How much does the incorporation of the additional data sources improve predictions?

3 STUDY AREA AND DATA DESCRIPTION

3.1 Study Area

The study area for this analysis spans between Northern and Southern California. In particular, the northern most monitoring stations for ozone are near San Francisco and Sacramento while the southern most stations are just northwest of Los Angeles. There is a diverse range of terrain including coastal, agricultural, mountain ranges, deserts, urban, and national forests. Figure 1 contains a map of the study area, monitoring locations, and terrain. The monitoring stations tend to be in densely-populated areas.

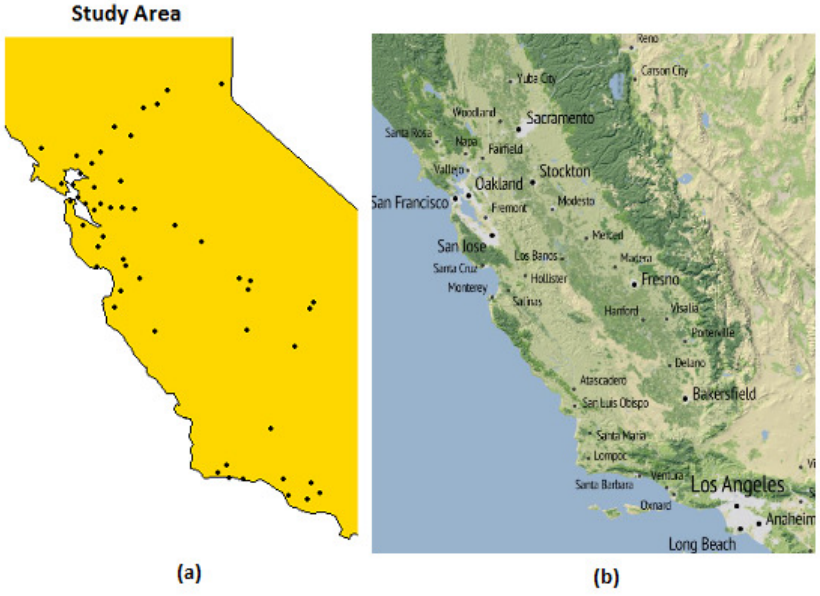

Figure 1: (a) Study Area and Monitoring Locations
(b) Study Area Terrain and Major Cities

3.2 Data Description

3.2.1 Ozone monitoring

Ozone data for the analysis come from the EPA at the monitor level in the form of an 8-hour average due to the fact that the national ambient air quality standards (NAAQS) for ozone are in this form. That is, for every monitoring station an 8-hour average is calculated for every clock hour. Only complete data (e.g., with 6 or more valid hourly samples in the 8 hour block, or 75% completeness) is included. The data are for the year 2015. The units for the ozone measurements are in parts per million (ppm). The NAAQS for ozone is that the annual fourth-highest daily maximum 8 hour average concentration, averaged over 3 years can not exceed 0.070 ppm. Each observation is associated with a latitude and longitude, date, etc.

- Dataset download link: http://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/8hour_44201_2015.zip
- Codebook: http://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/FileFormats.html#_8_hour_average_files

3.2.2 Elevation data

Elevation data for the analysis come from The Google Maps Elevation API, which provides elevation data for all locations on the surface of the earth. In those cases where Google does not possess exact elevation measurements at the precise location you request, the service will interpolate and return an averaged value using the four nearest locations. If the reader is interested in the API documentation, please visit https://developers.google.com/maps/documentation/elevation/intro.

3.2.3 Population density data

Population density (persons per square mile) data come from the 2010 Census conducted by the U.S. Census Bureau at a "places" level (incorporated cities and Census Designated Places). The "2010 Census Summary File 1" dataset can be downloaded directly from the American Fact Finder's website. The location

of observations in the data is given by a character variable containing the city or town name.

3.2.4 Meteorology data

Meteorology data come from the Weathersource.com API. Each query returns a collection of weather history data for a given postal code. Specifically, average cloud cover, average dew point, total precipitation, average relative humidity, average surface pressure, average specific humidity, minimum temperature, average temperature, max temperature, and average wind speed are returned. Unfortunately, the API restrictions limit users to 10 requests a minute, and 1000 per day. If the reader is interested in the API documentation, please visit https://developer.weathersource.com/documentation/resources/get-history_by_postal_code/.

4 METHODOLOGY

4.1 Data Preprocessing

4.1.1 Ozone data

First, the ozone dataset was subset only to observations in the study area. Next, a decision was made to only use the 8-hour average from 8a.m. to 4p.m. This seemed reasonable as it accounts for most of the sunlight during each day and simplified the analysis. Finally, the data was subset to observations in August for reasons mentioned in section 4.1.4.

4.1.2 Elevation data

The Google Maps Elevation API allows location requests in the form of a latitude and longitude pair. Thus, for every unique pair of latitudes and longitudes in the ozone data, a request was made to the API and the elevation value was extracted from the returned JSON object.

4.1.3 Population density data

To obtain population density data at every monitoring station, first a script was written to get the city or town name at each monitoring station location. At four percent of locations, a town or city name could not be identified and thus observations with these locations were deleted. For each remaining location, the observation in the population density dataset that had the corresponding town/city name was found and the population density value for that observation was integrated into the valid locations dataset. There were two special cases when performing this integration. Some observations had a city name that appeared in more than one row of the population density data. In this situation, the observation with the highest total population count was used. The second special case was for observations whose city name didn't appear at all in the population density data. This was a very small percentage and thus observations with these locations were deleted.

4.1.4 Meteorology data

The meteorology API restrictions (10 requests per minute, 1000 requests per day) made an analysis for all of 2015 infeasible in the time frame of this research. Thus, the scope was reduced down to August as this is a hot summer month where significant levels of ozone should be occurring. Since the meteorology API accepts request for locations in terms of their zip code, a script was written to obtain the zip code of each unique ozone monitoring station. Then, the meteorology data was integrated as time permitted. Lastly, near-zero variance meteorology predictors were removed (precipitation) as were highly-correlated variables. Specific humidity was used instead of relative humidity and dew point average. Average temperature was used instead of minimum or maximum temperature.

4.1.5 Data integration for prediction locations

For making predictions, a grid of 265 latitude and longitude pairs was created in the study area. The ideal number of prediction locations would have been in the thousands, but due to the meteorology API restrictions the number 265 was chosen as a trade-off for the number of days the data would be able to be collected in time. Then, the exact same data preprocessing/integration scripts and methods were used on these new locations.

4.2 Exploratory Data Analysis

4.2.1 General

EDA and quantification of the data were handled by quantile summaries, histograms, box plots, scatter plots, etc. Geostatistical models assume a normal distribution of the response variable at each monitoring station, and thus the ozone values for each station for all 31 days were collected separately and the Shapiro-Wilk normality test was performed on each. Approximately 44% of the stations didn't pass the normality test, and thus a new variable containing the log transform of the ozone values was created for each dataset.

4.2.2 Variograms

Variograms are an exploratory tool used to describe the degree of spatial dependence and/or estimate the covariance parameters in a geostatistical model. Formally, the theoretical variogram is defined as:

$$2\Upsilon(h) = E[(Y(x) - Y(x+h))^2]$$

Where x is a vector $\in \Re^2$ of spatial locations, $Y(x)$ is some quantity (in this case 8-hour ozone average in parts per million) available at all spatial locations, and h is the lag distance between the pairs of locations. Υ itself is the semivariogram. The empirical variogram (from the observed data) is an unbiased estimator of the theoretical variogram and is defined as:

$$\gamma(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} |z_i - z_j|^2$$

where z_i for i = 1,...,k is the response value (ozone) at locations $x_i, ..., x_k$, $N(h)$ denotes the set of pairs of observations i, j such that $|x_i - x_j| = h$, and $|N(h)|$ is the number of pairs in the set.

In this analysis, the empirical variogram for each day of August was calculated to analyze spatial dependence (see Figure 5).

4.3 Geostatistical Model

The geostatistical model can be written as:

$$Y(X) = \mu(X) + Z(X)$$

- X = the spatial locations ($x_i \in \Re^2$)
  - In particular, x_i represents (lon,lat) pairs for $i = 1, ..., n$
- $Y(X)$ = Ozone value at any spatial location in the data (a random quantity)
- $\mu(X) = \beta_0 + \beta_1 a_1(X) + \beta_2 a_2(X) + \beta_3 a_3(X) + \beta_4 a_4(X) + \beta_5 a_5(X) + \beta_6 a_6(X) + \beta_7 a_7(X) + \beta_8 a_8(X)$ is the non-random linear mean function of longitude(a_1), elevation (a_2), population density (a_3), average cloud cover (a_4), average specific pressure (a_5), average specific humidity (a_6), average temperature (a_7), and average wind speed (a_8) at any spatial location in the data. The β_i terms are constant coefficients to be estimated
- $Z(X)$ = Random spatially correlated deviation from the mean
  - $E(Z(X)) = 0$, $var(Z(X)) = \sigma^2$ (partial sill)
  - cov(Z(x),Z(x+h)) = c(h) is the generic stationary covariance function. This is a symmetric and positive definite function, and h represents the lag distance between a particular pair of points

– Parameters of the covariance function are the partial sill σ^2 and range parameter α. Additional possible parameters are the nugget τ^2, smoothness parameter ν and the anisotropy parameters ψ_R and ψ_A

4.3.1 Max-likelihood estimation

Maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model. In particular, for a fixed set of data, MLE works by selecting values for the model parameters that maximize the likelihood function. Thus, MLE is essentially maximizing the agreement of the model with the empirical data.

For the geostatistical model used in this research, it's assumed the $Y(X)$ follows a multivariate normal distribution. Specifically, $Y \sim N(A\vec{\beta}, \Sigma)$ where A denotes the design matrix for the mean function and Σ denotes the covariance matrix. Then, the log-likelihood of the data $\vec{y} = (y(x_1), ..., y(x_n))$ given the parameters $\vec{\omega} = (\vec{\beta}, \vec{\theta})^T$ is:

$$l(\vec{\beta}, \vec{\theta}) = -\frac{n}{2}log(2\pi) - \frac{1}{2}logdet(\Sigma) - \frac{1}{2}(\vec{y} - A\vec{\beta})^T\Sigma^{-1}(\vec{y} - A\vec{\beta})$$

Taking the derivative of the log-likelihood with respect to $\vec{\beta}$, setting it to zero, and solving for $\vec{\beta}$ yields the MLE and GLS estimator of $\vec{\beta}$:

$$\hat{\vec{\beta}}(\vec{\theta}) = (A^T\Sigma^{-1}A)^{-1}A^T\Sigma^{-1}\vec{y}$$

One can then plug this estimate into the log-likelihood expression, and maximize it with respect to $\vec{\theta}$ to get the MLE estimate of $\vec{\theta}$. This is typically done using numerical methods, as it's usually not possible to get a closed form solution.

In this research, the parameters were estimated using MLE separately for each day of August. The covariance functions used were the exponential

$$c(h) = \sigma^2 exp(-\frac{||h||}{\alpha})$$

as well as the Matern

$$c(h) = \sigma^2 \frac{1}{2^{\nu-1}\Gamma(\nu)}(||h||)^\nu \ K_\nu(||h||)$$

where K_ν is the modified Bessel function of the second kind. For each covariance function, MLE was used for various combinations of the additional covariance parameters τ^2, ψ_R, ψ_A, and ν (for the Matern). These combinations were also implemented separately for the raw ozone values and log-ozone values.

Lastly, since MLE was applied separately for each day, $\hat{\vec{\omega}} = (\hat{\vec{\beta}}, \hat{\vec{\theta}})^T$ is different for each day. For prediction of ozone values later on at unobserved locations, a constant $\hat{\vec{\omega}}$ is desired across all days. Denote this new parameter estimate as $\tilde{\vec{\omega}}$. Assuming days to be independent, the desired $\tilde{\vec{\omega}}$ can be "theoretically" found by setting

$$\tilde{\vec{\omega}} = max_{\vec{\omega}} \ \sum_{i=1}^{j} l(\vec{\omega}; Y_i)$$

where j is the number of days, and $l(\vec{\omega}; Y_i)$ is the log-likelihood value using the theoretical parameter vector $\vec{\omega}$ and ozone data Y_i for the i^{th} day. Since $\vec{\omega}$ is unknown, the MLE estimates from the previous techniques mentioned above are used to set

$$\tilde{\vec{\omega}} = argmax_{\hat{\vec{\omega}}_j} \ \sum_{i=1}^{j} l(\hat{\vec{\omega}}_j; Y_i)$$

where the only difference now is $\hat{\vec{\omega}}_j$ is the MLE estimate of $\vec{\omega}$ for the j^{th} day. Thus, j sums are calculated and whichever $\hat{\vec{\omega}}_j$ produces the largest sum gets set to $\tilde{\vec{\omega}}$ and is used across all days for predictions later on.

4.3.2 Inference

This analysis uses 95% confidence intervals to deduce properties about the parameters in the mean function ($\vec{\beta}$). $\hat{\beta}$ is both the Generalized Least Squares estimate and Max-Likelihood estimate of $\vec{\beta}$. It turns out that $\hat{\beta}$ is an unbiased estimator. That is, $E[\hat{\beta}] = \vec{\beta}$. Additionally, $\text{cov}(\hat{\beta}) = \sigma^2(A^T\Sigma^{-1}A)^{-1}$ and $Var(\hat{\beta}_j)$ is the $(j+1, j+1)$ entry of $\text{cov}(\hat{\beta})$. A 95% confidence interval for β_j then is:

$$[\hat{\beta}_j - 1.96 * sqrt(Var(\hat{\beta}_j)) \ , \ \hat{\beta}_j + 1.96 * sqrt(Var(\hat{\beta}_j))]$$

where the plug-in estimate used for σ^2 is $\hat{\sigma^2} = \frac{(\vec{y} - A\hat{\vec{\beta}})^T \Sigma^{-1} (\vec{y} - A\hat{\vec{\beta}})}{n}$

Confidence intervals for the covariance parameters (as well as for β) can also be obtained by maximizing the log-likelihood function, inverting the negative of the resulting Hessian matrix, and obtaining the standard errors by taking the square roots of the diagonal of it.

4.4 Kriging / Prediction Mapping

Kriging is an interpolation technique where the interpolated values are modeled by a Guassian Process governed by covariances. Essentially, Kriging makes predictions for the value of a function at a given point by using a weighted average of the known values of that function in the neighborhood of that point. Universal Kriging applies when a general polynomial trend (mean) model is assumed. This is appropriate for this analysis, due to the fact the mean function in the geostatistical model is a linear mean function of several covariates. If the reader is interested in the mathematical details of Kriging, they are directed elsewhere. Note that, for each competing model, while Kriging is performed separately for each day in August to obtain predictions for each day, the estimated parameters used for the Kriging are constant across all days ($\tilde{\vec{\omega}}$ mentioned at the end of section 4.3.1). That is, the data and covariate values change from day to day when performing Kriging, but the estimated parameters in the model do not.

4.5 Cross Validation

Cross Validation (CV) is primarily a method for measuring the predictive performance of a statistical model, and thus is used in this analysis to determine which model is the "best" fit among competing models. The CV strategy used was to leave out approximately 20% of the observed data (call this the "test" data), and repeat the model fitting procedures described in section 4.3.1 on the

remaining "train" data. Then, the same Kriging procedures described in section 4.4 were used to predict the test data ozone values. It's then possible to compare competing models using the Root Mean Squared Error (RMSE):

$$\text{RMSE} = sqrt(\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n})$$

where $\hat{y}_t$ represents the t^{th} predicted ozone value (for $t = 1, ..., n$), and y_t is the t^{th} observed ozone value. Thus, RMSE represents the sample standard deviation of the differences between predicted values and observed values and a model with a lower RMSE can be viewed as a better model or "fit" for the data.

One final RMSE is calculated to answer research question 4. This is done by fitting a model whose mean function is a linear trend of only longitude since all other models had their mean function in terms of longitude plus the integrated covariates.

4.6 Software Used

The R programming language was used for the duration of the analysis. In particular, the primary external packages used were "dplyr" for data preprocessing/integration, "jsonlite" for making API requests, "fields" for the empirical variograms and prediction maps, "geoR" for MLE and Kriging, and "ggmap"/"maps" for various plots.

5 RESULTS AND ANALYSIS

5.1 Data Preprocessing

The result of the data preprocessing/integration was a separate dataset for each day of August containing the latitude/longitude, elevation, population density, average cloud cover, average specific pressure, average specific humidity, average temperature, average wind speed, and the 8-hour ozone average at every monitoring station. The number of monitoring stations for each is 52.

Figure 2 contains a plot of the distribution of elevation values at the ozone monitoring stations, which are assumed to be constant across all days.

Figure 2: Elevation in Meters at Ozone Monitoring Stations

Figure 3 contains a plot of the distribution of population density values at the ozone monitoring stations, which are assumed to be constant across all days.

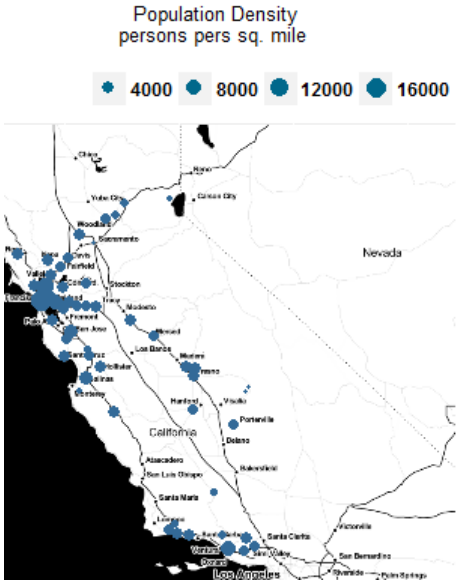

Figure 3: Population Density at Ozone Monitoring Stations

For the prediction locations, the meteorology API restrictions allowed the same data integration for only the first 16 days of August and thus predictions will only be made for these days. Figure 4 contains a plot of the prediction locations. 45 out of 265 of the prediction locations had to be deleted due to the fact that either population density or meteorology data was unobtainable there. These occurred primarily in the mountainous regions west of Central Valley.

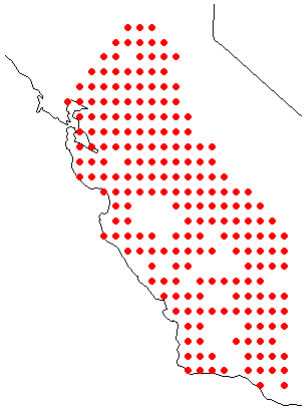

Figure 4: Prediction Locations

5.2 Exploratory Data Analysis

Longitude showed high correlation with ozone values and this is why it was included as a covariate in the linear mean function during modelling.

Figure 5 contains the empirical variogram for the log-ozone values for each day of August.

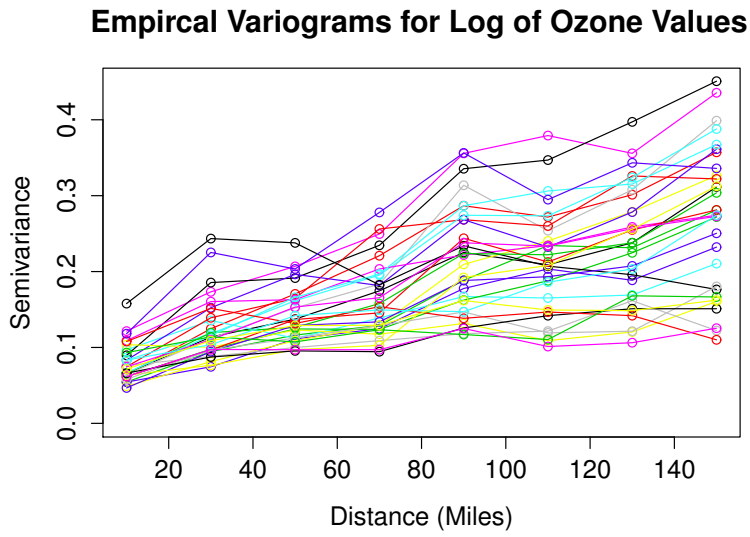

Figure 5: Empirical Variograms for August 2015

Indeed, the data demonstrate spatial dependence across August as the semivariance between pairs of locations grows on average as the lag distance h grows.

5.3 Modelling

5.3.1 Inference

The resulting 95% confidence intervals using the GLS approach for the geostatistical model with an exponential covariance function with nugget, linear mean function of the 8 covariates, and raw (untransformed) ozone response are:

- β_0 (intercept): [0.2541845 , 1.2528432]

- β_1 (longitude): [0.00135961 , 0.00956880]
- β_2 (elevation): [-1.257018e-05 , -3.859610e-06]
- β_3 (population density): [-5.465256e-07 , 3.682445e-07]
- β_4 (avg. cloud cover): [-1.683967e-04 , 9.008621e-05]
- β_5 (avg. specific pressure): [-1.723298e-04 , -3.486774e-05]
- β_6 (avg. specific humidity): [-0.0035479630 , 0.0004307692]
- β_7 (avg. temp): [0.0003929067 , 0.0013522826]
- β_8 (avg. wind speed): [-0.0004349296 , 0.0011848533]

The Hessian approach for constructing confidence intervals of the covariance parameters (and possibly β) was not implemented due to time constraints.

5.3.2 Cross validation

The lowest RMSE's for the competing models were:

- 0.01340138 for the model with an exponential covariance function with nugget, linear mean function of the 8 covariates, and raw (untransformed) ozone response
- 0.0129951 for a model with an exponential covariance function with nugget, linear mean function of longitude and temperature, and log-ozone response
- 0.01166284 for the model with an exponential covariance function with nugget, only longitude as a covariate in the linear mean function, and raw (untransformed) ozone response

Due to these results, the modeling process was iteratively repeated but this time the models were simplified by dropping integrated covariates in the linear mean function whose confidence intervals contained 0 in section 5.3.1. That is, population density, average cloud cover, average specific humidity, and average wind speed were dropped and the mean functions became a linear trend

of only longitude, elevation, average specific pressure, and average temperature. The parameters were re-estimated using the same techniques, and cross validation was repeated on these simplified models. This resulted in new RMSE's of 0.0131347 for the raw (untransformed) ozone response and 0.01380162 for the log-ozone response. Parameters were also re-estimated using only temperature in the mean function but no significant improvements resulted. Thus, the result of cross validation was that the model with only longitude as a covariate in the linear mean function had the best predictive performance.

5.3.3 Parameter estimates

The MLE parameter estimates for August 12th for the model with only longitude as a covariate in the linear mean function, exponential covariance function with nugget, and raw (untransformed) ozone response values are reported next. The reason for reporting August 12th's MLE parameter estimates is because they are the ones used across all days when performing Kriging (see the end of section 4.3.1.).

- β_0 (intercept): 1.40252766
- β_1 (longitude): 0.01132922
- τ^2 (nugget): 0.000008338151
- σ^2 (partial sill): 0.0000679366
- α (range): 3.0861453870

and the new GLS-based 95% confidence intervals for this β_0 and β_1 are:

- β_0 (intercept): [0.8666955 , 1.9383599]
- β_1 (longitude): [0.006882404 , 0.015776037]

5.4 Prediction maps

The following prediction map in Figure 6 is a result of Kriging using August 12th's MLE parameter estimates across all days, and the model with only longi-

tude as a covariate in the linear mean function, exponential covariance function with nugget, and raw (untransformed) ozone response values. Note that while it may be tempting to predict for all 31 days of August at all unobserved locations using this model, the predictions are only made for the first 16 days and the same unobserved locations to ensure consistency and validity.

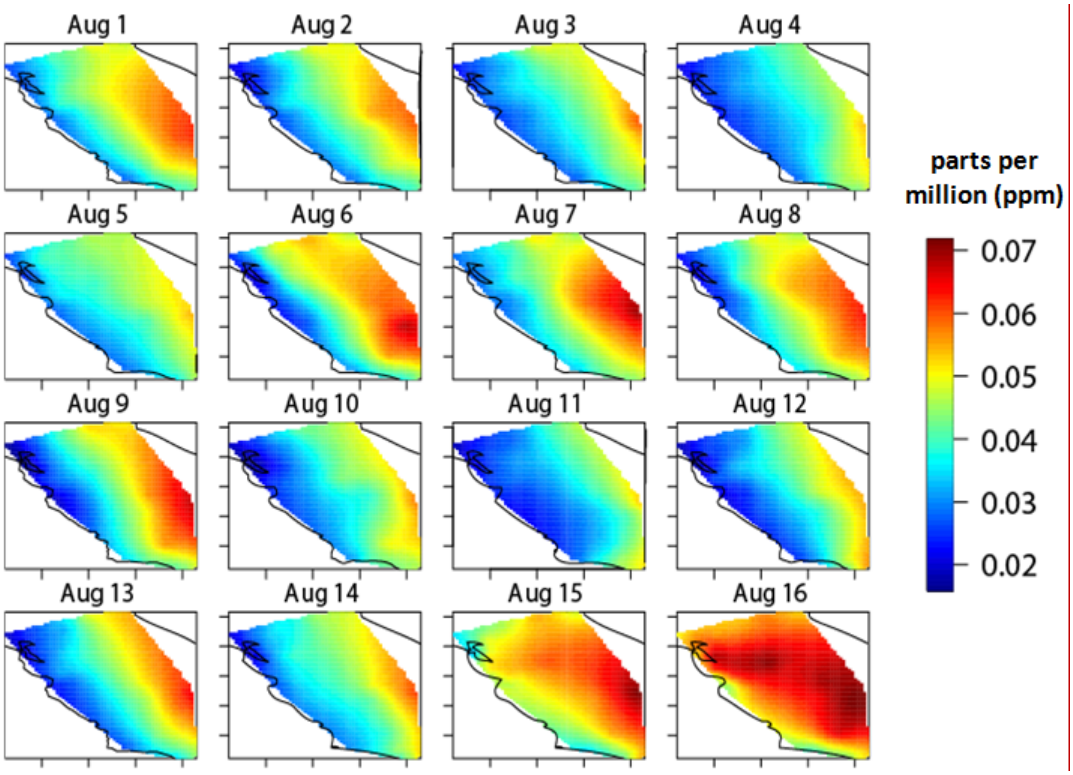

Figure 6: Ozone level predictions across the entire study area

In Figure 7, the same is produced for the standard deviations of the predictions to fully answer research question 3. Note that the map is only shown for a single day. This is due to the fact that the model being used provides the same uncertainty map for each day and it would be redundant to show the same map for every single day.

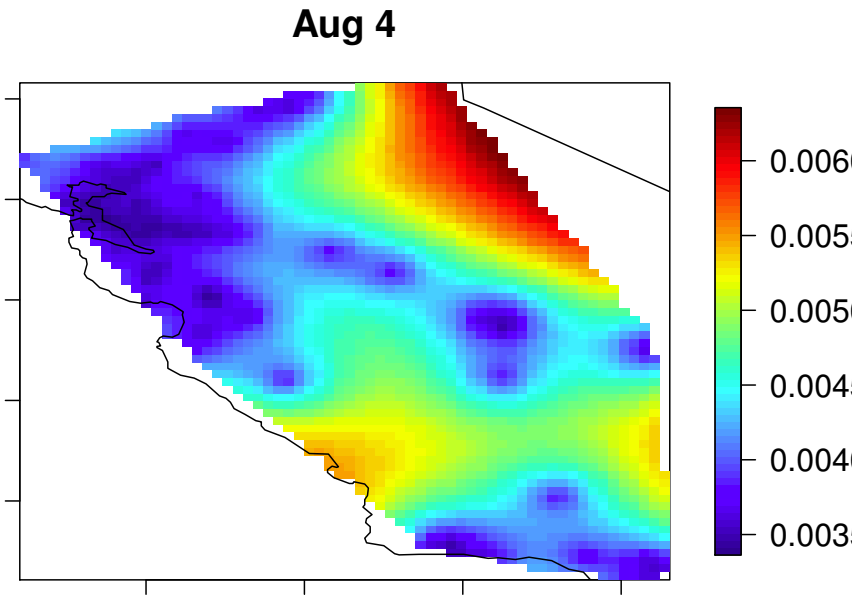

Figure 7: Standard Deviation Map for Ozone Predictions

6 DISCUSSION

6.1 Discussion

Several interesting results were found from the analysis. The RMSE of the models with integrated data covariates being higher could be attributed to several reasons such as not enough data, or too much noise being added from them. Even then, simplified models of the integrated covariates didn't perform nearly as well for prediction. Thus, the results suggest a "spatial-only model" is most appropriate for predicting ozone at unobserved locations, and that the integrated covariates don't improve prediction. It was also surprising that, more times than not, models using the raw (untransformed) ozone values predicted better than a log-ozone response. I believe this can be attributed to the MLE strategy of using a single day's MLE estimates across all days when Kriging.

The spatial distribution of ozone is readily apparent from the prediction map. Ozone levels tend to be low near the coast, mild near Central Valley, and high near the desert region of Death Valley. Thus, it's easy to see why longitude has such predictive power. I believe that temperature is the main reason for these trends. Exploratory analysis showed that temperature is highly correlated with ozone levels. Specifically, lower temperatures correspond to lower ozone levels (and are often near the coast) while higher temperatures correspond to higher ozone levels (and are further inland near the deserts). This speculation is strengthened further by looking at the distribution of temperature across the days where ozone predictions were high across the entire state (August 15 and August 16). August 15th and 16th had the highest mean, median, and max temperatures. Since ozone is created in the presence of sunlight, this seems plausible.

While it may appear as if there was an error made when producing the standard deviation maps, I believe this is not the case due to the model being used. All days have almost the exact same standard deviation distribution due to the fact that longitude is the only variable used in the mean function , and each column of points in the predictions have the same longitude. It's clear that the largest uncertainties in prediction are at locations far from any ozone monitoring stations (or by very few monitoring stations) and vice versa for the lowest uncertainties.

6.2 Recommendations

- Perform a spatiotemporal analysis instead of just a spatial analysis
- Obtain more data
- Investigate the data quality of the integrated data sources
- Experiment with more covariance functions, as they have a very large

impact on predictions and their uncertainties

- Try to integrate covariates such as land cover/use
- Perform likelihood based inference on all parameters
- Maximize $\tilde{\vec{\omega}}$