# ENHANCING OZONE PREDICTION USING ENVIRONMENTAL AND SOCIAL JUSTICE DATA*

ANTONIO GAMBOA†, LUCRETIA GANT ‡, TIMOTHY GANT §, ANDREA GLOVER¶, AND JULIANNA LONG‖

**Abstract.** Preliminary results suggest a potential correlation between the ground-level ozone and socially vulnerable neighborhoods. By including the Center for Disease Control and Prevention/Agency for Toxic Substances and Disease Registry's (CDC/ATSDR) Environmental Justice Index (EJI) as an additional principal component in the model we are able to produce a more accurate and socially informative distribution of ground-level ozone. Our prediction function is an effective tool to investigate and guide policies aimed to improve social justice.
This research project demonstrates that using bi-variate splines over a triangulation and an adaptive principal component selection is an effective method for daily ozone level predictions in a city. We demonstrate this model using the previous days' ozone concentration together with the EJI, a prediction was made for the ozone levels around Atlanta for a day.

**Key words.** Ground Level Ozone, Ozone, Tropospheric Ozone, Environmental Justice Index,Social Vulnerability Index, Health Vulnerability Index, Adaptative Principal Component Analysis, Auto-regression

## 1. Introduction.

Ground-level ozone (O3), is a greenhouse gas that is highly reactive. High levels of ozone in the tropospheric layer of the atmosphere can cause respiratory issues and environmental damage.
Ozone is formed in the troposphere by photochemical reactions between oxides of nitrogen (NOx ) and volatile organic compounds (VOCs) in the presence of sunlight. The combined effects of ozone and its precursors exacerbate adverse health outcomes and mortality. Therefore, it is essential to identify sources, monitor concentrations, and develop forecasting systems for O3. Current monitoring systems are limited and provide an incomplete story about the presence of O3. The ability to precisely measure and forecast ozone concentrations within a city is necessary to better study ozone's adverse health effects.
Marginalized and vulnerable communities often experience higher rates of adverse health outcomes, including respiratory diseases which are known to be aggravated by elevated ozone exposure. An improved method of measuring and forecasting O3 will enable researchers to identify which communities are disproportionately affected by ozone pollution.
This study aims to utilize the Environmental Protection Agency (EPA) data on hourly O3 to model ozone levels, and the CDC/ATSDR Environmental Justice Index (EJI) which ranks census tracts in the United States according to variables that measure that environmental burden, social vulnerability and health vunerability within communities. It will then identify correlations between elevated ozone and environmental justice and social vulnerability.

By measuring and understanding the distribution of O3 and its health impacts,

*Submitted to the editors DATE.
**Funding:** This work was funded by US National Science Foundation award DMS-2051019.
†Garey High School,Pomona, CA (antonio.gamboa@pusd.org).
‡Dekalb County Schools, Atlanta, GA (lucretia_gant@dekalbschoolsga.org).
§Dekalb County Schools, Atlanta, GA (timothy_gant@dekalbschoolsga.org).
¶Druid Hills High School, Atlanta, GA (andrea_glover@dekalbschoolsga.org).
‖Stone Ridge School, Bethesda, MD (jlong@stoneridgeschool.org).

we can identify communities that are disproportionately affected by high levels of O3. Addressing the unequal burden of O3 exposure promotes social justice by ensuring that everyone receive an equal and fair opportunity to live in a healthy environment. The goal of this study is to contribute to social justice by addressing and rectifying the impacts of O3 faced by marginalized and vulnerable populations.

**2. Background Information.** What is Environmental Justice? Environmental justice is the fair treatment and meaningful involvement of all people regardless of race, color, national origin, or income to develop, implement, and enforce environmental laws, regulations, and policies. This goal will be achieved when everyone enjoys the same degree of protection from environmental and health hazards, and equal access to the decision making process to live, learn, and work in a healthy environment. Environmental justice issues are often divided into issues of "procedural justice" and issues of "distributive justice" (Kuehn, 2000). Procedural justice seeks the equitable involvement of all people in environmental decision-making, with a focus on addressing unequal power structures. Distributive justice seeks to address place-based disparities in exposures to environmental hazards and access to environmental amenities and other resources. Distributive environmental injustice can have profound cumulative impacts on human health and well-being. Addressing these cumulative impacts is a key part of promoting health equity.

The Environmental Justice Index (EJI) is a ranking of Census tracts (subdivisions of counties) in every state of the US based on 36 environmental, social and health factors that are grouped into 3 themes/modules. Each Module is an indication of the cumulative effects on communities in terms of environmental burden, social vulnerability and health vulnerability. An EJI score is a sum of scores from all three modules and is based on methods and measures of data from the Center for Disease Control, the US Mine Safety and Health Administration and the Environmental Protection Agency.

**3. Methods Two Different Approaches to Predicting Ozone.** Over the summer, we will explore how to quantify the uncertainty of two different ways of prediction low level ozone values using bivariate splines over triangulations. Here we will assume

- the ozone concentration over $\mathcal{D}$ is a random surface, $X$, with known values at N locations.
- We consider the highest ozone concentration in one day for one city to be a functional, $f(X)$, of the ozone concentration surface, $X$, over $\mathcal{D}$ on the previous day.
- We use the bivariate spline space $S_d^r(\triangle)$ with smoothness $r > 0$ and degree $d > r$ over a triangulation, $\triangle$, of $\mathcal{D}$ to approximate $X$ using the given measurement values at the N locations.

To fit the surface to the data, we will use the Penalized Least Squares (PLS) Method. For PLS, we let $\{(x_i, y_i, f(x_i, y_i)),\ i = 1, \ldots, N\}$ be a scattered data set where $N$ *is a relatively large integer*. Then the penalized least squares method is to find $s_f \in \mathcal{S}$ such that for a positive weight $\rho > 0$

$$P_\rho(s_f) := \min_{s \in \mathcal{S}} P_\rho(s)$$

where

$$P_\rho(s) := \sum_{i=1}^{N} |s(x_i, y_i) - f(x_i, y_i)|^2 + \rho E(s)$$

and

$$E(s) := \sum_{T \in \triangle} \int_T \left(s_{xx}^2 + 2s_{xy}^2 + s_{yy}^2\right) dxdy.$$

We will use the ozone surface splines as the input to a Functional Linear Model. This is similar to a how we think of a linear model except instead of data point we are feeding the model surfaces. We approximate the functional $f$ by a linear functional $Y = \langle \alpha, X\rangle + \epsilon$, where $E(\epsilon) = 0$ or let $\alpha$ be the solution of the following:

$$\alpha = \arg\min_{\beta \in H} E\left[(f(X) - \epsilon - \langle \beta, X\rangle)^2\right]. \tag{3.1}$$

where $H = L_2(\mathcal{D})$ be a standard Hilbert space of all square integrable functions over $\mathcal{D}$, $\langle f, g\rangle$ denotes the standard inner product of $f, g$ on $H$, $\|f\|$ the norm of $f$, $(X, Y)$ be a pair of random variables defined on the same probability space $\Omega$, with $X$ valued in $H$ and $Y = f(X)$ valued in $\mathbf{R}$, and $\mathcal{X} \subset H$ be a given set of random surface distributions.

**3.1. Autoregressive Process.** Similar to [1], to find a solution to the minimization problem (3.1) we want to write the equation in terms of the Covariance and Cross Covariance functions.

Covariance:

$$\Gamma x(s) = \int_{\mathcal{D}} \mathcal{E}(X(s)X(t))x(t)dt, \quad \forall x \in H.$$

Cross Covariance:

$$\Delta x = \int_{\mathcal{D}} \mathcal{E}(X(s)Y)x(s)ds, \quad \forall x \in H.$$

Hence we have,

$$\begin{aligned}
Y &= \langle \alpha, X \rangle + \epsilon \\
\langle X, x \rangle Y &= \langle X, x \rangle \langle \alpha, X \rangle + \langle X, x \rangle \epsilon \\
\mathcal{E}\left(\langle X, x \rangle Y\right) &= \mathcal{E}\left(\langle X, x \rangle \langle \alpha, X \rangle + \langle X, x \rangle \epsilon\right) \\
\langle \mathcal{E}(XY), x \rangle &= \langle \alpha, \mathcal{E}(\langle X, x \rangle X) \rangle + 0 \\
\Delta(x) &= \langle \alpha, \Gamma(x) \rangle.
\end{aligned}$$

In practice, we do not observe the continuous random surface $X_i$ but we only observe the random surface at design points $s_k \in \mathcal{D},\ k = 1, \ldots, N$:

$$\{z_{i,k},\ k = 1, \ldots, N\}.$$

So we approximate $X_i$ by a smooth bivariate spline denoted $S_{X_i}$ and the empirical estimators can be approximated by

$$\widetilde{\Gamma_n}(x) = \frac{1}{n}\sum_{i=1}^{n} \langle S_{X_i}, x \rangle S_{X_i} = \sum_{j=1}^{m} \widetilde{\lambda_j} \langle \widetilde{v_j}, x \rangle \widetilde{v_j}$$

$$\widetilde{\Delta_n}(x) = \frac{1}{n}\sum_{i=1}^{n} \langle S_{X_i}, x \rangle Y_i,$$

where $\widetilde{\lambda_j}$ and $\widetilde{v_j}$ are a pair of eigenvalue and eigenvector of $\widetilde{\Gamma_n}$ and $m$ is the dimension of the spline space $S_d^r(\triangle)$. It then follows that

$$\widetilde{\Delta_n}(x) = \langle \alpha_n, \widetilde{\Gamma_n} x \rangle$$

for some $\alpha_n \in H$.

Assume that the first $k_n$ largest eigenvalues $\widetilde{\lambda_j}, j = 1, \cdots, k_n$ are nonzero. Then the principal component regression estimator of $\alpha_n$ is

$$\widetilde{\alpha}_{PCR} = \sum_{j=1}^{k_n} \frac{\Delta_n(\widetilde{v_j})}{\widetilde{\lambda_j}} \widetilde{v_j}. \tag{3.2}$$

**Computational Method:** Now we would like to solve $\hat{A}c = \hat{b}$ for $c$ which will yield a coefficient vector for a spline $S_\alpha$ such that

$$f(X) \approx \langle S_\alpha, S_X \rangle.$$

1. Fit each $X_i$ with a PLS spline $S_{X_i}$.
2. Use $S_{X_i}$ to calculate $\widetilde{\Gamma_n}$ and $\widetilde{\Delta_n}$.
3. Compute the SVD of $\widetilde{\Gamma_n}$.
4. Select appropriate number of non-zero eigenvalues $k_n$.
5. Compute $\widetilde{\alpha}_{PCR}$ in (3.2).

**3.2. Brute Force Approach .** We want to find a solution solving the following minimization problem:

$$\alpha = \arg\min_{\beta \in H} E\left[(f(X) - \epsilon - \langle \beta, X \rangle)^2\right].$$

Since bivariate spline space $S_d^r(\triangle)$ can be dense in a Hilbert space, $H$, as $|\triangle| \to 0$, we look for $S_\alpha \in S_d^r(\triangle)$ of $\alpha$ such that

$$S_\alpha = \arg\min_{\beta \in S_d^r(\triangle)} E\left[(f(X) - \epsilon - \langle \beta, X \rangle)^2\right]. \tag{3.3}$$

Let $\{\phi_1, \cdots, \phi_m\}$ be a basis for $S^r_d(\triangle)$. We write $S_\alpha = \sum_{j=1}^m c_j \phi_j$. Then its coefficient vector $\mathbf{c} = (c_1, \cdots, c_m)^T$ satisfies the following relation:

$$A\mathbf{c} = \mathbf{b}$$

with $A$ being a matrix of size $m \times m$ with entries

$$E(\langle \phi_i, X\rangle \langle \phi_j, X\rangle)$$

for $i, j = 1, \cdots, m$ and $\mathbf{b}$ being a vector of length $m$ with entries $E(f(X)\langle \phi_j, X\rangle)$ for $j = 1, \cdots, m$.

We want to consider the empirical estimate of $S_{\alpha_D}$ based on discrete observations of random surfaces $X_i, i = 1, \cdots, n$. The empirical estimate $\widetilde{S_{\alpha_D,n}} \in S^r_d(\triangle)$ is the solution of

$$\widetilde{S_{\alpha_D,n}} = \arg \min_{\beta \in S^r_d(\triangle)} \frac{1}{n} \sum_{i=1}^n (f(X_i) - \epsilon_i - \langle \beta, S_{X_i}\rangle)^2. \tag{3.4}$$

In fact the solution of the above minimization is given by

$$\widetilde{S_{\alpha_D,n}} = \sum_{i=1}^m \widetilde{c_{n,i}} \phi_i \tag{3.5}$$

with coefficient vector $\widetilde{\mathbf{c}_n} = (\widetilde{c_{n,i}}, i = 1, \cdots, m)$ satisfying $\widetilde{A_n}\widetilde{\mathbf{c}_n} = \widetilde{b_n}$,

$$\widetilde{A_n} = \left[ \frac{1}{n} \sum_{\ell=1}^n \langle \phi_i, S_{X_\ell}\rangle \langle \phi_j, S_{X_\ell}\rangle \right]_{i,j=1,\cdots,m}$$

$$\widetilde{b_n} = \left[ \frac{1}{n} \sum_{\ell=1}^n f(X_\ell)\langle \phi_j, S_{X_\ell}\rangle + \frac{1}{n} \sum_{\ell=1}^n \langle \phi_j, \epsilon_\ell S_{X_\ell}\rangle \right]_{j=1,\cdots,m}.$$

**3.3. Computational Method.** For a time series over a regional domain we use a bivariate spline to approximate a surface over a bounded region. We call the resulting spline $S_X$ it is an approximation of a functional random variable $X$. We also collect the desired quantity at location of interest. This yields another real random variable $Y = f(X)$. Next we compute:

$$\hat{A} = E(\langle \phi_i, S_X\rangle \langle \phi_i, S_X\rangle) = \frac{1}{n} \sum_{l=1}^n \langle \phi_i, S_{X_l}\rangle \langle \phi_j, S_{X_l}\rangle$$

$$\hat{b} = E\left((f(X) - \epsilon)\langle \phi_i, S_X\rangle\right) = \frac{1}{n} \sum_{l=1}^n f(X_l)\langle \phi_j, S_{X_l}\rangle$$

Now we would like to solve $\hat{A}c = \hat{b}$ for $c$ which will yield a coefficient vector for a spline $S_\alpha$ such that

$$f(X) \approx \langle S_\alpha, S_X\rangle.$$

Then we implement the two options we have explored for finding the coefficients of $S_\alpha$: Brute Force (3.5), and Principle Component Analysis (3.2).

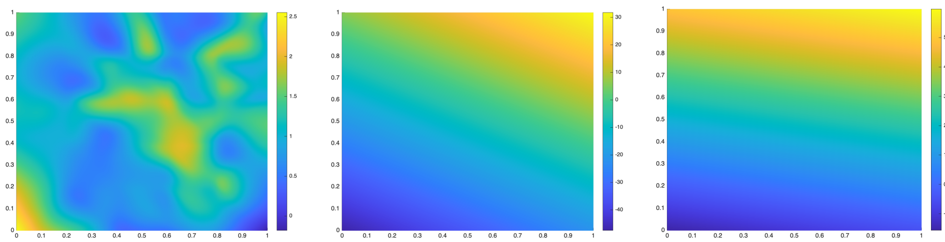

FIG. 1. *Left: Environmental Justice Index for Atlanta, Center:Ozone-based Prediction function for Atlanta using only 03, Right:*

It should be detailed enough to guide someone who wants to reproduce your study. It should not, however, be structured as a chronology of what you did (for example, "First we tried this, but that didn't work, so then we did this."). If you encountered problems, first discuss the methodology you finally settled on. The main question that readers expect to be answered is

*How was this work done?*

Here we state our main result as Theorem 3.1; the proof is deferred to Appendix A.

THEOREM 3.1 ($LDL^T$ Factorization [**?**]). *If $A \in \mathbb{R}^{n\times n}$ is symmetric and the principal submatrix $A(1:k,1:k)$ is nonsingular for $k = 1:n-1$, then there exists a unit lower triangular matrix $L$ and a diagonal matrix*

$$D = \operatorname{diag}(d_1, \ldots, d_n)$$

*such that $A = LDL^T$. The factorization is unique.*

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Our analysis leads to the algorithm in Algorithm 3.1.

**Algorithm 3.1** Build tree

Define $P := T := \{\{1\}, \ldots, \{d\}\}$
**while** $\#P > 1$ **do**
  Choose $C' \in \mathcal{C}_p(P)$ with $C' := \operatorname{argmin}_{C\in\mathcal{C}_p(P)} \varrho(C)$
  Find an optimal partition tree $T_{C'}$
  Update $P := (P\backslash C') \cup \{\bigcup_{t\in C'} t\}$
  Update $T := T \cup \{\bigcup_{t\in\tau} t : \tau \in T_{C'}\backslash\mathcal{L}(T_{C'})\}$
**end while**
**return** $T$

**4. Main Results.** ***Our team has enhanced a day predictive function for the concentration of ground level ozone. The computational model was written using

@Matlab software. The predictive model includes a learning dataset comprised of thirty previous known daily hour ozone concentration. It also includes an adaptive Eigen value selection utilized for the prediction based on percent variance of the known values. We then added a principle component consisting of the EJI in order decrease the mean percent error. To validate our model we defined a quadrant geolocation of Atlanta's vicinity. The coordinates utilized were the following: latMax = -84.05; latMin = -85.05; longMax = 34.02;and longMin = 33.40. In addition, we selected the 2016 EPA hourly and daily Ozone dataset to be used as our learning model. The rationale behind selecting the 2016 dataset included two major determinants. The first determinant was that the EPA dataset for 2016 contains a larger number of ozone sensor stations than a more current year dataset. The second determinant was that the CDC-EJI index from the 2020 census census tract was based on Ozone data collected from 2014-2016. To correlate the ozone data with that of EJI an additional dataset from the US Census Bureau ***TIGER/Line Shapefiles*** (is this correct?Is anyone aware of this being done versus the EJI merged dataset?)*** was utilized. This dataset allowed matching the coordinates within our Atlanta quadrant and the latitude/longitude relationships to the US census bureau census tract EJI areas.***

The main question that readers expect to be answered is

*What were the results? What do those results mean ?*

Try to think about what your reader wants to know about your results. Here are some questions to guide you:

- What, specifically, did you learn from comparing these algorithms or data structures?
- What do your results say about the problem or question you were investigating?
- Was your hypothesis confirmed or disproved?
- Are the results what you expected?
- If you obtained anomalies or other unexpected results, can you explain them? If not, how could you set about in the future to identify what caused them?
- How do your results compare to past findings? Are they consistent? Different? Why?
- How would you respond to objections or questions that other researchers might have about your methods, results, or interpretations?

**5. Experiments.** The computational experiments should be reported in a reproducible way, which would require that programs, data and results are made available and documented.

For our experiments, we used the TOAR surface ozone database, which is a publicly available global database of surface-level ozone and related metrics which can be found at https://toar-data.org/surface-data/. TOAR does not collect ozone data, but aggregates measurements taken by a variety of organizations world wide. The MATLAB code used in our experiments to access the TOAR data can be found in appendix IDK. The TOAR database can be queried using the REST interface. This allows the user to set parameters on the desired data such as latitude and longitude limits, sampling frequency and type of data. In order to access ozone data, is essential to set the `variable_id` flag to 5. Other `variable_ids` will return other ozone-related values such as particulate matter, volatile organic compound, and NOx levels. As of June 2023, TOAR has not implemented a method for limiting your search by date. Querying the TOAR database occurs in two stages. First the search query will return a list of the time-series files that meet your parameters. Second, the user must download each time-series file. When downloading the time-series files, it is helpful to use the 'AllOK' flag which means that the data has been verified by both the organization that collected the data and TOAR.

We pre-process the data by filling in gaps using spline interpolation. Only gaps 12 hours or smaller are filled (Apendix A). Then we reduce the dataset to only the dates of interest (Apendix B). Finally, we reformat the table into a matrix with latitude, longitude, and ozone columns(Apendix C).

For the EJI experiment we used merged datasets from CDC/ from the Environmental Justice Index

To evaluate the efficacy and accuracy of the model, we chose one location (Emory University) and one day (July 10, 2016), to compare the actual measurement and the predictions made with and without EJI. Figure 2 shows a comparison between the actual measurement and two predictions for each hour of the day. Here we see that the inclusion of the EJI values increased the accuracy of our predictive model. Table 1.

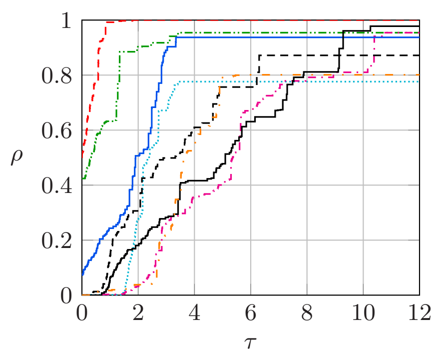


FIG. 2. *Example figure using external image files.*

Table 1 shows additional supporting evidence.

TABLE 1
*Example table.*

| Species | **Mean** | **Std. Dev.** |
|---|---|---|
| 1 | 3.4 | 1.2 |
| 2 | 5.4 | 0.6 |
| 3 | 7.4 | 2.4 |
| 4 | 9.4 | 1.8 |

In reviewing the analyses, it is notable that the first hour of each day is always predicted to have no ozone. Additionally, the shape of the hourly prediction always closely resembles the shape of the true data from the previous day. It would be interesting to run an experiment testing if our predictions are any improvement over simply assuming each day will have the same ozone levels as the day before it.

**6. Conclusions.** The conclusion answers the readers' question: "So what?" It should give your readers points to "take home" from your paper. It should state clearly what your results demonstrate about the hypothesis you were testing in the paper. It should also generalize your findings, perhaps suggesting how others can use them in future research. All generalizations should be supported by your data, however; the discussion should have proved these points, so that when the reader gets to the conclusion, the statements are logical and seem self- evident. No new evidence should be introduced in the conclusion.

Future work: state your ideas of what you estimate could be done in the future in order to further improve on the addressed problem. You may also state further problem areas you identified to be researched in the future. This section is not mandatory, but may be useful for others in identifying interesting new research problems.

**Appendix A. An example appendix.** Aenean tincidunt laoreet dui. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Integer ipsum lectus, fermentum ac, malesuada in, eleifend ut, lorem. Vivamus ipsum turpis, elementum vel, hendrerit ut, semper at, metus. Vivamus sapien tortor, eleifend id, dapibus in, egestas et, pede. Pellentesque faucibus. Praesent lorem neque, dignissim in, facilisis nec, hendrerit vel, odio. Nam at diam ac neque aliquet viverra. Morbi dapibus ligula sagittis magna. In lobortis. Donec aliquet ultricies libero. Nunc dictum vulputate purus. Morbi varius. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In tempor. Phasellus commodo porttitor magna. Curabitur vehicula odio vel dolor.

LEMMA A.1. *Test Lemma.*

**Acknowledgments.** We would like to thank the National Science Foundation (NSF) DMS-2051019, The Tropospheric Ozone Assessment Report (TOAR) , The Emory Math and Computer Science Departments and Dr. Bree Ettinger, online resources (url is commented out in LaTeX) for their help in putting together this manuscript template.

## REFERENCES

[1] B. E. S GUILLAS, MJ LAI, *Bivariate splines for ozone concentration forecasting*, Environmetrics, 23 (2012), pp. 317–328.