MACHINE LEARNING OF BIG DATA

A Gaussian Regression Model to Predict the Spatiotemporal Distribution of Ground Ozone

Abstract

Tracking pollution levels on the ground is important to the environment and public health. One of the pollutants of concern is ozone, which, at high concentrations, can cause respiratory and cardiovascular problems. The National Center for Atmospheric Research (NCAR) has published valuable ozone data obtained from ground-based sensors installed at selected locations. Because it is unfeasible to measure the exact ozone levels everywhere at any time, it would be valuable to predict the temporal-spatial distributions of ozone concentration based on existing data. This would help us better understand the patterns and trends in the data and make better decisions to reduce pollution. Motivated by this, the objective of this paper was to build predictive models to illustrate the temporal-spatial structure of the large amount of ozone data. The training data included measurements of ozone in 513 locations in the eastern states of the United States spanning five years. We used a machine-learning method called Gaussian process regression (GPR) with a covariance function that describes the temporal-spatial relationship between data points. With this method, we were able to observe the trends and dynamics of ozone formation. Additionally, maps were created to visualize the spatial and temporal distribution of ozone concentrations. The results demonstrate that the GPR method with the Matérn covariance function was able to give a reliable estimate of the uncertainty as well as the mean ozone concentration at various locations and times, which helps us better understand the dynamics of ozone formation.

Keywords

Gaussian process regression, Matérn covariance function, spatio-temporal modeling, ground-level ozone concentrations, air quality monitoring, spatial statistics, geostatistics

https://doi.org/10.7771/2158-4052.1613

Student Author

Jerry Gu is a junior at West Lafayette Jr./Sr. High School in Indiana and a nondegree student at Purdue University, where he has been taking courses from the Departments of Mathematics and Computer Science. He has been working with Professor Lin on machine learning of big data since August 2021. In the summer of 2022, he interned at the Crane Naval Surface Warfare Center, where he developed an image processing method with a convolutional neural network to identify and classify obstacles and street signs for ground military drones. He also participated in research on unmanned aerial systems (UAS), creating a simulation environment to predict the optimum trajectory of autonomous drones. These research experiences have deepened his interest in applying mathematics to the design and development of artificial intelligence, enabling the creation of intelligent systems that can learn, reason, and make decisions.

Mentor

Guang Lin is a Full Professor in the Department of Mathematics and the School of Mechanical Engineering at Purdue University. His research interests include diverse topics in computational and data science both on algorithms and applications. His main current thrust is machine learning, data-driven modeling, stochastic simulation, and multiscale modeling of interconnected, physical, and biological systems. He is the Director of Data Science Consulting Service, which performs cutting-edge research on data science and provides hands-on consulting support for data analysis and business analytics in all areas to overcome data science challenges arising in research, education, and business and organization management. Professor Lin is currently also Co-Chair of the Purdue Engineering Initiative in Data Engineering and Application.

INTRODUCTION

Ground-level ozone is a major component of air pollution and can be detrimental to human health.[1] High concentrations of ozone can cause respiratory and cardiovascular problems, as well as damage to vegetation, crops, and other natural resources.[2] According to the Environmental Protection Agency's (EPA) National Ambient Air Quality Standards (NAAQS), the acceptable ozone concentration is 70 parts per billion (ppb), which is calculated based on the daily maximum 8-hour average.[3] Monitoring ground-level ozone concentrations can help to identify areas with high ozone levels, which can then be tackled through public health initiatives, air quality regulations, and other methods to reduce air pollution. For this reason, EPA established ground sensors at 513 monitoring stations in the eastern United States to measure the daily surface ozone concentrations over a period of several years, and the data are publicly accessible.[4]

Ground-based sensors can be costly, so they are usually only installed in select areas with greater distances between them in order to reduce expenses. Additionally, these sensors may not be able to collect data continuously or at frequent intervals, leading to coarser spatial and temporal resolutions of the recorded data. This can make it difficult to identify the true source and location of the pollution and make decisions about how to address the issue. As a result, there is increasing interest in analyzing massive data to detect meaningful spatiotemporal dependence patterns and to subsequently smooth and predict in the space-time domain. Various algorithms have been developed to construct more flexible, accurate, and computationally efficient models for large spatio-temporal data.[5,6] Ma et al. provided a review of three common statistical spatial-temporal models for ambient ozone exposure in environmental epidemiology studies: the land use regression model, the random forest model, and the artificial neural network model.[7] The latter two methods, both of which are machine-learning algorithms, provide nonlinear mapping tools for large datasets, though they can be prone to overfitting with high-dimensional features and have a slow convergence rate, resulting in a high computational cost.

Motivated by the need to reduce the computational cost, this study seeks to apply a Gaussian regression process model with the Matérn covariance function to predict

the spatial and temporal distribution characteristics of ground-level ozone concentration in the eastern states of the United States. We will describe the data source and statistical model used, then show the trend and dynamics of ozone formation based on the predicted daily, monthly, and yearly ozone concentrations at a higher spatiotemporal resolution. Additionally, we will discuss the uncertainty analysis of the simulation results, as well as the factors influencing ozone patterns and trends.

MATERIALS AND METHODS

Training Data

The training data included measurements of the daily maximum average 8-hour ozone in 513 locations during the period 1995–1999.[4] There were 920 total samples for each location over the course of 5 years, indicating a total number of 471,960 data samples. The data spans 23 states in the eastern half of the United States. Figure 1 shows the distribution of the 513 ozone monitoring sites. It is clear that the sites are more concentrated in the big cities with large populations, whereas they are more sparse in rural areas and inner states. Neighboring sites can be as close as a few miles, or as distant as a thousand miles.

Statistical Method

In this study, we focused on the ground level ozone concentration at various locations, which was the output of the machine-learning model (dependent variable). The input data (or independent variables) were the location and time (longitude and latitude of the location and date).

We began by cleaning up the original recorded data, then performing statistical analysis on the datasets. To analyze the data, we used a cross-validation procedure whereby 70% of randomly selected data points were used as the training set and the remaining 30% as the testing set for assessing the model's predictive performance.

After examining the collected data, we found a weak relationship between spatial and temporal data, which allowed us to separate temporal and spatial components. Consequently, we decided to conduct the spatiotemporal prediction in two stages: (1) the use of Gaussian process regression for the modeling and prediction of the spatial distribution at each time step; and (2) the modeling and prediction of the temporal distribution based on the observed temporal covariance from the measurements.

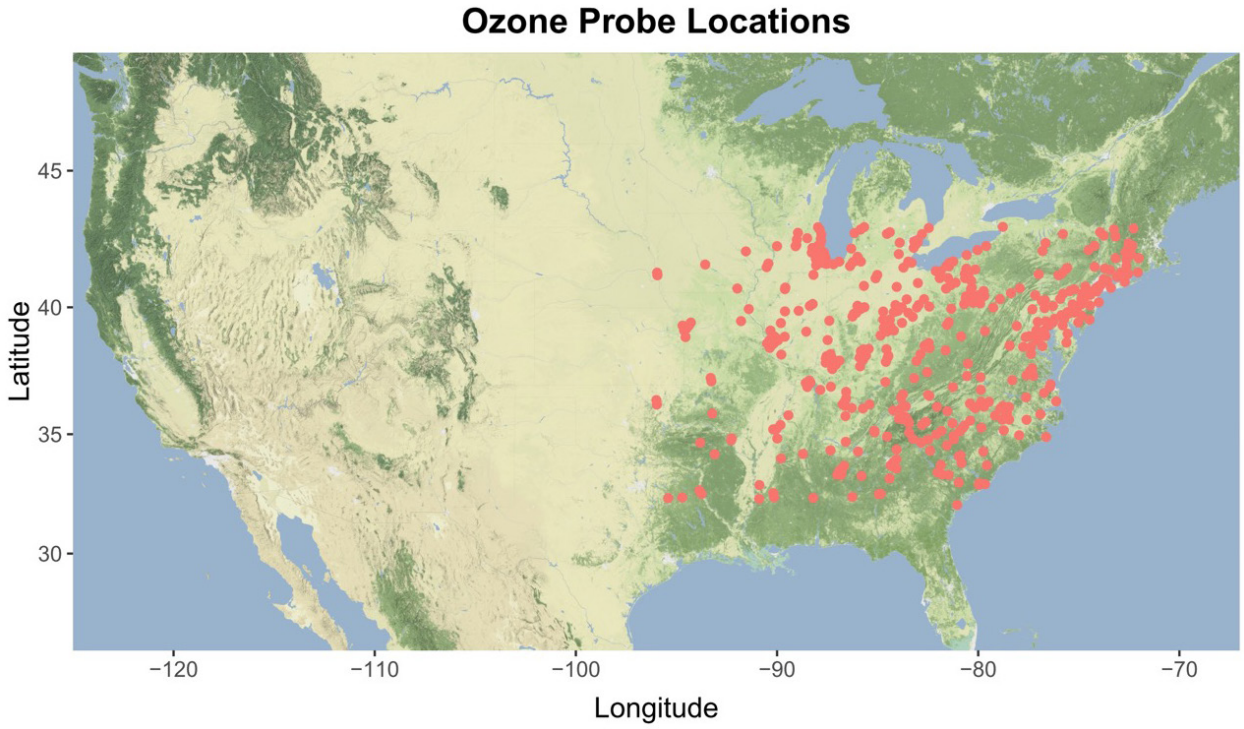

FIGURE 1. The distribution of ozone monitoring sites in the 23 states of the eastern United States.

This decoupled approach proved to be highly efficient, while still providing satisfactory results.

Using a machine-learning method called Gaussian process regression (GPR) to predict the temporal spatial dependence structure of the data, we assumed that the output variable was a random function of the input variable and that this random function was drawn from a Gaussian process. We selected a covariance function that describes the temporospatial relationship between data points, determining how quickly the data values change as a function of time and space. Then, we fitted the model to the training data set by maximizing the likelihood of the data under the model. Subsequently, we used the testing data set to evaluate the model by comparing the model-predicted ozone concentration with measured data. If the model met the success criteria, we could use it to make predictions about the ozone concentration at new input locations. For each new input location, we computed the mean and variance of the predicted output distribution. The mean represents the most likely value of the output variable at that location, while the variance represents our uncertainty about that prediction.

The Matérn kernel was chosen for this study due to its ability to calculate the covariance solely based on the distance between data points:

$$k(x_i,x_j)=\frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}}{l}d(x_i,x_j)\right)^{\nu}K_{\nu}\left(\frac{\sqrt{2\nu}}{l}d(x_i,x_j)\right)$$

where is the distance between two location vectors and ; is the gamma function; is the smoothness parameter, which controls the level of differentiability in the function; l is the length-scale parameter, which determines how quickly the covariance decays with distance; and is a modified Bessel function of the second kind.

We first organized the data into data frames in R and removed missing values from the dataset. Then we chose the Matérn kernel as the covariance function based on the distance between points, which is commonly used in spatial statistics. Finally, we used Ski-Kit Learning, a machine-learning library in Python, to conduct GRP. The output of the model is a prediction of the temporo-spatial structure of the data, which can be used to predict future values of the data, such as at different locations and times. Plots and animations were created to visualize ozone predictions in the eastern states.

Finally, it is important to note that even though the data utilized in this study is not the most up-to-date, the method and approach employed can still be used for more current data and other pollutants.

RESULTS

Descriptive Statistics

We first chose the city of Chicago to investigate the ground ozone concentrations over daily, seasonal, and annual timescales. Figure 2 demonstrates the ozone data for the months of May to October from 1995 to 1999, with no data available for the months of November to April due to the inactivity of the ozone sensors during this period. The boxplot for each year is included, which is a method used to show the distribution of the data through their quartiles. For all of the boxplots in this study, the middle, upper, and lower lines of the box represent the mean, 75%, and 25% percentiles of the data, and the upper and lower legs of each boxplot represent 95% and 5%; the rest of the dotted data are the outliers of the statistics. The yearly average ozone concentration does not change drastically with values of 48.6, 44.2, 43.7, 45.6, and 43.0 ppb, respectively. However, there are large changes within each year, with August having the highest peak value and October having the lowest. On top of this, the annual peaks and lows are consistent over the years, with the average being less than half of the peak value.

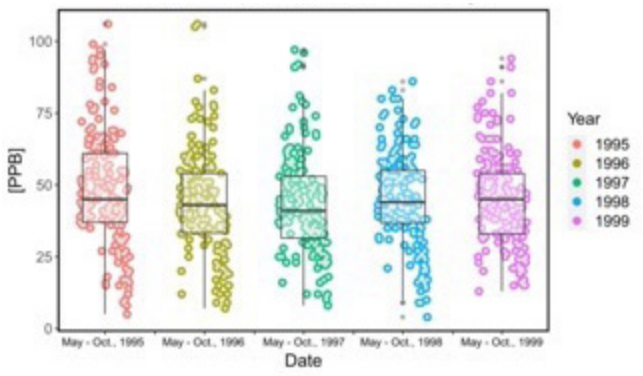

FIGURE 2. Statistics of the ground ozone concentration data in Chicago from May to October during the years 1995–1999.

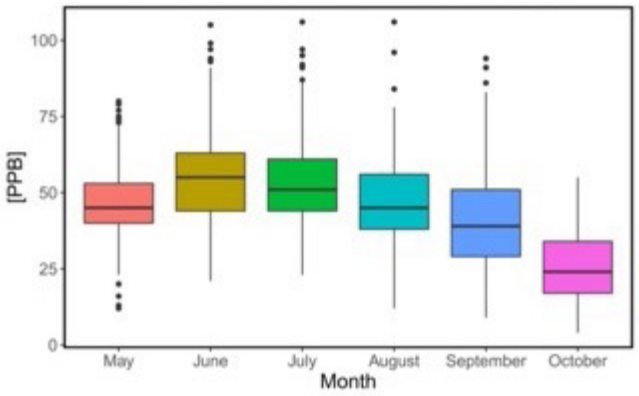

FIGURE 3. Ozone distribution in each month from May to October in 1999.

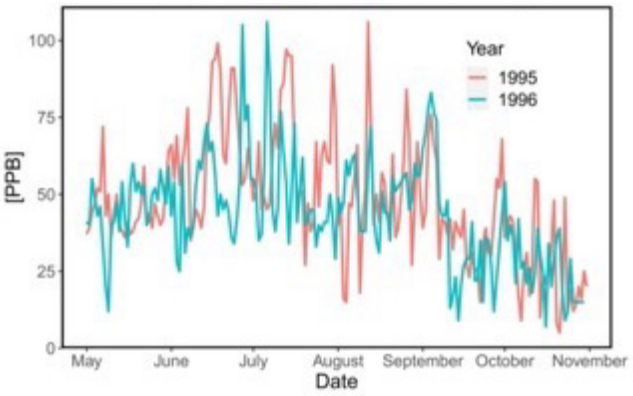

FIGURE 4. Daily variation of ozone concentration from May to October in Chicago in 1995 and 1996.

Figure 3 shows the average monthly ozone concentration of Chicago in 1999. It is evident that June has the highest ozone level, followed by July, while October is the lowest. The data points outside of the boxplots for June, July, and August indicate that the daily peaks for these months are nearly identical, around 110 ppb, whereas May has a slightly lower daily peak of 80 ppb. Additionally, May is the only month that has quite a few data points falling below the boxplot (outside of the 25% range), indicating the lowest daily ozone concentrations in May are far below the average. On the other hand, October has no outliers as all data points are within the boxplot. This implies that the daily ozone concentration changes significantly from May to August, but is relatively stable in October, which is likely due to the decrease in temperature. It is well documented that ground ozone concentrations are high in summer and low in winter because of the interactions between sunlight, temperature, and air pollutants. In summer, the intense sunlight and high temperatures create favorable conditions for ozone formation from air pollutants like nitrogen oxides and volatile organic compounds. In fall and winter, the decreased sunlight and lower temperatures limit ozone formation and reduce the concentrations of ozone.

Figure 4 illustrates the daily variation of ozone concentration in Chicago from May to October in 1996 (blue) and 1995 (red). It is clear that ozone concentration increases from May to June, July, and August, then decreases in September and October. The most noteworthy aspect of this figure, however, is the large fluctuations in ozone concentration on a daily basis. For

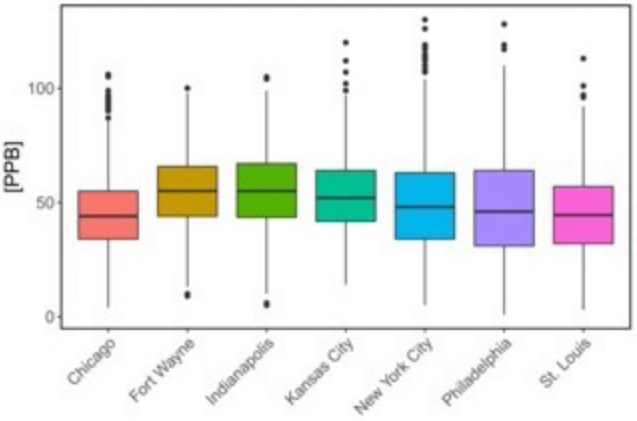

FIGURE 5. Annual average ozone concentration in different cities, including Chicago, Fort Wayne, Indianapolis, Kansas City, New York City, Philadelphia, and St. Louis.

instance, in 1995, the highest daily ozone concentration was recorded on August 12 with a value of 106 ppb, while the lowest was recorded on October 21 with a value of 5 ppb. In 1996, the highest daily ozone concentration was recorded on June 27 and July 6 with a value of 106 ppb, and the lowest was on October 16 with a value of 7 ppb. Within just a few days in the same month, drastic changes in ozone concentration can occur, which is likely due to significant daily fluctuations in temperature.

Next, we examined several other major cities in addition to Chicago: Fort Wayne, Indianapolis, Kansas City, New York City, Philadelphia, and St. Louis. Figure 5 shows the average yearly ozone concentration in these cities. Fort Wayne and Indianapolis had the highest average concentration while Chicago had the lowest. Interestingly, New

York City had the highest peaks, indicating that it may experience more severe air pollution at certain times of the year.

Extensive studies have suggested that ground ozone is primarily caused by emissions from motor vehicle exhaust, industrial facilities, and chemical solvents. While cities such as New York City and Chicago may have heavier traffic than Indianapolis and Fort Wayne, other factors also contribute to levels of ground ozone, such as geography and local industries. For example, although Chicago's population dominates in precursor emissions leading to high ozone formation, winds from the lake tend to blow the ozone and precursor emissions away from the city. Conversely, Fort Wayne and Indianapolis have less wind and traffic than Chicago, but more heavy industrial facilities, such as chemical plants, petroleum refineries, and power plants, which emit nitrogen oxides (NOx) and volatile organic compounds (VOCs), resulting in higher yearly average ground ozone formation. Indeed, according to the Environmental Protection Agency (EPA), the State of Indiana ranked 19th in ground ozone concentrations among all states in recent years.[8]

Spatial and Temporal Covariance

In Figures 6 and 7 present the spatial and temporal covariance of ozone daily exposure data, providing information on the spatiotemporal variation of the ozone residual concentration. As observed, the covariance decreases with increasing distance or time. Figure 6 depicts the monthly spatial covariance, showing the covariance between two locations within a 1,500-mile distance. Significant covariance (> 0.25) is observed when the distance is less than 500 miles. However, differences are evident among the months, with October displaying more data points outside of the 5%–95% range. This suggests that local weather fluctuations or extreme weather may have a more significant impact on ozone generation during this month. Figure 7 presents the monthly temporal covariance, which spreads across a narrower range than the spatial covariance, with fewer

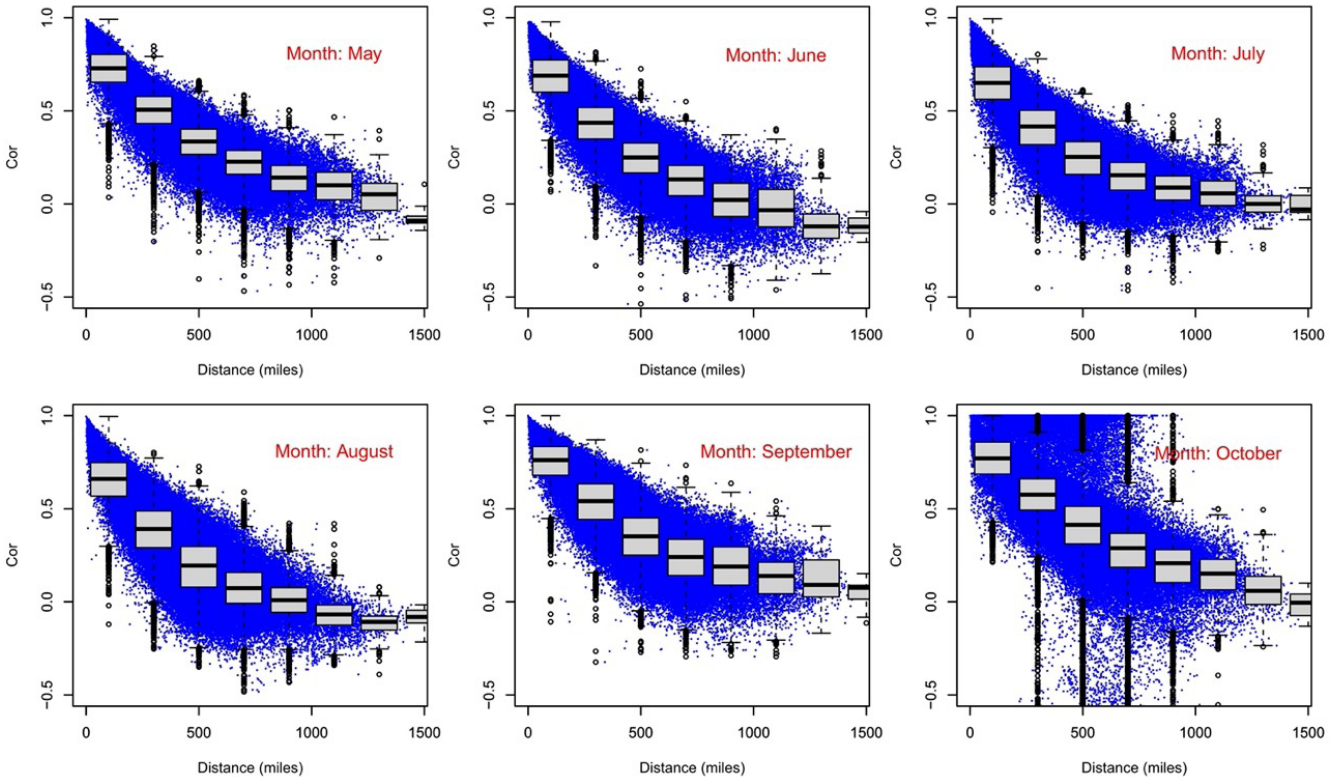

FIGURE 6. Observed covariance of the ozone residual concentrations on a daily time scale shown as a function of spatial lag from May to October for all years.

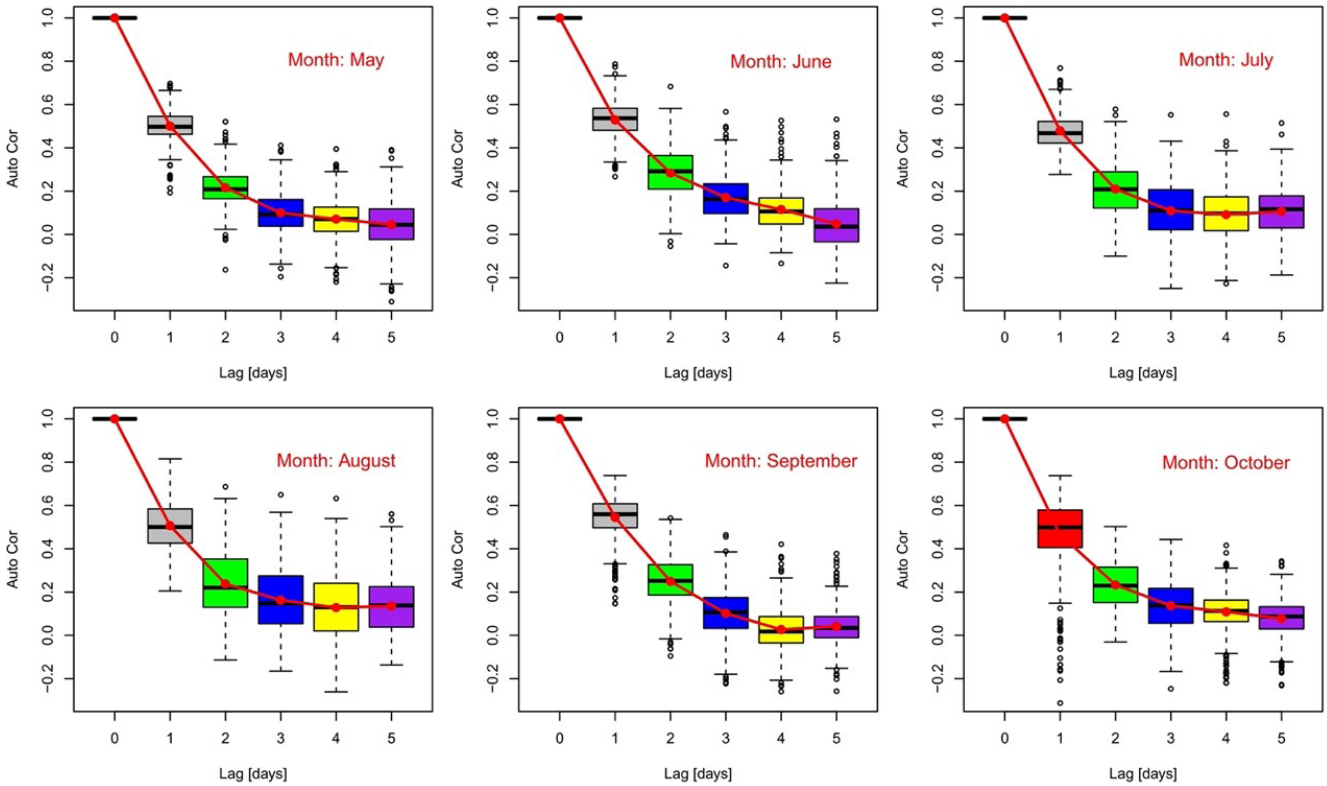

FIGURE 7. Observed covariance of the ozone residual concentrations on a daily time scale shown as a function of temporal lag from May to October for all years.

TABLE 1. GPR model quality with testing dataset.

R^2	RMSE	MAE	MAPE
0.7–0.89	12–20	11–14	2.5%–3.7%

outliers observed in each boxplot. Furthermore, it indicates that the temporal covariance is only affected by 3–4 lags, making temporal prediction computationally efficient.

To predict the spatial distribution of ozone concentration, a GPR machine-learning approach was employed, and the testing dataset (30% of the original data points) was utilized to assess the model's performance. Table 1 presents the statistical metrics of the ozone concentration prediction in comparison with the measurements. The model's statistical quality is generally acceptable, indicating its potential for predicting ozone concentration on uniform longitude-latitude grids in areas where no measurements are available.

Temporospatial Distribution of Ozone Concentrations

As discussed earlier, the uneven distribution of ozone monitoring stations, concentrated mainly in urban areas, does not fully reveal the overall spatial distribution characteristics of ozone. Moreover, ground monitoring stations cannot monitor changes in ozone concentrations over extended periods of time. To address this, models must be developed that are capable of predicting the temporospatial structure of the data so that estimated ozone levels can be determined in areas and times lacking monitoring data. This will facilitate a greater understanding of ozone distribution in the United States, as well as provide a scientific basis for air pollution monitoring and management.

To investigate the spatial distribution, we picked the first day of each month from May to October in 1999 and used the Gaussian regression process model with a covariant function to calculate the ozone concentration

at any location on the map, which is a function of the data from the 512 surrounding monitoring stations. Figure 8 shows the resulting spatial distribution of ozone concentrations, which were generated based on one-day data from May to October. We observed a clear trend that the overall ozone level increases from May, reaches a maximum in August, and then decreases again in October. Additionally, the spatial distribution of ozone varied significantly. For example, on June 1, the midwestern and southern states of Ohio, Kentucky, Tennessee, Georgia, and South Carolina had higher ozone levels than the east coast states such as Connecticut, New York, New Jersey, and Delaware. On July 1, northern states of Illinois, Indiana, Ohio, Pennsylvania, and New York seemed to have higher ozone concentrations than the southern states. However, on August 1, the highest ozone concentration areas were focused on the big cities along the east coast, starting from New York, to Philadelphia, to Washington, DC, to Atlanta. This correlates with the traffic of these big cities, where mobile cars emit Nox, which interacts with the strongest sunlight in August to produce the most ozone. Furthermore, on August 1, the warning threshold of ground ozone concentration, which is 70 ppm, was exceeded in most of the 25 eastern states, suggesting that precautionary measures should be taken

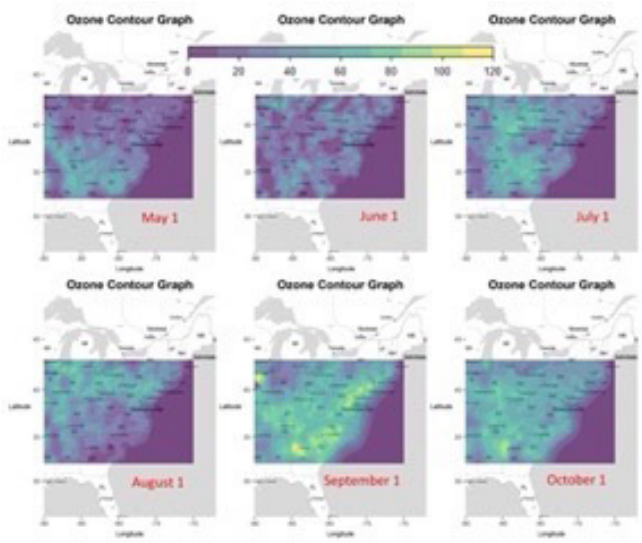

FIGURE 8. Ozone concentration distribution map on the first day of each month from May to October in 1999.

such as staying indoors and using air filters during the hottest summer season in these regions.[3]

While Figure 8 demonstrates the spatial characteristics of the data, Figure 9 illustrates both the temporal and spatial structure. It shows the average monthly ozone concentration in the 25 states. The data showed an

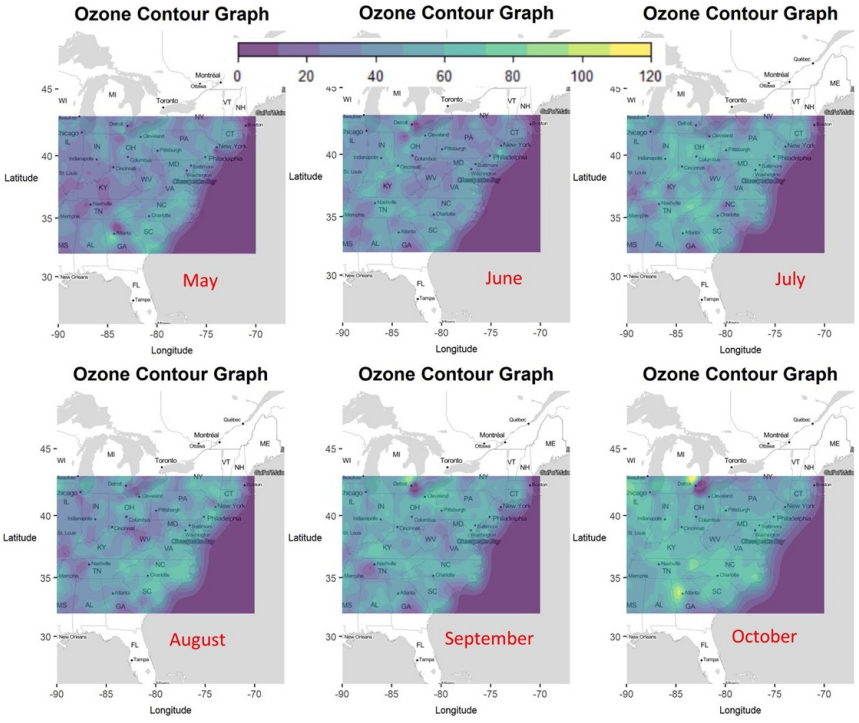

FIGURE 9. Monthly averaged ozone concentration distribution map from May to October in 1999.

overall increase in ozone concentration from May to August and September, followed by a significant decrease in October. While Figures 8 and 9 have many similarities, it is critical to consider the monthly (or seasonal and yearly) average when it comes to overall air quality control and management. Both temporal and spatial characteristics are thus of great importance.

DISCUSSION AND CONCLUSION

Examining the temporospatial distribution of ground ozone in the 25 states, it is evident that ozone levels generally increase from May to July, August, and September, and then decrease again in October. This is closely linked to temperature and sunlight exposure, which are the two most important parameters for ozone formation chemistry. Extensive research has established that ozone concentration is low at night due to lack of photons and increases during the day as temperature rises and sunlight exposure increases. Additionally, it is well known that in the summertime, intense solar radiation and prolonged high temperatures result in increased photochemical reactions caused by NOx and VOCs in the atmosphere, and the low relative humidity exacerbates ozone pollution.[9]

Upon examining the maps more closely, some interesting trends can be identified. Besides significant spatial variations across the 25 eastern states, there are huge daily fluctuations. Overall, the daily fluctuation is smallest in October and largest in the summer months of July and August. Additionally, in August, high ozone concentrations are concentrated around metropolitan cities along the east coast with high population density and traffic. However, in June, the highest ozone levels are found in midwestern cities. This shows that there is a change in the highest ozone areas between the months. Additionally, there are some parts of the Midwest and South that generally have low ozone levels, which tend to be suburban areas. Furthermore, comparing the average annual ozone levels in several major cities, it was revealed that population size and traffic are not the only factors that affect ozone concentration.

The spatial trends that we have seen cannot be solely explained by temperature and sunlight exposure. Other elements such as meteorological factors (e.g., wind speed and direction, and air humidity) may contribute as well. For example, researchers established a positive correlation between ozone concentration, wind speed, and temperature, and a negative correlation between ozone concentration and relative humidity.[10,11] Japanese researchers examined the increase of annual ozone in the western parts of Japan, despite the decrease in the ambient levels of NOx and organic compounds—precursors of photochemical reactions—from 1990 to 2010. They concluded that the increase in ozone was likely due to transboundary transport from the Asian continent during this time period.[12] The meteorological factors could explain why there is a higher ozone concentration in the Midwest compared to the east coast in June, as windy conditions enhance mixing between the high ozone in higher altitude atmospheres and the low ozone on the ground, leading to a higher ground ozone concentration.

Finally, spatial covariance of greater than 0.25 is observed when the distance between data points is less than 500 miles. The temporal covariance is only influenced by 3–4 lags. The model's statistical quality is generally satisfactory, suggesting its potential for predicting ozone concentration on uniform longitude-latitude grids in areas where no measurements are present. Additionally, the proposed method can be applied to predict the temporospatial structure of large datasets in other fields than atmospheric research. For example, it can be used to predict the temporospatial distribution of the unemployment rate and median household income, both of which are important economic indicators.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation. The authors would like to thank Prof. Emily Kang of the University of Cincinnati for insightful discussions.

REFERENCES

1. Nuvolone, D., Petri, D., & Voller, F. (2018). The effects of ozone on human health. *Environmental Science and Pollution Research, 25,* 8074–8088.
2. McKee, D. (1993). Tropospheric ozone: Human health and agricultural impacts. Lewis Publishers.

3. https://www.epa.gov/criteria-air-pollutants/naaqs-table
4. https://www.image.ucar.edu/Data/Ozmax/
5. Ma, P., Konomi, B. A., & Kang, E. L. (2019). An additive approximate Gaussian process model for large spatio-temporal data. *Environmetrics*, *30*(8). https://doi.org/10.1002/env.2569
6. Xu, S., Cui, C., Shan, M., Liu, Y., Qiao, Z., Chen, L., Ma, Z., Zhang, H., Gao, S., & Sun, Y. (2022). Spatio-temporal prediction of ground-level ozone concentration based on Bayesian maximum entropy by combining monitoring and satellite data. *Atmosphere*, *13*, 1568. https://doi.org/10.3390/atmos13101568
7. Ma, R., Ban, J., Wang, Q., & Li, T. (2020). Statistical spatial-temporal modeling of ambient ozone exposure for environmental epidemiology studies: A review. *Science of the Total Environment, 701*, 134463.
8. https://www.iqair.com/us/usa/indiana
9. Allu, S. K., Srinivasan, S., Maddala, R. K., Reddy, A., & Anupoju, G. R. (2020). Seasonal ground level ozone prediction using multiple linear regression (MLR) model. *Modeling Earth Systems and Environment*, *6*, 1981–1989.
10. Dueñas, C., Fernández, M. C., Cañete, S., Carretero, J., & Liger, E. (2002). Assessment of ozone variations and meteorological effects in an urban area in the Mediterranean coast. *Science of the Total Environment*, *299*, 97–113.
11. Toh, Y. Y., Lim, S. F., & Glasow, R. V. (2013). The influence of meteorological factors and biomass burning on surface ozone concentrations at Tanah rata. *Atmospheric Environment, 70*, 435–446, 2013.
12. Akimoto, H., Mori, Y., Sasaki, K., Nakanishi, H., Ohizumi, T., & Itano, Y. (2015). Analysis of monitoring data of ground-level ozone in Japan for long-term trend during 1990–2010: Causes of temporal and spatial variation. *Atmospheric Environment*, *102*, 302–310.