# Spatiotemporal Analysis of COVID-19 Incidence in High-Prevalent Counties in the United States

## Introduction and Data Description

This report presents a comprehensive spatiotemporal analysis of COVID-19 cases in high-prevalent counties across the United States up to October 31, 2020. The study utilizes a dataset that incorporates daily meteorological variables alongside confirmed cases and deaths due to COVID-19, as reported by reliable sources such as the New York Times COVID GitHub, Weather Underground, and the American Community Survey (Chien, Chen, & Lin, 2022) (<https://www.kaggle.com/datasets/jojochien/meteorological-measures-and-covid19-cases>). The data spans from March to October 2020, covering a critical period when the pandemic evolved through several stages of spread and containment efforts.

The motivation for this study stems from the necessity to understand the dynamics of COVID-19 spread in relation to meteorological factors, which have been suggested to influence virus transmission and mortality rates (Sajadi et al., 2020; Ma et al., 2020). Previous research has explored various aspects of the pandemic, such as the effects of non-pharmaceutical interventions (NPIs) on case numbers (Hsiang et al., 2020), the role of environmental factors in virus survival and transmission (Bashir et al., 2020), and the application of machine learning models for predicting case trends (Jiang et al., 2020).

The dataset under analysis comprises 16 meteorological variables and COVID-19 cases (confirmed and deaths) across 203 high-prevalent counties. The dataset’s granularity allows for detailed spatiotemporal analysis, facilitating the identification of patterns and correlations between meteorological factors and the incidence of COVID-19.

The data processing and cleaning part began with the use of Python’s Pandas library to inspect and understand the dataset’s structure. The initial examination revealed a total of 49985 entries and 37 columns, encompassing both meteorological and epidemiological data, the structure is shown in Table 1. Each record indicates the number of cases and deaths in a certain county on a certain date, and the geographical coordinates are the coordinates of the corresponding county.

Before simple visualization, necessary data processing and cleaning processes are performed. The data set is generally complete. There are some null values in the Environmental Factors section. I use the data of the last two days to fill in the missing values. Additionally, the date format is adjusted to the standard datatime format.

Table 1 Data structure table

|  |  |
| --- | --- |
| Category | Details |
| Epidemiological Data | Cases, Deaths, Rank (Risk ranking based on number of cases and deaths) |
| Demographic Information | Age, Gender (% Male), Race (% White, % Black, % Hispanic) |
| Socio-Economic Indicators | Poverty (%), No Insurance (%), Education Level (% High School) |
| Environmental Factors | Temperature (Max, Avg, Min), Dew Point (Max, Avg, Min), Relative Humidity (Max, Avg, Min), Wind Speed (Max, Avg, Min), Sea Level Pressure (Max, Avg, Min), Daily Precipitation |
| Public Health Interventions | Stay-at-home Orders |
| Location and Time | FIPS Code, County, State, Date, Day of the Week (Dow) |

After that, I used Python's Matplotlib and Seaborn libraries for simple data visualization, converting complex data sets into interpretable graphical formats. It begins by plotting a line chart on the entire data set to illustrate trends in daily COVID-19 cases and deaths, showing the trajectory of the epidemic over time. Histograms are used to represent the distribution of meteorological variables such as mean temperature and relative humidity, providing a simple understanding of environmental conditions during the pandemic. At the same time, plotting pairwise plots and correlation heat maps can investigate potential correlations between socioeconomic indicators or environmental variables and COVID-19 outcomes (cases, deaths).

The expected goal of this task is to explore the role of meteorological and social factors in the spread of the COVID-19 epidemic through some methods of spatiotemporal data mining. Finally, a model can be built to predict cases, the pros and cons of different models can be compared, and certain explanations can be given.

## Exploratory Spatiotemporal Data Analysis

The key to this study lies in the exploratory spatiotemporal data analysis (ESTDA), which aims to uncover the underlying patterns within the COVID-19 and meteorological data. This provides a reference for selecting variables for subsequent establishment of prediction models. ESTDA is pivotal in identifying clusters, anomalies, trends, and relationships that are not immediately apparent (Liu et al., 2023). This section delineates the methodologies employed and the insights derived from the ESTDA conducted on the provided dataset.

## Spital Analysis

The spatial analysis began with the generation of point geometries using longitude and latitude data, facilitated by the Geopandas library. The spatial distribution of COVID-19 cases on four key dates was visualized through GIS technology: March 31(Outbreak period), June 30 (secondary outbreak), September 30, and October 31, 2020 (plateau), to demonstrate the spatiotemporal trends of cases. The spread of the epidemic can be seen to spread over time, from specific epicenters to a wider national distribution. In addition, combined with the overall heat map, it can be seen that New York, San Francisco, Chicago, and some areas of Florida were the most severely affected. Shown in Figure 1.

The result of spatial autocorrelation yielded a Moran index of 0.67 at the confidence level of P=0.001, shown in Table 2, indicating that there is a positive correlation between epidemic distribution and space. However, since the time series is not considered here, the overall number of cases is used for statistics. The uneven temporal distribution of cases may have affected the correlation.

Figure 1 Time slice case distribution map and global heat map.

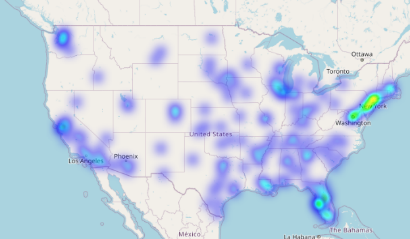
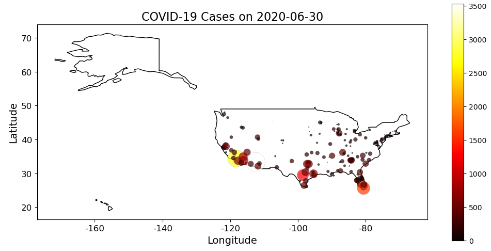
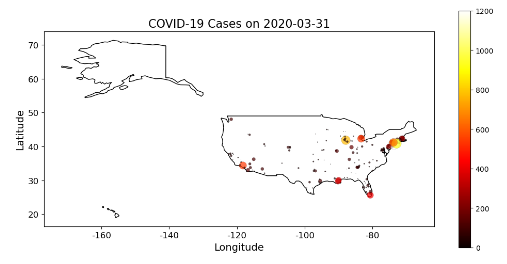


Table 2 Spatial autocorrelation results

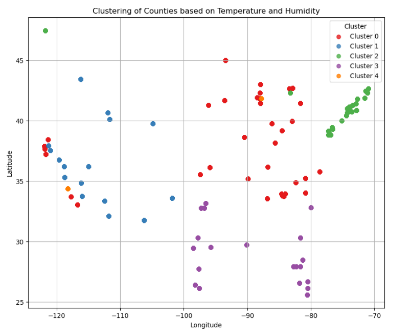
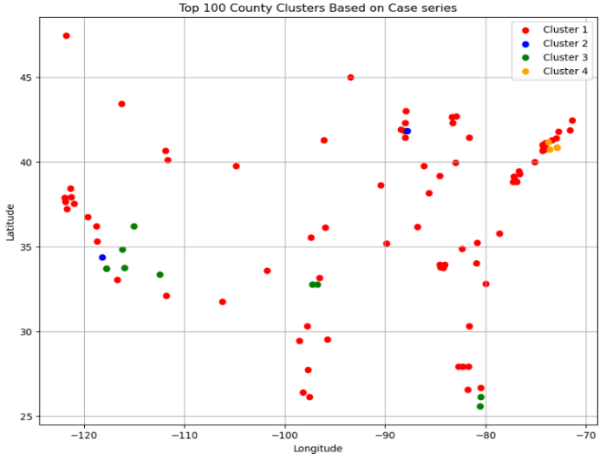
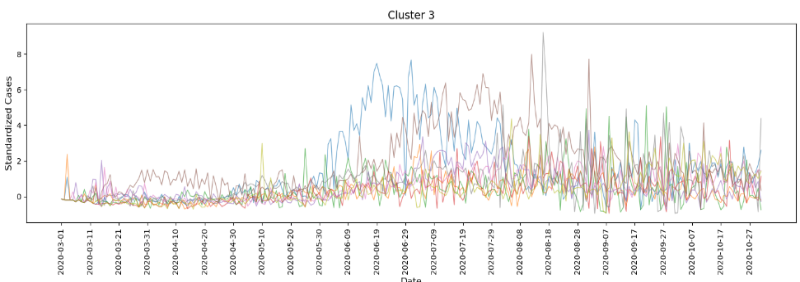
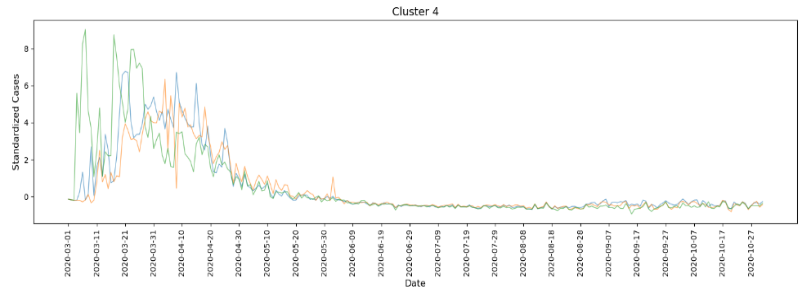
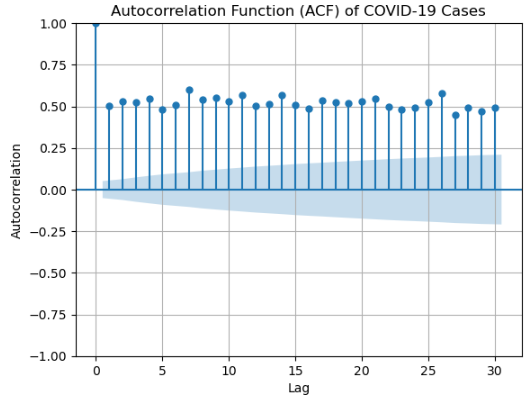
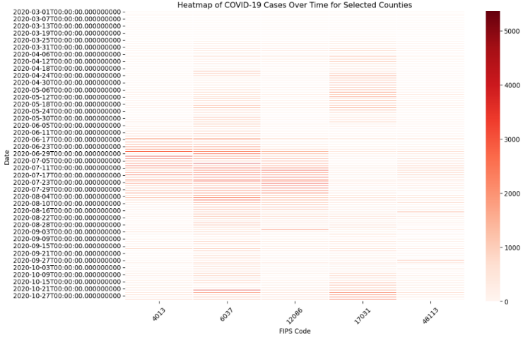
|  |  |
| --- | --- |
| Statistics | Spatial Autocorrelation |
| Moran's I | 0.6738 |
| P-Value | 0.001 |

## Temporal Analysis

By establishing the autocorrelation function (ACF) of the number of cases, it can be seen from the ACF graph that the lag term has maintained a stable and high correlation, which may indicate a long-term trend in epidemic cases.

To further delve into the temporal patterns, a subset of the data was used to create a heatmap displaying COVID-19 cases. The heatmap visualization offered a different perspective, showcasing the temporal progression of cases across selected counties (five counties with the top five risks in the data set). The intensity of color in the heatmap corresponded to the case counts, with darker shades signifying higher numbers. Most counties have a significant increase in cases from June to August. Based on this exploratory analysis, I guess that the cases may be related to seasonality, and more deeply, may be related to the increased temperature and humidity in summer. This provides a reference for subsequent variable selection in building prediction models. The results of the timing analysis are shown in Figure 2.

Figure 2 The timing exploratory analysis results diagram (They are the ACF chart of cases, time series heat map, temperature and humidity cluster map, case cluster map and time series cluster line chart.)



(Explanation of clustering: *I defined four clusters, two of which are shown here, each containing multiple time series curves. Each curve represents the standardized number of COVID-19 cases in a county over time. The difference between a single curve and multiple curves in a graph can usually be explained by the following factors:*

*The number of members within a cluster: If a cluster has only one curve, it means that only one county is assigned to this cluster. If there are multiple curves, it means that multiple counties have similar time series patterns assigned to the same cluster.*

*The result of standardization processing: If the standard deviation of certain values in the data is very small or close to zero during the standardization process, these values may become very large after standardization, which may also lead to single or few curves dominating the clustering.)*

The spatiotemporal analysis also involved the application of clustering techniques to discern patterns across different counties. The KMeans clustering algorithm was implemented on a subset of the data, standardizing the cases over time to identify counties with similar COVID-19 case trajectories. This clustering yielded groups of counties with similar temporal trends in case numbers, which could be indicative of similar underlying factors or responses to the pandemic.

Combining the results of time series heat maps, time series clustering, spatial heat maps, and spatial autocorrelation, I initially guessed that the distribution of epidemic cases is related to urban attributes (socioeconomic factors) and time changes (possibly seasonal climate factors or changes in related policies) is relevant. The results of the exploratory analysis make it possible to further explore the data set to build regression and predictive models.

It is worth mentioning that during the ESTDA process, I found that there were many anomalies in the data of many cities, especially small cities, such as an extreme increase in cases on a certain day, so using the entire data set for further analysis was obviously unreliable. Therefore, I decided to combine the results of ESTDA and select large city data subsets in four hotspot areas for modeling analysis. The four area selections are marked with four boxes on the cluster map.

## Methodology and Results

**Methodological Framework:** To analyze the relationship between multiple variables in complex data sets and the development of the epidemic, and to build models to predict the future development of the epidemic. The methodological framework adopted for this analysis involved several stages:

1. **Data Preprocessing:** Data preprocessing was conducted using Python’s Pandas library. The initial step included cleaning the data, handling missing values by imputation, and transforming the date column into a datetime format for temporal analysis. The geographical coordinates were transformed into a GeoDataFrame to facilitate spatial visualization. The data needs to be further preprocessed for different operations. Examples include converting datasets into pivot tables when clustering or using feature extraction to process high-dimensional datasets. Standardization must also be processed first when clustering and building models.
2. **Exploratory Spatiotemporal Data Analysis (ESTDA):** EDA included visualizing the data distribution, identifying outliers, and understanding the temporal trends and spatial distribution of COVID-19 cases. This stage laid the groundwork for subsequent in-depth analysis.
3. **Statistical Analysis:** The study employed statistical methods to evaluate the correlations between meteorological variables and COVID-19 cases. For example, paired plots are used to explore correlations between variables.
4. **Time Series Prediction Model:** The models used include ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal AutoRegressive Integrated Moving Average). These models are used to predict the development of urban cases from a time series level. Because in the EDA part I found that the climate variables in the data set have high temporal correlation, climate factors were also considered in the reforecast model.
5. **Machine Learning Models:** Advanced machine learning models were utilized to identify patterns and predict future trends. Models such as Decision Trees, Random Forest, Support Vector Regression, and Gradient Boosting were trained on the data. Because the data set used covers several hotspot cities, simply using climate variables is insufficient to explain the differences in real cities. Moreover, urban differences (such as social, demographic, and economic data) have a certain impact on the development of the epidemic (Xiao et al., 2023). Therefore, in the machine learning model, in addition to using Environmental Factors, Demographic Information and Socio-Economic Indicators are also introduced for joint analysis.

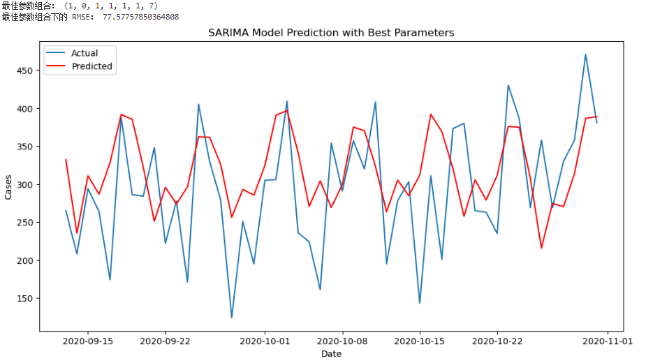
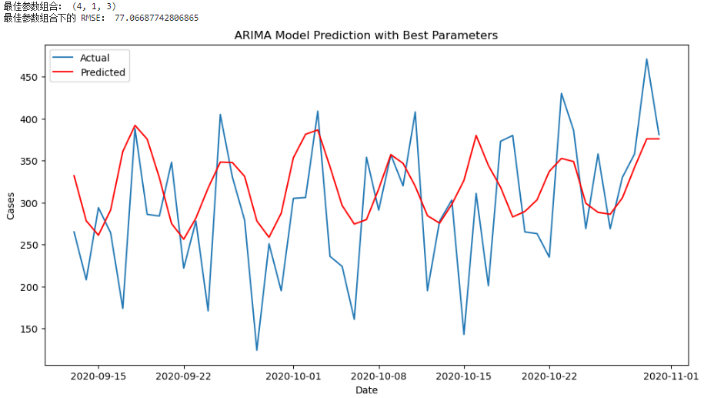
**Experimental Setup:** The dataset was partitioned into training and testing sets, with an 80-20 split, to evaluate the performance of the machine learning models. This distribution ratio ensures sufficient training data and reduces the risk of overfitting.When building a machine learning model, normalization is first applied to the features to standardize the data distribution. This is because the scale of features may have an impact on the performance of the model. If features have different scales or ranges, it may cause some features to be overweighted or underweighted during model training. But for time series models, because there is a differential operation in the model, the difference in scale is reduced, so there is no need to standardize the data set. Parameters in the integrated model method are determined through grid search. Hyperparameters for each model were fine-tuned using cross-validation techniques to ensure optimal performance. In this task, due to time cost, only coarse-fine search-based cross-validation was performed on the random forest model and Gradient Boosting. Parameters used by random forest: {'max\_depth': 11, 'max\_features': 22, 'min\_samples\_leaf': 2, 'min\_samples\_split': 8, 'n\_estimators': 35}, the model score is 0.7331. Parameters used by Gradient Boosting: {'learning\_rate': 0.2, 'max\_depth': 8, 'max\_features': 'log2', 'min\_samples\_leaf': 16, 'min\_samples\_split': 16, 'n\_estimators': 81}, and the model score is 0.7780.

The variables passed in by the model are mentioned in the Framework. Specifically, meteorological factors (Temperature, Dew Point, Relative Humidity, Wind Speed, Sea Level Pressure and Daily Precipitation) are introduced into the time series model. The above parameters are all passed into the model using the average value. For the machine learning model, in addition to the above-mentioned meteorological factors, variables such as Demographic Information and Socio-Economic Indicators are also considered. I created a new data set for building the model from the original data set by outlier processing and selecting hot city agglomerations through the ESTDA process. To a certain extent, it avoids the problems of small cities not striving for statistics and missing data (mainly because the cases in some small cities in the data set have abnormally high values, considering that when statistics are collected, the cases in a certain period are summarized and counted on a certain day. Socio-demographic data are also missing). In addition, the time series clustering labels and geographical hotspot area labels obtained during the ESTDA process are added to the new data set as variables. Descriptive characters like county and state can be encoded first and then passed into the model.

**Performance Evaluation:** Root Mean Squared Error (RMSE) is used as an indicator to evaluate the predictive ability of time series models. In this task, the ARIMA and SARIMA models for predicting cases in San Diego were established. The results of the two models are as follows.

Table 3 Timing model results table

Figure 3 ARIMA and SARIMA model forecast charts.



|  |  |  |
| --- | --- | --- |
| Statistics | RMSE | Cross-validation optimal parameters |
| ARIMA | 77.06 | 4,1,3 |
| SARIMA | 77.58 | 1,0,1,1,1,1,7 |

The performance of machine learning models was assessed using Mean Squared Error (MSE) and R-squared (R2) metrics. These metrics provided a quantitative measure of the models' accuracy and the variance explained by the models, respectively. The Random Forest model demonstrated superior performance with the lowest MSE and highest R2, suggesting its robustness in capturing the complex patterns of the data. The Decision Tree and Gradient Boosting models also showed promising results, with relatively low MSE and satisfactory R2 values. Support Vector Regression, although useful in certain contexts, did not perform as well as the ensemble methods in this analysis. The performance metrics for each model were as follows:

* **Random Forest Regressor:** MSE = 13276.05, R2 = 0.81
* **Gradient Boosting Regressor:** MSE = 13932.50, R2 = 0.80
* **Decision Tree Regressor:** MSE = 30725.60, R2 = 0.56
* **Support Vector Regression:** MSE = 63047.31, R2 = 0.10

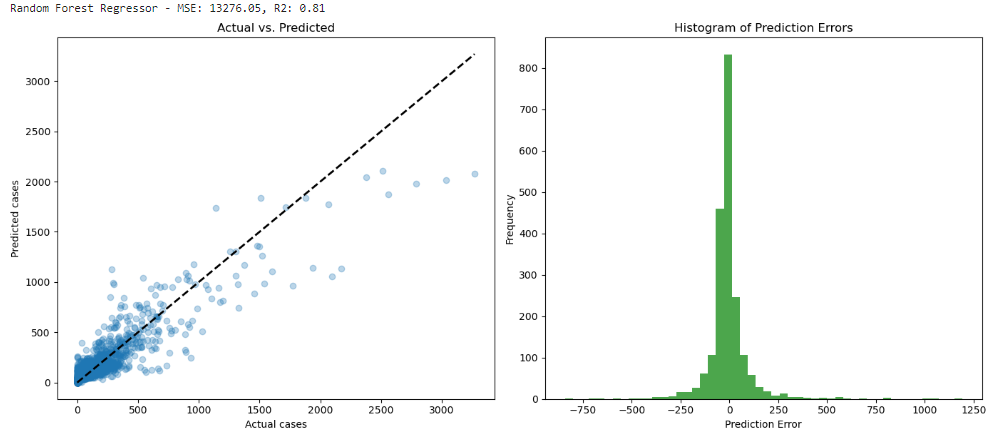
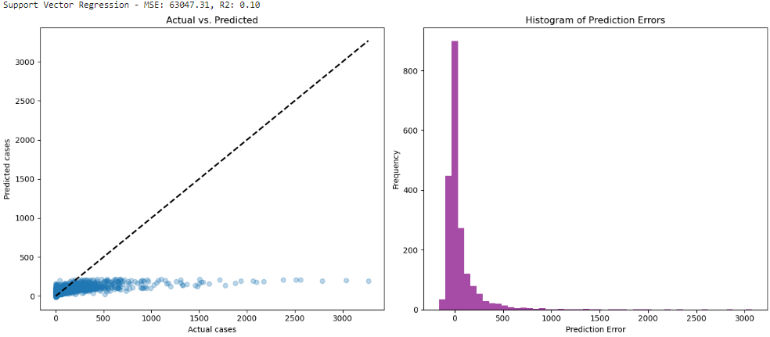


Figure 4 Random Forest and support vector model results plot.

## Discussion and Conclusions

The models employed in this study to analyze the spread of COVID-19 in relation to meteorological variables yielded intriguing insights, which are discussed in detail below.

**Model Performance Comparison:** In machine learning models, the Random Forest Regressor emerged as the top-performing model, exhibiting the lowest Mean Squared Error (MSE) and highest R-squared (R2) values. Its success can be attributed to its ability to handle non-linear relationships and interactions between variables effectively. Random Forest, being an ensemble method, benefits from the aggregation of decisions from multiple decision trees, reducing the risk of overfitting and improving generalization to unseen data.

In contrast, the Support Vector Regression (SVR) model displayed the poorest performance among the tested models. The likely reason for this is SVR's sensitivity to high-dimensional data and its tendency to underperform when the number of features is much greater than the number of observations, as is the case with our complex dataset. The Gradient Boosting Regressor and Decision Tree models showed promising results but did not match the performance of the Random Forest. This could be due to Gradient Boosting's sequential approach, where each tree attempts to correct the errors of the previous one, which can be prone to overfitting if not carefully tuned. Decision Trees are simple to interpret but can become complex and less interpretable as the depth increases, and they are also prone to overfitting unless pruned appropriately.

For time series modeling, because cases have obvious cyclicality, the forecast results of SARIMA may be slightly better than ARIMA because it achieves the prediction of some peaks and troughs.

In terms of time cost, since the data set is not large, there is not much difference between the models when the parameters are known. Considering the iterative method, Gradient Boosting will be slightly slower than other models. Gradient Boosting and SVR are the slowest when cross-validation are needed to determine parameters. Random forests and decision trees are relatively fast.

**Advantages in Interpretability and Implementation:** Random Forest and Decision Trees inherently offer interpretability through the hierarchy of features in their branching structure. However, as the complexity of a Random Forest model increases, its interpretability may decrease compared to a single Decision Tree, which can often be visualized and understood even by non-experts.

For time series modeling, a SARIMA model may be more complex than a simple ARIMA model, with more parameters, so more data may be required to fit the model, and it may be more difficult to interpret.

**Limitations and Improvement of the Methodology:** The limitations of the study arise primarily from the nature of the data and the chosen methods. The dataset, while comprehensive, may still omit factors that significantly influence the spread of COVID-19, such as policy interventions, public behavior, and healthcare capacity. Moreover, the assumption of linearity in regression models may not capture the true complexity of the disease spread. The agent-based simulation method may be more suitable for analyzing the development of the epidemic. At the national level, counties can be considered as agents. However, the lack of connections between counties in the data set makes this method difficult to apply in the current data set. Further improvements can be made by combining social media data and traffic connection data between cities to establish Agent based simulation. To conduct in-depth analysis of the spread of the epidemic:

Recommendations for Methodological Improvement: To enhance the methodology, several steps could be taken:

* Incorporating additional data sources, such as mobility data and social media trends, could provide more explanatory variables and improve model accuracy.
* Employing cross-validation techniques more extensively would help in better hyperparameter tuning and model selection. But accordingly, the time cost will increase. For Random Forest and Gradient Boosting, which performed better in this task, both models may benefit from further hyperparameter tuning, feature engineering and incorporating more diverse data sources. For gradient boosting, how you adjust the learning rate and depth of the tree is critical to balancing bias and variance. For random forests, exploring the optimal number of trees and depth can help achieve the best performance.
* Deep learning methods could be explored, which may capture complex nonlinear relationships better.

## References

* Chien, L. C., Chen, L. W. A., & Lin, R. T. (2022). Lagged meteorological impacts on COVID-19 incidence among high-risk counties in the United States—a spatiotemporal analysis. *Journal of exposure science & environmental epidemiology*, *32*(5), 774-781.
* Sajadi, M. M., Habibzadeh, P., Vintzileos, A., Shokouhi, S., Miralles-Wilhelm, F., & Amoroso, A. (2020). Temperature, humidity, and latitude analysis to predict potential spread and seasonality for COVID-19. *Social Science Research Network*.
* Ma, Y., Zhao, Y., Liu, J., He, X., Wang, B., Fu, S., ... & Luo, B. (2020). Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. *Science of the total environment*, *724*, 138226.
* Pairo-Castineira, E., Clohisey, S., Klaric, L., Bretherick, A. D., Rawlik, K., Pasko, D., ... & Princess of Wales Hospital, Llantrisant, UK Sathe S. 133 Davies E. 133 Roche L. 133. (2021). Genetic mechanisms of critical illness in COVID-19. *Nature*, *591*(7848), 92-98.
* Bashir, M. F., Ma, B., Komal, B., Bashir, M. A., Tan, D., & Bashir, M. (2020). Correlation between climate indicators and COVID-19 pandemic in New York, USA. *Science of the Total Environment*, *728*, 138835.
* Jiang, P., Fu, X., Van Fan, Y., Klemeš, J. J., Chen, P., Ma, S., & Zhang, W. (2021). Spatial-temporal potential exposure risk analytics and urban sustainability impacts related to COVID-19 mitigation: A perspective from car mobility behaviour. *Journal of cleaner production*, *279*, 123673.
* Liu, X., Jin, X., Luo, X., & Zhou, Y. (2023). Quantifying the spatiotemporal dynamics and impact factors of China's county-level carbon emissions using ESTDA and spatial econometric models. *Journal of Cleaner Production*, *410*, 137203.
* Xingyu, X. I. A. O., Shiyu, M. E. I., Ruiqi, L. I. U., Kuo, W. A. N. G., Qing, D. E. N. G., Lida, H. U. A. N. G., & Feng, Y. U. (2023). Impact of the COVID-19 pandemic on the development of core cities in the Beijing-Tianjin-Hebei urban agglomeration—Taken Beijing, Tianjin, and Shijiazhuang as examples. *Journal of Tsinghua University (Science and Technology)*, *63*(6), 994-1002.

**Description of the format:** 11-point Arial font is used for the main text, and 9-point Arial font is used for figure titles, labels, quotations, and additional explanations. The citation format is unified into APA format.

**Description of the picture:** Due to space limitations, some data visualization pictures are not shown in the first part. Similarly, some model results are not shown in the third part. These images can be viewed in the code notebook.

**Description of the code:** All code functions are implemented using PYTHON. From data visualization, ESTDA to the final model code process and results are covered in STDM\_Code\_HRGZ4.ipynb. Proofs about parts of the code that can be implemented. Because the notebook file will retain the running order and results, I cleared all the results and ran it completely. The running sequence and results in the notebook shows that the entire code can run smoothly. In addition, regarding the instructions for running the code, if you want to re-run the code, you only need to ensure that the STDM\_Code\_HRGZ4.ipynb file and the data set covid\_weather\_dlnm.csv are in the same path to run.