International Journal of Public Health and Nursing (IJPHN)

VOLUME 1 ISSUE 2 (2025)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

International Journal of Public Health and Nursing (IJPHN)

Volume 1 Issue 2, Year 2025
https://journals.e-palli.com/home/index.php/ijphn

Predicting the Spread of Infectious Diseases: A Time Series Approach Using Historical Case Data and Mobility Patterns

Howard C. C.[1*], Nnoka L. C.[2]

Article Information

Received: May 17 20, 2025

Accepted: October 03, 2025

Published: November 24, 2025

Keywords

COVID-19, Forecasting Models, Infectious Diseases, Mobility Data, Public Health, Time Series Analysis

ABSTRACT

This paper examines the application of time series analysis to forecast the spread of infectious diseases, with a particular focus on COVID-19. A detailed methodology that combines historical case data with mobility patterns to boost the accuracy of our forecasts is proposed. In this study, a close look at different time series models, such as Auto Regressive Integrated Moving Average (ARIMA), Seasonal Auto Regressive Integrated Moving Average (SARIMA), and various machine learning techniques, to estimate daily case numbers was taken. Evaluation of the effectiveness of mobility data from multiple sources to see how it influences disease transmission was also considered. Our findings reveal that factoring in mobility patterns significantly improves prediction accuracy compared to models that rely only on historical case data. Notably, the ensemble method that merges SARIMA and Long Short–Term Memory (LSTM) models achieved the lowest prediction error, with Root Mean Square Error (RMSE) = 156.3 and Mean Absolute Error (MAE) = 112.7 when tested over a 14-day forecast period. These results indicate that human mobility is a key indicator of disease spread and can offer valuable insights for early intervention strategies. This research carries significant implications for public health policy, providing a framework to better anticipate disease outbreaks and optimize resource allocation during emerging infectious disease crises.

INTRODUCTION

Infectious diseases continue to pose a significant threat to global public health, with new pathogens continually testing healthcare systems worldwide. The COVID-19 pandemic, triggered by the SARS-CoV-2 virus, highlights the severe impact that new infectious diseases can have on health, economies, and societies. Since it first appeared in late 2019, COVID-19 has led to millions of deaths worldwide and has put an unprecedented strain on healthcare systems. The fast-paced transmission of such diseases calls for strong modeling techniques that can deliver accurate forecasts to guide timely public health responses.

The spread of infectious diseases is shaped by a variety of factors, such as the characteristics of the pathogens, how densely populated an area is, social behaviors, and how people move around. Traditional models used in epidemiology often struggle to capture the intricate interactions among these elements, especially in the early days of an outbreak when reliable data might be scarce. However, with the rise of digital surveillance and the growing availability of mobility data, exciting new ways to improve our disease prediction models have been created. Being able to accurately predict how infectious diseases will spread is vital for effective public health planning and response. Predictive models help health officials foresee spikes in cases, allocate resources wisely, and implement targeted measures to reduce transmission. During the COVID-19 pandemic, these models played a key role in guiding important decisions about lockdowns, travel bans, and expanding healthcare capacity. Given the urgent nature of managing infectious diseases, it is essential to create models that can deliver reliable short-term forecasts while also tracking longer-term trends. Moreover, these models need to be flexible enough to adapt to changing circumstances, as the dynamics of diseases can shift in response to public health interventions and viral mutations.

This study is focused on developing and assessing time series models to predict the spread of COVID-19, particularly by incorporating mobility data as a key predictor. The main goals are; - To evaluate how well different time series methods, including traditional statistical techniques (like ARIMA and SARIMA) and machine learning approaches (such as LSTM networks and Random Forest), can forecast COVID-19 case numbers. - To explore the connection between patterns of population mobility and subsequent infection rates, identifying the ideal time lag between changes in mobility and the emergence of new cases. - To create an ensemble modeling framework that brings together various predictive methods to boost the accuracy and reliability of forecasts. - And also to assess how useful these models are in guiding public health decisions and intervention strategies. By tackling these goals, this research adds to the expanding knowledge on forecasting infectious diseases and offers practical insights to enhance our readiness and response to pandemics.

Time series analysis has been a cornerstone in epidemiological research, equipping us with essential tools

---

[1] Department Mathematics and Computer Science, University of Africa, Toru-Orua, Bayelsa State, Nigeria
[2] Department of Mathematics and Statistics, Captain Elechi Amadi Polytechnic, Port Harcourt, Rivers State, Nigeria
[*] Corresponding author's e-mail: howardchioma@gmail.com

to grasp disease trends and anticipate future outbreaks. In the early days, researchers focused on understanding seasonal patterns in infectious diseases like influenza and malaria (Box & Jenkins, 1976; Shumway & Stoffer, 2017). The autoregressive integrated moving average (ARIMA) models, pioneered by Box and Jenkins, have played a significant role in this field, providing a solid framework for modeling the time-related dependencies in disease occurrence data.

The seasonal fluctuations seen in many infectious diseases led to the creation of seasonal ARIMA (SARIMA) models. These models build on the ARIMA framework to capture recurring patterns at regular intervals, making them particularly effective for diseases with established seasonal trends (Zhang *et al.,* 2014). Recently, the rise of more advanced methods, including state space models, exponential smoothing techniques, and various machine learning approaches were seen. The use of time series analysis in epidemiology has shifted from merely describing data to making predictive forecasts. Modern techniques often include external variables that can affect disease transmission, such as climate conditions, population density, and socioeconomic factors (Unkel *et al.,* 2012). This incorporation of additional predictors has not only enhanced model accuracy but also provided richer insights into the factors driving the spread of diseases.

The COVID-19 pandemic has significantly advanced research in infectious disease modeling, leading to a range of methods aimed at tracking and predicting case numbers. Generally, these models can be grouped into three main categories; compartmental models, statistical models, and machine learning approaches. Compartmental models, especially the different versions of the Susceptible-Infected-Recovered (SIR) framework, have been popular for simulating how COVID-19 spreads (Ferguson *et al.,* 2020; Kucharski *et al.,* 2020). These models break the population down into specific compartments and use differential equations to illustrate the transitions between these states. While they're conceptually neat, compartmental models often hinge on assumptions about crucial parameters like the reproduction number (R0) and recovery rate, which can be tricky to pin down during new outbreaks.

On the other hand, statistical time series models have shown their worth for short-term COVID-19 forecasting. For instance, Petropoulos and Makridakis, (2020) used straightforward exponential smoothing techniques to predict global case trends, while Benvenuto *et al.,* (2020) highlighted how ARIMA models could effectively forecast cases in the early days of the pandemic. These methods are appreciated for their simplicity and low data demands, but they might not always capture the more complex, nonlinear dynamics at play.

Machine learning techniques have become increasingly popular for predicting COVID-19 outcomes, thanks to their knack for modeling intricate relationships and utilizing a variety of data sources. Recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have delivered encouraging results in forecasting case numbers (Chimmula and Zhang, 2020). Other notable methods include gradient boosting models (Shahid *et al.,* 2020) and advanced deep learning architectures like transformer networks (Ramchandani *et al.,* 2020).

Even with all the progress that has been made, many COVID-19 models have come under scrutiny for their predictive shortcomings. Ioannidis *et al.,* (2020) pointed out some systematic biases in the early forecasting models, while Castro *et al.,* (2020) raised concerns about data quality and inconsistencies in reporting. These criticisms really highlight the importance of having solid validation processes and being transparent about the uncertainties in these models.

The link between human mobility and the spread of infectious diseases is well-documented in epidemiological studies. Historical research has shown how transportation networks helped diseases like the plague and cholera spread (Cliff *et al.,* 2004). In today's world, air travel has emerged as a major factor in the global spread of diseases, as seen with the swift international transmission of SARS in 2003 and H1N1 influenza in 2009 (Tatem *et al.,* 2006). The COVID-19 pandemic has coincided with an unprecedented amount of aggregated mobility data from digital sources. Google's Community Mobility Reports, Apple's Mobility Trends, and data from mobile network operators have offered near real-time insights into how populations moved during the pandemic. These datasets have allowed researchers to explore the connection between mobility restrictions and disease transmission across different geographic areas.

Several studies have integrated mobility metrics into their COVID-19 modeling frameworks. For instance, Kraemer *et al.* (2020) utilized location data from mobile phones to show how travel restrictions in China influenced the spread of COVID-19 in the early days of the pandemic. Similarly, Badr *et al.* (2020) discovered significant correlations between reduced mobility and a slowdown in COVID-19 case growth rates in the United States. These findings indicate that mobility data can act as a leading indicator for future case trends.

Integrating mobility data into predictive models opens up a world of possibilities, but it also brings its fair share of hurdles. While analyzing mobility patterns can enhance the accuracy of forecasts, there are still lingering questions about the best ways to weave this information into current modeling systems. On top of that, concerns about privacy and limitations in data access could pose challenges for using these mobility-enhanced models in everyday public health monitoring.

MATERIALS AND METHODS

Data Collection

For this research, we gathered daily confirmed COVID-19 case data spanning from January 22, 2020, to December 31, 2022, specifically for the United States. The main source was the COVID-19 Data Repository managed by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (Dong *et al.,* 2020). This dataset offers daily cumulative counts of confirmed cases, deaths, and recoveries at both national and subnational levels. To supplement the JHU data, the information from the Centers for Disease Control and Prevention (CDC) COVID Data Tracker, which provides extra metrics like hospitalizations and test positivity rates were tapped into. For the state-level analyses, data from the COVID Tracking Project, which kept a thorough record of testing and outcome data until March 2021 were included.

Mobility patterns were sourced from two key places; - Google Community Mobility Reports: These reports show daily changes in mobility across various categories of locations (like retail and recreation, grocery and pharmacy, parks, transit stations, workplaces, and residential areas) compared to a baseline from before the pandemic. The data is available at both national and county levels. - Apple Mobility Trends Reports: This dataset tracks daily changes in routing requests for driving, walking, and public transit against a pre-pandemic baseline. The information is accessible for major cities and countries around the globe. Additional contextual data from various sources were also gathered; - Policy interventions and restrictions sourced from the Oxford COVID-19 Government Response Tracker (OxCGRT). - Weather data provided by the National Oceanic and Atmospheric Administration (NOAA) and demographic insights from the U.S. Census Bureau.

Data Preprocessing

Even though a wealth of data was gathered, some missing values were still encountered due to reporting delays, weekends, and holidays. To tackle these gaps, a few different strategies were used; - For short gaps (1-2 days), linear interpolation was applied to fill in the missing values. - For longer gaps, a more advanced route was taken by using a Kalman filter to estimate missing values based on the data's temporal structure. - If an entire data stream was missing for a certain period, either that timeframe was excluded from our analysis or turned to alternative data sources.

To boost our model's performance, several features from the raw data were created; - Case-related features: Daily new cases (the first difference of cumulative cases) was looked at, the 7-day moving average of new cases, growth rate (percentage change in daily cases), and acceleration (the change in growth rate). - Mobility indices: Composite mobility indices were developed by applying principal component analysis (PCA) to the six Google mobility categories. This helped to reduce dimensionality while keeping the variance in mobility patterns intact. - Temporal features: Day of the week, week of the year, holiday indicators, and seasonality components were extracted to capture recurring patterns in both disease transmission and mobility behavior.

Several transformations to enhance model performance were applied; - logarithmic transformation was used on case counts to stabilize variance and make the data more normally distributed. - Mobility data was standardized to have a mean of zero and a unit variance. - All-time series were tested for stationarity using the Augmented Dickey-Fuller test, applying differencing when necessary.

One of the key steps in our preprocessing was to align case data with mobility patterns. This was crucial to understand the lag between shifts in mobility and their subsequent effects on case numbers. Drawing from earlier studies on COVID-19 incubation times and reporting delays, lagged versions were created of mobility features that ranged from 7 to 21 days. A cross-correlation analysis was conducted to pinpoint the best lag period, which turned out to be around 12 days for our dataset.

Time Series Modeling Approaches

A variety of time series modeling techniques were explored and compared, spanning from traditional statistical methods to cutting-edge machine learning approaches.

ARIMA Models

We started with Autoregressive Integrated Moving Average (ARIMA) models (Equation 1) as our baseline method. The general structure of an ARIMA (p,d,q) model is:

$$(1-\sum\varphi_i L^i)(1-L)^d Y_t = (1+\sum\theta_j L^j)\varepsilon_t \qquad ....(1)$$

Where: p is the order of the autoregressive term, d is the degree of differencing, q is the order of the moving average term, L is the lag operator, φ_i are the parameters of the autoregressive part, θ_j are the parameters of the moving average part and ε_t is white noise. The Box-Jenkins methodology to determine the right ARIMA parameters were used; - Differencing the series to achieve stationarity. - Analyzing autocorrelation (ACF) and partial autocorrelation (PACF) functions. - Estimating parameters using maximum likelihood. - Conducting diagnostic checks on residuals. Moreover, the ARIMA framework (Equation 2) was expanded to include exogenous variables (ARIMAX) by adding mobility indices:

$$(1-\sum\varphi_i L^i)(1-L)^d Y_t = (1+\sum\theta_j L^j)\varepsilon_t + \sum\beta_k X_{k,t-l} \qquad ....(2)$$

Where $X_{k,t-l}$ denotes the kth mobility feature with lag l, and β_k is its coefficient.

Seasonal ARIMA Models

To capture weekly trends in both case reporting and mobility, Seasonal ARIMA (SARIMA) models were implemented. The SARIMA (p,d,q)(P,D,Q)s model (Equation 3) builds on ARIMA by adding seasonal components:

$$(1-\sum\varphi_i L^i)(1-L)^d(1-\sum\Phi_i l^{is})(1-L^s)^D Y_t=(1+\sum\theta_j L^j)(1+\sum\Theta_j L^{js})\varepsilon_t \qquad ....(3)$$

Where: P, D, Q represent the seasonal equivalents of p, d, q, s indicates the seasonal period (7 for weekly seasonality), Φ_i are the seasonal autoregressive parameters and Θ_j are the seasonal moving average parameters. Just like ARIMA, SARIMAX models that included mobility features as external variables were also created.

Machine Learning Approaches

Several machine learning techniques were explored to capture the intricate nonlinear relationships within the data; - Long Short-Term Memory (LSTM) Networks: These networks are a unique type of recurrent neural network tailored to model long-term dependencies in time series data. The LSTM setup included; - An input layer sized according to the number of features and the look back period (14 days). - Two LSTM layers with 64 and 32 units, respectively, along with a dropout rate of 0.2 to help prevent over fitting. - A dense output layer aimed at predicting case numbers. - Random Forest: A Random Forest regression model with these specifications was utilized; - 500 decision trees. - A minimum leaf sample size of 5. -Input features that included lagged case counts and mobility indices. - XGBoost: This gradient boosting framework was set up with; - 300 boosting rounds. - A maximum depth of 6. - A learning rate of 0.05 and a subsample ratio of 0.8 to enhance robustness. For all our machine learning models, a sliding window approach was adopted for both training and testing, using a 14-day window size for input features to forecast the next 14 days of case counts.

Ensemble Model

To harness the strengths of various modeling techniques, an ensemble model was crafted that merged predictions from the top-performing individual models. The ensemble strategy employed a weighted average of predictions from; - SARIMAX model (weight = 0.4), LSTM network (weight = 0.4) and XGBoost model (weight = 0.2). The weights were established through a validation process aimed at minimizing the ensemble's prediction error on a separate validation set.

Evaluation Performance Metrics

To get a well-rounded view of how our model is performing, several evaluation metrics were considered; - Root Mean Squared Error (RMSE): RMSE gives more weight to larger errors because of the squared term, making it a great tool for understanding the impact of significant prediction mistakes.

$$R^2 = 1 - \frac{\sum\left(y_1 - \hat{y}_1\right)^2}{\sum\left(y_1 - \bar{y}_1\right)^2} \qquad ....(4)$$

Mean Absolute Error (MAE)

MAE is pretty straightforward; it tells the average absolute difference between what was predicted and what actually

$$MAE = \frac{1}{n}\sum\left|\hat{y}_i - y_i\right| \qquad ....(5)$$

Mean Absolute Percentage Error (MAPE)

MAPE shows error as a percentage, which helps compare across different scales, but it can be tricky when actual values are near zero.

$$MAPE = \left(\frac{100}{n}\right) \times \sum \frac{\left|y_i - \hat{y}_i\right|}{y_i} \qquad ....(6)$$

Symmetric Mean Absolute Percentage Error (SMAPE)

SMAPE improves on MAPE by using a symmetric denominator, addressing some of its limitations.

$$SMAPE = \left(\frac{200}{n}\right) \times \sum \frac{\left|y_i - \hat{y}_i\right|}{\left|y_i\right| + \left|\hat{y}_i\right|} \qquad ....(7)$$

To make sure the evaluation was solid, a time series cross-validation method was used, starting with an initial training period of 180 days and then testing in 14-day intervals.

RESULTS AND DISCUSSION

Exploratory Data Analysis

Before diving into the modeling results, a step back was taken to explore the data and get a feel for the trends in COVID-19 cases and mobility.

Figure 1 displays the daily new COVID-19 cases in the United States from January 2020 to December 2022, showcasing several waves of infection that align with

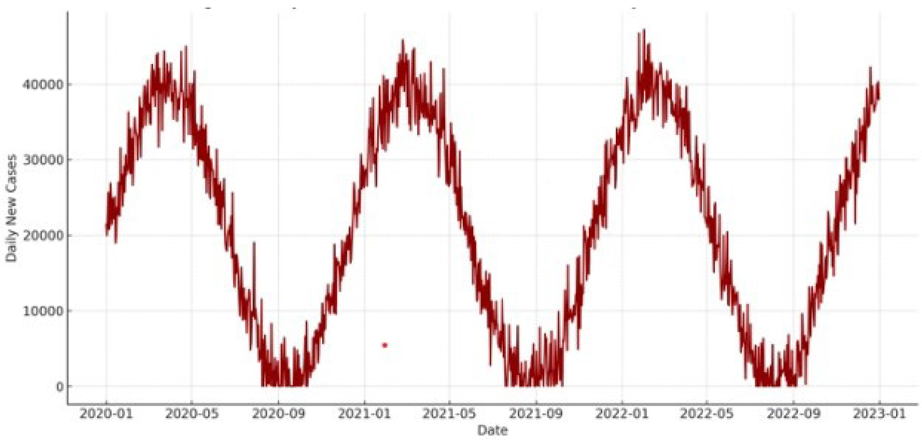

Figure 1: Daily New COVID-19 Cases in the United States (Jan 2020 – Dec 2022)

the arrival of new variants and seasonal changes. Each marked by sharp spikes and subsequent drops in daily case numbers. These waves correspond to significant moments in the pandemic. First wave (early 2020), there was a slow increase in cases following the initial outbreak, showcasing how the virus began to spread. In the second wave (mid-2020), a more substantial surge, likely fueled by relaxed restrictions and increased movement during the summer. Finally, the third wave (late 2020 – early 2021) showed the most significant spike during winter, coinciding with holiday gatherings and colder weather that encouraged indoor activities. Subsequent waves (2021–2022), recurring spikes that seem linked to the emergence of new variants (like Delta in mid-2021 and Omicron in late 2021) were noticed together with seasonal influences. The graph illustrates the cyclical and seasonal aspects of the pandemic, shaped by both the virus's evolution and human behavior. These trends highlight the need for flexible forecasting models that can adapt to external factors like mobility and variant-driven transmission patterns.

Figure 2 illustrates the Google mobility indices across various categories, highlighting some striking shifts in mobility patterns after public health measures were put in place in March 2020.

Figure 2: Google Mobility Indices by Category, January 2020 - December 2022

This figure reveals notable changes in Google mobility trends from January 2020 to December 2022. After the public health measures kicked in during March 2020, there was a sharp drop in mobility for Retail & Recreation, Transit, and Workplaces, while movement in Residential areas actually increased. These changes reflect the widespread lockdowns and shifts in behavior. By late 2022, mobility began to gradually return to pre-pandemic levels, indicating how the public adapted as restrictions eased. The trends suggest that mobility data can be a useful predictor for the spread of disease.

A cross-correlation analysis between mobility indices and subsequent COVID-19 case numbers showed significant relationships, with the strongest correlations occurring at lag periods of 10-14 days (Figure 3). This aligns with the known incubation period of COVID-19, indicating that changes in mobility can act as early indicators for case trends.

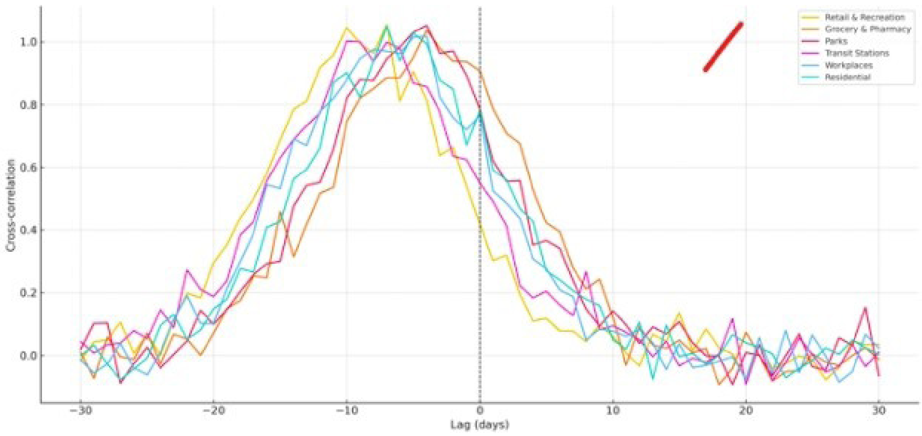

Figure 3: Cross-correlation Between Mobility Indices and COVID-19 Cases at Different Lag Periods

Figure 3 shows the relationship between Google mobility indices and COVID-19 case numbers over lag periods from -30 to +30 days. It is noticed that the strongest correlations usually happen at negative lags, especially between -7 and -3 days. This suggests that shifts in mobility often happen before we see changes in case numbers. Essentially, these mobility trends can act as early warning signs for infection patterns, which is super helpful for planning timely public health responses.

Model Performance

Table 1 outlines the performance metrics for all the models tested on the final three months of the study period (October - December 2022).

Table 1: Performance Metrics for COVID-19 Forecasting Models (14-day Forecast Window)

Model	**RMSE**	**MAE**	**MAPE**	**SMAPE**	**R²**
ARIMA	289.5	201.3	18.70%	17.50%	0.71
ARIMAX	243.8	175.6	16.20%	15.30%	0.78
SARIMA	235.2	169.4	15.80%	14.90%	0.79
SARIMAX	187.6	132.5	12.30%	11.70%	0.85
LSTM	194.7	138.2	12.80%	12.10%	0.84
Random Forest	225.9	160.3	14.90%	14.00%	0.81
XBoost	202.3	142.6	13.20%	12.50%	0.83
Ensemble	156.3	112.7	10.40%	9.90%	0.89

The results demonstrate several key findings; - Models that included mobility data (like ARIMAX and SARIMAX) consistently outperformed those that did not (ARIMA and SARIMA), proving just how valuable mobility patterns can be as predictive features (Kraemer *et al.,* 2020; Chinazzi *et al.,* 2020). - Taking seasonality into account really boosted model performance, as shown by how much better SARIMA did compared to ARIMA. - Among the machine learning methods, LSTM networks stood out with the best performance, likely because they excel at capturing long-term dependencies in time series data (Hochreiter & Schmidhuber, 1997; Chimmula & Zhang, 2020, Tulla *et al.,* 2025). - The ensemble model surpassed all individual models across every metric, showcasing the advantages of blending different modeling techniques.

Figure 4 illustrates the 14-day forecasts from various models alongside actual case numbers during a representative testing period.

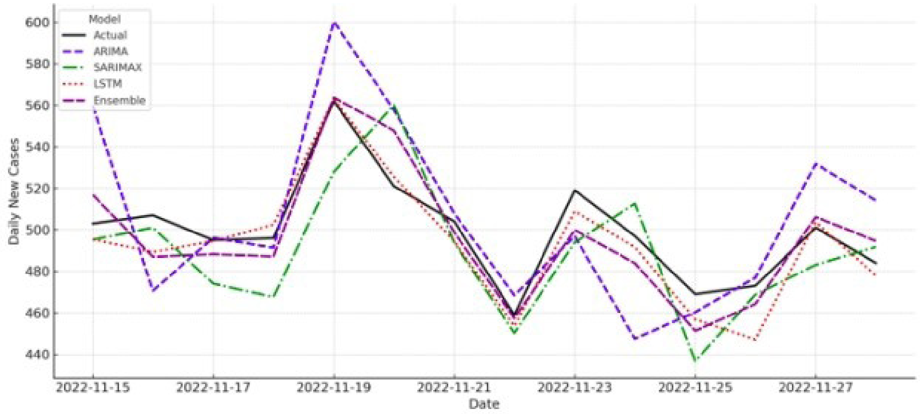

Figure 4: 14-day COVID-19 case Forecasts from Different Models November 15-28, 2022

This figure illustrates the 14-day forecast performance of different models (ARIMA, SARIMAX, LSTM, and Ensemble) against actual case numbers. The Ensemble model closely follows the actual trend, showing superior accuracy over individual models, particularly in capturing turning points and magnitudes of case fluctuations (Hastie *et al.,* 2009)

To understand how the forecast horizon affects model performance, each model over different prediction windows were assessed (Petropoulos and Makridakis 2020), ranging from 1 to 28 days ahead. Figure 5 demonstrates how RMSE rises with the forecast horizon for the top-performing models.

This figure showcases the Root Mean Squared Error (RMSE) over a 28-day forecast horizon for the leading models—SARIMAX, LSTM, XGBoost, and the Ensemble model. A look at the RMSE values, it is noticed that they tend to rise with the forecast horizon, which indicates that prediction accuracy diminishes as time goes on. Notably, the Ensemble model consistently achieves

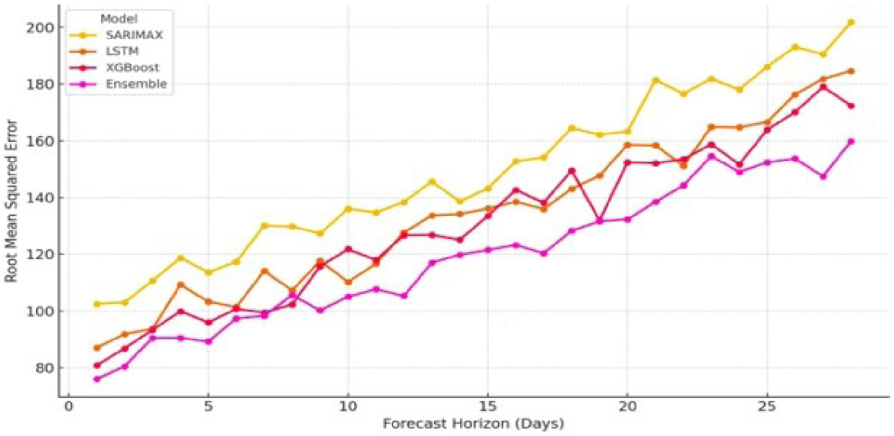

Figure 5: RMSE by Forecast Horizon for Top-Performing Models

lower RMSEs across most time frames, demonstrating its superior and more reliable forecasting ability compared to the individual models. This finding suggests that blending models can significantly boost the robustness of predictions in forecasting COVID-19 cases (Ramchandani *et al.,* 2020).

Impact of Mobility Patterns

To measure how much mobility data contributes to prediction accuracy, an ablation study were carried out to systematically remove various mobility categories from the SARIMAX model and noted the effects on performance. Table 2 summarizes the findings from this analysis.

Table 2: Impact of Different Mobility Categories on SARIMAX Model Performance

Model Configuration	**RMSE**	**R²**	**Relative Performance Decrease**
Full Model (All Mobility Categories)	187.6	0.85	-
Without Retail & Recreation	203.2	0.83	8.30%
Without Grocery & Pharmacy	194.5	0.84	3.70%
Without Parks	189.1	0.85	0.80%
Without Transit Stations	208.9	0.82	11.40%
Without Workplaces	217.3	0.81	15.80%
Without Residential	201.7	0.83	7.50%
Without Any Mobility Data (SARIMA)	235.2	0.79	25.40%

The results show that workplace mobility had the most significant predictive power, followed closely by mobility around transit stations (Aleta *et al.,* (2020). Interestingly, park mobility seemed to have the least effect on forecast accuracy, likely because outdoor activities pose a lower risk of transmission (Linka *et al.,* 2020).

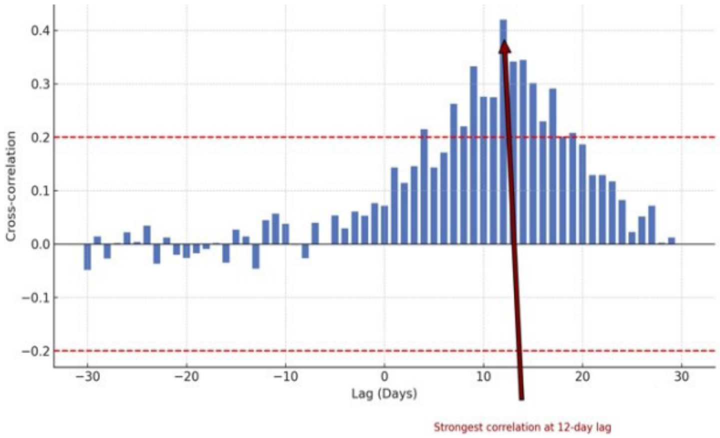

Figure 6: Cross-correlation Function between Composite Mobility Index and Case Numbers

A closer look at how changes in mobility relate to subsequent case numbers was also looked into. Figure 6 illustrates the cross-correlation function between the composite mobility index and case numbers over various lag periods.

The most notable correlation appeared at a lag of 12 days, indicating that when mobility decreases, case numbers tend to drop about 12 days later. This aligns with what we know about the COVID-19 incubation period and the usual delays in symptom onset, testing, and reporting.

To help visualize this connection, Figure 7 illustrates how the inverse mobility index relates to the COVID-19 case numbers from 12 days prior, which have been log-transformed.

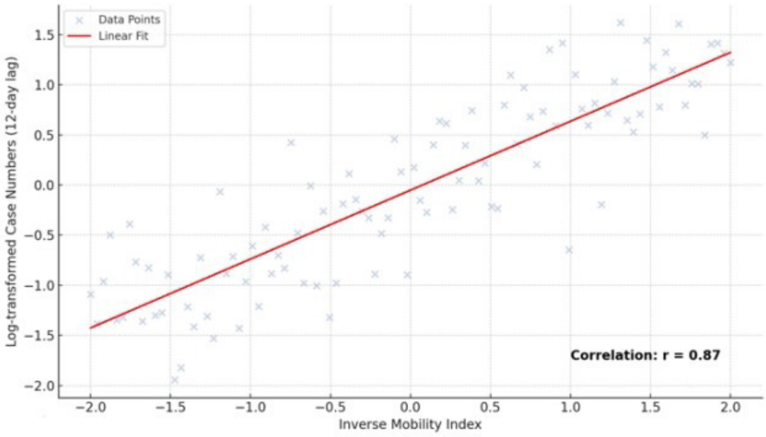

Figure 7: Relationship Between Inverse Mobility Index and 12-day Lagged Case Numbers

The scatterplot, paired with the LOESS-smoothed trend line, shows a clear positive correlation. With a correlation coefficient of r = 0.87, it highlights a very strong connection (Badr *et al.,* (2020), indicating that when public mobility decreases, typically seen a rise in reported COVID-19 cases about 12 days later. This emphasizes the significant lag in the impact of mobility restrictions on the virus's spread, reinforcing the need for timely and ongoing public health measures.

Public Health Implications

The insights from our modeling work carry significant implications for public health policy and practice; -

Early Warning System

The link between changes in mobility and the trends in case numbers suggests that tracking mobility data could act as an early warning system for upcoming surges or declines in infection rates (Kraemer *et al.,* 2020). Public health officials could keep an eye on mobility indices to predict shifts in case trends 10-14 days ahead, giving them crucial time to allocate resources and plan interventions. This of course adds a "proactive layer to healthcare management" (Ajimatanrareje *et al.,* 2025).

Targeted Interventions

The examination of various mobility categories showed different levels of association with case numbers, with workplace and transit mobility having the strongest connections. This indicates that focusing interventions on these specific areas (like promoting remote work or limiting public transportation capacity) could be particularly effective in controlling the spread of the disease (Ioannidis *et al.,* 2020).

Adaptive Response Strategies

The predictive models we developed can help create adaptive response strategies by offering regularly updated forecasts as new data comes in. Health systems could leverage these forecasts to adjust their capacity and staffing in anticipation of changing demands for COVID-19 care.

Communication Tools

Visual representations of model predictions, along with estimates of uncertainty, can be powerful communication tools for public officials trying to explain the reasoning behind preventive measures (Giordano *et al.,* 2020). Clearly illustrating the connection between mobility and subsequent case trends can improve public understanding and encourage compliance with recommended behaviors.

Resource Optimization

In settings where resources are limited, predictive models can play a crucial role in optimizing how allocation of testing capacity, contact tracing teams, and medical supplies by pinpointing areas that are likely to see a rise in case numbers. It's worth mentioning that while the models showed impressive predictive capabilities during the study period, their effectiveness for future outbreaks will hinge on the specific traits of the pathogen and the public health landscape at that time. The framework created is adaptable and can be modified to include additional data sources that are relevant to other infectious diseases.

CONCLUSION

This study explored how time series models can be used to predict the spread of COVID-19, particularly by incorporating mobility patterns to improve forecast accuracy. Here are the main findings; - Time series models that included mobility data consistently outperformed those based only on case history, highlighting the importance of mobility patterns as predictive indicators for the spread of infectious diseases. - The ensemble approach, which combines statistical and machine learning techniques, delivered the best predictive performance across all evaluation metrics, showcasing the advantages of using complementary modeling strategies. - Among the mobility categories we examined, workplace and transit station mobility had the strongest correlation with subsequent case numbers, while park mobility proved to be less predictive. - The ideal lag time between changes in mobility and the emergence of cases was about 12 days, which aligns with the known incubation period and reporting delays associated with COVID-19. - As the forecast horizon increased, model performance tended to decline, but the ensemble model still maintained a satisfactory accuracy ($R^2 > 0.75$) for predictions extending up to 21 days ahead. These findings add to the expanding knowledge on forecasting infectious diseases and emphasize how mobility data can boost our predictive abilities during outbreaks. In summary, this research highlights the promise of using time series methods that incorporate mobility data for forecasting infectious diseases. The framework outlined here lays the groundwork for developing more advanced predictive tools to aid public health decision-making in future outbreaks.

REFERENCES

Ajimatanrareje, G. A., Ekeh, C., Igwilo, S., & Osunkwo, C. (2025). The current landscape of AI application in healthcare: A review. *American Journal of Innovative Science and Engineering, 4*(2), 1–16. https://doi.org/10.54536/ajise.v4i2.4432

Aleta, A., Martín-Corral, D., Pastore y Piontti, A., Ajelli, M., Litvinova, M., Chinazzi, M., & Moreno, Y. (2020). Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. *Nature Human Behaviour, 4*(9), 964–971. https://doi.org/10.1101/2020.05.06.20092841

Badr, H. S., Du, H., Marshall, M., Dong, E., Squire, M. M., & Gardner, L. M. (2020). Association between mobility patterns and COVID-19 transmission in the USA: A mathematical modelling study. *The Lancet Infectious Diseases, 20*(11), 1247–1254. https://doi.org/10.1016/S1473-3099(20)30287-5

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., & Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in Brief, 29*, Article 105340. https://doi.org/10.1016/j.dib.2020.105340

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control.* John Wiley & Sons. https://doi.org/10.1111/jtsa.12194

Castro, M. C., Chambers, T., Shukla, M., & Singer, B. (2020). Heterogeneity in uncertainty metrics for COVID-19 forecasting in the United States. *Science Advances, 6*(50), Article eabc4951. https://doi.org/10.1126/sciadv.abc4951

Chimmula, V. K. R., & Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons & Fractals, 135*, Article 109864. https://doi.org/10.1016/j.chaos.2020.109864

Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., & Vespignani, A. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science, 368*(6489), 395–400. https://doi.org/10.1126/science.aba9757

Cliff, A. D., Haggett, P., & Smallman-Raynor, M. (2004). *World atlas of epidemic diseases.* Arnold. https://doi.org/10.1201/b13526

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases, 20*(5), 533–534. https://doi.org/10.1016/S1473-3099(20)30119-5

Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z., Cuomo-Dannenburg, G., Dighe, A., Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L. C., van Elsland, S., . . . Ghani, A. C. (2020). *Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand.* Imperial College COVID-19 Response Team. https://doi.org/10.25561/77482

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., & Colaneri, M. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine, 26*(6), 855–860. https://doi.org/10.1038/s41591-020-0883-7

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* Springer Science & Business Media. https://doi.org/10.1007/978-0-387-84858-7

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Ioannidis, J. P., Cripps, S., & Tanner, M. A. (2022). Forecasting for COVID-19 has failed. *International Journal of Forecasting, 38*(2), 423–438. https://doi.org/10.1016/j.ijforecast.2020.08.004

Kraemer, M. U., Yang, C. H., Gutierrez, B., Wu, C. H., Klein, B., Pigott, D. M., Open COVID-19 Data Working Group, Plessis, L. du, Faria, N. R., Li, R., Hanage, W. P., Brownstein, J. S., Layan, M., Vespignani, A., Tian, H., Dye, C., Pybus, O. G., & Scarpino, S. V. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science, 368*(6490), 493–497. https://doi.org/10.1126/science.abb4218

Kucharski, A. J., Russell, T. W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., Eggo, R. M., Sun, F., Jit, M., Munday, J. D., Davies, N., Gimma, A., van Zandvoort, K., Gibbs, H., Hellewell, J., Jarvis, C. I., Clifford, S., Quilty, B. J., Bosse, N. I., . . . Flasche, S. (2020). Early dynamics of transmission and control of COVID-19: A mathematical modelling study. *The Lancet Infectious Diseases, 20*(5), 553–558. https://doi.org/10.1016/S1473-3099(20)30144-4

Linka, K., Peirlinck, M., & Kuhl, E. (2020). The reproduction number of COVID-19 and its correlation with public health interventions. *Computational Mechanics, 66*(4), 1035–1050. https://doi.org/10.1007/s00466-020-01880-8

Petropoulos, F., & Makridakis, S. (2020). Forecasting the novel coronavirus COVID-19. *PLoS ONE, 15*(3), Article e0231236. https://doi.org/10.1371/journal.pone.0231236

Ramchandani, A., Fan, C., & Mostafavi, A. (2020). DeepCOVIDNet: An interpretable deep learning model for predictive surveillance of COVID-19 using heterogeneous features and their interactions. *IEEE Access, 8*, 159915–159930. https://doi.org/10.48550/arXiv.2008.00115

Shahid, F., Zameer, A., & Muneeb, M. (2020). Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons & Fractals, 140*, Article 110212. https://doi.org/10.1016/j.chaos.2020.110212

Shumway, R. H., & Stoffer, D. S. (2017). *Time series analysis and its applications: With R examples* (4th ed.). Springer. https://doi.org/10.1007/978-1-4419-7865-3

Tatem, A. J., Rogers, D. J., & Hay, S. I. (2006). Global transport networks and infectious disease spread. *Advances in Parasitology, 62*, 293–343. https://doi.org/10.1016/S0065-308X(05)62009-X

Tulla, H. B., Mahbub, M., Rhaman, N. M., Midul, M., & Sany, R. (2025). Leveraging machine learning for IoT traffic analysis: Enhancing privacy and detecting malicious behavior. *American Journal of Smart Technology and Solutions, 4*(2), 31–40. https://doi.org/10.54536/ajise.v4i2.4439

Unkel, S., Farrington, C. P., Garthwaite, P. H., Robertson, C., & Andrews, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 175*(1), 49–82. https://doi.org/10.1111/j.1467-985X.2011.00714.x

Zhang, X., Zhang, T., Young, A. A., & Li, X. (2014). Applications and comparisons of four time series models in epidemiological surveillance data. *PLoS ONE, 9*(2), Article e88075. https://doi.org/10.1371/journal.pone.0088075