ACCSCIENCE PUBLISHING

Microbes & Immunity

ORIGINAL RESEARCH ARTICLE

A comprehensive statistical analysis of COVID-19 trends: Global and United States insights through ARIMA, regression, and spatial models

Zhihao Lei[1,2*]

[1]School of Mathematics, University of Edinburgh, Edinburgh, Scotland, United Kingdom

[2]Department of Biostatistics, School of Public Health, Brown University, Providence, Rhode Island, United States of America

Abstract

The COVID-19 pandemic has driven the need for accurate data analysis and forecasting to support public health decision-making. This study applied autoregressive integrated moving average (ARIMA) models and ARIMA models with exogenous variables to predict short-term trends in confirmed COVID-19 cases across several regions, including the United States of America, Asia, Europe, and Africa. Model performance was compared between ARIMA and the automated model selection function, auto.arima, and anomaly detection was performed to investigate discrepancies between predicted and observed case numbers. Additionally, the study explored the relationship between vaccination rates and new case trends while also examining the influence of socioeconomic factors—such as gross domestic product per capita, human development index, and healthcare resources availability—on COVID-19 incidence across countries. The findings provide valuable insights into the effectiveness of predictive models and highlight the significant role of socioeconomic factors in the spread of the virus, thereby contributing to the development of more effective strategies for future epidemic prevention and control.

Keywords: Autoregressive integrated moving average model; COVID-19; Public health; Socioeconomic factors; Time series forecasting; Vaccination rates

***Corresponding author:**
Zhihao Lei
(Z.Lei-6@sms.ed.ac.uk)

Citation: Lei Z. A comprehensive statistical analysis of COVID-19 trends: Global and United States insights through ARIMA, regression, and spatial models. *Microbes & Immunity*. 2025;2(3):108-129. doi: 10.36922/MI025040007

Received: January 22, 2025

Revised: April 9, 2025

Accepted: May 12, 2025

Published online: June 18, 2025

1. Introduction

Since the onset of the COVID-19 pandemic in late 2019, the pandemic has had profound and widespread effects on global public health, economies, and daily life. As of 2024, it continues to pose challenges to healthcare systems worldwide, underscoring the ongoing need for accurate forecasting of case trends for effective policy-making decisions and intervention strategies. Statistical modeling, particularly time series analysis, has proven to be a valuable tool in predicting the trajectory of the pandemic and supporting the development of effective public health responses.[1]

Among the range of statistical models, the autoregressive integrated moving average (ARIMA) model has been widely employed in epidemiological studies for short-term forecasting due to its simplicity and effectiveness in modeling temporal data.[2] ARIMA

models have been utilized to predict COVID-19 case trends across different countries, often producing reasonably accurate forecasts over limited time horizons.[3] However, the accuracy of ARIMA-based forecasts can vary significantly depending on the region and time period, influenced by factors such as viral mutations, government interventions, and changes in population behavior.[4] One notable limitation of ARIMA models is their exclusive reliance on historical data, without considering external factors that might influence future trends—such as vaccination rates, policy changes, or behavioral adaptations—which introduces greater uncertainty in long-term predictions.

To address these limitations, the ARIMA with exogenous variables (ARIMAX) model incorporates external variables—such as vaccination rates—to enhance its predictive capabilities. By incorporating vaccination data, the model enables researchers to assess the potential impact of vaccination campaigns on future case trends, providing a more comprehensive understanding of epidemic dynamics.[5] Although previous studies have shown that vaccination plays a crucial role in mitigating the spread of COVID-19—leading to significant reductions in new case numbers following mass immunization efforts[6]—most of these studies are region-specific or limited to particular periods and do not fully capture the complex interactions between vaccination efforts, virus mutations, and policy interventions.

In addition to time series forecasting, examining the relationship between COVID-19 incidence and socioeconomic factors is crucial. Previous research has highlighted the influence of indicators—such as gross domestic product (GDP) per capita, healthcare infrastructure, and other socioeconomic variables—in shaping the impact of the pandemic across different regions.[7] For example, countries with greater healthcare spending and more robust medical systems tend to manage the crisis more effectively, resulting in lower mortality rates and more effective containment strategies.[8] However, most of the existing research relies on single-variable analyses and does not fully capture the complex, multifaceted interactions among socioeconomic factors, which may contribute to significant disparities in COVID-19 outcomes across different countries.

This study aims to advance existing research by applying both ARIMA and ARIMAX models to predict short-term COVID-19 case trends in the United States (US) and globally. In the ARIMAX model, vaccination rates are incorporated as an exogenous variable to enhance predictive accuracy and provide deeper insights into the relationship between vaccination efforts and new case trends. Discrepancies between predicted and actual case numbers are examined to investigate potential causes for forecast anomalies, such as policy changes and virus mutations. Additionally, the study explores the socioeconomic factors—including GDP per capita, healthcare resources, and human development index (HDI)—on COVID-19 case numbers across countries. This multidimensional approach allows for a more comprehensive comparison of ARIMA and ARIMAX models performance, while also offering valuable perspectives on the broader determinants of the pandemic's spread, contributing to future epidemic prevention and control strategies.

2. Data collection

To conduct a comprehensive analysis of the COVID-19 pandemic and its associated factors, a diverse set of datasets was obtained from reputable sources, including the World Health Organization (WHO), Centers for Disease Control and Prevention, World Bank, and other national and international agencies. These datasets were selected based on their relevance, comprehensiveness, and frequency of updates to ensure that the analysis reflects the most accurate and current information available. As shown in Table 1, the data include daily and weekly reports of COVID-19 cases, deaths, and vaccination trends, alongside key socioeconomic indicators such as GDP per capita, HDI, Gini index, healthcare expenditures, and healthcare infrastructure data. These variables were essential for modeling the progression of the pandemic and for evaluating the impact of various factors on infection rates.

All statistical analyses were conducted using R version 4.4.3.

3. Methodology

3.1. Theoretical basis of the ARIMA model

The ARIMA model is a statistical method commonly used for analyzing and forecasting time series data. The general form of an ARIMA model with order (p, d, q) is represented by the following equation:

$$\phi(B)\nabla^d x_t = \theta(B)\varepsilon_t \quad \text{(I)}$$

where:

(i) $\nabla^d = (1-B)^d$ is the differencing operator, with B representing the backshift operator,[9]

(ii) $\phi(B) = 1-\phi_1 B - \cdots - \phi_p B^p$ is the autoregressive (AR) coefficient polynomial,[9]

(iii) $\theta(B) = 1-\theta_1 B - \cdots - \theta_q B^q$ is the moving average (MA) coefficient polynomial,[9]

(iv) ε_t denotes white noise error terms that satisfy the following properties: $E(\varepsilon_t) = 0$, Var $(\varepsilon_t) = \sigma^2$,

(v) $E(\varepsilon_t \varepsilon_s) = 0$ for $s \neq t$,[9]

Table 1. Overview of key datasets

Data source	Data description	Link
WHO	Daily COVID-19 case and death data reported to the WHO, updated weekly and incorporating corrections to historical records based on newly received information	https://data.who.int/dashboards/covid19/data?n=c
United Nations Development Programme	The HDI measures key dimensions of human development, such as a long and healthy life, access to education, and a decent standard of living	https://hdr.undp.org/data-center/human-development-index#/indices/HDI
CDC	COVID-19 vaccination trends in the US at both national and jurisdictional levels, regularly updated to reflect changes over time	https://data.cdc.gov/Vaccinations/
	Archived weekly data on COVID-19 cases and deaths in the US, providing state-level trends and historical records	https://data.cdc.gov/Case-Surveillance/Weekly-United-States-COVID-19-Cases-and-pwn4-m3yp/about_data
	County-level data on COVID-19 vaccinations in the US, including detailed trends and demographic breakdowns	https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-Stat8×kx-amqh/about_data
World Bank	GDP per capita (in current US$), representing the monetary value of all final goods and services produced per person in a given year	https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
	The Gini index measures the distribution of income across a population, representing inequality	https://data.worldbank.org/indicator/SI.POV.GINI
	Current health expenditure per capita (in current US$) represents the average national spending on healthcare per individual	https://data.worldbank.org/indicator/SH.XPD.CHEX.PC.CD
	Hospital beds per 1,000 people, indicating the availability of healthcare infrastructure	https://data.worldbank.org/indicator/SH.MED.BEDS.ZS
	Population density, measured as the number of people per km^2 of land area, reflects how concentrated a population is within a specific area	https://data.worldbank.org/indicator/EN.POP.DNST
Bureau of Economic Analysis	GDP and personal income by state for the first quarter of 2024, providing insights into regional economic performance	https://www.bea.gov/data/gdp/gdp-state
US Census Bureau	Historical population density data (1910 – 2020) provides population density trends over the past century	https://www.census.gov/data/tables/time-series/dec/density-data-text.html
	State population totals and components of change from 2020 to 2023, highlighting trends in population growth, decline, and migration patterns	https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html
Kaiser Family Foundation	Healthcare expenditures per capita by state of residence, providing detailed insights into state-level spending on healthcare	https://www.kff.org/other/state-indicator/health-spending-per-capita/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22%22sort%22:%22asc%22%7D
	Number of hospital beds per 1,000 people by ownership type, providing insights into the distribution and availability of healthcare resources	https://www.kff.org/other/state-indicator/beds-by-ownership/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22%22sort%22:%22asc%22%7D

Abbreviations: GDP: Gross domestic product; HDI: Human development index; US: United states; WHO: World health organization.

(vi) $E(x_s \varepsilon_t) = 0$ for all $s < t$, ensuring that the noise terms are uncorrelated with past values of the series.[10]

The ARIMA model consists of three main components:

(i) Autoregressive: Captures the relationship between a current observation and its past (lagged) values.[9]

(ii) Integrated: Represents the differencing of observations required to achieve stationarity in the time series.[9]

(iii) Moving average: Models the relationship between a current observation and the residual errors from an MA model applied to past (lagged) observations.[9]

The ARIMA modeling process began with a stationarity test, commonly conducted using the augmented Dickey–Fuller test.[11] If the time series is found to be non-stationary, transformations such as differencing or logarithmic scaling are typically applied to achieve stationarity.[10] Model identification involved determining the orders of the model, specifically the values of p and q, which represent the AR and MA terms, respectively. This step is usually performed by analyzing the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots.[9] The differencing order (d) was determined based on the transformations applied during the stationarity testing phase.

Following model identification, the parameters ϕ_i and θ_j were estimated, typically using maximum likelihood estimation.[10] Model validation was then performed using statistical tests such as the Ljung-Box test to ensure that the residuals exhibited white noise behavior, indicating that the model adequately captured the time series structure.[12] Model selection was based on information criteria such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC), with preference given to the model with the lowest criterion value.[13] Finally, once validated, the model was employed to forecast future values of the time series.[9]

Figure 1 summarizes this process, illustrating the sequence of steps from stationarity assessment to forecasting.

3.2. Rolling window cross-validation and comparison with auto.arima

In this study, rolling window cross-validation was used to evaluate the performance of ARIMA models for time series forecasting. The primary goal was to identify the optimal ARIMA model parameters by minimizing the root mean squared error (RMSE) and to compare the results with those obtained from the automated model selection function, auto.arima.[14]

Rolling window cross-validation is a method specifically designed for time series data as it preserves the temporal order of the data. In each iteration, the model was trained on a fixed-length window of historical data and validated on the subsequent observation. This approach ensures that the evaluation reflects real-world forecasting conditions, where future values must be predicted using only past data.[15] For each ARIMA model evaluated, the one-step-ahead forecast errors were calculated, and RMSE was used as the primary evaluation metric. RMSE is given:

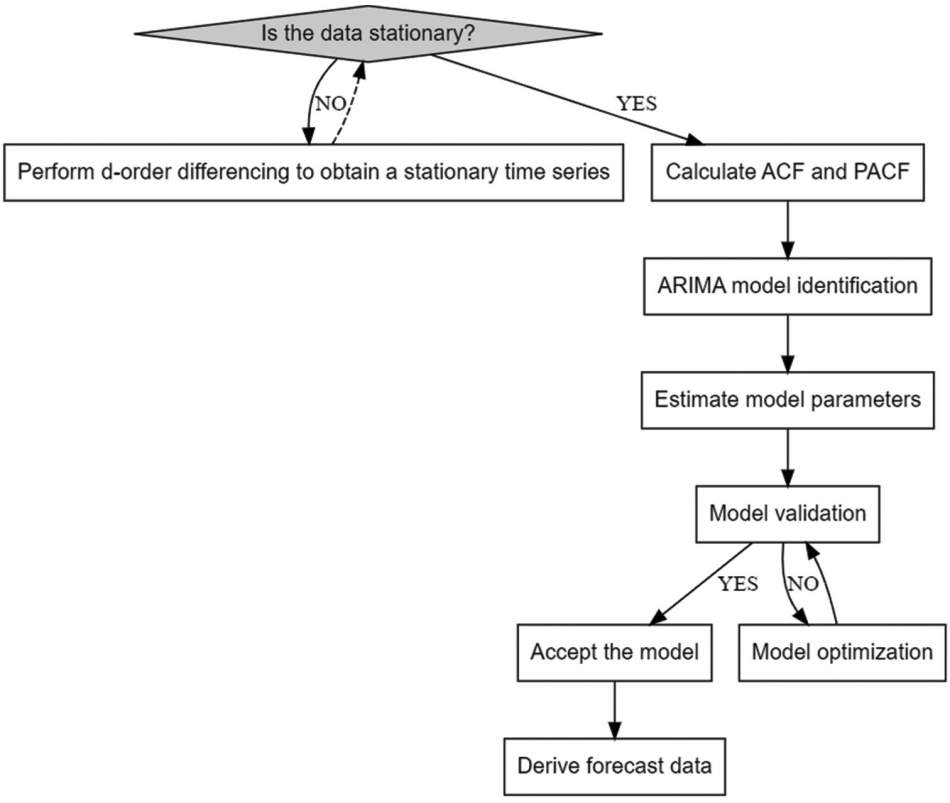

Figure 1. The autoregressive integrated moving average model construction flow chart
Abbreviations: ACF: Autocorrelation function; ARIMA: Autoregressive integrated moving average; PACF: Partial autocorrelation function.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \breve{x}_i\right)^2} \quad \text{(II)}$$

where x_i is the actual value, and $\hat{x}_i$ is the predicted value. Lower RMSE values indicate better model performance.[16]

A grid search was conducted across various combinations of the p and q parameters, with the d-order fixed at 1. This process was parallelized to efficiently explore the parameter space.[14] RMSE was used as the evaluation metric due to its effectiveness in quantifying average prediction errors while placing greater emphasis on larger errors.[17] This characteristic makes RMSE particularly useful in contexts where significant forecasting errors could lead to significant consequences, as it penalizes large discrepancies more heavily than other metrics, such as mean absolute error (MAE). Additionally, since RMSE is measured in the same units as the original data, it provides results that are interpretable in practical applications.

To compare the performance between manual and automated model selections, the auto.arima function was employed. This function automatically identifies the optimal ARIMA model by optimizing information criteria such as the AIC or BIC.[18] While auto.arima provides a rapid and efficient global fit over the entire dataset, rolling window cross-validation offers a more robust evaluation by assessing the model's predictive performance across different time periods.[19] This approach enabled a detailed comparison of the consistency and reliability of automated versus manually selected models.

By visualizing the RMSE values across various parameter combinations, the best-performing model identified through rolling window cross-validation was compared with the model selected by auto.arima. This comparison provided insights into the trade-offs between automated selection and manual tuning in ARIMA-based time series forecasting.

3.3. Anomaly detection

Anomaly detection in time series data is crucial for identifying irregular patterns, such as sudden spikes in COVID-19 case numbers. In this study, a statistical approach was employed to detect anomalies directly from the time series data without fitting a complex model like ARIMA. This method, known as residual-based anomaly detection, identifies outliers based on their deviation from expected behavior within the data.[14]

The anomaly detection approach relies on statistical rules that identify observations as anomalies when they significantly deviate from surrounding values. Specifically, outliers are detected by analyzing the residuals after accounting for typical patterns in the time series.[20] In this study, the residuals were examined, and points were classified as outliers if their deviation from the local mean exceeded a certain threshold.

The theoretical basis for this method involves identifying points that significantly deviate from the local mean or expected value of the time series. Mathematically, a data point x_t is considered an outlier if it satisfies the condition:

$$|x_t - \mu| > k \times \sigma \quad \text{(III)}$$

where:

(i) μ represents the local mean,

(ii) σ is the standard deviation of the surrounding data points,

(iii) k is a threshold factor that determines the sensitivity of the detection.[21]

Typically, k is set to values such as 2 or 3, corresponding to confidence intervals commonly used in outlier detection.[20]

This method is particularly effective for detecting additive outliers, which appear as sudden spikes or drops in the time series—events that may result from external shocks such as the emergence of a new COVID-19 variant.[14] Identifying and analyzing these outliers provides valuable insights into how unexpected events influence overall trends, enabling more informed adjustments to forecasting models.

The detected anomalies were then visualized in a time series plot, highlighting points of significant deviation to facilitate further investigation and model adjustments.[21]

3.4. Theoretical basis of the ARIMAX model

To improve the accuracy of time series forecasting, the ARIMAX model was employed, integrating external factors into the standard ARIMA model. This extension allows the model to account for factors beyond the inherent patterns in the target time series.[14] In this study, vaccination rates were included as an exogenous variable to determine whether they would improve forecast accuracy compared to the ARIMA model, which relied solely on historical time series data.

The ARIMAX model expanded upon the ARIMA framework by introducing exogenous regressors—external variables believed to influence the dependent variable. Mathematically, the ARIMAX model is expressed as:

$$y_t = \phi_0 + \sum_{i=1}^{p}\phi_i y_{t-i} + \sum_{j=1}^{q}\theta_j \epsilon_{t-j} + \epsilon_t + \sum_{k=1}^{r}\beta_k X_{t-k} \quad \text{(IV)}$$

where:
(i) y_t represents the value of the dependent variable at time t,
(ii) ϕ_i are the AR coefficients,
(iii) θ_j are the MA coefficients,
(iv) ϵ_t is the error term,
(v) X_{t-k} represents the exogenous variable lagged by k periods.[22]

Exogenous variables were incorporated to capture additional influences on the time series that could not be explained solely by its historical values.[9]

The ARIMAX model was fitted using an automated selection of ARIMA parameters (p, d, and q) while incorporating the selected exogenous variable. Model performance was evaluated by comparing the ARIMAX model against the ARIMA model using standard evaluation metrics, such as AIC, RMSE, and MAE.[14,16]

Model forecasts were generated for a holdout period to assess predictive accuracy. The inclusion of exogenous variables in the ARIMAX model enables an assessment of whether incorporating external factors can improve the forecasting performance and provide a more comprehensive understanding of the dynamics affecting the time series. This comparison between ARIMA and ARIMAX models provided insights into the benefits and limitations of incorporating external factors into the forecasting process.[22]

3.5. Evaluating the impact of vaccination on new COVID-19 cases

To analyze the relationship between vaccination rates and the number of new COVID-19 cases, several statistical methods were employed, including Granger causality testing, segmented regression, and regression discontinuity design (RDD). These methods support a clearer understanding of both the temporal relationships and potential causal effects of vaccination on the incidence of new cases.[9,14]

3.5.1. Granger causality test

The Granger causality test was employed to evaluate whether past vaccination rates provided predictive information for future new COVID-19 case numbers. This test determines whether one time series provides statistically significant information for forecasting another time series, suggesting a potential causal relationship.[23] In this context, the null hypothesis states that vaccination rates do not Granger-cause new COVID-19 cases—implying that past vaccination rates do not provide additional predictive value for future case numbers after accounting for past cases. A detailed mathematical formulation of the model is provided in the Supplementary File.

It is important to note that the Granger causality test does not confirm true causality in a philosophical or structural sense but rather indicates that past values of one series are useful in predicting another.

3.5.2. Segmented regression analysis and Chow test

Segmented regression analysis was employed to quantify the impact of vaccination on the trend of new COVID-19 cases. This method estimates changes in trends before and after an intervention, such as the introduction of a vaccination program.[24] The resulting coefficients provide estimates of the immediate change in case numbers and the change in the trend following the intervention.

To validate these findings, a Chow test was conducted to assess the presence of a structural break at the intervention point. This test evaluates whether the relationship between time and new COVID-19 cases differs significantly before and after the intervention.[25] Rejecting the null hypothesis indicates a statistically significant change in the trend post-intervention. A detailed mathematical formulation of both the segmented regression model and the Chow test is provided in the Supplementary File.

3.5.3. RDD

An RDD was employed to estimate the causal effect of vaccine introduction on new COVID-19 cases, using the start of mass vaccination as the cutoff point.[26] RDD assumes that observations just before and after the cutoff are comparable except for the treatment effect. This effect is estimated by comparing new COVID-19 cases immediately before and after vaccination introduction. The parameter of interest (β) represents the effect of the intervention at the cutoff. A non-parametric approach was used to flexibly model the relationship between time and new cases on either side of the cutoff. Detailed mathematical formulation and implementation of the RDD model are provided in the Supplementary File.

While RDD strengthens causal inference through a quasi-experimental design, it remains dependent on the assumption that other confounding factors vary continuously at the cutoff. As such, it does not provide definitive proof of causality.

3.6. Regression analysis of COVID-19 infection rates and determinants

3.6.1. Linear regression analysis of COVID-19 infection rates and economic development

To investigate the relationship between COVID-19 infection rates and economic development, a linear regression analysis was conducted with the infection rate as the dependent variable and GDP per capita as

the independent variable. This analysis aimed to assess whether a country's economic development is associated with its COVID-19 infection rate. Additionally, Pearson's and Spearman's correlation coefficients, along with the maximal information coefficient (MIC), were calculated to evaluate the strength of linear relationships.[27] Pearson's correlation measured the strength of linear relationships, Spearman's assessed the monotonic relationships, and MIC captured both linear and nonlinear associations. Detailed mathematical formulations for these analyses are provided in the Supplementary File.

3.6.2. Multiple regression analysis with additional socioeconomic and health variables

To further investigate the determinants of COVID-19 infection rates, a multiple regression model was employed, incorporating additional variables such as the HDI, Gini coefficient, health expenditure per capita, number of hospital beds per 1,000 people, and population density. This analysis was used to evaluate the relative influence of various socioeconomic and healthcare-related factors on COVID-19 infection rates across various countries.[28] Interaction terms were included to explore potential synergistic effects between variables.[29] The detailed mathematical formulation of the expanded regression model is provided in the Supplementary File.

3.6.3. Addressing multicollinearity: Stepwise regression, principal component regression (PCR), and partial least squares (PLS)

Given the potential for multicollinearity among socioeconomic and healthcare-related predictors, several strategies were implemented to improve model interpretability and estimation stability. First, stepwise regression was employed to refine the linear model by iteratively adding or removing predictors based on their statistical significance. The selection process aimed to minimize the AIC, balancing model fit with complexity.[30] To further evaluate multicollinearity, the variance inflation factor (VIF) was calculated for each predictor. Variables with VIF values exceeding 10 were considered to exhibit significant multicollinearity, which can inflate the variance of coefficient estimates and reduce model reliability.[31]

To address multicollinearity more robustly, two-dimensionality reduction techniques were applied: PCR and PLS regression. Both methods transformed the original set of correlated predictors into a smaller set of uncorrelated components, which were then used in place of the original variables in regression analysis.[32]

The PCR analysis constructed components solely based on the variance structure of the predictor variables, identifying principal components that captured the largest proportion of variance in the input space.[33] These components were then used to predict the dependent variable, regardless of their relevance to it. In contrast, PLS regression incorporated information from both the predictors and the response variable during component extraction, enabling it to select components most relevant for predicting the outcome by maximizing the covariance between predictors and response.[34]

The appropriate number of components for each method was determined using cross-validation procedures, and model performance was assessed based on the mean squared error of prediction (MSEP). Full mathematical formulations and implementation details for both PCR and PLS are provided in Supplementary File.

3.7. Spatial autocorrelation and hotspot analysis of COVID-19 cases

In this study, spatial analysis techniques were applied to examine the distribution of COVID-19 infection rates across various regions. Moran's *I* was calculated to assess global spatial autocorrelation, and the Getis-Ord Gi* statistic was performed to identify local hotspots and coldspots. The results were visualized using traditional red-blue color schemes, effectively highlighting areas with significant spatial clustering of high or low infection rates.[35,36]

3.7.1. Spatial autocorrelation: Moran's I

Moran's *I* is a widely used measure of global spatial autocorrelation that quantifies the degree of spatial clustering of a variable across geographical regions.[37] It identifies whether similar values (e.g., infection rates) tend to cluster spatially. A positive Moran's *I* indicates that similar values clustered together, while a negative value indicates that dissimilar values are adjacent. For this analysis, a spatial weights matrix was generated based on shared boundaries between geographic regions, and Moran's *I* was calculated to assess the overall spatial autocorrelation of COVID-19 infection rates.[35] The detailed mathematical formulation of Moran's *I* is provided in the Supplementary File.

3.7.2. Hotspot analysis: Getis-ord Gi statistic*

The Getis-Ord Gi* statistic is a local spatial statistic used to identify geographic hotspots and coldspots, representing areas with significant clustering of high or low values, such as COVID-19 infection rates. Hotspots indicate clusters of high values, while coldspots indicate clusters of low values. The significance of these clusters is determined by comparison with a reference distribution under the null hypothesis of spatial randomness.[38] For this analysis, the Getis-Ord Gi* statistic was calculated using a spatial

weights matrix to identify regions with statistically significant clustering of high or low infection rates.[36] The detailed mathematical formulation of the Getis-Ord Gi* statistic is provided in the Supplementary File.

4. Results and discussion

4.1. Short-term forecasting and anomaly detection in COVID-19 case counts using ARIMA models

To evaluate the short-term predictive performance of ARIMA models on COVID-19 case counts, forecasts were generated for four distinct time periods using training data from prior months. Predictive accuracy was assessed by comparing these forecasts with actual observed data.

The first forecast, covering September 27 – December 27, 2020, utilized data from January 5 to September 27, 2020. As shown in Figure 2A, the forecast generally follows the actual case trajectory, though deviations near the end of the period highlight the model's limitations in capturing sudden changes in the data. The ACF and PACF plots (Figure 2B and C) reveal some residual autocorrelation, highlighting potential areas for model improvement. The Ljung-Box test yields a p=0.3746, indicating no significant residual autocorrelation.

The second forecast, covering December 27, 2020 – March 28, 2021, utilized data up to December 27, 2020. Figure 2D illustrates a closer alignment between the predicted and actual observed cases, with only minor deviations. The ACF and PACF plots (Figure 2E and F) further support the model's adequacy, though some residual correlations persist. The Ljung-Box test for this period yields a p=0.6327, further indicating that residual autocorrelation is not a concern.

The third forecast, covering March 28 – June 27, 2021, was generated using data up to March 28, 2021. As illustrated in Figure 2G, the model closely aligns with the actual case counts throughout the period, demonstrating strong predictive capability. The corresponding ACF and PACF plots (Figure 2H and I) show that the model effectively captures the data's temporal structure, though the Ljung-Box test yields a p=0.0728, suggesting the presence of minor residual autocorrelation.

In the final forecast period, covering September 26 – December 26, 2021, the model included data from January 3, 2021, to September 26, 2021. As illustrated in Figure 2J, the model maintains strong performance, with forecasts closely aligning with the actual case counts. The ACF and PACF plots (Figure 2K and L) indicate that the model has successfully captured the underlying patterns, with the Ljung-Box test demonstrating a p=0.2876, indicating minimal residual autocorrelation.

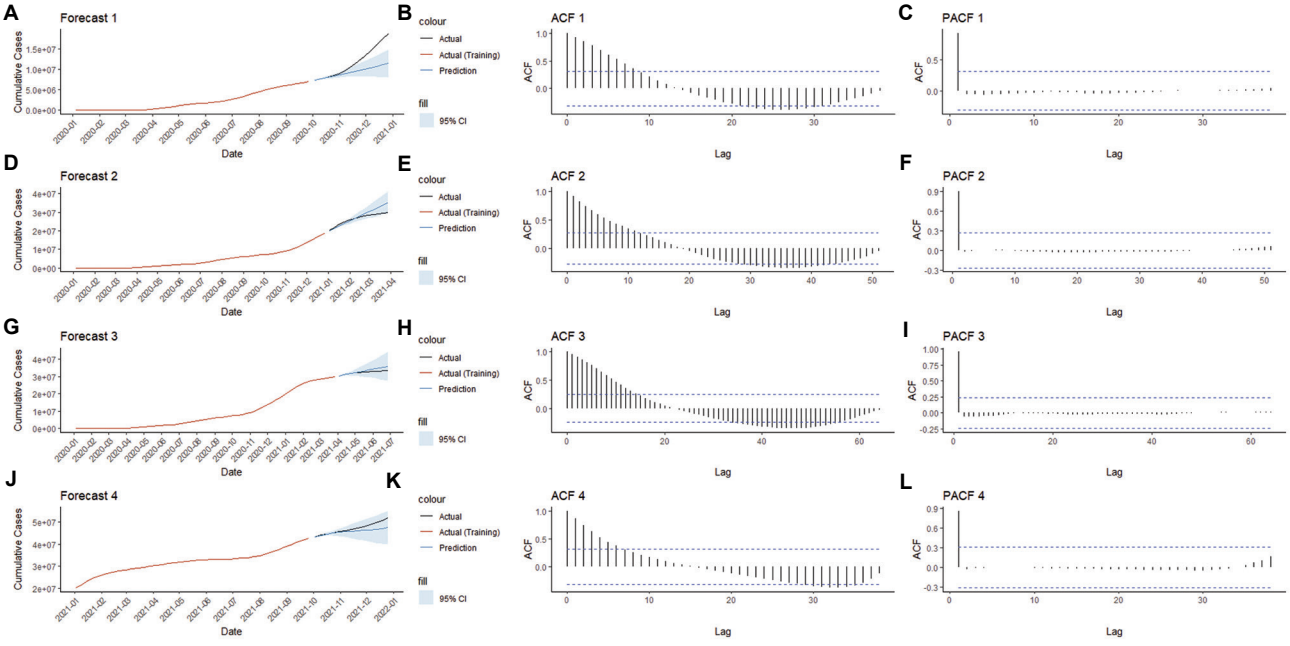

Figure 2. ARIMA model analysis of COVID-19 case forecasts in the United States across four time periods. First forecast period: Actual versus predicted (A), ACF (B) and PACF (C); second forecast period: Actual versus predicted (D), ACF (E) and PACF (F); third forecast period: Actual versus predicted (G), ACF (H) and PACF (I); and fourth forecast period: Actual versus predicted (J), ACF (K) and PACF (L).
Abbreviations: ACF: Autocorrelation function; CI: Confidence interval; PACF: Partial autocorrelation function; USA: United States of America.

Throughout these periods, the ARIMA model demonstrates consistent predictive accuracy, although the residual autocorrelation observed in the ACF and PACF plots highlights areas for further refinement to improve model performance. These findings indicate that while the ARIMA model effectively captures overall trends, it does not fully account for short-term dependencies or sudden structural changes in the data. The presence of residual autocorrelation—especially mild positive lags at short intervals—suggests the presence of unmodeled impacts, such as seasonal effects or external shocks. To address this, ARIMAX models incorporating vaccination rates as exogenous variables were then explored, with the findings discussed in Section 4.5, demonstrating improved performance in certain forecasting scenarios.

Among the four forecast periods analyzed using ARIMA models, the first forecast period demonstrates the lowest predictive accuracy. Several factors may contribute to this discrepancy between the predicted and actual observed data. One possibility is the inherent limitation of the ARIMA model itself—a linear model designed to predict future values based on past data. This model may struggle to capture sudden nonlinear changes or external shocks that occur during the forecast period. ARIMA models assume a degree of stationarity in the data. Therefore, structural breaks or sudden shifts in the underlying time series can reduce the reliability of the model's predictions.

Additionally, significant outliers or unexpected spikes in COVID-19 cases during the forecast period can affect predictive accuracy. Such anomalies may result from the emergence of new virus variants, changes in public health policies, or sudden shifts in public behavior. These rapid increases in case numbers reduce the effectiveness of models trained solely on historical data.

To investigate this hypothesis, outlier detection analysis was conducted on data from January 5 to December 27, 2020. The identified outliers, shown in Table 2 and illustrated in Figure S1, highlight key dates where significant anomalies were observed. These anomalies correspond to periods with sharp increases in case counts, suggesting that forecast discrepancies may be linked to these sudden and unexpected changes.

As shown in Table 2, significant outliers were detected on November 8, November 15, and December 13, 2020, corresponding to sharp rises in cumulative cases. These dates likely reflect specific events or conditions that triggered case surges, such as the emergence of more transmissible variants or changes in testing or reporting practices.

To explore potential anomalies in COVID-19 case trends, an outlier detection analysis was performed on cumulative case data for both the US and global datasets. This analysis aimed to identify time points where actual case numbers significantly deviated from expected trends, potentially indicating periods associated with the emergence and spread of new COVID-19 variants.

Table 2. The detected outliers in COVID-19 cases in the United States of America from January 5 to December 27, 2020

Date reported (year 2020)	Cumulative cases
November 8	9,920,253
November 15	10,925,098
December 13	16,012,396

Figure 3A displays the detected outliers in COVID-19 cases in the US from January 5 to December 27, 2020, with a summary of these outliers provided in Table S1. Notably, several of these dates align with the emergence of significant COVID-19 variants—such as the Omicron variant (B.1.1.529)—which was first identified in November 2021 in South Africa and Botswana.[39] Other variants—such as BQ.1 and BQ.1.1—spread rapidly in late 2022, contributing to the increased number of cases that may have reduced predictive accuracy.[40] Figure 3B presents the time series plot of COVID-19 cases in the US, highlighting the detected outliers.

Further analysis was conducted on a global scale, with the results presented in Figure S2. The corresponding dates and case numbers for the detected global outliers are summarized in Table S1. Similar to the US data, these global outliers correspond to key dates when emerging variants—such as XBB, CH.1.1, and BF.7—were identified and began spreading across various regions, leading to significant increases in case numbers.[41] These variants, first reported in late 2022 and early 2023, significantly impacted regions such as Asia and Europe, leading to significant deviations from the predicted trends.[42]

The detected outliers in both the US and global datasets highlight the significant impact of emerging COVID-19 variants on the spread of the virus. Although the Alpha (B.1.1.7) and Gamma (P.1) variants were not explicitly captured by the outlier detection process—possibly due to their emergence near the end of 2020—the trend illustrated in Figure 3A (US outlier detection plot) exhibits a marked increase in cases during this period.[39] This surge aligns with the period when Alpha and Gamma variants began to spread rapidly, suggesting that their enhanced transmissibility and potential for immune evasion contributed to the surge in case numbers. Consequently, almost all significant surges in the data correspond with the emergence of new variants.

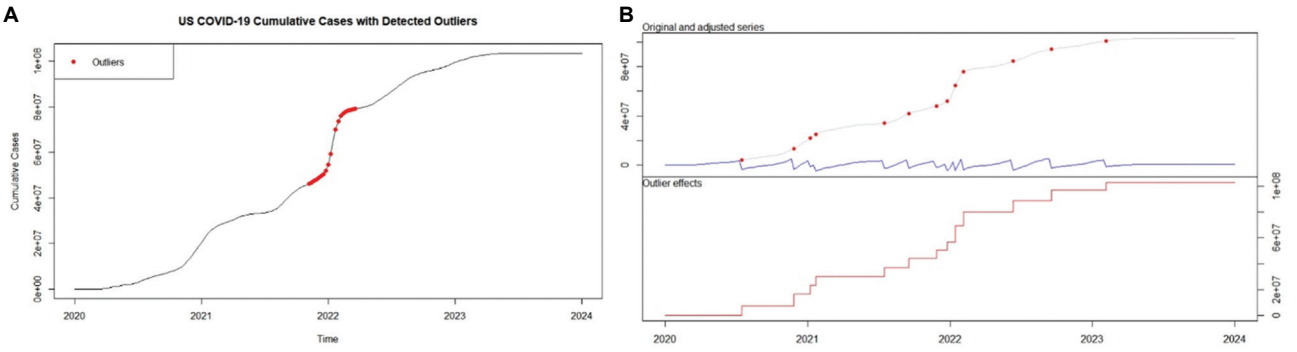

Figure 3. Outlier detection and time series plot of cumulative COVID-19 cases in the United States from January 2020 to December 2023. (A) Detected outliers in the cumulative COVID-19 case data. (B) Time series plot of cumulative COVID-19 cases with detected outliers.
Abbreviation: US: United States.

The outlier dates listed in Tables 1 and 2 closely align with the timelines of variant emergence and global spread. For example, the sharp rise in the US case counts observed between November 2021 and February 2022 aligns with the emergence of the Omicron variant and its subvariants.[40] Similarly, the global spikes identified from late 2021 through 2022 align with the spread of Omicron and its subvariants, further supporting the notion that these variants had a significant impact on the accuracy of predicted versus actual case numbers.

4.2. Regional COVID-19 forecasting across continents

The ARIMA model was employed to forecast COVID-19 cases across various continents, including Asia, Europe, Africa, the Americas, and South America. To isolate trends specific to South America, the Americas dataset excluded Canada, the US, and Mexico. Figure S3 illustrates the forecast results for each continent, with predictions covering the period from January 2020 to early 2021.

In Asia, the ARIMA model's predictions closely align with the actual observed data, effectively capturing the overall upward trend in COVID-19 cases. The prediction intervals encompass the actual case numbers, indicating the model's robustness in this region.

In Europe, the ARIMA model's predictions are less accurate, as the predicted cases significantly deviate from the actual observed data. This discrepancy is particularly evident toward the end of 2020 when a sharp and sudden increase in COVID-19 cases occurred—an event the ARIMA model failed to predict effectively. Based on the forecast's patterns, the earlier anomaly detection for the US and global data, and reports from the WHO on emerging variants, it is plausible to attribute this rapid rise to the Alpha variant (B.1.1.7). First detected in September 2020 in the United Kingdom, this variant was the first to be classified by the WHO as a "variant of concern." Its high transmissibility likely contributes to the significant increase in cases, which is not fully captured by the ARIMA model, thereby highlighting the challenges of forecasting during periods of rapid epidemiological change.

In Africa, the model demonstrates a good fit with the actual observed data, although the sharp rise in cases toward the end of the year pushes the limits of the prediction interval—similar to the pattern observed in Europe. In the Americas, the ARIMA model demonstrates good performance, with predictions closely matching the rapid increase in case numbers. Despite this region experiencing one of the most significant surges in cases, the predictions remain within the confidence intervals, indicating the model's robustness in capturing the trend.

In South America, after excluding the northern countries, the ARIMA model continues to show good model performance. The predicted cases remain within reasonable bounds compared to the observed data, similar to the other continents.

Across all regions, the Ljung-Box test p-values remain well above the 0.05 threshold, indicating no significant autocorrelation in the residuals. This suggests that the ARIMA models successfully capture the temporal patterns of COVID-19 case progression in each region. Occasional underestimations, particularly during rapid case surges, highlight the challenges posed by the pandemic's dynamic nature and the emergence of new variants that earlier model training data may not fully capture. Nonetheless, the ARIMA models demonstrate robust overall performance across various regions, providing valuable insights into the transmission of COVID-19 during the forecast periods.

4.3. Rolling window cross-validation and comparison with auto.arima

In the previous ARIMA forecasting efforts, the auto.arima function was used to automatically select the model parameters *p, d*, and *q*. This function optimizes the model by minimizing the AIC, which balances the model fit and complexity by penalizing excessive parameters. This approach offers several advantages—including speed, automation, and generally reliable results. However, relying solely on AIC may not always produce the most accurate forecasts, especially when working with nonstationary time series or for long-term predictions.

To explore whether other parameter selection methods could improve forecast accuracy, a rolling window cross-validation technique was applied to optimize the *p* and *q* parameters, while the *d* parameter remains fixed as determined by the auto.arima function. The differencing order *d* is fixed because it addresses the time series' stationarity by removing trends or seasonality—a concept well-supported by statistical theory. For example, once a time series is made stationary through differencing, the order of *d* generally remains unchanged to maintain that stationarity, even as *p* and *q* are adjusted.

In this analysis, the period where ARIMA predictions significantly diverged from the actual observed data—such as in the US and Europe from January 5 to December 27, 2020—was examined. These discrepancies are primarily due to sudden surges in cases associated with the emergence of new variants, highlighting the limitations of traditional ARIMA models in capturing such sudden changes.

The rolling window cross-validation approach was employed to evaluate different combinations of *p* and *q* based on the RMSE metric. This approach, which assesses out-of-sample performance across multiple training windows, is particularly valuable for forecasting nonstationary time series with evolving patterns. Table 3 summarizes the RMSE values for the US's ARIMA model using parameters selected through cross-validation, compared to those obtained using auto.arima, while Figure 4A provides a heatmap visualizing RMSE across different *p* and *q* combinations.

As illustrated in Figure 4A, the RMSE heatmap shows that the cross-validated ARIMA parameters (p=2, q = 2) achieve better performance compared to the auto.arima parameters (p=1, q = 0). The heatmap provides a comprehensive view of how different combinations of *p* and *q* affect forecast accuracy, with lower RMSE values indicating better performance.

Furthermore, Figure 4B compares the forecasted COVID-19 cases in the US using ARIMA models with parameters selected by auto.arima and cross-validation. While both models exhibit significant deviations from the actual observed data due to the sudden surge in cases, the cross-validated model's predictions are more closely aligned with the actual observed data than those of auto.arima. This suggests that the cross-validation approach can improve forecast accuracy under certain conditions.

Table 3. Comparison of RMSE values for ARIMA models with parameters selected by auto.arima and cross-validation for COVID-19 case data in the United States

Model	ARIMA parameters			RMSE
	p	*d*	*q*	
auto.arima	1	2	0	27,648.12
Cross-validation-based ARIMA	2	2	2	22,949.3

Abbreviations: ARIMA: Autoregressive integrated moving average; RMSE: Root mean squared error.

A similar approach was employed in the European ARIMA model. Table S2 presents the RMSE values comparing parameters selected by cross-validation and auto.arima, while the RMSE heatmap in Figure S4A visualizes the model performance across different combinations of *p* and *q*.

Figure S4B compares the forecasted COVID-19 cases in Europe using ARIMA models with parameters selected by auto.arima and cross-validation. The forecast line generated by the cross-validated model aligns more closely with the actual observed data than that of auto.arima, although both models show notable deviations from the actual trajectory. These findings are consistent with the results observed in the US, highlighting the potential advantages of using cross-validation for parameter selection in ARIMA models when dealing with highly volatile and non-stationary time series data.

4.4. The effect of vaccination on new COVID-19 cases

Beginning in December 2020, global vaccination efforts against COVID-19 raised a critical question of whether the vaccination campaigns effectively reduce the number of new COVID-19 cases. To address this issue, several statistical methods were applied, including the Granger causality test, segmented regression analysis, the Chow test, and RDD.

The Granger causality test was performed to evaluate whether the number of vaccinated individuals could predict future new COVID-19 cases while accounting for past case counts. Two models were compared: One incorporating lags of both new cases and vaccination counts, and another including only lags of new cases.

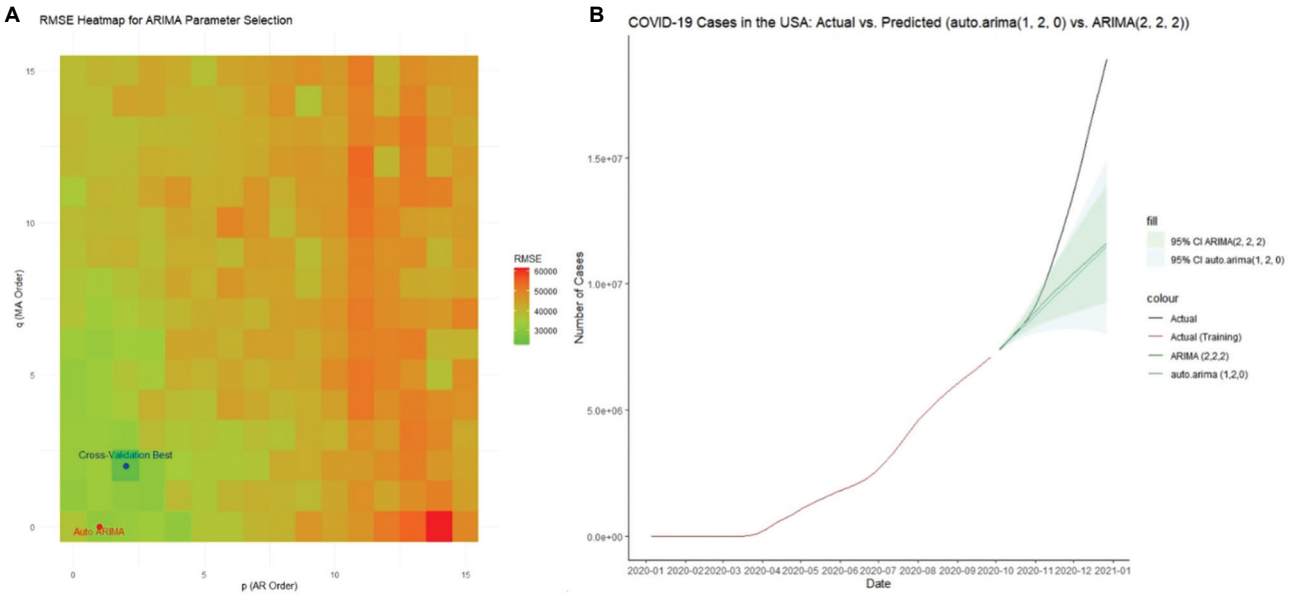

Figure 4. Performance and forecasts of ARIMA models for COVID-19 cases in the United States of America. (A) Heatmap of RMSE for ARIMA models with parameters selected by auto.arima and cross-validation for COVID-19 cases in the United States. (B) Forecast comparison of COVID-19 cases in the United States for ARIMA models with parameters selected by auto.arima and cross-validation.
Abbreviations: ARIMA: Autoregressive integrated moving average; CI: Confidence interval; MA: Moving average; RMSE: Root mean squared error; USA: United States of America.

The results demonstrate no significant causal relationship between vaccination numbers and a reduction in new cases, as evidenced by an *F*-statistic of 0.24 and a p=0.9746. This suggests that the inclusion of vaccination data does not improve the predictive power of the model within the tested lags. Specifically, with a lag of 7 (equivalent to 49 days), the Granger causality test demonstrates no significant effect of vaccination on new cases during this period, as shown in Table 4.

To further investigate the potential impact of vaccination on the trend in COVID-19 cases, segmented regression analysis was employed by introducing a breakpoint at the onset of the vaccination campaign. The regression model included time—an indicator for the post-intervention period—and the interaction between time and the post-intervention phase. The analysis shows that while the overall trend in new cases has a positive slope (β = 18,987, p=0.0214), the interaction term (time–post-intervention) is negative and significant (β = −24,115, p=0.0044), indicating a reduction in the growth rate of new cases following the intervention. However, the post-intervention indicator itself is not statistically significant (p=0.31), which is consistent with the results of the Granger causality test, further suggesting that the immediate effect of vaccination on reducing new cases is not significant. Regardless, the significant negative interaction term suggests that vaccination has a significant long-term effect in reducing new cases, indicating a beneficial impact over time. Figure 5A illustrates the segmented regression results, showing how the predicted number of cases diverges from the actual cases over time. As shown in Table S3, the segmented regression results clearly reflect these trends.

Table 4. The Granger causality test results

Model	Lags	Res. Df	Df	*F*	Pr(> *F*)
Model 1 (new cases and vaccination counts)	New cases=1:7; vaccination counts=1:7	151	-	-	-
Model 2 (new cases only)	New cases=1:7	158	−7	0.24	0.9746

Notes: Df: Degrees of freedom; F: F-statistic; Pr(> F): *P*-value; Res.Df: Residual degrees of freedom.

Additionally, the Chow test was conducted to assess the presence of a structural break at the intervention point. The test provides strong evidence of a structural change, with a $p=6.437\times10^{-6}$, indicating that the introduction of the vaccination campaign significantly alters the underlying relationship between time and new cases. This finding is consistent with the results of the segmented regression analysis, suggesting that vaccination leads to a structural shift in the trend of new cases.

Finally, RDD analysis was applied to further validate these findings. This method focused on the sharp change in the trend of new cases at the intervention point, yielding a conventional coefficient estimate of 76,662.154 with a

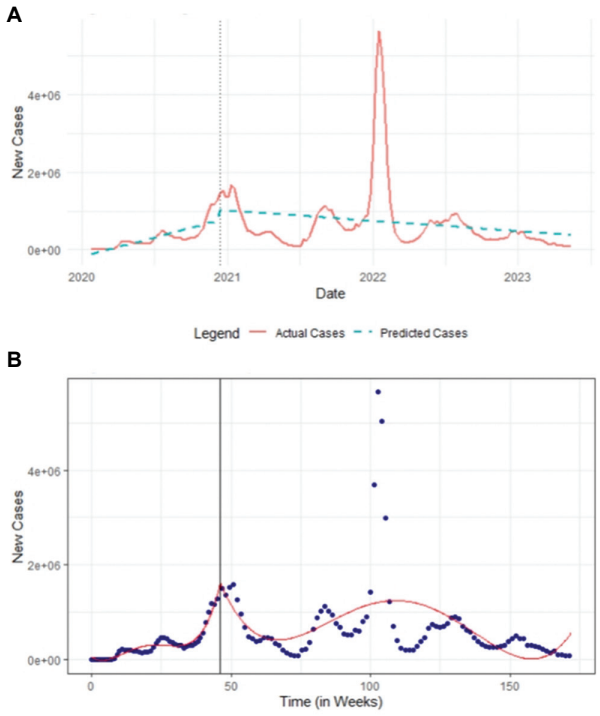

Figure 5. Segmented and regression discontinuity analysis of COVID-19 cases. (A) Segmented regression analysis of COVID-19 cases. (B) Regression discontinuity plot, where blue points represent the observed data and the red line indicates the fitted regression discontinuity model. The vertical black line marks the intervention date (December 13, 2020).

non-significant p=0.636. This suggests that while there may be an observable shift in the trend of new cases at the intervention point, it is not statistically significant at conventional levels. As shown in Table S4, the RDD results further support the idea that the immediate impact of vaccination is not statistically significant. Figure 5B provides a visualization of the RDD results, highlighting the discontinuity at the intervention point. The non-significant result from the RDD analysis is consistent with the findings from both the Granger causality test and the segmented regression analysis, indicating that the immediate impact of vaccination is not significant.

4.5. Forecast on COVID-19 cases using ARIMAX model, with vaccination rates as the exogenous variable

Based on the results in the previous section, where vaccination demonstrates a significant long-term impact on the reduction of new COVID-19 cases, a logical extension was made to incorporate the number of vaccinations as an exogenous variable in ARIMAX models. It was hypothesized that the inclusion of this variable could improve forecast accuracy compared to the standard ARIMA model, which does not account for such external factors.

To test this hypothesis, the dataset starting from December 13, 2020—the beginning of the vaccination campaign—was employed. Given the evidence that the impact of vaccination is more pronounced over time, the first forecast period selected spanned from January 5, 2020, to June 27, 2021—approximately 6 months after vaccination began. This period was used to predict the cumulative number of cases for the subsequent 3 months. Following this, the training data period was gradually extended across 3 time periods:

(i) January 5, 2020 – June 27, 2021,
(ii) January 5, 2020 – December 26, 2021,
(iii) January 5, 2020 – September 25, 2022.

Figures 6, S5, and S6 compare the ARIMA and ARIMAX model predictions across these periods, while Tables 5, S5, and S6 present the evaluation metrics (AIC, RMSE, and MAE) for both models.

These results indicate that the ARIMAX model generally produces forecasts that are closer to the actual data than those of the ARIMA model, as evidenced by lower RMSE and MAE values in certain time periods. However, in some cases, the ARIMAX model exhibits greater deviation from the actual data, resulting in higher RMSE values. Notably, improvements in RMSE and MAE do not always correspond to lower AIC values. For example, in the third period, although the ARIMAX model provides more accurate predictions (reflected by lower RMSE and MAE values), its AIC value is higher than that of the ARIMA model. This highlights the trade-off between model complexity and goodness-of-fit inherent in the AIC calculations.

4.6. Multivariate regression analysis of global COVID-19 infection rates

To investigate the factors influencing COVID-19 infection rates across different countries, it was initially hypothesized that countries with advanced healthcare systems and greater access to medical resources would exhibit lower infection rates. However, an analysis of the top 10 countries by infection rate as of December 31, 2023 (Figure S7) contradicts this assumption. Several highly developed countries, including Luxembourg, Denmark, and Austria, appear among those with the highest infection rates, challenging the initial hypothesis.

To further examine this relationship, a linear regression analysis was conducted using GDP per capita as an indicator of a country's level of development and the COVID-19 infection rate as the outcome variable. The scatterplot with the fitted regression line is illustrated in

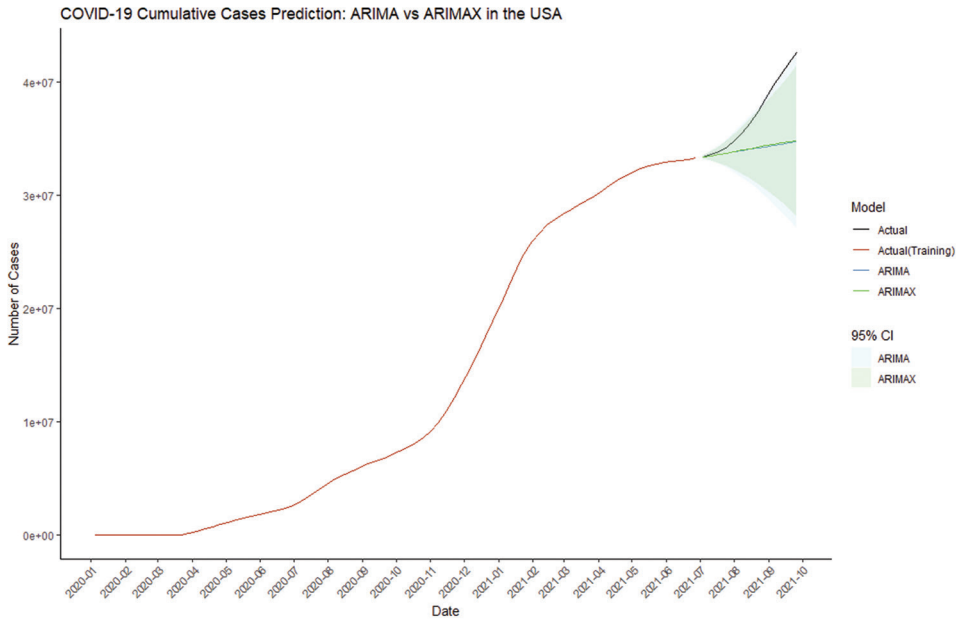

Figure 6. Comparison between ARIMA and ARIMAX models for the first forecast period
Abbreviations: ARIMA: Autoregressive integrated moving average; ARIMAX: Autoregressive integrated moving average with exogenous variables; CI: Confidence interval; USA: United States of America.

Table 5. Comparison between ARIMA and ARIMAX models for the first forecast period

Model	AIC	RMSE	MAE
ARIMA	1,919.556	4,082,257	3,063,789
ARIMAX	1,919.935	4,011,124	3,004,951

Abbreviations: AIC: Akaike information criterion; ARIMA: autoregressive integrated moving average; ARIMAX: autoregressive integrated moving average with exogenous variables; MAE: Mean absolute error; RMSE: Root mean squared error.

Figure 7. The analysis (Table 6) demonstrates a statistically significant positive relationship between GDP per capita and infection rate, with the regression coefficient for GDP per capita being positive and highly significant ($p<2\times10^{-16}$). This suggests that countries with higher GDP per capita tend to have higher reported infection rates. However, the model yields a relatively low R^2 value of 0.4763, suggesting that GDP per capita accounts for approximately 47.63% of the variance in infection rates.

In addition to the regression analysis, three correlation metrics were calculated to further assess the relationship between GDP per capita and COVID-19 infection rate. Pearson's correlation coefficient yields a value of 0.6902, which indicates a moderately strong positive linear relationship between the two variables. Spearman's rank correlation coefficient is higher, at 0.8593, suggesting a strong monotonic relationship. Additionally, the MIC yields a value of 0.7256, reflecting a strong association that may capture nonlinear relationships between GDP per capita and infection rate. Collectively, these correlation measures suggest that higher GDP per capita is associated with increased infection rates, although other factors likely contribute to the remaining unexplained variance.

The relatively low infection rates observed among low-GDP per capita countries may reflect underreporting due to limited testing capacity rather than true differences in transmission. Testing data is sparse and inconsistent across countries, especially in low-income regions, making it difficult to correct this effect quantitatively. Nonetheless, this under-detection is a plausible contributor to the observed pattern.

Given the relatively low R^2, it is evident that factors beyond GDP per capita may influence infection rates. Therefore, the model was expanded to incorporate additional variables that could plausibly affect infection rates, including HDI, Gini coefficient, health expenditure per capita, the number of hospital beds per 1,000 people, and population density. The resulting multivariate regression model incorporated both main effects and interaction terms among these variables.

The analysis reveals a more complex relationship between the predictors and the infection rate. While GDP per capita remains a significant factor (p=0.0065), other variables like health expenditure and certain interaction terms also emerge as significant predictors.

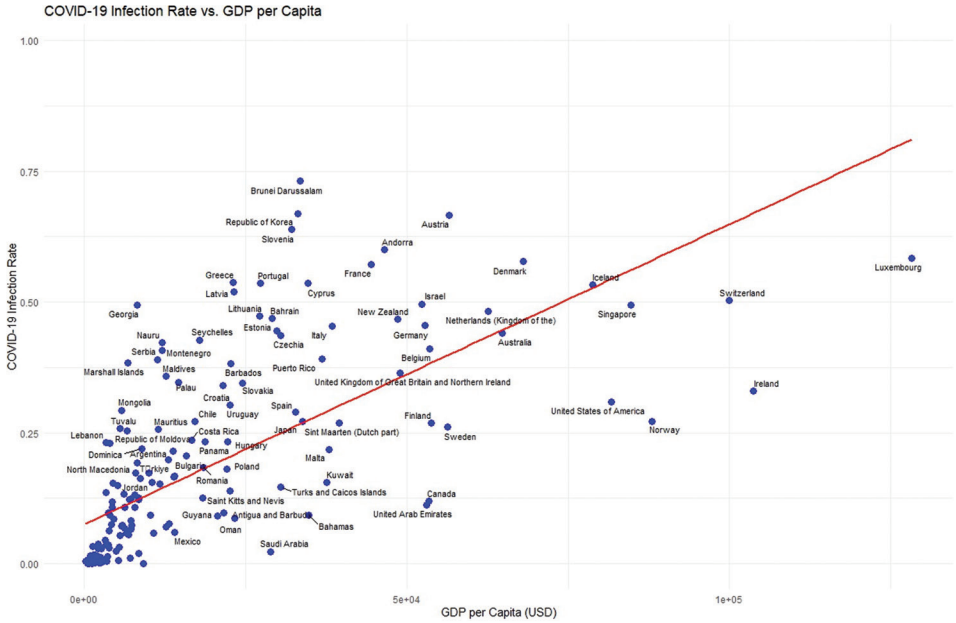

Figure 7. The scatterplot with fitted regression line
Abbreviations: GDP: Gross domestic product; USD: United States dollars.

Table 6. Linear regression results: Infection rate versus gross domestic product (GDP) per capita

Coefficients	Estimate	Standard error	*t*-value	Pr(>|*t*|)
(Intercept)	7.523×10^{-2}	1.254×10^{-2}	6.001	1.05×10^{-8}***
GDP per capita	5.736×10^{-6}	4.470×10^{-7}	12.831	$<2\times10^{-16}$***

Notes: Residuals: Minimum=−0.3404; first quartile: −0.0784; median=−0.0503; maximum=0.4639. Residual standard error=0.1352 on 181 degrees of freedom. Multiple R^2=0.4763; adjusted R^2=0.4734; *F*-statistic=164.6 on 1 and 181 degrees of freedom; $p=2.2\times10^{-16}$. Three asterisks (***) represent $p<0.001$.

For example, the interaction between GDP per capita and HDI (p=0.0064), as well as between GDP per capita and the Gini coefficient (p=0.0297), are both statistically significant. These findings suggest that the effect of GDP per capita on infection rates is moderated by a country's level of HDI and income inequality. Additionally, the interaction between HDI and health expenditure (p=0.0007) is also significant, suggesting that their combined effect significantly influences infection rates. The detailed results of the regression analysis, including coefficients, standard errors, *t*-values, and *p*-values, are provided in Table S7.

Despite these findings, the model's R^2 increased significantly to 0.8179, indicating that approximately 81.79% of the variance in infection rates can be explained by the expanded set of predictors and their interactions. However, residual plots (Figure 8A) reveal potential issues with model fit, including non-constant variance (heteroscedasticity) and deviations from normality, as indicated by the *Q*–*Q* plot.

The scatterplot matrix (Figure 8B) and coefficient plot (Figure S8) further illustrate the complexity of the relationships among the predictors. The scatterplot matrix shows the correlations between variables, with some expected relationships, such as a positive correlation between GDP per capita and HDI (0.729) and a negative correlation between GDP per capita and the Gini coefficient (−0.330). The coefficient plot shows the magnitude and direction of the effects, with GDP per capita, health expenditure, and certain interaction terms having the most pronounced impacts on infection rates.

4.7. Addressing multicollinearity in the regression model

The initial multivariate regression model, which incorporated interaction terms, significantly improved the model's explanatory power, as indicated by a significant increase in the R^2 value. However, this complexity introduced severe multicollinearity, as evidenced by extremely high VIF values. Predictors such as GDP per capita, HDI, and health expenditure, along with their interaction terms, exhibited VIF values in the tens of thousands, indicating that multicollinearity is indeed a significant problem. This multicollinearity can destabilize regression coefficients and complicate their interpretation, thereby necessitating a more rigorous approach to model simplification and stabilization.

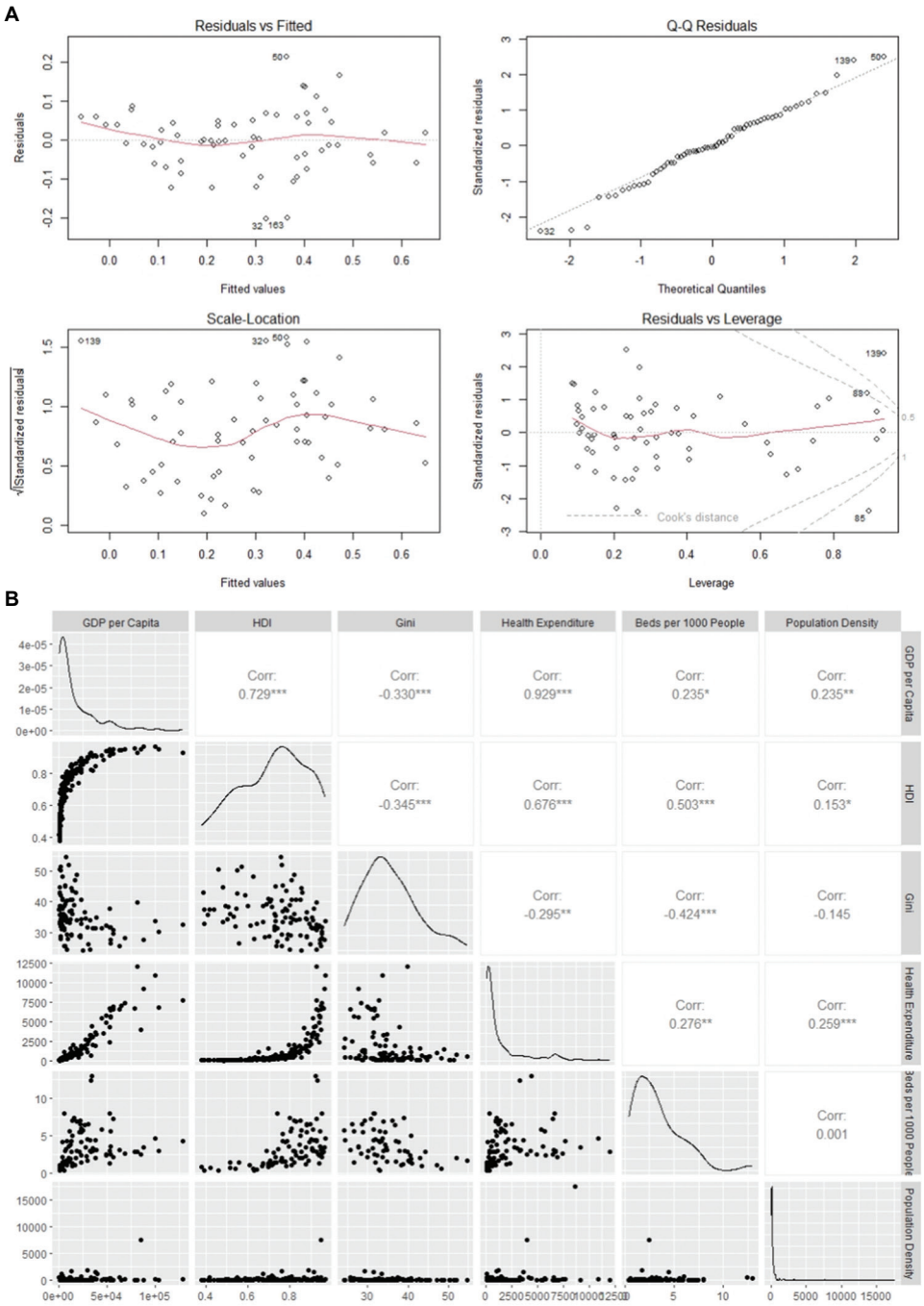

Figure 8. Diagnostic residual plots and correlation analysis of model predictors. (A) Residual plots and (B) scatterplot matrix.
Note: Asterisks (*) indicate levels of statistical significance: $p<0.05$ (*), $p<0.01$ (**), and $p<0.001$ (***).
Abbreviations: GDP: Gross domestic product; HDI: Human development index.

To address these issues, stepwise regression was employed as an initial strategy to reduce model complexity by removing less significant predictors. The simplified model retained key variables and interaction terms that contributed meaningfully to the model's explanatory power while excluding those with minimal impact. Although the resulting model was more manageable, notable multicollinearity persisted, with several VIF values remaining high, albeit reduced from their initial levels.

To further address the issue of multicollinearity, alternative approaches such as PLS and PCR were

employed. These techniques are specifically designed to mitigate multicollinearity by transforming the predictor variables into a set of uncorrelated components.

The PLS and PCR models were applied to the dataset, with each method aiming to reduce the dimensionality of the predictor variables while maximizing the explained variance in the response variable—namely, the infection rate. In particular, the PLS analysis—as shown in Table 7—is effective, demonstrating 67.22% of the variance in infection rates using five components—identified as the optimal number of components through cross-validation (Figure 9). Beyond these components, the model's MSEP begins to increase, suggesting that additional components may introduce noise rather than improve predictive accuracy.

In addition to the error plot, Table S9 provides detailed cross-validation results for each model. This table presents the MSEP for different numbers of components, highlighting how the error decreases as the number of components increases up to five and then rises with the inclusion of additional components. These results support the findings illustrated in Figure 9, which identify five components as optimal.

The component loadings from the PLS model, as illustrated in the heatmap (Figure S9), highlight the contribution of each variable to the principal components. Predictors such as GDP per capita, HDI, and health expenditure demonstrate significant loadings on the first few components, indicating their strong influence on the model. Additionally, more complex interactions—such as those between GDP per capita and population density or health expenditure and population density—play critical roles in the later components.

In addition to the heatmap, detailed PLS loadings are provided in Table S10. This table lists the specific loading values for each variable across the first five components, further illustrating the contributions and interactions among variables in shaping the principal components.

By reducing the predictors into principal components, the PLS model provides a more stable set of coefficients, as shown by the reduced VIF values and improved interpretability of the regression coefficients. The final regression coefficients obtained from the PLS model (Table S11) reveal both the direct and interaction effects of the predictor variables on infection rates, offering a clearer insight into the complex underlying relationships.

In contrast, PCR (Table S8) demonstrates similar results but with slightly lower explained variance for the same number of components. While PCR effectively reduces multicollinearity, this comes at the cost of reduced predictive power compared to PLS. Specifically, PCR explains 59.43% of the variance in infection rates using six components, increasing to 75.81% with 17 components, but it does not exceed the overall performance of the PLS model.

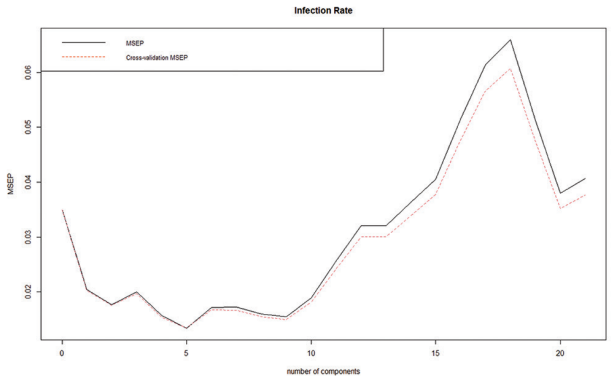

Figure 9. Cross-validation mean squared error plot for the partial least squares model analysis
Abbreviation: MSEP: Mean squared error of prediction.

Table 7. The partial least squares analysis results showing variance explained by the number of components

Number of components	Predictor variables (%)	Infection rates (%)
1	47.98	48.67
2	66.98	55.79
3	81.77	59.75
4	88.76	65.34
5	94.81	67.22
6	97.28	70.14
7	98.05	72.68
8	98.92	73.76
9	99.48	73.93
10	99.78	74.10
11	99.85	74.56
12	99.90	75.47
13	99.95	75.98
14	99.97	76.27
15	99.98	76.49
16	99.99	76.84
17	99.99	77.45
18	100.00	77.79
19	100.00	78.81
20	100.00	81.62
21	100.00	81.79

4.8. Spatial autocorrelation and hotspot analysis of COVID-19 cases across states in the United States

A spatial autocorrelation and hotspot analysis of COVID-19 cases across states in the US was conducted. Spatial autocorrelation assessed the degree to which COVID-19 cases are geographically clustered, while hotspot analysis identified regions with significantly high or low case counts.

Using the most recent COVID-19 case data across states in the US, the number of cases per 100,000 people was calculated to account for population differences. Figure 10A illustrates the spatial distribution of these normalized case counts. States with higher case rates per 100,000 people are shown in red, while those with lower rates are shown in blue. Notably, Alaska and several southern states exhibit particularly high case rates.

In addition to the spatial visualization, Table S12 presents the detailed rankings of the top 10 and bottom 10 states based on COVID-19 cases/100,000 people. Alaska reports the highest case rates, with 40,576.16 cases/100,000 people, followed by Rhode Island and Kentucky. Conversely, New York, Maryland, and Oregon report the lowest case rates, with New York ranking lowest with 18,251.51 cases/100,000 people.

To further understand the spatial pattern, Moran's *I* test—a commonly used measure of spatial autocorrelation—was conducted. The results demonstrate a Moran's *I* value of 0.1578 with a p=0.0317, indicating a significant positive spatial autocorrelation. This suggests that states with high COVID-19 case rates tend to be geographically clustered rather than randomly distributed.

To identify specific clusters of high or low case rates, the Getis-Ord Gi* statistic was employed, providing a measure of local spatial clustering. As shown in Figure 10B, the hotspot analysis reveals several hotspots and coldspots. Southern states—such as Arkansas, Georgia, and Mississippi, as well as parts of Texas and Ohio—are identified as hotspots, with high Gi* values, indicating significant clustering of high case rates. Conversely, states like Alaska, Delaware, New Hampshire, and Vermont are identified as coldspots, suggesting significant clustering of low case rates.

In addition, Table S13 lists the top 10 hotspot and coldspot states along with their corresponding Gi* values. For instance, Arkansas has the highest Gi* value of 1.0004, indicating that it is a significant hotspot, while Alaska has the lowest Gi* value of −1.2373, making it a prominent coldspot. These detailed Gi* values provide a quantitative basis for understanding the spatial clustering patterns observed in the map.

For Alaska, given that it does not share borders with any other state, Washington is designated as its sole neighbor. Despite this adjustment, Alaska—which has the highest infection rate among all states—is classified as a coldspot in the Getis-Ord Gi* hotspot analysis. This counterintuitive result may be due to the isolation of Alaska, where the absence of adjacent states reduces the influence of its high case rate on surrounding areas. Additionally, its high infection rate does not align with a broader regional trend, causing the Gi* statistic to categorize it as a coldspot rather than a hotspot. This highlights the importance of considering geographic and relational context in spatial analyses, particularly for isolated regions.

Figure 10. Geographic distribution and hotspot analysis of COVID-19 cases/100,000 people across the United States of America. (A) COVID-19 cases/100,000 people by state. (B) Getis-Ord Gi* hotspot analysis of COVID-19 cases/100,000 people.

5. Conclusion

The comprehensive statistical analysis of COVID-19 trends—employing ARIMA, ARIMAX, multiple regression, and spatial autocorrelation models—provides valuable insights into the dynamics of the pandemic both globally and within the US. These findings highlight the strengths and limitations of different modeling approaches and the complexity of factors influencing COVID-19 case numbers.

The ARIMA models demonstrate robust performance in predicting short-term COVID-19 trends, particularly when case dynamics follow relatively stable patterns.[14] However, the models show limitations when sudden changes occur in infection rates, such as those caused by sudden policy shifts or the emergence of new virus variants.[43] These situations often reduce predictive accuracy, suggesting that while ARIMA models effectively capture general trends, they may require augmentation or combination with other models to better account for sudden, non-linear changes.[44]

The ARIMAX models, which incorporate exogenous variables such as vaccination data, provide a more nuanced analysis by accounting for external influences on COVID-19 case numbers.[22] However, the effectiveness of the ARIMAX model depends heavily on the specific characteristics of the time period and the data. For instance, during periods when the impact of vaccination on case numbers is delayed or less pronounced, the model struggles to accurately capture the true relationship between variables.[45] This is particularly evident when vaccine uptake is gradual or when vaccination effects take time to appear in the population. Under these conditions, the model may overestimate or underestimate the influence of vaccination, leading to skewed forecasts.[46]

Several challenges arise in applying the ARIMAX model. Firstly, the model assumes a direct and linear effect of the exogenous variable (vaccination rates) on the dependent variable (COVID-19 cases), which may not fully capture the complex, non-linear relationships involved.[14] Factors such as varying vaccine efficacy, the emergence of new virus variants, shifts in public behavior, and policy interventions (e.g., lockdowns, mask mandates) influence the effectiveness of vaccination efforts in reducing case numbers.[47] If these factors are not properly incorporated, the ARIMAX model may incorrectly attribute changes in case numbers to vaccination, leading to inaccurate predictions.

Moreover, including vaccination data as an exogenous variable introduces the risk of multicollinearity, particularly if the vaccination rates correlate with other factors influencing the spread of COVID-19. Multicollinearity causes instability in coefficient estimates, reducing the reliability of the model's predictions.[31] In some cases, this instability leads the ARIMAX model to perform worse than the simpler ARIMA model, which does not encounter this complication.

Timing also plays a crucial role in the performance of the ARIMAX model. The effects of vaccination on COVID-19 cases often involve variable and unpredictable lags.[45] If the model fails to capture the appropriate lag structure, it could lead to inaccurate predictions. For example, the time required for immunity to develop post-vaccination or differences in response across population groups can cause mismatches between vaccination data and observed case changes, further complicating the accuracy of ARIMAX predictions.

Additionally, the ARIMAX model carries a risk of overfitting, especially when it becomes overly complex in relation to the available data. Overfitting occurs when the model captures noise or random fluctuations in the training data as meaningful patterns, reducing its predictive accuracy on new data.[32] This issue becomes more pronounced when vaccination data are included, as the added complexity could reduce the model's generalizability.

In the multiple regression analysis, several socioeconomic factors emerge as significant predictors of COVID-19 case numbers. For example, previous research indicates that factors such as population density, median income, and access to healthcare services demonstrate strong correlations with case numbers.[48] These findings highlight the unequal impact of the pandemic across various demographic groups and regions. Specifically, areas with higher population density and lower income levels tend to report higher case numbers, likely due to the challenges in practicing social distancing and the limited access to healthcare services.[8]

The regression analysis further emphasizes the importance of incorporating a broad range of socioeconomic factors when assessing the spread of COVID-19. However, the model also reveals certain limitations. The relationships between the independent variables and COVID-19 case numbers are not always linear, suggesting the need for more advanced modeling approaches that can capture these complexities.[29] Moreover, the presence of interaction effects among the variables, such as the combined impact of income and healthcare access, suggests that future models should explore these interactions to better understand the pandemic's dynamics.

Spatial autocorrelation analyses provide additional insights, particularly regarding the geographic clustering

of COVID-19 cases. The results reveal significant spatial clusters of high infection rates, suggesting that local factors—such as public health policies, population density, and mobility patterns—play crucial roles in the spread of the virus.[35] These findings suggest that a one-size-fits-all approach is insufficient for managing the pandemic, highlighting the necessity of region-specific strategies.

In conclusion, while the ARIMA and ARIMAX models serve as valuable tools for understanding and predicting COVID-19 trends, their limitations underscore the need for more complex models that can effectively capture the dynamic and non-linear nature of the pandemic.[47] The multiple regression analysis highlights the crucial role of socioeconomic factors in determining COVID-19 case numbers, suggesting that public health interventions should be tailored to address these disparities. The spatial autocorrelation analysis further emphasizes the importance of region-specific strategies in controlling the spread of the virus. Future research should focus on refining these models, incorporating more real-time data, and improving the granularity of spatial analyses to enhance their predictive accuracy and applicability in public health decision-making. Additionally, the effect of vaccination on COVID-19 case numbers, as explored through various statistical techniques, highlights the critical role of timely and effective vaccination efforts in controlling the pandemic. However, the variability in outcomes across different regions suggests that a tailored, region-specific approach is essential for optimizing public health responses.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Dr. Wen Zhang, for his invaluable guidance, insightful feedback, and continuous support throughout the entire process of my research and thesis writing. His expertise, patience, and encouragement have been instrumental in the successful completion of this work.

Funding

None.

Conflict of interest

The author declares no conflicts of interest.

Author contributions

This is a single-authored article.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

All datasets used in this study are listed in Table 1.

References

1. Lai CC, Shih TP, Ko WC, Tang HJ, Hsueh PR. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. *Int J Antimicrob Agents*. 2020;55(3):105924.

   doi: 10.1016/j.ijantimicag.2020.105924

2. Adhikari R, Agrawal RK. An introductory study on time series modeling and forecasting. *arXiv [Preprint]*; 2013.

   doi: 10.48550/arXiv.1302.6613

3. Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data Brief*. 2020;29:105340.

   doi: 10.1016/j.dib.2020.105340

4. Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. *PLoS One*. 2020;15(3):e0231236.

   doi: 10.1371/journal.pone.0231236

5. Bontempi E, Vergalli S, Squazzoni F. Understanding COVID-19 diffusion requires an interdisciplinary, multi-dimensional approach. *Environ Res*. 2020;188:109814.

   doi: 10.1016/j.envres.2020.109814

6. Paltiel AD, Zheng A, Schwartz JL. Speed versus efficacy: Quantifying potential tradeoffs in COVID-19 vaccine deployment. *Ann Intern Med*. 2021;174(4):568-570.

   doi: 10.7326/M20-7866

7. Islam N, Khunti K, Dambha-Miller H, Kawachi I, Marmot M. COVID-19 mortality: A complex interplay of sex, gender and ethnicity. *Eur J Public Health*. 2020;30(5):847-848.

   doi: 10.1093/eurpub/ckaa150

8. Bambra C, Riordan R, Ford J, Matthews F. The COVID-19 pandemic and health inequalities. *J Epidemiol Community Health*. 2020;74(11):964-968.

   doi: 10.1136/jech-2020-214401

9. Box GE, Jenkins GM, Reinsel GC, Ljung GM. *Time Series Analysis: Forecasting and Control*. 5th ed. United States: John Wiley and Sons; 2015.

10. Hamilton JD. Time Series Analysis. United States: Princeton University Press; 1994.

11. Dickey DA, Fuller WA. Distribution of the estimators for autoregressive time series with a unit root. *J Am Stat Assoc*. 1979;74(366a):427-431.

    doi: 10.1080/01621459.1979.10482531

12. Ljung GM, Box GE. On a measure of lack of fit in time series models. *Biometrika*. 1978;65(2):297-303.

doi: 10.1093/biomet/65.2.297

13. Akaike H. A new look at the statistical model identification. In: Parzen E, Tanabe K, Kitagawa G, editors. *Selected Papers of Hirotugu Akaike*. Germany: Springer; 1994. p. 215-232.

doi: 10.1007/978-1-4612-1694-0_16

14. Hyndman RJ, Athanasopoulos G. *Forecasting: Principles and practice*. Australia: OTexts; 2018.

15. Bergmeir C, Benítez JM. On the use of cross-validation for time series predictor evaluation. *Inform Sci.* 2012;191: 192-213.

doi: 10.1016/j.ins.2011.12.028

16. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. *Geosci Model Dev*. 2014;7(3):1247-1250.

doi: 10.5194/gmd-7-1247-2014

17. Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Res*. 2005;30(1):79-82.

18. Hyndman RJ, Khandakar Y. Automatic time series forecasting: The forecast package for R. *J Stat Softw*. 2008;27(3):1-22.

doi: 10.18637/jss.v027.i03

19. Tashman LJ. Out-of-sample tests of forecasting accuracy: An analysis and review. *Int J Forecasting*. 2000;16(4):437-450.

doi: 10.1016/S0169-2070(00)00065-0

20. Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. *ACM Comput Surv.* 2009;41(3):1-58.

doi: 10.1145/1541880.1541882

21. Aggarwal CC. *Outlier Analysis*. 2nd ed. Germany: Springer; 2017.

22. Pankratz A. *Forecasting with Dynamic Regression Models*. United States: John Wiley and Sons; 1991.

23. Granger CWJ. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica.* 1969;37(3):424-438.

doi: 10.2307/1912791

24. Wagner AK, Soumerai SB, Zhang F, Ross-Degnan D. Segmented regression analysis of interrupted time series studies in medication use research. *J Clin Pharm Ther*. 2002;27(4):299-309.

doi: 10.1046/j.1365-2710.2002.00430.x

25. Chow GC. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*. 1960;28(3):591-605.

doi: 10.2307/1910133

26. Imbens GW, Lemieux T. Regression discontinuity designs: A guide to practice. *J Econometrics*. 2008;142(2):615-635.

doi: 10.1016/j.jeconom.2007.05.001

27. Mukaka MM. A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J*. 2012;24(3):69-71.

28. Kutner MH, Nachtsheim CJ, Neter J, Li W. *Applied Linear Statistical Models*. 5th ed. United Staes: McGraw-Hill/Irwin; 2005.

29. Montgomery DC, Peck EA, Vining GG. *Introduction to Linear Regression Analysis.* 5th ed. United States: John Wiley and Sons; 2012.

30. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. Germany: Springer; 2004.

31. O'Brien RM. A caution regarding rules of thumb for variance inflation factors. *Qual Quant.* 2007;41(5):673-690.

doi: 10.1007/s11135-006-9018-6

32. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With Applications in R*. Germany: Springer; 2013.

33. Jolliffe IT. *Principal Component Analysis*. 2nd ed. Germany: Springer; 2002.

34. Wold S, Sjöström M, Eriksson L. PLS-regression: A basic tool of chemometrics. *Chemometr Intell Lab Syst.* 2001;58(2):109-130.

doi: 10.1016/S0169-7439(01)00155-1

35. Anselin L. Local indicators of spatial association-LISA. *Geograph Anal*. 1995;27(2):93-115.

doi: 10.1111/j.1538-4632.1995.tb00338.x

36. Ord JK, Getis A. Local spatial autocorrelation statistics: Distributional issues and an application. *Geograph Anal.* 1995;27(4):286-306.

doi: 10.1111/j.1538-4632.1995.tb00912.x

37. Cliff AD, Ord JK. *Spatial Processes: Models and Applications.* Billerica, MA: Pion; 1981.

38. Getis A, Ord JK. The analysis of spatial association by use of distance statistics. *Geograph Anal.* 1992;24(3):189-206.

doi: 10.1111/j.1538-4632.1992.tb00261.x

39. World Health Organization. *Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern*. Available from: https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern [Last accessed on 2021 Nov 26].

40. World Health Organization. *Tracking SARS-CoV-2 Variants*; 2022. Available from: https://www.who.int/en/activities/tracking-sars-cov-2-variants [Last accessed on 2025 Jun 16].

41. World Health Organization. *Update on Omicron Subvariants and the Global COVID-19 Situation*; 2022. Available from:

https://www.who.int/news-room/feature-stories/detail/update-on-omicron-subvariants-and-the-global-covid-19-situation [Last accessed on 2024 Sep 24].

42. World Health Organization. *Weekly Epidemiological Update on COVID-19*; 2023. Available from: https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19 [Last accessed on 2023 Jan 11].
43. Chowell G, Hyman JM, Castillo-Chavez C. *Mathematical and Statistical Estimation Approaches in Epidemiology.* Germany: Springer; 2021.
44. Liu Z, Magal P, Seydi O, Webb G. Predicting the cumulative number of cases for the COVID-19 epidemic in China from early data. *Math Biosci Eng*. 2020;17(4):3040-3051.

    doi: 10.3934/mbe.2020172
45. Li Q, Guan X, Wu P, *et al*. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N Engl J Med*. 2020;382(13):1199-1207.

    doi: 10.1056/NEJMoa2001316
46. Hernández-Orallo E, Chiner-Oms Á, Rubio-Soler M, *et al*. The importance of considering the impact of COVID-19 variants in forecasting models. *J Ambient Intell Hum Comput*. 2022;13(7):3285-3298.

    doi: 10.1007/s12652-021-03113-x
47. Gao Q, Hu Y, Dai H, *et al*. Modeling COVID-19 with ARIMA and ARIMAX models: A case study in China. *IEEE Access*. 2022;10:55089-55102.

    doi: 10.1109/ACCESS.2022.3182134
48. Wooldridge JM. *Introductory Econometrics: A Modern Approach*. 6th ed. United States: Cengage Learning; 2016.