ACCSCIENCE PUBLISHING

Microbes & Immunity

ORIGINAL RESEARCH ARTICLE

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

Supplementary File

1. Evaluation parameters

Evaluating autoregressive integrated moving average (ARIMA) models involves selecting the most suitable model and measuring forecast accuracy using several key metrics. Common evaluation criteria include the Akaike information criterion (AIC), Bayesian information criterion (BIC), root mean squared error (RMSE), mean absolute error (MAE), and mean squared error (MSE). Each metric offers different insights into the model's performance, making them suitable for various aspects of model evaluation.[1]

1.1. AIC

The AIC is widely used for model selection. It balances model fit and complexity by penalizing models with more parameters to avoid overfitting.[2] AIC is calculated as:

$$\mathrm{AIC} = 2k - 2\ln(L) \quad \text{(SI)}$$

where k is the number of parameters in the model, and L is the likelihood of the model. A lower AIC value indicates a better model performance, as it reflects a good trade-off between model complexity and fit. However, AIC tends to favor slightly more complex models compared to BIC, as it imposes a lighter penalty on the number of parameters.[3]

1.2. BIC

The BIC is another model selection criterion that penalizes model complexity more strongly than AIC.[4] It is calculated as:

$$\mathrm{BIC} = k\ln(n) - 2\ln(L) \quad \text{(SII)}$$

where n is the number of observations. Similar to AIC, a lower BIC value indicates a better model performance; however, BIC tends to favor simpler models, especially when applied to large datasets. Due to its conservative nature, BIC is particularly useful for avoiding overfitting and ensuring generalizability to new data.[1]

1.3. RMSE

RMSE measures the average magnitude of the forecast errors by squaring the differences between the actual and predicted values before averaging them.[1] It is calculated as follows:

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

RMSE is sensitive to large errors due to the squaring of the residuals, making it particularly useful when larger deviations are more critical. However, this sensitivity also means that RMSE can be heavily influenced by outliers.[5]

1.4. MAE

MAE measures the average magnitude of prediction errors without considering their direction. Unlike RMSE, MAE uses the absolute value of the errors, making it less sensitive to outliers.[6] MAE is calculated as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \breve{x}_i\right| \quad \text{(SIV)}$$

MAE provides a linear score, where all errors contribute equally to the metric, making it easier to interpret than RMSE. It is particularly useful when focusing on the average error magnitude, regardless of the size of the deviations.[6]

1.5. MSE

MSE is another common method that measures the average of the squared differences between the actual and predicted values. It is calculated as:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \breve{x}_i\right)^2 \quad \text{(SV)}$$

Similar to RMSE, MSE emphasizes larger errors due to the squaring of the residuals; however, it does not take the square root, making it less interpretable in the same units as the original data.[1] Despite this, MSE is particularly valuable when the goal is to penalize larger errors more severely.

Each of these evaluation metrics offers unique characteristics that make them suitable for different scenarios. AIC and BIC focus on balancing model fit with complexity, with BIC being more conservative.

RMSE and MSE are sensitive to larger errors, making them appropriate when outliers are important, while MAE provides a more robust measure against outliers by treating all errors equally. Depending on the objectives of the analysis, a combination of these metrics is often used to comprehensively assess and compare ARIMA models.[1]

2. Mathematical formulations and detailed analysis

2.1. Granger causality model formulation

The Granger causality test was used to evaluate whether the number of vaccinated individuals could predict future new COVID-19 case numbers, suggesting a potential causal relationship. Generally, a time series, y_t, is hypothesized to be Granger-caused by another series, x_t, if past values of x_t provide statistically significant information about y_t in the presence of past values of y_t.[7] The model is presented as follows:

$$y_t = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{j=1}^{q} \gamma_j x_{t-j} + \epsilon_t \quad \text{(SVI)}$$

where:
(i) y_t represents the number of new COVID-19 cases,
(ii) x_t represents the number of vaccinated individuals,
(iii) ϵ_t represents the error term,
(iv) α represents the intercept,
(v) β_i are the coefficients for the lagged values of y_t,
(vi) γ_j are the coefficients for the lagged values of x_t.

The null hypothesis of the Granger causality test states that the coefficients γ_j are jointly zero, implying that x_t does not Granger cause y_t. In other words, if the null hypothesis is rejected, it indicates that past vaccination rates provide statistically significant predictive power for future COVID-19 cases.

2.2. Segmented regression analysis and chow test formulation

Segmented regression analysis was employed to assess the impact of vaccination on COVID-19 case trends. The model is expressed as:

$$y_t = \beta_0 + \beta_{1\ \text{time}t} + \beta_2\ \text{post-intervention}t + \beta_{3\ \text{time}} - \text{post-intervention}t + \epsilon_t \quad \text{(SVII)}$$

where:
(i) y_t is the number of new COVID-19 cases at time t,
(ii) time_t is the time since the start of the study,
(iii) post-intervention$t$ is a binary variable indicating whether the observation falls after the intervention (e.g., initiation of vaccination),
(iv) time–post-intervention$t$ is the time since the intervention began,
(v) ϵ_t is the error term.

The key coefficients of interest are:
(i) β_2: Represents the immediate level change in new cases after the intervention,
(ii) β_3: Represents the change in trend following the intervention.

To further validate the segmented regression analysis, a Chow test was performed to detect any structural breaks at the point of intervention. The test compares the sum of squared residuals (SSR) from three different models:
(i) The full model, which includes all data across the entire study period,
(ii) The pre-intervention model, which includes only data collected before the initiation of vaccination,
(iii) The post-intervention model, which includes data collected after the vaccination started.

The test statistic is given by:

$$F = \frac{\dfrac{\text{SSR}_{full} - \left(\text{SSR}_{pre} + \text{SSR}_{post}\right)}{k}}{\dfrac{\text{SSR}_{pre} + \text{SSR}_{post}}{n_1 + n_2 - 2k}} \quad \text{(SVIII)}$$

where:
(i) SSR_{full} is the sum of squared residuals from the full model,
(ii) SSR_{pre} and SSR_{post} are the SSR from the pre-intervention and post-intervention models, respectively,
(iii) k is the number of parameters in the model,
(iv) n_1 and n_2 are the number of observations before and after the intervention, respectively.

The null hypothesis of the Chow test states that there is no structural break at the intervention point, indicating that the coefficients remain consistent before and after the intervention. Rejecting the null hypothesis indicates a significant structural break, suggesting that the intervention (e.g., vaccination) caused a change in the trend of new COVID-19 cases.

2.3. Regression discontinuity design (RDD) model formulation

The RDD analysis was employed to estimate the causal effect of vaccine introduction on the number of new COVID-19 cases. The model is expressed as:

$$y_t = \alpha + \beta\ \text{treatment}_t + f(\text{time}_t) + \epsilon_t \quad \text{(SIX)}$$

where:
(i) y_t is the number of new COVID-19 cases at time t,
(ii) treatment_t is the indicator variable equal to 1 if the observation occurred after the cutoff (e.g., after the start of mass vaccination), and 0 otherwise,

(iii) $f(\text{time}_t)$ is a smooth function of time, allowing for flexible trends on either side of the cutoff,
(iv) α is the intercept term,
(v) β represents the treatment effect of vaccination at the cutoff point,
(vi) ϵ_t is the error term.

The RDD approach relies on the assumption that observations near the cutoff are comparable, except for the treatment effect induced by the introduction of vaccines. The non-parametric approach used in this study allows for a flexible functional form for $f(\text{time}_t)$, avoiding restrictive assumptions about the relationship between time and new COVID-19 cases on either side of the cutoff.[8]

2.4. Regression and correlation analyses formulation

The linear regression analysis was used to investigate the relationship between COVID-19 infection rates and economic development. The model is formulated as follows:

$$\text{Infection rate} = \beta_0 + \beta_1 \text{ GDP per capita} + \epsilon \quad \text{(SX)}$$

where:
(i) β_0 is the intercept,
(ii) β_1 is the regression coefficient representing the effect of GDP per capita on the infection rate,
(iii) ϵ is the error term.

In addition to regression analysis, Pearson's, Spearman's, and maximal information coefficient (MIC) were calculated to further and evaluate the strength and direction of the association between GDP per capita and COVID-19 infection rates. The Pearson's correlation coefficient is calculated using the following formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad \text{(SXI)}$$

where:
x_i and y_i represent the individual observations of the two variables being analyzed, $\bar{x}$ and $\bar{y}$ denote the means (averages) of each variable.

The Spearman's rank correlation coefficient is defined as Pearson's correlation between the ranked values of the variables. It assesses the strength of a monotonic relationship between two variables.

In addition, the MIC was measured to capture potential nonlinear relationships between GDP per capita and COVID-19 infection rates. MIC is based on mutual information and measures the strength of the association between two variables without assuming any linear relationship. It is designed to detect both linear and nonlinear dependencies, and its value ranges from 0 (no association) to 1 (perfect association).

2.5. Expanded multiple regression model formulation

The expanded multiple regression model used to investigate the determinants of COVID-19 infection rates is specified as follows:

$$\text{Infection rate} = \beta_0 + \beta_1 \cdot \text{GDP per capita} + \beta_2 \cdot \text{HDI} + \beta_3 \cdot \text{Gini} + \beta_4 \cdot \text{Health expenditure} + \beta_5 \cdot \text{Beds per 1,000 people} + \beta_6 \cdot \text{Population density} + \epsilon \quad \text{(SXII)}$$

Where:
(i) β_0 is the intercept,
(ii) β_1 GDP per capita represents the coefficient for GDP per capita,
(iii) β_2 HDI represents the coefficient for HDI,
(iv) β_3 Gini represents the coefficient for the Gini coefficient,
(v) β_4 Health expenditure represents the coefficient for per capita health expenditure,
(vi) β_5 Beds per 1,000 represents the coefficient for the number of hospital beds per 1,000 people,
(vii) β_6 Population density represents the coefficient for population density,
(viii) ϵ is the error term

Interaction terms were included to investigate the potential synergistic effects between these variables. For instance, the interaction between health expenditure and GDP per capita was examined to assess whether healthcare investment influences the relationship between economic development and infection rates. Additionally, the interactions between population density and other socioeconomic factors were analyzed to assess the impact of urbanization on infection spread.[9]

2.6. Principal component regression (PCR) and partial least squares (PLS) regression model formulation

PCR involves applying principal component analysis to the predictor variables. The resulting principal components are then used as predictors in the regression model. The PCR model is formulated as follows:

$$\textit{Infection Rate} = \alpha_0 + \sum_{i=1}^{k} \alpha_i \cdot \text{PC}_i + \dot{o} \quad \text{(SXIII)}$$

where:
(i) α_0 is the intercept term,
(ii) PC_i is the principal component derived from the original predictor variables,
(iii) α_i is the regression coefficient corresponding to the i-th principal component,

(iv) k is the number of principal components included in the model,
(v) ϵ is the error term.

Principal components are uncorrelated, with the first few components capturing the maximum variance in the predictor variables. Cross-validation was employed to select the optimal number of components to include in the model, balancing model complexity and prediction accuracy.[10]

PLS regression is similar to PCR but extends the approach by incorporating the covariance between predictors and the dependent variable when determining the components. Unlike PCR, PLS typically requires fewer components as it identifies components that are more directly related to the dependent variable.[11] The PLS model can be expressed as:

$$Infection\,Rate = \beta_0 + \sum_{i=1}^{m} \beta_i \cdot \mathrm{PLS}_i + \varepsilon \quad \text{(SXIV)}$$

where:
(i) β_0 is the intercept term,
(ii) PLS_i is the i-th PLS component,
(iii) β_i is the coefficient corresponding to the i-th PLS component,
(iv) m is the number of PLS components included in the model,
(v) ϵ is the error term.

Similar to PCR, cross-validation was performed to determine the optimal number of components for PLS. The models were evaluated based on the MSE of prediction, and component loadings were analyzed to interpret the contribution of the original variables to the extracted components.[8]

2.7. Mathematical formulation of Moran's *I*

Moran's I is a measure of global spatial autocorrelation that quantifies the degree of spatial clustering of a variable across geographic space. Mathematically, Moran's I is expressed as:

$$I = \frac{N}{W} \cdot \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2} \quad \text{(SXV)}$$

where:
(i) N is the total number of spatial units (e.g., states or regions),
(ii) x_i and x_j are the values of the variable of interest (e.g., COVID-19 infection rates) at locations i and j,
(iii) x^- is the mean value of the variable,
(iv) w_{ij} is the spatial weight between locations i and j, indicating the strength of the spatial relationship (e.g., based on shared boundaries),

- W is the sum of all spatial weights, i.e., $W = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}$.[12]

Moran's I range from −1 to 1, where:
(i) $I > 0$ indicates positive spatial autocorrelation, meaning similar values are spatially clustered,
(ii) $I < 0$ indicates negative spatial autocorrelation, meaning dissimilar values are adjacent,
(iii) $I = 0$ suggests a random spatial distribution of values.

For this analysis, the spatial weights matrix was generated based on shared boundaries between geographic regions. Moran's I was measured to assess the overall spatial autocorrelation of COVID-19 infection rates, using boundary-based spatial relationships to understand the clustering behavior of infection rates.[13]

2.8. Mathematical formulation of the Getis-Ord Gi* statistic

The Getis-Ord Gi* statistic is a local spatial measure used to identify hotspots (areas with high-value clustering) and coldspots (areas with low-value clustering) within a geographic region. The Gi* statistic for a location i is calculated as:

$$G_i^* = \frac{\sum_{j=1}^{N} w_{ij}x_j - \bar{X}\sum_{j=1}^{N} w_{ij}}{S\sqrt{\frac{\sum_{j=1}^{N} w_{ij}^2 - \left(\sum_{j=1}^{N} w_{ij}\right)^2 / N}{N-1}}} \quad \text{(SXVI)}$$

where:
(i) x_j represents the value of the variable of interest (e.g., COVID-19 infection rates) at location j,
(ii) X^- is the mean value of the variable across all locations,
(iii) S is the standard deviation of the variable,
(iv) w_{ij} is the spatial weight between locations i and j, indicating the strength of their spatial relationship,
(v) N is the total number of spatial units (e.g., regions or states).[14]

A significantly positive G_i^* value indicates the presence of a hotspot (i.e., clustering of high values), while a significantly negative G_i^* value indicates a coldspot (i.e., clustering of low values). The significance of the G_i^* values is determined by comparing the observed statistic to a reference distribution under the null hypothesis of spatial randomness.[38]

For this analysis, the same spatial weights matrix was used to compute the Getis-Ord Gi* statistic, which identifies geographic regions with significant clustering of COVID-19 infection rates. These regions were classified as hotspots or coldspots based on the sign and significance of the Gi* values.

3. Detected outliers in global and United States COVID-19 cases

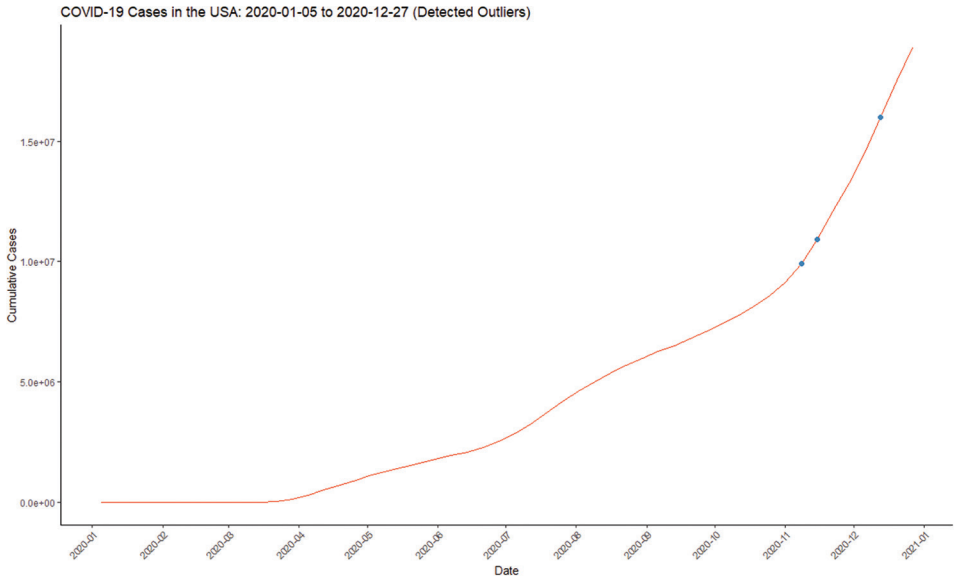

Figure S1. Detected outliers in COVID-19 cases in the United States of America from January 5, 2020, to December 27, 2020
Abbreviation: USA: United States of America.

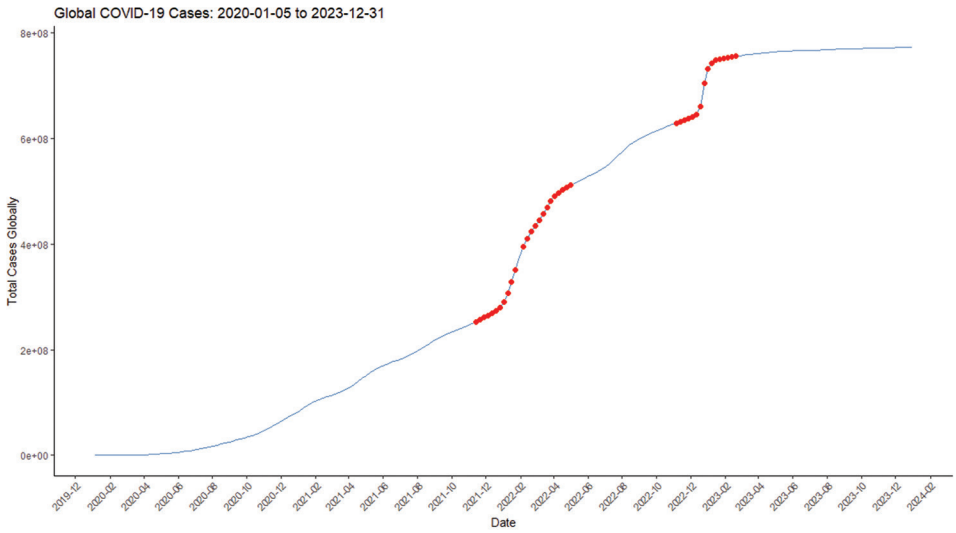

Figure S2. Global COVID-19 cases outliers detected from January 5, 2020, to December 31, 2023

Table S1. Detected outlier points in the United States of America and globally

Year	Date	Cumulative cases (millions)	
		United States of America	Global
2021	November 7	46.1	-
	November 14	46.7	253.3
	November 21	47.4	257.2
	November 28	48.0	261.2
	December 5	48.7	265.5
	December 12	49.6	269.8
	December 19	50.5	274.5
	December 26	51.9	280.6
2022	January 2	54.6	291.1
	January 9	59.3	307.7
	January 16	-	328.3
	January 23	69.9	351.9
	January 30	73.8	-
	February 6	75.9	395.2
	February 13	77.2	411.1
	February 20	77.9	424.0
	February 27	78.3	435.1
	March 6	78.7	446.0
	March 13	78.9	457.8
	March 20	79.1	470.2
	March 27	-	481.2
	April 3	-	490.5
	April 10	-	497.7
	April 17	-	503.4
	April 24	-	508.1
	May 1	-	512.1
	November 6	-	629.4
	November 13	-	631.9
	November 20	-	634.6
	November 27	-	637.6
	December 4	-	640.8
	December 11	-	645.4
	December 18	-	660.4
	December 25	-	704.6
2023	January 1	-	732.4
	January 8	-	743.3
	January 15	-	748.2
	January 22	-	750.7
	January 29	-	752.3
	February 5	-	753.7
	February 12	-	754.9
	February 19	-	756.1

4. ARIMA model analysis across different continents

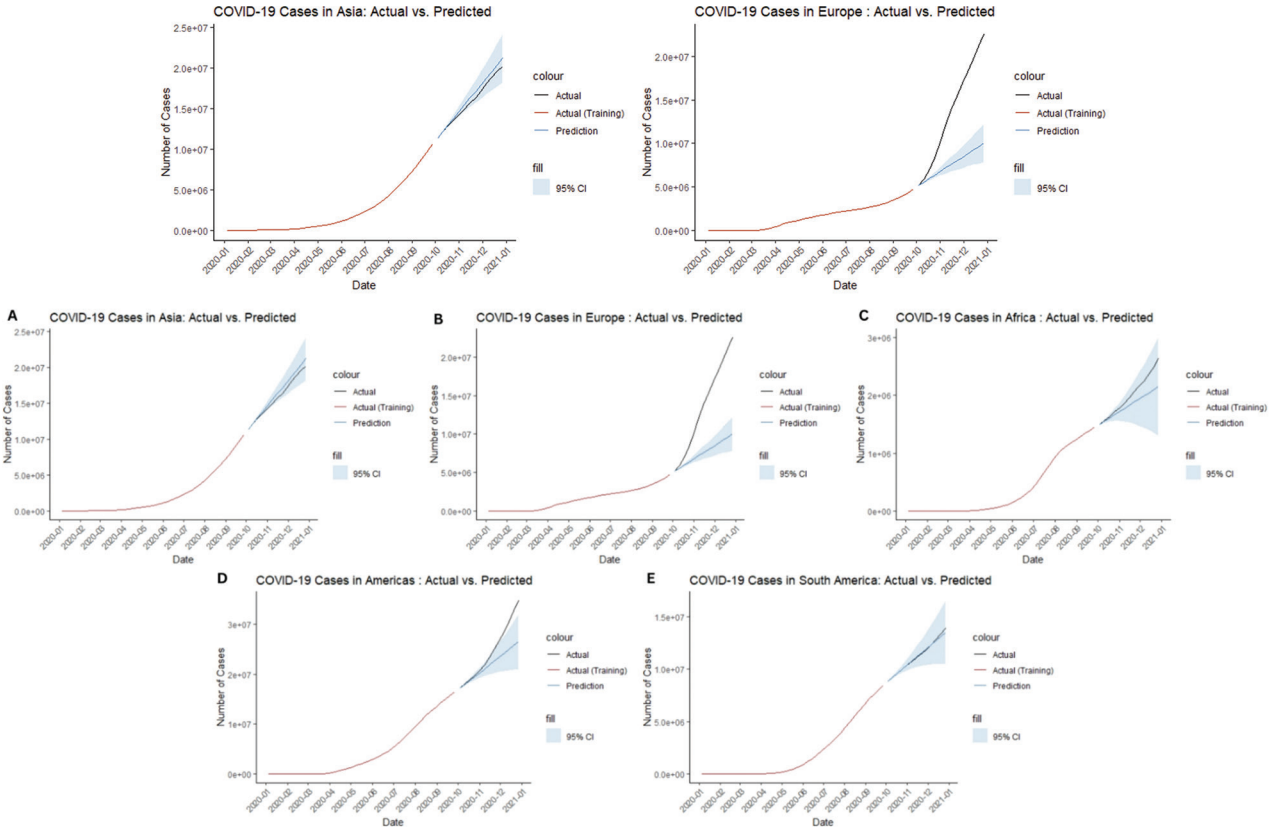

Figure S3. Autoregressive integrated moving average model analysis across different continents. (A) Asia; (B) Europe; (C) Africa; (D) Americas; (E) South America.
Abbreviation: CI: Confidence interval.

5. Comparison of RMSE values and forecast analysis for ARIMA models for COVID-19 cases in Europe

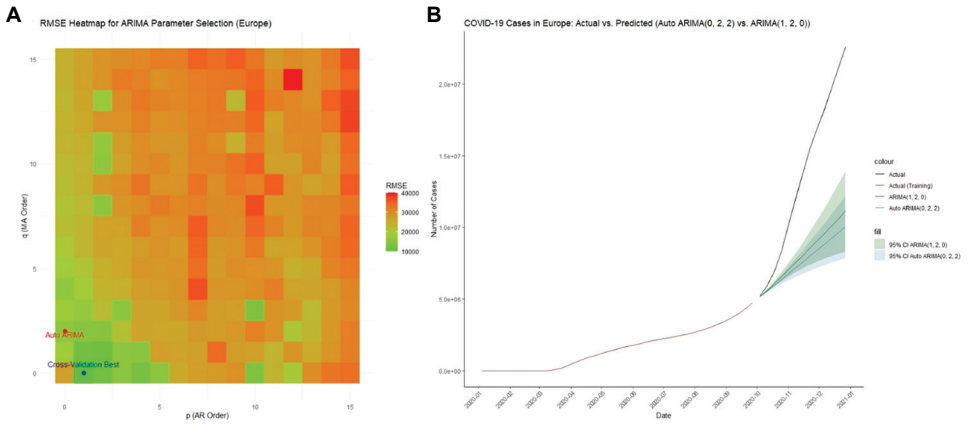

Figure S4. Performance and forecasts of ARIMA models for COVID-19 cases in Europe. (A) Heatmap of RMSE for ARIMA models with parameters selected by auto.arima and cross-validation for COVID-19 cases in Europe. (B) Forecast comparison of COVID-19 cases in Europe for ARIMA models with parameters selected by auto.arima and cross-validation.
Abbreviations: ARIMA: Autoregressive integrated moving average; CI: Confidence interval; MA: Moving average; RMSE: Root mean squared error.

Table S2. Comparison of RMSE values for autoregressive integrated moving average models with parameters selected by auto.arima and cross-validation for COVID-19 case data in Europe

Model	ARIMA parameters			RMSE
	p	*d*	*q*	
auto.arima	0	2	2	14,479.46
Cross-validation-based ARIMA	1	2	0	9,981.78

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

6. Detailed results of segmented regression and RDD analysis

Table S3. Results of segmented regression analysis

Coefficients	Estimate	Standard error	*t*-value	Pr(>|t|)
Intercept	−11,1033	213,511	−0.520	0.6037
Time	18,987	8,173	2.323	0.0214*
Post-intervention	260,639	25,5961	1.018	0.3100
Time–post-intervention	−24,115	8,365	−2.883	0.0044**

Note: Residuals: Minimum=−803,922; first quartile: −274,147; median=−105,598; third quartile=81,517; maximum=4,920,221. Residual standard error=735,900 on 169 degrees of freedom. Multiple R^2=0.1192; adjusted R^2=0.1036; *F*-statistic=7.623 on 3 and 169 degrees of freedom; $p=8.242\times10^{-5}$. Asterisks (*) indicate levels of statistical significance: $p<0.05$ (*) and $p<0.01$ (**).

Table S4. Results of regression discontinuity design analysis

Left of cutoff Right of cutoff				
Number of observations		46		127
BW type			mserd	
Kernel			Triangular	
VCE method			NN	
Effect of number of observations		5	6	
Order estimate (p)		1	1	
Order bias (q)		2	2	
BW estimate (h)		5.520	5.520	
BW bias (b)		9.645	9.645	
rho (h/b)		0.572	0.572	
Unique observations		46	127	
Method (95% CI; XXX)	Coefficient	Standard error	*z*	*p*>|*z*|
Conventional (−2,411,040.87; 394,465.18)	76,662.15	162,147.38	0.473	0.636
Robust (−315,644.73; 470,854.24)	-	-	0.387	0.699

Abbreviations: BW: Bandwidth; CI: Confidence interval; mserd: Mean squared error optimal bandwidth selector for regression discontinuity; NN: Nearest-neighbor (used as a variance-covariance estimation method); rho: Ratio of estimation bandwidth to bias bandwidth (i.e., $\rho=h/b$ρ=h/b); VCE: Variance-covariance estimation.

7. Detailed results of ARIMAX model

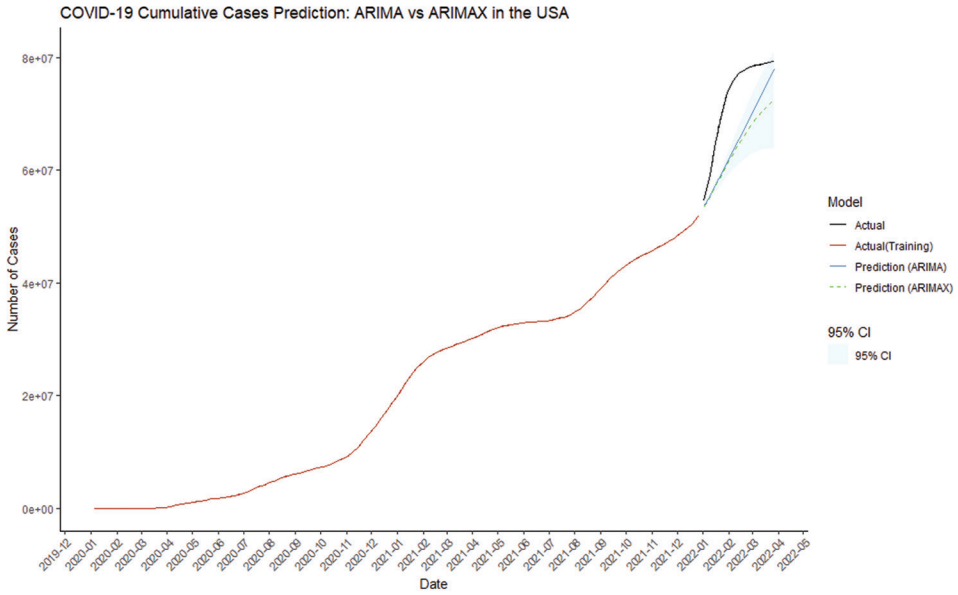

Figure S5. Comparison between ARIMA and ARIMAX models for the second forecast period
Abbreviations: ARIMA: Autoregressive integrated moving average; ARIMAX: Autoregressive integrated moving average with exogenous variables; CI: Confidence interval; USA: United States of America.

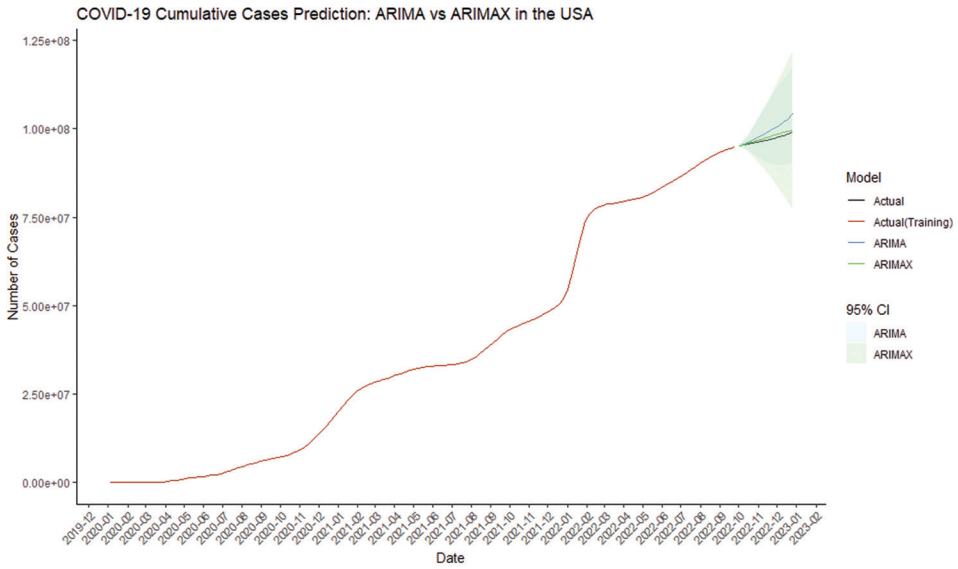

Figure S6. Comparison between ARIMA and ARIMAX models for the third forecast period
Abbreviations: ARIMA: Autoregressive integrated moving average; ARIMAX: Autoregressive integrated moving average with exogenous variables; CI: Confidence interval; USA: United States of America.

Table S5. Comparison between ARIMA and ARIMAX models for the second forecast period

Model	AIC	RMSE	MAE
ARIMA	2,633.136	8,288,456	7,301,748
ARIMAX	2,657.777	9,578,389	8,950,037

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.

Table S6. Comparison between ARIMA and ARIMAX models for the third forecast period

Model	AIC	RMSE	MAE
ARIMA	2,434.937	2,648,795.3	2,178,441.3
ARIMAX	3,794.536	617,398.9	545,843.7

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.

8. Top 10 countries by COVID-19 infection rate as of December 31, 2023

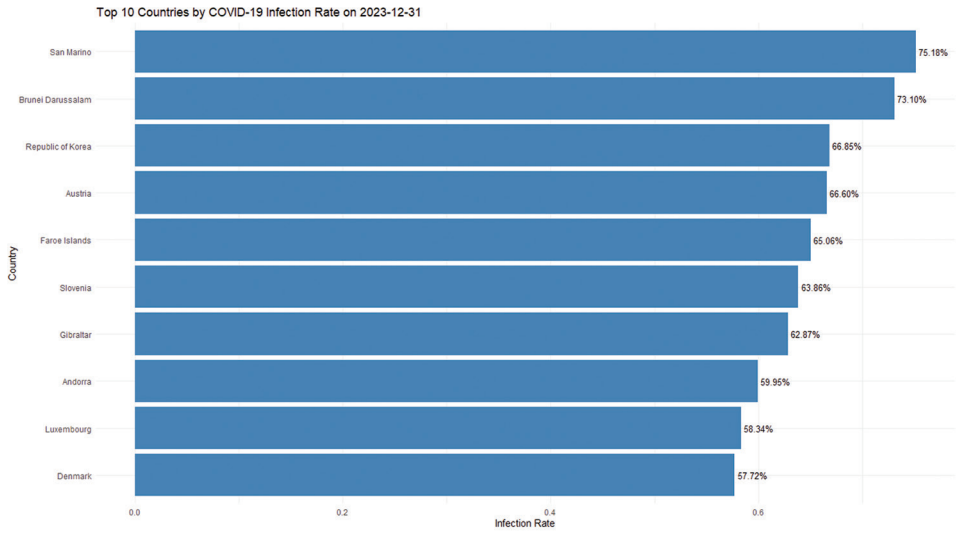

Figure S7. Top ten countries by COVID-19 infection rate as of December 31, 2023, identified through multiple regression analysis

9. Coefficient plot for multiple regression analysis

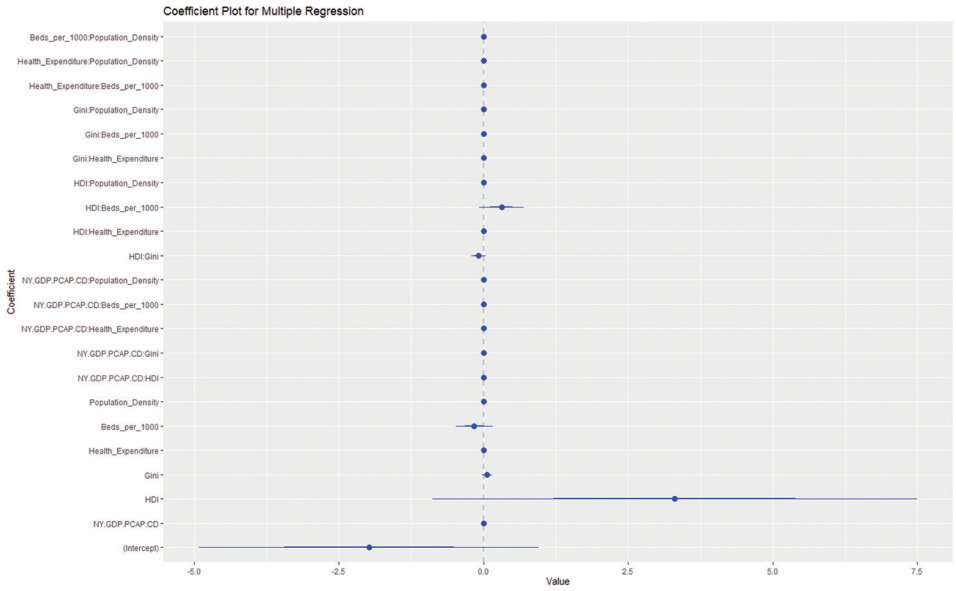

Figure S8. Coefficient plot for multiple regression analysis
Abbreviations: HDI: Human development index; NY.GDP.PCAP.CD: GDP per capita.

10. Multiple regression results

Table S7. Results of multiple regression analysis: Infection rate versus socioeconomic and health factors

Coefficients	Estimate	Standard error	*t*-value	Pr(>|t|)
Intercept	-1.983×10	1.466×10	−1.353	0.1838
GDP per capita	-2.556×10^{-4}	8.909×10^{-5}	−2.869	0.0065**
HDI	3.305×10	2.095×10	1.578	0.1225
Gini	5.595×10^{-2}	4.314×10^{-2}	1.297	0.2021
Health expenditure	3.226×10^{-3}	9.324×10^{-4}	3.460	0.0013**
Number of beds per 1,000 people	-1.624×10^{-1}	1.589×10^{-1}	−1.022	0.3129
Population density	-1.753×10^{-3}	2.132×10^{-3}	−0.822	0.4160
GDP per capita–HDI	1.937×10^{-4}	6.730×10^{-5}	2.879	0.0064**
GDP per capita–Gini	2.490×10^{-6}	1.104×10^{-6}	2.255	0.0297*
GDP per capita–health expenditure	9.444×10^{-10}	5.735×10^{-10}	1.647	0.1075
GDP per capita– number of beds per 1,000 people	-2.560×10^{-6}	2.457×10^{-6}	−1.042	0.3037
GDP per capita–population density	-5.759×10^{-9}	2.454×10^{-8}	−0.235	0.8156
HDI–Gini	-8.932×10^{-2}	6.185×10^{-2}	−1.444	0.1565
HDI–health expenditure	-2.918×10^{-3}	7.967×10^{-4}	−3.662	0.0007***
HDI–number of beds per 1,000 people	3.109×10^{-1}	1.936×10^{-1}	1.606	0.1162
HDI–population density	3.461×10^{-3}	2.981×10^{-3}	1.161	0.2525
Gini–health expenditure	-1.901×10^{-5}	8.145×10^{-6}	−2.333	0.0247*
Gini–number of beds per 1,000 people	-1.251×10^{-3}	1.789×10^{-3}	−0.699	0.4885
Gini–population density	$-1.178\times10-^{5}$	$3.779\times10-^{5}$	−0.312	0.7569
Health expenditure–number of beds per 1,000 people	$1.944\times10-^{5}$	$1.622\times10-^{5}$	1.198	0.2378
Health expenditure–population density	$5.495\times10-^{8}$	$2.320\times10-^{7}$	0.237	0.8140
Number of beds per 1,000 people–population density	$-2.616\times10-^{4}$	$1.166\times10-^{4}$	−2.244	0.0304*

Note: Residuals: Minimum=−0.2023; first quartile: −0.0443; median=−0.0028; third quartile = 0.0474; maximum = 0.2142. Residual standard error = 0.0977 on 40 degrees of freedom. Multiple $R^2 = 0.8179$; adjusted $R^2 = 0.7223$; *F*-statistic = 8.554 on 21 and 40 degrees of freedom; *P*-value = 4.634×10^{-9}. 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.

11. Results for PCR and PLS analyses: RMSE of prediction values

Table S8. The principal component regression analysis results showing variance explained by the number of components

Number of components	Predictor variables (%)	Infection rates (%)
1	49.17	40.39
2	71.73	44.89
3	86.52	52.61
4	94.13	52.86
5	96.52	58.47
6	97.94	59.43
7	98.77	64.55
8	99.45	65.58
9	99.66	69.19
10	99.80	73.38
11	99.89	73.68
12	99.95	73.87
13	99.97	73.89
14	99.98	73.90
15	99.99	73.96
16	99.99	74.09
17	100.00	75.81
18	100.00	75.98
19	100.00	76.00
20	100.00	76.56
21	100.00	81.79

Table S9. Cross-validation results of Partial Least Squares

Number of components	Cross-validation	Adjusted cross-validation
Intercept	0.1868	0.1868
1	0.1430	0.1423
2	0.1331	0.1325
3	0.1416	0.1402
4	0.1253	0.1242
5	0.1158	0.1159
6	0.1309	0.1293
7	0.1313	0.1290
8	0.1266	0.1246
9	0.1243	0.1225
10	0.1377	0.1348
11	0.1601	0.1558
12	0.1791	0.1733
13	0.1791	0.1734
14	0.1905	0.1841
15	0.2013	0.1943
16	0.2269	0.2184
17	0.2478	0.2379
18	0.2568	0.2464
19	0.2265	0.2179
20	0.1950	0.1877
21	0.2016	0.1941

12. PLS loadings and final regression coefficients

Table S10. Partial least squares loadings

Variable	Number of components				
	1	2	3	4	5
NY.GDP.PCAP.CD	0.297	-	−0.200	0.171	-
HDI	0.263	0.242	0.366	-	−0.218
Gini	−0.172	−0.106	−0.174	-	−0.840
Health expenditure	0.293	-	−0.255	0.130	-
Number of beds per 1,000 people	0.160	0.364	0.320	−0.375	-
Population density	-	−0.386	0.518	−0.265	-
NY.GDP.PCAP.CD–HDI	0.297	-	−0.206	0.170	-
NY.GDP.PCAP.CD–Gini	0.289	−0.100	−0.240	0.172	−0.115
NY.GDP.PCAP.CD–health expenditure	0.274	−0.128	−0.384	-	-
NY.GDP.PCAP.CD–number of beds per 1,000 people	0.306	-	-	-	0.125
NY.GDP.PCAP.CD–population density	0.186	−0.336	0.407	−0.117	-
HDI–Gini	-	-	0.365	-	−1.086
HDI–health expenditure	0.293	-	−0.259	0.129	-
HDI–number of beds per 1,000 people	0.192	0.341	0.316	−0.321	-
HDI–population density	-	−0.375	0.536	−0.241	-
Gini–health expenditure	0.282	−0.104	−0.298	0.115	−0.175
Gini–number of beds per 1,000 people	0.128	0.365	0.287	−0.388	−0.345
Gini–population density	-	−0.399	0.494	−0.270	−0.105
Health expenditure–number of beds per 1,000 people	0.298	-	−0.122	-	-
Health expenditure–population density	0.188	−0.330	0.414	−0.126	-
Number of beds per 1,000 people–population density	0.125	−0.280	0.569	−0.321	-

Abbreviations: HDI: Human development index; NY.GDP.PACP.CD: GDP per capita.

Table S11. Final regression coefficients

Variable	Coefficient
NY.GDP.PCAP.CD	0.0130
HDI	0.0794
Gini	−0.0442
Health expenditure	−0.0045
Number of beds per 1,000 people	0.0043
Population density	−0.0197
NY.GDP.PCAP.CD–HDI	0.0092
NY.GDP.PCAP.CD–Gini	0.0143
NY.GDP.PCAP.CD–health expenditure	−0.0607
NY.GDP.PCAP.CD–number of beds per 1,000 people	0.0276
NY.GDP.PCAP.CD–population density	0.0430
HDI–Gini	0.0126
HDI–health expenditure	−0.0099
HDI–number of beds per 1,000 people	0.0231
HDI–population density	−0.0092
Gini–health expenditure	−0.0067
Gini–number of beds per 1,000 people	−0.0164
Gini–population density	−0.0267
Health expenditure–number of beds per 1,000 people	0.0223
Health expenditure–population density	0.0404
Number of beds per 1,000 people–population density	−0.0369

Abbreviations: HDI: Human development index; NY.GDP.PCAP.CD: GDP per capita.

13. PLS model component analysis

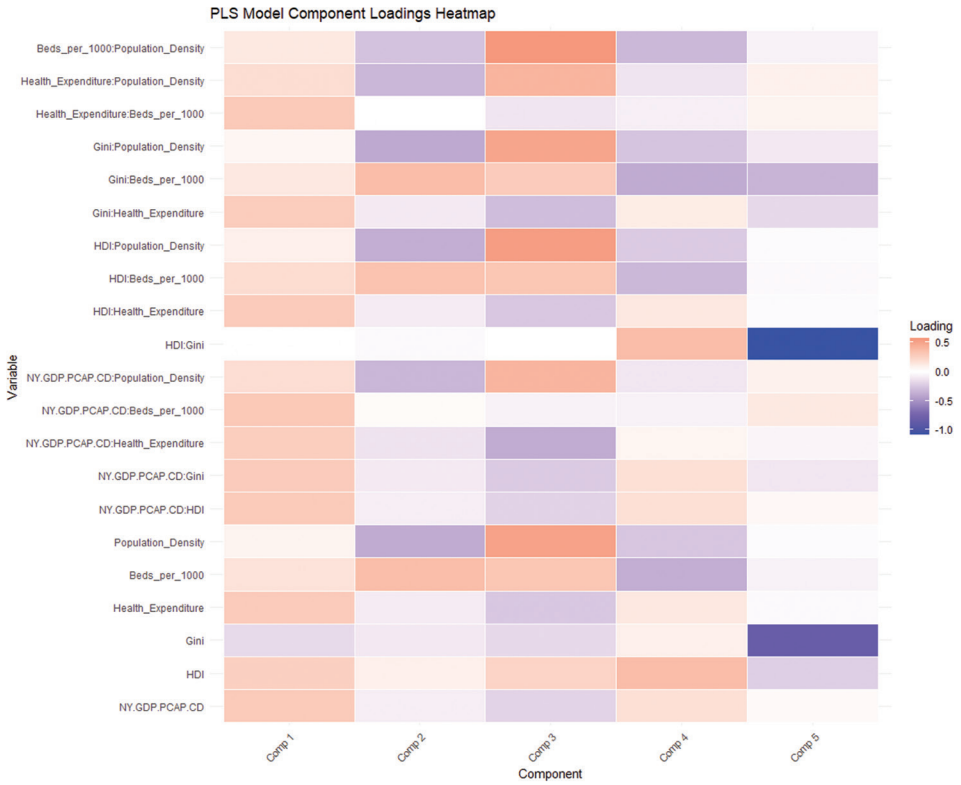

Figure S9. Heatmap of component loadings from the partial least squares (PLS) analysis Abbreviations: HDI: Human development index; NY.GDP.PCAP.CD: GDP per capita.

14. State-level analysis of COVID-19 cases, hotspots and coldspots

Table S12. Top ten and bottom ten states by COVID-19 cases per 100,000 people

Rank	Top 10 states		Bottom 10 states	
	State	Cases per 100,000	State	Cases per 100,000
1	Alaska	40,576.16	New York	18,251.51
2	Rhode Island	40,281.14	Maryland	22,319.23
3	Kentucky	38,512.10	Oregon	23,051.58
4	North Dakota	37,132.71	Maine	23,140.85
5	West Virginia	36,753.10	Vermont	23,822.64
6	Tennessee	35,672.03	Washington	25,058.10
7	Louisiana	34,995.58	Hawaii	26,078.61
8	South Carolina	34,465.43	District of Columbia	26,349.25
9	Wisconsin	34,355.14	Virginia	26,513.61
10	Mississippi	34,031.31	Idaho	26,778.19

Table S13. Hotspot and coldspot states with their corresponding Gi* values

Rank	Hotspot states		Coldspot states	
	State	Gi* value	State	Gi* value
1	Arkansas	1.0004	Alaska	−1.2373
2	Georgia	1.5828	Delaware	−1.1393
3	Mississippi	1.5567	New Hampshire	−1.6934
4	Missouri	1.3959	Idaho	−1.3462
5	Indiana	1.060	Pennsylvania	−1.1330
6	Alabama	1.0328	New Jersey	−1.7651
7	Texas	1.0711	New York	−1.1789
8	Ohio	1.1268	Vermont	−1.9462
9	-	-	District of Columbia	−1.9877
10	-	-	Oregon	−1.3481

References

1. Hyndman RJ, Athanasopoulos G. *Forecasting: Principles and Practice*. Australia: OTexts; 2018.
2. Akaike H. A new look at the statistical model identification. In: Parzen E, Tanabe K, Kitagawa G, editors. *Selected Papers of Hirotugu Akaike*. Berlin: Springer; 1994. p. 215-232.
   doi: 10.1007/978-1-4612-1694-0_16
3. Burnham KP, Anderson DR. Multimodel inference: Understanding AIC and BIC in model selection. *Soc Method Res*. 2004;33(2):261-304.
   doi: 10.1177/0049124104268644
4. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461-464.
   doi: 10.1214/aos/1176344136
5. 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):12471250.
   doi: 10.5194/gmd-7-1247-2014
6. Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim Res*. 2005;30(1):79-82.
   doi: 10.3354/cr030079
7. Granger CWJ. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*. 1969;37(3):424-438.
   doi: 10.2307/1912791
8. Jolliffe IT, Cadima J. Principal component analysis: A review and recent developments. *Philos Trans Math Phys Eng Sci*. 2016;374(2065):20150202.
   doi: 10.1098/rsta.2015.0202
9. Montgomery DC, Peck EA, Vining GG. *Introduction to Linear Regression Analysis*. 5th ed. United States: John Wiley and Sons; 2012.
10. Jolliffe IT. *Principal Component Analysis*. 2nd ed. Berlin: Springer; 2002.
11. Wold S, Sjöström M, Eriksson L. PLS-regression: A basic tool of chemometrics. *Chemom Intell Lab Systems*. 2001;58(2):109-130.
    doi: 10.1016/S0169-7439(01)00155-1
12. Cliff AD, Ord JK. *Spatial Processes: Models and Applications*. Billerica, MA: Pion; 1981.
13. Anselin L. Local indicators of spatial association (LISA). *Geogr Anal*. 1995;27(2):93-115.
    doi: 10.1111/j.1538-4632.1995.tb00338.x
14. Ord JK, Getis A. Local spatial autocorrelation statistics: Distributional issues and an application. *Geogr Anal*. 1995;27(4):286-306.
    doi: 10.1111/j.1538-4632.1995.tb00912.x