



Comparison between simulated scenarios and Swedish COVID-19 cases throughout the pandemic

Downloaded from: <https://research.chalmers.se>, 2026-05-18 18:50 UTC

Citation for the original published paper (version of record):

Darabi, H., Galanis, I., Benzi, F. et al (2025). Comparison between simulated scenarios and Swedish COVID-19 cases throughout the pandemic. *Scientific Reports*, 15(1).

<http://dx.doi.org/10.1038/s41598-025-08682-z>

N.B. When citing this work, cite the original published paper.



OPEN Comparison between simulated scenarios and Swedish COVID-19 cases throughout the pandemic

Hatef Darabi¹✉, Ilias Galanis¹, Federico Benzi¹, Gerard Farré Puiggali¹, Philip Gerlee², Torbjörn Lundh² & Lisa Brouwers¹

This study assesses the accuracy of COVID-19 scenarios for new infections produced by the Swedish Public Health Agency (PHAS) from December 1, 2020, to March 20, 2023. We introduce a Similarity Error (*SEr*), which evaluates the dissimilarity between simulated and observed case time series using the following attributes: area under the curves, peak timings, and growth/decline rates before and after peaks. Rather than using an arbitrary cut-off, we used a threshold determined through Receiver Operating Characteristic (ROC) analysis, with performance evaluated using the Area Under the Curve (AUC), based on true positives identified by visual inspection for categorization. To further evaluate *SEr*'s effectiveness, we conducted a sensitivity analysis across the full range of possible threshold values within the unit interval. Applying *SEr* with an optimal threshold determined through ROC-analysis 7 rounds out of 11 rounds were classified as having one or more similar scenarios, including the 6 rounds identified by visual inspection. Our findings indicate that, despite the challenges of a rapidly evolving epidemic, PHAS delivered simulations that reflected real-world trends in most of the rounds.

Keywords COVID-19, Scenario analysis, Simulation similarity, Time series comparison

The COVID-19 pandemic profoundly impacted societies across the globe^{1,2}. The rapid and widespread transmission of the virus underscored the urgent need for effective public health resource management and the ability to anticipate and respond to emerging challenges. Understanding the development of the spread became crucial for governments and health agencies to efficiently allocate resources, implement timely interventions, and mitigate the public health impact. Numerous dedicated modelling teams from diverse disciplines were established to provide insights for policymaking^{3–7}. Consequently, the prediction of COVID-19 outcomes has been extensively examined utilizing a range of methodologies, including time series modelling techniques, machine learning algorithms, statistical modelling, and compartmental models, either individually or in combination thereof^{8–28}.

The Swedish Public Health Agency (PHAS) was tasked early in the pandemic to simulate scenarios for how COVID-19 spread might develop in the future, resulting in overall seventeen simulation efforts, amongst these, thirteen focused on new infections covering the period from December 1, 2020, to March 20, 2023. Utilizing epidemiological modelling of disease progression over an extended period for reflecting true case counts is inherently challenging²⁹. Nevertheless, such models are essential for illustrating potential trends, including peaks and lows in case numbers based on key assumptions and uncertainties³⁰. In doing so, they help anticipate and prepare for a range of possible futures and ultimately assisting in proactive decision-making and effective intervention planning for disease control.

Our objective is to conduct a retrospective accuracy evaluation of epidemiological models that have contributed to policy-making in Sweden, by comparing simulated cases to observed cases, across different PHAS scenarios for new infections over time, in order to determine which scenarios are similar and which are not^{3,4,31}. Since PHAS simulations provide only point estimates without confidence intervals, error measures like the Weighted Interval Score (WIS) are not directly applicable^{3,31–33}. To address this limitation, we proceeded by relying on traditional error measures like Dynamic Time Warping (DTW), Euclidean distance, and Mean Absolute Percentage Error (MAPE) for assessing accuracy. However, these may overlook important epidemiological characteristics such as area under the curve (representing total disease burden over time), peak timings and growth/decline rates (representing outbreak dynamics) and thus reduce their effectiveness in

¹The Public Health Agency of Sweden, Solna 171 82, Sweden. ²Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Gothenburg 412 96, Sweden. ✉email: hatef.darabi@folkhalsomyndigheten.se

scenario comparison, therefore we introduce a new error measure based on a set of specific attributes designed to capture the epidemiological similarity between simulated and observed case time series.

Methods

Data

The PHAS assignment led to the creation of seventeen distinct simulation efforts aimed at continuously updating scenarios for the spread of the virus that causes COVID-19. These consist of thirteen primary and two interim simulation rounds exploring various scenarios for new infections, and additional two targeted simulations designed to support decision-making on vaccination strategies. Current analysis utilizes data from the thirteen primary simulation rounds in conjunction with updated daily COVID-19 case data retrieved from the PHAS website (www.folkhalsomyndigheten.se) as of August 18, 2023. In each simulation round scenarios were consistently labelled as Scenario 0, Scenario 1, and Scenario 2 associated with unique underlying assumptions and increased level of severity in disease spread. In text for simplicity, we will index each round (Rn) and scenario (Sn) by respective number (n) e.g., R3-S0 for Round 3 Scenario 0 or simply R3 if we refer to all simulation scenarios of round 3. Except for the first two rounds, all remaining eleven rounds provided daily number of simulated cases, and the total number of scenarios across all considered simulation rounds sum to twenty-seven, as not all rounds included three scenarios. The core simulations were conducted at the national level, while regional projections were derived by dividing the national simulations according to each region's population relative to the total Swedish population. The baseline curve, the Smoothed Daily Case Count (SDCC), is derived from the daily-recorded count of observed cases, specifically presented in its smoothed 7-day rolling average form to eliminate the effect of periodic volatility and outliers³⁴. Figure 1 displays the unprocessed daily national case count curve along with its smoothed 7-day rolling average (SDCC) in which observed sudden peak during week 39, 2022 is due to late reporting and retroactive registration for some regions. This figure highlights the intervals for the various simulation rounds, their publication dates, and periods dominated by different variants. Notably, not all simulations rounds covered the same number of scenarios, and some even overlapped in their timeframes. Certain rounds provided scenario estimates a longer period before their publication dates, these are confined to their respective training periods, in order to highlight if a recent peak before publication was modelled adequately by the specific scenarios or not. Figure 2 displays each rounds scenario estimates in addition to observed SDCC.

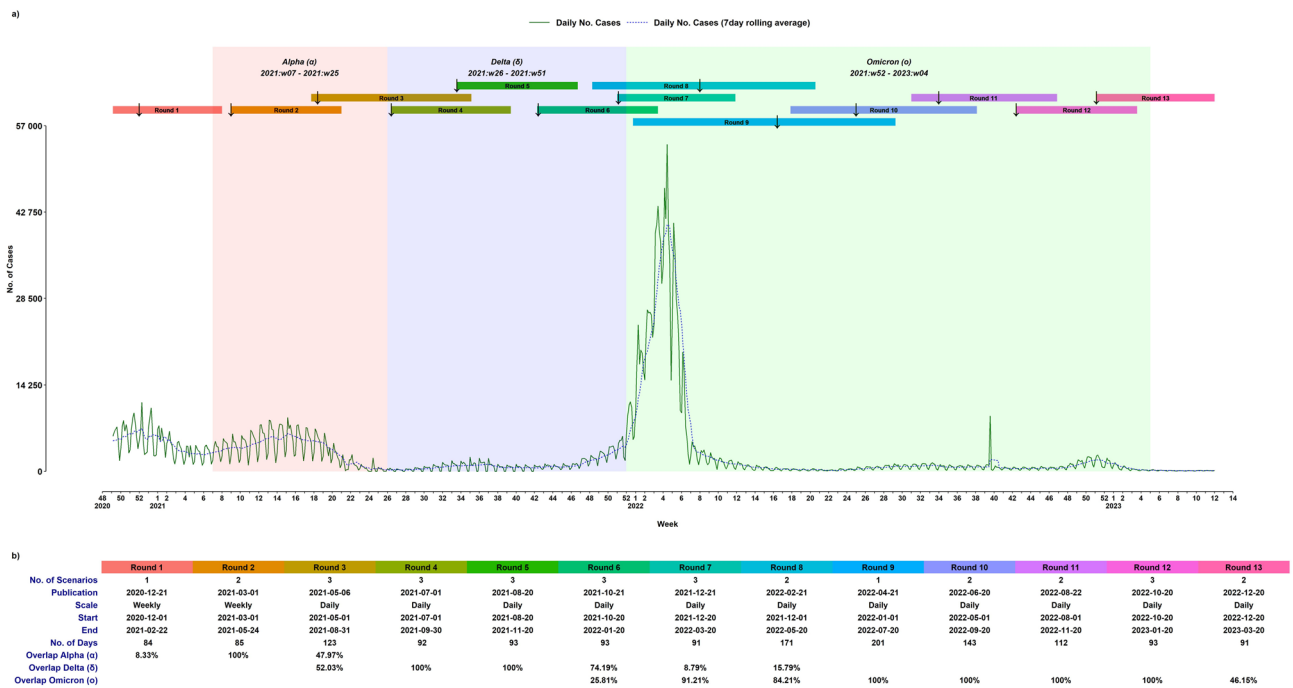


Fig. 1. PHAS released thirteen rounds of simulations of new infections, from December 1, 2020, to March 20, 2023, with each round featuring one to three scenarios. (a) Each round is represented by a unique colour. The figure displays the raw daily national case counts (in green) and the smoothed 7-day rolling averages (SDCC in blue). Intervals for simulation rounds are highlighted as bars, publication dates are marked with downward arrows, and shaded areas indicate periods dominated by different variants. (b) The table provides details for each round, including the number of scenarios, publication date, reporting scale, period covered, duration in days, and the extent of overlap between the simulation period and the dominance of various variants.

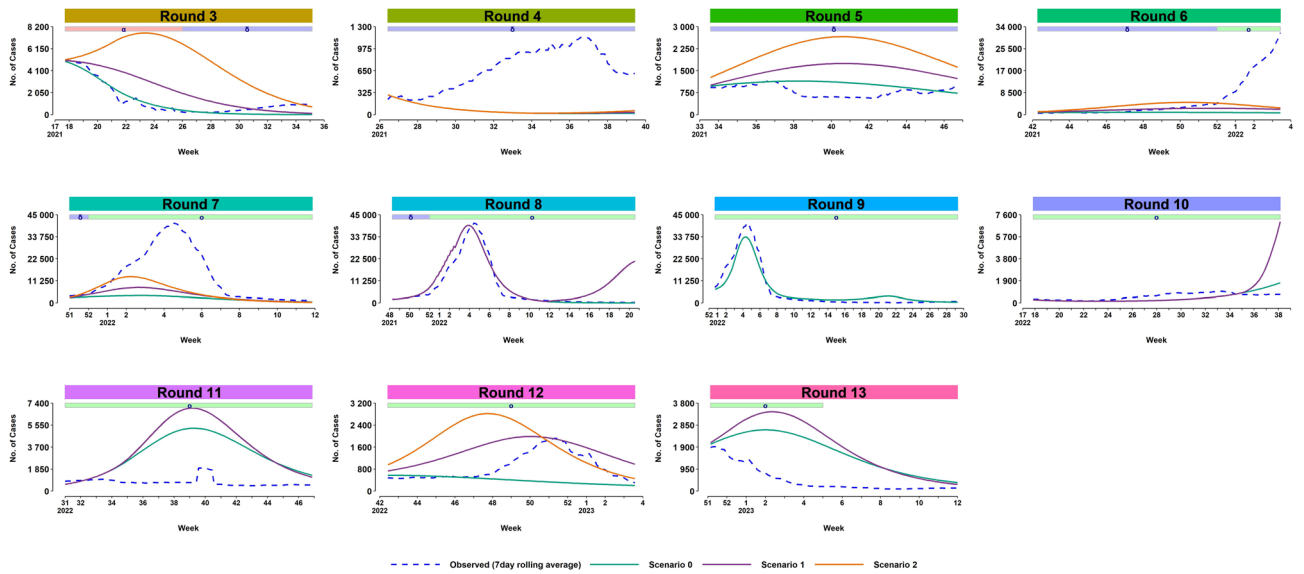


Fig. 2. The smoothed 7-day rolling averages daily national (Riket) case count (dashed blue line) across different simulation rounds, along with the simulated case numbers for each specific scenario.

| Attribute | Short Description | Long Description |
|-----------|-------------------|---|
| A1 | Peak timing Prior | Comparing if scenario curves peak timing is occurring at the same time or prior to the SDCC curves peak timing |
| A2 | Peak timing Post | Comparing if scenario curves peak timing is occurring post the SDCC curves peak timing |
| A3 | AUC | Comparing scenario curves AUC with the SDCC curves AUC |
| A4 | Growth rate Prior | Comparing prior growth rate of the scenario curve (from its start to its peak) with SDCC curves prior growth rate |
| A5 | Decline rate Post | Comparing post decline rate of the scenario curve (from its peak to its end value) with SDCC curves post decline rate |

Table 1. Considered attributes.

Methods Dissimilarity measure

To effectively compare disease scenario projections of new infections with the SDCC curve of observed cases from an epidemiological perspective under the assumption of a single peak during the projection period, we assess several key attributes (Table 1), including A1-A2: the timing of the peak, A3: the Area Under the Curve (AUC), and A4-A5: the growth/decline rates of the scenario curves in relation to the SDCC curve. Evaluating the AUC, representing the total disease burden over time, provides critical insights into the scenarios overall performance and accuracy. Another crucial factor is the peak timing as it measures how accurately the scenario estimates the timing of disease surges. Furthermore, examining the growth/decline rate of the curve before and after the peak reveals additionally whether the scenario approximately captures the dynamics of disease spread, including phases of acceleration and deceleration.

To assess the accuracy of these attributes, we quantify the error for each by calculating its respective Absolute Percentage Error (APE), defined as

$$APE(X, Y) = \left| \frac{X - Y}{X} \right| \tag{1}$$

where X is the reference (true or actual) value and Y is the comparator value. We modify the APE formulation to ensure that errors are well-defined and stable, even when the true denominator value is zero. The modified APE formulation denoted as $APE_M(X, Y)$, where the subscript M indicates the modification, is defined as

$$APE_M(X, Y) = \left| \frac{(X^{I(X \neq 0)} - I(X = 0)) - (Y^{I(Y \neq 0)} - I(Y = 0))}{X^{I(X \neq 0)}} \right| \tag{2}$$

Where $I(\bullet)$ denotes the indicator function. This transformation ensures that, when $X \neq 0$ and $Y \neq 0$ then $APE_M(X, Y)$ reduces to the standard APE. Adhering to the combinatorial convention that zero to the power of zero is one, if $X = 0$ and $Y \neq 0$, the entire magnitude of Y is set as an error which subsequently is capped at one (see below), and when $X \neq 0$ and $Y = 0$, the error is set to one. Finally, if $X = Y$, including when $X = 0$ and $Y = 0$, the $APE_M(X, Y)$ is zero, as there is no difference between the X and Y values.

Among the attributes, A1 and A2 include an additional parameter, a window size $\Delta^{prior} = \Delta^{post} = \Delta = 14$ days (chosen arbitrarily) centred around the peak location of the SDCC curve. For A1 and A2, we set the APE to zero if the peak of the scenario curve falls within the specified window, otherwise, the error is scaled relative to Δ as follows

$$\begin{aligned} APE_{A1} &= APE_M(\Delta^{prior}, t_{SDCC}^p - t_{scenario}^p)(1 - \mathbf{I}(t_{scenario}^p \in [t_{SDCC}^p - \Delta^{prior}, t_{SDCC}^p])), \\ APE_{A2} &= APE_M(\Delta^{post}, t_{scenario}^p - t_{SDCC}^p)(1 - \mathbf{I}(t_{scenario}^p \in]t_{SDCC}^p, t_{SDCC}^p + \Delta^{post}]), \end{aligned} \tag{3}$$

where t_{SDCC}^p is the peak timing of the SDCC curve, and $t_{scenario}^p$ is the peak timing of the scenario curve. For A3, we apply the trapezoidal rule to approximate the area under each curve (AUC), for which then APE is

$$APE_{A3} = APE_M(AUC_{SDCC}, AUC_{scenario}). \tag{4}$$

To calculate APE for A4 and A5, the respective linear growth/decline rates before and after the peaks are evaluated, if defined, to construct the APEs as

$$\begin{aligned} K_{SDCC}^1 &= \frac{Y_{SDCC}^p - Y_{SDCC}^s}{t_{SDCC}^p - t^s}, K_{SDCC}^2 = \frac{Y_{SDCC}^e - Y_{SDCC}^p}{t^e - t_{SDCC}^p}, \\ K_{scenario}^1 &= \frac{Y_{scenario}^p - Y_{scenario}^s}{t_{scenario}^p - t^s}, K_{scenario}^2 = \frac{Y_{scenario}^e - Y_{scenario}^p}{t^e - t_{scenario}^p}, \\ APE_{A4} &= APE_M(K_{SDCC}^1, K_{scenario}^1), \\ APE_{A5} &= APE_M(K_{SDCC}^2, K_{scenario}^2), \end{aligned} \tag{5}$$

where $\{t^s, Y_{SDCC}^s\}, \{t_{SDCC}^p, Y_{SDCC}^p\}, \{t^e, Y_{SDCC}^e\}$ represent the start, peak, and end time points within the comparison window for the SDCC curve, along with their corresponding values. Similarly, $\{t^s, Y_{scenario}^s\}, \{t_{scenario}^p, Y_{scenario}^p\}, \{t^e, Y_{scenario}^e\}$ denote the same points for the scenario curve.

Subsequently we cap each $APE_{A_j}, j \in \{1, 2, 3, 4, 5\}$ at one to prevent extreme values. These are then combined using specific weights w_i to form the asymmetric Similarity Error (*SEr*)

$$SEr = \frac{\sum_{i=1}^5 w_i \mathbf{I}(APE_{A_i} \text{ is defined}) \min(1, APE_{A_i})}{\sum_{i=1}^5 w_i \mathbf{I}(APE_{A_i} \text{ is defined})}. \tag{6}$$

that ranges between 0 and 1. *SEr* is never based on all the defined attributes, as A4 and A5 may not always be well-defined due to potential alignment issues (e.g., if peak timing coincides with either starting or end timing for each curve), and for each comparison either A1 (when $t_{scenario}^p \leq t_{SDCC}^p$) or A2 (when $t_{scenario}^p > t_{SDCC}^p$) is present not both.

Other measures

We also assess *SEr* performance alongside other traditional key measures such as the Dynamic Time Warping algorithm (DTW), Euclidean distance and Mean Absolute Percentage Error (MAPE). The Euclidean distance calculates the square root of the sum of the squared distance between actual (X) and simulated values (Y) by

$$ED(X, Y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \tag{7}$$

for vectors $X = \{x_1, \dots, x_n\}$ and $Y = \{y_1, \dots, y_n\}$ while the asymmetric MAPE measures the average absolute percentage difference between actual and simulated values relative to the actual values by

$$MAPE(X, Y) = \frac{1}{n} \sum_{i=1}^n \left| \frac{x_i - y_i}{x_i} \right|. \tag{8}$$

DTW, more flexible, extends these approaches by finding an optimal alignment between series X and Y, allowing for nonlinear comparisons by warping the time axis^{35,36}. The minimal DTW distance and corresponding optimal warping path are computed using a dynamic programming algorithm. DTW computes a distance measure between X and Y by constructing a cost matrix $C \in \mathbb{R}^{n \times n}$, where each element $C(i, j)$ represents the local cost of aligning elements x_i and y_j usually defined as $C(i, j) = (x_i - y_j)^2$ ^{35,36}. Using C an accumulated cost matrix $D \in \mathbb{R}^{n \times n}$ is evaluated by

$$D(i, j) = C(i, j) + \min \begin{cases} D(i-1, j-1) \\ D(i-1, j) \\ D(i, j-1) \end{cases} \tag{9}$$

with boundary conditions imposed on the first row and column. Once the accumulated cost matrix is computed, the overall DTW distance is given by $DTW(X, Y) = D(n, n)$. The optimal warping path is obtained by backtracking from $D(n, n)$ to $D(1, 1)$ along the minimum-cost neighbours^{35,36}.

Optimal configuration and threshold

To select between a 3-attribute (A1–A3) and full 5-attribute (A1–A5) versions of *SEr*, we conduct a Receiver Operating Characteristic (ROC) analysis using three different weighting schemes where performance is evaluated using the AUC. Subsequently we choose the configuration that achieves the highest AUC^{ROC} and identify an optimal classification threshold, $\epsilon \in [0,1]$, which distinguishes “Similar” ($SEr \leq \epsilon$) from “Not Similar” ($SEr > \epsilon$) sequences (Supplementary ROC analysis).

All data processing, calculations, and plotting in this paper are done using the statistical software R (version 4.4.3)³⁷.

Results

Similarity assessment on National level

To avoid relying on arbitrary thresholds we visually assessed which scenario curves we deem to be similar (R3-S0, R5-S0, R8-S0, R9-S0, R10-S0, R12-S1) to the SDCC on a national level and followed up with ROC-analysis and different weighting schemes to select between a 3-attribute (A1–A3) and full 5-attribute (A1–A5) versions of *SEr* by choosing the one that achieved highest AUC. The ROC analysis pointed towards the full 5-attribute *SEr* ($AUC^{ROC} = 97\%$) with following weights $w_1 = 1, w_2 = 2, w_3 = f(AUC_{SDCC}, AUC_{Scenario}), w_4 = 2$ and $w_5 = 1$ estimating an optimal threshold of $\epsilon_{SEr} = 0.54$ for classification (Supplementary ROC analysis). The function f considers the scale of AUC_{SDCC} and the over or underestimation of $AUC_{Scenario}$ by measuring the absolute difference to AUC_{SDCC} and is defined as

$$f(AUC_{SDCC}, AUC_{Scenario}) = \max(1, \lfloor \log_{10}(|AUC_{SDCC}|) \rfloor) * \max(1, \lfloor \log_{10}(|AUC_{Scenario} - AUC_{SDCC}|) \rfloor) \tag{10}$$

with $\lfloor x \rfloor = floor(x)$. Using the full 5-attribute *SEr* and the estimated optimal threshold for classification, curves R6-S2, R8-S1, and R10-S1 were classified as similar, contradicting original classification, alongside the 6 curves that were visually identified as similar. Resulting in 9 out of 27 scenario comparisons (33%) were classified as similar, represented by 7 out of 11 simulation rounds (64%) with at least one scenario classified as similar (Fig. 3). In R4, R7, R11 and R13 no *SEr*-similar scenario was observed.

Repeating the ROC analysis with more traditional comparison error measures resulted in AUC^{ROC} values of 82% (DTW, $\epsilon_{DTW} = 62069.31$), 79% (Euclidean, $\epsilon_{Euclidean} = 6325.49$), and 77% (MAPE, $\epsilon_{MAPE} = 0.52$) all of which are lower than *SEr*'s AUC^{ROC} (Supplementary ROC analysis). DTW classified 11 scenarios across 6 rounds (54%) as similar: R3-S0*, R3-S1, R5-S0*, R8-S0*, R10-S0*, R10-S1*, R12-S0, R12-S1, R12-S2, R13-S0, and R13-S1 (Fig. 3). Of these five were also *SEr*-similar (*-marked). The Euclidean distance classified 4 scenarios across 4 (36%) rounds: R3-S0*, R5-S0*, R10-S0*, and R12-S1*, all of which were *SEr*-similar as well. The MAPE classified 7 scenarios across 7 (64%) rounds: R3-S0*, R5-S0*, R6-S0, R7-S2, R8-S0*, R10-S0*, and R12-S0, of which four were also *SEr* similar. Jointly R3-S0, R5-S0, and R10-S0 were



Fig. 3. Similarity classification of the national (Riket) simulation rounds. **(a)** The four error measures classified based on thresholds determined through ROC analysis. The specific thresholds are as follows: $\epsilon_{SEr} = 0.54$, $\epsilon_{DTW} = 62069.31$, $\epsilon_{Euclidean} = 6325.49$, $\epsilon_{MAPE} = 0.52$. **(b)** Upset figure displaying how many distinct rounds each error measures or combination of error measures exhibited similarity simultaneously.

classified as similar by all considered error measures. In R3, R5, R10, R12 either error measures identified at least one scenario to be classified as similar (Fig. 3).

Examining the *SEr* decomposition (Table 2) reveals that scenarios classified as similar displayed consistent patterns of low feature-level errors that contributed to their reduced *SEr* values. Specifically, the timing before peak (A1) frequently fell within the admissible window, resulting in an $APE_{A1} = 0$, while other features also showed relatively low APEs. This alignment led to a lower weighted error sum in the *SEr* numerator, bringing the overall value below the similarity threshold. In contrast, non-similar scenarios were characterized by high APEs across multiple features, often reaching 1, resulting in a substantially larger numerator and pushing the *SEr* well above the threshold.

Complementing with a sensitivity analysis for *SEr* based on a range of epsilon values from 0 to 1 in incremental steps of 0.05 manifest intriguing patterns (Fig. 4). The lower the threshold epsilon, the more scenarios are classified as non-similar by the *SEr* metric. At low epsilon values ($0 < \epsilon \leq 0.1$), only a negligible fraction of rounds (9%) are identified as *SEr*-Similar. For moderate epsilon values ($0.15 \leq \epsilon \leq 0.50$), the proportion of *SEr*-Similar rounds rise steadily, reaching 55% when ϵ goes towards the upper range. For higher epsilon values ($0.55 \leq \epsilon \leq 0.85$), *SEr*-Similar rounds reach 73%. At very high epsilon values ($\epsilon \geq 0.90$), *SEr*-Similar rounds overwhelmingly take precedence. For the scenarios the increase in the number of *SEr*-similar is slower and at $\epsilon = 0.55$ about a third of the scenarios are classified as similar. This trend is broken at around $\epsilon = 0.8$ after which the number of *SEr*-similar scenarios increase faster. The fact that the number of *SEr*-similar rounds follow a constant linear trend (dashed line in Fig. 4) suggests that the minimal *SEr* within rounds are evenly distributed in the unit interval, whereas lower rate of increase among the scenarios (the curve falls below the dashed line) implies that *SEr*-values for the scenarios are unevenly distributed and biased towards higher *SEr*-values.

Similarity assessment on regional level

Assessments of the various regions are also presented in Fig. 5 under same configuration of *SEr* as in the national comparison. Results show variability in the similarity percentages across different regions and simulation rounds

| <i>SEr</i> | | | | | | | | | | | | | |
|----------------------------------|------------|------------|------------|------------|------------|-------|-------|-------|-------|-------|----|---------|------------|
| Scenario | APE_{A1} | APE_{A2} | APE_{A3} | APE_{A4} | APE_{A5} | w_1 | w_2 | w_3 | w_4 | w_5 | D | N | <i>SEr</i> |
| <i>SEr</i> similar scenarios | | | | | | | | | | | | | |
| R3-S0 | 0 | | 0.132 | | 0.2261 | 1 | | 20 | | 1 | 22 | 2.8661 | 0.13 |
| R5-S0 | | 0 | 0.2826 | 0.4659 | 1 | | 2 | 16 | 2 | 1 | 21 | 6.4534 | 0.31 |
| R6-S2 | 1 | | 0.3696 | 0.8123 | | 1 | | 25 | 2 | | 28 | 11.8646 | 0.42 |
| R8-S0 | 0 | | 0.1238 | 0.0491 | 0.0542 | 1 | | 30 | 2 | 1 | 34 | 3.8664 | 0.11 |
| R8-S1 | 0 | | 0.4596 | 0.0491 | 0.5621 | 1 | | 30 | 2 | 1 | 34 | 14.4483 | 0.42 |
| R9-S0 | 0 | | 0.0125 | 0.1509 | 0.1684 | 1 | | 24 | 2 | 1 | 28 | 0.7702 | 0.03 |
| R10-S0 | | 1 | 0.2354 | 0.6914 | | | 2 | 16 | 2 | | 20 | 7.1492 | 0.36 |
| R10-S1 | | 1 | 0.2 | 1 | | | 2 | 16 | 2 | | 20 | 7.2 | 0.36 |
| R12-S1 | 0 | | 0.6459 | 0.0155 | 0.5237 | 1 | | 16 | 2 | 1 | 20 | 10.8891 | 0.54 |
| <i>SEr</i> non-similar scenarios | | | | | | | | | | | | | |
| R3-S1 | 0 | | 0.6774 | | 0.2206 | 1 | | 25 | | 1 | 27 | 17.1556 | 0.64 |
| R3-S2 | | 1 | 1 | | 1 | | 2 | 25 | | 1 | 28 | 28 | 1 |
| R4-S0 | 1 | | 0.9012 | | 0.8914 | 1 | | 16 | | 1 | 18 | 16.3106 | 0.91 |
| R4-S1 | 1 | | 0.8975 | | 0.897 | 1 | | 16 | | 1 | 18 | 16.257 | 0.9 |
| R4-S2 | 1 | | 0.8921 | | 0.9068 | 1 | | 16 | | 1 | 18 | 16.1804 | 0.9 |
| R5-S1 | | 1 | 0.8956 | 0.5132 | 1 | | 2 | 16 | 2 | 1 | 21 | 18.356 | 0.87 |
| R5-S2 | | 0.9286 | 1 | 1 | 1 | | 2 | 20 | 2 | 1 | 25 | 24.8572 | 0.99 |
| R6-S0 | 1 | | 0.8447 | 0.9918 | | 1 | | 25 | 2 | | 28 | 24.1011 | 0.86 |
| R6-S1 | 0.9286 | | 0.6337 | 0.9318 | | 1 | | 25 | 2 | | 28 | 18.6347 | 0.67 |
| R7-S0 | 0 | | 0.8236 | 0.9621 | 0.9328 | 1 | | 36 | 2 | 1 | 40 | 32.5066 | 0.81 |
| R7-S1 | 0 | | 0.71 | 0.8003 | 0.8466 | 1 | | 30 | 2 | 1 | 34 | 23.7472 | 0.7 |
| R7-S2 | 0.0714 | | 0.5998 | 0.5333 | 0.7426 | 1 | | 30 | 2 | 1 | 34 | 19.8746 | 0.58 |
| R11-S0 | 0 | | 1 | 1 | 1 | 1 | | 20 | 2 | 1 | 24 | 23 | 0.96 |
| R11-S1 | 0 | | 1 | 1 | 1 | 1 | | 20 | 2 | 1 | 24 | 23 | 0.96 |
| R12-S0 | 1 | | 0.5493 | | 0.9253 | 1 | | 16 | | 1 | 18 | 10.7141 | 0.6 |
| R12-S2 | 0.7857 | | 0.9635 | 1 | 0.2052 | 1 | | 16 | 2 | 1 | 20 | 18.4069 | 0.92 |
| R13-S0 | | 0.3571 | 1 | 0.2155 | 0.616 | | 2 | 20 | 2 | 1 | 25 | 21.7612 | 0.87 |
| R13-S1 | | 0.5714 | 1 | 0.4815 | 1 | | 2 | 20 | 2 | 1 | 25 | 23.1058 | 0.92 |

Table 2. Decomposition table of *SEr* for the National comparison. D = Denominator, N = Numerator, *SEr* = N/D rounded to two decimal points

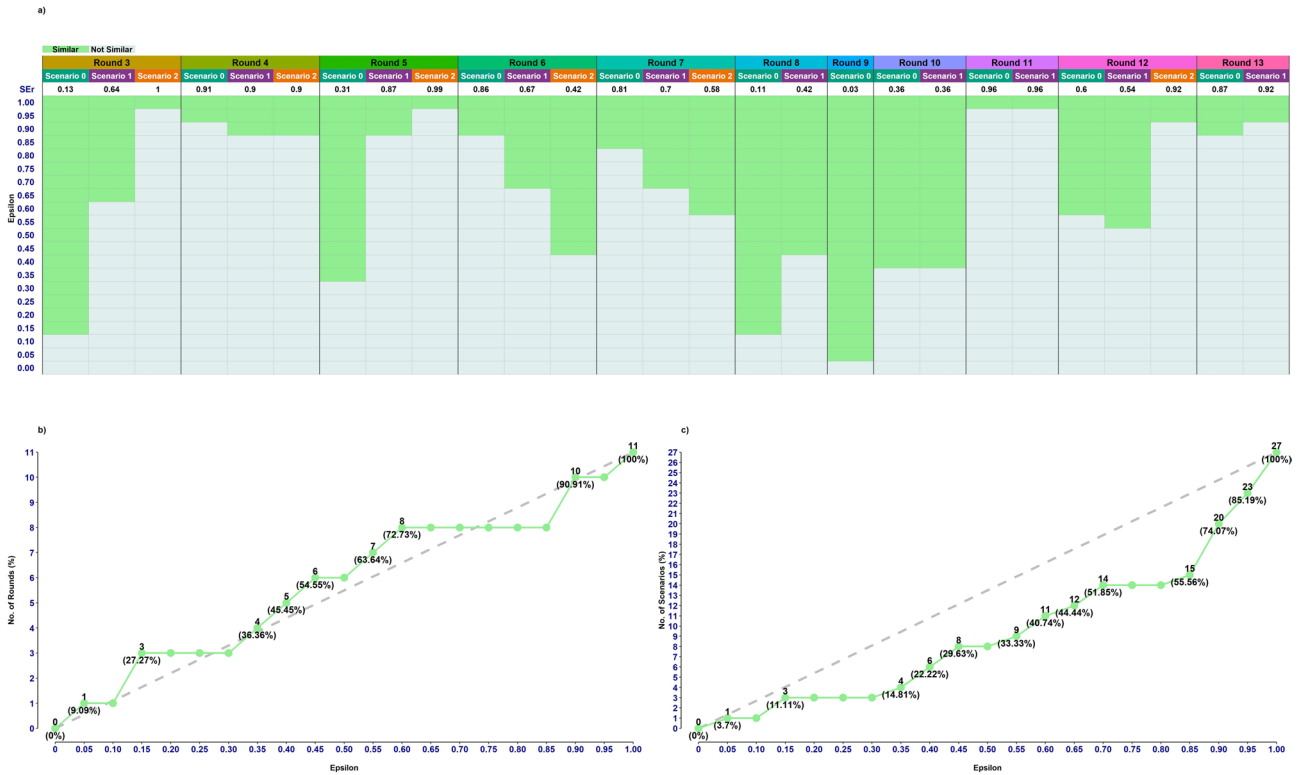


Fig. 4. Classification performance of *SEr*. **a)** Over a range of ϵ values where each of the national (Riket) scenarios per round are classified and colour coded accordingly. **b)** Number of rounds and **c)** Number of scenarios classified as *SEr* Similar over the range of ϵ values.

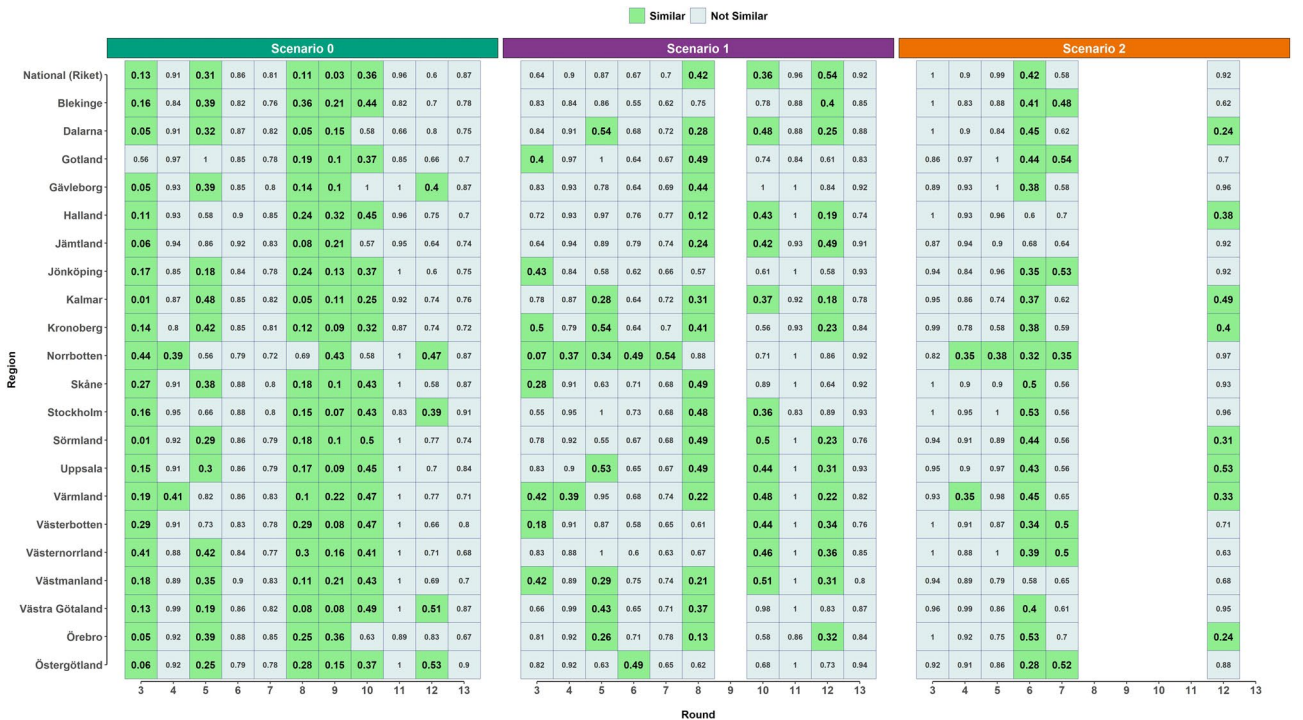


Fig. 5. The *SEr* categorization of all regions and scenarios for all simulation rounds and considered regions. Based on the configuration of *SEr* determined by the national ROC analysis (Riket, representing the national level). Respective *SEr*-value is displayed in each tile and subsequently colour coded according to classification.

(Supplementary Fig. 3). Regions *Halland* and *Jämtland* had the lowest number of *SEr* similarity rounds, with 5 (46%) rounds, whereas regions *Blekinge*, *Västernorrland*, and *Östergötland* had the highest, with 8 (73%) rounds. The regions *Skåne*, *Stockholm*, and *Västra Götaland* are the most populous regions in Sweden. For *Skåne*, 8 scenarios across 6 (55%) rounds were *SEr* similar (R3-S0, R3-S1, R5-S0, R6-S2, R8-S0, R8-S1, R9-S0, R10-S0). Similarly, for *Stockholm*, 8 scenarios across 6 (55%) rounds were *SEr* similar (R3-S0, R6-S2, R8-S0, R8-S1, R9-S0, R10-S0, R10-S1, R12-S0). For *Västra Götaland*, 9 scenarios across 7 (64%) rounds were *SEr* similar (R3-S0, R5-S0, R5-S1, R6-S2, R8-S0, R8-S1, R9-S0, R10-S0, R12-S0).

Discussion

This study evaluates the accuracy of various COVID-19 simulation scenarios developed by the Public Health Agency of Sweden (PHAS) at both national and regional levels utilizing both traditional error measures as well as introducing a new asymmetric measure. We applied traditional error measures, including DTW, Euclidean distance, and MAPE, though each has limitations in capturing essential aspects of epidemic spread and are better suited for evaluating mathematical similarity or model fit. For instance, MAPE and Euclidean distance do not account for the cumulative scale of disease spread and primarily focus on pointwise differences and may miss structural differences, such as shifted peaks or diverging slopes, while DTW may fall short in representing differences in rates of increase or decrease and obscure meaningful epidemiological discrepancies. Moreover, traditional error measures often lack interpretability for epidemiological evaluation, reducing their effectiveness in comparing scenarios^{35,36}.

Attempting to address these gaps, we developed a new, interpretable error measure *SEr* designed to capture relevant epidemiological attributes, such as AUC, peak timing, and pre- and post-peak growth/decline rates. Using *SEr* with threshold $\epsilon_{SEr} = 0.54$, 7 (64%) out of 11 national comparison rounds had at least one scenario classified as similar to the SDCC, including five S0 (R3, R5, R8, R9, R10), three S1 (R8, R10, R12), and one S2 (R6). This includes all 6 scenarios that were visually identified as similar five S0 (R3, R5, R8, R9, R10), and one S1 (R12). *SEr* decomposition (Table 2) shows that similar scenarios had low APEs, resulting in *SEr* values below the similarity threshold. Notably, R3-S0, R8, R9-S0, and R12-S1 exhibited synchronized peaks relative to the SDCC, while R5-S0, though peaking after the SDCC, remained within the admissible range. In contrast, non-similar scenarios had high APEs, leading to elevated *SEr* values, even in cases where peak timing appeared synchronized with the SDCC.

A key strength of *SEr* lies in its multi-dimensional assessment, integrating diverse trajectory characteristics such as timing, magnitude, and shape condensed into a single, interpretable score. This comprehensive approach enables *SEr* to detect discrepancies that traditional single-metric error measures—such as MAPE, Euclidean distance, or DTW—may overlook. Another important advantage of *SEr* is its dynamic weight assignment, which allows modelers to emphasize specific attributes based on the objectives of the analysis. This flexibility makes *SEr* highly adaptable: by adjusting the weights w_i , one can prioritize the accuracy of more critical features. For example, when optimizing a model for early pandemic detection, the *SEr* framework allows for assigning greater weight to attribute A4, a distinct advantage over traditional metrics. Despite its strengths and valuable insights, *SEr* has limitations. The sensitivity analysis demonstrates strong performance, particularly at extreme epsilon values, where the model aligns closely with expectations. While *SEr* showcases robust classification abilities, fine-tuning is needed to better handle intermediate epsilon values. Its weighting flexibility allows for customization but introduces subjectivity. Additionally, capping extreme APE values improves stability but may mask some modelling issues. Future refinements could include adjusting attribute weights, different window sizes (Δ^{prior} and Δ^{post}), and exploring different thresholds (ϵ) determination methods without the need of initial visual assessment. Furthermore, the linear approximations of the growth and decline phases implemented in *SEr* are practical for ensuring comparability and provide a crude but interpretable estimate of factual trends. However, they may underrepresent the non-linear dynamics that often characterize critical phases of an outbreak, particularly during rapid acceleration or deceleration phases. Additionally, although AUC reflects the overall epidemic burden, it may obscure important variations in peak intensity—two scenarios might have comparable total cases yet differ substantially in their peak healthcare demands. Explicitly including peak magnitude would increase *SEr*'s sensitivity to these critical differences, thereby improving its value for public health decision-making and preparedness planning. Addressing multiple peaks and developing a symmetric version of *SEr* would also enhance its robustness.

In conclusion, PHAS has provided long-term simulation rounds that aligned with observed trends according to the proposed *SEr*, facilitating proactive decision-making during the pandemic. However, effective disease modelling remains challenging^{29,33} due to evolving non-pharmaceutical interventions (NPIs), new virus variants, and reporting adjustments to name a few. As an example, focusing on R3 and R4 in Figs. 1 and 2, we note that R4 was published shortly after the period dominated by the Alpha variant. While the preparation and evaluation of R4 was carefully monitored, the rapid emergence and dominance of the Delta variant was impossible to predict given the knowledge then available. As a result, the assumptions that underpinned R4 did not fully align with the evolving reality, requiring a prompt reassessment, in particular peak timing and overall case burden was off for the scenarios of the round as evident visually and by the *SEr* decomposition (Table 2). To swiftly respond to the changing epidemic trends, assumptions were readapted and R5 was released within just one month, in contrast to the two-month intervals between other simulation rounds. Scenarios generated during phases of relatively stable transmission patterns, without significant epidemiological shifts, were more likely to align with observed data. Consequently, effectively addressing the complexities of epidemic modelling necessitates continuous reassessment of foundational assumptions in each simulation round to capture the dynamic nature of the disease^{3,31}. Continuous improvement of error measures, like our proposed *SEr*, which emphasizes epidemiological assessment by incorporating key features like timing, magnitude, and shape based on dynamic weight assignment can significantly enhance the evaluation of both retrospective and prospective

models. By identifying trajectory components that are challenging to predict accurately via its decomposition, *SER* offers actionable insights. These insights can guide future modeling efforts by encouraging the inclusion of uncertainty bounds around critical features and the exploration of alternative weighting schemes to prioritize specific attributes according to the analysis objectives. This may aid the development of more diverse scenario assumptions, enabling a better representation of the full spectrum of plausible epidemic trajectories.

Data availability

The datasets analysed and source codes used to produce the findings of this study are available from the corresponding author upon reasonable request.

Received: 10 January 2025; Accepted: 23 June 2025

Published online: 02 July 2025

References

1. Sawicka, B. et al. Elsevier, in *Coronavirus Drug Discovery* Vol. 1 (ed Chukwuebuka Egbuna) 267–311 (2022).
2. McKibbin, W. & Fernando, R. The global economic impacts of the COVID-19 pandemic. *Econ. Model.* **129**, 106551. <https://doi.org/10.1016/j.econmod.2023.106551> (2023). <https://doi.org/https://doi.org/>
3. Howerton, E. et al. Evaluation of the US COVID-19 scenario modeling hub for informing pandemic response under uncertainty. *Nat. Commun.* **14**, 7260. <https://doi.org/10.1038/s41467-023-42680-x> (2023).
4. Starck, T. & Langevin, M. Retrospective analysis of Covid-19 hospitalization modelling scenarios which guided policy response in France. *MedRxiv* **2023.2012.2016.23300086** <https://doi.org/10.1101/2023.12.16.23300086> (2023).
5. Loo, S. L. et al. The US COVID-19 and influenza scenario modeling hubs: delivering long-term projections to guide policy. *Epidemics* **46**, 100738. <https://doi.org/10.1016/j.epidem.2023.100738> (2024).
6. Bicher, M. et al. Supporting COVID-19 policy-making with a predictive epidemiological multi-model warning system. *Commun. Med. (Lond)*. **2**, 157. <https://doi.org/10.1038/s43856-022-00219-z> (2022).
7. Crawford, M. M. & Wright, G. The value of mass-produced COVID-19 scenarios: A quality evaluation of development processes and scenario content. *Technol. Forecast. Soc. Change*. **183**, 121937. <https://doi.org/10.1016/j.techfore.2022.121937> (2022).
8. Capistran, M. A., Capella, A. & Christen, J. A. Forecasting hospital demand in metropolitan areas during the current COVID-19 pandemic and estimates of lockdown-induced 2nd waves. *PLoS One*. **16**, e0245669. <https://doi.org/10.1371/journal.pone.0245669> (2021).
9. Farcomeni, A., Maruotti, A., Divino, F., Jona-Lasinio, G. & Lovison, G. An ensemble approach to short-term forecast of COVID-19 intensive care occupancy in Italian regions. *Biom J.* <https://doi.org/10.1002/bimj.202000189> (2020).
10. Gecili, E., Ziady, A., Szczesniak, R. D. & Forecasting COVID-19 confirmed cases, deaths and recoveries: revisiting established time series modeling through novel applications for the USA and Italy. *PLoS One*. **16**, e0244173. <https://doi.org/10.1371/journal.pone.0244173> (2021).
11. Goic, M., Bozanic-Leal, M. S., Badal, M. & Basso, L. J. COVID-19: Short-term forecast of ICU beds in times of crisis. *PLoS One*. **16**, e0245272. <https://doi.org/10.1371/journal.pone.0245272> (2021).
12. Harun Yonar, A. Y. Mustafa Agah Tekindal, Melike Tekindal Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods. *EJMO* **4**, 160–165 (2020). <https://doi.org/10.14744/ejmo.2020.28273>
13. Liu, M., Thomadsen, R. & Yao, S. Forecasting the spread of COVID-19 under different reopening strategies. *Sci. Rep.* **10**, 20367. <https://doi.org/10.1038/s41598-020-77292-8> (2020).
14. Maleki, M., Mahmoudi, M. R., Wraith, D. & Pho, K. H. Time series modelling to forecast the confirmed and recovered cases of COVID-19. *Travel Med. Infect. Dis.* **101742** <https://doi.org/10.1016/j.tmaid.2020.101742> (2020).
15. Nikolopoulos, K., Punia, S., Schafers, A., Tsinopoulos, C. & Vasilakis, C. Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions. *Eur. J. Oper. Res.* <https://doi.org/10.1016/j.ejor.2020.08.001> (2020).
16. Papastefanopoulos, V. L. & Kotsiantis, P. S. COVID-19: A comparison of time series methods to forecast percentage of active cases per population. *Applied Sciences* **10** (2020).
17. Ribeiro, M., da Silva, R. G., Mariani, V. C. & Coelho, L. D. Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil. *Chaos Solitons Fractals*. **135**, 109853. <https://doi.org/10.1016/j.chaos.2020.109853> (2020).
18. Rivera-Rodriguez, C. & Urdinola, B. P. Predicting hospital demand during the COVID-19 outbreak in Bogotá, Colombia. *Front. Public Health*. **8** <https://doi.org/10.3389/fpubh.2020.582706> (2020).
19. Sahin, U. & Sahin, T. Forecasting the cumulative number of confirmed cases of COVID-19 in Italy, UK and USA using fractional nonlinear grey Bernoulli model. *Chaos Solitons Fractals*. **138**, 109948. <https://doi.org/10.1016/j.chaos.2020.109948> (2020).
20. Salgotra, R., Gandomi, M. & Gandomi, A. H. Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming. *Chaos Solitons Fractals*. **138**, 109945. <https://doi.org/10.1016/j.chaos.2020.109945> (2020).
21. Schweigler, L. M. et al. Forecasting models of emergency department crowding. *Acad. Emerg. Med.* **16**, 301–308. <https://doi.org/10.1111/j.1553-2712.2009.00356.x> (2009).
22. Shinde, G. R. et al. Forecasting models for coronavirus disease (COVID-19): A survey of the State-of-the-Art. *SN Comput. Sci.* **1**, 197. <https://doi.org/10.1007/s42979-020-00209-9> (2020).
23. Singh, R. K. et al. Prediction of the COVID-19 pandemic for the top 15 affected countries: advanced autoregressive integrated moving average (ARIMA) model. *JMIR Public Health Surveill.* **6**, e19115. <https://doi.org/10.2196/19115> (2020).
24. Sujath, R., Chatterjee, J. M. & Hassanien, A. E. A machine learning forecasting model for COVID-19 pandemic in India. *Stoch. Environ. Res. Risk Assess.* 1–14. <https://doi.org/10.1007/s00477-020-01827-8> (2020).
25. Tang, Y. & Wang, S. Mathematic modeling of COVID-19 in the united States. *Emerg. Microbes Infect.* **9**, 827–829. <https://doi.org/10.1080/22221751.2020.1760146> (2020).
26. Yonar, H., Tekindal, Y. A. & Tekindal, M. M. Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods. *EJMO* **4**, 160–165 (2020).
27. Team, I. C. F. Author Correction: Modeling COVID-19 scenarios for the United States. *Nat Med* **26**, (1950). (2020) <https://doi.org/10.1038/s41591-020-01181-w>
28. Srivastava, A., Singh, S. & Lee, F. Shape-based Evaluation of Epidemic Forecasts. *IEEE International Conference on Big Data (Big Data)*, 1701–1710 (2022), 1701–1710 (2022). (2022). <https://doi.org/10.1109/BigData55660.2022.10020895>
29. Thomas, S. & Maxime, M. Retrospective analysis of Covid-19 hospitalization modelling scenarios which guided policy response in France. *MedRxiv* **2023.2012.2016.23300086** <https://doi.org/10.1101/2023.12.16.23300086> (2023).
30. Gerlee, P. et al. Evaluation and communication of pandemic scenarios. *Lancet Digit. Health.* **6**, e543–e544. [https://doi.org/10.1016/S2589-7500\(24\)00144-4](https://doi.org/10.1016/S2589-7500(24)00144-4) (2024).
31. Sherratt, K. et al. Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. *Elife* **12** <https://doi.org/10.7554/eLife.81916> (2023).

32. Bracher, J., Ray, E. L., Gneiting, T. & Reich, N. G. Evaluating epidemic forecasts in an interval format. *PLoS Comput. Biol.* **17**, e1008618. <https://doi.org/10.1371/journal.pcbi.1008618> (2021).
33. Jit, M. et al. Reflections on epidemiological modeling to inform policy during the COVID-19 pandemic in Western Europe, 2020–23. *Health Aff (Millwood)*. **42**, 1630–1636. <https://doi.org/10.1377/hlthaff.2023.00688> (2023).
34. Luo, Z., Zhang, L., Liu, N. & Wu, Y. Time series clustering of COVID-19 pandemic-related data. *Data Sci. Manage.* **6**, 79–87. <https://doi.org/10.1016/j.dsm.2023.03.003> (2023).
35. Cassisi, C., Montalto, P., Aliotta, M., Cannata, A. & Pulvirenti, A. in *Advances in Data Mining Knowledge Discovery and Applications* (ed Adem Karahoca) Ch. 3 IntechOpen, (2012).
36. Rojas-Valenzuela, I., Valenzuela, O., Delgado-Marquez, E. & Rojas, F. Estimation of COVID-19 Dynamics in the Different States of the United States during the First Months of the Pandemic. *Engineering Proceedings* **5**, 53 (2021).
37. Team, R. C. R: A Language and Environment for Statistical Computing. (2022).

Author contributions

H.D, I.G., F.B. and G.P conceptualized the study, developed the methods, evaluated results, and wrote the first draft of the manuscript. H.D. wrote the code and executed analyses. H.D, I.G, F.B, G.P, P.G, T.L and L.B were involved in writing and approved the final version of the manuscript.

Funding

Open access funding provided by Public Health Agency of Sweden.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-08682-z>.

Correspondence and requests for materials should be addressed to H.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025