



# Temporal Analysis of Infectious Diseases: A Case Study on COVID-19

Jinyang Liu<sup>1,2</sup>, Boping Tian<sup>1\*</sup>, Jiaxuan Wu<sup>2</sup><sup>1</sup> School of Mathematics, Harbin Institute of Technology, 150006 Harbin, China<sup>2</sup> School of Statistics, Chengdu University of Information Technology, 610103 Chengdu, China\* Correspondence: Boping Tian ([bopingt361147@hit.edu.cn](mailto:bopingt361147@hit.edu.cn))

Received: 04-12-2023

Revised: 05-15-2023

Accepted: 05-25-2023

**Citation:** J. Y. Liu, B. P. Tian, and J. X. Wu, “Temporal analysis of infectious diseases: A case study on COVID-19,” *Acadlore Trans. Appl Math. Stat.*, vol. 1, no. 1, pp. 1–9, 2023. <https://doi.org/10.56578/atams010101>.



© 2023 by the authors. Published by Acadlore Publishing Services Limited, Hong Kong. This article is available for free download and can be reused and cited, provided that the original published version is credited, under the CC BY 4.0 license.

**Abstract:** Historically, infectious diseases have greatly impacted human health, necessitating a robust understanding of their trends, processes, and transmission. This study focuses on the COVID-19 pandemic, employing mathematical, statistical, and machine-learning methods to examine its time-series data. We quantify data irregularity using approximate entropy, revealing higher volatility in the U.S., Italy, and India compared to China. We employ the Dynamic Time Warping algorithm to assess regional similarity, finding a strong correlation between the U.S. and Italy. The Seasonal Trend Decomposition using the LOESS algorithm illuminates strong trend degrees in all observed regions, but China’s prevention measures show marked effectiveness. These tools, whilst already valuable, still present opportunities for development in both theory and practice.

**Keywords:** Approximate entropy; Complexity analysis; Dynamic time warping; Seasonal-trend decomposition; Epidemiology; Public health

## 1 Introduction

Throughout history, infectious diseases have persistently posed significant threats to human life. Despite societal advancement during the 20th century, numerous infectious diseases continue to jeopardize global health. In particular, the first documented case of COVID-19, a novel coronavirus, emerged in Wuhan, Hubei Province, China, in December 2019. Characterized by their high transmissibility, severe health implications, and unpredictable epidemic timing, coronaviruses represent a constant menace to human well-being.

Amidst the COVID-19 pandemic, the role of data science has become increasingly prominent, serving as an invaluable tool in mitigating the crisis. This study aims to harness advanced methodologies to deeply explore the temporal characteristics of infectious diseases and discern their propagation dynamics. Gaining these insights is critical for early intervention and the establishment of effective preventive measures against similar infectious diseases in the future.

Analytical tools such as approximate entropy, the Dynamic Time Warping (DTW) algorithm, and Seasonal Trend Decomposition using Loess (STL) decomposition offer promising avenues for time-series analysis. However, their application in infectious disease time-series data remains limited. In this study, we adopt the COVID-19 pandemic as a case study, employing approximate entropy to measure the complexity of infectious disease time series. Additionally, we utilize the DTW algorithm to evaluate similarity characteristics and the STL algorithm for time series decomposition and trend feature extraction.

The analysis considers the number of confirmed COVID-19 cases and associated fatalities in China, the United States, Italy, and India, as well as nine severely affected cities within China. Data were sourced from the World Health Organization, the Johns Hopkins University real-time surveillance system, and the Chinese Health Commission. This comprehensive and nuanced approach allows us to delve into the heart of infectious disease dynamics, offering a valuable foundation for future prevention and control measures.

## 2 Related Works

A variety of quantitative methodologies from statistics, mathematics, infectious disease modeling, and other fields have been employed by researchers to analyze time series data on the spread of infectious diseases. Two core components of such analyses are curve fitting to capture the trajectory of an outbreak and time series forecasting to

predict future case numbers. For example, Aryatama et al. [1] proposed using the SPCIRD model to anticipate the spread of COVID-19 in Indonesia while accounting for public compliance with containment measures, employing the Levenberg Marquardt curve fitting approach to develop an accurate model. De Silva and Abeysundara [2] utilized functional data analysis techniques to model and examine the spread dynamics of the first wave of COVID-19 globally and in Asia. Based on daily reported data from Iran, Gholami et al. [3] compared three continuous distributions—normal, lognormal, and Weibull—to model the distributions of COVID-19 cases and fatalities.

Epidemiological compartmental models, autoregressive models, Kalman filtering, and artificial intelligence have also been applied for real-time prediction of COVID-19 [4]. Kogan et al. [5] proposed an early warning system using multiple digital traces of COVID-19 activity for close to real-time monitoring. Some have successfully predicted the spread of COVID-19 using ARMA models or a combination of ARMA and other models [6, 7].

Many researchers have employed machine learning and deep learning techniques rather than traditional time series analysis approaches for prediction. According to Tamang et al. [8], artificial neural networks are well suited for handling large data sets and can be computationally analyzed to uncover patterns, trends, and predictions. We forecast new COVID-19 cases and deaths in India, the US, France, and the UK using artificial neural network-based curve fitting, accounting for patterns seen in China and South Korea. Comparative studies have also been conducted [9]. Chandra et al. [10] found that cyclic neural networks, a type of deep learning model, are ideal for simulating spatiotemporal sequences. They used popular recurrent neural networks like LSTMs to conduct multi-step prediction of COVID-19 spread in Indian states [11–13].

For statistical analysis of epidemiological characteristics, some have calculated peak and trough amplitudes, disease selectivity, oscillation intensity, and other metrics before and after epidemics, while others have performed straightforward correlation analyses. Yang et al. [14] noted that China's approach to infectious disease prevention and control has changed significantly since the 2003 SARS outbreak, though few studies had examined trends and epidemiological characteristics. Ho et al. [15] observed that exponential growth culminating in a peak is typical of infectious disease epidemics, including coronavirus outbreaks. They used item response theory to identify changes in each country or region, then described inflection point characteristics on a time bar graph to locate the COVID-19 inflection point. Zhao et al. [16] calculated peak-valley amplitudes, disease selectivity, preferred outbreak time, and oscillation intensity for 23 notifiable diseases in China from 2017 to 2021, finding changes before and after epidemics [16, 17].

Approximate entropy, DTW, and STL have rarely been used in infectious disease research. Schneckenreither et al. [18] proposed an effective aggregation dispersion index using approximate entropy for disease research. Rodriguez et al. [19] used deep neural networks and approximate entropy as a tool for early diagnosis of Chagas disease and cardiac damage [20]. Regarding DTW, Dallas et al. [21] hypothesized that geographically proximate locations may have similar infectious disease dynamics. They used DTW to analyze the importance of distance, population size, and age structure in determining similarities between U.S. counties' COVID-19 epidemics. STL has mainly been applied in meteorology, environment, and finance. He et al. [22] combined STL and machine learning to predict rainfall time series. Jainonthee et al. [23] used STL to determine the seasonality of two diseases. Zhao et al. [24] used STL to analyze influenza seasonality in China, then compared SARIMA, SARIMA-LSTM, and SSA-SARIMA-LSTM models for prediction.

In summary, researchers have studied the time series attributes of infectious diseases using qualitative and quantitative techniques, providing crucial guidance for effective pandemic prevention and control. The use of approaches like approximate entropy, DTW, and STL for infectious diseases remains conceptually and empirically underexplored. Further research on integrating these methods with infectious disease modeling is warranted.

### 3 Method

In this study, the complexity of time series data associated with infectious diseases was explored, with a particular focus on the COVID-19 pandemic, using an entropy approximation approach. This method, introduced by Steven M. Pincus in 1991, provides a measure of complexity in signal sequences through the concept of approximate entropy. A key strength of this approach is its robustness in dealing with small data sets. This study hence benefits from this feature, as most measured time series are capable of satisfying the requirements, and the resultant findings exhibit a strong resilience and credibility.

To ascertain the similarity between time series data, we employ the DTW algorithm. This algorithm uses dynamic programming to nonlinearly align time series data, thereby facilitating the accurate computation of similarity between distinct time sequences. The DTW methodology used in this research was provided by the DTAI Research Group at the University of Leuven and is available open-source.

Beyond measuring complexity and similarity in time series data, this study also investigates trends, periodicity, and seasonality. To achieve this, the Seasonal-Trend decomposition based on the STL method was applied to decompose time series data related to COVID-19 from various countries and regions. This analysis facilitates a thorough exploration of trend characteristics inherent in the outbreak patterns of COVID-19.

Collectively, these three methodologies offer valuable insights into the temporal transmission characteristics and patterns of infectious diseases, such as COVID-19. The data used in this study primarily consist of time series of daily new confirmed COVID-19 cases in different countries and regions, from the onset of the outbreak to when each respective area had the pandemic under control.

The aforementioned methodologies were implemented through Python, underscoring the specific computational process. This blend of techniques underscores the innovative and multifaceted approach of the study in comprehending the temporal dynamics of infectious diseases, and in particular, the ongoing COVID-19 pandemic.

### 3.1 Approximate Entropy

Approximate entropy serves as a critical metric in assessing the complexity and irregularity inherent in time series data. It operates on the principle of conditional probability to ascertain the propensity for the emergence of new information within time series data. This effectively encapsulates the likelihood of developing novel subseries within the overall time series data set [25]. The procedural steps to calculate approximate entropy are as follows:

**Step 1:** Suppose we have a time series of length  $N : u(1), u(2), u(3), \dots, u(N)$ . Set a threshold value  $r$  which represents the similarity comparison. Then determine a measure  $m$  that divides the sequence length into subsequences.

**Step 2:** By reconstructing the original sequence, obtain the subsequences  $X(1), X(2), X(3), \dots, X(N-m+1)$ , where each subsequence is denoted by  $X(i)$ .

$$X(i) = u(i), u(i+1), u(i+2), \dots, u(i+m-1) \quad (1)$$

**Step 3:** Calculate the distance  $d_m[X(i), X(j)]$  between any two reconstructed vectors  $X(i)$  and  $X(j)$ , where,  $d_m$  represents the distance between two reconstructed vectors  $X(j) (1 \leq j \leq N-m+1)$  and  $X(i)$ . The distance  $d_m$  is calculated by the maximum difference of the corresponding elements in the two vectors, this includes the distance when  $i$  equals  $j$ ;

**Step 4:** Count the number of vectors that satisfy certain conditions and determine the ratio to the total number of statistics.

$$C_i^m(r) = \frac{\text{num}[d_m(X(i), X(j)) < r]}{N-m+1} \quad (2)$$

This process is called the template matching process of  $X(i)$ ,  $C_i^m(r)$  represents the matching probability between any  $X(j)$  and the template;

**Step 5:** Define the average similarity rate when the number of molecular sequences is  $m$ .

$$\Phi_m(r) = \frac{\sum_{i=1}^{N-m+1} \log(C_i^m(r))}{N-m+1} \quad (3)$$

**Step 6:** Based on Steps 1 through 5, calculate the average similarity rate  $\Phi_{m+1}(r)$  when the number of molecular sequences is  $m+1$ .

**Step 7:** Approximate entropy is calculated.

$$ApEn = \Phi_m(r) - \Phi_{m+1}(r) \quad (4)$$

For the ApEn.alpha.R program, it is the equality transformation of Step 7, then there is

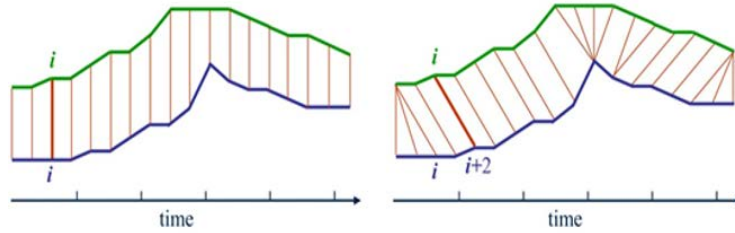
$$\begin{aligned} ApEn &= \Phi_m(r) - \Phi_{m+1}(r) \\ &= \frac{\sum_{i=1}^{N-m+1} \ln(C_i^m(r))}{N-m+1} - \frac{\sum_{i=1}^{N-m} \ln(C_i^{m+1}(r))}{N-m} \\ &= \frac{\sum_{i=1}^{N-m} [\ln C_i^m(r) - \ln C_i^{m+1}(r)]}{N-m} \\ &= \frac{\sum_{i=1}^{N-m} \left[ \ln \frac{C_i^m(r)}{C_i^{m+1}(r)} \right]}{N-m} \end{aligned} \quad (5)$$

It's important to note that for the selection of  $m$  and  $r$ ,  $m$  is typically chosen as 2 or 3, while  $r$  is chosen based on the actual application scenario. In this study,  $r$  is selected to be 0.2, meaning it is 0.2 times the standard deviation of the original time series.

### 3.2 Dynamic Time Warping (DTW)

Each point in a time series must be mapped to one or more points in another time series for the series to be aligned along the timeline. This is the underlying mechanism of the dynamic time-warping (DTW) algorithm. Dynamic programming is a technique that can be used to determine the optimal mapping [26].

As shown in Figure 1, the mapping between time series points is no longer one-to-one; rather, it allows a one-to-many or many-to-one relationship. The right panel of Figure 1 depicts two time series represented by the green and blue lines; similar points between the two sequences are connected by the red lines. The DTW algorithm measures the similarity between time series using the total distance between these matched points, known as the integration path distance.

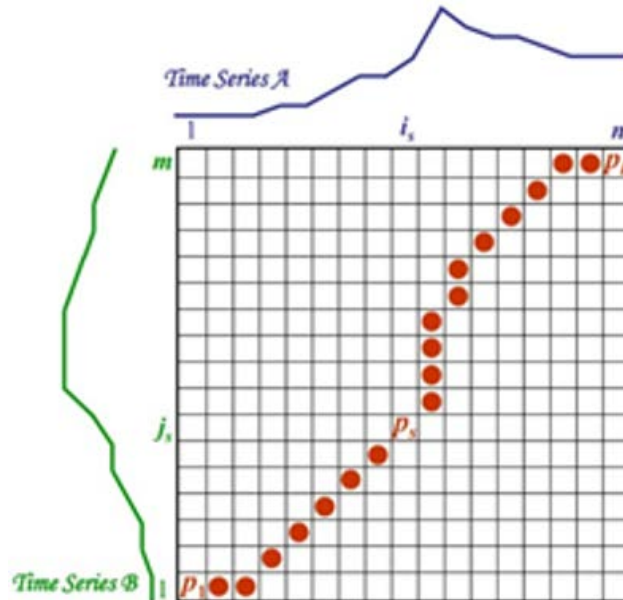


**Figure 1.** The way Euclidean distance (left) corresponds to DTW distance (right)

It should be noted that relevant parameters, such as time series length, window size, etc., should be chosen according to the actual data when utilizing the DTW method for time series similarity analysis. Due to the high computational complexity of the DTW algorithm, it may be necessary to optimize or choose different algorithms for large-scale data sets.

Suppose there are two time series  $A$  and  $B$ , with lengths  $n$  and  $m$ , respectively. Comparing the two sequences with an  $n \times m$  matrix, the warping path  $P = \{p_1, \dots, p_s, \dots, p_S\}$  will pass through this matrix. The  $s$ -th element of the warping path is represented by  $p_s = (i_s, j_s)$ , where  $i$  and  $j$  represent the corresponding points of the two sequences, respectively.

The goal of the DTW algorithm is to find an optimal warping path  $p$ . This is essentially an optimization problem and can be expressed mathematically as  $DTW(A, B) = \min \sum_{s=1}^S (p_s)$ . To solve this optimization problem, three conditions must be met:



**Figure 2.** The solution path of the DTW algorithm

**Condition 1: Boundary conditions.**  $p_1 = (1, 1)$  and  $p_S = (n, m)$  mean that the beginning and end of the two sequences must match. The path starts from the bottom left and ends at the top right, ensuring that the entire sequence is taken into account.

**Condition 2: Continuity.** If  $p_s = (a, b)$  and  $p(s-1) = (a', b')$ ,  $a - a' \leq 1$  and  $b - b' \leq 1$  must be satisfied. This constraint indicates that in the matching process, many-to-one and one-to-many cases can only match the surrounding time step. That is to say, it is impossible to cross a certain point for matching. You can only align points adjacent to you. This keeps the warping path free of jumps and every coordinate in A and B appears in the warping path.

**Condition 3: Monotonicity.** If  $p_s = (a, b)$ , and  $p(s-1) = (a', b')$ ,  $a - a' \geq 0$  and  $b - b' \geq 0$  must be satisfied, indicating that the warping path does not take a backward path. This ensures that each coordinate will not be repeated in the path and that the warping path increases monotonically over time.

The monotonicity and continuity of the warping path  $p$  mean that  $p$  can only proceed in three ways: one space to the right, one space up, and one space diagonally up to the right. In addition to the boundary conditions of  $p$ , the solution of the optimal  $p$  becomes a dynamic programming problem. Let's call this dynamic programming problem  $\gamma$ , then

$$\gamma(i, j) = d(a_i, b_j) + \min(\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)) \quad (6)$$

Therefore, the similarity between two-time series A and B is obtained by the DTW algorithm (Figure 2).

### 3.3 STL Decomposition

Seasonal Trend decomposition procedure based on Loess (STL) is a popular algorithm used in time series decomposition. It breaks down data  $Y_v$  into a trend component, a seasonal component, and a residual component [27]. In other words,  $Y_v = T_v + S_v + R_v, v = 1, \dots, N$ . STL comprises an inner and outer cycle, with the inner cycle primarily used for trend fitting and seasonal component calculation.

At the end of the  $(k-1)$ th pass in the inner cycle,  $T_v^{(k)}$  and  $S_v^{(k)}$  are assumed to be the trend and seasonal components respectively.  $T_v^{(k)} = 0$  is initialized to zero. Several parameters are defined as follows:

- $n_{(i)}$  is the number of inner cycles;
- $n_{(0)}$  is the outer cycle number;
- $n_{(p)}$  is the number of samples in a period;
- $n_{(s)}$  is the smoothing parameter described in Step 2;
- $n_{(l)}$  is the smoothing parameter described in Step 3;
- $n_{(t)}$  is the smoothing parameter described in Step 6.

Sample points at the same position in each cycle form a subsequence, referred to as a cycle-subseries. There are a total of  $n_{(p)}$  such subsequences. The inner cycle is mainly divided into the following 6 steps:

**Step1** Detrend by subtracting the trend component of the previous iteration,  $Y_v - T_v^{(k)}$ ;

**Step2** Use the LOESS ( $q = n_{(s)}, d = 1$ ) to do the regression for each subsequence and extend one cycle forward and backward. Smooth the result of temporary seasonal sequence, remember to  $C_v^{(k+1)}, v = -n_{(p)} + 1, \dots, -N + n_{(p)}$ ;

**Step3** Filter the low flux of the periodic subsequence. Do the sliding average of the result sequence  $C_v^{(k+1)}$  in turn, and then do the regression of LOESS ( $q = n_{(l)}, d = 1$ ). The result sequence  $L_v^{(k+1)}, v = 1, \dots, N$  is equivalent to extracting the low flux of the periodic subsequence;

**Step4** removes the subsequence trend, smooth cycle  $S_v^{(k+1)} = C_v^{(k+1)} - L_v^{(k+1)}$ ;

**Step5** Decycle trend, subtract the periodic component,  $Y_v - S_v^{(k+1)}$ ;

**Step6** Trend smoothing is described. For the sequence after removing the period, do the regression of LOESS ( $q = n_{(t)}, d = 1$ ), and obtain the trend component  $T_v^{(k+1)}$ .

The outer cycle is primarily used to adjust the robustness weight. If there are outliers in the data series, the residual will be larger. Define  $h = 6 * \text{median}(|R_v|)$ , for the location of  $v$  data points, its robustness weighting for  $\rho_v = B(|R_v|/h)$ , which functions as the bisquare function  $B$ :

$$B(u) = \begin{cases} (1 - u^2)^2 & \text{for } 0 \leq u < 1 \\ 0 & \text{for } u \geq 1 \end{cases}$$

Then in the inner cycle of each iteration, when doing the regression in Step2 and Step6, the neighborhood weight needs to be multiplied by  $\rho_v$  to reduce the influence of outliers on the regression.

## 4 Result and Discussion

### 4.1 Complexity Characteristics of Infectious Disease Time Series

Table 1 reveals that among the four countries, the approximate entropy calculation for the time series of daily new COVID-19 cases in the United States (0.5104) is significantly higher than that of China (0.0640). Likewise, the

entropy values for Italy (0.2201) and India (0.3144) are also significantly higher than China's. This suggests that the course of COVID-19 in these three countries has been relatively more unpredictable and volatile.

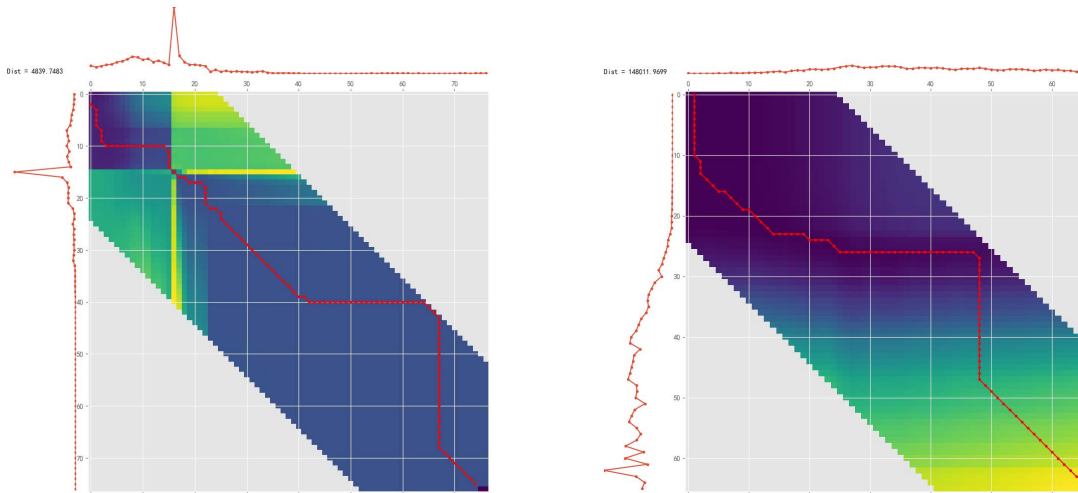
Looking at the nine cities in China, only Chengdu (0.2523) and Harbin (0.2588) have higher approximate entropy values than the other seven cities, though still lower than those of the United States and India. This indicates that the time series of daily new COVID-19 cases in China is considerably less complex or irregular than in countries such as the United States, Italy, and India. This is likely due to China's timely implementation of effective, scientific, and reasonable preventive measures, which have yielded significant results.

**Table 1.** Approximate entropy of time series of newly confirmed COVID-19 cases in selected countries or regions

Country or city	$\Phi_m(r)$	$\Phi_{m+1}(r)$	$ApEn$
China	-0.7531	-0.8171	0.0640
America	-2.2338	-2.7442	0.5104
Italy	-1.7777	-1.9978	0.2201
India	-2.0425	-2.3569	0.3144
Chengdu	-1.0668	-1.3191	0.2523
Wuhan	-0.2869	-0.3154	0.0285
Hangzhou	-0.7877	-0.9403	0.1526
Guangzhou	-0.3794	-0.3981	0.0187
Urumqi	-0.7994	-0.9064	0.1070
Harbin	-1.2781	-1.5369	0.2588
Xi'an	-0.7038	-0.8648	0.1610
Beijing	-0.4857	-0.5227	0.0369
Shanghai	-0.3608	-0.3866	0.0259

#### 4.2 The Similarity Feature of Time Series of Infectious Diseases

This study employs the DTW algorithm to perform pairwise calculations on epidemic time series data from various nations and regions. The findings suggest that Wuhan shows a strong correlation with the national epidemic in China, as depicted in Figure 3(left), which is due to the city being the initial epicenter of the outbreak. Among the four countries, only the United States shows a strong correlation with the Italian outbreak (Figure 3(right)).



**Figure 3.** The results of the DTW algorithm in America and Italy (left), China overall, and Wuhan (right)

From a historical standpoint, no event occurs without cause or warning. A case in point is Italy, which became severely affected by COVID-19. Some may attribute this to the government's inefficiency or societal factors, but these are common across Europe and the Western world. As per real-time statistics from Johns Hopkins University, as of 11 a.m. Beijing time on April 2, 2020, the U.S. had 215,417 confirmed cases and 5,116 deaths. The U.S., being the first country to surpass 200,000 confirmed cases, is likely the source of this global disaster.

### 4.3 Trend Characteristics of Infectious Disease Time Series

The STL algorithm is used to decompose the time series of infectious diseases, effectively revealing trend, periodic, and seasonal characteristics. This study selected four countries and nine cities in China severely affected by COVID-19 for STL decomposition of time series data for daily new cases and deaths.

Time series data decomposes into trend (T), seasonal (S), and residual (R) components:  $y_t = T_t + S_t + R_t$ . For strongly trending data, the variance of residuals should be smaller than seasonally adjusted data, thus  $\text{Var}(R_t) / \text{Var}(T_t + R_t)$  is small. For time series with weak or no trend, these variances should be comparable. Therefore, the trend strength is defined as:

$$F_T = \max\left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(T_t + R_t)}\right) \quad (7)$$

This yields a measure of trend strength between 0 and 1. Similarly, the intensity of seasonality is defined using data after trend adjustment:

$$F_S = \max\left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(S_t + R_t)}\right) \quad (8)$$

When  $F_S$  was close to 0, it indicated that there was almost no seasonality in the sequence, and when  $F_S$  was close to 1, it indicated that  $\text{Var}(R_t)$  of the sequence was much smaller than  $\text{Var}(S_t + R_t)$ .

Table 2's calculations indicate that for the time series of new COVID-19 cases, China, the United States, Italy, India, and eight Chinese cities show strong trends, while Wuhan shows a weak trend. This can be attributed to China's consistent and effective preventive measures. When considering the trend strength of new death numbers, it is clear that China's control measures have been more effective than those in the United States, Italy, and India. In terms of seasonal intensity, the United States and Italy show strong seasonal characteristics for new cases and deaths.

Time series with more trends and seasonality produces more predictable data. Hence, it is important to decompose the data with STL and calculate its trend and seasonal intensity before forecasting the progression of infectious diseases. This will aid in better model selection and improve prediction accuracy.

**Table 2.** Approximate entropy of time series of newly confirmed COVID-19 cases in selected countries or regions

Country or city	Number of newly confirmed cases		Number of new deaths	
	$F_T$	$F_S$	$F_T$	$F_S$
China	0.8653	0.3114	0.3621	0.3492
America	0.9486	0.7690	1.0000	0.8438
Italy	0.9863	0.8463	1.0000	0.5270
India	0.9899	0.5133	1.0000	0.2505
Chengdu	0.7004	0.1556	0.2155	0.3685
Wuhan	0.5874	0.3115	0.3066	0.3483
Hangzhou	0.7822	0.3214	0.2200	0.3337
Guangzhou	0.9247	0.3372	0.3741	0.3945
Urumqi	0.8725	0.2898	0.2178	0.2887
Harbin	0.8569	0.3998	0.5291	0.5654
Xi'an	0.9547	0.2368	0.3865	0.6783
Beijing	0.9767	0.3951	0.2155	0.3685
Shanghai	0.9259	0.4397	0.3066	0.3483

## 5 Conclusion

In light of the profound impact and the multifaceted challenges the COVID-19 pandemic has exerted on global health systems, this study leverages a blend of approximate entropy, Dynamic Time Warping (DTW) algorithm, and Seasonal Trend decomposition procedure based on Loess (STL) to draw pivotal insights into infectious disease research.

Firstly, approximate entropy serves as an effective tool in assessing the complexity of time series data. The computational findings of this study provide an objective and quantifiable gauge of the volatility of the infectious disease time series across various countries and regions. This methodology lends a new perspective to understanding and interpreting the course of epidemic progression.

Secondly, the DTW algorithm was employed to ascertain the congruity of COVID-19 spread between four countries and nine cities in China. This calculation holds significant potential in facilitating comparative analysis of infectious disease transmission across different geographical entities, thus providing valuable insights into disease transmission correlations.

Finally, the STL algorithm was utilized to decompose the time series of infectious diseases and determine their trend characteristics. This information offers a practical basis to analyze and evaluate the efficacy of epidemic prevention and control measures implemented across different countries or regions.

Despite the notable findings and conclusions, several thought-provoking questions emanate from this study that warrants further exploration. A fascinating point of discussion is the potential implications of high and low entropy on the course of the epidemic. Further inquiry could shed light on the causes of the observed disparities between countries and cities.

Moreover, using the DTW algorithm's calculation results, deeper discussions could be initiated to unravel the relationship between places exhibiting strong disease transmission similarities. Regarding STL decomposition, future discourse could aim to relate the findings of this study with disease control efforts more directly.

In summation, this study invites more nuanced and comprehensive investigations into these intriguing issues, contributing to the ongoing dialogue surrounding global public health crises.

### Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

### Conflicts of Interest

The authors declare that they have no conflicts of interest.

### References

- [1] F. Y. Aryatama, F. I. Kurniadi, and N. I. Manik, "Mathematical model estimation and prediction application of COVID-19 infection in Indonesia using Levenberg-Marquardt Algorithm based on Python," *Procedia Comput. Sci.*, vol. 216, pp. 120–127, 2023. <https://doi.org/10.1016/j.procs.2022.12.118>
- [2] J. W. E. W. De Silva and S. P. Abeyesundara, "Functional data analysis on global COVID-19 data," *Asian J. Probab. Stat.*, pp. 12–28, 2023. <https://doi.org/10.9734/ajpas/2023/v21i1453>
- [3] E. Gholami, K. Mansori, and M. Soltani-Kermanshahi, "Statistical distribution of novel coronavirus in Iran," *Research Square*, 2020. <https://doi.org/10.21203/rs.3.rs-19181/v1>
- [4] J. A. L. Marques, F. N. B. Gois, J. Xavier-Neto, and S. J. Fong, *Predictive Models for Decision Support in the COVID-19 Crisis*. Springer International Publishing, 2021. <https://doi.org/10.1007/978-3-030-61913-8>
- [5] N. E. Kogan, L. Clemente, P. Liautaud, J. Kaashoek, N. B. Link, A. T. Nguyen, and M. Santillana, "An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time," *Sci. Adv.*, vol. 7, no. 10, p. eabd6989, 2021. <https://doi.org/10.1126/sciadv.abd6989>
- [6] C. Barría-Sandoval, G. Ferreira, K. Benz-Parra, and P. López-Flores, "Prediction of confirmed and death cases of Covid-19 in Chile through time series techniques: A comparative study," *Cold Spring Harbor Laboratory*, 2021. <https://doi.org/10.1101/2020.12.31.20249085>
- [7] C. Katris, "A time series-based statistical approach for outbreak spread forecasting: Application of COVID-19 in Greece," *Expert Syst. Appl.*, vol. 166, p. 114077, 2021. <https://doi.org/10.1016/j.eswa.2020.114077>
- [8] S. K. Tamang, P. D. Singh, and B. Datta, "Forecasting of Covid-19 cases based on prediction using artificial neural network curve fitting technique," *Glob. J. Environ. Sci. Manage.*, vol. 6, no. Special Issue (Covid-19), 2020. <https://doi.org/10.22034/GJESM.2019.06.SI.06>
- [9] S. Balli, "Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods," *Chaos Solitons Fractals*, vol. 142, p. 110512, 2021. <https://doi.org/10.1016/j.chaos.2020.110512>
- [10] R. Chandra, A. Jain, and D. Singh Chauhan, "Deep learning via LSTM models for COVID-19 infection forecasting in India," *PLoS ONE*, vol. 17, no. 1, p. e0262708, 2022. <https://doi.org/10.1371/journal.pone.0262708>
- [11] H. Abbasimehr and R. Paki, "Prediction of COVID-19 confirmed cases combining deep learning methods and Bayesian optimization," *Chaos Solitons Fractals*, vol. 142, p. 110511, 2021. <https://doi.org/10.1016/j.chaos.2020.110511>
- [12] I. H. Aslan, M. Demir, M. M. Wise, and S. Lenhart, "Modeling COVID19: Forecasting and analyzing the dynamics of the outbreaks in Hubei and Turkey," *Math. Meth. Appl. Sci.*, vol. 45, no. 10, pp. 6481–6494, 2022. <https://doi.org/10.1002/mma.8181>



- [13] F. Saleem, A. S. A. M. AL-Ghamdi, M. O. Alassafi, and S. A. AlGhamdi, "Machine learning, deep learning, and mathematical models to analyze forecasting and epidemiology of COVID-19: A systematic literature review," *Int. J. Environ. Res. Public Health*, vol. 19, no. 9, p. 5099, 2022. <https://doi.org/10.3390/ijerph19095099>
- [14] S. Yang, J. Wu, C. Ding, Y. Cui, Y. Zhou, Y. Li, M. Deng, C. Wang, K. Xu, J. Ren, B. Ruan, and L. Li, "Epidemiological features of and changes in incidence of infectious diseases in China in the first decade after the SARS outbreak: An observational trend study," *Lancet Infect. Dis.*, vol. 17, no. 7, pp. 716–725, 2017. [https://doi.org/10.1016/s1473-3099\(17\)30227-x](https://doi.org/10.1016/s1473-3099(17)30227-x)
- [15] S. Y. C. Ho, T. W. Chien, Y. Shao, and J. H. Hsieh, "Visualizing the features of inflection point shown on a temporal bar graph using the data of COVID-19 pandemic," *Medicine*, vol. 101, no. 5, p. e28749, 2022. <https://doi.org/10.1097/md.00000000000028749>
- [16] X. Zhao, M. Li, N. Haihambo, J. Jin, Y. Zeng, J. Qiu, M. Guo, Y. Zhu, Z. Li, J. Liu, J. Teng, S. Li, Y. N. Zhao, Y. Cao, X. Wang, Y. Li, M. Gao, X. Feng, and C. Han, "Changes in temporal properties of notifiable infectious disease epidemics in China during the COVID-19 pandemic: Population-based surveillance study," *JMIR Public Health Surveill.*, vol. 8, no. 6, p. e35343, 2022. <https://doi.org/10.2196/35343>
- [17] J. M. Kang, J. Jung, Y. E. Kim, K. Huh, J. Hong, D. W. Kim, M. Y. Kim, S. Y. Jung, J. H. Kim, and J. G. Ahn, "Temporal correlation between Kawasaki disease and infectious diseases in South Korea," *JAMA Netw. Open*, vol. 5, no. 2, p. e2147363, 2022. <https://doi.org/10.1001/jamanetworkopen.2021.47363>
- [18] G. Schneckreither, L. Herrmann, R. Reisenhofer, N. Popper, and P. Grohs, "Assessing the heterogeneity in the transmission of infectious diseases from time series of epidemiological data," 2022. <https://doi.org/10.1101/2022.02.21.22271241>
- [19] M. F. Rodriguez, A. Ravelo-Garcia, E. Alvarez, L. A. Diaz, D. Rodrigo Cornejo, V. Andres Cabrera-Caso, D. Condori-Merma, and M. Vizcardo, "Approximate entropy and densely connected neural network in the early diagnostic of patients with chagas disease," in *Computing in Cardiology Conference (CinC)*. Computing in Cardiology, 2022. <https://doi.org/10.22489/cinc.2022.313>
- [20] A. Makani, A. Akhavan, F. Shahbazi, M. Noruzi, and M. Zare, "Age-related complexity of the resting state MEG signals: A multiscale entropy analysis," 2022. <https://doi.org/10.1101/2022.09.14.507986>
- [21] T. A. Dallas, G. Foster, R. L. Richards, and B. D. Elder, "Epidemic time series similarity is related to geographic distance and age structure," *Infect. Dis. Modell.*, vol. 7, no. 4, pp. 690–697, 2022. <https://doi.org/10.1016/j.idm.2022.09.002>
- [22] R. He, L. Zhang, and A. W. Z. Chew, "Modeling and predicting rainfall time series using seasonal-trend decomposition and machine learning," *Knowl.-Based Syst.*, vol. 251, p. 109125, 2022. <https://doi.org/10.1016/j.knosys.2022.109125>
- [23] C. Jainonthee, Y. L. Wang, C. W. Chen, and K. Jainonthee, "Air pollution-related respiratory diseases and associated environmental factors in Chiang Mai, Thailand, in 2011–2020," *Trop. Med. Infect. Dis.*, vol. 7, no. 11, p. 341, 2022. <https://doi.org/10.3390/tropicalmed7110341>
- [24] Z. Zhao, M. Zhai, G. Li, X. Gao, W. Song, X. Wang, H. Ren, Y. Cui, Y. Qiao, J. Ren, L. Chen, and L. Qiu, "Study on the prediction effect of a combined model of SARIMA and LSTM based on SSA for influenza in Shanxi Province, China," *BMC Infect. Dis.*, vol. 23, no. 1, 2023. <https://doi.org/10.1186/s12879-023-08025-1>
- [25] G. N. Reissig, T. F. de Carvalho Oliveira, A. G. Parise, Á. V. L. Costa, D. A. Posso, C. V. Rombaldi, and G. M. Souza, "Approximate entropy: A promising tool to understand the hidden electrical activity of fruit," *Commun Integr Biol*, vol. 16, no. 1, 2023. <https://doi.org/10.1080/19420889.2023.2195236>
- [26] Z. Tao, Q. Xu, X. Liu, and J. Liu, "An integrated approach implementing sliding window and DTW distance for time series forecasting tasks," *Appl Intell*, pp. 1–12, 2023. <https://doi.org/10.1007/s10489-023-04590-9>
- [27] X. He, Y. Li, J. Tan, B. Wu, and F. Li, "OneShotSTL: One-Shot Seasonal-Trend decomposition for online time series anomaly detection and forecasting," *arXiv*, 2023. <https://doi.org/10.48550/ARXIV.2304.01506>