https://africanjournalofbiomedicalresearch.com/index.php/AJBR
Afr. J. Biomed. Res. Vol. 28(2s) (February 2025); 1063- 1082
Research Article

A Nonparametric Approach to Early Warning Signs using COVID-19 data from South Africa

Claris Shoko[1*] and Caston Sigauke[2]

[1*]Department of Statistics, University of Botswana, Private Bag 0022, Gaborone, Botswana.
[2]Department of Mathematical and Computational Science, University of Venda, Private Bag X5050, Thohoyandou 0950, Limpopo, South Africa

Abstract
Introduction: Predicting outbreaks in epidemic events is crucial to developing and implementing effective mitigation measures by the public health sector.
Methods: This study examines the metric-based approach and the model-based approach (nonparametric Diff-Diffusion-Jump (DDJ) model) for early-warning signals (EWS) using the COVID-19 spread data from South Africa.
Results: The standard deviation is the most effective generic EWS. An increase in the standard deviation towards the critical points is shown with a positive Kendall-tau on all waves. The BDS test gives strong evidence of nonlinearity in all five waves. Conditional heteroskedasticity confirmed the periods of critical transitions that were captured by the generic EWS around day 140 and day 165 of wave 1, between day 110 and 120 of the second wave, and from day 120 onwards from wave 2. The results from the DDJ model are also consistent with results from the metric-based EWS. Sensitivity analysis for the robustness of the fitted model is based on the standard deviation and shows positive and increasing Kendall-tau estimates across all rolling windows. Results from these different approaches are consistent.
Conclusion: This study provides insights into pandemics that have similar characteristics as COVID-19 and informs policy making and implementation.

Keywords: Early-warning signals, Generic earlywarning signals, Drift-Diffusion-jump model, COVID-19.

***Author of Correspondence Email:** Shokoc@ub.ac.bw*

Received 01/01/2025, Acceptance 20/01/2025

DOI: https://doi.org/10.53555/AJBR.v28i2S.6261

1 Introduction

1.1 Background

Near a tipping point, small changes in a certain parameter cause an irreversible shift in the behavior of a system, called critical transitions. Critical transitions can be observed in a variety of complex dynamical systems, ranging from ecology to financial markets, climate change, molecular bio-systems, health, and disease. As critical transitions can occur suddenly and are hard to manage, it is important to predict their occurrence.

Rapid and often unexpected punctuated shifts characterise the spread of the COVID-19 pandemic and these have severe consequencies to human well-being. The ability to identify and predict outbreaks during epidemics and pandemic events is crucial to developing and implementing effective mitigation measures by the public health sector. Knowing the number of states that describe the dynamics of COVID-19 is crucial because it helps to model the system's behaviour and understand the underlying dynamics [2]. Livina et al. [2] further argued that with some time series data, it is possible to visualise the behaviour of the data. However, when two

or more states are present, appropriate models are used to determine a precise number of states. Dakos et al. [3] assert that critical transitions can occur unexpectedly and therefore, difficult to manage, hence the need to develop methods that can be used to identify when a critical transition is approaching. Sudden transitions to high COVID19 incidence come with costs that include overburdened health resources resulting in loss of lives. Detection of early-warning signals could provide important insights for surveillance of pandemics such as COVID-19 and to inform policy making and implimentation [4].

Mathematical models have primarily been used to track the number of infection cases and to predict trajectories of the number of cases. Although the standard mathematical models are useful in evaluating the trajectories of disease spread, their main focus is primarily on disease spread after outbreak [4].

1.2 Literature review

In performing early warning signal (EWS) assessments, O'Brien [10] stated that data showing seasonality or cycles should be detrended before fitting the models. This is done to avoid models interpreting the cycles as transitions. The author also argues that in performing EWS assessments, expanding windows have a stronger evidence base compared to rolling windows for what constitutes a warning. In another study, O'Brien and Clements [11] argue that composite early warning signals consisting of autocorrelation, skewness, and variance can predict with varying degrees of accuracy between waves of COVID-19 daily confirmed cases. The study was done across twenty-four countries of the world. The authors state that using sequential EWS assessment could assist policymakers in improving the accuracy of urgent intervention decisions.

Through simulations, Dablander et al. [12] show a lack of time scale separation during the second European epidemic wave of COVID-19. They concluded that the theory of critical slowing down only applies when the external forcing of the system across critical points is slow relative to the internal system dynamics.

Using a trait-based ecological model and analysis of real-world fisheries data, Clements et al. [13] assessed whether early warning signals are present before overexploited marine systems recover. Results from this study show that trait-based and abundance signals were detectable independently prior to the recovery of stocks. It was also observed that combining these two signals resulted in the best predictions of the recovery.

Using different scenarios, Weinans et al. [14] evaluated the performance of different multivariate indicators of resilience loss. Empirical results showed that there was no single method which outperformed the others. The authors then proposed guidelines on which multivariate resilience indicator best suits a particular system.

In their study, Deb et al. [15] developed a deep neural network, the early warning signal network (EWSNet), for detecting critical transitions. The authors argue that this developed model can be used on a wide range of complex systems without the need for information on the system's structure being monitored. In another related study, Bury et al. [16] argue that generic early warning indicators designed to work across all systems do not provide information on the state beyond the tipping point. Empirical results from the study show how deep learning algorithms can provide EWS of tipping points in real-world systems. Their developed algorithm predicts that certain qualitative aspects of a new state are more sensitive and generate fewer false positives than generic indicators.

Southall et al. [17] used a variety of well-known early warning signal detection methods in a simulated study of an infectious disease system undergoing disease elimination. A comparative analysis was done with alternative algorithms from the change-point analysis literature. The authors argue that EWSNet captures features indicative of approaching transitions and characteristics the generic CSD-based EWSs do not capture.

In their study, Dessavre et al. [20] show that the challenge faced by analysing warning signals in time series data is accurately detrending the signal to preserve the statistical properties of the fluctuations. Results from the study showed that detrending using the mean of just four realisations processes can significantly improve when compared to using a moving window average.

The study of early warning signals in the field of COVID-19 has not been widely explored. Li et al [4] examined two early-warning signals, the increase in autocorrelation at lag-1 and standard deviation in the time series of population visits to points of interest from 17 metropolitan cities of the United States of America. The limitation of their study is that it was carried out in a short time frame which might lead to false positives.

1.3 Contribution and research highlights

Based on the literature review in Section 1.2 the study provides critical early warning signals for transitions in the spread of COVID-19 across multiple pandemic waves, using statistical analysis of autocorrelation, standard deviation, skewness, and conditional heteroskedasticity to identify patterns of disease escalation and decline. The study also applies the nonparametric drift-diffusion-jump (DDJ) model to identify early warning signals of critical transitions across multiple COVID-19 waves, highlighting robust trends in variance, diffusion, and jump intensity. Sensitivity analysis is used to assess the robustness of the fitted models at various rolling windows. The research highlights of this study are:

Autocorrelation at lag-1 (ACF(1)) increased during wave 1, with moderate trends in both the original and residual series as indicated by Kendall's tau values (τ = 0.455 and τ = 0.461, respectively).

The standard deviation for wave 1 increased initially but decreased after day 150, with a strong positive correlation between standard deviation and time (τ = 0.924).

Skewness for wave 1 showed a sharp increase around day 80, signaling a transition to higher disease spread, followed by a decrease after day 90 indicating a lower spread.

Wave 2 saw a slower increase in standard deviation and a more flickering skewness, with a significant state

change observed after day 100, marked by increasing skewness and return rates.

In wave 3, early warning signals indicated significant changes in standard deviation, skewness, and return rates between days 100 and 150, leading to a peak in COVID-19 cases around day 140.

Conditional heteroskedasticity was significant at key transitions across all three waves, with notable signals detected on days 80 and 140–165 for wave 1, around day 100 for wave 2, and consistently from day 125 in wave 3.

The nonparametric drift-diffusion-jump (DDJ) model applied to wave 1 showed a decrease in both conditional and total variance, signaling a critical transition toward the end of the wave.

In wave 2, an increase in conditional variance between days 60 and 100, followed by a rise in the diffusion term, signaled an upcoming transition after day 100.

Wave 3, the longest, exhibited clearer trends after detrending, with increases in conditional variance from day 90 to 120 and again from day 150 to 210, indicating transitions into wave 4.

Sensitivity analysis confirmed significant trends in standard deviation for all rolling windows across waves, except for wave 3, which was significant only for windows below 90.

The DDJ model and sensitivity analysis revealed the robustness of early warning signals, with consistent results across multiple pandemic waves.

The rest of the paper is organized as follows: Section 2 presents the models, empirical results are discussed in Section 3. Section 4 concludes.

2 Materials and Methods

2.1 Data

The study uses confirmed daily cases of COVID-19 from South Africa for the period 6 March 2020 to 27 November 2022. The data is accessible on https://github.com/csigauke

2.2 Metric-based early warning signals for critical transitions

Early warning signals for critical transitions can be estimated using methods that can be classified into two broad categories: metric-based and model-based methods. The ultimate goal of both classifications is to capture changes in the memory (correlation structure) and variability of a time series as it approaches a transition into an alternative dynamic regime. These two classifications quantify changes in the properties of the observed time series of a system that is generated by a general process given in (1):

$$dx_t = f(x_t,\theta_t)dt + g(x_t,\theta_t)dW, \quad (1)$$

where x_t is the state of the system, $f(x_t,\theta_t)$ describes the deterministic part of the system, and $g(x_t,\theta_t)$ is the stochastic part and dW is a white-noise process at time t. A slow change in the underlying condition driver, θ, moves the system close to a threshold where a transition may occur.

In this study, we use the metric-based approaches to detect the early warning signals of COVID-19 for each wave. Metric-based methods quantify changes in memory and variance properties of a time series without attempting to fit the data with a specific model structure. The metric-based approaches considered here are the generic early warning signals with Gaussian filtering, BDS test, and conditional heteroskedasticity.

2.2.1 Generic early warning signals

This approach makes use of the generic early warning indicators, namely, lag-1 autocorrelation (AR(1)), variance, skewness, and kurtosis. Some pattern in the correlation and the variation of the time series data are observed and these patterns are used to predict critical transitions [9]. [1] argued that autocorrelation is the most appropriate indicator for a critical transition and an increasing lag-1 autocorrelation signals that the system is approaching a transition. The lag-1 autocorrelation is calculated according to [1] as follows (2):

$$\rho_1 = \frac{E\left[(x_t-\mu)(x_{t+1}-\mu)\right]}{\sigma^2}, \quad (2)$$

where E is the expectation and μ and σ^2 are the mean and variance of the state varible x_t, respectivel.

On the other hand, slowing down and flickering increase the variance before a critical transition. Flickering refers to the strong noises that push the system across boundaries of alternative states whereas slow recovery rate prior to a transition makes the system state drift widely around the stable state to [1]. The variance is defined as (3):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_t-\mu)^2, \quad (3)$$

where N is the number of observations. Along the same vein, slowing down and flickering can also increase skewness and kurtosis before a transition.

2.2.2 BDS test

The BDS test (after the initials of W. A. Brock, W. Dechert and J. Scheinkman) is a powerful tool for detecting nonlinear, (an indication of chaos) serial dependence in time series [23]. The BDS test helps to avoid false detection of critical transitions due to mopdel specification making it a useful tool for detecting early warning signals. Detrending (or first-differencing) of the time series is done to remove linear dependencies from the data since the BDS test cannot test chaos directly but only nonlinearity. Dtrending is done by fitting any linear model (for example using the traditional ARMA(p,q), ARCH(q) or GARCH(p,q) models). The BDS then tests the null hypothesis that the remaining residuals are independent and identically distributed (i.i.d.) against an unspecified alternative hypothesis [34].

If the i.i.d. hypothesis is rejected, then the time series must still contain structure, such as hidden nonlinearity, hidden nonstationarity, or another kind of structure that detrending or model fitting missed. The BDS test is anticipated to reject the i.i.d. hypothesis in the residual time series from a system that is close to a critical transition since large nonlinear responses are thought to be what initiates critical transitions. As an ad hoc diagnostic test, the BDS test can be useful for identifying nonlinearities in time series before

transitions. An additional strong leading indication increases the likelihood that the identified early warning is not a false positive if the BDS test rejects the i.i.d. hypothesis. The BDS test's mathematical specifications can be found in [27]

2.2.3 Conditional heteroskedasticity

Conditional heteroskedasticity is another metric for assessing how a time series' variability pattern has changed. Conditional heteroscedasticity was defined by [18] as a persistent variance feature of time series with clustered volatility. This indicates that there is a positive correlation between variance at one time step and variance at one or more prior time steps, as per [34]. It is implied that significant variations in variability tend to be followed by significant variations in variability and vice versa. [19]. Conditional heteroskedasticity can be used as a leading indicator since variability tends to rise before a transition, and the part of a time series close to an approaching shift will show up as a cluster of high variability. On the other hand, the part of the time series that is not near the shift will show up as a cluster of low variability [34].

The foundation of conditional heteroskedasticity is the Lagrange multiplier test [22], which is computed by first taking the residuals from a model that has been fitted to the time series. Usually, an autoregressive model is selected according to a measure of the relative quality of fit (e.g. the Akaike Information Criterion (AIC)). After being squared, the residuals are regressed on themselves, one-time step lag included. To determine the significance of the regression R^2, the Lagrange multiplier test compares its coefficient of determination with a χ^2 distribution with one degree of freedom. The χ^2 can be directly compared to the R^2 value by dividing it by the sample size.

2.3 Model based indicators

In order to quantify changes in the time series, model-based indicators attempt to fit the data to a model based on the general structure of (1). Among the several model-based indicators are the time-varying AR(p), threshold AR(p), non-parametric drift-diffusion-jump models (DDJ), and others [3]. This study will employ the DDJ model, which can simulate a range of nonlinear processes without requiring the specification of an explicit equation.

2.3.1 The Drift–Diffusion–Jump (DDJ) model

In the DDJ model, drift is the speed at which local changes occur over time. Diffusion, on the other hand, records the comparatively small and constant shocks that happen at every distinct time step. Conversely, jumps are large, erratic shocks that occasionally cause the system to malfunction. Combining the impacts of diffusion and jumps generates the total variance, which provides a full assessment of variability. Therefore, the statistics for drift, diffusion, and leap act as early warning signs of changes in a time series [34].

Let x represent the dynamics of daily reported COVID-19 cases in South Africa over time steps of length $\triangle t$ indexed by $t \in \mathbb{N}$ [28]. The DDJ model is defined by [29] as a surrogate for the unknown true data-generating process and the model is an extension of (1) given in (4).

$$dx_t = f(x_t,\theta_t)dt + g(x_t,\theta_t)dW + dJ_t, \quad (4)$$

where x is the state variable, $f(x_t,\theta_t)$ describes the deterministic part of the system, and $g(x_t,\theta_t)dW$ determines how stochasticity interacts with the state variable, dW is a white noise process, and J_t is a jump process. These jumps are large, one-step, positive, or negative shocks uncorrelated in time. A slow change in the underlying conditions, θ_t, moves slowly towards a threshold where a transition may occur, θ_C (a critical parameter value), relative to the speed of the change in x. The assumption is that x_t is observed at discrete intervals of time, $\triangle t$. All these parameters are unknown; however, the drift, diffusion, and jump statistics can be estimated from the data, which may serve as early warnings of the bifurcation.

The assumption is that high-frequency observations of the system in equation (4) can be approximated by the DDJ model given in equation (5).

$$dx_t = \mu(x_{t-},\theta_t)dt + \sigma_D(x_{t-},\theta_t)dW + d\left(\sum_{n=1}^{N_t} Z_n\right) \quad (5)$$

In equation (5), the drift, diffusion, and jump functions track the slow changes in θ_t. The drift function $\mu(x_{t-},\theta_t)$ measures the instantaneous deterministic change in the time series. The diffusion function $\sigma_D(x_{t-},\theta_t)$ measures the standard deviation of the relatively small shocks that occur at each time step. The last part of equation (5) represents jumps, that is, relatively large shocks that occur intermittently and are characterised by $\sigma_Z(\theta_t)$, where $Z_n \sim N(0,\theta_t)$, and the probability of a jump arriving in a small time increment dt is $\lambda(x_t,\theta_t)dt$. The drift, diffusion, and jump functions are estimated using nonparametric regression.

For n observations at time step $\triangle t$ over interval period of time $H(H = n \triangle t)$, the parameters of equation (5) are estimated as follows; compute the kernel-weighted moment estimates for $j = 1,2,4,6$:

$$\hat{M}_n^j(x) = \frac{n\sum_{i=1}^{n-1} K[(x_{i\triangle t} - x)/h](x_{(i+1)\triangle t} - x_{i\triangle t})^j}{H\sum_{t=1}^{n-1} K[(x_{i\triangle t} - x)/h]} \quad (6)$$

In this study we used the Gaussian kernel with a bandwidth h_j and the Loess kernel of bandwidth h_j. The subscript j indicates that, in some cases, it is useful to estimate each moment with a different bandwidth. The Gaussian kernel function is given in equation (7).

$$K[(x_{i\triangle t} - x)/h] = \frac{1}{\sqrt{2\pi h_j^2}}\exp\left(-\frac{(x_{i\triangle t} - x)^2}{2h_j^2}\right) \quad (7)$$

We estimate the drift function at each x as ([30]:

$$\hat{\mu}(x,\theta_t) \approx \hat{M}_T^1(x), \quad (8)$$

where the subscript T refers to the entire sample of n time steps of length $\triangle t$. The local total variance at each x is estimated in expression (9).

$$\hat{\sigma}_T^2(x,\theta_t) \approx \hat{M}_T^2(x), \quad (9)$$

The local jump variance function at each x is estimated using equation (10).

$$\hat{\sigma}_Z^2(x,\theta_t) \approx \frac{1}{5n}\sum_{i=1}^{n}\frac{\hat{M}_i^6(x_{i\Delta t})}{\hat{M}_i^4(x_{i\Delta t})} \quad (10)$$

The ratio of the kernel estimators of the sixth moment to the fourth moment is averaged over the time points to smooth out the variability [30], and it should be constant under the null hypothesis of constant jump magnitude. The jump frequency function at each x is estimated as:

$$\hat{\lambda}(x,\theta_t) \approx \frac{\hat{M}_i^6(x_{i\Delta t})}{3[\hat{\sigma}_Z^2(x,\theta_t)]} \quad (11) \qquad (11)$$

Finally, the difference in equation (12) estimates the diffusion variance function.

$$\hat{\sigma}_D^2(x,\theta_t) \approx \hat{\sigma}_T^2(x,\theta_t) - \hat{\lambda}(x,\theta_t)\hat{\sigma}_Z^2(x,\theta_t) \quad (12)$$

2.4 Sensitivity analysis

To verify the reliability of our findings, we performed sensitivity analysis [5]. Two factors, the rolling window size and the Gaussian filtering bandwidth, would influence the outcomes of the two computed early-warning signals, according to [5]. As a result, we performed sensitivity analysis using various rolling window size and bandwidth combinations. With increments of 1% of the data length, we selected rolling window sizes ranging from 25% to 50% of the time series data's total length. With a 20% data length increment, the bandwidth of the Gaussian filtering ranged from 20% to 90% of the time series data length. The ordinal link between the estimated rolling window statistics for each combination of the rolling window size and the Gaussian filtering bandwidth was measured using the Kendall τ rank correlation coefficient. Thus, the patterns of autocorrelation at lag-1 and standard deviation were assessed using the Kendall τ rank correlation coefficient. An improving trend of the computed measures is shown by a positive Kendall τ rank correlation coefficient.

3 Empirical results and discussion

This section presents the study's empirical results, focusing on the key findings derived from the data analysis. The results of the study's objectives and research questions are discussed, providing insights into the observed patterns and trends. We also explore how these findings compare with previous research in the field, highlighting their implications, limitations and potential areas for further investigation.

3.1 Exploratory data analysis

Exploratory Data Analysis (EDA) is a critical initial step in data analysis, providing a comprehensive overview of the dataset's underlying structure. This section uses various statistical tools and visualisation techniques to summarise key characteristics, detect patterns, spot anomalies, and test our initial hypotheses.

3.1.1 Summary statistics and graphical summary

A time series plot of the confirmed daily cases of COVID-19 for the period 6 March 2020 to 27 November 2022 is given in Figure 1. Visual inspection of the data shows that there are five waves. Following the work of O'Brien and Clements [11], who used sequential EWS assessment, this study split the data into five waves (See Figure 1).

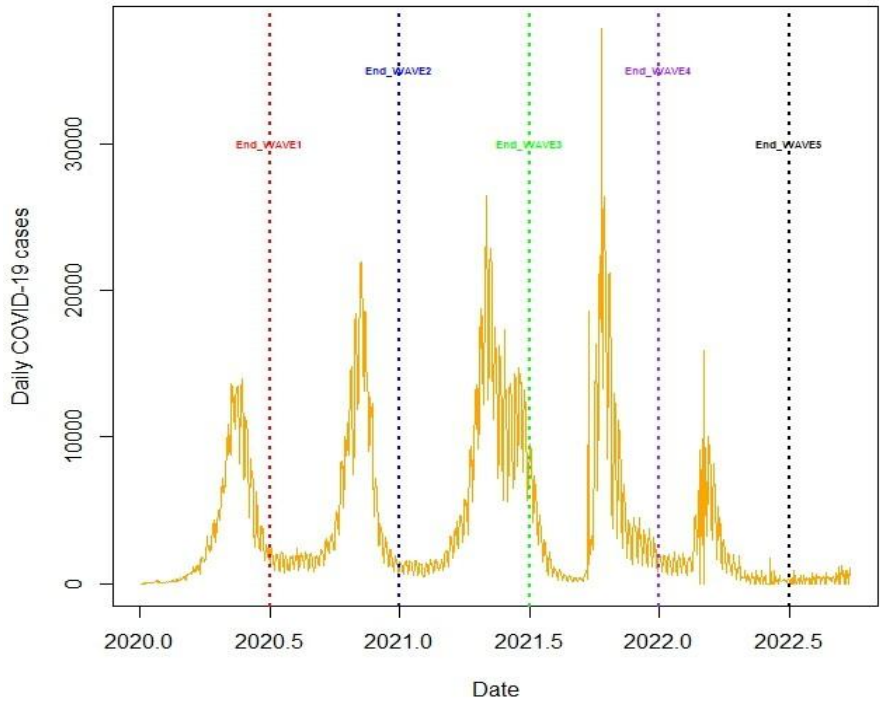

Figure 1: COVID-19 waves for South Africa.

Figure 1 shows that the five waves for the number of COVID-19 cases in South Africa have different peaks. The first wave had a maximum number of daily recorded cases of 13,944. The second phase had a maximum of 21,980 daily cases. In the third wave, the maximum number of daily cases was 26,485. The fourth phase had the highest maximum number of daily cases, equal to 37,875, and lastly, wave 5 had a maximum of 15,927 daily recorded cases. Despite having different peaks, the lengths of these waves are not the same. This makes it difficult to use the seasonal autoregressive moving average models.

Table 1 summarises the important statistics calculated on the five datasets for the five waves. The positive high skewness suggests that the data are right-skewed. In all cases, the kurtosis is greater than three, showing that the distribution for each data set has leptokurtic characteristics, meaning it has heavier tails and a sharper peak compared to a normal distribution. This implies that the data are more prone to producing extreme values, which occur more frequently than in a normal distribution. This suggests that heavy-tailed distributions best describe the data.

Table 1: Summary statistics for the data for the five waves.

Statistic	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
Minimum	0	725	437	106	0
Q1	168.8	1752	1422	473.5	1174
Median	1714.5	2337	3202	2409	1877
Mean	3483.6	5338	6024	5899	2636
Q3	5798.5	8161	9913	9139	3226
Maximum	13944	21980	26485	37875	15927
Skewness	1.149	1.414	1.148	1.579	2.298
Kurtosis	3.074	4.006	3.475	5.194	10.247

In Figure 2, we further investigate the distribution and skewness of each of the five waves by fitting the Box and Whiskers plots. The upper whiskers for each wave are longer than the lower ones. In addition, the median positions are closer to the bottom of the boxes. This confirms that the data is non-normal and positively skewed. Waves 2, 3,4, and 5 show the presence of outliers in the data set. These represent the extreme cases.

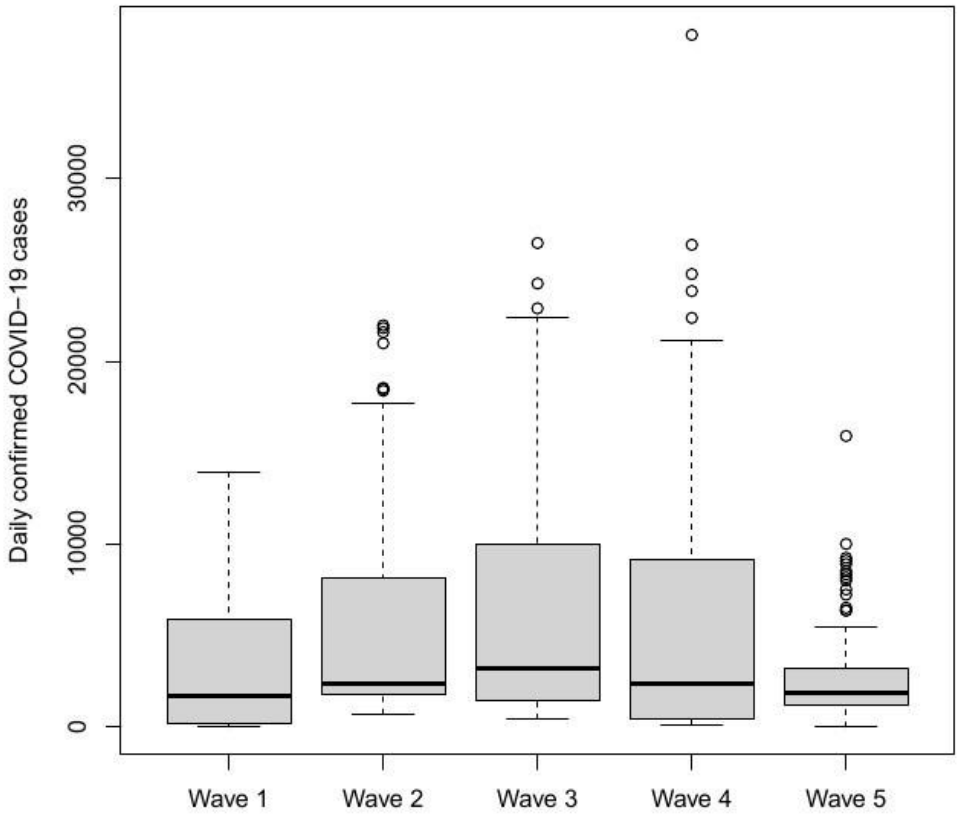

Figure 2: Box plots for the five waves.

Figure 3 shows box plots of the monthly distribution of confirmed COVID-19 daily cases from 2020 to 2022. It can be seen that the incidences of COVID-19 are high during the winter and summer seasons, and its values are low during the autumn and spring seasons. High COVID-19 cases are also experienced during the festive season in December, and the effect is also experienced in January.

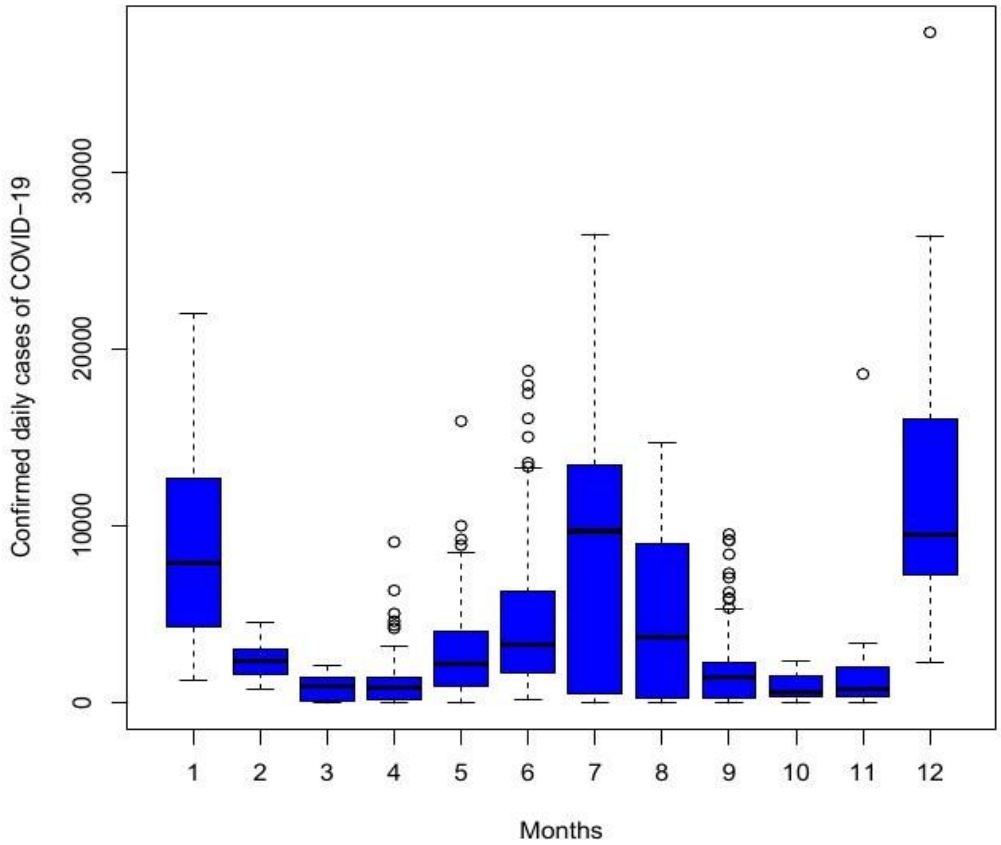

Figure 3: Distribution of monthly confirmed COVID-19 cases for the sampling period 6 March 2020 to 27 November 2022.

3.2 Metric Based Indicators

Results of the metric based indicators discussed in Section 2.2 are presented in this section.

3.2.1 Generic early warning signal with Gaussian filtering

We estimated each wave's autocorrelation, return rate, standard deviation, and skewness. For each statistic, trends are estimated using the nonparametric Kendall tau correlation. Kendall's tau, like Spearman's correlation ([38]), measures the strength of the relationships between the estimated metrics with time. However, Kendall's tau statistic has the advantage that its distribution has better statistical properties. This was done after detrending the dataset of each wave using Gaussian smoothing with a bandwidth size of 5%.

Figure 4 presents the generic early warning signals for wave 1. Autocorrelation at lag-1 (ACF(1)) increased to the transition with a moderate trend as indicated by Kendall's τ (rank correlation) both for the original series (τ = 0.455) and the residual (after detrending) dataset (τ=0.461). The standard deviation for the original series increased. However, only until transitions across the two attractors persisted (approximately up to day 150 of the first wave), and after that, it started to decrease. The Kendall's tau for the standard deviation with respect to the time variable is 0.924, which shows a strong increase up to the transition. Skewness decreased for both the original series (τ = −0.141) and the residuals after detrending (τ = −0.494), although somewhat irregular. A sharp increase in the skewness is observed from day 80 of the first wave and started to decrease around day 90. This shows that after day 80, the number of daily COVID-19 cases was transitioning toward a higher state of disease spread than the previous state. After day 90 of wave 1, the decrease in skewness signals a transition towards a state of low spread of the disease. A rise in the skewness at day 80 also indicates that the COVID-19 pandemic was more likely to reach extreme values close to the transition. However, according to [1], with the irregular pattern in the skewness and kurtosis, these metrics can not successfully warn us of an upcoming transition. The return rate for the original series decreased moderately (τ = −0.359), while the density ratio for the residuals after detrending increased moderately (τ = 0.461). The behaviours of all the residuals were according to our expectations for systems approaching a critical transition. A decline in the spread of COVID-19 after day 90 was mainly due to stringent measures imposed by the South African government. During wave one, the COVID-19 vaccine was not yet administered.

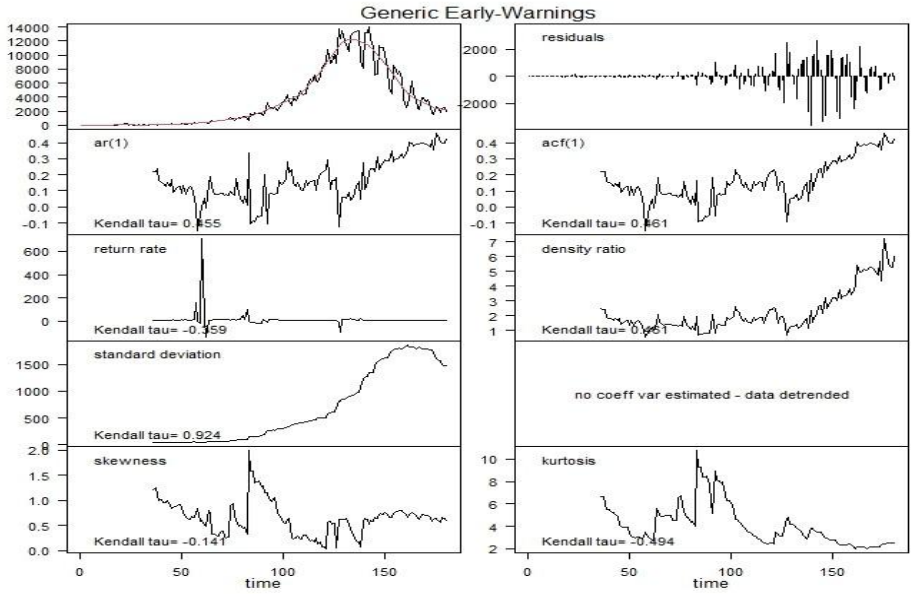

Figure 4: Generic early warning signals for wave 1: the Gaussian detrending, window size =20, bandwidth=5.

In Figure 5, we present the generic early warning signals for wave 2. Just like in wave1, a moderate increase in the ACF(1), as indicated by a positive Kendall's τ = 0.441 for the detrended series and Kendall's τ = 0.425 for the residuals. The standard deviation for the detrended series increased at a slower rate compared to that of wave 1. However, the rate of decrease in the skewness for the detrended series is faster in wave two than in wave 1. The skewness for wave 2 is more shakey (flickering) than wave 1. The highest skewness value is around day 95 of the second wave, and around this day, the standard deviation started to increase as well as the return rate. This signals the state change observed after day 100 of the second wave, as indicated by the original time series.

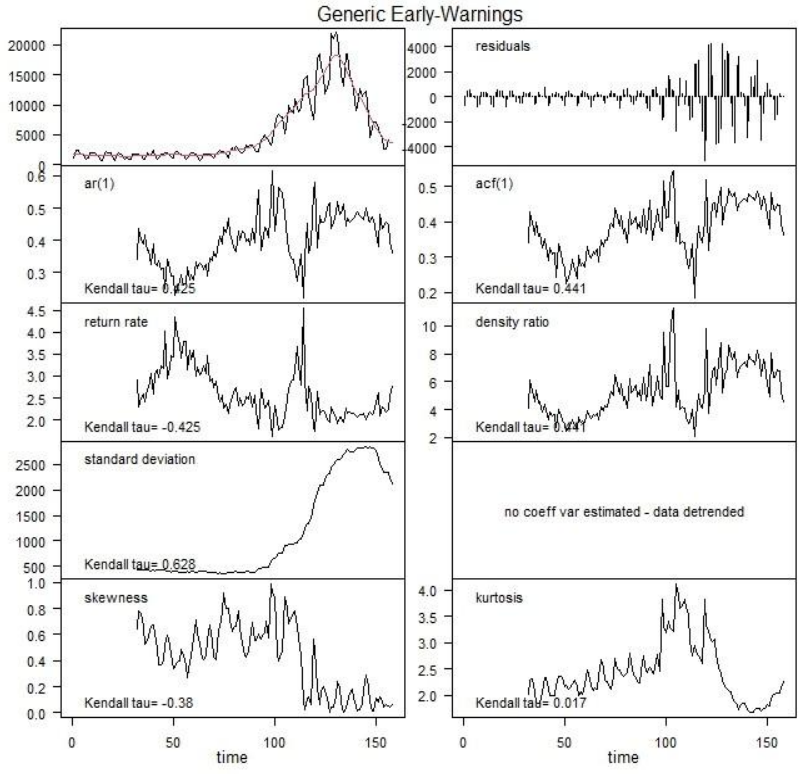

Figure 5: Generic early warning signals for wave 2: the Gaussian detrending, window size =20, bandwidth=5.

Wave 3 was one of the longest. It took close to 300 days. The generic early warnings for wave three are presented in Figure 6. For this wave, skewness, return rate, and AR(1) changes were close to 0. The standard deviation is the only metric that increased with time, as indicated by Kendall's $\tau = 0.59$. However, the metrics plots show shifts between day 100 and day 150 of wave 3. The standard deviation, skewness, and return rates for the detrended series increased from day 100. From day 125 to day 155, the kurtosis for the residuals was above 3. This shows that between these days, the pandemic was at its extremes. Based on the kurtosis and skewness, the number of daily COVID-19 cases changed from a lower transition state to a higher transition state on day 125. This state reached its maximum around day 130 and started transitioning towards fewer COVID-19 cases. These results were an early warning signal for the peak in COVID-19 cases observed around day 140 of the third wave. Generic early warning signals for waves four and five are presented in the supplementary material which can be accessed on https://github.com/csigauke

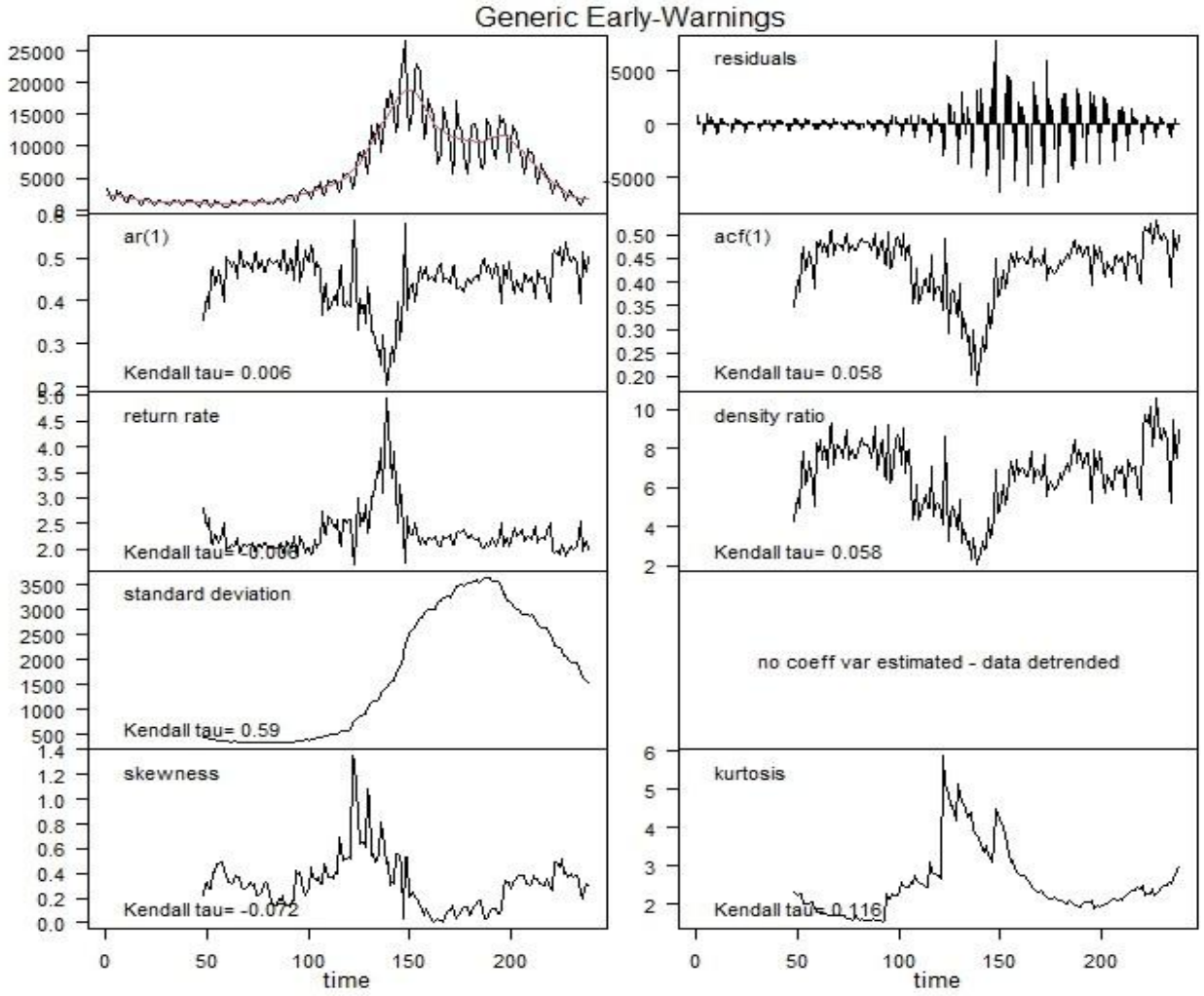

Figure 6: Generic early warning signals for wave 3: the Gaussian detrending, window size =20, bandwidth=5.

3.2.2 BDS Test at $\epsilon = 0.5, 0.75, and\ 1)$

We present the BDS method applied to a simulated time series in which a critical transition is approaching.

We removed the underlying linear structure by (a). first-differencing, (b). fitting an AR(1) and (c). fitting a GARCH(0,1)) to the entire time series after log-transforming. The remaining residuals were used to estimate the BDS statistic for embedding dimensions (ED) 2 and 3, and $\epsilon = 0.5, 0.75, and\ 1)$ times the observed standard deviation of the time series, as suggested by [3].

The results are presented in Table 2. The significance of the BDS statistics was calculated using 200 bootstrap iterations. The results show the significance of the BDS tests at 5% level. Thus, the BDS statistic provides strong evidence of nonlinearity.

Table 2: BDS statistic estimated on the daily reported COVID-19 dataset for South Africa with measurement error.

$*pv < 0.001$	ED	1st diff/detrend			AR(1) residuals			GARCH(0,1) resid		
	ε	0.5	0.75	1	0.5	0.75	1	0.5	0.75	1
Wave1	2	7.36*	6.15*	4.48*	8.63 *	4.74*	2.39*	91.5*	139.9*	113.8*
	3	11.95*	9.05*	6.35*	13.88*	7.59*	4.67*	153.78*	201.18*	145.10*
Wave2	2	5.86*	5.83*	5.07*	6.76*	6.15*	4.38*	30.47*	40.68*	42.1*
	3	8.35*	7.88*	6.91*	9.92*	8.58*	6.24*	42.62*	53.60*	51.00*
Wave3	2	12.68*	10.91*	8.65*	10.31*	10.07*	9.37 *	80.85*	83.19*	61.42 *
	3	16.52*	13.48 *	10.65*	12.88 *	11.91*	10.69*	130.67*	115.16*	76.34*
Wave4	2	5.71*	4.78*	5.61*	7.63 *	6.07 *	6.16 *	32.55 *	24.33*	18.50*
	3	8.70*	6.44*	6.18 *	8.83 *	6.63 *	5.81*	53.25*	29.8*	21.05*
Wave5	2	7.44*	7.19*	6.36*	9.13 *	9.59 *	9.42 *	8.59*	7.48*	6.19 *
	3	9.24*	8.85*	7.99*	9.40*	9.56*	9.22*	8.51*	6.74*	5.73*

Results from Table 2 give enough evidence to conclude that there is a remaining structure in the time series, including hidden nonlinearity, hidden nonstationarity or other type of structure missed by detrending or model fitting.

The underlying structure of the series is removed by first differencing, fitting an AR(1) or fitting a GARCH(0,1) to the entire dataset after log-transforming. The plots for the BDS tests are given in the supplementary material https://github.com/csigauke

3.2.3 Conditional heteroskedasticity

Conditional heteroskedasticity (CH) was applied to the residuals of the best fit AR(p) on the original time series based on the AIC. CH was estimated in rolling windows of 50% of the size of the time series, and the results are shown in Figure 7.

In Figure 7, we present plots of the estimated conditional heteroskedasticity for each of the five waves. Values above the dashed red line are significant at the 10% level. For wave one, conditional heteroskedasticity is significant around day 80, corresponding to the transition detected by the generic early warning signals. Another significant CH for wave one is noted between day 140 and day 165. During wave two, a significant CH is noted around day 100 and between days 110 and 120 of the second wave. The third wave showed a significant CH from day 125 and was consistently significant up to the last part of the wave. The generic early warning also picked shifts in transitions around these periods for wave 3. These findings are consistent with the findings from the generic early warning signals.

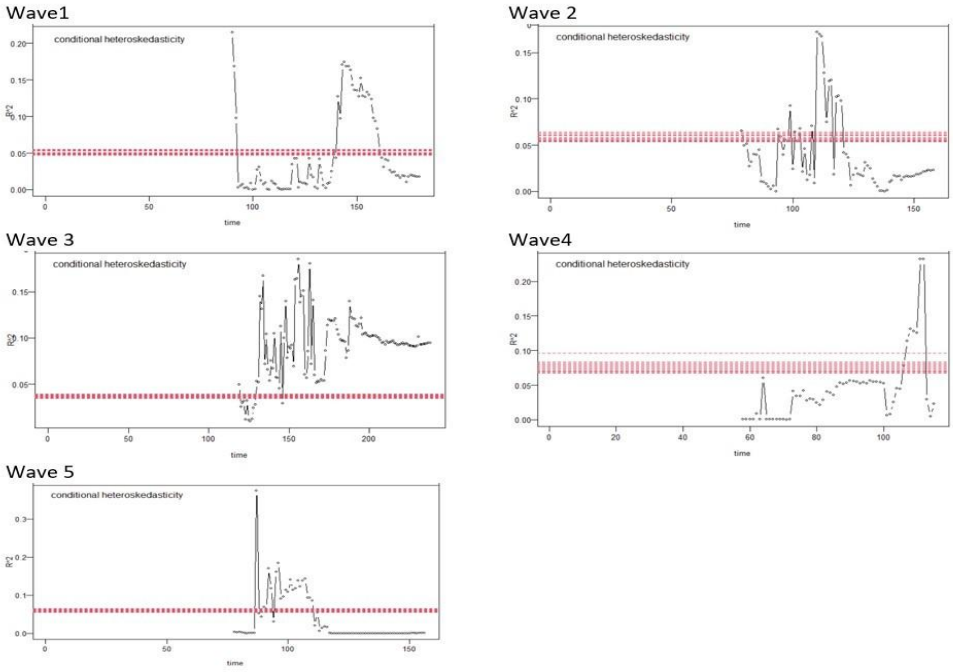

Figure 7: Conditional heteroskedasticity estimated on the daily COVID-19 cases in South Africa for the five waves.

3.3 Model-based Indicators

3.3.1 Nonparametric drift-diffusion-jump models

We often need to learn the underlying processes that generate the time series we are analysing for early warnings. After log transformation, the nonparametric drift-diffusion-jump (DDJ) model was applied to the five waves.

In Figure 8, we present the DDJ method applied to wave 1 of the COVID-19 pandemic, in which a critical transition is observed towards the end. The plots show a decrease in conditional and total variance, jump intensity in the time series, and a decrease followed by a slow increase in the diffusion term. The trends were noisy, but they became very clear after smoothing). For log-transformed values for the conditional variance and the jump density signals the upcoming transition between 4 and 8.

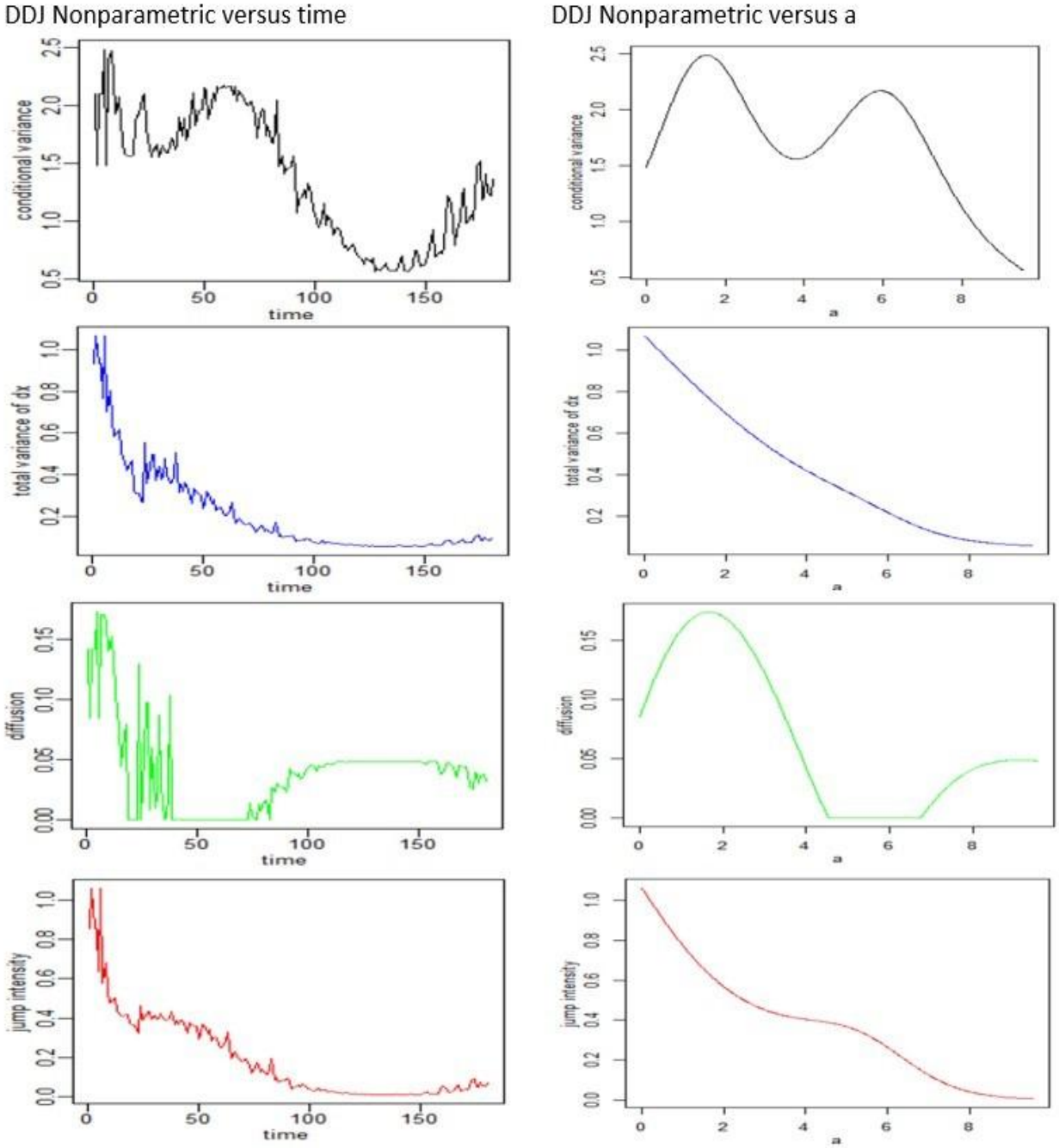

Figure 8: Nonparametric drift-diffusion-jump metrics in the COVID-19 dataset. (A, E) Conditional variance versus time and daily COVID-19 cases, respectively. (B, F) Total variance versus time and daily COVID-19 cases, respectively. (C, G) Diffusion versus time *a* for wave1.

Next, we present the nonparametric drift-diffusion-jump plots for wave 2 in Figure 9. We found an increase in conditional variance from day 60 to day 100 of the second wave. The diffusion term remained constant until day 100 when it started to increase to day 130 of wave 2. The trends were noisy, but they became very clear due to smoothing. For log-transformed values

between 8.0 and 9.0, the indicators started to signal the upcoming transition.

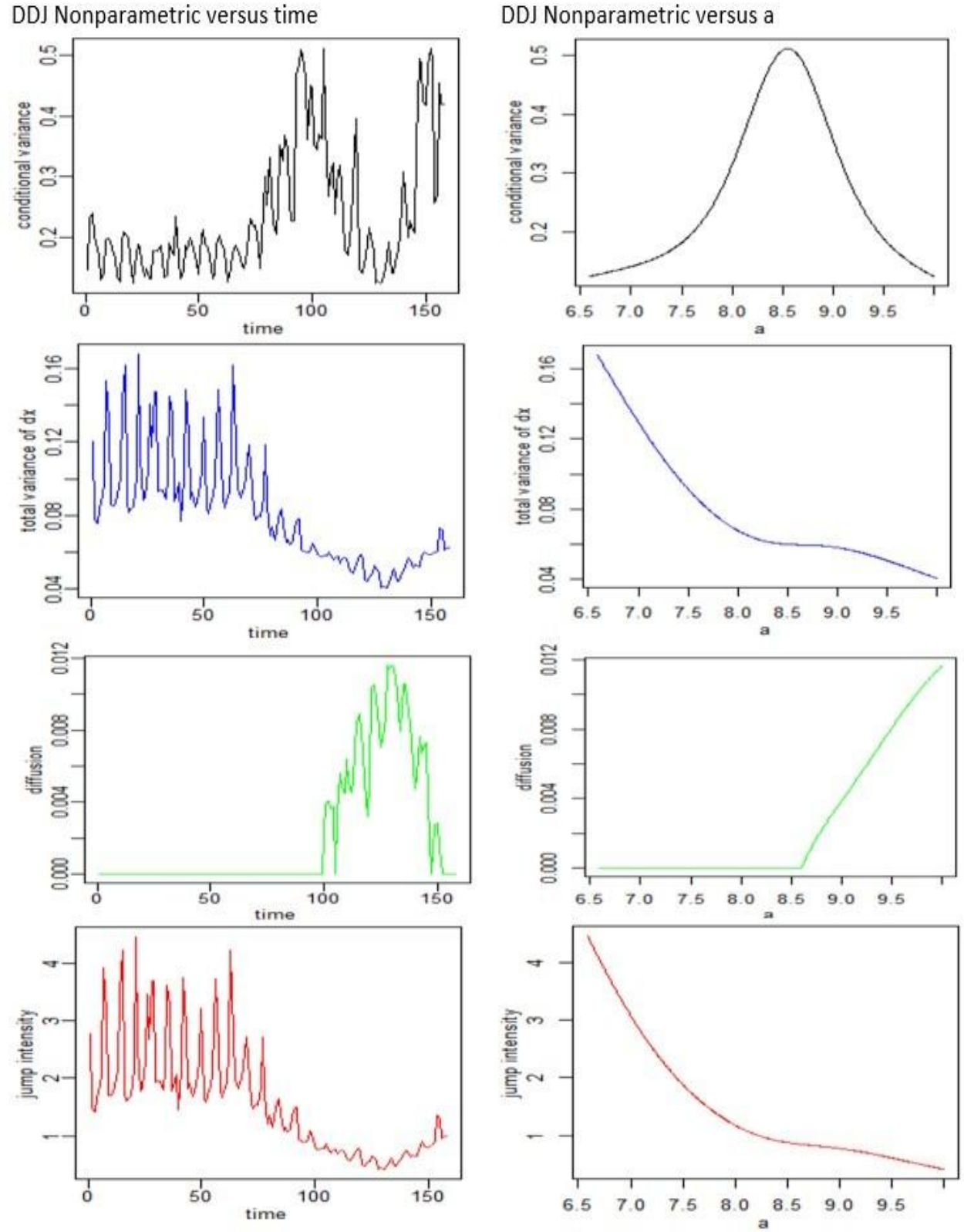

Figure 9: Nonparametric drift-diffusion-jump metrics in the COVID-19 dataset. (A, E) Conditional variance versus time and daily COVID-19 cases, respectively. (B, F) Total variance versus time and daily COVID-19 cases, respectively. (C, G) Diffusion versus time *a* for wave2.

The results for the DDJ nonparametric model for wave three are presented in Figure 10. Wave 3 is the longest. The conditional variance, total variance, diffusion and jump density were noisy before detrending. However, after detrending, the trends became clearer. The conditional variance is increased from day 90 to day 120 and from day 150 to day 210. Thus, the conditional

variance signals the upcoming transition within wave three and then into wave 4.

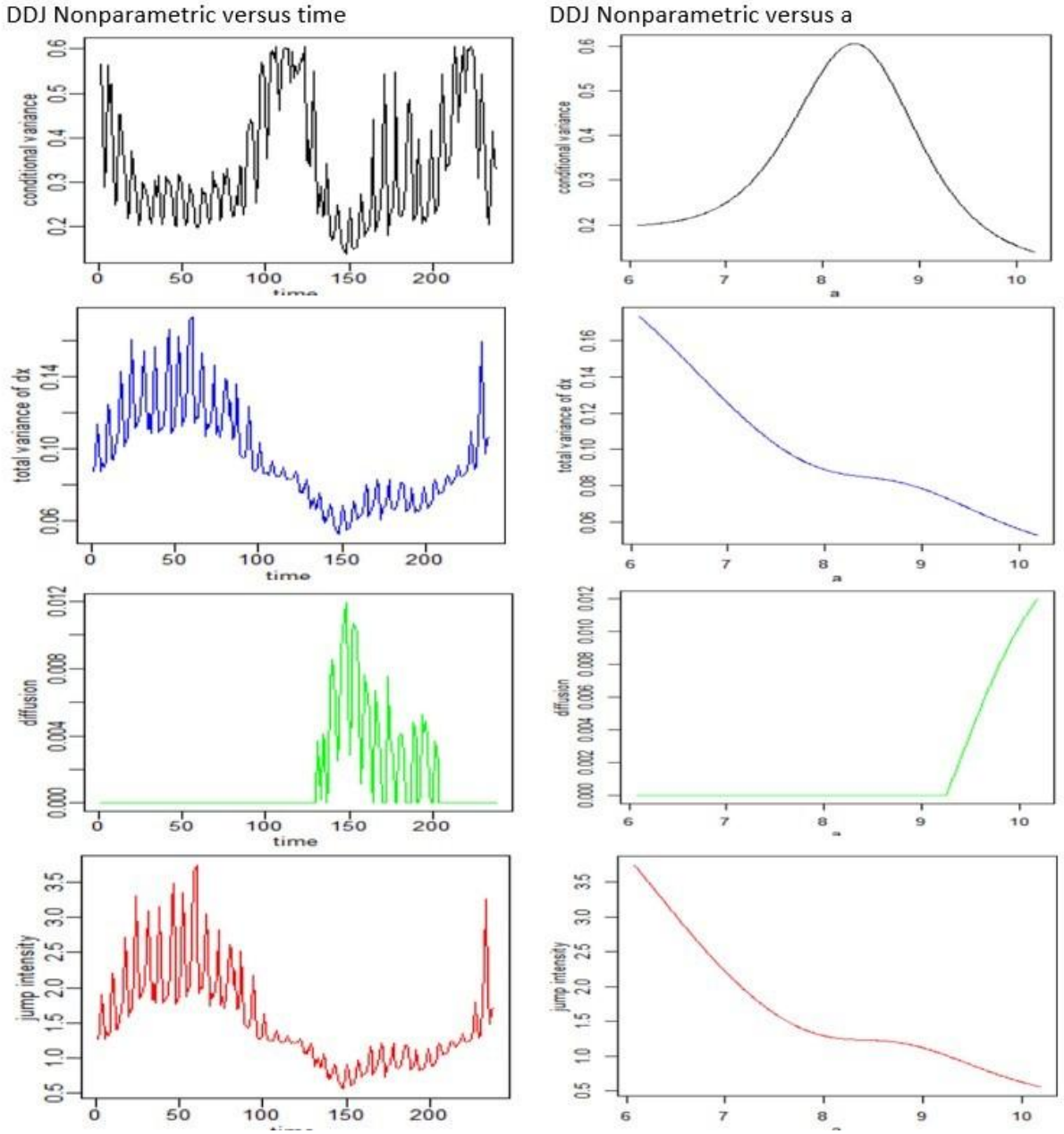

Figure 10: Nonparametric drift-diffusion-jump metrics in the COVID-19 dataset. (A, E) Conditional variance versus time and daily COVID-19 cases, respectively. (B, F) Total variance versus time and daily COVID-19 cases, respectively. (C, G) Diffusion versus time *a* for wave3.

Wave 4 and 5 plots are presented in the supplementary material https://github.com/csigauke.

Wave 5 has nothing interesting to offer since this was the last wave before the pandemic started approaching extinction. This was also due to an increased uptake of vaccines.

3.3.2 Quick Detection Analysis of Generic Early Warning Signals

We present plots for the quick detection analysis of Generic early warning Signals for each of the five waves. The first plot contains the original data, the detrending/filtering applied and the residuals, autocorrelation, and variance. For each statistic, trends are estimated using the nonparametric Kendall tau correlation. The second plot returns a histogram of the distributions of the Kendall trend statistic for autocorrelation and variance estimated on the surrogate data. Vertical lines represent the significance level, whereas the black dots are the trends in the time series. The third plot is the reconstructed potential landscape in 2 dimensions. We used a significant level of 5%, a detection threshold of 0.002, cut-off = 0.05, a window size of 50 for the time series length in increments of 10 points, and null bandwidth gave the ARMA(4,4) model as the best for all waves based on the AICs. Red vertical lines indicate the significance level and the black dot is the actual Kendall trend estimated on the original residual dataset for rolling window size and Gaussian filtering. As presented in Figure 11 below, a Kendall τ

estimate of -0.025 for the autocorrelation shows that autocorrelation at lag-1 decreased rapidly. A Kendall τ of 0.927 shows that the variance increased rapidly regardless of the bandwidth choice and the size of the rolling window. For wave 1, the significance autocorrelation is $p = 0.2$ and the significance variance is $p = 0.1$ and these were estimated from 10 surrogate ARMA time series. Plots for wave two and wave three are presented in Figure 12 and 13, respectively. The Kendall τ of the variances for waves one, two and three are both above 0.9. Waves four and five results are presented in the supplementary material https://github.com/csigauke.

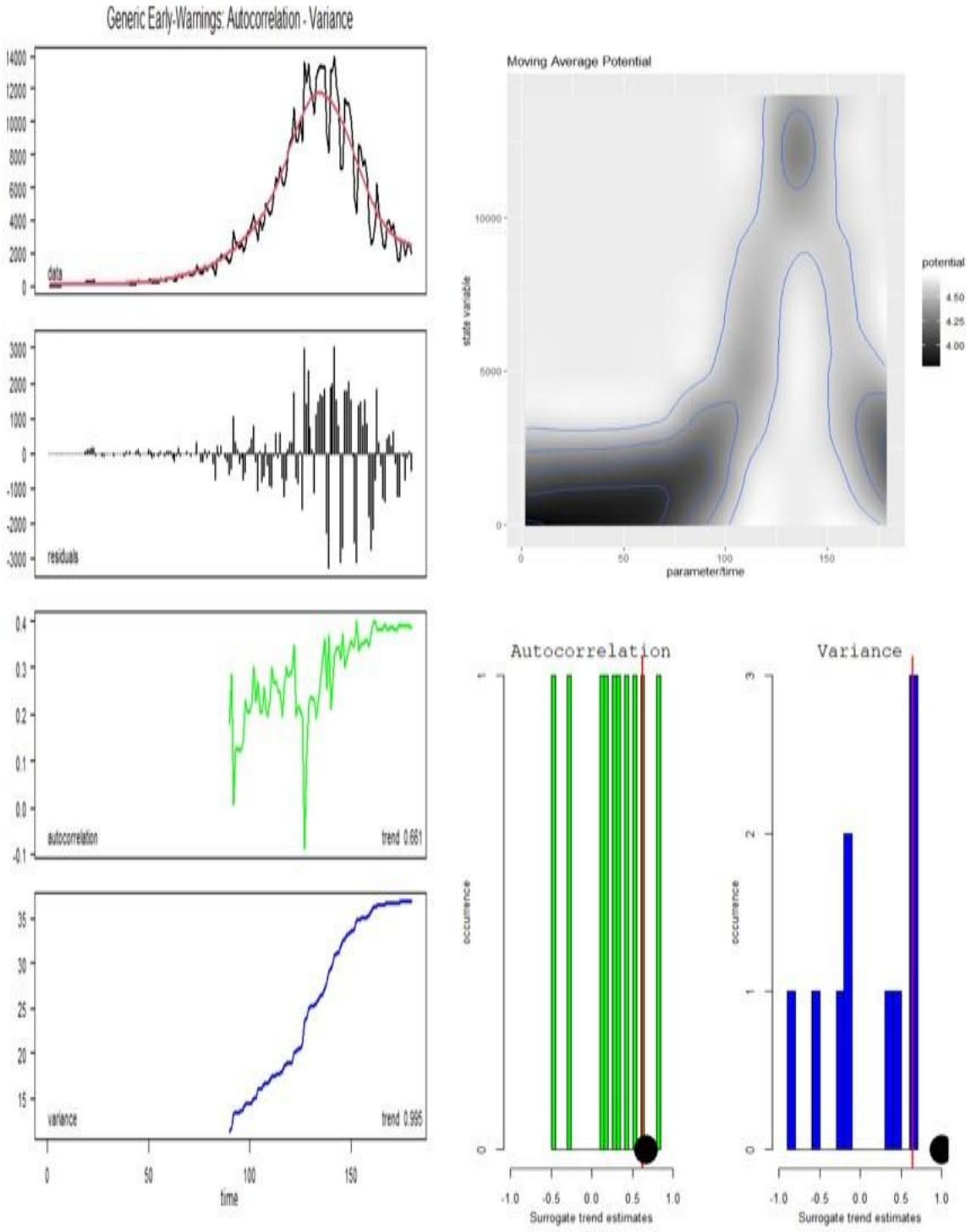

Figure 11: plots for the quick detection analysis of Generic early warning Signals. The first plot contains the wave one series, the detrending/filtering applied and the residuals, autocorrelation and variance.

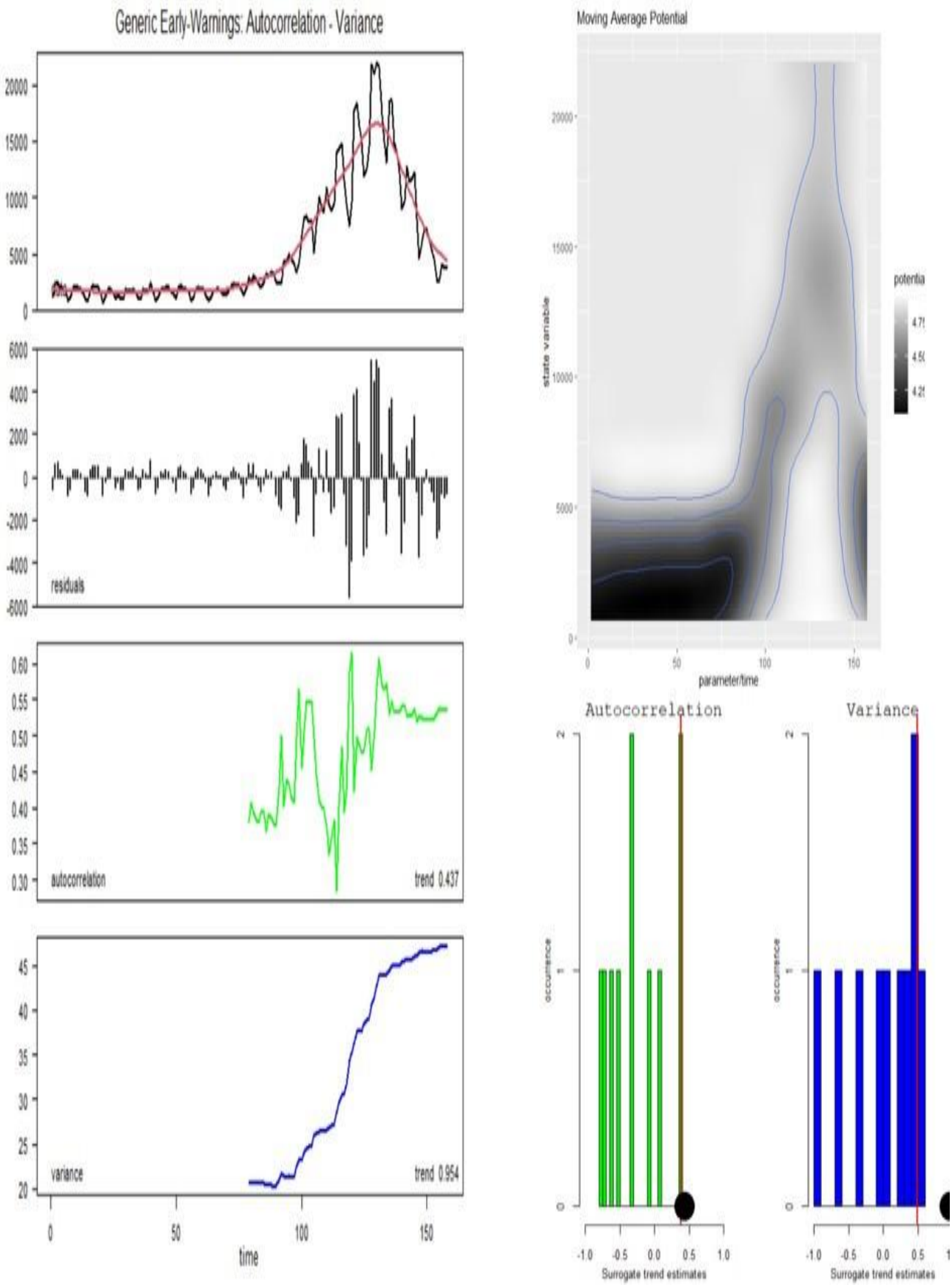

Figure 12: plots for the quick detection analysis of Generic early warning Signals. The first plot contains the wave two series, the detrending/filtering applied and the residuals, autocorrelation and variance.

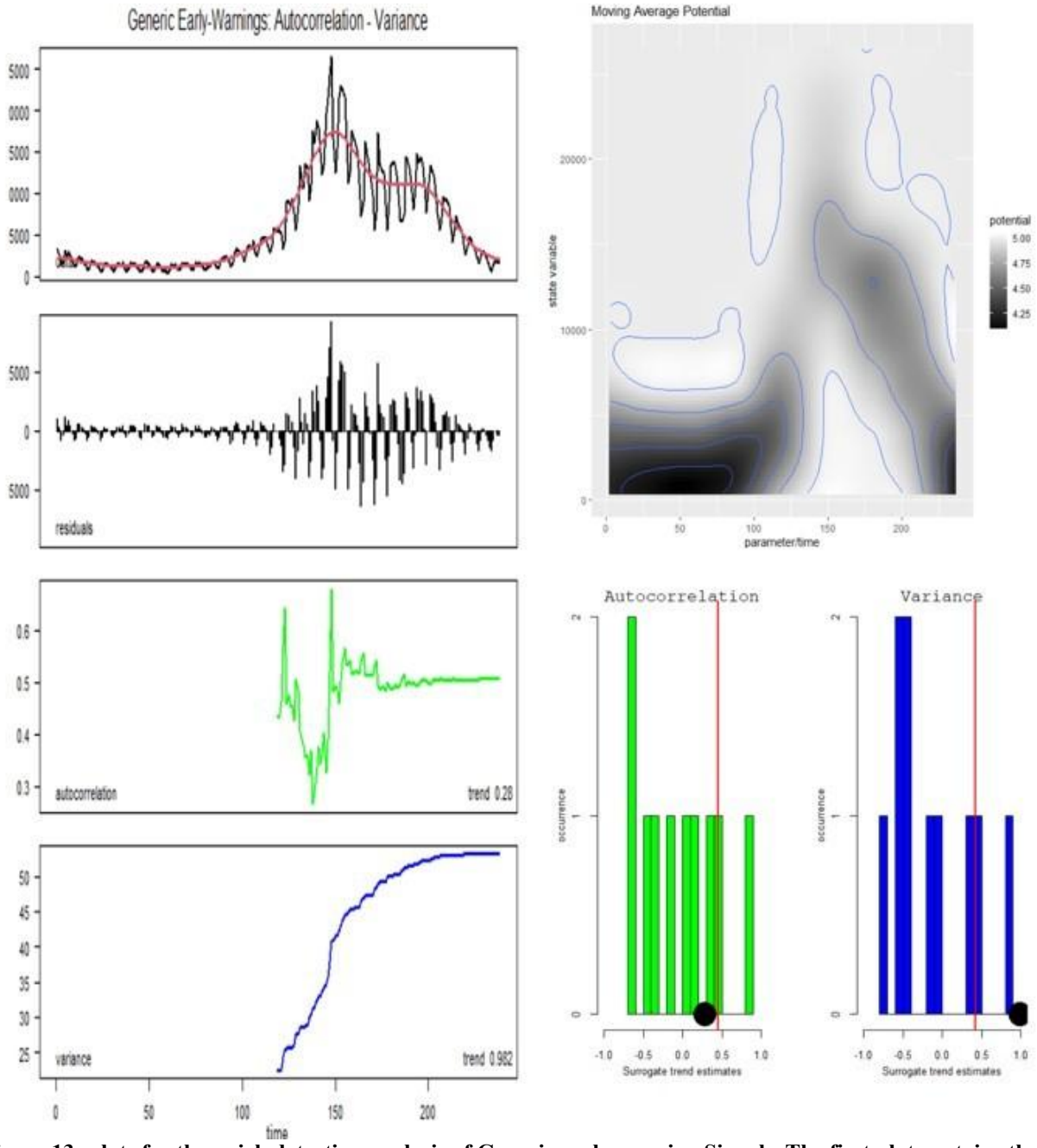

Figure 13: plots for the quick detection analysis of Generic early warning Signals. The first plot contains the wave three series, the detrending/filtering applied and the residuals, autocorrelation and variance.

3.4 Sensitivity analysis

In Figures 14, 15, and 16 we present contour plots to illustrate the results of sensitive analysis for wave 1 wave 2, and wave 3, respectively: Kendall τ values of different combinations of the rolling window size and the bandwidth of Gaussian filtering. The histograms that illustrate the distribution of Kendall τ estimates of the sensitivity analysis are presented.

We can observe from Figure 14, 15, and 16 that the three waves have clear early-warning signals for an approaching epidemic. This is shown by the standard deviation which depicts an increasing trend. The results show similar patterns for wave 4 and wave 5 (supplementary materials) https: //github.com/csigauke. The early-warning signals for critical high do not only demonstrated that the number of reported COVID-19 cases started to show a different pattern, but also indicated that the system of reported COVID-19 cases were approaching a phase transition point (high level of COVID19 cases). Also, the sensitivity analysis demonstrated the robustness of the results, as Kendall τ rank correlation coefficients stayed positive for the examined combinations of rolling window sizes and bandwidths of Gaussian filtering. In all waves, significant trends were estimated for the standard deviation for all the rolling windows ($p-$value < 0.0001). The standard deviations for wave 2 and wave 3 were also significant for all the rolling windows except for wave 3, which was significant for a rolling window below 90. Sensitivity analysis plots for waves 4, and 5 are presented in the supplementary material https://github.com/csigauke.

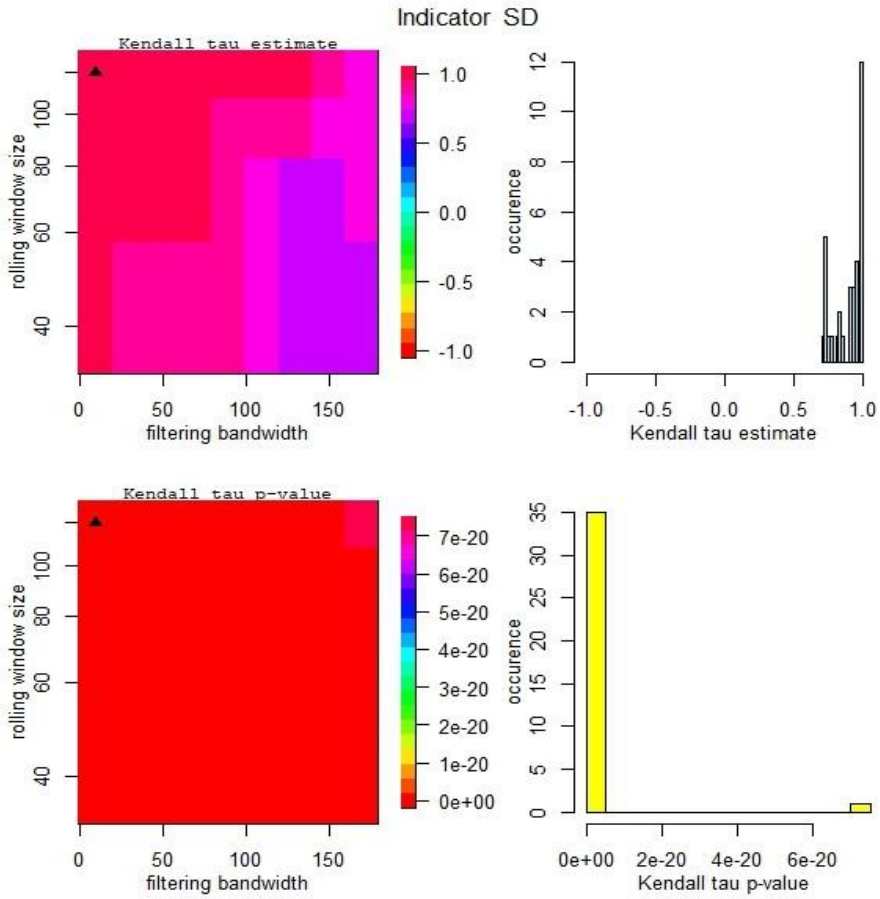

Figure 14: Sensitivity analysis for wave 1. Sensitivity analysis for rolling window metric, standard deviation, for the first wave of the COVID-19 pandemic.

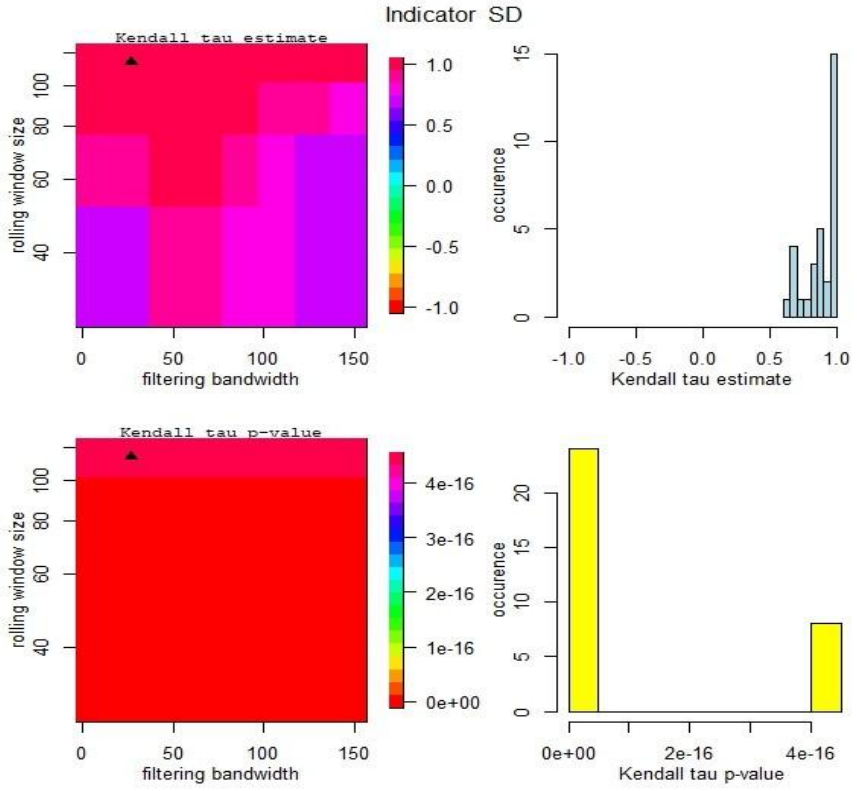

Figure 15: Sensitivity analysis for wave 2. Sensitivity analysis for rolling window metric, standard deviation, for the first wave of the COVID-19 pandemic.

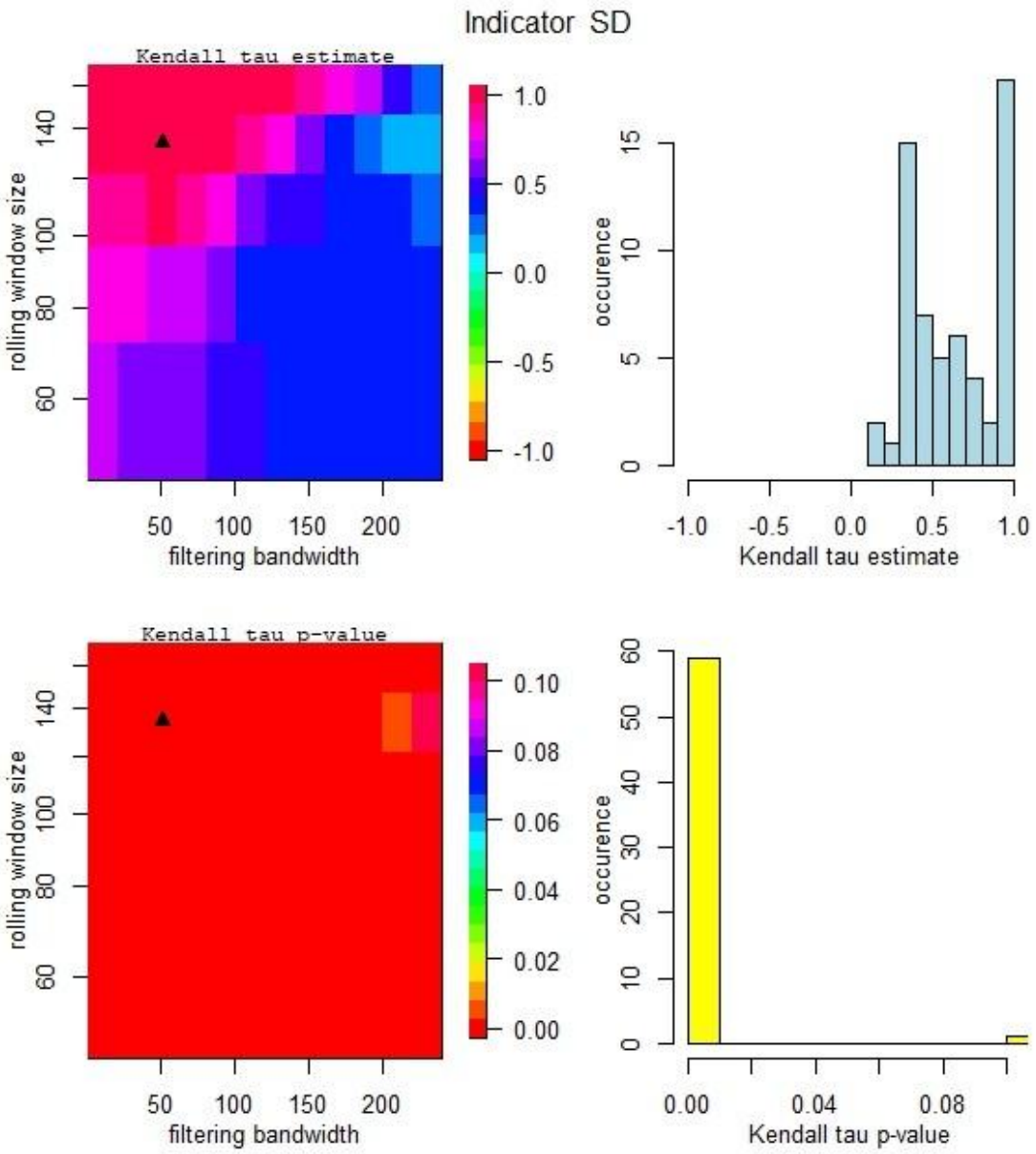

Figure 16: Sensitivity analysis for wave 3. Sensitivity analysis for rolling window metric, standard deviation, for the first wave of the COVID-19 pandemic.

4 Conclusion

This study applied a range of robust frameworks for early warning signals for the critical transitions to daily COVID-19 cases in South Africa. The spread of COVID-19 in South Africa is divided into five waves. For each of the waves, metric- and model-based approaches are used to signal an approaching rise in the number of daily COVID-19 cases. The metric-based approaches used include the return rates, standard deviation, ACF(1), skewness, kurtosis, among others. The drift-diffusion-jump (DDJ) is the model-based approach applied in this study. The DDJ model combines drift (the speed at which local changes occur), diffusion (for recording the comparatively small and constant shocks that happen at every distinct time step), and jumps (large, erratic shocks that occasionally cause systematic malfunction). A combination of these provides a full assessment of variability. The metric-based and model-based methods give similar signals for the upcoming changes in transitions for each wave. A sensitivity analysis was also performed for each wave, showing the significance of the metrics, the standard deviation in particular, across the rolling windows.

Acknowledgements

The authors are grateful to the numerous people who have provided helpful comments on this paper.

Data Availability Statement

Data Availability Statement: The analytic data can be downloaded from https://github.com/ csigauke (accessed on 31 October 2023).

Conflicts of Interest

The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACF	Autocorrelation Function
AIC	Akaike Information Criterion
AR	Autoregressive
ARCH	Autoregressive Conditional Heteroskedasticity
DDJ	Drift Diffusion Jump
EDA	Exploratory Data Analysis
EWS	Early Warning Signal
GARCH	Generalised Autoregressive Conditional Heteroskedasticity

References

[1] Koushik Garain and Partha Sarathi Mandal. Stochastic sensitivity analysis and early warning signals of critical transitions in a tri-stable prey–predator system with noise. Chaos 32, 033115 (2022); doi: 10.1063/5.0074242

[2] Livina, V. N., Kwasniok, F., and Lenton, T. M.: Potential analysis reveals changing number of climate states during the last 60 kyr, Climate of the Past *2010*, **6**, 77–82. https://doi.org/10. 5194/cp-6-77-2010

[3] Dakos V, Carpenter SR, Brock WA, Ellison AM, Guttal V, et al. (2012) Methods for Detecting Early Warnings of Critical Transitions in Time Series Illustrated Using Simulated Ecological Data. PLoS ONE 7(7): e41010. https://doi.org/10.1371/journal.pone.0041010

[4] Q. Li, Z. Tang, N. Coleman and A. Mostafavi, "Detecting Early-Warning Signals in Time Series of Visits to Points of Interest to Examine Population Response to COVID-19 Pandemic," in IEEE Access, vol. 9, pp. 27189-27200, 2021, doi: 10.1109/ACCESS.2021.3058568.

[5] T. M. Lenton, V. N. Livina, V. Dakos, E. H. Van Nes and M. Scheffer, "Early warning of climate tipping points from critical slowing down: Comparing methods to improve robustness", Philos. Trans. Roy. Soc. A Math. Phys. Eng. Sci., vol. 370, no. 1962, pp. 1185-1204, 2012.

[6] Ana Filipa Duarte1 • Am´ılcar Soares1 • Maria Jo˜ao Pereira1 • Andr´e Peralta-Santos3,4 • Pedro Pinto Leite2 • Leonardo Azevedo. A Stochastic Model of an Early Warning System for Detecting Anomalous Incidence Values of COVID-19. Mathematical Geosciences (2024) 56:41–54. https: //doi.org/10.1007/s11004-023-10096-4

[7] Oraya Srimokla1 , Wirichada Pan Ngum2,3, Amnat Khamsiriwatchara4 , Chantana Padungtod5 , Rungrawee Tipmontree5 , Noppon Choosri6 and Sompob Saralamba2*. Early warning systems for malaria outbreaks in Thailand: an anomaly detection approach. Srimokla et al. Malaria Journal (2024) 23:11 https://doi.org/10.1186/s12936-024-04837-x

[8] V. N. Livina1 , F. Kwasniok2 , and T. M. Lenton1. Potential analysis reveals changing number of climate states during the last 60 kyr. Clim. Past, 6, 77–82, 2010. www.clim-past.net/6/77/2010/.

[9] J. Venegas, T. Winkler, G. Musch, M. V. Melo, D. Layfield, N. Tgavalekos, A. Fischman, R. Callahan, G. Bellani, and R. Harris, "Self-organized patchiness in asthma as a prelude to catastrophic shifts," Nature 434, 777 (2005).

[10] O'Brien, D. Performing early warning signal assessments. Available online. https://cran. r-project.org/web/packages/EWSmethods/vignettes /ews_assessments.html Accessed on 22 July 2024.

[11] O'Brien DA, Clements CF. 2021 Early warning signal reliability varies with COVID-19 waves. Biology Letters **2021**, *17* 1–6, 20210487. https://doi.org/10.1098/rsbl.2021.0487

[12] Dablander F, Heesterbeek H, Borsboom D, Drake JM. Overlapping timescales obscure early warning signals of the second COVID-19 wave. Proceedings of the Royal Society B **2022**, *289*, 20211809, 1–11. https://doi.org/10.1098/rspb.2021.1809

[13] Clements, C.F., McCarthy, M.A. & Blanchard, J.L. Early warning signals of recovery in complex systems. Nature Communications **2019**, 10, 1681, 1–9. https://doi.org/10.1038/ s41467-019-09684-y

[14] Weinans, E.; Quax, R.; van Nes, E.H.; van de Leemput, I.A. Evaluating the performance of multivariate indicators of resilience loss. Scientific Reports **2021**, 11, 9148, 1–11. https://doi.org/10.1038/s41598-021-87839-y

[15] Deb, S., Sidheekh S, Clements CF, Krishnan NC, Dutta PS. 2022 Machine learning methods trained on simple models can predict critical transitions in complex natural systems. Royal Society Open Science **2022**, 9, 211475, 1–12. https://doi.org/10.1098/rsos.211475

[16] Bury, T.M.; Sujithc, R.I.; Pavithrand, I.; Scheffere, M.; Lentonf, T.M.; Anandb, M.; Baucha,

C.T. Deep learning for early warning signals of tipping points. Applied Mathematics Ecology **2021**, 118(39) e2106140118, 1–9. https://doi.org/10.1073/pnas.2106140118

[17] Southall, E.; Tildesley, M.J.; Dyson, L. How early can an upcoming critical transition be detected?

MedRxiv preprint **2022** Available online https://doi.org/10.1101/2022.05.27.22275693 Accessed on 25 July 2024.

[18] deYoung B, Barange M, Beaugrand G, Harris R, Perry RI, Scheffer M, Werner F. Regime shifts in marine ecosystems: detection, prediction and management. Trends Ecol Evol. 2008 Jul;23(7):402-9. doi: 10.1016/j.tree.2008.03.008. Epub 2008 May 22. PMID: 18501990.

[19] Seekell DA, Carpenter SR, Pace ML. Conditional heteroscedasticity as a leading indicator of ecological regime shifts. Am Nat. 2011 Oct;178(4):442-51. doi: 10.1086/661898. Epub 2011 Aug 25. PMID: 21956023.

[20] Dessavre, A.G.; Southall, E.; Tildesley, M.J.; Dyson, L. The problem of detrending when analysing potential indicators of disease elimination. Journal of Theoretical Biology **2019**, 481, 183–193. https://doi.org/10.1016/j.jtbi.2019.04.011

[21] Dakos V, Carpenter SR, Brock WA, Ellison AM, Guttal V, Ives AR, K´efi S, Livina V, Seekell DA, van Nes EH, Scheffer M. Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. PLoS One. 2012;7(7):e41010. https://doi.org/10.1371/journal.pone.0041010 Epub 2012 Jul 17. PMID: 22815897; PMCID: PMC3398887.

[22] Furno, M. (2008). Misspecification and estimation effect in the Lagrange multiplier tests for heteroskedasticity. Journal of Statistical Computation and Simulation, 78(4), 299–313. https://doi.org/10.1080/10629360601052374

[23] Brock, W. A., W. Dechert, and J. Scheinkman. (1987). A test for independence based on the correlation dimension. Working paper, University of Winconsin at Madison, University of Houston, and University of Chicago.

[24] Vasilis Dakos, Marten Scheffer, Egbert H. van Nes, Victor Brovkin, Vladimir Petoukhov , and Hermann Held. Slowing down as an early warning signal for abrupt climate change. PNAS, September 23, 2008, vol. 105, no. 38, 14308–14312

[25] Marina Hirota,1 Milena Holmgren,2 * Egbert H. Van Nes,1 Marten Scheffer1. Global Resilience of Tropical Forest and Savanna to Critical Transitions. Science, 2011 VOL 334 10.1126/science.1210657. www.sciencemag.org/cgi/content/full/334/6053/232/DC1

[26] Charlotte Dion, Sarah Lemler. Nonparametric drift estimation for diffusions with jumps driven by a Hawkes process. Statistical Inference for Stochastic Processes, 2020, ff10.1007/s11203-020-09213-5ff. ffhal-02094627v2f

[27] Baum C. F., Hurn S. and Lindsay K. The BDS test for independence. The Stata Journal (2021). 21(2), 279-294. DOI: 10.1177/1536867X211025796

[28] Carpenter SR, Brock WA (2006) Rising variance: a leading indicator of ecological transition. Ecol Lett 9: 311–318.

[29] Horsthemke W, Lefever R (2006) Noise-Induced Transitions : Theory and Applications in Physics, Chemistry, and Biology. Berlin: Springer.

[30] Bandi F, Nguyen T (2003) On the functional estimation of jump-diffusion models. J Econom 116: 293–328.

[31] Ana Filipa Duarte1 • Am´ılcar Soares1 • Maria Jo˜ao Pereira1 • Andr´e Peralta-Santos3,4 • Pedro Pinto Leite2 • Leonardo Azevedo. A Stochastic Model of an Early Warning System for Detecting Anomalous Incidence Values of COVID-19. Mathematical Geosciences (2024) 56:41–54. https: //doi.org/10.1007/s11004-023-10096-4

[32] Oraya Srimokla1 , Wirichada Pan Ngum2,3, Amnat Khamsiriwatchara4 , Chantana Padungtod5 , Rungrawee Tipmontree5 , Noppon Choosri6 and Sompob Saralamba2*. Early warning systems for malaria outbreaks in Thailand: an anomaly detection approach. Srimokla et al. Malaria Journal (2024) 23:11 https://doi.org/10.1186/s12936-024-04837-x

[33] Dakos V, Carpenter SR, Brock WA, Ellison AM, Guttal V, et al. (2012) Methods for Detecting Early Warnings of Critical Transitions in Time Series Illustrated Using Simulated Ecological Data. PLoS ONE 7(7): e41010. doi:10.1371/journal.pone.0041010

[34] Dakos V, Carpenter SR, Brock WA, Ellison AM, Guttal V, Ives AR, K´efi S, Livina V, Seekell DA, van Nes EH, Scheffer M. Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. PLoS One. 2012;7(7):e41010. https://doi.org/10.1371/journal.pone.0041010 Epub 2012 Jul 17. PMID: 22815897; PMCID: PMC3398887.

[35] Vasilis Dakos, Marten Scheffer, Egbert H. van Nes, Victor Brovkin, Vladimir Petoukhov , and Hermann Held. Slowing down as an early warning signal for abrupt climate change. PNAS, September 23, 2008, vol. 105, no. 38, 14308–14312

[36] Marina Hirota,1 Milena Holmgren,2 * Egbert H. Van Nes,1 Marten Scheffer1. Global Resilience of Tropical Forest and Savanna to Critical Transitions. Science, 2011 VOL 334 10.1126/science.1210657. www.sciencemag.org/cgi/content/full/334/6053/232/DC1

[37] Charlotte Dion, Sarah Lemler. Nonparametric drift estimation for diffusions with jumps driven by a Hawkes process. Statistical Inference for Stochastic Processes, 2020, ff10.1007/s11203-02009213-5ff. ffhal-02094627v2f

[38] Crichton N.J. (1999). Information point: Spearman's rank correlation. Journal of Clinical Nursing, 8, 763