Computers in Biology and Medicine 184 (2025) 109367

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A Gaussian Process Regression and Wavelet Transform Time Series approaches to modeling Influenza A

Edmund Fosu Agyemang *

School of Mathematical and Statistical Science, College of Sciences, University of Texas Rio Grande Valley, USA
Department of Statistics and Actuarial Science, College of Basic and Applied Sciences, University of Ghana, Ghana
Department of Computer Science, Ashesi University, No. 1 University Avenue, Berekuso, Eastern Region, Ghana

ARTICLE INFO

Keywords:
Influenza A
Discrete wavelet transform
Continuous wavelet transform
Gaussian process regression

ABSTRACT

The global spread of Influenza A viruses is worsening economic and social challenges. Various mechanistic models have been developed to understand the virus's spread and evaluate intervention effectiveness. This study aimed to model the temporal dynamics of Influenza A using Gaussian Process Regression (GPR) and wavelet transform approaches. The study employed Continuous Wavelet Transform (CWT), Discrete Wavelet Transform (DWT) and Wavelet Power Spectrum to analyze time-series data from 2009 to 2023. The GPR model, known for its non-parametric Bayesian nature, effectively captured non-linear trends in the Influenza A data, while wavelet transforms provided insights into frequency and time-localized characteristics. The integration of GPR with DWT denoising techniques demonstrated superior performance in forecasting Influenza A cases compared to traditional models like Auto Regressive Integrated Moving Averages (ARIMA) and Exponential Smoothing (ETS) using Holt–Winter method. The study identified significant anomalies in Influenza A cases, corresponding to known pandemic events and seasonal variations. These findings highlight the effectiveness of combining wavelet transform analysis with GPR in understanding and predicting infectious disease patterns, offering valuable insights for public health planning and intervention strategies. The research recommends extending this approach to other respiratory viruses to assess its broader applicability.

1. Introduction

Influenza A commonly known as swine flu, presents a substantial public health challenge due to its potential for rapid transmission and mutation [1]. Accurate modeling of this disease's spread is critical for effective prevention and control measures. Recent advances in statistical learning provide sophisticated tools for understanding and predicting the behavior of such infectious diseases. This study focuses on employing a Gaussian Process Regression (GPR) Model, Continuous Wavelet Transform (CWT), Discrete Wavelet Transform (DWT) and Wavelet Power Spectrum (WPS) Time Series approaches to model the incidence of Influenza A. The GPR model, a non-parametric Bayesian approach, offers a flexible framework for capturing the complex, non-linear patterns inherent in epidemiological data. Its probabilistic nature allows for the quantification of uncertainty in predictions, which is crucial for planning public health interventions [2]. By considering the full distribution of possible outcomes, the GPR model enables the identification of underlying trends and seasonal effects in disease prevalence, providing a deeper understanding of the temporal dynamics of the Influenza A virus. Complementing this, the Wavelet Power Spectrum analysis serves as a potent tool for uncovering the frequency and time-localized characteristics of the Influenza A time series data. By decomposing the data across various scales, it reveals the intricate oscillatory modes that traditional methods may overlook. This multiscale analysis is especially pertinent given the cyclical nature of influenza outbreaks, which may be influenced by factors such as environmental changes, population mobility, and social behavior [3]. Integrating these two approaches, the study aims to offer a comprehensive insight into the spread of Influenza A. The GPR model provides the predictive foundation, while the Wavelet Power Spectrum contributes a significant analysis of periodicity and variance in the data over time. In lieu of this, the primary objective of this study is to highlight the role of wavelet transform such as DWT and CWT in time series analysis and also to demonstrate the advantages in modeling and forecasting time series data when denoised using wavelet transform. Moreover, the study seeks to identify and understand significant deviations, or anomalies, from the expected trend in Influenza A cases. These anomalies, which represent periods where the actual number of Influenza A cases significantly exceeded or fell below the predicted values, offer critical

* Corresponding author at: School of Mathematical and Statistical Science, College of Sciences, University of Texas Rio Grande Valley, USA.
E-mail address: edmundfosu6@gmail.com.

https://doi.org/10.1016/j.compbiomed.2024.109367
Received 6 June 2024; Received in revised form 18 October 2024; Accepted 4 November 2024
Available online 16 November 2024

insights into the underlying factors influencing Influenza transmission and reporting.

Influenza A disease has wreaked havoc worldwide with its capacity to incite widespread morbidity and mortality, destabilize health systems, and inflict substantial economic damage [4,5]. Its highly mutable nature leads to frequent epidemics and pandemics, challenging global preparedness and response efforts. Particularly vulnerable populations, including the young, elderly, and those with chronic health conditions, face significant risks. The disease's impact extends beyond health, affecting economies, workplaces, and schools, highlighting the need for effective vaccines, rapid diagnostics, and robust public health strategies to mitigate its far-reaching consequences [6]. It also disrupts global supply chains, diminishes productivity through workforce absenteeism, and imposes a heavy load on healthcare resources, leading to billions in economic losses [7]. Its zoonotic nature, jumping between animals and humans, amplifies its threat, complicating control measures. The disease pinpoints critical gaps in global health security, emphasizing the urgency for advanced surveillance, rapid response capabilities, and global cooperation to enhance vaccine development and distribution, aiming to safeguard global health and economic stability. The devastation of Influenza A extends to straining international relations, as countries implement travel restrictions and border controls, often leading to diplomatic tensions. The pandemic readiness disparities between countries call attention to global inequality, emphasizing the necessity for international cooperation in health technology transfer and capacity-building initiatives [8]. Furthermore, the psychological impact on populations, from fear and anxiety to social isolation, cannot be understated. These crises necessitate a unified, global response that includes scientific research, public health preparedness, economic support, and compassionate community actions to navigate the challenges posed by Influenza A [9,10].

[11] offers a comprehensive review estimating the reproduction number (R) for different types of influenza providing insight into the transmissibility of the virus. The study noted that the reproduction number varied significantly across different settings and pandemic waves, illustrating the variability in Influenza's spread and the challenges in predicting and managing outbreaks. This review emphasizes the importance of understanding R values in public health planning and response strategies as Influenza surveillance plays a crucial role in public health by providing early warning of outbreaks and guiding control measures [12]. Traditional methods, such as the Serfling regression models, have been used for decades to estimate baseline Influenza activity and identify epidemic periods. However, the challenge lies in accurately excluding predefined epidemic periods to establish a reliable baseline. [13] however introduced an Adjusted Serfling Regression Model to enhance early detection of peak timing of Influenza in Beijing. The Adjusted Model addresses this issue by iteratively fitting the baseline and making adjustments in the rule of epidemic-period exclusion based on actual observations, leading to improved performance in peak timing prediction. The study climax the effectiveness of the Adjusted Serfling Regression Model compared to traditional methods. By incorporating adjustments in the rule of epidemic-period exclusion and iteratively fitting the baseline based on actual observations, the Adjusted model demonstrated superior performance in detecting peak timing of Influenza outbreaks. This improvement was evident in terms of sensitivity, specificity, and lead time, highlighting the model's enhanced accuracy in peak timing predictions. Moreover, the flexibility in baseline establishment and the focus on peak timing rather than epidemic onset provided a more objective and accurate measure for evaluating alternative approaches. The study also emphasized the Adjusted Model's success in prospective forecasting, indicating its potential for real-time surveillance and early detection of Influenza events. Their findings suggest that the Adjusted Serfling Regression Model offers a more adaptive and data-driven approach to influenza surveillance, leading to improved early warning systems for influenza outbreaks in Beijing.

[14] explores the application of wavelet coherence analysis to understand the coupling of intermuscular signals during stroke recovery. The study provides insights into how myoelectric-controlled interfaces can modulate muscle coordination, which is crucial for developing advanced rehabilitation technologies. [15] likewise applied a novel wavelet-based metric for detecting specific patterns in time series data collected through telemonitoring platforms. The methodology combines wavelet transforms with state space models to predict future events like hypertension in patients. [16] presents a method to analyze nonstationary time-series data using wavelet transforms. The study is particularly focused on plasma physics, where the behavior of charged particles creates complex, fluctuating signals. Traditional analysis methods struggle with nonstationary data, but wavelet transforms offer a solution by decomposing the signals into time–frequency components, allowing for the detection and analysis of transient events and fluctuations. This methodology can be applied to other fields, including medical data analysis, where similar nonstationary characteristics are common.

The integration of wavelet approaches (DWT and CWT) and GPR model in modeling infectious diseases is key as they represent a robust methodology for capturing the complex behavior of Influenza outbreaks, aiding in the anticipation of future outbreak and the reinforcement of public health preparedness, hence, their adoption in this study. The remainder of the paper is organized as follows: Section 2 discusses the data and methods used for the study. Section 3 provides the mathematical framework of the study. Section 4 discusses the results and findings of the study whilst Section 5 discusses, concludes and provides recommendations for further studies.

2. Data and methods

Secondary Influenza A data from January 2009 to December 2023 retrieved from Our World in Data and can be assessed at https://ourworldindata.org/influenza was used for the analysis of the study. The data was analyzed using DWT, CWT and Gaussian Power Regression (GPR) model to explore the temporal dynamics and frequency characteristics of Influenza A cases in the United States of America (USA). Monthly Influenza cases from January 2009 to December 2022 (168 months) were used in training the respective models under consideration, while the monthly Influenza cases from January 2023 to December 2023 (12 months) were used in testing the accuracy of the models. For the GPR model, the predictor X was created by transforming dates into a numerical sequence, indicating days elapsed since the earliest date. The response variable Y consists of the observed Influenza cases, allowing the model to analyze Influenza trends over time. Likewise, in the GPR model, a composite kernel was employed, combining a Constant Kernel (C) and a Radial Basis Function (RBF) kernel. The Constant Kernel is initialized with a scaling factor of 1.0, with its value allowed to vary within a range from 10^{-3} to 10^{3}. This kernel acts as a scaling factor for the model's output. The RBF kernel, initialized with a length scale of 10, permits its length scale parameter to adjust within the range from 10^{-2} to 10^{2}. This flexibility in the length scale allows the model to adapt to the smoothness of the data's underlying function. The GPR model is configured with `n_restarts_optimizer=10`, indicating that the optimization process, aimed at finding the best parameters for the kernel, is executed ten times from different initial conditions. This approach enhances the likelihood of locating a global optimum in the parameter space. First, the GPR model is fitted to the data and afterwards DWT is applied to the Influenza case data using the four variants (Daubechies, Coilflet, Discrete Meyer and Biorthogonal wavelets). This step decomposes the signal into approximation and detail coefficients while also denoising the signal. This denoised version of the original Influenza time series data retains the essential trends and patterns while minimizing the impact of noise, making it more suitable for accurate modeling. The fitted DWT-GPR models are then used to make predictions and their performance were measured against

two standard traditional statistical models Auto Regressive Integrated Moving Averages (ARIMA) with a seasonal component and Exponential Smoothing (ETS) using the Holt–Winters method.

For the CWT analysis, the process begins by ensuring that the data is in the correct format: the 'Date' column is converted to a datetime format, which allows for proper time-based indexing, and the 'FluAcases' column is converted to a float type to facilitate numerical computations. The signal, which represents the Influenza case data, is then extracted from the 'FluAcases' column and prepared for the wavelet transform. The CWT analysis employs the `cmor1.5-1.0` wavelet (a Complex Morlet wavelet) for its ability to provide a good balance between time and frequency localization. The choice of this wavelet is motivated by its efficacy in capturing the sinusoidal shapes common in time series data, making it suitable for analyzing seasonal patterns and trends in Influenza cases. A range of 1 to 128 was chosen to cover a broad spectrum of frequencies. These scales correspond to different levels of detail, allowing the analysis to capture both high-frequency (short-term) and low-frequency (long-term) patterns in the data. The Continuous Wavelet Transform is then applied to the signal using the defined wavelet and scales. The CWT produces coefficients that represent the magnitude of the wavelet's correlation with the signal at each scale and time point, effectively providing a time–frequency representation of the data. This process involves discretizing both the scales and the time axis, allowing the continuous nature of the wavelet transform to be implemented in a computationally feasible manner. Finally, the wavelet coefficients are visualized using a heatmap, where the x-axis represents time (from the earliest to the latest date in the dataset), and the y-axis represents the different scales. The color intensity in the heatmap corresponds to the magnitude of the wavelet coefficients, with different colors indicating areas of high or low activity in the time–frequency domain. This visualization allows for the identification of patterns in Influenza A cases over time, revealing how the frequency content of the signal evolves and highlighting significant trends or periodic behaviors. The study's analysis was conducted using the Python and R programming languages.

3. Methodology

3.1. Gaussian process regression

Gaussian Process Regression (GPR) is a non-parametric Bayesian approach that models the distribution of functions using a Gaussian process (GP) [17]. A GP is defined by its mean function $m(x)$ and covariance function $k(x, x')$, where x and x' are inputs in the domain. Given the training Influenza A data $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ with corresponding target values $\mathbf{Y} = \{y_1, y_2, \ldots, y_n\}$, we model the target values as being generated from a true function $f(x)$ corrupted by Gaussian noise ϵ, such that $y = f(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$. The GP prior over the function $f(x)$ in this study is expressed in (1) as:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')). \tag{1}$$

where the mean function $m(x)$ is often assumed to be zero, and $k(x, x')$ is the kernel or covariance function. The joint distribution of observed targets $\mathbf{Y}$ and predictions $\mathbf{f}_*$ at new input points $\mathbf{X}_*$ is given in (2) by:

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{pmatrix}\right) \tag{2}$$

where $K(\mathbf{X}, \mathbf{X})$ is the covariance matrix of the training points, $K(\mathbf{X}, \mathbf{X}_*)$ is the covariance between the training points and prediction points, $K(\mathbf{X}_*, \mathbf{X}_*)$ represents the covariance matrix computed between all pairs of prediction points and I is the identity matrix scaled by noise variance σ_n^2.

The posterior predictive distribution for $\mathbf{f}_*$, given $\mathbf{Y}$, is Gaussian with mean and covariance given in (3) and (4) represented by:

$$\mathbb{E}[\mathbf{f}_* | \mathbf{X}_*, \mathbf{X}, \mathbf{Y}] = K(\mathbf{X}_*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I]^{-1} \mathbf{Y} \tag{3}$$

$$\mathrm{Var}[\mathbf{f}_* | \mathbf{X}_*, \mathbf{X}, \mathbf{Y}] = K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I]^{-1} K(\mathbf{X}, \mathbf{X}_*). \tag{4}$$

These formulations allow for predictions at new Influenza A cases along with an estimation of the uncertainty of these predictions, making GPR a powerful tool for regression tasks where quantifying prediction confidence is of essence.

3.2. Continuous and Discrete Wavelet Transforms

The CWT of an input signal $x(t)$ is defined by the integral transformation given in (5) by:

$$CWT(s, \tau) = \frac{1}{\sqrt{|s|}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t - \tau}{s}\right) dt \tag{5}$$

where s represents the scale factor, τ is the translation parameter, $\psi(t)$ denotes the mother wavelet, and $\psi^*(t)$ indicates the complex conjugate of the mother wavelet. This formation allows for the decomposition of the signal at various frequencies and times by adjusting the scale and translation of the wavelet.

The Discrete Wavelet Transform (DWT) is a mathematical tool used for signal processing and can be particularly used in the analysis of time-series data such as the Influenza A cases under investigation. It allows a signal to be decomposed into different frequency components, enabling the examination of both the temporal (time-based) and spectral (frequency-based) characteristics of the signal. The DWT decomposes a signal $x(t)$ into a set of wavelet coefficients that represent the signal at various scales and positions. The DWT can be expressed mathematically in (6) as:

$$\mathrm{DWT}(s, \tau) = \sum_t x(t) \cdot \psi_{s,\tau}(t) \tag{6}$$

where s is the scale parameter that controls the width of the wavelet, τ is the translation parameter that controls the position of the wavelet, $\psi_{s,\tau}(t)$ is the wavelet function at scale s and translation τ, and $x(t)$ is the original signal. The wavelet function $\psi_{s,\tau}(t)$ is derived from a mother wavelet $\psi(t)$ given in (7) as:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{2^s}} \psi \left(\frac{t - 2^s \tau}{2^s}\right) \tag{7}$$

where 2^s controls the dilation (scaling) of the wavelet and $t - 2^s\tau$ controls the translation (shift) of the wavelet. The DWT operates by recursively applying a pair of filters to the signal: a low-pass filter $g[\tau]$ that captures the low-frequency components (approximations) and a high-pass filter $h[\tau]$ that captures the high-frequency components (details). This process is described in (8) and (9) as:

$$a_{j+1}[\tau] = \sum_k g[k - 2\tau] \cdot a_j[k] \tag{8}$$

and

$$d_{j+1}[\tau] = \sum_k h[k - 2\tau] \cdot a_j[k] \tag{9}$$

where $a_j[\tau]$ is the approximation at level j, $d_j[\tau]$ is the detail at level j, $g[k]$ and $h[k]$ are the coefficients of the low-pass and high-pass filters, respectively. By iterating this process, the DWT provides a multi-resolution analysis of the signal, breaking it down into various levels of detail and approximation that can be analyzed separately.

The original signal was reconstructed from the wavelet coefficients in this study using the Inverse Discrete Wavelet Transform (IDWT), which is given in (10) by:

$$x(t) = \sum_s \sum_\tau \mathrm{DWT}(s, \tau) \cdot \psi_{s,\tau}(t) \tag{10}$$

The IDWT was employed as it recombines the approximations and details at each level to recover the original signal, ensuring that no information is lost during the time series decomposition process. For more insight on wavelet methodology, the reader is directed to [18] and see [19] for details on wavelet variants such as Daubechies and Biorthogonal.

3.3. Wavelet power spectrum

The variance of the signal at each scale, or the Wavelet Power Spectrum, is given (11) by:

$$Power(s,\tau) = |CWT(s,\tau)|^2. \tag{11}$$

This spectrum presents a two-dimensional view of the signal's power across time and frequency scales, pinpointing regions with significant variance which often correspond to notable features in the data. Through the CWT, the study analyzed various signals in components associated with different scales of the Influenza cases, facilitating a comprehensive examination of the disease structure. The Wavelet Power Spectrum quantifies these components' strengths, providing insights into the signal dynamics and characteristics across various scales. This approach is especially beneficial for analyzing non-stationary data as it is the case of the Influenza A cases, where the signal's frequency content evolves over time, making it invaluable in diverse scientific and engineering fields. Hence, its usage in this study is crucial.

3.4. Model assessment metrics

In order to evaluate the competing models, Theil U1 Statistic (τ), root mean square error (RSME), mean absolute error (MAE) and mean absolute percentage error (MAPE) are employed in this study. The computation of the model assessment metrics are given by (12)–(15).

$$\tau = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2} \div \left(\sqrt{\frac{1}{n}\sum_{t=1}^{n}y_t^2} + \sqrt{\frac{1}{n}\sum_{t=1}^{n}\hat{y}_t^2}\right). \tag{12}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}}. \tag{13}$$

$$MAE(MAD) = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|. \tag{14}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{y_t} \times 100\%. \tag{15}$$

where y_t are the actual values, $\hat{y}_t$ are the forecast values and $y_t - \hat{y}_t = \epsilon_t$ are the forecast errors, $\bar{y}_t$ is the mean of the actual values. $0 \le \tau \le 1$, for $\tau \approxeq 0$ implies good fit of model to data and $\tau \approxeq 1$ implies poor fit of model to data. Lower values of these error metrics are indicative of a good model.

4. Results and discussion

This section presents the discussion of the findings of the study.

4.1. Wavelet decomposition of Influenza A cases

The wavelet decomposition of the Influenza cases data provides a multi-resolution analysis, allowing us to examine the underlying trends and fluctuations at various scales. This approach is particularly advantageous for identifying patterns that may not be immediately evident in the raw time series data.

Fig. 1 shows a five-level multi-resolution wavelet synthesis of the monthly Influenza A series from January 2009 to December 2023 using Daubechies wavelet of order 1. The first component of the wavelet decomposition, referred to as the approximation (low-frequency) component, captures the overall trend of Influenza A cases across the observed period. This component filters out short-term fluctuations and emphasizes the long-term behavior of the data. From the approximation plot, we observe a general upward trend in Influenza A cases, particularly noticeable in certain years (for example, 2019 and 2023), reflecting potentially seasonal peaks or outbreak events. This trend suggests a periodic increase in Influenza A cases, which could be associated with seasonal changes, variations in population immunity, or changes in the virus strain. The smooth nature of the approximation curve indicates that the wavelet transform effectively isolates the significant, slow-varying patterns within the data. These patterns may correspond to macro-level epidemiological factors, such as vaccination campaigns, public health interventions, or the introduction of new Influenza A strains. The gradual increase in the approximation curve during specific periods may also suggest years of higher Influenza A activity, warranting further investigation into the corresponding external factors during those times.

The detail components represent the high-frequency variations in the flu cases data, capturing the short-term fluctuations and abrupt changes that the approximation component smooths out. Each subsequent detail component analyzes the data at increasingly finer resolutions. The first detail component, which captures the highest level of high-frequency fluctuations, reveals significant variability in the Influenza A cases. These fluctuations could be indicative of sudden outbreaks, possibly linked to specific epidemiological events such as the emergence of a particularly virulent strain of the Influenza A virus or breakdowns in public health measures. The presence of sharp peaks in the detail components suggests periods of acute increases in Influenza A cases, possibly corresponding to localized outbreaks or anomalies in reporting practices. As we move to higher detail levels (lower in frequency but still high relative to the approximation), the fluctuations become less pronounced, reflecting more localized variations in the data. These finer details might represent random noise or smaller-scale events that contribute to the overall Influenza A activity but are less impactful than the major outbreaks captured in the higher detail components.

The wavelet decomposition reveals that the Influenza cases data exhibits both strong seasonal trends and significant short-term fluctuations. The low-frequency trend highlights a periodic pattern that could be predictive of future Influenza A activity, while the high-frequency components highlight the unpredictable nature of Influenza A outbreaks. The analysis also suggests that while long-term patterns are relatively stable, the short-term variability is considerable, pointing to the complexity of factors influencing Influenza A transmission. These factors may include not only biological elements, such as virus mutation rates and vaccine efficacy, but also social determinants, such as population density, mobility patterns, and public health response effectiveness.

4.2. Analysis of Continuous Wavelet Transform

The CWT was applied to the Influenza A time series data to obtain a time–frequency representation, enabling the identification of both localized and extended patterns in the data.

In Fig. 2, the color intensity indicates the magnitude of the wavelet coefficients. The prominent feature is a large red area at the top right corner of the plot, indicating a high magnitude of coefficients in recent years at higher scales. This suggests the presence of a strong low-frequency component in the Influenza A cases, which could correspond to a long-term trend or change in the data. The nature of the trend could be an increase in the number of cases, a shift in the pattern of cases, or the emergence of a new cycle in the recent years. Moving down the scale, there is a repeating pattern of ridges, particularly visible from around 2010 to 2020. These ridges suggest periodicity in the Influenza A cases data. The consistent appearance of these patterns at specific scales indicates that there are regular cycles in the data, which could correspond to seasonal variations in Influenza A cases. These might represent the typical seasonal Influenza cycles known to occur annually. The color intensity varies across different years, which shows that the strength of these seasonal cycles changes over time. Some years show brighter colors, indicating higher magnitudes of wavelet coefficients, suggesting more pronounced seasonal effects or potentially more severe Influenza A seasons in those years. The areas of the plot with less intense color, particularly at lower scales, suggest less activity in the corresponding frequencies at those times, which could represent

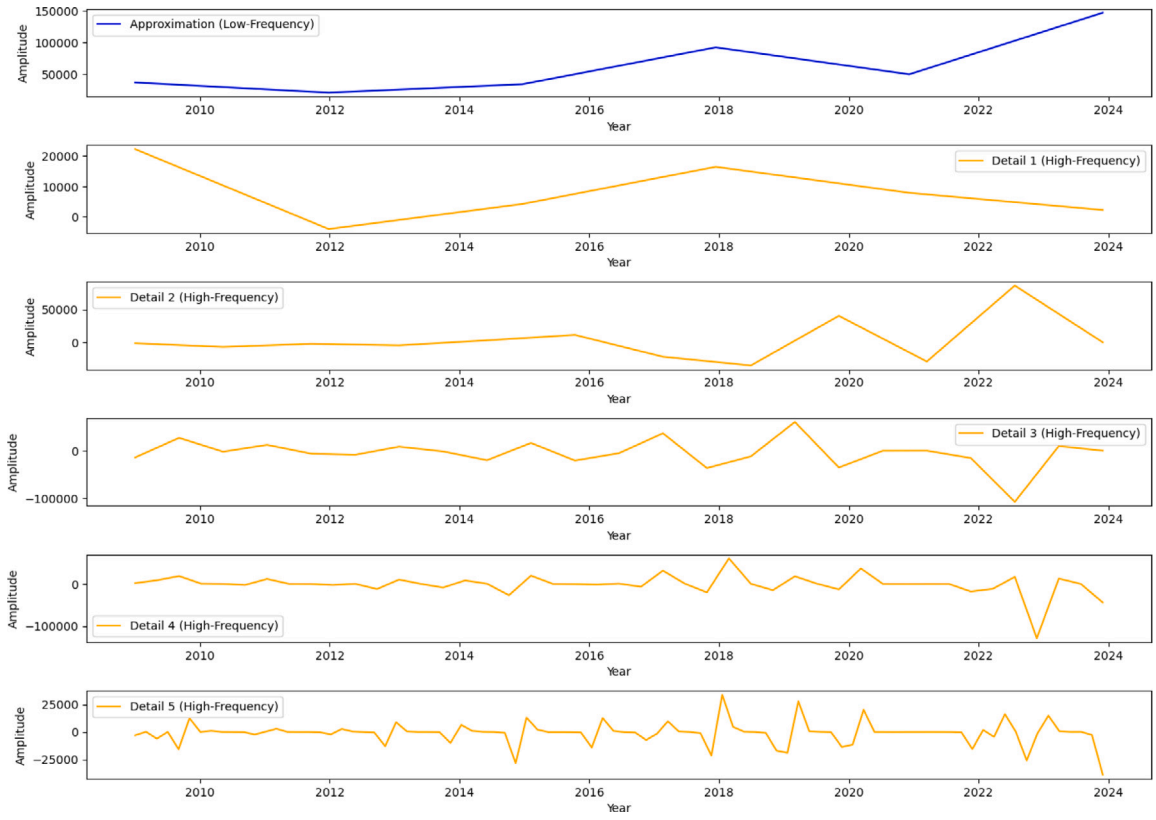


Fig. 1. 5-level wavelet transform for the monthly Influenza A cases using Daubechies filter of order 1.

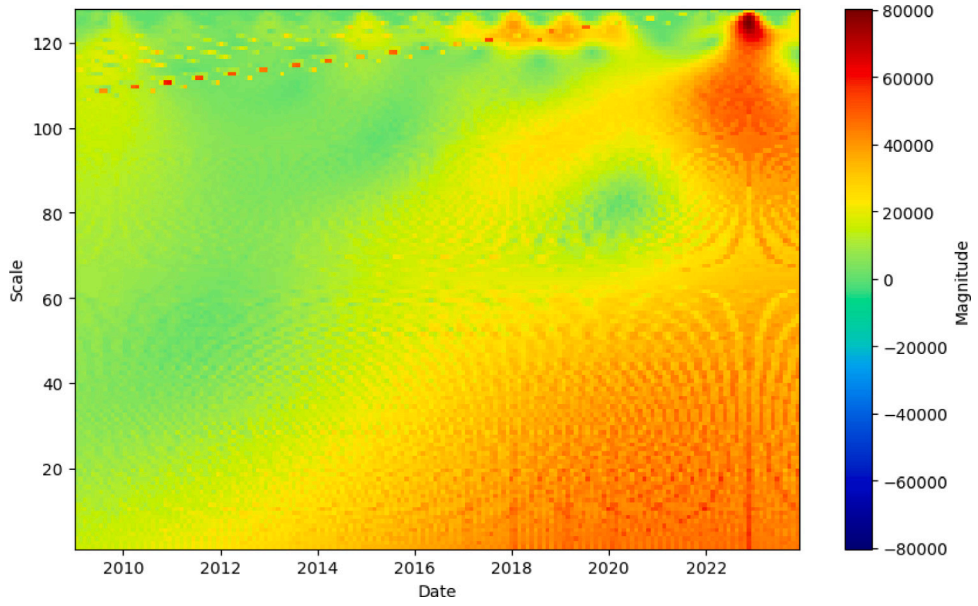


Fig. 2. Continuous Wavelet Transform (CWT) of Influenza cases.

periods with fewer cases or less variation in the number of cases. The CWT analysis of Influenza cases from 2009 onwards reveals both long-term trends and seasonal patterns. The presence of a strong trend in recent years may warrant further investigation to determine its cause, which could be related to external factors such as changes in the virus, vaccine effectiveness, or healthcare practices. The seasonal patterns align with expected behavior for Influenza cases but show variation in intensity, which may correlate with different strains of the virus or public health responses. Fig. 2 provides a powerful tool for public health officials and epidemiologists to understand the dynamics of Influenza A cases over time and to plan for future seasons accordingly.

4.3. Analysis of Wavelet Power Spectrum

The WPS was derived from the coefficients obtained via CWT, providing a visualization of the variance of the Influenza cases across

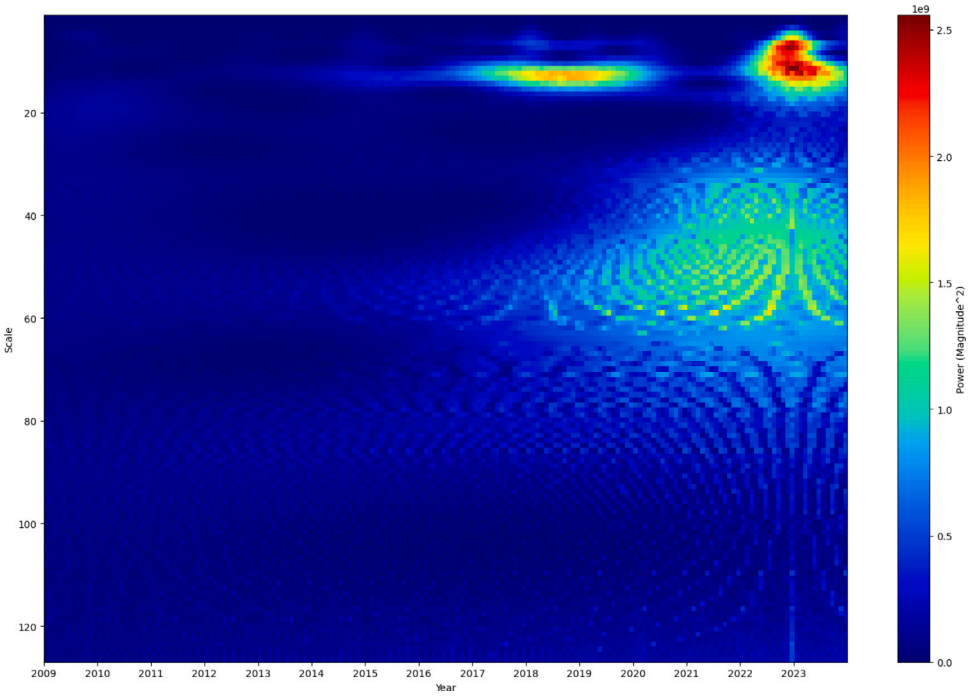


Fig. 3. Wavelet Power Spectrum of Influenza cases.

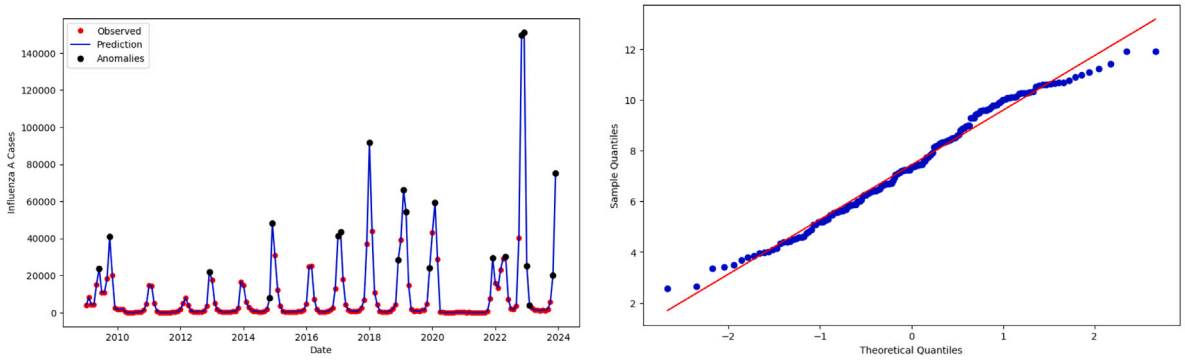
different scales (frequencies) over time. This spectrum illuminates areas of increased activity or variability, indicating potential outbreaks or significant changes in the Influenza cases.

The Wavelet Power Spectrum depicted in Fig. 3 also provides a different variant time–frequency analysis of Influenza A case counts from 2009 to 2023. The x-axis denotes consecutive years, providing a chronological perspective, while the y-axis represents the scales of wavelet analysis, with larger scales correlating with lower frequency components at the top and smaller scales with higher frequency components at the bottom. The color gradient, signifying the wavelet power (the square of the wavelet coefficient magnitude), reveals the intensity of variations in Influenza A cases across different frequencies over time. Darker regions indicate lower power, suggesting minimal variability in the time series data at those scales, while brighter regions denote higher power, reflecting significant activity or fluctuations. Upon close observation, the spectrum exhibits localized bright regions, which correspond to periods of high variability in Influenza case counts. Such regions could signify outbreaks or surges in Influenza cases. From Fig. 3, we can discern areas of heightened activity corresponding to significant periodic components within the time series. For example, the horizontal band of intense color around the scales of 20–40 suggests a dominant cycle in the Influenza case data within this frequency band over the span of several weeks. This could potentially correspond to a biweekly or monthly oscillation in Influenza cases. Towards the top of the spectrum, we observe a concentrated region of high power in the last quarter of the time axis. This abrupt and pronounced increase in power at lower frequencies may signify a notable upward trend or shift in the long-term pattern of Influenza A cases, indicative of a potential outbreak or an atypical increase in cases. It warrants further epidemiological investigation to ascertain the underlying causes. Furthermore, the spectrum intermittently shows horizontal banding at specific scales across various years. These consistent patterns could be indicative of seasonal cycles in Influenza A transmission, aligning with the known seasonal behavior of Influenza A outbreaks. The variation in intensity of these bands over the years might suggest fluctuating severity of seasonal cycles, potentially due to factors such as viral mutations, vaccination rates, or public health interventions.

4.4. Analysis of Gaussian process regression model

Fig. 4 gives the summary of the predictions of the GPR model and its anomaly detection schema. The GPR model captures the general trend in the data, including some of the variability in the number of Influenza A cases over time. However, due to the inherent noise and the possibly complex underlying patterns in the data (such as seasonality and outliers), some discrepancies between the observations and the model's predictions are expected. Fig. 4(a) showcases a non-linear relationship between time and Influenza case counts, pinpointing the complexity of Influenza spread dynamics that cannot be aptly captured by simpler linear models. The GPR model identified several key anomalies (21 anomalies) in the Influenza A cases data. Notably, anomalies were detected during the following periods: June 2009, October 2009, December 2012, November–December 2014, January–February 2017, January 2018, February–March 2019, December 2019, February 2020, December 2021, May 2022, November–December 2022, and November–December 2023. These anomalies are characterized by substantial deviations from the predicted trend, suggesting periods of unexpected increases or decreases in Influenza A cases.

The anomaly detected in June 2009 aligns with the emergence of the H1N1 influenza pandemic, often referred to as the 'swine flu.' During this time, there was a notable spike in Influenza A cases, as the new H1N1 virus strain spread rapidly across populations with little prior immunity. Similarly, the anomaly in October 2009 coincides with the second wave of the H1N1 pandemic, further emphasizing the impact of this event on Influenza A cases. The anomalies observed in the winter months of December 2012, November–December 2014, and January–February 2017 may correspond to particularly severe flu seasons. These periods often see heightened flu activity due to factors such as the circulation of more virulent strains or lower vaccine efficacy. For instance, the anomaly in January 2018, which recorded an exceptionally high number of flu cases, coincides with one of the most severe flu seasons in recent history, attributed to the prevalence of the H3N2 strain. The anomalies detected in February–March 2019 and December 2019–February 2020 suggest unusual flu activity leading up to the COVID-19 pandemic. This period may have seen fluctuations in flu reporting and transmission dynamics due to the emerging global health crisis.



(a) Anomaly Detection using Gaussian Process Regression (b) Q-Q Plot of Influenza A Cases

Fig. 4. Anomaly Detection in Influenza A Cases using Gaussian Process Regression and Normality Diagnostic.

The December 2021 anomaly, following the initial waves of COVID-19, could reflect changes in public health behavior, such as reduced social distancing and mask usage, leading to an unexpected rise in flu cases. The anomalies in May 2022, November–December 2022, and November–December 2023 might indicate changes in flu transmission patterns in the post-pandemic era. These periods could be associated with the relaxation of public health measures, waning immunity, or the emergence of new flu variants. The May 2022 anomaly, in particular, is noteworthy as it occurs outside the typical flu season, suggesting an unusual event or reporting change. See [20–23] for further details on anomaly detection.

The Shapiro–Wilk test was applied to statistically determine the normality of the data. The test yielded a p-value of 0.067. This p-value is significantly higher than the typical alpha level of 0.05, indicating that the null hypothesis (which posits that the data follow a normal distribution) cannot be rejected. This suggests that the Influenza A data is approximately normally distributed. In Fig. 4(b), the Q–Q plot further visualizes the possibly non-deviation from normality. Ideally, if the data were normally distributed, the points would lie along the red diagonal line. As seen in the Q–Q plot, the blue points closely follow the red line for most of the data, indicating that the data might be approximately normally distributed, which further corroborates the finding from the Shapiro–Wilk test. The combination of the Shapiro–Wilk test results and the Q–Q plot clearly indicates that the Influenza A data follows a normal distribution.

4.5. Modeling Influenza A using DWT-GPR models

Firstly, the GPR was fitted to the Influenza A data before the four variants of DWT (Daubechies, Coilflet, Discrete Meyer and Biorthogonal wavelets) were employed as a denoising tool to remove the noise in the Influenza A series data before modeling with the GPR model. The GPR and DWT-GPR models performance were evaluated against two traditional statistical time series models namely Auto Regressive Integrated Moving Averages (ARIMA) via seasonality and Exponential Smoothing (ETS) via the Holt–Winters Method. Table 1 presents a comprehensive comparison of actual Influenza A cases for 2023 with forecasted values generated using various models, including the Gaussian Process Regression (GPR) model and its variants combined with Discrete Wavelet Transform (DWT), alongside traditional models such as Auto Regressive Integrated Moving Averages (ARIMA) and Exponential Smoothing (ETS). The model evaluation metrics showcase the efficacy of these models by offering insights into their relative performance and the advantages of using wavelet-based denoising techniques in time series forecasting.

From the table, it is evident that the GPR model and its discrete wavelet-enhanced variants (GPR-DBI, GPR-Coif, GPR-DM, and GPR-Bior) consistently produce forecasts that are closer to the actual values compared to the traditional ARIMA and ETS models. This is particularly noticeable in the significant reduction of forecast errors, as indicated by the lower Mean Absolute Percentage Error (MAPE) and Theil's U1 statistics for the GPR models. For instance, the GPR-DBI model, which utilizes the Daubechies wavelet, achieves a MAPE of 53.5046 and a Theil's U1 statistic of 0.2221, both of which are substantially lower than the corresponding metrics for the ARIMA (MAPE: 1262.1027, Theil's U1: 0.4184) and ETS (MAPE: 5769.9103, Theil's U1: 0.7871) models. This indicates that the GPR-DBI model not only fits the data more accurately but also provides more reliable forecasts with less bias and variance.

Furthermore, the Root Mean Square Error (RMSE) values across the GPR models remain relatively consistent and significantly lower than those of the traditional models, leveraging the robustness of the discrete wavelet-based GPR models in handling the inherent variability in Influenza A cases data. The RMSE for the GPR-DBI model, for example, is 10537.6785, which contrasts sharply with the RMSE of 31331.9012 for the ARIMA model and a staggering 1131437 for the ETS model. This substantial difference suggests that the wavelet-enhanced data better capture the underlying trends and seasonal fluctuations.

5. Discussion and conclusion

The analysis conducted using a Gaussian Process Regression (GPR) model on Influenza A data revealed several critical insights into the trends, seasonality, and anomalies within the dataset. This systematic and comprehensive exploration provides a better understanding of Influenza A case drift, instrumental for public health monitoring and intervention planning. The GPR model captured the non-linear trend in Influenza cases over time, highlighting the complex nature of Influenza A transmission dynamics. The GPR model's ability to fit the historical data with a non-linear function and provide a predictive trend over time leverages GPR's flexibility and adaptability in modeling time-series data with inherent variability. Utilizing the GPR model's for anomaly detection, data points significantly deviating from expected trends were identified. These anomalies, potentially indicative of unusual Influenza A outbreaks or data collection errors, highlight the need for continuous monitoring and investigation of Influenza A case data to understand and respond to emerging public health threats. The anomalies identified in this analysis provide valuable insights into the factors that can cause significant deviations in Influenza A cases. By understanding these periods of anomalous activity, public health officials and researchers can better anticipate and respond to future outbreaks. The context surrounding each anomaly, such as the 2009 H1N1 pandemic or the ongoing impacts of COVID-19, highlights the importance of considering both biological and societal factors in flu transmission and reporting.

The model evaluation metrics clearly demonstrate the superiority of the discrete wavelet-enhanced GPR models (GPR-DBI, GPR-Coif,

Table 1
Actual and forecast for 2023 Influenza A cases with model evaluation metrics.

Date	Actual	GPR	GPR-DBI	GPR-Coif	GPR-DM	GPR-Bior	ARIMA	ETS
2023-01-31	24 955	6335	6335	6339	6394	6341	73 487	141 720
2023-02-28	3947	4746	4748	4745	4736	4746	30 077	155 289
2023-03-31	2376	849	852	848	784	847	32 007	157 984
2023-04-30	1378	1757	1759	1759	1703	1756	33 424	144 208
2023-05-31	1341	1089	1090	1090	1080	1090	33 753	129 951
2023-06-30	1156	1210	1207	1209	1211	1214	26 981	112 367
2023-07-31	1377	1588	1585	1587	1589	1586	25 501	113 712
2023-08-31	1168	571	572	573	431	572	25 431	118 415
2023-09-30	1703	4582	4583	4583	4343	4583	25 930	123 645
2023-10-31	5535	7602	7603	7605	7528	7602	36 679	143 343
2023-11-30	20 126	49 430	49 429	49 427	49 654	49 430	68 314	187 241
2023-12-31	75 147	64 613	64 610	64 610	64 909	64 613	66 776	170 518
MAPE		53.5581	53.5046	53.5325	53.2532	53.5528	1262.1027	5769.9103
Theil's U1		0.2229	0.2221	0.2228	0.2219	0.2220	0.4184	0.7871
MAE		5602.0589	5601.6667	5601.6667	5577.5833	5601.5833	29 574.4278	129 848
RMSE		10 537.8011	10 537.6785	10 536.8101	10 551.8311	10 536.9265	31 331.9012	1 131 437

DBI: Daubechies; Coif: Coiflet; DB: Discrete Meyer; Bior: Biorthogonal; ARIMA: ARIMA (0,1,3)(0,0,1); ETS: Exponential Smoothing.

GPR-DM, and GPR-Bior) in forecasting time series data with complex, non-linear patterns, such as those observed in the Influenza A cases. The application of DWT as a denoising tool before modeling with GPR allows these models to focus on the most relevant features of the data, effectively filtering out noise and reducing the impact of outliers. This results in more precise forecasts, particularly in periods with abrupt changes or anomalies, where traditional models tend to falter. This hybrid approach effectively combines the strengths of DWT for signal denoising and GPR for flexible, non-parametric modeling of time series data. The DWT denoising techniques employed in the study enhanced the quality of the Influenza A data by reducing noise, which, in turn, improves the accuracy and reliability of the wavelet-enhanced GPR model predictions. This approach is particularly valuable for modeling complex, noisy datasets such as Influenza A cases, where capturing both the trend and uncertainty is critical for public health forecasting and decision-making. From a public health perspective, the ability to decompose the Influenza cases data into these different components using DWT is invaluable. Through DWT decomposition, clear patterns of seasonality in Influenza cases were observed, with cases typically peaking in colder months. This analysis not only validated expected seasonal trends in Influenza A activity but also quantified the extent and timing of these fluctuations, which are crucial for effective public health planning. The approximation component can guide long-term planning and resource allocation, ensuring that healthcare systems are prepared for expected seasonal peaks. Meanwhile, the detail components can assist in real-time monitoring and response, enabling health authorities to react swiftly to emerging outbreaks. Moreover, the findings suggest that while it is possible to anticipate the general direction of Influenza A activity, the high level of short-term variability pinpoints the need for flexible and adaptive public health strategies. This approach would allow for rapid adjustments to vaccination campaigns, public advisories, and resource deployment in response to the sudden changes in flu cases revealed by the high-frequency analysis [24].

The analysis employing the Wavelet Power Spectrum and Continuous Wavelet Transform (CWT) approaches on Influenza A data spanning from 2009 to 2023 unveils critical insights into the temporal dynamics and frequency of Influenza A outbreak. Through the Wavelet Power Spectrum, we discern areas of heightened variability, represented by brighter colors, which highlight periods of increased activity or potential outbreaks, alongside consistent patterns indicative of seasonal cycles in Influenza cases. This spectral view is pivotal for identifying not only the cyclic behavior inherent in Influenza A outbreaks but also for pinpointing long-term trends or shifts in Influenza A case patterns, particularly noted in the lower frequency bands of recent years. The CWT analysis complements this by offering a better view into how these frequencies evolve over time, enabling the detection of both abrupt changes and gradual trends within the data. Together, these analyses furnish a comprehensive understanding of the Influenza's behavior over time, emphasizing the importance of both seasonal influences and longer-term shifts in the epidemiology of Influenza A [25]. These integrated hybrid approaches provide invaluable insights for public health planning and response strategies, highlighting the utility of wavelet analysis in capturing the complex dynamics of disease spread.

The integration of wavelet transforms in epidemiological modeling is an emerging field. This study's application of CWT and DWT to analyze Influenza A cases aligns with recent research that uses wavelet analysis to uncover periodicity and anomalies in health data. For example, [14] utilized wavelet coherence analysis to study muscle coordination during stroke recovery, emphasizing the utility of wavelet methods in capturing complex biological signals. Similarly, [25] applied wavelet-based methods to understand regional and seasonal variations in Influenza-related health outcomes across mainland China, highlighting the value of wavelet analysis in public health surveillance. The advantages of wavelet-based methods in handling non-stationary data are well-documented in the literature, as seen in the work of [16], where wavelet transforms were used to analyze non-stationary plasma fluctuations. Likewise, [15] applied wavelet-based approaches to time series pattern detection, showing the benefits of wavelet transforms in improving prediction accuracy by capturing transient events and fluctuations in data. This study therefore recommended the application of Gaussian Process Regression and its discrete wavelet-enhanced variants (GPR-DBI, GPR-Coif, GPR-DM, and GPR-Bior), and CWT to other respiratory viruses, such as SARS-CoV-2 (COVID-19), Influenza B, and RSV (Respiratory Syncytial Virus) to assess the models' versatility and adaptability across different pathogens.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The first, corresponding and sole author acknowledges the enormous support of the University of Texas Rio Grande Valley (UTRGV) Presidential Research Fellowship fund.

Data availability

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

References

[1] José Carlos Mancera Gracia, Douglas S. Pearce, Aleksandar Masic, Monica Balasch, Influenza A virus in swine: epidemiology, challenges and vaccination strategies, Front. Vet. Sci. 7 (2020) 647.

[2] Matteo Guardiani, Philipp Frank, Andrija Kostić, Gordian Edenhofer, Jakob Roth, Berit Uhlmann, Torsten Enßlin, Causal, Bayesian, & non-parametric modeling of the SARS-CoV-2 viral load distribution vs. patient's age, PLoS One 17 (10) (2022) e0275011.

[3] Konstantinos Sitaropoulos, Salvatore Salamone, Lina Sela, Frequency-based leak signature investigation using acoustic sensors in urban water distribution networks, Adv. Eng. Inform. 55 (2023) 101905.

[4] George F. Gao, William J. Liu, Let's get vaccinated for both flu and COVID-19: On the World Flu Day 2021, China CDC Wkly. 3 (44) (2021) 915.

[5] Ariful Islam, Sarah Munro, Mohammad Mahmudul Hassan, Jonathan H. Epstein, Marcel Klaassen, The role of vaccination and environmental factors on outbreaks of high pathogenicity avian influenza H5N1 in Bangladesh, One Health 17 (2023) 100655.

[6] Jennifer D. Roberts, Shadi O. Tehrani, Environments, behaviors, and inequalities: reflecting on the impacts of the influenza and coronavirus pandemics in the United States, Int. J. Environ. Res. Public Health 17 (12) (2020) 4484.

[7] Caroline de Courville, Sarah M. Cadarette, Erika Wissinger, Fabián P. Alvarez, The economic burden of influenza among adults aged 18 to 64: A systematic literature review, Influenza Other Respir. Viruses 16 (3) (2022) 376–385.

[8] Stephen G. Waller, The impact of pandemics on national and international security, in: The Nature of Pandemics, CRC Press, 2022, pp. 365–380.

[9] Serena Barello, Anna Falco-Pegueroles, Debora Rosa, Angela Tolotti, Guendalina Graffigna, Loris Bonetti, The psychosocial impact of flu influenza pandemics on healthcare workers and lessons learnt for the COVID-19 emergency: a rapid review, Int. J. Public Health 65 (7) (2020) 1205–1216.

[10] Emanuele Preti, Valentina Di Mattei, Gaia Perego, Federica Ferrari, Martina Mazzetti, Paola Taranto, Rossella Di Pierro, Fabio Madeddu, Raffaella Calati, The psychological impact of epidemic and pandemic outbreaks on healthcare workers: rapid review of the evidence, Curr. Psychiatry Rep. 22 (2020) 1–22.

[11] Matthew Biggerstaff, Simon Cauchemez, Carrie Reed, Manoj Gambhir, Lyn Finelli, Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature, BMC Infect. Dis. 14 (1) (2014) 1–20.

[12] Hyunju Lee, Heeyoung Lee, Kyoung-Ho Song, Eu Suk Kim, Jeong Su Park, Jongtak Jung, Soyeon Ahn, Eun Kyeong Jeong, Hyekyung Park, Hong Bin Kim, Impact of public health interventions on seasonal influenza activity during the COVID-19 outbreak in Korea, Clin. Infect. Dis. 73 (1) (2021) e132–e140.

[13] Xiaoli Wang, Shuangsheng Wu, C. Raina MacIntyre, Hongbin Zhang, Weixian Shi, Xiaomin Peng, Wei Duan, Peng Yang, Yi Zhang, Quanyi Wang, Using an adjusted serfling regression model to improve the early warning at the arrival of peak timing of influenza in Beijing, PLoS One 10 (3) (2015) 1–14.

[14] Hairong Yu, Weiling Xu, Yu Zhuang, Kaiyu Tong, Rong Song, Wavelet coherence analysis of muscle coupling during reaching movement in stroke, Comput. Biol. Med. 131 (2021) 104263.

[15] Teresa Rocha, Simão Paredes, Paulo Carvalho, Jorge Henriques, A wavelet-based approach for time series pattern detection and events prediction applied to telemonitoring data, in: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2011, pp. 6037–6040.

[16] S. Santoso, E.J. Powers, Roger D. Bengtson, A. Ouroua, Time-series analysis of nonstationary plasma fluctuations using wavelet transforms, Rev. Sci. Instrum. 68 (1) (1997) 898–901.

[17] M.T. Alodat, Mohammed K. Shakhatreh, Gaussian process regression with skewed errors, J. Comput. Appl. Math. 370 (2020) 112665.

[18] Rodrigo Capobianco Guido, Wavelets behind the scenes: Practical aspects, insights, and perspectives, Phys. Rep. 985 (2022) 1–23.

[19] Rumaih M. Alrumaih, Mohammad A. Al-Fawzan, Time series forecasting using wavelet denoising an application to saudi stock index, J. King Saud Univ., Eng. Sci. 14 (2) (2002) 221–233.

[20] Edmund Fosu Agyemang, Ezekiel NN Nortey, Richard Minkah, Kwame Asah-Asante, The unfolding mystery of the numbers: First and second digits based comparative tests and its application to Ghana's elections, Model Assist. Stat. Appl. 18 (2) (2023) 183–192.

[21] Edmund F. Agyemang, Ezekiel N.N. Nortey, Richard Minkah, Kwame Asah-Asante, Baseline comparative analysis and review of election forensics: Application to Ghana's 2012 and 2020 presidential elections, Heliyon (2023).

[22] Ezekiel Nii Noi Nortey, Edmund Fosu Agyemang, Richard Minkah, Kwame Asah-Asante, Bayesian estimation of presidential elections in Ghana: A validation approach, Afr. J. Appl. Stat. 9 (1) (2022) 1297–1317.

[23] Edmund Fosu Agyemang, Anomaly detection using unsupervised machine learning algorithms: a simulation study, Scientific African 26 (2024) e02386.

[24] Matthew Newland, David Durham, Jason Asher, John J. Treanor, Jonathan Seals, Ruben O. Donis, Robert A. Johnson, Improving pandemic preparedness through better, faster influenza vaccines, Expert Rev. Vaccin. 20 (3) (2021) 235–242.

[25] Charlie Diamond, Hui Gong, Fiona Yueqian Sun, Yang Liu, Billy J. Quilty, Mark Jit, Juan Yang, Hongjie Yu, W. John Edmunds, Marc Baguelin, Regional-based within-year seasonal variations in influenza-related health outcomes across mainland China: a systematic review and spatio-temporal analysis, BMC Med. 20 (1) (2022) 58.