International Journal of Biomathematics
(2024) 2450132 (24 pages)

DOI: 10.1142/S1793524524501328

Analyzing and forecasting dengue fever incidence in Thailand: A comprehensive study for public health preparedness

Sayooj Aby Jose[*,†,§,‖], Karuna Mathew[‡], Hamna Mariyam K. B.[†], Anuwat Jirawattanapanit[*,¶,‖] and Anurak Wereprasertsakun[*]

[*]*Department of Mathematics, Faculty of Education*
Phuket Rajabhat University, Phuket, Thailand

[†]*School of Data Analytics, Mahatma Gandhi University*
Kottayam, Kerala, India

[‡]*Faculty of Engineering Environment and Computing*
Coventry University, Coventry, UK
[§]*sayooaby999@gmail.com*
[¶]*anuwat.j@pkru.ac.th*

Received 6 May 2024
Revised 25 August 2024
Accepted 14 September 2024
Published 23 November 2024

Communicated by Bo Zheng

Thailand is currently grappling with a severe dengue fever outbreak, with a rising threat to public health as the rainy season and El Niño draw near. This year has witnessed a troubling surge in dengue cases, prompting the Ministry of Public Health (MoPH) to issue warnings that the numbers may hit a three-year peak. Dengue outbreaks in Thailand have historically followed a cyclical pattern, excluding COVID-19 years. This research employs data analysis and predictive modeling to forecast the forthcoming dengue case numbers in Thailand, facilitating better public health preparedness. It also incorporates data visualization for enhanced data exploration. Various forecasting models, including Exponential Smoothing, Polynomial Fitting and Random Forest, are deployed to predict dengue cases within the constraints of our data. This study offers valuable insights into the potential trajectory of dengue cases in Thailand, aiding proactive measures to combat the outbreak.

Keywords: Dengue fever; exponential smoothing model; forecast; mathematical biology; polynomial fitting; public health; random forest; statistical analysis.

[‖]Corresponding authors.

1. Introduction

Mathematical models serve as invaluable tools in portraying and elucidating real-world scenarios, particularly in the domains of ecology, epidemiology and population systems [3, 15–19, 29]. These models play a pivotal role by translating complex situations into a set of variables and equations, with the intention of encapsulating the essential elements of an event while making simplifications or assumptions about specific details. This abstraction allows for a more manageable and comprehensible representation of intricate real-world phenomena. However, the efficacy of these models relies on their ability to accurately mirror reality. Hence, the validation of a mathematical framework becomes imperative. This validation process serves to authenticate the model's representation of real-life scenarios, ensuring that the predictions and insights derived from the mathematical model align closely with the observed phenomena, thereby enhancing the model's reliability and applicability in practical situations [4, 6]. The global impact of the dengue virus, transmitted by mosquitoes, has raised significant concerns in international public health. According to a World Health Organization (WHO) report, an estimated 2.5 billion people live with the virus, with 50–100 million dengue fever cases occurring annually [8]. The first documented dengue outbreak was in Africa in 1779–1780, followed by Asia and North America, underscoring a history of over two centuries in which mosquitoes have been the culprits behind dengue fever [14]. Mosquitoes predominantly inhabit tropical and subtropical regions worldwide, with female mosquitoes, specifically *Aedes albopictus* and *Aedes aegypti*, known as the primary carriers of the dengue epidemic within the Flavivirus family. Numerous mathematical models have emerged to comprehend the transmission patterns of Dengue fever. Esteva and Vargas introduced an SIR model to investigate dengue fever transmission dynamics in scenarios with both constant [12] and varying human populations [27]. The duration of incubation periods in both hosts and vectors plays a crucial role in shaping the transmission patterns of dengue disease. Consequently, various mathematical investigation [9] have been undertaken to examine the transmission dynamics of dengue, with a specific focus on the impact of these incubation periods. Dengue fever, shaped by various climatic and biological factors, necessitates diverse modeling approaches for effective control. Zheng *et al.* [32] highlight the significant roles of temperature and precipitation in dengue transmission in China, emphasizing the need for integrated climate considerations. Olayiwola and Alaje [21] studied host immune responses, showing the importance of CTLs and B-cells in enhancing immunity. Pandey *et al.* [23] analyze dengue transmission dynamics in Nepal, identifying mosquito biting and death rates as critical factors. Naaly *et al.* [20] advocate for a combined strategy of vector control, treatment and mass awareness. Din *et al.* [10] use stochastic modeling to understand dengue's persistence and extinction thresholds.

Thailand, a humid tropical region, is characterized by the presence of *Aedes* mosquitoes, facilitating the transmission of dengue fever. Occasional outbreaks

occur throughout the year, particularly during the rainy season from May to September when there are numerous flood basins and heavy precipitation, creating ideal conditions for mosquito reproduction. The Department of Disease Control of Thailand's Ministry of Public Health (MoPH) has reported a significant dengue fever pandemic in the past two years: 34,467 cases with 41 fatalities in 2022 and 56,547 cases with 101 deaths in January to August 2023. Bangkok, the most populous city in the densely populated metropolitan area, centralizes the disease's propagation. Consequently, dengue fever is widespread in Thailand, falling under the category of surveillance diseases characterized by high morbidity and mortality rates, extensive transmission, and a statistical tendency for further expansion. Cases have been documented in every province and territory, indicating nationwide spread. Prompt identification and effective management can reduce dengue fever-related fatalities among both children and adults, as suggested by the Centers for Disease Control and Prevention [7, 24]. In this context, Aguiar *et al.* emphasized the importance of accurately modeling dengue fever datasets to minimize false predictions and enhance the reliability of forecasting models [2]. Additionally, Steindorf *et al.* explored the effects of general cross-immunity protection and antibody-dependent enhancement on dengue dynamics, offering valuable insights into the complex factors influencing disease transmission and prediction [28]. Wongkoon *et al.* developed temporal modeling techniques for predicting dengue infections in Northeastern Thailand, underscoring the importance of temporal dynamics in forecasting dengue outbreaks [30]. Rahman *et al.* employed machine learning approaches to map the spatial distribution and predict the abundance of the dengue vector *A. aegypti* in Northeastern Thailand, highlighting the role of advanced computational techniques in vector control [25]. Gangula *et al.* utilized ensemble machine learning methods to enhance the accuracy of dengue disease predictions, demonstrating the potential of integrating multiple models to improve forecasting performance [13]. Finally, Aguiar *et al.* provided a comprehensive review of mathematical models for dengue fever epidemiology, reflecting on the evolution and advancements in modeling approaches over the past decade [5].

In this analysis, we have utilized a dataset containing records of dengue cases in Thailand over multiple years, documenting the number of reported cases. Our objective is to conduct a comprehensive data exploration by employing various data visualization techniques. The ensuing visual representations provide diverse perspectives on the dataset, offering insightful and informative ways to interpret the information. As the field of quantitative research continues to evolve, the integration of statistical software (SS) has become indispensable for proficient data analysis. Researchers are transitioning from traditional manual paper-based analyses to more efficient digital/electronic methods utilizing SS. This shift is imperative to meet the requirements for conducting high-quality studies using contemporary SS solutions [1]. In this study, we employ various SS tools, including Excel, MATLAB and Python packages, for exploration, visualization and predictive purposes. Despite its limited capabilities and occasional confusing outputs, Microsoft Excel

is utilized for statistical analysis and visualization, proving suitable for preliminary analysis due to data constraints [11]. It holds the distinction of being the most commonly used software for quantitative data analysis, with 92% awareness among respondents [22]. In the realm of mathematical modeling, MATLAB assumes a crucial role, particularly when handling large datasets for analysis, processing and tasks that are challenging to complete manually. Its robust numerical calculation capabilities and diverse graphics toolkit functionalities efficiently address mathematical modeling challenges across various domains [31]. Python, recognized as an advanced programming language, has experienced substantial growth, particularly in the fields of data science and analytics. Its accessibility, extensive library support and overall user-friendliness have established it as the preferred language for data analysts, scientists and academics. Python's adaptability, simplicity and stability position it as a prime language for developing cutting-edge applications and implementing machine learning techniques [26]. In this study, Python packages such as NumPy, Pandas, Matplotlib, Seaborn and scikit-learn are employed for data analysis, visualization and forecasting purposes. This paper is divided into three sections. The first introduces the origin and impact of dengue fever, data sources, analysis methods and objectives. Section 2 analyzes the data, uncovering insights using visualization tools. Section 3 explores the application of Exponential Smoothing, Polynomial Fitting and Random Forest models as predictive tools within the context of limited data, which lacks seasonality and clear patterns. Finally, the last section concludes the paper by summarizing the key insights and contributions along with suggestions for future enhancements.

Main contributions of this paper are as follows:

- This paper scrutinizes real-life data, specifically concentrating on dengue cases in Thailand, with a focus on visualizing the data using Python language libraries, including Seaborn.
- Extensive data comparisons within the dataset are presented through the utilization of violin plots, offering a detailed examination of dengue cases in Thailand.
- Three distinct forecasting techniques were explored for predicting dengue cases, including the introduction of the Exponential Smoothing model. The upper and lower ranges of predicted dengue cases were determined to enhance forecasting accuracy.
- The paper advocates for the adoption of Polynomial Fitting and Random Forest models as additional forecasting techniques for dengue cases, contributing to a diversified and comprehensive approach to predictive modeling in this context.

2. Analyzing Dengue Case Data: Uncovering Insights and Patterns

In this section, we embark on a comprehensive Analysis of Dengue Case Data, unveiling crucial patterns and insightful revelations. This endeavor involves the succinct summarization of fundamental dataset characteristics, employing a blend

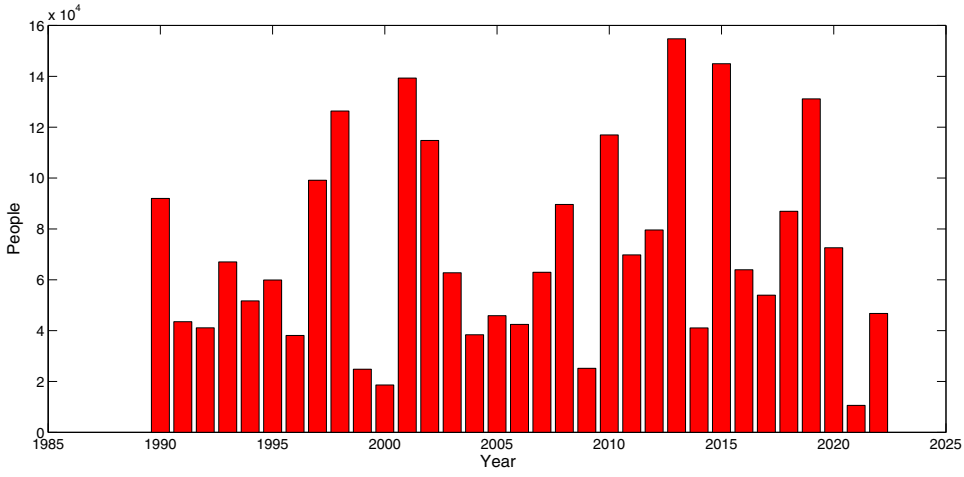

Fig. 1. Dengue cases in Thailand.

of visual and statistical approaches. Figure 1 showcases a bar graph, a remarkably effective and user-friendly way to depict our data. This particular bar chart is employed to illustrate the incidence of dengue cases in Thailand from 1990 to 2022. Notably, the graph reveals that the highest number of dengue cases occurred in 2013, while the number of cases dropped significantly in 2021. In this visual presentation, each individual bar corresponds to the total count of dengue fever cases recorded for a specific year. What sets the bar graph apart is its straightforwardness and precision in conveying this information. This format allows for an immediate and uncomplicated evaluation of which years saw higher or lower case counts, facilitating the identification of significant spikes or declines. The graph unequivocally highlights that 2013 stands out with a significantly higher number of cases compared to all other years in the dataset. In stark contrast, 2021 reports the lowest number of cases of all. This observation underscores a marked disparity in the incidence of dengue fever over the years.

Figure 2 displays below serves to highlight the maximum points, or peaks, within a dataset. The conspicuous red crosses in the illustration pinpoint these noteworthy peaks. Peaks in the data are of particular significance as they typically signify extreme values or points of exceptional interest within the dataset, often holding substantial implications or warranting further analysis.

Our primary objective is to enrich our understanding of the data and the intricate patterns it holds, while also identifying outliers and anomalies that can provide invaluable insights for subsequent analyses and modeling efforts. To achieve this, we harness a diverse range of visualization techniques, including the robust Box Plots, illuminating Empirical Cumulative Distribution Function (ECDF) Plots and versatile Violin Plots. These visual aids serve as powerful tools in our exploration, empowering us to dissect the data and extract meaningful knowledge to inform

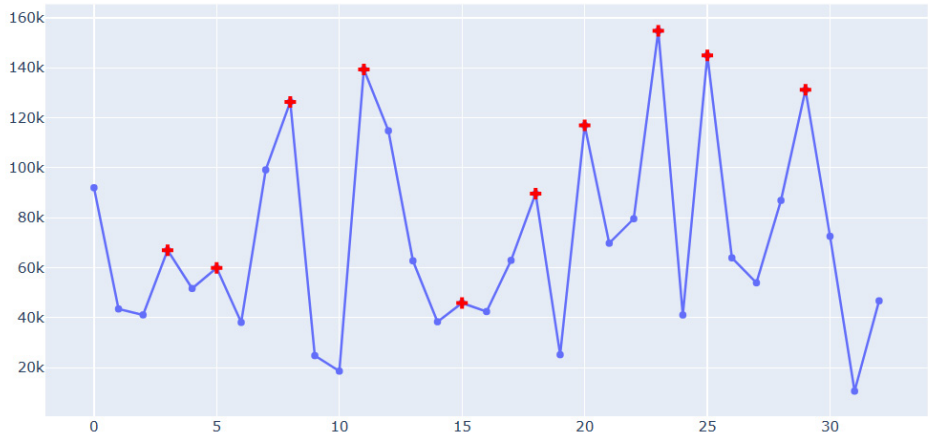

Fig. 2. Visualization of dengue cases' peak values.

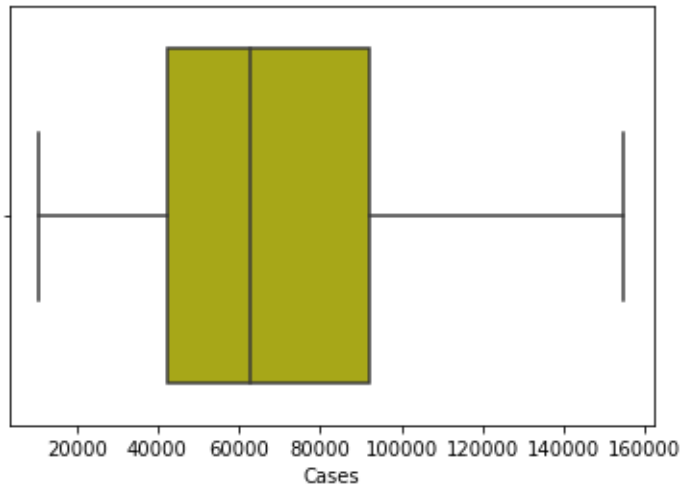

Fig. 3. Visualizing dengue case distribution with box plots.

decision-making and further research. Figure 3 showcases dengue cases using a box plot visualization. A box plot, also referred to as a box and whisker chart, is a valuable tool for understanding key characteristics of a statistical dataset, including its shape, variability and center, typically represented by the median. This visualization technique is especially beneficial when dealing with skewed data distributions. In the context of Fig. 3, the dengue cases data exhibits a right-skewed distribution, where a few extreme values are pulling the overall data to the right. This is evident because the median number of cases, which is 62,949, is lower than the mean number of cases, which stands at 71,417.

The box plot provides additional insights into the distribution of cases. The left side of the box, representing the lower number of cases, is notably shorter than the right side, which corresponds to the higher number of cases. This discrepancy indicates that the lower number of cases are clustered more closely together, while the higher number of cases are more spread out. This is a characteristic feature of right-skewed data. Furthermore, the position of the median relative to the center of the box is another important aspect of interpretation. In symmetric datasets, the median would be approximately in the center of the box. In this case, the off-center positioning of the median within the box further confirms that the data distribution is asymmetric, aligning with our earlier observation of a right-skewed distribution. This asymmetry in the data is essential information for researchers and analysts as it helps guide more nuanced statistical and epidemiological assessments, which may be critical for addressing and managing dengue cases effectively.

A violin plot combines the characteristics of a box plot and a kernel density plot, allowing for a comprehensive representation of data distribution, especially highlighting peaks in the dataset. Its primary purpose is to provide a visual depiction of the numerical data's distribution. In contrast to a traditional box plot, which primarily offers summary statistics, violin plots not only convey these statistics but also present the data's density, revealing more detailed insights into variable distributions. Violin plots encompass several key summary statistics akin to those found in box plots. Within the violin plot, the white dot serves as a visual indicator for the median, while the prominent, central gray bar graphically signifies the interquartile range. Adjacent to this bar, a delicate gray line extends to showcase the broader data distribution. On either side of this central gray line, we observe kernel density estimations illustrating the shape of the data's distribution. Notably, the violin plot's width dynamically conveys the probability associated with specific data values. Wider segments correspond to a higher likelihood that data points within the population will assume the given value, whereas narrower sections suggest a lower probability.

Figure 4 presents data categorized by the annual count of dengue cases. It's essential to observe that only one axis, specifically the x-axis, is populated, displaying the total number of dengue cases for each year. The distribution's shape, with pronounced narrowness at both extremes and considerable width in the central region, signifies a high concentration of dengue cases around the median value. To deepen our understanding, we will organize the data into three distinct sets: Set A, Set B and Set C. The first 10 rows will be assigned to Set A, the next 10 rows to Set B and the remaining rows will comprise Set C. This deliberate segmentation creates the foundation for a comprehensive analysis that harmoniously blends the data with meaningful visual representations. Through this synergy of data and figures, we equip ourselves with the tools to delve into the complex network of relationships and connections embedded within this multifaceted dataset. This approach nurtures a holistic exploration, allowing us to unravel the intricate interplay and influence among the various elements.

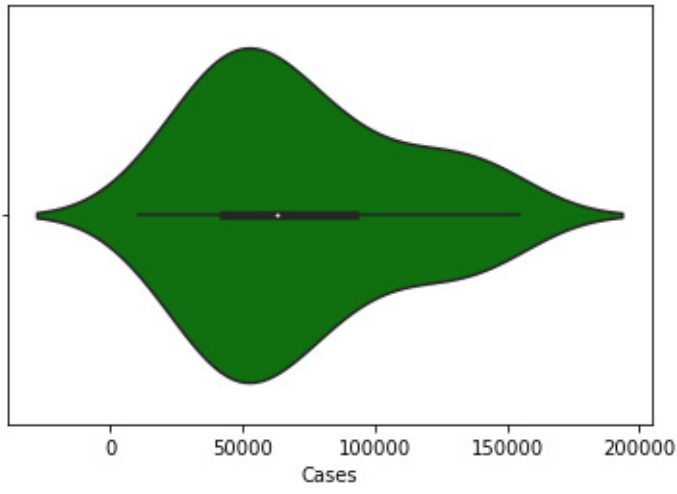

Fig. 4. Dengue cases distribution illustrated with violin plot.

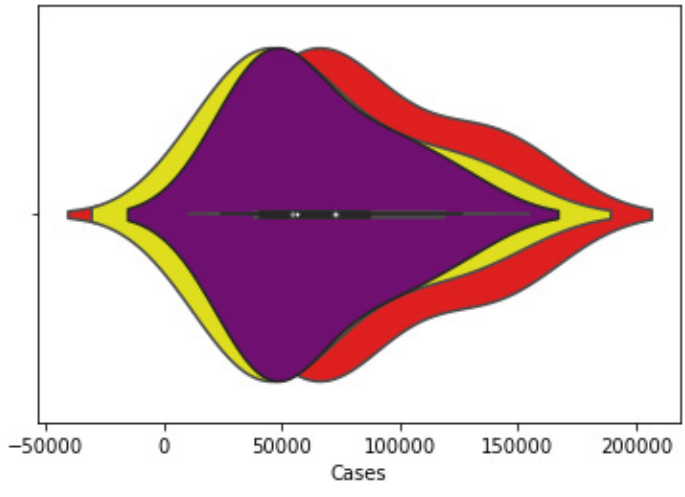

Fig. 5. Visualizing dengue cases across different Sets A, B, C using violin plots.

The depicted Fig. 5 provides a unified visual presentation of violin plots for Sets A, B and C, where each plot overlays the others. What immediately catches the eye is the marked divergence of Set C, represented in red, from the characteristic patterns observed in Sets A and B. Set C's representation exhibits a notably elongated distribution, noticeably distinct from the other two sets. This elongation results in a more pronounced and defined peak in the data distribution for Set C. Furthermore, it's essential to note that Set C stands out not only due to its unique peak but also because it features long tails in the distribution. These extended tails imply that Set C contains a greater range of data values, both higher and lower,

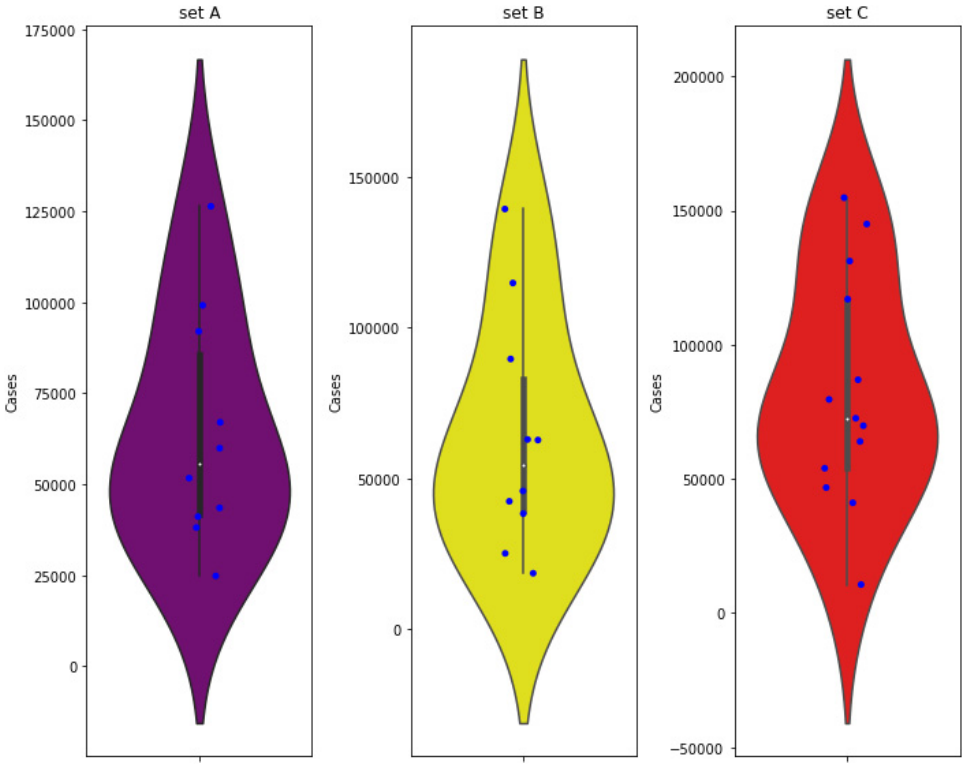

Fig. 6. Separate visualization of dengue cases across distinct sets with violin plots.

compared to Sets A and B. Given these distinctive characteristics of Set C, isolating it in separate visualizations would enable a more comprehensive examination of its individual properties, providing deeper insights into the dataset's complexities.

The provided visual representation showcases violin plots for three distinct sets: Set A, Set B and Set C, each uniquely color-coded. In Fig. 6 representation, the purple violin signifies Set A, the yellow represents Set B and the red corresponds to Set C. These visualizations also include blue data points within the violins, which denote the annual dengue case counts for each set. One notable observation pertains to the medians. Set C stands out as it subtly deviates from Sets A and B. Sets A and B appear to share a similar range of data values, as evidenced by their violin plots. This similarity suggests that the data in Sets A and B falls within a comparable range. On the other hand, Set C exhibits a broader range of data values, spanning both higher and lower values when compared to Sets A and B. This implies that Set C encompasses a more extensive spectrum of data, making it more diverse in terms of the data values it contains. As a result, Set C is notably distinct from Sets A and B in both the range and quantity of data values it represents, with Set A containing the fewest data values among the three sets. This additional context provides a deeper understanding of the dataset's composition and variations.

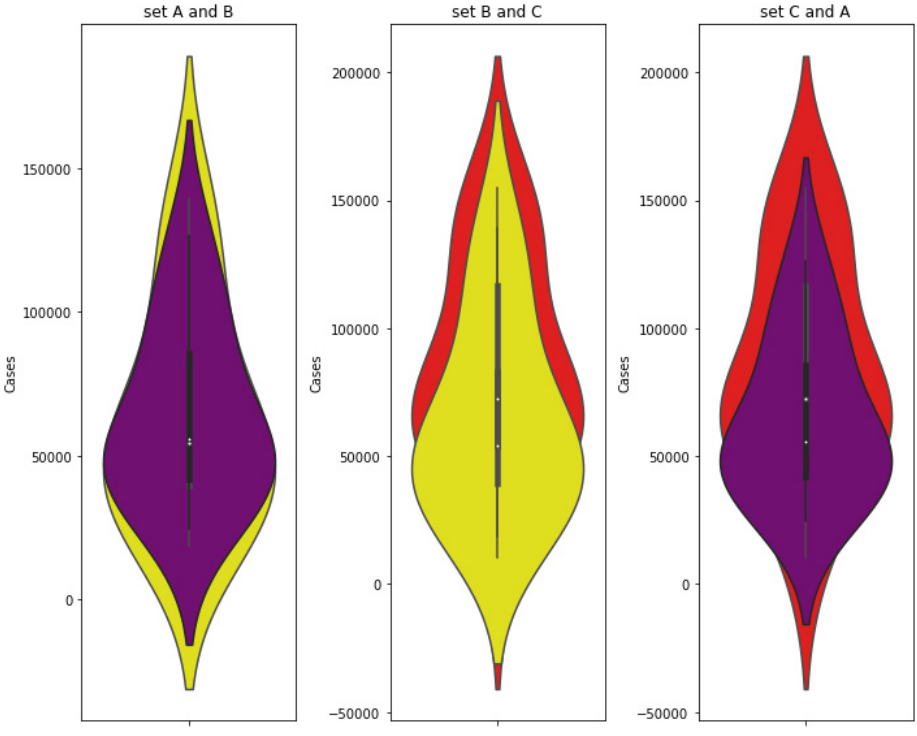

Fig. 7. Comparing two sets with violin plots: A visual analysis.

The distribution shapes within Sets A and B exhibit a pronounced similarity, revealing nearly identical patterns in their data distributions. An evident indicator of this similarity can be gleaned from Fig. 7, where the two white dots representing the medians in Sets A and B are aligned in close proximity to each other. This alignment indicates that the median values for these sets fall within the same range. Conversely, when comparing Sets A and C, as well as Sets B and C, a striking difference in median values is readily discernible. The visual presentation clearly illustrates that Set C stands apart from both Sets A and B in terms of median values, emphasizing the considerable variation in central tendencies between these three sets.

Figure 8 is a visual representation aimed at comparing the total cases within each set. A discernible pattern emerges from this comparison: Sets A and B closely resemble the total case distribution, indicating a similar structural composition. In contrast, Set C stands out as distinct from the others, suggesting that it possesses a unique data structure not mirrored in Sets A and B. In this context, it is noteworthy that Sets A and B play a pivotal role in shaping and contributing to the comprehensive distribution of the total dataset. The data within Sets A and B, when combined, serves as the cornerstone upon which the overall distribution of the dataset is built. Their collective contributions significantly influence and define

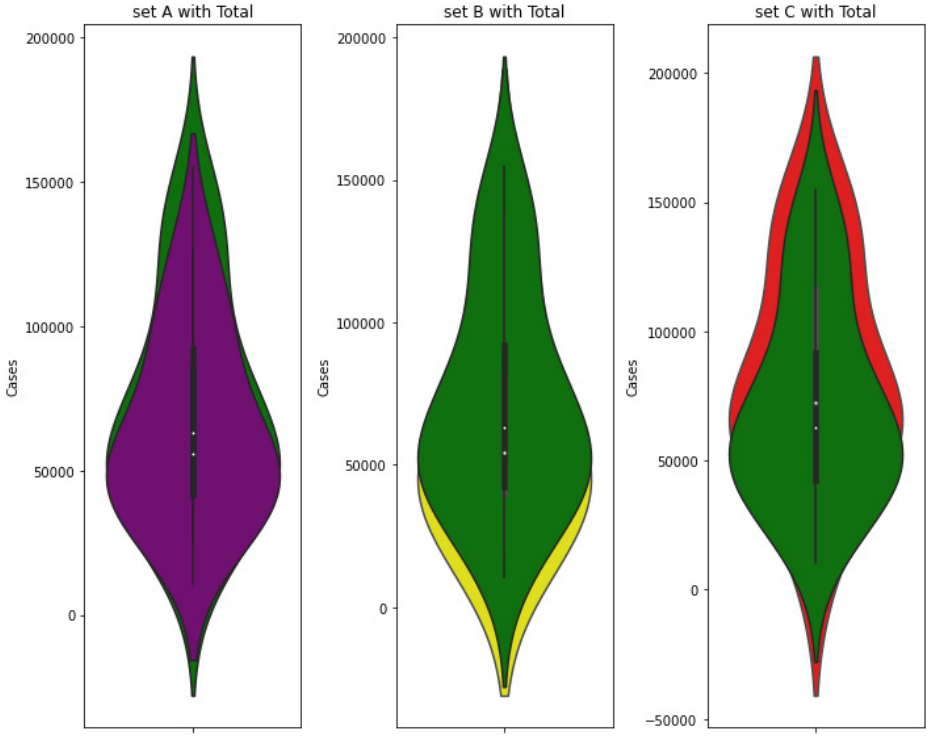

Fig. 8. Comparative violin plots for total dengue cases across different sets.

the fundamental characteristics of the entire dataset's distribution, showcasing the central role played by Sets A and B in representing the dataset's overall structure and patterns.

The ECDF is a data visualization tool that estimates the Cumulative Distribution Function (CDF). It arranges data in ascending order, providing a clear depiction of how the data feature is distributed within the dataset. The ECDF illustrates the proportion or count of observations below each unique value without requiring adjustments like binning or smoothing. Its direct visualization of individual observations makes it a powerful tool for understanding data distributions and patterns. Visualizing the data through an ECDF plot can be a more accessible way to comprehend the variations in the case counts. The ECDF's slope offers valuable insights into the range of values within the dataset: a steeper curve indicates a narrower range, while a gentler slope signifies a broader span. The presented Fig. 9 serves as an invaluable tool for directly discerning fundamental features of the distribution of dengue cases. A prominent feature of this plot is its gentle slope, which suggests a broader span of cases, indicative of a wide range in the counts of reported dengue cases. In simpler terms, the data displays a substantial variation in the number of cases. For example, as we examine the plot, we notice that approximately 60% of

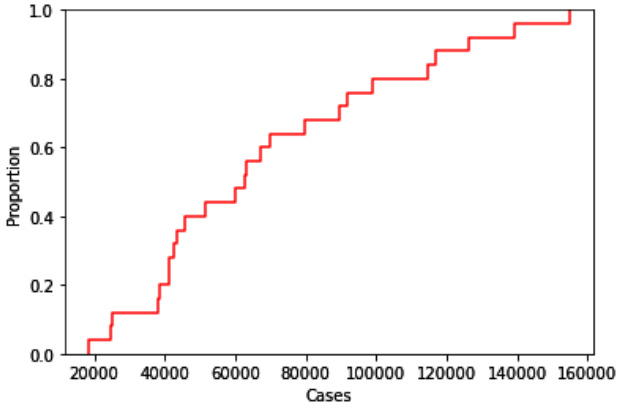

Fig. 9. ECDF plot for dengue cases.

the cases in the dataset register at numbers below 80,000. This implies that the majority of observations fall within this range, and it gives us a sense of the central tendency in the data. The median, which corresponds to the point where 50% of the data falls below and 50% above, is noted at 62. This serves as a key reference point in understanding the dataset's center. Furthermore, it's interesting to note that only about 20% of the cases have values under 40,000. This provides a glimpse into the portion of cases that have relatively lower counts. In summary, this plot allows for the immediate extraction of critical insights, providing an overview of the distribution's span, central tendency and the proportion of cases within specific count ranges.

3. Forecasting Dengue Cases: Employing Three Distinct Approaches

The process of forecasting is a vital technique that revolves around the prediction of future trends, occurrences or outcomes by drawing insights from historical data and conducting in-depth pattern and information analysis. This methodology serves as a fundamental tool employed across diverse industries, including finance, weather forecasting, economics, business management and epidemiology, among others. In this particular context, our focus lies in Forecasting Dengue Cases through the application of three distinct and innovative methods. Our chosen approaches encompass the utilization of the Exponential Smoothing model, Polynomial Fitting and Random Forest models. By combining these methods, we aim to address the diverse characteristics of dengue case data, enhancing both the precision and robustness of our predictions. These techniques collectively empower us to navigate the complexities of dengue case predictions, leveraging the strengths of each method to enhance the accuracy and reliability of our forecasts. The fusion of these methods represents a significant step forward in our pursuit of understanding and anticipating dengue trends, which can have far-reaching implications for public health

and policy decisions. Furthermore, by integrating these diverse forecasting methods, we aim to capture different aspects of the data's behavior, thereby offering a more comprehensive and nuanced prediction model. This multifaceted approach not only enhances our ability to forecast dengue cases but also provides valuable insights that can guide targeted interventions and resource allocation in affected regions.

3.1. *Exponential Smoothing model*

In Fig. 10, forecasting is done using Excel software. Excel software uses Exponential Smoothing model to deal with prediction of time-related data. Exponential Smoothing is a forecasting method for univariate time series data. Input data is number of dengue cases over the years from 1990 to 2022. The objective is to predict the number of dengue cases in future years. Here, we used data up to 2021 and included prediction of 2022, to compare it with actual data in 2022. Figure 10 illustrates the forecasted number of dengue cases over the years from 2022 to 2029. The 95% confidence interval of forecasted values is also shown it.

Table 1 compares the actual and predicted number of dengue cases in 2022. Here, the actual number of dengue cases is 46,755 and predicted number is 45,248. The residual is 1507.

The forecast Table 2 lists out the forecasted number of dengue cases over the years from 2023 to 2029 along with their 95% confidence intervals.

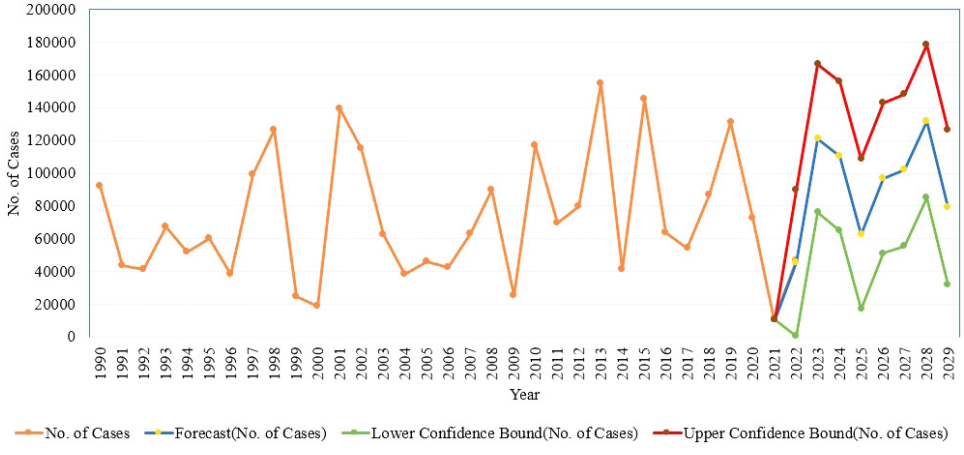

Fig. 10. Exponential Smoothing model.

Table 1. Comparison.

Year	No. of cases	Forecast	Lower confidence bound	Upper confidence bound
2021	10,617	10,617	10, 617	10, 617
2022	46,755	45,248	690.91	89, 804.26

Table 2. Forecasted number of dengue cases.

Year	Forecast	Lower confidence bound	Upper confidence bound
2023	121, 269	76,354	166,184
2024	110, 430	65,155	155,705
2025	62, 787	17,148	108,425
2026	97, 009	51,004	143,014
2027	101, 957	55,583	148,330
2028	131, 477	84,732	178,222
2029	79, 182	32,063	126,301

3.2. *Polynomial fitting*

Polynomial Fitting is a statistical method used to model the relationship between a dependent variable and one or more independent variables using polynomial functions. Polynomial equation is used to approximate the relationship between variables.

The general form of polynomial equation is

$$y = f(x) = q_n x^n + \cdots + q_2 x^2 + q_1 x + q_0,$$

where

- y is dependent variable or response variable.
- x is independent variable or predictor.
- $q_0, q_1, q_2, \ldots, q_n$ are the coefficients. Value of coefficients is estimated through Polynomial Fitting process.
- n is the degree of the polynomial.

The objective of Polynomial Fitting is to find the values of coefficients that minimize the residual of dependent variable. Residual is the difference between observed values of the dependent variable and the predicted values from the polynomial equation. Polynomial Fitting is useful when the relationship between variables exhibits a nonlinear pattern. In this section, Polynomial Fitting is done using MATLAB software. Input data is number of dengue cases over the years from 1990 to 2022. Here, the dependent variable is number of dengue cases and independent variable is corresponding year. In Polynomial Fitting, it is essential to choose appropriate degree of the polynomial to obtain more accurate predictions. Here, Polynomial Fitting is done up to degree 4 and compare the most accurate fitting.

The measures of goodness of fit evaluated in this section are Sum of Squared Errors (SSEs), coefficient of determination (R-square), Degrees of Freedom for Error (DFE) and Root Mean Squared Error (RMSE).

• SSE: SSE is a metric to evaluate the goodness of fit in Polynomial Fitting. It measures the difference between observed and predicted values of dependent variable based on the model.

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2,$$

where

- n is the number of data points.
- y_i represents the actual value of the dependent variable for the ith data point.
- $\widehat{y}_i$ represents the predicted value of the dependent variable for the ith data point.

SSE is used to quantify how well the polynomial model fits the observed data. A lower SSE means predictions by model are closer to actual data points. Conversely, a higher SSE indicates model does not fit the data well and there is huge difference between predicted and actual values.

• Coefficient of Determination: Coefficient of determination, denoted as R^2 is a tool for assessing goodness of fit of model. It represents the proportion of variance in the dependent variable explained by the independent variable. R^2 value ranges from 0 to 1. R^2 value close to 0 indicates poor fit whereas R^2 close to 1 indicates good fit.

• DFE: It represents the difference between the number of data points and number of parameters estimated in the model.

DFE = No. of data points–No. of parameters estimated in model.

DFE is used in the calculation of evaluation metrics like Mean Squared Error (MSE) with lower MSE indicating best fit. The equation of MSE is as follows:

$$\text{MSE} = \frac{\text{RSS}}{\text{DFE}},$$

where RSS stands for Residual Sum of Squares

$$\text{RSS} = \sum_{i=1}^{n} ((\text{Residual})_i)^2.$$

DFE stands for Degrees of Freedom for Error. Lower MSE indicates better fit of model to data.

• RMSE: RMSE is an evaluation metric used to asses the goodness of fit of model to the data. It is the square root of MSE. RMSE provides an idea of how well the polynomial model predicts future data. A good prediction model has lower RMSE along with considering cautions to avoid overfitting.

$$\text{RMSE} = \sqrt{\text{MSE}},$$

where MSE is Mean Squared Error.

Figure 11 illustrates Polynomial Fitting of degree 1 and corresponding residuals. The Polynomial Fitting is performed to analyze the relationship between the variables, showcasing the models of varying complexity. The residuals provide insight into the differences between the fitted curves and the actual data points, aiding in the assessment of the model's accuracy and appropriateness for the given dataset. Original data points are way more scattered from fitted curve.

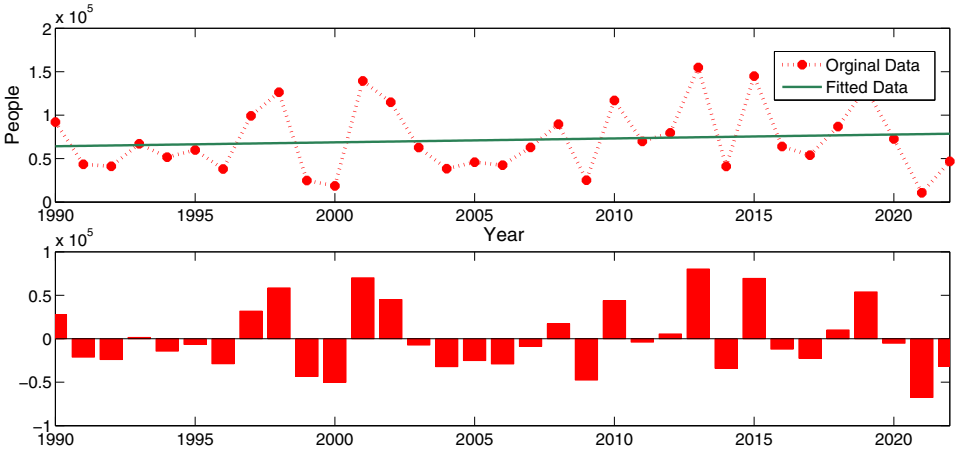

Fig. 11. Polynomial Fitting of degree 1 and residuals.

Table 3. Coefficients and 95% confidence bounds of linear polynomial.

Coefficients	Value	Lower	Upper
q_1	449.7911	−1.0097e+03	1.9093e+03
q_2	−8.3086e+05	−3.7586e+06	2.0968e+06

Corresponding polynomial equation is

$$f(x) = q_1 x + q_2. \tag{1}$$

Coefficients q_1 and q_2 are estimated through Polynomial Fitting. Table 3 shows the estimated values of coefficients and their confidence intervals. Figure 12 depicts the polynomial fit of degree 1 along with 95% confidence interval. The polynomial equation (1) with estimated values of coefficients is then used to predict the number of dengue cases in future years.

The goodness of fit for the polynomial model (1) is assessed using several metrics. SSE is 4.7496×10^{10} which indicates predictions made by polynomial model (1) has large deviations with actual data. R-square value = 0.0126 shows that only 1.26% of variations in the number of dengue cases is explained by year. DFE is 31 since two parameters are estimated in this model (1). RMSE value is 3.9142×10^4. It clearly indicates the model (1) provided by Polynomial Fitting of degree 1 is not a good prediction model.

Figure 13 illustrates Polynomial Fitting of degree 2.

Corresponding polynomial equation is

$$f(x) = q_1 x^2 + q_2 x + q_3. \tag{2}$$

Table 4 shows the estimated values of coefficients and their confidence intervals. Figure 14 visualizes the polynomial degree 2 fit with 95% confidence bounds.

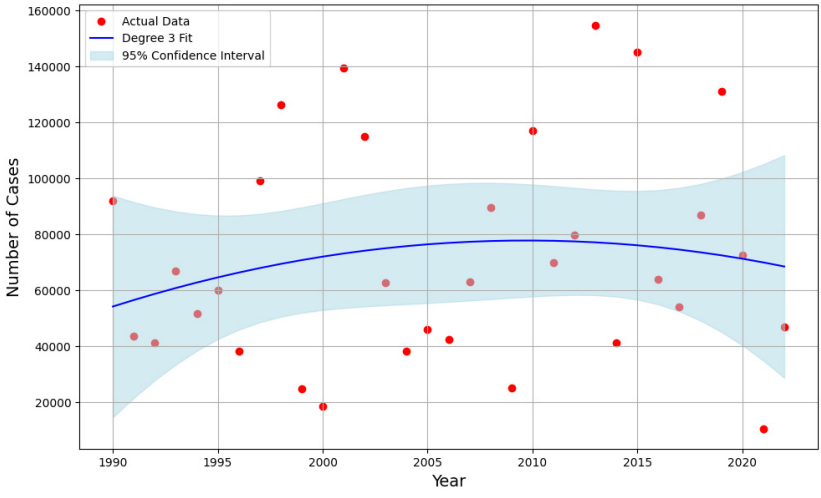

Fig. 12. Polynomial degree 1 fit with 95% confidence bounds.

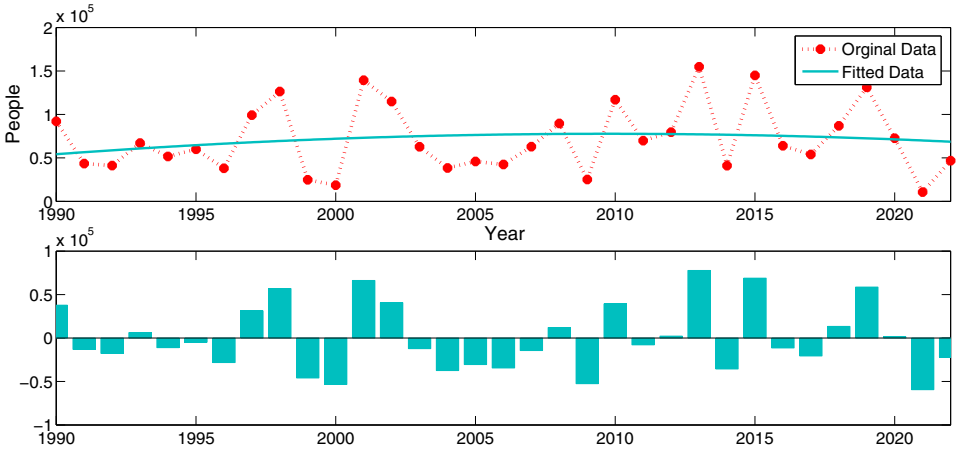

Fig. 13. Polynomial Fitting of degree 2 and residuals.

The polynomial equation (2) with estimated values of coefficients is then used to predict the number of dengue cases in future years.

The goodness of fit for the polynomial model (2) is assessed using several metrics. SSE is 4.6704×10^{10} which indicates predictions made by polynomial model (2) has large deviations with actual data. R-square value = 0.0290 shows that only 2.9% of variations in the number of dengue cases is explained by year. DFE is 30 since three parameters are estimated in the model (2). RMSE value is 3.9457×10^4. It clearly indicates the model (2) provided by Polynomial Fitting of degree 2 is not a good prediction model.

Figure 15 illustrates Polynomial Fitting of degree 3.

Table 4. Coefficients and 95% confidence bounds of quadratic polynomial.

Coefficients	Value	Lower	Upper
q_1	−60.4673	−233.6810	112.7464
q_2	2.4304e+05	−4.5189e+05	9.3798e+05
q_3	−2.4415e+08	−9.4116e+08	4.5286e+08

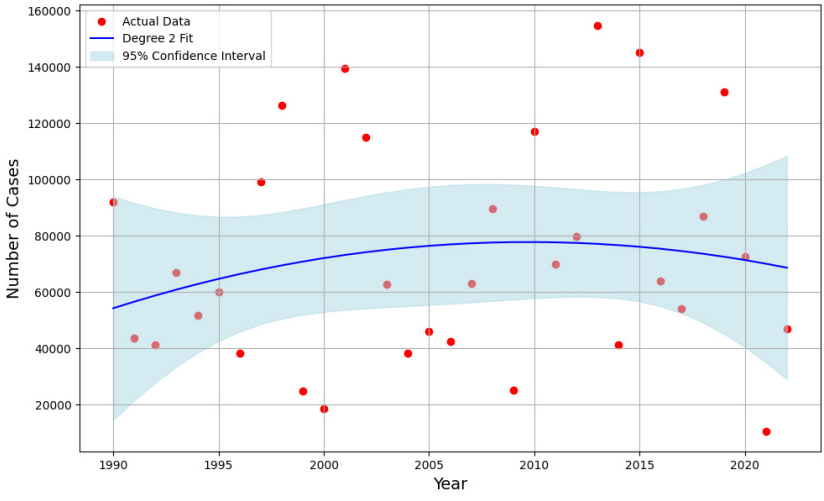

Fig. 14. Polynomial degree 2 fit with 95% confidence bounds.

Corresponding polynomial equation is

$$f(x) = q_1 x^3 + q_2 x^2 + q_3 x + q_4. \tag{3}$$

Table 5 shows the Estimated values of coefficients and their confidence intervals. Figure 16 portrays the polynomial Fitting of degree 3 with 95% confidence interval. The polynomial equation (3) with estimated values of coefficients is then used to predict the number of dengue cases in future years.

The goodness of fit for the polynomial model (3) is assessed using several metrics. SSE is 4.4750×10^{10} which indicates predictions made by polynomial model (3) has slightly less error compared to predictions made by models (1) and (2). However, since SSE is still a large value so that the model does not fits the data well. R-square value = 0.0697 shows that only 6.97% of variations in the number of dengue cases is explained by year.

After fitting polynomials of degrees 1, 2 and 3, we discuss the forecasting using these models. Figure 17 illustrates the polynomial forecasts, with the corresponding results summarized in Table 6. The performance of each polynomial model

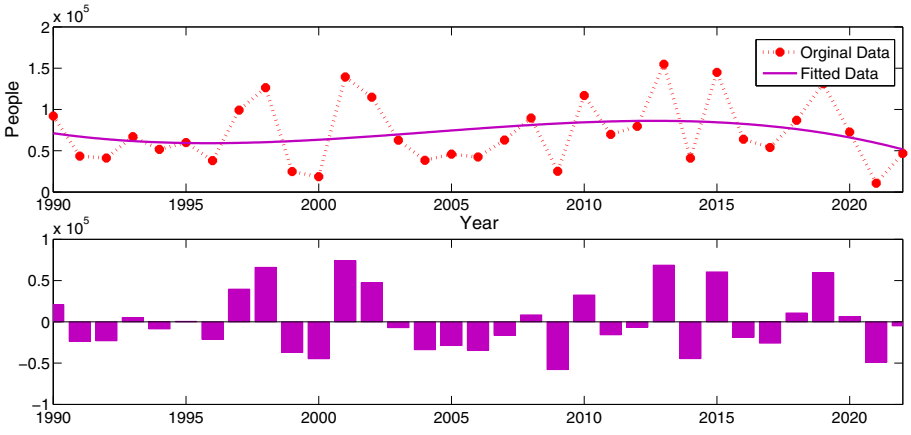

Fig. 15. Polynomial Fitting of degree 3 and residuals.

Table 5. Coefficients and 95% confidence bounds of cubic polynomial.

Coefficients	Value	Lower	Upper
q_1	−11.4047	−32.1309	9.3216
q_2	6.8573e+04	−5.6158e+04	1.9330e+05
q_3	−1.3743e+08	−3.8764e+08	1.1277e+08
q_4	9.1813e+10	−7.5488e+10	2.5911e+11

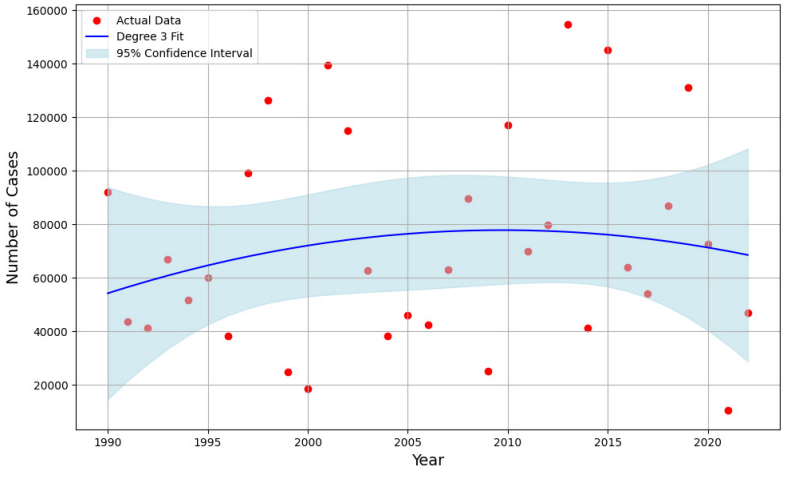

Fig. 16. Polynomial degree 3 fit with 95% confidence bounds.

is assessed based on its accuracy and ability to capture trends in the data. By comparing the residual errors and the goodness-of-fit metrics, we determine which polynomial degree provides the most reliable forecasts. Additionally, the limitations

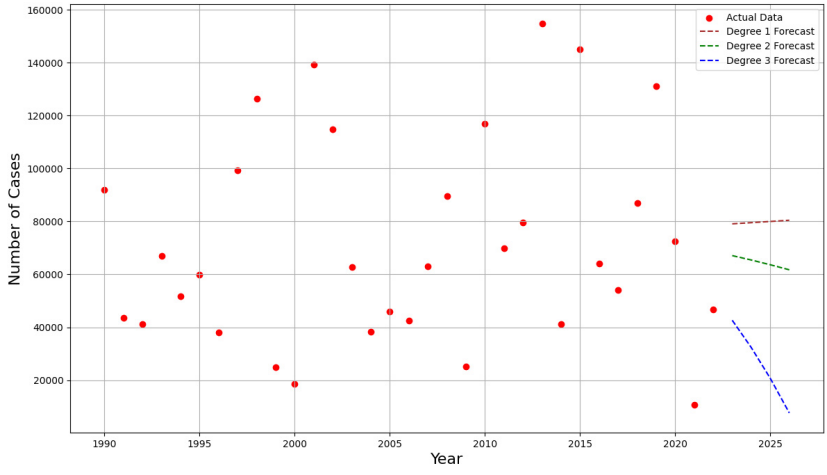

Fig. 17. Polynomial forecast.

Table 6. Polynomial forecast.

Year	Degree 1	Degree 2	Degree 3
2023	79,063	67,071	42,642
2024	79,513	65,404	32,353
2025	79,963	63,616	20,712
2026	80,413	61,708	7650

of higher-degree polynomials, such as overfitting, are considered in the context of forecasting. This analysis provides insights into the trade-offs between model complexity and predictive performance, guiding the selection of the most appropriate model for future predictions.

3.3. *Random Forest*

In this section, Random Forest model is used to make predictions on the number of dengue cases in upcoming years. Random Forest is an ensemble learning method that builds multiple decision trees during training. Each tree is constructed from a subset of the data and features, and the final prediction is derived by averaging the predictions of all individual trees. This technique enhances predictive accuracy and mitigates overfitting by aggregating the results of multiple trees. The Random Forest model's ability to handle nonlinear relationships and interactions between variables makes it particularly useful for forecasting tasks in complex datasets such as those involving disease incidence. By leveraging historical data on dengue cases and associated features, the model generates forecasts that account for historical trends and potential variability.

Figure 18 illustrates the forecast generated by the Random Forest model. The graph displays the predicted number of dengue cases from 2023 to 2026.

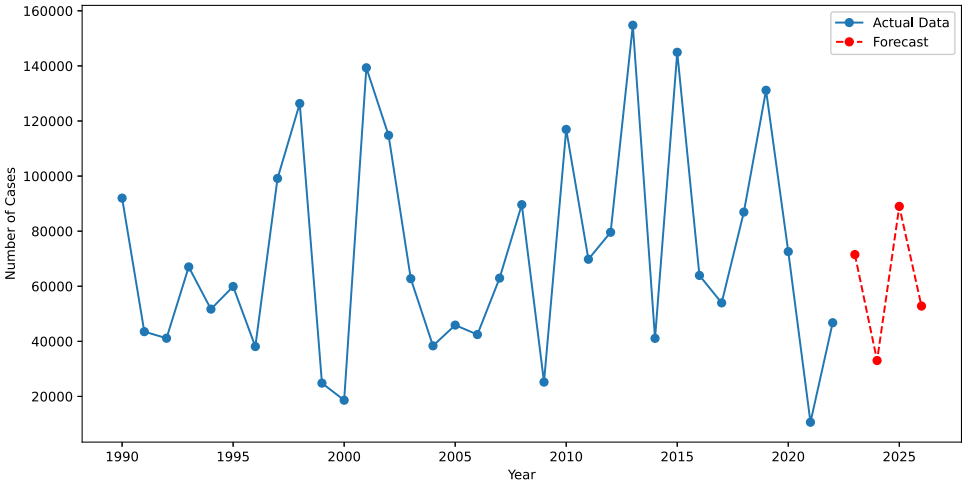

Fig. 18. Random Forest forecast.

Table 7. Random Forest forecast.

Year	Forecasted Value
2023	71,514
2024	33,028
2025	88,995
2026	52,804

The forecasted values show a range of potential outcomes, reflecting the model's consideration of historical trends and inherent variability in the data. This range provides a nuanced view of future dengue case numbers, indicating periods of both high and low incidence.

Table 7 provides the forecasted values for the number of dengue cases from 2023 to 2026. The predictions suggest significant fluctuations in dengue incidence over these years. Specifically, the forecast indicates a potential increase in cases in 2025, with a decrease in 2024 and a resurgence in 2026. These variations highlight the model's capacity to predict both peaks and troughs in dengue cases, which can be crucial for public health planning and intervention strategies.

4. Conclusion and Future Direction

In this study, a comprehensive exploration of predictive models was conducted to anticipate the future dengue case numbers in Thailand. The models evaluated include Exponential Smoothing model, Polynomial Fitting and Random Forest model, each offering a unique approach to prediction under the constraints of limited and patternless data. The effectiveness of these models was rigorously evaluated

through the use of various evaluation metrics. Our research leveraged multiple software tools, including MATLAB, Python and Excel, to both explore the dataset and to develop the forecasting models. Recognizing the constraints posed by our limited data, we made earnest efforts to construct a dependable predictive model. While our models represent a significant step toward understanding and predicting dengue outbreaks, it is crucial to acknowledge the inherent uncertainties and complexities associated with infectious disease dynamics. Further research and data collection will be essential to refine and enhance the accuracy and reliability of these predictive models, ultimately contributing to more effective public health preparedness and response strategies in Thailand and beyond.

In our future work, we envision enhancing the depth and scope of our dengue case analysis by incorporating additional parameters into the dataset. This expansion aims to facilitate a more comprehensive analysis, ultimately leading to improved predictions. One prospective avenue involves obtaining region-specific data, enabling a nuanced examination of the geographical distribution of dengue cases in Thailand. By stratifying the data regionally, we anticipate gaining insights into the areas most severely impacted by dengue. This regional breakdown can serve as a valuable tool in identifying key factors influencing the prevalence of dengue cases. Understanding the localized dynamics of the disease will aid in pinpointing specific risk factors and potentially contribute to the development of targeted interventions and preventive measures. This future work holds promise for not only refining our predictive models but also for contributing valuable information to public health efforts. The identification of specific factors driving dengue outbreaks in particular regions can inform targeted interventions and preventive measures, thus fostering a more effective approach to dengue control and mitigation in Thailand.

Acknowledgments

The researcher expresses deep gratitude for the valuable help extended by the Research and Development Institute. The research fund and substantial support were received from Phuket Rajabhat University, for which the researcher is greatly appreciative.

ORCID

Sayooj Aby Jose https://orcid.org/0000-0003-4437-1623
Karuna Mathew https://orcid.org/0009-0007-4397-4904
Hamna Mariyam K. B. https://orcid.org/0009-0006-2207-0765
Anuwat Jirawattanapanit https://orcid.org/0000-0002-6319-0214
Anurak Weraprasertsakun https://orcid.org/0009-0007-3783-848X

References

[1] S. M. Abatan and M. Olayemi, The role of statistical software in data analysis, *Int. J. Appl. Res. Stud.* **3**(8) (2014), https://ssrn.com/abstract=2532326.

[2] M. Aguiar *et al.*, Are we modelling the correct dataset? Minimizing false predictions for dengue fever in Thailand, *Epidemiol. Infect.* **142**(11) (2014) 2447–2459.

[3] N. Anggriani, H. S. Panigoro, E. Rahmi, O. J. Peter and S. A. Jose, A predator–prey model with additive Allee effect and intraspecific competition on predator involving Atangana–Baleanu–Caputo derivative, *Results Phys.* **49** (2023) 106489.

[4] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*, 1st edn. (Oxford University Press, Oxford, 1991).

[5] M. Aguiar, V. Anam, K. B. Blyuss, C. D. S. Estadilla, B. V. Guerrero, D. Knopoff, B. W. Kooi, A. K. Srivastav, V. Steindorf and N. Stollenwerk, Mathematical models for dengue fever epidemiology: A 10-year systematic review, *Phys. Life Rev.* **40** (2022) 65–92, doi:10.1016/j.plrev.2022.02.001.

[6] F. Brauer, P. Driessche and J. Wu, *Mathematical Epidemiology*, 1st edn. (Springer-Verlag, Berlin, Heidelberg, 2008).

[7] Bureau of Epidemiology, Department of Disease Control, Ministry of Public Health, Dengue fever (2019), https://ddc.moph.go.th/disease.php (accessed 13 January 2022).

[8] Centers for Disease Control and Prevention, Dengue, https://www.cdc.gov/dengue/index.html.

[9] M. Chan and M. A. Johansson, The incubation periods of dengue viruses, *PLoS One* **7**(11) (2012) e50972.

[10] A. Din, T. Khan, Y. Li, H. Tahir, A. Khan and W. A. Khan, Mathematical analysis of dengue stochastic epidemic model, *Results Phys.* **20** (2021) 103719.

[11] A. C. Elliott, L. S. Hynan, J. S. Reisch and J. P. Smith, Preparing data for analysis using Microsoft Excel, *J. Investig. Med.* **54**(6) (2006) 334–341, doi:10.2310/6650.2006.05038.

[12] L. Esteva and C. Vargas, A model for dengue disease with variable human population, *J. Math. Biol.* **38**(3) (1999) 220–240.

[13] R. Gangula, L. Thirupathi, R. Parupati, K. Sreeveda and S. Gattoju, Ensemble machine learning-based prediction of dengue disease with performance and accuracy elevation patterns, *Mater. Today Proc.* **80**(3) (2023) 3458–3463, doi:10.1016/j.matpr.2021.07.270.

[14] D. J. Gubler and G. G. Clark, Dengue hemorrhagic fever: The emergence of the global health problem, *Emerg. Infect. Dis.* **1**(2) (1995) 55–57.

[15] S. A. Jose, R. Raja, J. Dianavinnarasi, D. Baleanu and A. Jirawattanapanit, Mathematical modeling of chickenpox in Phuket: Efficacy of precautionary measures and bifurcation analysis, *Biomed. Signal Process. Control* **84** (2023) 104714.

[16] S. A. Jose, R. Raja, B. I. Omede, R. P. Agarwal, J. Alzabut, J. Cao and V. E. Balas, Mathematical modeling on co-infection: Transmission dynamics of Zika virus and Dengue fever, *Nonlinear Dynam.* **111** (2023) 4879–4914.

[17] S. A. Jose, R. Raja, J. Alzabut, G. Rajchakit, J. Cao and V. E. Balas, Mathematical modeling on transmission and optimal control strategies of corruption dynamics, *Nonlinear Dynam.* **109** (2022) 3169–3187.

[18] S. A. Jose, R. Raja, Q. Zhu, J. Alzabut, M. Niezabitowski and V. E. Balas, Impact of strong determination and awareness on substance addictions: A mathematical modeling approach, *Math. Methods Appl. Sci.* **45**(8) (2022) 4140–4160.

[19] D. Joseph, R. Ramachandran, J. Alzabut, S. A. Jose and H. Khan, Fractional order-density dependent mathematical model to find the better strain of Wolbachia, *Symmetry* **15**(4) (2023) 845.

[20] B. Z. Naaly, T. Marijani, A. Isdory and J. Z. Ndendya, Mathematical modeling of the effects of vector control, treatment, and mass awareness on the transmission dynamics of dengue fever, *Comput. Methods Programs Biomed. Update* **6** (2024) 100159.

[21] M. O. Olayiwola and A. I. Alaje, Mathematical analysis of intrahost spread and control of Dengue Virus: Unraveling the crucial role of antigenic immunity, *Franklin Open* **7** (2024) 100117.

[22] M. S. Olayemi, O. O. Olajide and O. F. Ajayi, Application of statistical software in quantitative data analysis, *ABUAD Int. J. Nat. Appl. Sci.* **2**(1) (2022) 29–34, doi:10.53982/aijnas.2022.0201.03-j.

[23] H. R. Pandey, G. R. Phaijoo and D. B. Gurung, Analysis of dengue infection transmission dynamics in Nepal using fractional order mathematical modeling, *Chaos Solitons Fractals X* **11** (2023) 100098.

[24] Phuket Provincial Public Health Office, *Phuket Health Development Strategic Plan* (Public Health Strategic Development Group, Phuket, 2022).

[25] M. S. Rahman, C. Pientong, S. Zafar, T. Ekalaksananan, R. E. Paul, U. Haque, J. Rocklöv and H. J. Overgaard, Mapping the spatial distribution of the dengue vector *Aedes aegypti* and predicting its abundance in northeastern Thailand using machine-learning approach, *One Health* **13** (2021) 100358, doi:10.1016/j.onehlt.2021.100358.

[26] M. Ranjan, K. Barot, V. Khairnar, V. Rawal, A. Pimpalgaonkar and S. Saxena, Python: Empowering data science applications and research, *J. Oper. Syst. Dev. Trends* **10** (2023) 27–33.

[27] E. Soewono and A. K. Supriatna, A two-dimensional model for the transmission of dengue fever disease, *Bull. Malays. Math. Sci. Soc.* **24**(1) (2001) 49–97.

[28] V. Steindorf, S. Oliva, J. Wu and M. Aguiar, Effect of general cross-immunity protection and antibody-dependent enhancement in dengue dynamics, *Comput. Math. Methods* **2022** (2022) 2074325.

[29] R. Thomas, S. A. Jose, R. Raja, J. Alzabut, J. Cao, V. E. Balas and M. Niezabitowski, Modeling and analysis of SEIRS epidemic models using homotopy perturbation method: A special outlook to 2019-nCoV in India, *Int. J. Biomath.* **15** (2022) 2250059, doi:10.1142/S1793524522500590.

[30] S. Wongkoon, M. Jaroensutasinee and K. Jaroensutasinee, Development of temporal modeling for prediction of dengue infection in Northeastern Thailand, *Asian Pac. J. Trop. Med.* **5**(3) (2012) 249–252, doi:10.1016/S1995-7645(12)60034-0.

[31] Y. Xu, Research on MATLAB and its application in mathematical modeling, *Appl. Mech. Mater.* **687–691** (2014) 1202–1205, doi:10.4028/www.scientific.net/amm.687-691.1202.

[32] Z. Zheng, N. Xu, M. Khan, M. Pedersen, T. Abdalgader and L. Zhang, Nonlinear impacts of climate change on dengue transmission in mainland China: Underlying mechanisms and future projection, *Ecol. Model.* **492** (2024) 110734.