Time Series Analysis and Forecasting on the Dengue Case Incidence in Iloilo using SARIMA Models

Sted Micah Cheng, Ivan Yuri De Leon, Annika Nicole Montemayor

Abstract

Dengue remains a major public health concern in the Philippines, with seasonal surges during the rainy season placing significant strain on public healthcare systems. This study models weekly dengue incidence in Iloilo, a province in Western Visayas, using time series analysis to identify short-term and multi-year seasonal patterns. After data pre-processing and statistical testing, several SARIMA models were fitted and evaluated using the Akaike Information Criterion (AIC). The best-fit model, SARIMA$(0,1,1) \times (0,1,1)_{156}$, suggests that weekly dengue incidence is influenced by recent case counts and three-year seasonal patterns. Forecasts from this model project surges around mid-2023 and 2024, aligning with known annual patterns despite the triannual seasonal specification. These results may inform the timing of seek-and-destroy operations on mosquito breeding sites and hospital resource allocation. However, wide forecast confidence intervals also highlight the limitations of relying solely on historical statistical models. Combining time series models with real-time surveillance and expert recommendations remains essential for effective and timely public health response.

1 Introduction

1.1 Background

Dengue is the fastest-spreading vector-borne viral disease globally and remains endemic in over 100 countries. It is caused by any of four closely related virus types (DENV-1 to DENV-4) and is primarily transmitted to humans through bites from infected *Aedes aegypti* and *Aedes albopictus* mosquitoes [2]. Symptoms typically include fever, body aches, nausea, and rashes, with warning signs such as fatigue, vomiting, and bleeding emerging 24–48 hours after the fever subsides. These symptoms usually manifest within two weeks after being bitten and last approximately one week. While most infections are mild, about 1 in 20 cases progresses to severe dengue, which can result in shock, internal bleeding, and death [1].

As a tropical nation, the Philippines experiences regular dengue outbreaks, specifically during the rainy season when environmental conditions such as high rainfall and cooler temperatures favor mosquito breeding [7]. In response, the Department of Health has implemented seek-and-destroy campaigns to eliminate breeding sites. However, following the easing of mobility restrictions in the post-pandemic era, a resurgence of dengue has been observed in recent years [3].

One notable hotspot in the Philippines is Iloilo province. By August 2024, over 5,000 dengue cases were recorded, prompting the local government to declare a state of calamity [6]. In 2025, although the rainy season had only begun in late May, dengue cases in the region had already exceeded 2,000 by the following month, prompting intensified community efforts for mosquito elimination and case control [5].

Given the persistent seasonal behavior of dengue transmission, it is pertinent to analyze temporal patterns in dengue cases. Accurately forecasting trends allows stakeholders to proactively intensify

Submission for MATH 271.2 Advanced Time Series and Forecasting, AY 2025-2026 Intersession

their seek-and-destroy operations and assist in optimal hospital resource allocation. As such, this study seeks to utilize time-series analysis through autoregressive and moving average models to model and forecast weekly dengue cases in Iloilo using historical data from the Department of Health - Philippine Integrated Disease Surveillance and Response (DOH-PIDSR).

1.2 Statement of the problem

This paper seeks to investigate the time series patterns and characteristics of dengue cases in Iloilo from 2008 to 2022. The analysis will focus on identifying trends and seasonality in the data, testing for stationarity and autocorrelation, determining the best-fit time series model among autoregressive (AR), moving average (MA), and autoregressive moving average (ARMA) processes with consideration for trends and seasonality (collectively SARIMA processes), and forecasting future surges in dengue cases for the next few years.

1.3 Scope and limitations

The study will assess the time series characteristics of weekly dengue cases in Iloilo from 2008 to 2022. The analysis will be based on historical data obtained from the DOH-PIDSR under Project Climate Change, Health, and Artificial Intelligence (Project CCHAIN). Since the dataset is open-source, the granularity and duration of its data collection period are limited. More importantly, since the time frame concludes in 2022, it may not have captured the resurgence of dengue cases outlined in previous studies [3].

Moreover, this research focuses on identifying the temporal patterns of dengue cases in Iloilo and predicting future incidences up to 2024. The data will be fit to SARIMA processes, but these do not account for potential volatile behavior that may be caused by significant epidemiological events during this period.

2 Methodology

2.1 Dataset description

The dataset was retrieved from Project CCHAIN [8], a compilation of climate, environmental, socioeconomic, and health data in twelve Filipino cities, one of which is Iloilo. The contributors include Thinking Machines (a data science consulting firm), EpiMetrics (a health research institution), Manila Observatory (an environmental research institution), and the Philippine Action for Community-led Shelter Initiatives (an NGO advocating for the urban poor).

The CSV file (`disease_pidsr_totals.csv`) is the total number of disease cases reported by the DOH-PIDSR, in weekly and city granularity. The modeled variable (`case_totals`) represents the total case counts of a particular disease in a particular week and city. Out of 74,880 rows in the dataset, 780 rows pertain to the dengue cases in Iloilo. The dataset documentation can be retrieved via [4].

A preliminary analysis of the graph (Figure 1) shows that the dengue cases in Iloilo are significantly higher than those in other cities. This contributes to the interest in modeling this time series.

2.2 Theorems and models

Before proceeding to model the datasets, the augmented Dickey-Fuller (ADF) and Ljung-Box (LB) tests are used to test for stationarity and autocorrelation. Stationarity implies that the mean and covariance do not change over time, while autocorrelation implies the existence of linear relationships between the current and past values of the time series. More details are described in Appendix A.

The functional form of the mixed SARIMA (seasonal autoregressive integrated moving average) process is

$$\Phi(B^s)\phi(B)\nabla_s^D\nabla^d X_t = \Theta(B^s)\theta(B)Z_t.$$

The factors $\phi(B)$ and $\theta(B)$ are the characteristic polynomials of the non-seasonal AR and MA components, referring to the non-seasonal dependence of the current cases on the cases (AR) or the errors (MA) of the past week(s), respectively. The variable s refers to the seasonality or the number

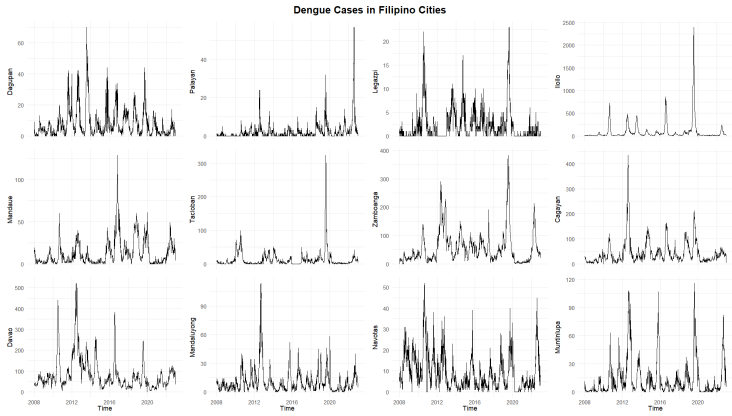

Figure 1: Plot of dengue cases in twelve Filipino cities from 2008 to 2022.

of time steps in each season. The factors $\Phi(B^s)$ and $\Theta(B^s)$ are the characteristic polynomials of the seasonal AR and MA components, referring to the seasonal dependence of the current cases on the cases (AR) or errors (MA) of the previous season(s), respectively. The ∇^d operator refers to performing d non-seasonal differences (between two consecutive time steps), and the ∇_s^D operator refers to performing D seasonal differences (between two time steps that are s apart, i.e., one season apart). A more detailed discussion of the SARIMA model can be found in Appendix B.

To develop a mixed SARIMA time series process, the `R` packages `tseries` and `forecast` were mainly employed. Once a suitable model exhibiting stationarity and autocorrelation was found, the correlograms of the sample partial autocorrelation function (PACF) and autocorrelation function (ACF) of the number of dengue cases in Iloilo were plotted. Each plot includes two horizontal lines representing the standard-error limits, and only lags that exceed the standard-error limit line of PACF and ACF will be considered as the AR and MA orders, respectively. Additionally, due to the inherent seasonality of the data, a seasonal differencing of $s = 52$ will be applied. If the time series is not stationary or not autocorrelated, (non-seasonal) differencing will be conducted until the new time series is stationary and autocorrelated.

To determine the best model, the Akaike Information Criterion (AIC) will be used. It imposes penalties for having too many parameters in the model and has the formula $\text{AIC} = 2p - 2\ln(\mathcal{L})$, where p is the number of model parameters and $\mathcal{L}$ is the maximum likelihood estimator of the model. The model with the lowest AIC will be chosen, since a smaller AIC indicates a better model.

2.3 Assumptions

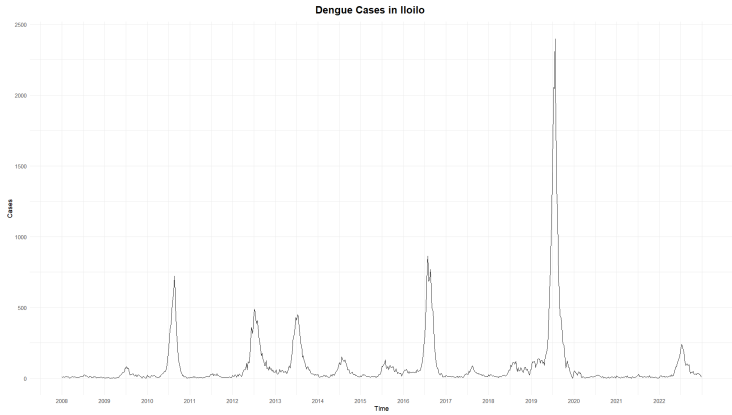

Figure 2: Plots of dengue cases in Iloilo from 2008 to 2022.

The case spike in 2019 (refer to Figure 2) and the lack of smaller spikes in 2020–2022 (probably due to lesser exposure and reporting during the pandemic) are assumed not to affect the modeling process. Additionally, it is assumed that any trends and seasonality present in the dataset can be removed through differencing and seasonal parameters. A significance level of $\alpha = 0.01$ will be used in all tests. For parsimony, only order parameters (p, q, P, Q) that are at most 5 will be considered.

3 Results and discussion

3.1 Preliminary data processing

Let X_t be the time series with no transformations. Due to the presence of large values, the log-transformation $\ln X_t$ was performed. From Table 2 (in the appendix), X_t and $\ln X_t$ are both stationary and autocorrelated, however, the ADF test is not as effective in detecting non-stationarity due to seasonality. This suggests performing seasonal differencing ($D = 1$) with $s = 52$, resulting in $\nabla_{52} \ln X_t$, which is autocorrelated but not stationary. Thus, another (non-seasonal) differencing ($d = 1$) is taken, resulting in $\nabla(\nabla_{52} \ln X_t)$, which is both stationary and autocorrelated.

In the plot of the dengue cases in Iloilo (Figure 2), there is a noticeable upward trend in the middle parts of the years 2010, 2013, 2016, and 2019. Again, this suggests performing seasonal differencing ($D = 1$) with $s = 156$ (three years), resulting in $\nabla_{156} \ln X_t$, which is autocorrelated but not stationary. Thus, another (non-seasonal) differencing ($d = 1$) is taken, resulting in the time series $\nabla(\nabla_{156} \ln X_t)$, which is both stationary and autocorrelated.

3.2 Model specification

We now seek to determine the potential orders (p, q, P, Q) of the model. $d = 1$ and $D = 1$ were determined by the differencing in Section 3.1, with $s = 52$ and $s = 156$ for the parameter D.

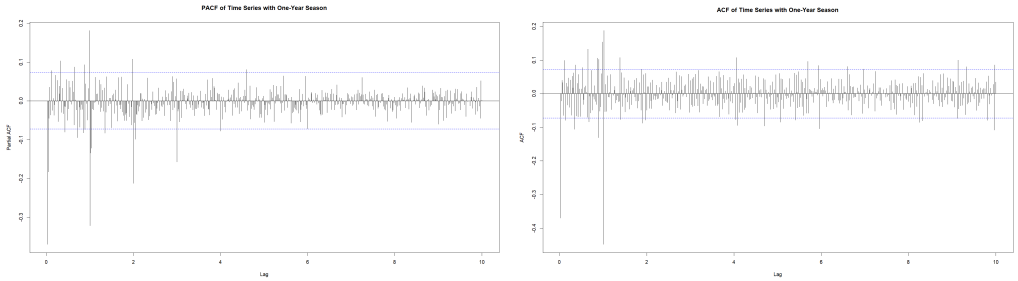

Figure 3: PACF (left) & ACF (right) for the log-transformed time series with $d = 1$, $D = 1$, $s = 52$.

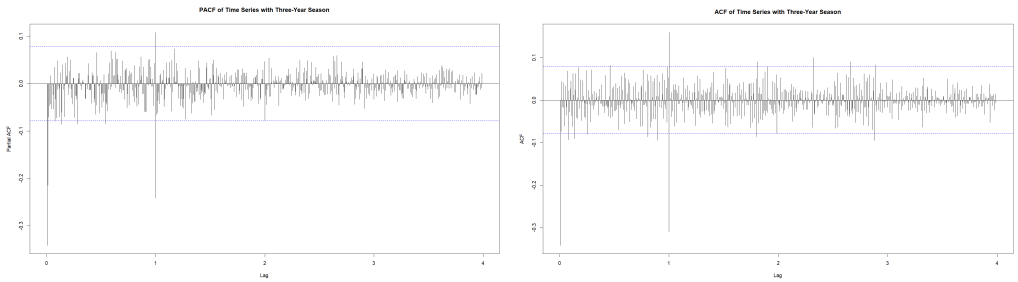

Figure 4: PACF (left) & ACF (right) for the log-transformed time series with $d = 1$, $D = 1$, $s = 156$.

Figures 3 and 4 show the sample PACFs and ACFs for the transformed time series with $s = 52$ and $s = 156$ respectively. For $s = 52$, the PACF tails off from $h = 0$ to $h = 3$, while the ACF cuts off at $h = 1$. This behavior may indicate a seasonal MA(1) model (i.e., $P = 0, Q = 1$), but the non-seasonal counterparts $p = 0, 1, 2, 3$ and $q = 0, 1$ are also tested. The possibility of $P = 0$ and $Q = 0$ is examined as well. Next, for $s = 156$, the PACF and ACF both cut off at $h = 1$, so every

combination of $p, q, P, Q = 0, 1$ is tested. The AIC of each possible combination of (p, q, P, Q) is assessed using `Arima()`. The results for the 32 potential models are outlined in Table 1.

Table 1: Potential models with their AICs for seasonal lags of 52 weeks and 156 weeks

Seasonal lag of 52			Seasonal lag of 156		
(p, d, q)	(P, D, Q)	AIC	(p, d, q)	(P, D, Q)	AIC
$(0, 1, 0)$	$(0, 1, 0)$	1392.31	$(0, 1, 0)$	$(0, 1, 0)$	1123.30
$(0, 1, 1)$		1263.52	$(0, 1, 1)$		1009.51
$(1, 1, 0)$		1286.95	$(1, 1, 0)$		1041.22
$(1, 1, 1)$		1265.10	$(1, 1, 1)$		1010.89
$(2, 1, 0)$		1263.68	$(0, 1, 0)$	$(0, 1, 1)$	991.56
$(2, 1, 1)$		1264.64	$(0, 1, 1)$		882.96
$(3, 1, 0)$		1264.24	$(1, 1, 0)$		913.17
$(3, 1, 1)$		1262.51	$(1, 1, 1)$		883.47
$(0, 1, 0)$	$(0, 1, 1)$	1010.34	$(0, 1, 0)$	$(1, 1, 0)$	1018.30
$(0, 1, 1)$		894.26	$(0, 1, 1)$		912.24
$(1, 1, 0)$		920.99	$(1, 1, 0)$		*
$(1, 1, 1)$		896.08	$(1, 1, 1)$		*
$(2, 1, 0)$		897.09	$(0, 1, 0)$	$(1, 1, 1)$	993.42
$(2, 1, 1)$		896.72	$(0, 1, 1)$		884.92
$(3, 1, 0)$		896.83	$(1, 1, 0)$		*
$(3, 1, 1)$		898.72	$(1, 1, 1)$		*

The asterisks (*) in Table 1 indicate that errors involving non-finite values have occurred when attempting to run `Arima()`. Upon examining the results, models with the same parameter structure $(p, d, q) = (P, D, Q)$ consistently produced lower AIC values when using a seasonal lag of 156 weeks compared to 52. This supports the observation that dengue outbreaks in Iloilo tend to exhibit larger spikes approximately every three years, and suggests that case patterns may be more strongly influenced by multi-year seasonal cycles rather than annual seasonality alone.

The best-fit model follows a SARIMA$(0, 1, 1) \times (0, 1, 1)_{156}$ structure, with $\theta_1 = -0.4680$ and $\Theta_1 = -0.7039$. The functional form of the model can be expressed as follows:

$$\left(1 - B^{156}\right)(1 - B)\, X_t = \left(1 - 0.7039B^{156}\right)(1 - 0.4680B)\, Z_t$$
$$\Rightarrow X_t = X_{t-1} + X_{t-156} - X_{t-157} + Z_t - 0.4680Z_{t-1} - 0.7039Z_{t-156} + 0.3294252Z_{t-157}$$

Hence, the weekly number of dengue cases in Iloilo is primarily influenced by the previous week's cases and by seasonal patterns from three years ago. The inclusion of lag terms such as X_{t-1} and X_{t-156} reflects the differenced structure of the data, allowing the model to account for recent values and periodic patterns that recur at roughly three-year intervals.

The model also includes lagged random fluctuations to capture short-term and seasonal variations affecting the current week's cases. The relatively large seasonal moving average coefficient at lag 156 (-0.7039) indicates that values from three years prior play a notable role in the model's structure. This is accompanied by a positive coefficient at lag 157 ($+0.3294$), which appears to moderate the influence from the preceding seasonal lag. Additionally, the non-seasonal moving average term at lag 1 (-0.4680) reflects short-term adjustments based on the prior week's case count.

Overall, the model incorporates both recent and seasonal information in a way that aligns with the observed temporal structure of the data. While specific interpretations of individual coefficients should be made with caution, the model provides a reasonable representation of the patterns present in the Iloilo dengue time series.

3.3 Forecasting

The best-fit model was used to generate a 104-week (two-year) forecast of weekly dengue incidence in Iloilo. The model was applied to the log-transformed series to maintain consistency with model training. The resulting forecasts were then back-transformed to the original scale for interpretation.

To account for the transformation applied to the data, we assumed that the forecast distribution on the log scale is approximately normal. This assumption is well supported by the near-normal distribution of residuals, as clearly shown in the QQ plot (see Figure 5).

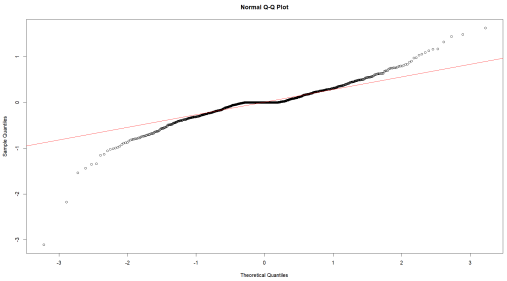

Figure 5: QQ plot of the residuals of the log-transformed series.

Under this assumption, the predicted dengue cases on the original scale follow a log-normal distribution. Accordingly, the forecasted mean and standard error (which were then converted into 99% confidence intervals) were derived using standard log-normal formulas, as detailed in Appendix D.

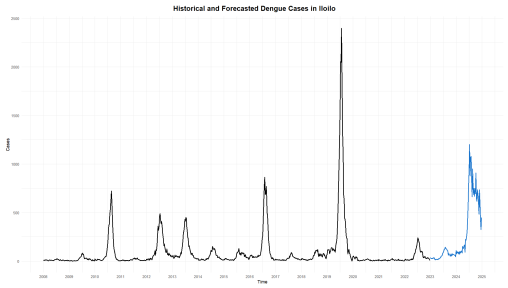

Figure 6: Historical values (black) and point estimates for the forecasted values (blue)

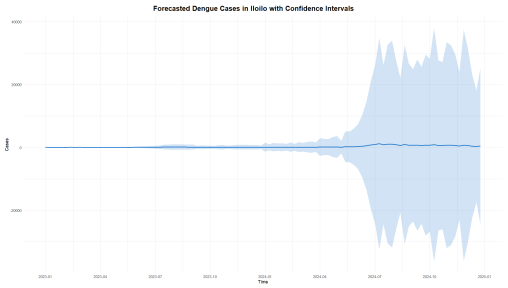

Figure 7: Forecasted values (line) with their corresponding 99% confidence intervals (shaded)

Figure 6 displays both historical and forecasted values. The forecast reveals recurring midyear spikes, with a more pronounced surge projected in 2024. This pattern suggests an intensification of seasonal outbreaks in the second year of the forecast horizon, consistent with the trend identified by the model.

Next, Figure 7 shows the forecasted values only, with the blue line indicating the expected number of dengue cases and the shaded band representing the 99% prediction interval. The width of this band reflects the inherent uncertainty of long-term forecasting, particularly for dengue, whose transmission is influenced by a complex interplay of environmental, biological, and social factors.

The forecast extends the observed multi-year pattern of dengue incidence, with a noticeable increase in cases expected around the midpoints of 2023 and 2024 within the forecast horizon. While the projected case counts remain within the historical range, they highlight periods of potential elevation, indicating a possible seasonal surge. This trend aligns with the annual cycle identified during model development. Notably, although the model was specified with triannual seasonality, it successfully captures the expected annual spike.

However, the presence of wide confidence intervals in some parts of the forecast underscores the limitations of relying solely on historical trends to predict future outcomes. The complexity of dengue transmission introduces uncertainty that cannot be fully addressed by a purely statistical model.

Nonetheless, the model's projected trajectory appears broadly consistent with historical records. As of mid-2024, dengue cases in the region had surpassed 5,000 [6], while model forecasts projected this number to be surpassed by the end of June (see Table 3 in the appendix), reinforcing the anticipated midyear surge. Overall, the forecast is best interpreted as a tool for identifying possible trends rather than producing precise outcomes. It may be most effective when combined with real-time surveillance data and local expertise to support timely public health responses.

References

[1] Centers for Disease Control and Prevention. Symptoms of dengue and testing, May 2025.

[2] Department of Health. Dengue prevention and control program.

[3] J. S. Interior, K. J. J. Bigay, R. A. A. Iringan, and M. B. F. Tanco. Resurgence of dengue in the Philippines. *World Journal of Virology*, 13(3), Sep 2024.

[4] Lacuna Fund Project Team. Project CCHAIN database, 2024.

[5] P. Lena. Intensified community action urged as Iloilo sees hike in dengue cases, Jul 2025.

[6] Republic of the Philippines, Province of Iloilo. Iloilo logs 5,836 dengue cases, Aug 2024.

[7] M. E. Subido and I. S. Aniversario. A correlation study between dengue incidence and climatological factors in the philippines. *Asian Research Journal of Mathematics*, 18(12):110–119, Dec 2022.

[8] Thinking Machines Data Science, Inc. Project CCHAIN, 2024.

A Hypothesis tests

The augmented Dickey-Fuller test evaluates the following hypotheses:

$$H_0\text{: The model has a unit root.}$$
$$H_a\text{: The model has no unit roots.}$$

The null hypothesis implies that the time series is non-stationary, and the alternative hypothesis implies that it is stationary. The implementation in `R` returns the p-value of the ADF test. If $p < 0.01$, the null hypothesis is rejected and thus the time series is stationary.

Let ρ_i for $i \in \{1, 2, \ldots, m\}$ be the value of the autocorrelation function (ACF) of the time series at lag h. The LB test examines the following hypotheses:

$$H_0\text{: } \rho_1 = \rho_2 = \ldots = \rho_m = 0$$
$$H_a\text{: } \rho_i \neq 0 \text{ for some } i \in \{1, 2, \ldots, m\}$$

The null (and alternative) hypothesis implies that the time series is not autocorrelated (and autocorrelated), respectively. The implementation in `R` returns the p-value of the LB test. If $p < 0.01$, the hypothesis is rejected and thus the time series is autocorrelated.

B Model preliminaries

This section discusses the different components of a mixed or multiplicative SARIMA process. For uniformity, Z_t always refers to a white noise process with mean 0 and variance σ^2. The time series X_t is assumed to have zero mean.

A time series X_t is an *autoregressive* process of order p, denoted by AR(p), if

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} + Z_t,$$

where $\text{Cov}(X_r, Z_s) = 0$ for each $r < s$. For a stationary Gaussian AR(p) process, the order p can be determined by the partial autocorrelation function (PACF), since for all $h > p$, the sample PACF $\hat{\phi}_{h,h}$ at lag h approaches 0. That is, the sample PACF "cuts off" at lag p.

A time series X_t is a *moving average* process of order q, denoted by MA(q), if

$$X_t = Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j}.$$

For an MA(q) process, the order q can be determined by the ACF ρ_h, since $\rho_h = 0$ for $h > q$. That is, the sample ACF "cuts off" at lag q.

A time series X_t is an *autoregressive moving average* process of orders p and q, denoted by ARMA(p, q), if it is stationary and can be expressed as

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j}.$$

We define the backward shift operator B as decreasing the subscript of a time series by 1. For example, $BX_t = X_{t-1}$ and $B^2 X_{t-1} = X_{t-3}$. From this, we can write the summations compactly using the characteristic polynomials of the AR and MA components, which are

$$\phi(B) = 1 - \sum_{i=1}^{p} \phi_i B^i \quad \text{and} \quad \theta(B) = 1 + \sum_{j=1}^{q} \theta_j B^j,$$

respectively. The functional form of the ARMA process can be rewritten as $\phi(B)X_t = \theta(B)Z_t$.

A time series X_t is a *seasonal ARMA* process with (seasonal) orders P and Q, denoted by SARMA(P, Q), if $\Phi(B^s)X_t = \Theta(B^s)Z_t$, where the characteristic polynomials of the seasonal AR and MA components are defined as

$$\Phi(B^s) = 1 - \sum_{i=1}^{P} \Phi_i B^{is} \quad \text{and} \quad \Theta(B^s) = 1 + \sum_{j=1}^{Q} \Theta_j B^{js}.$$

That is, the current value of the series depends on the values and errors that occurred ks steps ago, where $k \in \mathbb{N}$.

We combine the characteristic polynomials of the non-seasonal and seasonal AR and MA components to obtain the *mixed (or multiplicative) seasonal ARMA* process with seasonal orders P and Q and non-seasonal orders p and q, denoted by SARMA$(p, q) \times (P, Q)_s$. The functional form of this process is $\Phi(B^s)\phi(B)X_t = \Theta(B^s)\theta(B)Z_t$.

For the non-seasonal differencing component d, we first define $\nabla X_t = X_t - X_{t-1} = (1 - B)X_t$, which represents performing one difference between two consecutive time steps, and inductively define $\nabla^d X_t = \nabla(\nabla^{d-1} X_t) = (1 - B)^d X_t$, which represents performing d differences. For the seasonal differencing component $D = 1$, we first define $\nabla_s X_t = X_t - X_{t-s} = (1 - B^s)X_t$, which represents performing one difference between two time steps that are one season apart, and inductively define $\nabla_s^D X_t = \nabla_s(\nabla_s^{D-1} X_t) = (1 - B^s)^D X_t$, which represents performing D differences.

C Results of the ADF and LB hypothesis tests

Table 2: p-values of the ADF and LB hypothesis tests for the time series with different transformations

Time Series	ADF p-values	LB p-values
X_t	< 0.01	$< 2.2 \times 10^{-16}$
$\ln X_t$	< 0.01	$< 2.2 \times 10^{-16}$
$\nabla_{52} \ln X_t$	0.07989	$< 2.2 \times 10^{-16}$
$\nabla(\nabla_{52} \ln X_t)$	< 0.01	$< 2.2 \times 10^{-16}$
$\nabla_{156} \ln X_t$	0.01004	$< 2.2 \times 10^{-16}$
$\nabla(\nabla_{156} \ln X_t)$	< 0.01	$< 2.2 \times 10^{-16}$

D Log-normal distribution formulas

Let Y represent the time series on the original scale. Then the predicted mean and standard error of Y are calculated as follows:

$$\mathbb{E}[Y] = \exp\left(\mu + \frac{1}{2}\sigma^2\right),$$
$$\text{SE}[Y] = \mathbb{E}[Y] \cdot \sqrt{\exp(\sigma^2) - 1},$$

where μ and σ represent the predicted mean and standard error on the log scale, respectively.

E Cumulative predicted dengue cases in 2024

Table 3: Cumulative sum of the predicted dengue cases in 2024

Week	Current Cases	Cumulative Cases	Week	Current Cases	Cumulative Cases
1	110	110	27	925	5627
2	74	184	28	1201	6828
3	99	283	29	877	7705
4	84	367	30	1067	8772
5	90	457	31	1078	9850
6	72	529	32	856	10706
7	98	627	33	662	11368
8	70	697	34	950	12318
9	99	796	35	757	13075
10	85	881	36	686	13761
11	98	979	37	749	14510
12	105	1084	38	667	15177
13	83	1167	39	747	15924
14	152	1319	40	693	16617
15	136	1455	41	908	17525
16	130	1585	42	645	18170
17	155	1740	43	615	18785
18	166	1906	44	736	19521
19	92	1998	45	692	20213
20	224	2222	46	616	20829
21	214	2436	47	483	21312
22	248	2684	48	734	22046
23	295	2979	49	600	22646
24	403	3382	50	431	23077
25	556	3938	51	325	23402
26	764	4702	52	446	23848