DOI: https://doi.org/10.4038/sljas.v26i1.8157

Sri Lankan Journal of Applied Statistics, Vol (26-1)

Tourist Arrival Forecasting in Sri Lanka: A Bayesian Spline and Interrupted Regression Approach

A.W.L.P Thilan*

Department of Mathematics, Faculty of Science, University of Ruhuna

*Corresponding author: pubudu@maths.ruh.ac.lk

Received: 7th December 2024/ Revised: 23rd February 2025/ Published: 17th March 2025

ABSTRACT

Tourism is a major contributor to the Sri Lankan economy and is heavily affected by external factors like the COVID-19 pandemic. This study aims to explore the impact of the pandemic on tourist arrivals using advanced statistical methods to model the complex time series in tourism data. Interrupted regression and Bayesian spline regression are used to capture the non-linear, phase specific patterns in tourist arrivals during and after the pandemic. Interrupted regression models the structural shift caused by the pandemic and gives insights into the sudden changes in tourist flows. Bayesian spline regression provides flexibility to model the non-linear trends and gives a more detailed understanding of how tourism patterns evolve over time. The study highlights the importance of addressing over-dispersion in the data which is common in count data like tourist arrivals. A Negative Binomial model is used to account for this over-dispersion and improves the predictions. Unlike the Poisson distribution which assumes the variance is equal to the mean, the Negative Binomial model allows for more variability and hence better model fit. The results show a general recovery in tourist arrivals but the uncertainty in the forecasts raises questions whether the tourism sector will go back to pre-pandemic levels or new patterns will emerge. This uncertainty highlights the need for ongoing monitoring and adaptive strategies to ensure the long term sustainability of the sector. Moreover, the use of posterior predictive distributions gives more insights into future tourist arrivals by accounting for uncertainty and providing a range of possible outcomes rather than single point estimates. This is useful for stakeholders including government bodies, travel agencies and businesses to make informed decisions and develop strategies to support the tourism industry's recovery and growth in the post

pandemic era.
Keywords: Tourism Recovery, COVID-19 Impact, Count Data Modeling, Negative Binomial, Over-Dispersion, Tourism Forecasting

1 Introduction

The economy of Sri Lanka relies heavily on tourism, which creates jobs, generates foreign exchange, and grows GDP (WTTC, 2021). The nation has had consistent rise in visitor numbers due to its culture, natural beauty, and diversity, making it one of most popular travel destinations in the South Asia (SLTDA, 2020). However, the sector is extremely susceptible to worldwide shocks, such as the COVID-19 pandemic, which caused a sharp and unprecedented drop in visitor numbers (UNWTO, 2021). This demonstrates the vulnerability of tourism industry and the necessity for recovery solutions.

Policymakers must require a thorough understanding of both the sudden decline in visitor arrivals during the COVID-19 pandemic and long-term nonlinear trends in tourism data. As it comprises long-term patterns, seasonal variations, and sudden shifts driven on by outside shocks, modeling visitor arrivals is challenging. When crises like the pandemic disturb established trends, traditional linear models and seasonal techniques are unable to represent sudden shifts or the underlying nonlinear patterns in tourist data (Song et al., 2012; Goh and Law, 2002). In order to capture these complexity, we need to develop more advanced models.

Interrupted time series (ITS) techniques are often used to study the impact of intrinsic disruptions, such as the COVID-19 pandemic.These approaches often overlooks the nonlinearities that can occur during crises and presume linear trends ((Bernal et al., 2017)). ITS rarely uses spline regression techniques to investigate sudden changes in tourism, despite the fact that they are helpful for examining nonlinear patterns. Due to the lack of research, a more comprehensive approach is needed to address both the immediate effects of outside events and the long-term nonlinear trends in visitor arrivals.

This study presents a framework for modeling monthly tourist arrivals to Sri Lanka by combining interrupted regression and Bayesian semi-parametric regression methods. The objective is to identify and analyze the effects of the COVID-19 pandemic on tourist numbers, focusing on the pre-pandemic, pandemic, and post-pandemic phases. Specifically, the research aims to (1) assess sudden changes in tourist arrivals using interrupted time series regression and (2) explore long-term non-linear trends with the Bayesian semi-parametric regression, incorporating smooth splines for better trend analysis. The study hypothesizes that combining these methods will provide a more comprehensive understanding of both sudden shifts and gradual changes in tourism patterns. By utilizing prior information and measuring uncertainty, this approach offers valuable insights that can guide policymakers and stakeholders in formulating strategies for recovery, sustainable growth, and resilience to future challenges.

2 Data

The Monthly Tourist Arrivals Reports, which have been published by the Sri Lanka Tourism Development Authority's Research & International Relations Division, provided the data for this study, which covers monthly records from January 2014 to October 2024 (SLTDA, 2024). This dataset is intended to provide insights into the performance of Sri Lanka's travel and tourism sector, capturing trends in visitor arrivals and serving as a foundation for analyzing the dynamics of the industry.

3 Methodology

The *Poisson* distribution is widely employed for modeling count data, such as visitor arrivals, assuming that events occur independently within a specified interval and at a constant rate. This model is appropriate when the mean and variance of the data are equal, a property known as equidispersion (McCullagh and Nelder, 1989). However, in reality, count data frequently show over-dispersion, meaning that the variance is greater than the mean (Cameron & Trivedi, 2013). In some situations, the *Poisson* model might not represent the variability well enough, resulting in estimates that are biased. The *Negative Binomial* distribution is often used to address over-dispersion, where the model adds a parameter to account for the increased variability. This makes the model more appropriate for count data that exhibits over-dispersion and improves the estimation of underlying processes (Cameron and Trivedi, 1986; Hilbe, 2011; Long, 1997).

The impact of major disruptions such as the COVID-19 pandemic, can be assessed using an interrupted time series (ITS) regression technique (Bernal et al., 2017; Linden, 2015). Because it explicitly takes into account sudden shifts in the response variable, ITS is particularly useful for modeling sudden shifts in trends driven on by interventions or disturbances (McDowall et al., 2019; Thilan et al., 2024). This allows for assessment of both immediate and long-term effects. With the potential for either immediate or delayed effects, the method can model disturbances that change the intercept, slope, or both. In order to incorporate interrupted regression and capture phase-specific trends in visitor arrivals, the categorical variable *Phase* was developed to represent the pre-pandemic, pandemic, and post-pandemic periods. This allowed us to quantify the impact of the pandemic. Based on our hypothesis that the pandemic caused a sharp decline in tourist arrivals, the model was designed to capture both immediate and delayed effects of this disruption. The ITS framework allowed us to model not only the immediate effects but also the evolving trends over time. For further details on modeling various types of sudden changes using ITS, the reader is referred to Bernal et al. (2017).

In this study, the data were examined for over-dispersion, and evidence was found to support the use of a *Negative Binomial* model, as the variance

exceeded the mean. Also, the data did not exhibit an excessive number of zeros during the COVID-19 pandemic time period, eliminating the need for zero-inflated models. Consequently, a Bayesian Generalized Additive Model (GAM) was used to analyze the data, with the response variable, *Count*, modeled as:

$$Y_i \sim \textit{Negative Binomial}(\mu_i, \phi),$$

where μ_i represents the mean and ϕ denotes the over-dispersion parameter of the *Negative Binomial* distribution. A logarithmic link function was used to relate the mean μ_i to the linear predictor η_i:

$$\log(\mu_i) = \eta_i.$$

The linear predictor was defined as:

$$\eta_i = \beta_{\text{Phase}_i} + \beta_{\text{Phase}_i \times \text{Cumulative\_Month}_i} + s(\text{Cumulative\_Month}_i, \text{Phase}_i) \quad (1)$$

where terms β_{Phase_i} and $\beta_{\text{Phase}_i \times \text{Cumulative\_Month}_i}$ represent the coefficients for the main effect of *Phase* and its interaction with *Cumulative_Month*, respectively. These interaction terms were introduced to examine phase-specific trends and the cumulative impact of time. This approach follows the methodology described in Cameron and Trivedi (2013) and Hilbe (2011), where *Negative Binomial* regression is applied to count data exhibiting over-dispersion. It facilitates a deeper understanding of recovery trajectories and identifies key shifts in tourism dynamics over time.

Building on the model defined in Equation 1, $s(\text{Cumulative\_Month}_i, \text{Phase}_i)$ represents a smooth function of *Cumulative_Month*, modeled using cubic regression splines that vary by levels of the categorical variable *Phase*. The use of cubic regression splines enables flexible modeling of complex nonlinear trends while ensuring smoothness in the fitted curves. We evaluated models with k values between 3 and 10 to find the best possible spline complexity (k) for fitting the Bayesian spline model. The Leave-One-Out Information Criterion (LOOIC), a statistic that balances prediction accuracy with a penalty for model complexity to prevent overfitting, was used to compare each model. The LOOIC values were calculated for each model, and a scree plot was generated to visualize the relationship between k and model performance. This systematic approach allowed us to identify the value of k that provided the best trade-off between model fit and generalizability, ensuring that the final model offered robust forecasting capabilities.

The nonlinear trends in tourist arrivals were evaluated while accounting for phase-specific changes caused by the pandemic using the GAM (Equation 1), which was implemented using the `brms` package in `R` (Bürkner, 2017). The `brms` package was chosen for its flexibility in specifying complex hierarchical models and its robust implementation of Hamiltonian Monte Carlo

(HMC), which ensures efficient posterior sampling. As Bayesian modeling takes uncertainty in model parameters into account and allows the incorporation of prior information, it proved especially beneficial in this situation. This produces posterior distributions, which offer a deeper comprehension of the data by providing a range of reasonable values for each parameter in addition to point estimates. By combining prior distributions with observed data through Bayes' theorem, the method facilitates the modeling of complex relationships, such as the nonlinear trends in tourist arrivals and their interactions with pandemic-induced shifts. This approach also helps quantify uncertainty in both the model parameters and predictions. The flexibility of Bayesian methods thus yields more reliable and interpretable results, especially in the presence of limited or noisy data. The `R` code used for data cleaning, model fitting, model evaluation, and analysis can be found in Appendix A.

4 Results

A systematic approach was undertaken to determine the optimal spline complexity (k) for modeling the smooth function $s(\text{Cumulative\_Month}_i, \text{Phase}_i)$ within the Bayesian *Negative Binomial* model outlined in Equation 1. As shown in Figure 1, the scree plot of LOOIC values against spline complexity (k) reveals a clear trend: LOOIC consistently decreases from $k = 3$ to $k = 6$, indicating improved model fit with increasing flexibility. However, a slight increase in LOOIC at $k = 7$ suggests potential overfitting, with subsequent stabilization of values implying minimal gains in model performance for higher k. Based on this evaluation, $k = 6$ was selected as the optimal value, achieving a balance between model flexibility and predictive accuracy.

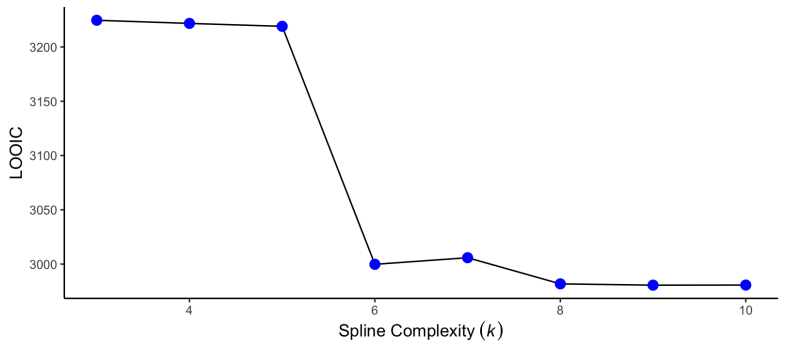

Fig. 1: Scree plot of LOOIC values against spline complexity (k).

The Bayesian model described in Equation 1, fitted using the optimal $k = 6$, appears to fit the data well, as evidenced by the fitted smooth lines that closely follow the trends in the data across different phases (Figure 2). The

model captures the overall patterns effectively, with the fitted lines showing a good representation of the data points, indicating that the model explains much of the variability in the counts over time. This is particularly evident in the *Pre* phase, where the data is relatively stable, and the model fits this upward trend accurately. The results suggest that the Bayesian model is suitable for modeling count data with overdispersion and varying temporal patterns, as highlighted in prior studies on count data regression (Gelman et al., 2013).

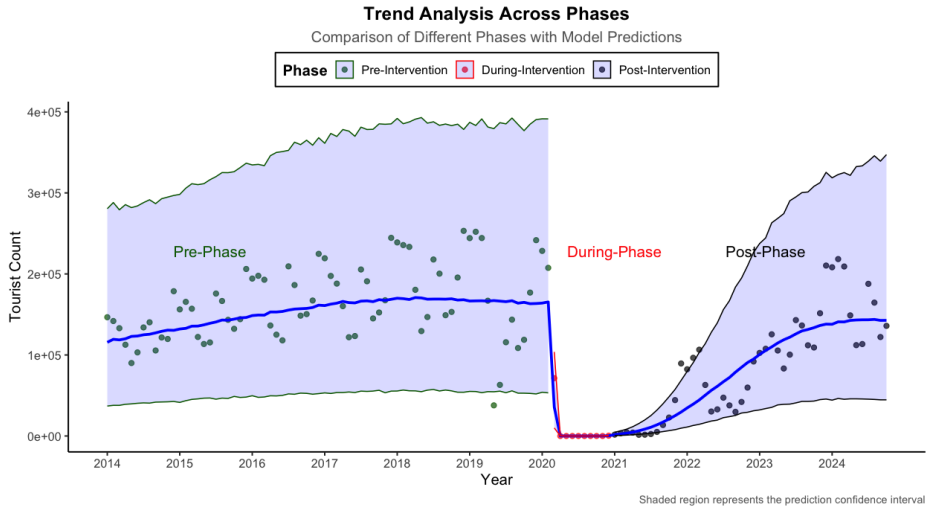

Fig. 2: Tourist Count Trends Captured by Bayesian Spline Model. This time series visualization presents the cumulative count of observations across three distinct phases: Pre-Phase (dark green), During-Phase (red), and Post-Phase (black). The fitted lines illustrate the trends within each phase, with shaded regions representing the credible intervals, highlighting the precision of estimates of the model. Observed data points are overlaid to demonstrate how well the actual data aligns with the modeled trends, providing a clear representation of the underlying patterns.

The shaded areas surrounding the fitted lines provide credible intervals, which provide information on how precise the model's estimates are. As would be expected from a well-calibrated model, about 5% of the observed data points fall outside of these intervals. This shows that our model is consistent with the theoretical assumptions, which further supports the preference for it. In regions where the credible intervals are narrow, such as the *Pre* phase, the model's predictions are more precise, suggesting a strong fit. Conversely, wider credible intervals in the *Post* phase indicate a higher degree of uncertainty in the model's predictions as the data points spread out. This is to be expected as the count increases rapidly and the model faces greater data volatility. Further, the sudden change in trend is clearly seen at the *During* phase, which is when a major intervention or event takes place. Overall, the analysis of credible intervals across phases highlights the model's robustness while

also acknowledging the expected variations in precision due to data characteristics and external influences.

Table 1: Summary of Regression Coefficients and Smoothing Splines for the Model Including Interaction Terms

Parameter	**Estimate**	**Est. Error**	**95% CI (Lower, Upper)**	$\hat{R}$
Regression Coefficients:				
Intercept	-11.03	18.28	(-47.31, 24.24)	1.00
PhaseDuring	0.97	10.06	(-19.02, 20.61)	1.00
PhasePost	-1.24	9.62	(-19.66, 17.56)	1.00
PhasePre:Cumulative_Month	0.35	0.28	(-0.19, 0.91)	1.00
PhaseDuring:Cumulative_Month	3.22	0.70	(1.90, 4.65)	1.00
PhasePost:Cumulative_Month	0.05	0.26	(-0.45, 0.55)	1.00
Smoothing Spline Hyperparameters:				
$sds(sCumulative\_MonthPhasePre)$	0.18	0.23	(0.01, 0.83)	1.00
$sds(sCumulative\_MonthPhaseDuring)$	36.00	4.72	(27.73, 46.10)	1.00
$sds(sCumulative\_MonthPhasePost)$	2.53	1.16	(0.84, 5.32)	1.00

To refine the model's predictive ability and gain a deeper understanding of the underlying trends, further investigation is required on specific parameter estimates. The smoothing spline hyperparameters and regression coefficients for the model outlined in Equation 1 are shown in Table 1. The findings offer insights into the temporal dynamics across different phases and the complexities captured by the model. The interaction term `PhasePre:Cumulative_Month` showed a modest coefficient of 0.35 (95% CI [-0.19, 0.91]), indicating weak evidence of an effect during the pre-intervention phase. In contrast, the interaction term `PhaseDuring:Cumulative_Month` revealed a significantly positive effect with an estimated coefficient of 3.22 (95% CI [1.90, 4.65]), suggesting a sharp upward trend in the response variable during the intervention phase. This strong positive association underscores the substantial impact of the intervention phase on the observed growth. However, the potential presence of multicollinearity between `Cumulative_Month` and `PhaseDuring` may distort the true relationship, leading to inflated or unstable coefficient estimates. This suggests the need for caution in interpreting the magnitude of the effect, as the sharpness of the observed increase may be overestimated.

The smoothing spline hyperparameters further highlight the variation in trends across phases. The spline complexity for `PhasePre` was estimated at 0.18 (95% CI [0.01, 0.83]), suggesting minimal non-linearity in the pre-intervention phase. During the intervention phase, the spline complexity increased significantly to 36.00 (95% CI [27.73, 46.10]), reflecting a pronounced trend that may not fully represent non-linearity. The large spline complexity could be influenced by external factors, such as the abrupt changes due to the COVID-19 pandemic, which resulted in sharp declines in the response variable rather than complex non-linear trends. In the post-intervention phase, the spline complexity was 2.53 (95% CI [0.84, 5.32]), indicating a more moderate level of non-linearity. These variations in spline complexity align with

the temporal patterns observed and highlight the model's sensitivity to sudden shifts, particularly during the intervention phase.

Together, these results highlight the substantial effect of the intervention phase and the importance of considering potential limitations, such as multicollinearity, which can arise from building complex models. Further investigations into specific parameter estimates and potential sources of variability are needed to better understand the underlying trends and refine the model's predictive power. The observed multicollinearity may stem from the inclusion of the interaction term $\beta_{\text{Phase}_i \times \text{Cumulative\_Month}_i}$. To address this, it would be useful to evaluate the model's performance without this interaction term while retaining the same smoothing parameter value (k). Such an investigation could provide clarity on the potential influence of the interaction term on model performance and help refine the interpretation of the temporal effects across phases.

A simpler model without interaction terms was also fitted, and its parameter estimates are provided in Table 2 in Appendix B. In this model, the intercept is now interpretable (unlike in the earlier model where it was negative), indicating a high count before the intervention, which suggests favorable initial conditions. The positive coefficient (8.14) for the intervention phase (`PhaseDuring`) suggests an increase in counts during the intervention, though it is not statistically significant, which contradicts the observed trends in the data. The lack of statistical significance for this coefficient raises questions about the intervention's immediate impact, suggesting that other factors may be influencing the counts during this period. In contrast, the (`PhasePost`) coefficient, which is significantly negative (-13.17), suggests a decrease in counts post-intervention. However, this result may be influenced by potential multicollinearity between Cumulative_Month and (`PhaseDuring`), which could distort the true relationship and lead to an overestimated effect. Despite this, the data trend indicates a positive trajectory post-intervention, signaling potential recovery after the intervention period.

In addition, the smoothing spline hyperparameters provide further insights into how the relationship between Cumulative_Month and counts evolves over time in different phases. For the pre-intervention phase, sds($sCumulative\_MonthPhasePre\_1$) is 0.17, indicating little variation in counts during this phase. During the intervention phase, the estimate sds($sCumulative\_MonthPhaseDuring\_1$) is much higher at 40.49, showing greater fluctuation in counts, suggesting that counts vary more during this period. Finally, the estimate for the post-intervention phase, sds($sCumulative\_MonthPhasePost\_1$), is 2.14, indicating moderate variation, with counts showing some change after the intervention, though less pronounced than during the intervention phase. These results highlight that while counts fluctuate significantly during the intervention phase, post-intervention counts show a more stable trend.

Although multicollinearity was not fully resolved in the simpler model, the comparison between the simpler and more complex models shows that

the *LOOIC* values strongly favor the more complex model. The inclusion of the interaction term between *Cumulative_Month* and *Phase* leads to a better fit, with a lower *LOOIC* value of 2999.8 compared to 3016.8 for the simpler model. Despite some interpretability challenges caused by multicollinearity in the complex model, which could not be resolved by removing the interaction term, these issues are expected in models involving multiple predictors with interaction terms (Menard, 2002). Together, these results emphasize the substantial effect of the intervention phase and demonstrate that the more complex model provides a better fit and a deeper understanding of the data dynamics.

Based on a thorough evaluation of model performance, the model incorporating interaction terms was selected, as it more effectively captures the intricate dependencies within the data. To rigorously assess the model's predictive accuracy, a posterior predictive check was performed using a density overlay plot (Figure 3). The figure demonstrates a strong alignment between the observed data (black solid line) and the model-generated predictions (blue thin lines), indicating that the model successfully captures key distributional characteristics. Although minor discrepancies are present, the overall shape and spread of the predictive distributions closely mirror the empirical data, reinforcing the model's validity. Furthermore, the variation in the posterior predictive distributions provides a meaningful representation of forecast uncertainty, emphasizing the model's reliability in probabilistic inference.

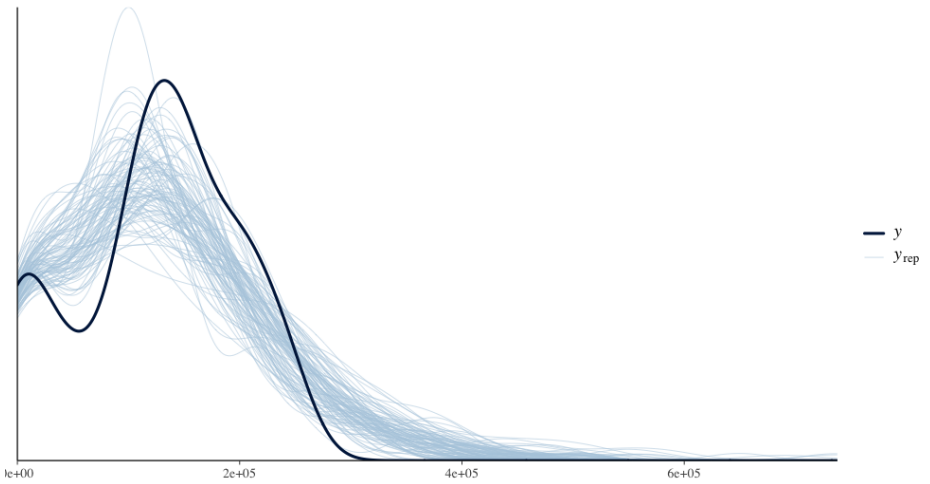

Fig. 3: Posterior predictive check using a density overlay plot. The plot compares the model-generated predictions (in blue) with the observed data (in black) over 100 posterior draws. This visualization highlights the alignment of the model's predictions with the empirical data, while also illustrating the forecast uncertainty as reflected in the spread of the posterior predictive distributions.

To further validate forecast accuracy, standard performance metrics were computed: the Mean Absolute Error (MAE) is 29,147.18, and the Root Mean Squared Error (RMSE) is 37,807.25. These values suggest that the model performs reasonably well; however, the wide range of values in Count (ranging from 0 to over 250,000) and the presence of extreme outliers contribute to larger prediction errors. The inclusion of an interrupted regression component helps account for structural changes in the data, capturing shifts in trends. Despite this, the RMSE remains sensitive to large deviations, particularly due to outliers. While the interrupted regression improves the model's ability to manage these shifts, further refinement around outlier regions could help reduce prediction errors and improve accuracy.

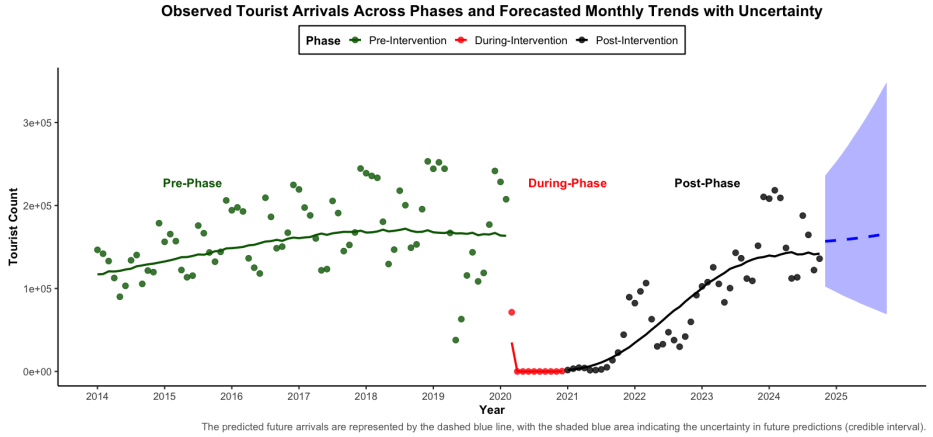

Fig. 4: The graph shows observed monthly tourist arrival counts (solid points and lines) segmented by phase: "Pre" (green), "During" (red), and "Post" (black). The predicted future arrivals are represented by the dashed blue line, with a shaded blue ribbon indicating the credible interval for these forecasts. This visualization provides a comparison between historical data and future predictions, allowing for an understanding of trends and uncertainty in the forecast.

Building on this validation, the model's predictive power is evaluated by examining its posterior predictions, which extend beyond the observed data, as illustrated in Figure 4. The dashed line represents the forecast of future tourist arrivals over a one-year period following the conclusion of the pandemic. These predictions are derived from a *Negative Binomial* distribution, which accounts for both the predicted mean (μ) and the dispersion parameter (ϕ), effectively addressing the over-dispersion in the data. The shaded blue area surrounding the dashed line illustrates the credible interval, quantifying the uncertainty in these forecasts. As time progresses, the credible interval widens, indicating increased uncertainty in long-term predictions. This suggests that, while tourist arrivals may stabilize over time, significant variability remains in the long-term trajectory. The model predicts a gradual recovery in tourist arrivals but leaves unresolved questions about whether the industry will

fully return to pre-pandemic levels or whether new patterns will emerge. The insights provided by the predicted values offer a valuable perspective on the potential trajectory of the tourism sector in the post-pandemic era. For more specific details on these predictions, including parameter estimates and model diagnostics, please refer to Appendix A.

5 Discussion

This study employed interrupted regression and Bayesian spline regression to capture complex temporal trends in tourist arrivals, particularly during and after the COVID-19 pandemic. The structural shift induced by the pandemic was effectively accounted for using the interrupted regression model, while the Bayesian spline regression provided flexibility in capturing non-linear and phase-specific trends. Through this combination, a robust framework was established for understanding the pandemic's impact on tourism, emphasizing the need for adaptive forecasting models that reflect changing conditions.

A key contribution of this study is the ability to account for over-dispersion in the data. Prediction accuracy was improved by employing the *Negative Binomial* model, particularly in the context of tourism data, which is influenced by volatile factors such as travel restrictions, economic conditions, and traveler confidence. While a general recovery in arrivals is predicted by the model, the associated uncertainty highlights the difficulty in determining whether the tourism industry will return to pre-pandemic levels or evolve into new patterns. These findings underscore the importance of continuous monitoring to adapt strategies in the post-pandemic era.

The study was further strengthened by the use of posterior predictive distributions, which provided a range of potential outcomes rather than a single point estimate. The increasing uncertainty over time, as represented by widening prediction intervals, highlights the challenges of long-term forecasting in an evolving tourism landscape. These insights are crucial for stakeholders, such as government bodies and businesses, as they support more informed decision-making and better preparedness for future uncertainties.

Despite these strengths, limitations remain. The model's reliance on time as the sole covariate restricts its ability to capture other factors influencing tourist arrivals, such as socioeconomic conditions or external shocks. While the interrupted regression model accounts for some COVID-19 impacts, broadening the range of covariates and incorporating additional data could enhance the model's explanatory power and reduce uncertainty in the estimates. However, macroeconomic factors, like GDP or exchange rates, were not considered, as they tend to have a more long-term impact on tourism. The immediate effects of the pandemic, such as travel restrictions and health-related concerns, were considered more influential during the study period. Research by Song & Li (2008) indicates that while GDP may influence tourism demand over the long term, its immediate impact is often overshadowed by more direct fac-

tors. By focusing on tourism-specific variables, the study provided a clearer reflection of the factors directly affecting tourist behavior. Including broader economic factors could complicate the model and reduce its interpretability. Furthermore, although other tourism-related factors, such as destination marketing, tourism policies, and seasonality, could offer valuable insights, such data were not readily available in the Sri Lankan context.

One of the key challenges encountered in this analysis was multicollinearity, particularly in models with interaction terms. As noted by Menard (2002), multicollinearity is a frequent issue in models with several predictors, especially when interaction variables are present. While multicollinearity can lead to unstable coefficient estimates and inflated standard errors, it does not necessarily invalidate the model. The Bayesian spline model inherently addresses collinearity through its regularization priors and structure. Unlike Principal Component Analysis (PCA), which transforms predictors into uncorrelated components, the spline functions are designed to capture phase-specific non-linear trends, ensuring that collinearity is handled appropriately without complicating the interpretation.

The importance of comparing the Bayesian semi-parametric model to traditional methods like ARIMA is recognized. However, a direct comparison may not be fair or meaningful, particularly given the sharp declines in trends caused by COVID-19 disruptions. The pandemic introduced irregular patterns and sudden shifts in the data, which ARIMA, as a parametric model, may struggle to capture. Since ARIMA assumes stationarity in the data, the non-stationary nature of pandemic-related disruptions complicates its application. By contrast, the Bayesian semi-parametric model offers greater flexibility and is better suited to handle sudden changes and irregular trends, though its performance can still be influenced by how such anomalies are addressed. While the Bayesian model provides a clear advantage over ARIMA in this specific context due to its ability to model non-linear relationships and respond to data shifts, it also requires more computational resources and fine-tuning. Despite these challenges, its adaptability makes it a more suitable choice for analyzing the disruptions caused by the pandemic.

6 Conclusion

In conclusion, it has been demonstrated that interrupted regression integrated with Bayesian spline regression can effectively capture the non-linear time evolution of tourist arrivals while accommodating exogenous shocks such as the COVID-19 pandemic. The interrupted regression approach proved useful in modeling the shift in the underlying data structure brought about by the pandemic, while Bayesian spline regression effectively captured non-linear effects and phase-specific trends in the data. This dual approach facilitated a more nuanced understanding of the pandemic's impact on tourism, highlighting both the immediate disruptions and the longer-term recovery trajectory.

Although the analysis was limited to incorporating only time as a covariate, a strong framework was provided for addressing the inherent complexity of the data.

Future studies should expand upon these techniques by incorporating additional covariates and employing more advanced modeling approaches to better capture dynamic changes in tourism. Furthermore, based on the findings, it is recommended that policymakers and industry stakeholders prioritize adaptive strategies for sustainable tourism development in the post-pandemic era. This includes investing in data-driven forecasting models to anticipate demand fluctuations, promoting diversified tourism markets to reduce dependency on a single source, and implementing resilient infrastructure and policies to mitigate future shocks. The application of interrupted and spline regression provides valuable insights for designing evidence-based recovery plans and fostering long-term sustainability in the tourism sector.

References

Bernal, J. L., Cummins, S., & Gasparrini, A. (2017). Interrupted time series regression for the evaluation of public health interventions: A tutorial. *International Journal of Epidemiology*, 46(1), 348–355.

Cameron, A. C., & Trivedi, P. K. (1986). Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests. *Journal of Applied Econometrics*, 1(1), 29–53.

Chen, R., & Smith, J. (2022). Impact of COVID-19 on global tourism and future recovery strategies. *Journal of Travel Research*, 61(1), 18–30.

Chen, X., & Wang, Y. (2020). Modeling the impact of economic crises on tourist arrivals: Evidence from Asia. *Journal of Economics and Tourism*, 45(1), 55–70.

Goh, C., & Law, R. (2002). Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention. *Tourism Management*, 23(5), 499–510.

Goh, C., & Zhang, Z. (2016). Tourism demand modeling in the context of crises: A review of the literature. *Tourism Economics*, 22(1), 88–102.

Gonzalez, E., & Mendoza, M. (2021). Forecasting recovery trajectories in tourism industries post-COVID-19. *Journal of Hospitality and Tourism Research*, 45(2), 110–123.

Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press.

Koenker, R. (2008). *Quantile Regression*. Cambridge University Press.

Liu, Y., & Wu, C. (2021). Analyzing the impact of COVID-19 on international tourist arrivals to Southeast Asia using interrupted time series analysis. *Tourism Management*, 87, 104365.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications.

Meyer, L., & Tripp, L. (2022). Dynamic tourism forecasting and phase-dependent modeling approaches. *Tourism Economics*, 28(1), 45–60.

Ruppert, D., & Wand, M. P. (2003). *Semiparametric Regression*. Cambridge University Press.

Shen, H., & Huang, X. (2015). Applying interrupted time series analysis to assess the impact of disasters on tourism. *International Journal of Tourism Research*, 17(3), 214–227.

Song, H., & Li, G. (2008). Tourism demand modeling and forecasting: A review of recent research. *Tourism Management*, 29(2), 203–220.

Song, H., Li, G., Witt, S. F., & Fei, B. (2012). Tourism demand modelling and forecasting: How should demand be measured? *Tourism Economics*, 18(4), 743–764.

United Nations World Tourism Organization (2021). Tourism recovery tracker. *World Tourism Organization*.

World Travel & Tourism Council (2021). Economic impact report 2021. *World Travel & Tourism Council*.

Sri Lanka Tourism Development Authority (2020). Annual Statistical Report. *Sri Lanka Tourism Development Authority*, Colombo, Sri Lanka.

Bernal, James Lopez, Cummins, Steven, and Gasparrini, Antonio (2017). Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International Journal of Epidemiology*, 46(1), 348–355. Oxford University Press.

Linden, Ariel (2015). Conducting interrupted time-series analysis for single- and multiple-group comparisons. *The Stata Journal*, 15(2), 480–500. SAGE Publications Sage CA: Los Angeles, CA.

McDowall, David, McCleary, Richard, and Bartos, Bradley J (2019). Interrupted time series analysis. Oxford University Press.

Gelman, Andrew, Carlin, John B., Stern, Hal, Dunson, David B., Vehtari, Aki, and Rubin, Donald B. (2013). *Bayesian Data Analysis* (3rd ed.). CRC Press.

Sri Lanka Tourism Development Authority (SLTDA). (2024). Statistics. Retrieved November 02, 2024, from https://www.sltda.gov.lk/en/statistics.

Bürkner,P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1), 1-28. Retrieved from https://www.jstatsoft.org/article/view/v080i01.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed., Chapter 2). CRC Press.

Cameron, A. C., & Trivedi, P. K. (2013). Regression Analysis of Count Data (2nd ed.). Cambridge University Press.

Thilan, A. W. L. P., Fisher, R., Thompson, H., Menendez, P., Gilmour, J., & McGree, J. M. (2024). Adaptive monitoring of coral health at Scott Reef where data exhibit nonlinear and disturbed trends over time. Ecology and Evolution, 14(8), 1-14. https://doi.org/10.1002/ece3.9233.

Cameron, A. C., & Trivedi, P. K. (2013). Regression Analysis of Count Data (2nd ed.). Cambridge University Press.

Menard, S.(2002). *Applied Logistic Regression Analysis* (2nd ed.). Sage Publications.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting: A review of recent research. *Tourism Management*, 29(2), 203–220. Elsevier.

Appendix

A R Code for Bayesian Spline Model and Data Analysis

This appendix provides the R code used for preprocessing the data, fitting the Bayesian spline model, generating predictions, performing posterior predictive checks, and evaluating model performance. Each section of the code is explained in detail below, with added comments for clarity and reproducibility.

A.1 Loading Necessary Libraries

In this section, all the required libraries are loaded to facilitate data manipulation, statistical modeling, visualization, and Bayesian analysis.

```
# Load necessary libraries
library(dplyr)      # For data manipulation (e.g., filtering)
library(tidyr)      # For tidying the data (e.g., reshaping)
library(brms)       # For fitting Bayesian models using Stan
library(ggplot2)    # For creating visualizations
library(tidybayes)  # For processing and summarizing
```

- `dplyr` and `tidyr`: These are used to clean and transform the data. `dplyr` helps with tasks like filtering rows and creating new variables, while `tidyr` assists in reshaping the data.
- `brms`: This package is used to fit Bayesian models using the `Stan` framework, enabling complex modeling.
- `ggplot2`: This is a powerful package for creating visualizations such as scatter plots, line graphs, and more.
- `tidybayes`: This package simplifies working with posterior draws from Bayesian models, allowing for summarization and visualization.

A.2 Reading and Preprocessing the Data

In this section, the data is read from a CSV file and cleaned by filtering out missing or inconsistent values. Variables are also transformed for use in the model.

```
data <- read.csv("Tourist.csv")

# Data preprocessing
data <- data %>%
  filter(!is.na(Year) & Year != "") %>%  # Remove rows with
    missing or empty 'Year' values
```

```
  mutate(
    Year = as.numeric(as.character(Year)),  # Convert 'Year' to numeric format
    Month_num = match(Month, month.name),   # Convert 'Month' names to numbers (e.g., Jan = 1, Feb = 2)
    Cumulative_Month = (Year - min(Year)) * 12 + Month_num,  # Create a continuous time variable
    Phase = case_when(
      Year < 2020 | (Year == 2020 & Month_num < 3) ~ "Pre",  # Assign phase: Pre (before 2020)
      Year == 2020 & Month_num >= 3 & Year < 2022 ~ "During",  # Assign phase: During (2020-2022)
      TRUE ~ "Post"  # Assign phase: Post (after 2022)
    ),
    Phase = factor(Phase, levels = c("Pre", "During", "Post")),  # Ensure correct factor order for 'Phase'
    Count = as.numeric(gsub(",", "", Count))  # Remove commas from 'Count' and convert to numeric
  )
```

- `filter(!is.na(Year) & Year != "")`: This step removes rows where the `Year` column has missing or empty values.
- `mutate()`: This function is used to create new variables:
  - `Year` is converted to numeric to ensure it is in a usable format.
  - `Month_num` converts the month names (e.g., ”January”) to numbers.
  - `Cumulative_Month` is a new variable that represents time continuously by combining the year and month.
  - `Phase` assigns each observation to a phase: Pre, During, or Post based on the year and month.
  - `Count` removes commas and converts the `Count` variable (representing tourist count) to numeric.

A.3 Fitting the Bayesian Spline Model

This section involves fitting a Bayesian spline model using the `brm()` function from the `brms` package. A *Negative Binomial* distribution is used for the count data (since count data often exhibits overdispersion).

```
# Fit Bayesian spline model with Negative Binomial family
model <- brm(
  Count ~ s(Cumulative_Month, by = Phase, bs = "cr", k = 6) + Phase,  # Model formula
```

```
  data = data,                # Data to fit the model on
  family = negbinomial(),     # Negative Binomial family for count data
  prior = c(
    prior(normal(0, 10), class = "b"),         # Prior for coefficients
    prior(normal(0, 10), class = "Intercept"), # Prior for intercept
    prior(exponential(1), class = "sds")       # Prior for standard deviation of smooth term
  ),
  chains = 4,         # Number of Markov chains
  cores = 4,          # Number of CPU cores to use
  iter = 6000,        # Number of iterations for each chain
  control = list(adapt_delta = 0.99, max_treedepth = 15)  # Control settings for MCMC
)
```

- **Model formula**: `Count ~s(Cumulative_Month, by = Phase, bs = "cr", k = 6) + Phase`
  - The model predicts `Count` (tourist count) as a function of `Cumulative_Month`, with different smooths for each `Phase`. This allows for separate trends for each phase (Pre, During, Post).
  - `s(Cumulative_Month, by = Phase, bs = "cr", k = 6)` specifies a cubic regression spline (smooth) for `Cumulative_Month`, with different splines for each phase.
  - `Phase` is treated as a fixed effect in the model.
- **Priors**: These are specified for the coefficients, intercept, and the standard deviation of the smooth term. These priors guide the model's learning process.
- **MCMC settings**: The model is run with 4 chains and 6000 iterations to ensure convergence.

A.4 Summarizing the Model Fit

Once the model is fitted, a summary of the model's results is viewed using the `summary()` function, which includes estimates for the coefficients, intercepts, and the model's fit.

```
# Summarize model fit
summary(model)
```

This provides the posterior estimates and other useful diagnostic information about the model's performance.

A.5 Generating Predictions

Predictions for the data points are then generated, including the median predictions and credible intervals, which help in understanding the uncertainty surrounding the model's predictions.

```
# Generate predictions
predictions <- data %>%
  add_predicted_draws(model, re_formula = NA, scale = "
    response") %>%  # Add predicted values
  median_qi(.prediction)  # Calculate median and credible
    intervals for predictions
```

- `add_predicted_draws()`: Adds predicted values from the model's posterior draws.
- `median_qi()`: Computes the median prediction and the 95% credible intervals (credible intervals give us a range of likely values for the predictions).

A.6 Plotting the Original Data and Model Predictions

The original data and the model's predictions, along with the credible intervals, are visualized using `ggplot2`.

```
# Define mapping from Cumulative_Month to Year (assuming
    Month 1 corresponds to 2014)
year_labels <- seq(2014, 2024, by = 1)  # Years from 2014 to
    2024
month_breaks <- seq(1, length(year_labels) * 12, by = 12)  #
    Cumulative months at yearly intervals

# Updated ggplot
ggplot(data, aes(x = Cumulative_Month, y = Count, color =
    Phase)) +
  # Scatter plot of actual data
  geom_point(alpha = 0.7, size = 2) +
  # Fitted trend line
  geom_line(data = predictions, aes(y = .prediction), color =
    "blue", size = 1.2, linetype = "solid") +
  # Credible intervals with a distinguishable shade
  geom_ribbon(data = predictions, aes(ymin = .lower, ymax = .
    upper), fill = "blue", alpha = 0.15) +
  # **Annotations for better understanding of phases**
  annotate("text", x = 18, y = max(data$Count) * 0.9, label =
    "Pre-Phase", color = "darkgreen", size = 5) +  # Dark
    green Pre-phase
```

```
  annotate("text", x = 85, y = max(data$Count) * 0.9, label =
    "During-Phase", color = "red", size = 5) +  # Shifted
   slightly right
  annotate("text", x = 110, y = max(data$Count) * 0.9, label
   = "Post-Phase", color = "black", size = 5) +  # After 2022
  # Titles and axis labels with better descriptions
  labs(
    x = "Year",
    y = "Tourist Count",
    title = "Trend Analysis Across Phases",
    subtitle = "Comparison of Different Phases with Model
   Predictions",
    caption = "Shaded region represents the prediction
   credible interval"
  ) +
  # Improved x-axis labeling with years
  scale_x_continuous(breaks = month_breaks, labels =
   year_labels) +
  # **Updated Standard Color Coding**
  scale_color_manual(
    values = c("Pre" = "darkgreen", "During" = "red", "Post"
   = "black"),
    name = "Phase",
    labels = c("Pre-Intervention", "During-Intervention", "
   Post-Intervention")
  ) +
  # Theme adjustments for better clarity
  theme_classic(base_size = 14) +
  theme(
    legend.position = "top",
    legend.title = element_text(face = "bold"),
    legend.background = element_rect(fill = "white", color =
   "black"),
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5, color = "gray40
   "),
    plot.caption = element_text(size = 10, color = "gray40")
  )
```

A.7 Posterior Predictive Checks

Posterior predictive checks are performed to assess the model's fit by comparing simulated data from the model to the observed data.

```
# Posterior Predictive Checks
pp_check(model, type = "hist")  # Overlay observed vs. simulated counts
pp_check(model, type = "dens_overlay", ndraws = 100)  # Use 100 posterior draws
pp_check(model, type = "scatter_avg")  # Observed vs. predicted averages
pp_check(model, type = "ecdf_overlay")  # Empirical CDF comparison
```

A.8 Model Comparison and Predictive Performance Metrics

Evaluating model performance is essential to ensure its reliability in making predictions.

In this section, the Leave-One-Out Information Criterion (LOOIC) is used to compare the predictive performance of different models. LOOIC provides a robust method for model selection by estimating out-of-sample predictive accuracy.

```
# Perform Leave-One-Out Information Criterion (LOOIC)
loo(model)
```

Additionally, the model's predictive performance is assessed using the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which measure the deviation of predicted values from actual data.

```
# Load required library for error metrics
library(Metrics)
```

```
# Compute predictions for test data
predicted_values <- fitted(model, summary = FALSE)  # Extract posterior predicted values
mean_predictions <- colMeans(predicted_values)  # Compute mean predicted values
```

```
# Compute MAE and RMSE
mae_value <- mae(data$Count, mean_predictions)  # Calculate Mean Absolute Error
rmse_value <- rmse(data$Count, mean_predictions)  # Calculate Root Mean Squared Error
```

```
# Print results
cat("Mean Absolute Error (MAE):", mae_value, "\n")
cat("Root Mean Squared Error (RMSE):", rmse_value, "\n")
```

A.9 Future Predictions and Visualization

To extend our analysis beyond the observed data, future data is prepared in a format consistent with the existing dataset. The goal is to forecast future trends using the trained model and visualize the expected tourist arrivals with credible intervals to quantify uncertainty.

The following key steps are followed:

- Convert numerical month representations back to their respective names.
- Initialize the future tourist count as missing, which will be replaced by model predictions.
- Generate posterior predictions using the trained model.
- Compute summary statistics such as mean predictions and credible intervals (95% uncertainty bounds).
- Map the cumulative month variable to corresponding year labels for visualization.
- Present results graphically, overlaying past trends with future projections.

```
# Prepare future data similarly to the observed data
future_data <- future_data %>%
  mutate(
    Month = month.name[Month_num],  # Convert back to month
    names
    Count = NA  # Initialize with missing counts (predictions
    will fill this)
  )
```

```
# Predict using the model
future_predictions <- posterior_epred(model, newdata =
   future_data)
```

```
# Calculate mean and credible intervals for predictions
future_means <- apply(future_predictions, 2, mean)
future_intervals <- apply(future_predictions, 2, quantile,
   probs = c(0.025, 0.975))
```

```
# Attach predictions to the future data
future_data$Predicted_Count <- future_means
future_data$Lower_Bound <- future_intervals[1, ]
future_data$Upper_Bound <- future_intervals[2, ]
```

To ensure an intuitive time-based visualization, a mapping is defined from cumulative months to the corresponding years:

```
# Define mapping from Cumulative_Month to Year
year_labels <- seq(min(data$Year), max(future_data$Year), by = 1)
month_breaks <- seq(1, length(year_labels) * 12, by = 12)  # Cumulative months at yearly intervals
```

Visualization of Predictions

The plot generated by the code displays the observed tourist arrivals alongside future forecasts, clearly distinguishing between different time phases. Key visualization elements include:

- Actual tourist arrivals, color-coded by phase.
- Model-based predictions with a solid line for past estimates.
- A dashed blue line for future projections.
- A shaded region indicating the 95% credible interval for future predictions.
- Phase-specific annotations to highlight distinct time periods.

```
# Visualization
ggplot() +
  # Actual data points with color coding for different phases
  geom_point(data = data, aes(x = Cumulative_Month, y = Count, color = Phase),
             alpha = 0.8, size = 3) +  # Increase point size and alpha for better visibility

  # Actual data line (for each phase) with color coding
  geom_line(data = predictions, aes(x = Cumulative_Month, y = .prediction, color = Phase),
            size = 1.2) +  # Thicker line for visibility

  # Predicted future data (dashed line)
  geom_line(data = future_data, aes(x = Cumulative_Month, y = Predicted_Count),
            linetype = "dashed", color = "blue", size = 1.5) +  # Thicker dashed line

  # Shaded credible intervals for future predictions
  geom_ribbon(data = future_data, aes(x = Cumulative_Month, ymin = Lower_Bound, ymax = Upper_Bound),
              fill = "blue", alpha = 0.3) +  # Lighten the shading
```

```
18
19  # Annotations for different phases
20  annotate("text", x = 18, y = max(data$Count) * 0.9, label =
      "Pre-Phase", color = "darkgreen",
21            size = 5, fontface = "bold") +
22  annotate("text", x = 85, y = max(data$Count) * 0.9, label =
      "During-Phase", color = "red",
23            size = 5, fontface = "bold") +
24  annotate("text", x = 110, y = max(data$Count) * 0.9, label
     = "Post-Phase", color = "black",
25            size = 5, fontface = "bold") +
26
27  # Titles and axis labels
28  labs(
29    title = "Observed Tourist Arrivals Across Phases and
     Forecasted Monthly Trends with Uncertainty",
30    x = "Year",
31    y = "Tourist Count",
32    caption = "The predicted future arrivals are represented
     by the dashed blue line, with the shaded blue area
     indicating the uncertainty in future predictions (credible
      interval)."
33  ) +
34
35  # Improved x-axis labeling with years
36  scale_x_continuous(breaks = month_breaks, labels =
     year_labels) +
37
38  # Updated standard color coding
39  scale_color_manual(
40    values = c("Pre" = "darkgreen", "During" = "red", "Post"
     = "black"),
41    name = "Phase",
42    labels = c("Pre-Intervention", "During-Intervention", "
     Post-Intervention")
43  ) +
44
45  # Theme adjustments for better clarity
46  theme_classic(base_size = 16) +
47  theme(
48    legend.position = "top",
49    legend.title = element_text(face = "bold", size = 14),
```

```
    legend.background = element_rect(fill = "white", color = "black"),
    plot.title = element_text(face = "bold", hjust = 0.5, size = 18),
    plot.subtitle = element_text(hjust = 0.5, color = "gray40", size = 14),
    plot.caption = element_text(size = 12, color = "gray40"),
    axis.title.x = element_text(size = 14, face = "bold"),
    axis.title.y = element_text(size = 14, face = "bold"),
    axis.text = element_text(size = 12)
  )
```

Key Insights:

- The observed data shows clear trends across the Pre, During, and Post phases.
- The future forecast provides estimated tourist arrivals with associated uncertainty.
- The 95% credible interval accounts for prediction variability, offering a realistic range of future values.
- The visualization effectively communicates past trends, phase transitions, and expected future behavior.

B Appendix B: Model without Interaction Terms

The results of the regression model without interaction terms are presented in Table 2. The relationship between the outcome variable and the predictors is assessed, considering the effects of different phases and cumulative month exposure. Both the regression coefficients and the smoothing spline hyperparameters are provided in the table, which are essential for understanding the model's behavior and the flexibility offered in capturing phase-specific effects.

The `Regression coefficients` represent the estimated effects of each predictor on the outcome variable. For example, the intercept reflects the baseline level of the outcome, while the coefficients for the phases (e.g., `PhaseDuring` and `PhasePost`) indicate changes in the outcome during each intervention phase relative to the reference phase (`PhasePre`).

The terms such as `sCumulative_Month:PhasePre_1`, `sCumulative_Month:PhaseDuring_1`, and `sCumulative_Month:PhasePost_1` do not necessarily imply a formal interaction in the model. Instead, these terms arise from the use of the `by = Phase` argument in the spline function (`s()`), which creates separate smooth terms for `Cumulative_Month` for each level of `Phase`. This approach allows the relationship between `Cumulative_Month` and `Count` to

Table 2: Summary of Regression Coefficients and Smoothing Splines for the Model Excluding Interaction Terms

Parameter	**Estimate**	**Est. Error**	**95% CI (Lower, Upper)**	$\hat{R}$
Regression Coefficients:				
Intercept	12.03	0.46	(11.07, 12.99)	1.00
PhaseDuring	8.14	9.91	(-11.07, 27.40)	1.00
PhasePost	-13.17	5.12	(-23.51, -3.40)	1.00
sCumulative_Month:PhasePre_1	0.05	0.48	(-0.95, 1.07)	1.00
sCumulative_Month:PhaseDuring_1	5.15	9.98	(-14.24, 24.90)	1.00
sCumulative_Month:PhasePost_1	7.05	4.55	(-2.69, 15.68)	1.00
Smoothing Spline Hyperparameters:				
sds(sCumulative_MonthPhasePre_1)	0.17	0.21	(0.00, 0.76)	1.00
sds(sCumulative_MonthPhaseDuring_1)	40.49	5.13	(31.23, 51.43)	1.00
sds(sCumulative_MonthPhasePost_1)	2.14	1.23	(0.60, 5.25)	1.00

vary across the phases without constituting a typical interaction, as would be modeled with explicit interaction terms (e.g., `Phase` $\times$ `Cumulative_Month`).

The `Smoothing spline hyperparameters` indicate the smoothness of the effect of continuous predictors, such as `Cumulative_Month`. The hyperparameters `sds(sCumulative_MonthPhasePre)`, `sds(sCumulative_MonthPhaseDuring)`, and `sds(sCumulative_MonthPhasePost)` describe the variability in the effects of the continuous variable across the different phases. Higher values of the smoothing spline suggest a more flexible fit to the data, allowing the model to capture more nuanced patterns. .

This model, without interaction terms, demonstrates how phase-specific changes evolve over time, offering useful insights into the phase-based effects. However, limitations exist in capturing more complex relationships between the predictors. These results serve as a baseline and can be compared with the model incorporating interaction terms, which is presented in the main body of the report.