Sains Malaysiana 53(11)(2024): 3817-3829
http://doi.org/10.17576/jsm-2024-5311-23

Modelling Malaysia Air Quality Data using Bayesian Structural Time Series Models

(Memodelkan Data Kualiti Udara Malaysia menggunakan Model Siri Masa Berstruktur Bayesian)

AESHAH MOHAMMED[1,2], MOHD AFTAR ABU BAKAR[1,*], MAHAYAUDIN M. MANSOR[3] & NORATIQAH MOHD ARIFF[1]

[1]Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
[2]Faculty of Science, University of Benghazi, AL Marje, Libya
[3]School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

Received: 9 February 2024/Accepted: 27 September 2024

ABSTRACT

Air pollution poses a significant threat to human health and the environment, especially in developing nations facing rapid industrialization, urbanization, and increased vehicle emissions. As cities and factories continue to grow, the air quality problem worsens, making it crucial to enhance the monitoring, testing, and forecasting of air quality. In this context, this study focuses on building air quality models using Bayesian Structural Time Series (BSTS) models to predict air quality levels in Malaysia. The BSTS model integrates three main techniques: The structural model, which employs the Kalman filter approach to model trend and seasonality components; spike and slab regression for variable selection; and Bayesian model averaging to estimate the best-performing prediction model while accounting for uncertainty. The study utilized air quality time-series data spanning two years, from June 2017 to July 2019, obtained from the Malaysian Department of Environment (DOE). The primary objective of this study was to forecast air quality and assess the effectiveness of the Bayesian structural time series analysis on air quality time-series data. The results indicated that the BSTS technique is capable of modeling air quality time-series data with high accuracy, effectively capturing seasonal and trend components. The seasonal component showed a repetition of weekly concentration patterns, while the local linear trend component showed a steady decline in PM_{10} and $PM_{2.5}$ concentration levels in most stations. Regression analysis demonstrated that humidity and ambient temperature significantly affected air quality in most locations in Malaysia.

Keywords: Air quality; Bayesian Structural Time Series; Monte Carlo Markov Chain (MCMC); spike and slab regression

ABSTRAK

Pencemaran udara menimbulkan ancaman besar kepada kesihatan manusia dan alam sekitar, terutamanya di negara membangun yang menghadapi perindustrian pesat, pembandaran dan peningkatan pelepasan kenderaan. Perkembangan bandar dan pertambahan kilang mengakibatkan masalah kualiti udara bertambah buruk, menjadikan pentingnya pemantauan, ujian dan ramalan kualiti udara. Dalam konteks ini, kajian ini tertumpu kepada pembinaan model kualiti udara menggunakan model Siri Masa Berstruktur Bayesian (BSTS) untuk meramalkan tahap kualiti udara di Malaysia. Model BSTS menyepadukan tiga teknik utama: Model struktur yang menggunakan pendekatan penapis Kalman untuk memodelkan komponen trend dan bermusim; regresi spike dan papak untuk pemilihan berubah; dan model Bayesian secara purata untuk menganggarkan model ramalan berprestasi terbaik sambil mengambil kira ketidakpastian. Kajian itu menggunakan data siri masa kualiti udara yang menjangkau dua tahun dari Jun 2017 hingga Julai 2019 yang diperoleh daripada Jabatan Alam Sekitar Malaysia (JAS). Objektif utama kajian ini adalah untuk meramal kualiti udara dan menilai keberkesanan analisis BSTS terhadap data siri masa kualiti udara. Keputusan menunjukkan bahawa teknik BSTS mampu memodelkan data siri masa kualiti udara dengan ketepatan yang tinggi, menangkap komponen bermusim dan trend dengan berkesan. Komponen bermusim menunjukkan pengulangan corak kepekatan mingguan, manakala komponen aliran linear tempatan menunjukkan penurunan yang stabil dalam tahap kepekatan PM_{10} dan $PM_{2.5}$ di kebanyakan stesen. Analisis regresi menunjukkan bahawa kelembapan dan suhu ambien menjejaskan kualiti udara dengan ketara di kebanyakan lokasi di Malaysia.

Kata kunci: Kualiti udara; Rantaian Markov Monte Carlo (MCMC); regresi pepaku dan papak; Siri Masa Berstruktur Bayesian

INTRODUCTION

Air quality is a critical global issue, with rising pollution levels impacting human health and the environment. According to the WHO, air pollution is a leading cause of premature deaths, contributing to respiratory and cardiovascular diseases. Rapid industrialization, urbanization, and increased vehicular emissions exacerbate these issues. In Malaysia, air pollution is managed by the Department of Environment under the Ministry of Environment and Water through the Air Pollutant Index of Malaysia (APIMS). The country faces challenges due to rapid economic development and urbanization, leading to frequent poor air quality episodes. Short-term air quality forecasting is essential for managing air pollution and protecting public health. It informs expected pollutant levels, allowing local governments to implement temporary measures and reduce emissions (Wen et al. 2024).

Many studies have been conducted to predict air quality and its severity. Bakar et al. (2022) used Network Temporal Convolution (TCN) to predict air pollution levels in Peninsular Malaysia. Based on suspended particle time series, TCN outperformed Long Short-Term Memory (LSTM) in terms of accuracy. Mun, Abd Rahman and Che Ilias (2022) evaluated Multi-Layer Perceptron (MLP) and AutoRegressive Integrated Moving Average (ARIMA) models to predict Air Pollution Index (API) in central Malaysia and found that MLP models performed better than ARIMA models. Zheng et al. (2023) found long-term relationships between climate parameters and air quality fluctuations in Peninsular Malaysia, using observational data from 2000 to 2019. Meanwhile, Ariff, Bakar and Lim (2023) predicted Malaysia's daily PM_{10} using a combination of k-means clustering and the LSTM model. Nasr Ahmed, Nurulkamal and Zamira Hasanah (2018) developed the compositional time series analysis method, which helps express air pollution data regarding the proportional term of each air pollutant component. Considering its structural and descriptive statuses, they proposed a mixed model corresponding to healthy and unhealthy conditions to characterize the distributional form of air pollution data. In contrast, Nurul Nnadiah et al. (2019) proposed a simple forecasting tool using the Markov chain model to evaluate the distribution of pollution levels in the long term. The model's probability indicates the probability of a good, moderate, or hazardous state. Similarly, Nurulkamal and Muhammad Aslam (2020) suggested using a five-state Markov chain model to study and characterize the dynamic fluctuation of air quality status with stochastic behaviours.

The models for predicting air quality have strengths and weaknesses. For instance, the Network Temporal Convolution (TCN) model is accurate but struggles with long-term dependencies and computational demands. The Multi-Layer Perceptron (MLP) models are adaptable to non-linear patterns but are data-intensive and prone to overfitting. The Auto-Regressive Integrated Moving Average (ARIMA) models fail with non-stationary data. The k-means clustering and Long Short-Term Memory (LSTM) combination improves performance but requires substantial tuning and computational resources. Compositional time series analysis, though structurally insightful, complicates direct prediction. Markov chain models offer simplicity and interpretability but fall short of capturing long-term dependencies and non-linear relationships. The BSTS offers a robust alternative by incorporating multiple components within a state-space framework, accommodating complex dependencies and providing interpretable results through a Bayesian approach, making it particularly suitable for the dynamic and uncertain nature of air quality data.

The Bayesian Structural Time Series (BSTS) models incorporate prior knowledge and data to provide accurate and probabilistic estimates. The BSTS model uses three main techniques to analyze time-series data: Kalman filtering for estimating trend and seasonality components, spike and slab regression for selecting relevant regressors, and Bayesian model averaging to determine the best prediction model while accounting for uncertainty. This comprehensive approach allows for the inclusion of multiple covariates and a more accurate description of stochastic behaviour, resulting in more robust parameter estimates and improved forecast accuracy (Durbin & Koopman 2002; George & McCulloch 1997; Madigan & Raftery 1994; Scott & Varian 2014; Volinsky et al. 1999).

The Bayesian structural time series (BSTS) model has been successfully applied in various fields. It outperformed the classical Auto-Regressive Integrated Moving Average (ARIMA) model in stock price forecasting (Almarashi & Khan 2020) and provided reliable long-term electricity demand forecasts for better planning in the energy sector (Mokilane et al. 2019). Additionally, Jun (2019) used the BSTS model and Bayesian regression to construct a sustainable technology structure for artificial intelligence, highlighting its reliability in analyzing and forecasting technology landscapes for improved decision-making.

This research aims to develop a reliable and accurate BSTS model for predicting air quality using historical data and variables like wind direction, wind speed, relative humidity, solar radiation, and ambient temperature. The process involves data collection, preprocessing, and Bayesian inference for parameter estimation. The goal was to improve air quality forecasting and support environmental management and policy decision-making. The remainder of this paper is organized as follows. The following section explains the Bayesian structural time-series method. The subsequent section presents an application to air quality data. Finally, the conclusions of this study are presented in the last section.

METHODS

BAYESIAN STRUCTURAL TIME SERIES

A Bayesian Structural Time Series (BSTS) model was introduced by Scott and Varian (2014) to predict economic time series. The model combines the Kalman filter for estimating trend and seasonality, spike and slab regression for identifying variables, and Bayesian model averaging to improve forecasting.

STRUCTURAL TIME SERIES

To utilize a structural time series model, it is essential to establish a set of equations that connect y_t to a vector of latent state variables α_t.

$$y_t = Z_t^T \alpha_t + \varepsilon_t \qquad \varepsilon_t \sim \mathrm{N}(0, H_t) \tag{1}$$

$$\alpha_{t+1} = T_t\, \alpha_t + R_t\, \eta_t \qquad \eta_t \sim N(0,\, Q_t) \tag{2}$$

Equation (1) is known as the observation equation, and Equation (2) is known as the state equation or transition equation which responsible for determining how the state variables will change over time. In this formulation, R_t is a rectangular matrix (control matrix) with dimensions $m \times q$, while T_t is a square transition matrix, Z_t is a fixed vector of size $m \times 1$. The observation error, ε_t , has a mean of zero and a variance of H_t (a positive scalar), following a normal distribution. An $m \times q$ state variance matrix is represented by Q_t (Brodersen et al. 2015). Structural time series, as defined by Durbin and Koopman (2002), are models in which the observations consists of trend, seasonal, and regression error variables.

The following state equations describe various components of the latent state. The local linear trend, the initial component of our model, is characterized by two equations. These equations capture the trend behaviour of the time series.

$$\mu_t = \mu_{t-1} + \delta_{t-1} + u_t \tag{3}$$

$$\delta_t = \delta_{t-1} + v_t \tag{4}$$

For seasonality, the most commonly used model is represented by

$$\gamma_t = -\sum_{s=1}^{S-1} \tau_{t-s} + w_t \tag{5}$$

where μ_t is the level of the time series; δ_t is the slope of the time series; γ_t is the seasonal component; and τ_{t-s} is the seasonal effect. The components $\eta_t = (u_t, v_t, w_t)$ are independent Gaussian random noise elements with a normal distribution and variances $(\sigma_u^2, \sigma_v^2, \sigma_w^2)$.

The seasonal component τ_t, comprises a series of S dummy variables with dynamic coefficients constrained to have zero expectation over an S-season period. In this study, we modelled daily data with an S of 52 and a season duration of 7. S=52 is chosen for modelling weekly seasonality over a year, assuming there are 52 weeks in a year. This choice helps capture the repeating weekly patterns in the data. State estimation is using the Kalman filter.

To incorporate external factors, explanatory variables x_t are added to the equation:

$$(y_t = \mu_t + \tau_t + \beta^T X_t + \varepsilon_t) \tag{6}$$

Parameters to estimate include regression coefficients (β) and error term variances ε_t.

KALMAN FILTER ESTIMATION

The Kalman Filter is a method used to estimate the state of a system using observed data and system dynamics. It operates through two main stages which are the prediction and update stages. The Kalman Filter can be implemented in the following steps (Kalman 1960):

Initialization The algorithm begins with an initial estimate of the state vector α_0 and its associated error covariance matrix P_0.

Prediction In the prediction step, the filter forecasts the state vector α_{t+1} and its associated error covariance matrix P_{t+1} using the provided equations, based on the current state vector estimate α_t and its error covariance matrix P_t.

$$\begin{aligned} \hat{\alpha}_{t+1|t} &= F\hat{\alpha}_{t|t} \\ P_{t+1|t} &= F\, P_t\, F' + Q \end{aligned} \tag{7}$$

where F is the state transition matrix; and Q is the process noise covariance matrix.

Update The filter integrates integration of the projected state with the new observation y_{t+1}, the filter enhances the estimation of the state vector α_{t+1} and its associated error covariance matrix P_{t+1} after the prediction step. The equations used to achieve this improvement are as follows:

$$\begin{aligned} K_{t+1} &= P_{t+1|t} H' (H\, P_{t+1|t} H' + R)^{-1} \\ \hat{\alpha}_{t+1|t+1} &= \hat{\alpha}_{t+1|t} + K_{t+1} (y_{t+1} - H\, \hat{\alpha}_{t+1|t}) \\ P_{t+1|t+1} &= (I - K_{t+1} H)\, P_{t+1|t} \end{aligned} \tag{8}$$

In these equations, R represents the measurement noise covariance matrix, which reflects the uncertainty related to the observation, and H represents the observation matrix that connects the state to the observations.

Repeat The prediction and update steps are iteratively applied at each time step to estimate the state vector and its associated error covariance matrix.

SPIKE-AND-SLAB REGRESSION

George and McCulloch (1997) and Madigan and Raftery (1994) proposed a method to select models based on spikes and slabs. This is used for estimating regression coefficients and selecting regressors, which can reduce the number of variables and incorporate prior opinions into the model. Spike and slab priors consist of two parts: A spike that determines the probability of a variable being chosen for the model, and a slab that shrinks these coefficients towards zero.

To set up the prior distributions, Zellner's g-prior is used, which supposes a normal distribution for the coefficients with mean zero and a variance-covariance matrix scaled by the scalar g times the error variance, represented as σ². The hyperparameter g determines the strength of the prior information: smaller values make the prior more informative, while larger values make it less informative (Brodersen et al. 2015; Zellner 1986). Zellner's g-prior is a preferred prior due to its balance between informativeness and flexibility. It allows for prior beliefs about coefficients to be incorporated while allowing data to influence the posterior distribution. This is useful in situations where prior knowledge about variables is needed.. Other priors might be too rigid or too diffuse, either overly constraining the coefficients or not providing enough regularization. Zellner's g-prior strikes a balance, making it a practical choice for many regression problems.

To joint distribution of $(\beta, \gamma, \sigma_\varepsilon^2)$ with a spike and slab prior is calculated using the following equation,

$$p\,(\beta, \gamma, \sigma_\varepsilon^2) = p\,(\beta_{\gamma} \mid \gamma, \sigma_\varepsilon^2)\; p\,(\sigma_\varepsilon^2 \mid \gamma)\; p(\gamma) \tag{9}$$

Spike Prior The spike is the marginal distribution $p(\gamma)$, which places the point mass at zero, representing a Bernoulli distribution.

$$\gamma \sim \prod_{k=1}^{k} \pi_k^{\gamma k} (1 - \pi_K)^{1-\gamma K} \tag{10}$$

where π_k is the probability of including k in the model, and it is often set to the ratio of the expected model size p (number of nonzero predictors) to the number of regressors K, $\pi_k = p/K$.

Slab Prior $p\,(\beta_{\gamma} \mid \gamma, \sigma_\varepsilon^2)$ For the slab prior, b is denoted as the vector of prior predictions about the regression coefficients› values β. In the case of the symmetric matrix Ω^{-1}, the rows and columns of Ω^{-1} that correspond to $\gamma_k = 1$ are denoted as Ω_{γ}^{-1}. The slab prior is given as

$$\beta_\gamma \mid \sigma_\varepsilon^2, \gamma \sim N(b_\gamma, \sigma_\varepsilon^2 (\Omega_\gamma^{-1})\text{-}1) \tag{11}$$

where Ω^{-1} is the symmetric full model, given by $\Omega^{-1} = k\,(w\mathbf{X}^{\mathrm{T}}\mathbf{X} + (1\text{-}w)\,\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X}))\,/n$. Here, $\mathbf{X}$ is the predictor matrix with row $\boldsymbol{x}_t$ corresponding to time t, w is typically set to 1/2, and k is set to 1. The matrix $\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})$ represents the diagonal matrix with diagonal elements equal to those of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Additionally, it is common to set the vector of prior means, b to zero. This simplifies the model by assuming no prior knowledge about the means of the predictors, allowing the data to fully inform the estimation process (Brodersen et al. 2015).

Error Variance Prior $p\,(\sigma_\varepsilon^2 \mid \gamma)$ The gamma distribution with mean r/s and variance r/s^2 is denoted by $Ga(r,s)$. Also, ss can be thought as a prior sum of squares, and df, which is the prior sample size in Equation (12). The prior sum can be calculated using the expected R^2 from the regression and df to determine the weight assigned to that guess. Then, $ss/\mathrm{df} = (1\text{-}R^2)s_y^2$, where s_y^2 is the marginal variance of the response. Meanwhile scaling by s_y^2 implies that our priors are data-determined. Scott and Varian (2014) acknowledge that this violates the Bayesian paradigm, but claim that it was effective in their applications.

$$\frac{1}{\sigma_\varepsilon^2} \mid \gamma \sim Ga\left(\frac{df}{2}, \frac{ss}{2}\right) \tag{12}$$

In summary, the spike and slab prior introduced here allow for significant flexibility in expressing prior beliefs through the parameters π_k, b, Ω^{-1}, ss, and df. For simplicity, prior information can be summarized using an expected model size, expected R^2, and a sample size df, or default values. This study used Bayesian structural time series with weakly informative priors (defaults) with values $R^2 = 0.5$, $\mathrm{df} = 0.01$, and $\pi_k = 0.5$. These values balance data influence and regularization, with R^2 indicating a moderate prior belief in explaining 50% of data variation, and df = 0.01 providing flexibility to accommodate unexpected data patterns. Using Bayesian inference with spike-and-slab priors facilitates effective Bayesian model averaging. Drawing from the posterior distribution provides multiple parameter sets for forecasting (y_{t+1}). Repeating this process yields an estimate of the posterior distribution.

RESULTS AND DISCUSSION

This study used structural time series analysis to investigate air quality, with air pollutant readings as the dependent variable and five explanatory variables: wind direction, wind speed, relative humidity, and temperature. Data was collected hourly from four Malaysian monitoring stations which comprises of rural, industrial, sub-urban,

and urban stations between July 2017 and June 2019. To create an accurate air quality model, relevant data is critical. The data was processed, and missing values were removed and averaged into daily measurements for four monitoring stations. It was split into training and test sets, with 90% in the former. A Bayesian structural time series model was fitted to the $PM_{2.5}$ and PM_{10} data using the 'bsts' package. The Monte Carlo Markov Chain (MCMC) was used to simulate posterior distributions over 1000 iterations, with the first 100 discarded as burn-in and the parameter estimates' uncertainty evaluated using posterior distributions derived from the MCMC simulation. By employing weakly informative default values (R^2 = 0.5, *df* = 0.01, and π_k = *0.5*), the Bayesian structural time series model can be initialized, and preliminary parameter estimates can be obtained.

As part of creating the state specification, a prior for the time series component is specified. The prior distributions for the variances (σ^2) are specified as gamma distributions. The prior sum of squares (*ss*) and prior sample size (*df*) were used to calculate the prior sum, based on the expected R^2 from the regression. Then, ss / df = (1-R^2) s_y^2 , where s_y^2 is the marginal variance of the response.

By scaling the prior sum of squares by $s_y^2 = \Sigma_t(y_t - \bar{y})^2 / (n-1)$ the prior variance becomes data-determined. This approach allows for the specification of the prior distribution for variables such as $PM_{2.5}$ and PM_{10} (Table 1). By using the prior sum of squares, a guess of the variance can be derived (*df/ss*) for the estimation process by using default prior for the variance σ^2, $\frac{1}{\sigma^2} \sim G(10^{-2}, 10^{-2}s_y^2)$.

The term 'prior guess' refers to the initial estimate for the standard deviation of a component before fitting a model. The 'prior.*df*' parameter represents the prior sample size or the degree of certainty associated with the prior guess. By default, the prior.*df* is set to 0.01, indicating a relatively low level of certainty or diffuse prior distribution.

The initial value is crucial for the model fitting process as it affects performance and convergence. To simplify the process, the initial values of the parameters are set to the prior guess values. An upper limit, calculated as (prior guess*/df*), is set to prevent unrealistic or unstable results during the model fitting process. If the estimated value exceeds the upper limit, it will be truncated to the upper limit value. This constraint helps ensure that the model produces reliable and interpretable results.

In Table 1, for each monitoring station and pollutant component, the model estimates three parameters: A level parameter (σ_u), a slope parameter (σ_v), and a seasonal parameter (σ_w). For example, The prior guess for each parameter in the CA13A station is 0.134 for $PM_{2.5}$ and 0.172 for PM_{10}. This pattern repeats for the CA05K, CA46D, and CA20B stations. noted that the time series components have the same standard deviation value before observing any data, this simplifies the initialization process and provides a convenient starting point for estimation. The prior degrees of freedom are the same (0.01) in all stations, accommodating a wider range of potential values. The initial values can be updated and refined during the estimation process as the model learns from the data. The upper limits for the Bayesian structural time series model parameters are set significantly higher than prior guesses to provide flexibility for capturing significant variations in $PM_{2.5}$ and PM_{10} levels. Historical data shows typical levels range from 5 to 50 μg/m³, and the chosen limits (13.4 for $PM_{2.5}$ in the CA13A station) are set 100 times higher than the prior guesses. This allows the model to handle extreme events and unexpected variations while balancing the risk of overfitting.

The specifications for the prior distribution of the initial value of the component in Table 2 are included in the model specifications that describe the state vector's prior distribution at time 1. It utilizes a Normal (scalar Gaussian) prior distribution. Table 2 displays the initial values and prior distributions for various components of a state space model that will be utilized to simulate air quality data from different monitoring stations. The parameter μ specifies the mean of the normal prior and sigma specifies the standard deviation. The initial value for the level in the state space model.

For instance, the first row shows that the monitoring station 'CA13A' has two pollutants: $PM_{2.5}$ and PM_{10}. The initial value of σ_u (level) is 32.18, and it follows a normal distribution with an average of 32.18 and a standard deviation of 13.4. Similarly, the initial value for σ_v (slope) is 0.014, and for σ_w (seasonal) is 0. Also, for PM_{10} the initial value of σ_u (level) is 41.24, and the prior for this parameter is a normal distribution with a mean of 41.24 and a standard deviation of 17.2. Similarly, the initial value for σ_v (slope) is 0.018, and for σ_w (seasonal) is 0. This pattern repeats for the CA05K, CA46D, and CA20B stations.

Noted that 'initial value' is equivalent to μ. In some cases, the initial value can be used as the prior mean (μ) because it is a good starting point for the prior distribution. This assumption assumes that the initial value is a reasonable estimate of the true underlying parameter. It is possible that in the specific context are referring to, the initial value is used as the prior mean (μ) for simplicity and as an initial assumption (Scott & Varian 2014).

In our modelling process, we will use Table 3 values to initialize the state space model and estimate the model parameters from the data. In this section, we set the spike and slab priors for the regression component so that all potential independent variables have an equal chance (50%) of being included in the model (π_k = *0.5*). We set

TABLE 1. The standard deviation prior distribution for each components and stations

Stations		Components	Prior guess	prior.*df*	initial.value	upper.limit
CA13A	$PM_{2.5}$	σ_u (level)	0.134	0.01	0.134	13.4
		σ_v(slope)	0.134	0.01	0.134	13.4
		σ_w(seasonal)	0.134	0.01	0.134	13.4
	PM_{10}	σ_u (level)	0.172	0.01	0.172	17.2
		σ_v(slope)	0.172	0.01	0.172	17.2
		σ_w(seasonal)	0.172	0.01	0.172	17.2
CA05K	$PM_{2.5}$	σ_u (level)	0.185	0.01	0.185	18.5
		σ_v(slope)	0.185	0.01	0.185	18.5
		σ_w(seasonal)	0.185	0.01	0.185	18.5
	PM_{10}	σ_u (level)	0.194	0.01	0.194	19.4
		σ_v(slope)	0.194	0.01	0.194	19.4
		σ_w(seasonal)	0.194	0.01	0.194	19.4
CA46D	$PM_{2.5}$	σ_u (level)	0.128	0.01	0.128	12.8
		σ_v(slope)	0.128	0.01	0.128	12.8
		σ_w(seasonal)	0.128	0.01	0.128	12.8
	PM_{10}	σ_u (level)	0.158	0.01	0.158	15.8
		σ_v(slope)	0.158	0.01	0.158	15.8
		σ_w(seasonal)	0.158	0.01	0.158	15.8
CA20B	$PM_{2.5}$	σ_u (level)	0.135	0.01	0.135	13.5
		σ_v(slope)	0.135	0.01	0.135	13.5
		σ_w(seasonal)	0.135	0.01	0.135	13.5
	PM_{10}	σ_u (level)	0.172	0.01	0.172	17.2
		σ_v(slope)	0.172	0.01	0.172	17.2
		σ_w(seasonal)	0.172	0.01	0.172	17.2

TABLE 2. Specification for the prior distribution of the initial values for each components and stations

Stations		Initial components	μ	Sigma	initial.value
CA13A	$PM_{2.5}$	σ_u (level)	32.18	13.4	32.18
		σ_v(slope)	0.014	13.4	0.014
		σ_w(seasonal)	0	13.4	0
	PM_{10}	σ_u (level)	41.24	17.2	41.24
		σ_v(slope)	0.018	17.2	0.018
		σ_w(seasonal)	0	17.2	0
CA05K	$PM_{2.5}$	σ_u (level)	16.32	18.5	16.32
		σ_v(slope)	0.029	18.5	0.029
		σ_w(seasonal)	0	18.5	0
	PM_{10}	σ_u (level)	25.92	19.4	25.92
		σ_v(slope)	0.028	19.4	0.028
		σ_w(seasonal)	0	19.4	0
CA46D	$PM_{2.5}$	σ_u (level)	3.808	12.8	3.808
		σ_v(slope)	0.009	12.8	0.009
		σ_w(seasonal)	0	12.8	0
	PM_{10}	σ_u (level)	7.215	15.8	7.215
		σ_v(slope)	0.020	15.8	0.020
		σ_w(seasonal)	0	15.8	0
CA20B	$PM_{2.5}$	σ_u (level)	41.99	13.5	41.99
		σ_v(slope)	-0.031	13.5	-0.031
		σ_w(seasonal)	0	13.5	0
	PM_{10}	σ_u (level)	59.89	17.2	59.89
		σ_v(slope)	-0.046	17.2	-0.046
		σ_w(seasonal)	0	17.2	0

TABLE 3. Spike and slab prior for each station

Stations		prior. inclusion.probabilities	μ	sigma. guess	prior. *df*	sigma. upper.limit
CA13A	$PM_{2.5}$	0.833	0	9.479	0.01	16.09
	PM_{10}	0.833	0	12.13	0.01	20.58
CA05K	$PM_{2.5}$	0.833	0	13.08	0.01	22.19
	PM_{10}	0.833	0	13.71	0.01	23.26
CA46D	$PM_{2.5}$	0.833	0	9.072	0.01	15.39
	PM_{10}	0.833	0	11.19	0.01	19.00
CA20B	$PM_{2.5}$	0.833	0	9.530	0.01	16.17
	PM_{10}	0.833	0	12.16	0.01	20.63

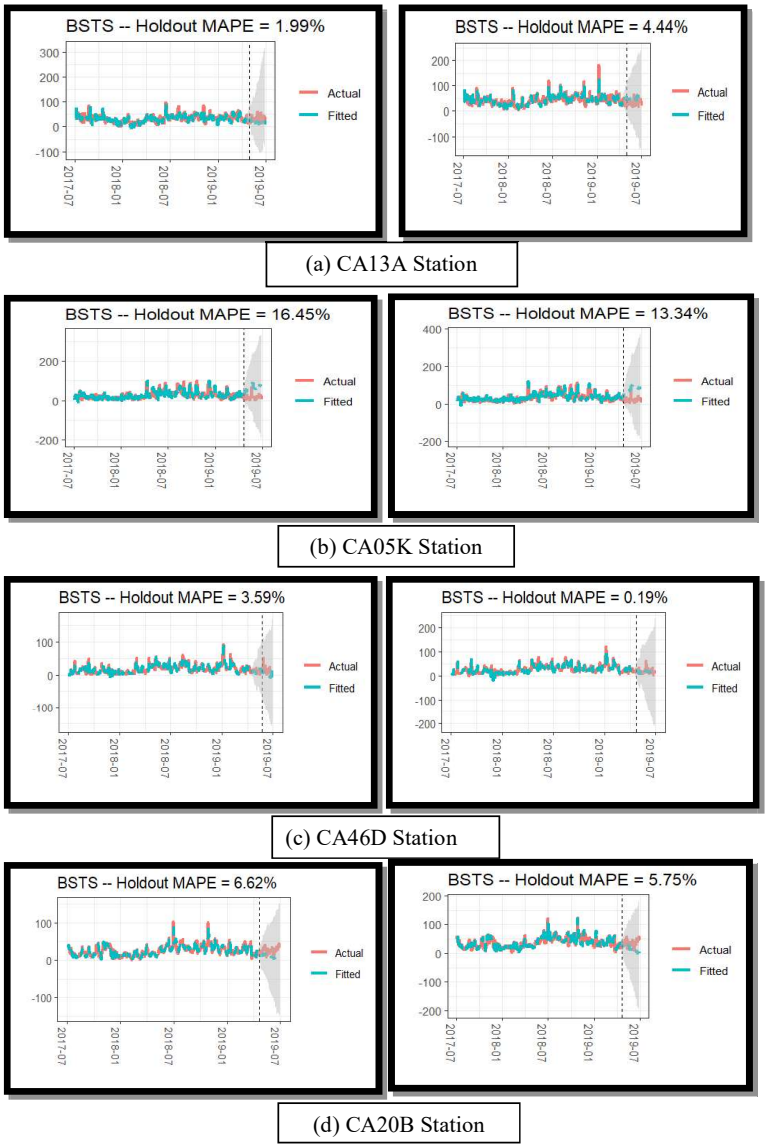

FIGURE 1. Comparison of actual with fitted values based on the BSTS odel for $PM_{2.5}$ and PM_{10} time series for each station

the model's prior for the overall variation explained to 0.5 ($R^2 = 0.5$), and the shrinkage parameter to 1% ($df = 0.01$). We employ the entire Bayes technique with slight variations as s_y^2 is determined by data (Scott & Varian 2014). Table 3 displays the results of the spike and slab regression model used to determine the prior distributions for each station. The priors will guide the estimation process by specifying the range of plausible values for each parameter.

The results for the spike and slab prior are displayed in Table 3. The table specifies the inclusion probabilities, means, and standard deviation of the regression coefficients for each pollutant and station in a Bayesian structural time series model. The inclusion probabilities indicate the prior probability that each regression coefficient is included in the model. In this case, all six regression coefficients have an equal inclusion probability of 0.833. The μ parameter is set to 0 for all regression coefficients, indicating a prior assumption that the predictors have no significant effect on the response variable. This is a conservative assumption when there is no prior information available about the coefficients' expected values. The sigma. guess parameter represents the prior guess for the error term's standard deviation (σ_ε) in the model. The prior.*df* parameter controls the shape of the gamma distribution for the error variance (σ_ε^2). In this case, prior.*df* is set to 0.01, indicating a diffuse prior distribution. The sigma. upper. limit parameter sets the upper limit for the standard deviation of the error term. Any values higher than this limit will be truncated at this value, preventing the prior from being too diffuse and allowing for unrealistic or uninformative values. These prior values inform the model and provide a starting point for the estimation of the posterior distribution of the regression coefficients and error variance.

After training the model, the actual and fitted values were compared to assess the model's performance. If the results are satisfactory, the model can be used to forecast the future values. The accuracy of the model's predictions were evaluated based on the mean absolute percentage error (MAPE). In Figure 1, the actual PM values were compared with the fitted values from the BSTS model. The shaded grey region is the 95% credible range for the forecast.

The performance of the Bayesian Structural Time Series (BSTS) model was evaluated in Figure 1. The BSTS model successfully captured seasonal patterns and trends in the data. Its predictions for the next 30 days were mostly accurate, with MAPE values ranging from 1.99% to 16.45%. The model showed a gradual decline in PM_{10} and $PM_{2.5}$ levels at CA05B stations, while fluctuating levels at CA13A and CA46D stations suggest instability in air quality.

By understanding how PM_{10} and $PM_{2.5}$ levels are likely to fluctuate at different locations, we can take targeted actions to reduce pollution and improve air quality in those areas. For more information about model evaluation, refer to Table 4. Table 4 compares the forecast accuracy of Bayesian Structural Time Series (BSTS) and Structural Time Series (STS) models for $PM_{2.5}$ and PM_{10} across multiple stations. Overall, BSTS generally outperforms STS, showing lower MAPE values, particularly at stations like CA13A and CA46D, and demonstrating better consistency with lower residual and prediction standard deviations at stations such as CA13A and CA05K. BSTS also tends to have a relative goodness of fit (GOF) closer to zero, indicating a better model fit. While STS occasionally outperforms BSTS, as seen at CA05K for $PM_{2.5}$, BSTS typically provides more accurate and reliable forecasts.

After training and making predictions, further analysis can be conducted to improve understanding of air quality time series data, including trend and seasonality analysis and regression analysis. This comprehensive approach aids in identifying factors impacting air quality and guiding the development of effective strategies for future improvement.

Figure 2 shows that air pollution levels have decreased at most stations in recent years due to better regulations and reduction efforts. Air quality data fluctuates seasonally, with higher pollution in summer and lower in winter. Random fluctuations in the residual component indicate unpredictable factors affecting air quality. Estimating

TABLE 4. Compare between measures of forecast accuracy for the BSTS and STS models

Station	model	MAPE		residual. *sd*		prediction. *sd*		Relative. gof		R^2	
		STS	BSTS	STS	BSTS	STS	BSTS	STS	BSTS	STS	BSTS
CA13A	$PM_{2.5}$	1.32	1.99	5.72	5.41	10.9	10.7	-0.27	-0.22	0.81	0.84
	PM_{10}	2.43	4.44	7.60	7.21	12.7	12.6	0.03	0.04	0.80	0.82
CA05K	$PM_{2.5}$	4.83	16.5	3.98	2.25	11.8	11.8	-0.06	-0.07	0.95	0.98
	PM_{10}	3.93	13.3	3.78	1.91	11.4	11.4	-0.08	-0.09	0.96	0.99
CA46D	$PM_{2.5}$	2.31	3.59	2.99	3.38	7.80	7.51	-0.14	-0.05	0.94	0.93
	PM_{10}	0.44	0.19	2.73	2.89	9.49	9.40	-0.12	-0.09	0.97	0.97
CA20B	$PM_{2.5}$	1.69	6.62	3,44	3.74	8.00	7.79	-0.07	-0.02	0.93	0.92
	PM_{10}	1.55	5.75	3.60	3.68	9.50	9.29	-0.08	-0.03	0.95	0.95

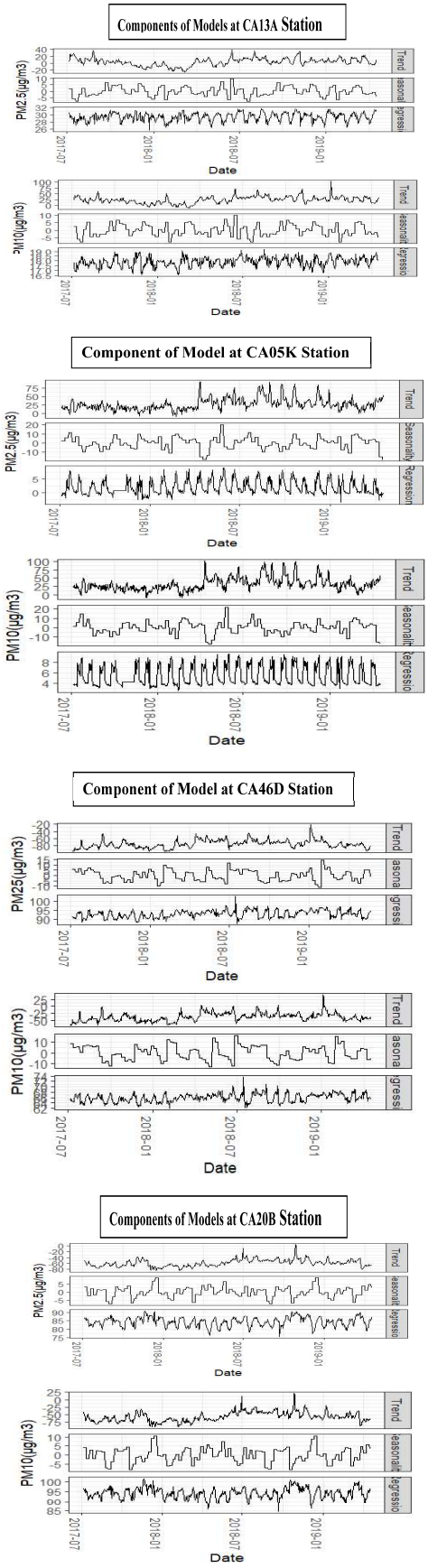

FIGURE 2. Components contribution to $PM_{2.5}$ and PM_{10} time series for each station

TABLE 5. Posterior distribution of the coefficient values (β) for the $PM_{2.5}$ and PM_{10} models for each station

Stations		Coefficients	mean	*Sd*	Incremental probability
CA13A	$PM_{2.5}$	Humidity	0.274	0.266	1
		Ambient.Temperature	0.907	1.084	0.69
		Solar.Radiation	-0.003	0.003	0.61
		Wind.speed	-0.119	0.231	0.23
		Wind.direction	0.000	0.000	0.00
		(Intercept)	0	0	0
	PM_{10}	Ambient.Temperature	0.445	0.796	0.64
		Humidity	0.067	0.172	0.52
		Wind. direction	0.003	0.006	0.27
		Wind.speed	-0.287	0.637	0.24
		Solar.Radiation	-0.0003	0.001	0.12
		(Intercept)	0	0	0
CA05K	$PM_{2.5}$	Wind.speed	-2.380	0.766	1
		Ambient.Temperature	0.425	0.766	0.84
		Solar.Radiation	-0.009	0.007	0.75
		Humidity	-0.079	0.093	0.73
		Wind.direction	-0.0002	0.001	0.11
		(Intercept)	0	0	0
	PM_{10}	Solar.Radiation	-0.015	0.006	0.93
		Humidity	0.004	0,093	0.74
		Ambient.Temperature	0.293	0.409	0.72
		Wind.speed	-0.498	0.915	0.31
		Wind.direction	0.0002	0.002	0.28
		(Intercept)	0	0	0
CA05B	$PM_{2.5}$	Humidity	0.512	0.146	0.99
		Ambient.Temperature	1.948	0.654	0.99
		Wind.direction	0.003	0.004	0.41
		Solar.Radiation	-0.0008	0.002	0.24
		Wind.speed	-0.114	0.324	0.18
		(Intercept)	0	0	0
	PM_{10}	Ambient.Temperature	1.928	0.684	1
		Humidity	0.531	0.141	1
		Solar.Radiation	-0.004	0.003	0.61
		Wind.direction	0.004	0.005	0.45
		Wind.speed	-0.109	0.332	0.19
		(Intercept)	0	0	0
CA46D	$PM_{2.5}$	Ambient.Temperature	2.183	0.399	1
		Humidity	5.086	0.099	1
		Wind.speed	8.872	0.187	0.12
		Wind.direction	-3.881	0.001	0.08
		Solar.Radiation	7.734	0.0004	0.06
		(Intercept)	0	0	0
	PM_{10}	Ambient.Temperature	1.378	0.571	0.99
		Humidity	2.738	0.156	0.94
		Wind.speed	4.931	0.719	0.41
		Wind.direction	-1.099	0.001	0.11
		Solar.Radiation	7.055	0.001	0.07
		(Intercept)	0	0	0

coefficients is key to understanding a model's dynamics, identifying influential variables, and making accurate predictions. These estimates show variable relationships and assess the model's effectiveness, as shown in Table 5.

Table 5 shows the posterior distribution of coefficient values (β) for models analyzing the impact of environmental variables on $PM_{2.5}$ and PM_{10} air pollution at stations CA13A, CA05B, CA05K, and CA46D. Each model has $PM_{2.5}$ and PM_{10} as dependent variables, with humidity, temperature, wind speed, wind direction, and solar radiation as independent variables. The table presents the estimated coefficients for each independent variable, along with their posterior means and standard deviations, and the probability that each coefficient is nonzero. The 'mean' column shows the average value for each coefficient. The '*sd'* column indicates the variation and uncertainty in these values, with larger standard deviations signifying greater uncertainty. The 'Incremental Probability' column displays the probability that each coefficient is non-zero, indicating its importance in predicting air pollution levels.

Table 5 presents an insightful analysis of how environmental variables affect $PM_{2.5}$ and PM_{10} levels across various stations. At most stations, variables like ambient temperature and humidity consistently show high inclusion probabilities, indicating their significant influence on air pollution levels. For example, at CA46D, ambient temperature and humidity have strong positive effects on $PM_{2.5}$, with high inclusion probabilities, emphasizing their crucial role. Conversely, some variables, like wind speed for $PM_{2.5}$ at CA13A, exhibit considerable uncertainty due to higher standard deviations. This variability highlights the complex and sometimes unpredictable nature of air pollution factors. Overall, the table shows that while certain variables are consistently important, their effects can vary significantly, reflecting the diverse influences on air quality.

CONCLUSIONS

This study aimed to create a dependable and precise model for predicting air quality in Malaysia using the Bayesian structural time series (BSTS) method. The study focused on two pollutants, PM_{10} and $PM_{2.5}$, and developed a model that included a regression component, a local linear trend component, and a seasonal component with a duration of 7 and a period of 52 weeks. The estimation of the model's parameters was done using Markov Chain Monte Carlo (MCMC) methods to calculate the posterior distribution. The model's effectiveness was evaluated using various measures.

The results of the study showed that the Bayesian structural time series approach was highly successful in modelling air quality in Malaysia. The model was able to accurately capture the seasonal and trend components of the air quality data and account for the impact of predictor variables on air quality, the study by Bakar et al. (2022) supports the finding that incorporating meteorological variables improves the performance of the model when used for LSTM models for predicting PM_{10}. The local linear trend component indicated a gradual decrease in PM_{10} and $PM_{2.5}$ concentrations over time in most stations, while the seasonal component showed weekly fluctuations in concentrations. The regression component showed that humidity and ambient temperature significantly affected air quality in most stations, whereas wind direction did not have a significant impact. Additionally, the model was able to accurately predict air quality for upcoming periods, making it a valuable tool for decision-making and planning aimed at improving air quality in Malaysia.

REFERENCES

Almarashi, A.M. & Khan, K. 2020. Bayesian structural time series. *Nanoscience and Nanotechnology Letters* 12(1): 54-61.

Ariff, N.M., Bakar, M.A.A. & Lim, H.Y. 2023. Prediction of PM_{10} concentration in Malaysia using k-means clustering and LSTM hybrid model. *Atmosphere* 14(5): 853.

Bakar, M.A.A., Ariff, N.M., Bakar, S.A., Chi, G.P. & Rajendran, R. 2022. Peramalan kualiti udara menggunakan kaedah pembelajaran mendalam rangkaian perlingkaran temporal (TCN). *Sains Malaysiana* 51(8): 2645-2654.

Bakar, M.A.A., Mohd Ariff, N.M., Mohd Nadzir, M.S., Wen, O.L. & Suris, F.N.A. 2022. Prediction of multivariate air quality time series data using long short-term memory network. *Malaysian Journal of Fundamental and Applied Sciences* 18(1): 52-59.

Brodersen, K.H., Gallusser, F., Koehler, J., Remy, N. & Scott, S.L. 2015. Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics* 9(1): 247-274.

Durbin, J. & Koopman, S.J. 2002. A simple and efficient simulation smoother for state space time series analysis. *Biometrika* 89(3): 603-615.

George, E. & McCulloch, R. 1997. Approaches for Bayesian variable selection. *Statistica Sinica* 7(2): 339-373.

Jun, S. 2019. Bayesian structural time series and regression modeling for sustainable technology management. *Sustainability* (Switzerland) 11(18): 4945.

Kalman, R.E. 1960. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering, Transactions of the ASME* 82(1): 35-45.

Madigan, D. & Raftery, A.E. 1994. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association* 89(428): 1535-1546.

Mokilane, P., Debba, P., Yadavalli, V. & Sigauke, C. 2019. Bayesian structural time-series approach to a long-term electricity demand forecasting. *Applied Mathematics and Information Sciences* 13: 189-199.

Mun, C.K., Abd Rahman, N.H. & Che Ilias, I.S. 2022. Performance of Levenberg-Marquardt neural network algorithm in air quality forecasting. *Sains Malaysiana* 51(8): 2645-2654.

Nasr Ahmed AL-Dhurafi, Nurulkamal Masseran & Zamira Hasanah Zamzuri. 2018. Compositional time series analysis for air pollution index data. *Stochastic Environmental Research and Risk Assessment* 32(10): 2903-2911. https://doi.org/10.1007/s00477-018-1542-0

Nurulkamal Masseran & Muhammad Aslam Mohd Safari. 2020. Modeling the transition behaviors of PM_{10} pollution index. *Environmental Monitoring and Assessment* 192: 441. https://api.semanticscholar.org/CorpusID:219729578

Nurul Nnadiah Zakaria, Mahmod Othman, Rajalingam Sokkalingam, Hanita Daud, Lazim Abdullah & Evizal Abdul Kadir. 2019. Markov chain model development for forecasting air pollution index of Miri, Sarawak. *Sustainability (Switzerland)* 11(19): 5190. https://doi.org/10.3390/su11195190

Scott, S.L. & Varian, H.R. 2014. Predicting the present with Bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation* 5(1-2): 4-23.

Volinsky, C.T., Raftery, A.E., Madigan, D. & Hoeting, J.A. 1999. David Draper and E.I. George, and a rejoinder by the authors. *Statistical Science* 14(4): 382-417.

Wen, Z., Ma, X., Xu, W., Si, R., Liu, L., Ma, M., Zhao, Y., Tang, A., Zhang, Y., Wang, K., Zhang, Y., Shen, J., Zhang, L., Zhao, Y., Zhang, F., Goulding, K. & Liu, X. 2024. Combined short-term and long-term emission controls improve air quality sustainably in China. *Nature Communications* 15(1): 5169.

Zellner, A. 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Stud. Bayesian Econometrics Statist*, edited by Goel, P.K. & Zellner, A. North-Holland Publishing Co., Amsterdam. 6: 233-243.

Zheng, Y., Ooi, M.C.G., Juneng, L., Wee, H.B., Latif, M.T., Nadzir, M.S.M., Hanif, N.M., Chan, A., Li, L., Ahmad, N. & Tangang, F. 2023. Assessing the impacts of climate variables on long-term air quality trends in Peninsular Malaysia. *Science of The Total Environment* 901: 166430.

*Corresponding author; email: aftar@ukm.edu.my