

University of Arkansas, Fayetteville
ScholarWorks@UARK

Theses and Dissertations


5-2019

A Bayesian Framework for Estimating Seismic Wave Arrival Time

Hua Zhong

University of Arkansas, Fayetteville

Follow this and additional works at: <https://scholarworks.uark.edu/etd>

 Part of the [Applied Statistics Commons](#), [Geology Commons](#), [Geophysics and Seismology Commons](#), [Longitudinal Data Analysis and Time Series Commons](#), [Statistical Models Commons](#), [Survival Analysis Commons](#), and the [Tectonics and Structure Commons](#)

Recommended Citation

Zhong, Hua, "A Bayesian Framework for Estimating Seismic Wave Arrival Time" (2019). *Theses and Dissertations*. 3282.
<https://scholarworks.uark.edu/etd/3282>

This Thesis is brought to you for free and open access by ScholarWorks@UARK. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of ScholarWorks@UARK. For more information, please contact ccmiddle@uark.edu.

A Bayesian Framework for Estimating Seismic Wave Arrival Time

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Statistics and Analysis

by

Hua Zhong
Wuchang University of Technology
Bachelor of Engineering in Environment, 2009

May 2019
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

John R. Tipton, Ph.D.
Thesis Director

Giovanni Petris, Ph.D.
Committee Member

Mark Arnold, Ph.D.
Committee Member

Abstract

Because earthquakes have a large impact on human society, statistical methods for better studying earthquakes are required. One characteristic of earthquakes is the arrival time of seismic waves at a seismic signal sensor. Once we can estimate the earthquake arrival time accurately, the earthquake location can be triangulated, and assistance can be sent to that area correctly. This study presents a Bayesian framework to predict the arrival time of seismic waves with associated uncertainty. We use a change point framework to model the different conditions before and after the seismic wave arrives. To evaluate the performance of the model, we conducted a simulation study where we could evaluate the predictive performance of the model framework. The results show that our method has acceptable performance of arrival time prediction with accounting for the uncertainty.

©2019 by Hua Zhong
All Rights Reserved

Acknowledgments

I want to thank Dr. John Tipton who is my thesis director for revising my thesis, for directing me to finish this project.

I also want to thank all my committee members, Dr. Giovanni Petris and Dr. Mark Arnold for giving me valuable advice.

This research is supported by the Arkansas High Performance Computing Center which is funded through multiple National Science Foundation grants and the Arkansas Economic Development Commission.

Table of Contents

1	Introduction.....	1
1.1	Background	1
1.2	Previous Work.....	3
2	Model Statement.....	6
2.1	Model Introduction	6
2.1.1	Moving Average Model	6
2.1.2	Autoregressive Model.....	7
2.1.3	Autoregressive-Moving-Average Model.....	8
2.1.4	Change Point Model for This Study	8
2.2	Bayesian Framework.....	10
2.2.1	Markov Chain Monte-Carlo.....	10
2.2.2	Metropolis-Hastings Algorithm	12
2.2.3	Stan Software	13
2.2.4	Hamiltonian Monte Carlo	13
2.2.5	Priors	15
2.3	Model Averaging and Estimation of τ	16
2.3.1	AIC Weight Model Averaging.....	16
2.3.2	Bayesian Model Averaging.....	18
3	Simulation Study	20
3.1	Simulation Method.....	20
3.2	Results	21
4	Discussion.....	30
	References.....	32

1 Introduction

1.1 Background

Earthquakes are incredibly powerful natural disasters that have large impacts on society. For example, a magnitude 8 earthquake occurred in 2008 in Sichuan, China and killed almost 69,225 people. In 2010, a 7.0 magnitude earthquake killed approximate 200,000 people in Haiti. Because the impact of some earthquakes is so large, there is a great need to learn as much as possible about these events.

From the earliest seismograph invented in China in 132 AD to the permanent global earthquake detection network stations built today, there have been many scientific advances. Today's earthquake researches can not only detect the orientation of the earthquake which the ancient seismograph did but can also detect the seismic waves and quantize them. In addition, there are many earthquake observatories around the world which monitor seismic activity today. Traditionally, the analysis of seismographs was done by hand. As these observatories are becoming more and more automated, observers are not required to stay in observatories all day. This automation is a result of the use of computers to record and analyze seismic data. Earthquakes produce two types of seismic waves: body waves and surface waves. Body waves travel through the interior of the earth whereas surface waves travel on the surface of the earth. In addition, body waves travel faster than surface waves. Body waves are composed of P waves and S waves (Shearer, 2003). Both P and S seismic wave signals are time series processes; however, P waves can travel through both liquids and solids, whereas S waves can only travel

P-waves and S-waves from a small (M4) earthquake that took place near Vancouver Island in 1997.

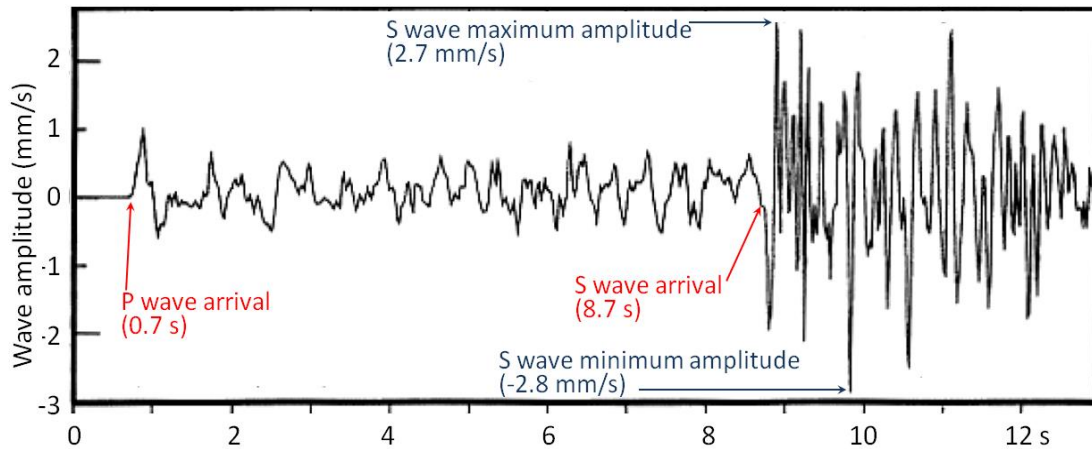


Fig. 1. Plot for a real seismic wave signal which contains S wave and P wave over a period of 12 seconds. The P wave and S wave arrival time are shown on the diagram.

through solids. P waves travel faster than S waves and do so with an example time series shown in Fig. 1 (Earle, 2016).

In order to detect the earthquake arrival time and locate where the earthquake occurred as quickly as possible, seismic waves are monitored with sensors at many locations around the world resulting in an increasing amount of seismic data. Because of the increase in data collection, there is a need for development of new statistical models to replace the traditional inference by eye. An important characteristic of seismic waves is arrival time. The arrival time is the time at which a seismic signal is first detected. If the arrival time of an earthquake is estimated with precision, the earthquake's location could be triangulated with high accuracy. Therefore, accurate and precise estimation of arrival time is vital for earthquake warning and attribution.

This thesis focuses on estimating the arrival time of seismic waves using a change point

time series model. Because we are uncertain about the model and parameters, we account for model uncertainty by fitting a number of models and then applying Bayesian model averaging (BMA) over the model set and parameter uncertainty is accounted for by using a Bayesian posterior. The final estimate of arrival time is estimated with uncertainty by fitting the models over a grid of candidate models and applying a second iteration of BMA over the range of possible change points.

1.2 Previous Work

To motivate the model, we introduce prior work on the statistical modeling of S wave arrival times using the multi-variate locally stationary autoregressive (Takanami & Kitagawa, 1991). Takanami and Kitagawa (1991) built models for describing background noise and the signal associated with an S wave arrival and define the arrival time as the change-point between these models.

The seismic signal data in Takanami and Kitagawa (1991) is the amplitude of two different seismic wave frequencies. The background noise model defines the seismic wave signal after the arrival of the faster P wave but before the S wave is detected. Because the background noise is the seismic signal before S wave is detected, the background model includes white noise and the tail of P wave.

The background noise model is:

$$\mathbf{y}_t = \sum_{i=1}^{p_1} \boldsymbol{\theta}_{i1} \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_{1t}, \quad (t=1, \dots, \tau) \quad (1)$$

where \mathbf{y}_t is the observation at time t . $\boldsymbol{\varepsilon}_{1t}$ is white noise which has mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_1$, $\boldsymbol{\theta}_{i1}$ is the autoregressive coefficient matrix of the model for the tail of P wave, τ is

the S wave arrival time, and p_1 is the order of the model for tail of P wave.

Because P wave moves faster than S waves, Takanami and Kitagawa (1991) assumed the tail of P wave is dominated by the S wave. Thus, the signal model combines S wave signal and an uncorrelated white noise giving rise to the signal model:

$$\mathbf{y}_t = \sum_{i=1}^{p_2} \boldsymbol{\theta}_{i2} \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_{2t}, \quad (t = \tau + 1, \dots, T) \quad (2)$$

where the white noise $\boldsymbol{\varepsilon}_{2t}$ has mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_2$, $\boldsymbol{\theta}_{i2}$ is an autoregressive coefficient matrix for the S wave, τ is the S wave arrival time, and p_2 is the order of the model for S wave.

Takanami and Kitagawa (1991) fit their model for each fixed τ , calculated the Akaike information criterion (AIC) value across the possible arrival times τ , and selected the arrival time based on the model with minimum AIC value. However, uncertainty exists in model selection and these authors selected a single model based on the lowest AIC value, not accounting for uncertainty. A possible approach to account for uncertainty is to apply AIC weighting over the arrival time. However, this has been shown to have poor performance in model selection (Link & Barker, 2006) and does not result in a probabilistic estimate of uncertainty. Similar to Takanami and Kitagawa (1991), we model the seismic activity using a change point model which includes an autoregressive and moving average model (ARMA) structure for the seismic wave process; however, we use an uncorrelated Gaussian white noise model for the background process. To properly account for model and arrival time uncertainty, we employ Bayesian model averaging (BMA), leading to a computationally efficient parallelizable frame work for estimation (Hoeting et al., 1999). Furthermore, if BMA estimation

were used to obtain three arrival time interval estimates from different locations, interval estimates could be used to locate the earthquake epicenter with associated uncertainty using triangulation methods. In contrast, methods that generate point estimates for the arrival time (like estimating the change point with Akaike information criterion (AIC)) have a low probability of producing a triangulated estimate that covers the true earthquake location due to a lack of formal uncertainty propagation. The contributions of this thesis are a modeling framework that accounts for uncertainty in the arrival time estimate while additionally accounting for model uncertainty and an improvement of triangulation for area estimation which uses BMA interval estimates.

In the section 2, we present the model framework and show how we use BMA to estimate the arrival time of seismic waves. In the section 3, we outline a simulation study that demonstrates model performance empirically. Finally, we discuss the result of our study and present avenues for future research.

2 Model Statement

2.1 Model Introduction

The statistical modeling of observations indexed in time is called time series (Chatfield, 2003). Based on this definition, the seismic signal collected by a seismograph is a time series. Takanami and Kitagawa (1991), Hayah and Kim (2013), and Colombelli (2014) have used methods from time series to study earthquakes, and we use these methods as inspiration for our work.

The models used in this study assume the time series is stationary. A second order stationary time series is a time series which has no trend, no seasonality and constant variance (Chatfield, 2003). Specifically, if a distribution $\{y_1, y_2, \dots, y_t\}$ is a second order stationary time series, the mean and variance of this distribution will be constant over time t , and the correlation between y_a and y_{a+h} ($1 < a, a + h < t$) only depends on the interval h . In a strictly stationary time series, the distribution of $\{y_1, y_2, \dots, y_a\}$ is the same as $\{y_{1+h}, y_{2+h}, \dots, y_{a+h}\}$ for any choice of a and h . Many models can be used in stationary time series analysis including the moving average model (MA), the autoregressive model (AR), and the autoregressive-moving-average model (ARMA).

2.1.1 Moving Average Model

The moving average (MA) model can be presented as the current white noise innovation plus a weighted sum of past innovations where the weights are called coefficients.

Moving average models of order q are given by:

$$y_t = \sum_{i=1}^q \beta_i \varepsilon_{t-i} + \varepsilon_t, \quad (3)$$

where ε_t is the current white noise where $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_q)$ are moving average coefficients, y_t is the observation at time t , and q is the order of MA model.

The autocorrelation function (ACF) is an important tool in evaluating the model order of a stationary time series. The empirical ACF estimates the correlation between points at lag k in a time series process by:

$$r_k = \frac{\sum_{t=k+1}^n (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^n (y_t - \bar{y})^2}, k = 1, 2, \dots \quad (4)$$

where y_t is the observation at time t , \bar{y} is the mean of observation, and n is the sample size.

In a time series process, the cut-off of the empirical ACF can be used to determine the order of MA model. For example, if the empirical ACF is between $-\frac{2}{\sqrt{n}}$ to $\frac{2}{\sqrt{n}}$ after lag q , we can use a MA(q) to model this process.

2.1.2 Autoregressive Model

The autoregressive (AR) model describes a time series model where the current observation is regressed onto past observed values with additional uncorrelated white noise.

The autoregressive models of order p is:

$$y_t = \sum_{i=1}^p \phi_i y_{t-i} + \eta_t, \quad (5)$$

where $\eta_t \sim N(0, \sigma_\eta^2)$ is uncorrelated white noise, $\boldsymbol{\phi} = (\phi_1, \phi_2, \dots, \phi_p)$ are autoregressive coefficients, y_t is the observation at time t , and p is the order of AR model.

The empirical partial autocorrelation (pACF) can be used as a tool to estimate the order of the AR model in a time series process. The pACF is the ACF between points separated at lag k in time conditional on the linear correlation for all observations with lag less than k . The empirical pACF at lag k is:

$$\rho_k = \frac{\text{Covariance}(y_t, y_{t-k} | y_{t-1}, y_{t-2}, \dots, y_{t-k+1})}{\sqrt{\text{Variance}(y_t | y_{t-1}, y_{t-2}, \dots, y_{t-k+1}) \text{Variance}(y_{t-k} | y_{t-1}, y_{t-2}, \dots, y_{t-k+1})}}, \quad (6)$$

where y_t is the observation at time t . If the empirical pACF is between $-\frac{2}{\sqrt{n}}$ to $\frac{2}{\sqrt{n}}$ after lag p , then the AR(p) model is a reasonable choice for the time series.

2.1.3 Autoregressive-Moving-Average Model

The Autoregressive-moving-average model (ARMA) is constructed by combining an autoregressive (AR) model and a moving average (MA) model (Cryer & Chan, 2008).

The ARMA model of order p and q is :

$$y_t = \sum_{i=1}^p \phi_i y_{t-i} + \sum_{i=1}^q \beta_i \xi_{t-i} + \xi_t, \quad (7)$$

where $\boldsymbol{\phi} = (\phi_1, \phi_2, \dots, \phi_p)$ are the autoregressive coefficients of the AR portion of the model and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_q)$ are the coefficients of the MA portion of the model, p is the order of AR portion, q is the order of MA portion, and ξ_t are uncorrelated white noise error terms with mean 0 and variance σ_ξ^2 .

2.1.4 Change Point Model for This Study

We assume that prior to the arrival of the seismic wave, the background process is a white noise process because the background process is a random process that has the same intensity at various frequencies. After the seismic signal arrives at the seismograph, the seismic signal becomes stronger and can be detected. Although the seismic signal is non-stationary during a long time interval, it is approximately stationary during a short time interval (Ozaki & Tong, 1975). Thus, we assume the seismic signal can be represented as a stationary time series process during a short time interval.

Our model combines two equations, the white noise process model which describes the seismic signal before the arrival time, and an ARMA model which describes the process after the arrival time. Therefore, we use a change point model to model the change of condition which are before and after the arrival time. Our model framework is the following:

Background Noise Model:

Before the arrival of the earthquake signal, we assume the signal is a white noise process:

$$y_t = \epsilon_t, \quad \text{if } t \leq \tau \quad (8)$$

Where the background noise observations are defined as $\{y_t\}$, $\epsilon_t \sim N(0, \sigma_\epsilon^2)$ is uncorrelated white noise, and τ is defined as the earthquake arrival time.

Signal Model:

Because we assumed the seismic signal is stationary for small timescales, an ARMA model can be used to describe the signal. The signal model describes the seismic wave process after the arrival time τ .

$$y_t = \sum_{i=1}^p \phi_i y_{t-i} + \sum_{i=1}^q \beta_i \delta_{t-i} + \delta_t, \quad \text{if } t > \tau \quad (9)$$

In the above equation, $\boldsymbol{\phi} = (\phi_1, \phi_2, \dots, \phi_p)$ are the parameters of the AR portion of the model and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_q)$ are the parameters of the MA portion of the model, p is the order of AR portion, q is the order of MA portion, and δ_t are uncorrelated Gaussian white noise error with mean 0 and variance σ_δ^2 .

In order to simplify the calculation when we fit this model, we noted that the ARMA(p, q) process can be represented as AR(∞) model (Chatfield, 2003). Because the AR

model is easier to fit than a ARMA model using Markov chain Monte Carlo (MCMC), we decided to use $AR(\infty)$ representation.

Using the $AR(\infty)$ representation, the signal model is:

$$y_t = \sum_{i=1}^{\infty} \phi_i y_{t-i} + \eta_t, \quad \text{if } t > \tau \quad (10)$$

where $\eta_t \sim N(0, \sigma_\eta^2)$ is uncorrelated white noise and $\{\phi_i\}$ are autoregressive coefficients.

In practice, we are unable to fit an $AR(\infty)$ model with an infinite number of parameters, thus we approximate the $AR(\infty)$ model with an $AR(p)$ model where the best choice of p is unknown.

The full change point model given τ and p is:

Change Point Model:

$$y_t = \begin{cases} \epsilon_t, & \text{if } t \leq \tau \\ \sum_{i=1}^p \phi_i y_{t-i} + \eta_t, & \text{if } t > \tau \end{cases} \quad (11)$$

2.2 Bayesian Framework

2.2.1 Markov Chain Monte-Carlo

To obtain estimates of the location of the change point with associated uncertainty, we used a Bayesian framework. In order to estimate the posterior distribution for each parameter, we used Markov chain Monte Carlo (MCMC) because MCMC is an efficient approach to draw samples from the posterior distributions which are intractable (Ravenszwaaij, Cassey & Brown, 2016).

A Markov Chain is a memoryless stochastic process where the current state only depends on values of the last state. For example, the generated random sequence of states up to iteration n in the stochastic process are $X = \{x_1, x_2, \dots, x_n\}$. Because of the Markov property, we have

$$p(x_n | x_{n-1}, x_{n-2}, \dots, x_1) = p(x_n | x_{n-1}). \quad (12)$$

The posterior probability density $p(X)$ is:

$$p(X) = p(x_1) \prod_{i=2}^n p(x_i | x_{i-1}). \quad (13)$$

In addition, the Markov Chain is stationary which means the transition probabilities are unchanged at different positions in the chain. For any different points n and n^* in the Markov Chain, $k(x_n | x_{n-1})$ is equal to $k(x_{n^*} | x_{n^*-1})$ which means that the probability x_{n-1} transitions to x_n is same as the probability x_{n-1^*} transitions to x_{n^*} . We only use the samples from the stationary Markov Chain to implement Monte-Carlo estimate, which in practice means we eliminate the initial MCMC iterations as burn-in.

The Monte-Carlo method allows for numeric calculation of functions of a probability distribution using a large number of samples when these functions are difficult to calculate analytically. For example, the mean μ of a probability density $p(x)$ is:

$$\mu = \int_x xp(x)dx. \quad (14)$$

Sometimes, the integral of (14) may be difficult to calculate analytically. In these cases, a Monte-Carlo method estimate can be calculated by drawing a large number of samples n from the probability density $p(x)$. Then, the samples are x_1, x_2, \dots, x_n , and the Monte Carlo estimate of mean $\hat{\mu}$ is:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (15)$$

As the number of Monte Carlo samples increases,

$$\lim_{n \rightarrow \infty} \hat{\mu} = \mu. \quad (16)$$

2.2.2 Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is a Markov Chain sampling method that produces stationary Markov chains (Robert & Casella, 2010). In order to converge the correct distribution, it is necessary for the Markov chain to be stationary when we implement MCMC sampling. A Markov chain satisfying the detailed balance relationship:

$$p(x^*)k(x | x^*) = p(x)k(x^* | x) \quad (17)$$

is stationary, where $p(x^*)$ and $p(x)$ are target densities, $k(x^* | x)$ and $k(x | x^*)$ are transition probabilities. If equation (17) is true, then the Markov chain is stationary. For many algorithms, equation (17) may not be satisfied. Because of that, the MCMC will converge to an incorrect distribution. In order to solve this problem, the Metropolis-Hastings algorithm was proposed by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) and promoted by Hastings (1970). This algorithm uses the acceptance rate $\alpha(x^*)$ to determine whether to move x to a new state x^* :

$$\alpha(x^*) = \min\left(1, \frac{p(x^*)}{p(x)} \times \frac{q(x|x^*)}{q(x^*|x)}\right), \quad (18)$$

where $q(x^* | x)$ and $q(x | x^*)$ are proposal distributions.

Thus, we have

$$\begin{aligned} p(x)q(x^* | x)\alpha(x^*) &= p(x)q(x^* | x)\min\left(1, \frac{p(x^*)}{p(x)} \times \frac{q(x | x^*)}{q(x^* | x)}\right) \\ &= \min(p(x)q(x^* | x), p(x^*)q(x | x^*)) \\ &= p(x^*)q(x | x^*)\min\left(1, \frac{p(x)}{p(x^*)} \times \frac{q(x^* | x)}{q(x | x^*)}\right) \\ &= p(x^*)q(x | x^*)\alpha(x), \end{aligned} \quad (19)$$

which satisfies the detailed balance. Therefore, the Metropolis-Hastings algorithm uses the acceptance rate to make the Markov chain stationary when implementing MCMC sampling to guarantee the Markov Chain converges to the correct distribution. To evaluate a proposed jump from state x to x^* , we draw a random number u from $\text{Unif}(0, 1)$ and compare with $\alpha(x^*)$,

$$\begin{cases} x = x^*, & (u \leq \alpha(x^*)) \\ x = x, & (u > \alpha(x^*)) \end{cases} \quad (20)$$

if $u \leq \alpha(x^*)$, then accept x move to x^* , where x^* is a new generated proposal state. If $u > \alpha(x^*)$, x is set to be x without movement.

2.2.3 Stan Software

In our study, we implemented MCMC sampling using the Stan software. Stan is a mature platform for statistical modeling and computation that is widely used in engineering, sociology, biology and business. There are various probability functions and algebra in Stan's math library. Stan also can interface with the current data analysis languages such as R, Python, Bash, MATLAB and Julia. In addition, Stan is an open-source software which can work on the current main platforms such as Linux, Mac and Windows (Stan Development Team, 2014).

2.2.4 Hamiltonian Monte Carlo

Stan uses Hamiltonian Monte Carlo (HMC) for MCMC sampling (Stan Development Team, 2014). HMC is effective in preventing random walk behavior and increasing the computational efficiency in terms of effective sample size per second (Hoffman & Gelman, 2014). The Hamiltonian method uses auxiliary momentum variables \mathbf{x} and draws from the joint density:

$$P(\mathbf{x}, \mathbf{Q}) = P(\mathbf{x} | \mathbf{Q})P(\mathbf{Q}), \quad (21)$$

where $P(\mathbf{x} | \mathbf{Q})$ is the density of the auxiliary momentum variables conditional on the current state \mathbf{Q} and $P(\mathbf{Q})$ is the probability density for parameter \mathbf{Q} . The joint density $P(\mathbf{x}, \mathbf{Q})$ can be defined as a Hamiltonian function:

$$\begin{aligned}
H(\mathbf{x}, \mathbf{Q}) &= -\log P(\mathbf{x}, \mathbf{Q}) \\
&= -\log P(\mathbf{x} | \mathbf{Q}) - \log P(\mathbf{Q}) \\
&= K(\mathbf{x} | \mathbf{Q}) + U(\mathbf{x}, \mathbf{Q}),
\end{aligned} \tag{22}$$

where $K(\mathbf{x} | \mathbf{Q}) = -\log P(\mathbf{x} | \mathbf{Q})$ is called the kinetic energy, $U(\mathbf{x}, \mathbf{Q}) = -\log P(\mathbf{Q})$ is called the potential energy (typically called the log-posterior in statistics), and $H(\mathbf{x}, \mathbf{Q})$ is called the total energy. In this study, $U(\mathbf{x}, \mathbf{Q})$ represents the likelihood. The \mathbf{x} are called momentum variables and following a multivariate normal distribution with mean $\mathbf{0}$ and variance Σ that is learned during the estimation algorithm.

In a dynamic system, $H(\mathbf{x}, \mathbf{Q})$ is defined by a current parameter \mathbf{Q} and new momentum \mathbf{x} . The dynamic Hamiltonian equations are:

$$\begin{cases} \frac{d\mathbf{Q}}{dt} = +\frac{\partial K}{\partial \mathbf{x}} \\ \frac{d\mathbf{x}}{dt} = -\frac{\partial U}{\partial \mathbf{Q}} \end{cases} \tag{23}$$

The leapfrog integrator algorithm that approximates Hamiltonian dynamics updates, the momentum and position by repeating the algorithm below n times:

$$\begin{cases} \mathbf{x} \leftarrow \mathbf{x} - \varepsilon \frac{1}{2} \frac{\partial U}{\partial \mathbf{Q}} \\ \mathbf{Q} \leftarrow \mathbf{Q} + \varepsilon \Sigma \mathbf{x} \\ \mathbf{x} \leftarrow \mathbf{x} - \varepsilon \frac{1}{2} \frac{\partial U}{\partial \mathbf{Q}}, \end{cases} \tag{24}$$

where ε is the step size of the leapfrog algorithm. If the Hamiltonian algorithm implements T leapfrog steps, the overall computation time will be $O(nT)$.

After T iterations of equation (23), the new state $H(\mathbf{x}^*, \mathbf{Q}^*)$ will be obtained. Similar to the Metropolis-Hastings' acceptance rate, the new state $H(\mathbf{x}^*, \mathbf{Q}^*)$ is accepted with probability:

$$\min(1, \exp(H(\mathbf{x}, \mathbf{Q}) - H(\mathbf{x}^*, \mathbf{Q}^*))). \quad (25)$$

If the new state $H(\mathbf{x}^*, \mathbf{Q}^*)$ is not accepted, the previous value of the parameter will be used for the iteration. Although the computation of HMC method is more complicated than MCMC method, HMC can generate less correlated samples than Metropolis-Hastings, resulting in a more efficient algorithm in terms of effective sample size per second.

2.2.5 Priors

In order to fit model (11) within a Bayesian framework, we defined priors for the unknown parameters $\boldsymbol{\phi}, \sigma_\varepsilon^2, \sigma_\eta^2$. We assigned σ_ε^2 and σ_η^2 inverse-Gamma ($\alpha_0 = 0.5, \beta_0 = 0.5$) priors. Because it is difficult to specify priors on the AR $\boldsymbol{\phi}$ that make the time series causal, we instead assign priors on a transformation of $\boldsymbol{\phi}$. If we apply uniform prior support on the partial autocorrelation function (pACF) parameters $\tilde{\boldsymbol{\phi}}$ and then transform the pACF parameters to the autocorrelation function (ACF) parameters, any $\boldsymbol{\phi}$ we obtain under that condition will provide a causal time series process. Therefore, the uniform priors on the pACF parameters induce a causal prior on $\boldsymbol{\phi}$ giving rise to a causal time series. The pACF to ACF formula for AR(p) model (Monahan, 1984) is:

$$\phi_i^{(k)} = \phi_i^{(k-1)} + \tilde{\phi}_k^{(k)} \phi_{k-i}^{(k-1)}, \quad (i = 1, \dots, k-1) \quad (26)$$

where $k = 1, \dots, p$, $\{\phi_i^{(k)}\} = \boldsymbol{\phi}$ are the coefficients of AR(p) model, $\tilde{\phi}_k^{(k)}$ is the partial autocorrelations at lag k where $|\tilde{\phi}_k^{(k)}| < 1$ and $\{\tilde{\phi}_k^{(k)}\} = \tilde{\boldsymbol{\phi}}$

2.3 Model Averaging and Estimation of τ

In practice, model uncertainty exists if the true model is unknown. For example, if the true data generating process is an AR(5) model but we fit an AR(3) model, there will be unaccounted for uncertainty due to an error in model choice. We could use AIC value to select the best model for a data set. However, if we use Akaike information criterion (AIC) to choose the best model from to from a set of candidate models (Banks & Joyner, 2017), an error in model selection still could be made.

The AIC value for each model is:

$$AIC_i = -2\hat{L}_i + 2k_i, \quad (27)$$

where \hat{L}_i are the log of maximum likelihood for each model and k_i are the number of parameters for each model.

In our study, the correct model for the seismic signal is unknown and model uncertainty needs to be accounted for. In addition, we are fitting a fixed change point model where the true change point is unknown. Thus, it is necessary to account for uncertainty when estimate the arrival time in our study. There are two methods that can be used to overcome the model uncertainty: AIC model averaging and Bayesian model averaging.

2.3.1 AIC Model Averaging

AIC model averaging is a method to account for uncertainty. AIC model averaging weights different models by AIC weight. In order to calculate the AIC weights, the AIC difference is computed:

$$\Delta AIC_i = AIC_i - AIC_{min}, \quad (28)$$

where AIC_{min} is the minimum of $\{AIC_1, AIC_2, \dots, AIC_n\}$ of the models under consideration, AIC_i is the AIC value for the i th model and ΔAIC_i is the difference in AIC between the i -th model and the model with the minimum AIC.

Using these model weights, the relative likelihood for model M_i is:

$$L(M_i | y) \propto \exp\left(-\frac{1}{2} \Delta AIC_i\right), \quad (29)$$

where y is the data. The AIC weight for model M_i is:

$$w_{r|\tau} = \frac{\exp\left(-\frac{1}{2} \Delta AIC_r\right)}{\sum_{i=1}^R \exp\left(-\frac{1}{2} \Delta AIC_i\right)}, \quad (30)$$

where R is the total number of models under consideration.

Therefore, in our study, if we used AR(2) to AR(20) models to fit the data and used AIC weight model averaging to account for model uncertainty, the estimate of arrival time $\tilde{\tau}$ is:

$$\tilde{\tau} = \sum_{\tau=1}^{T=500} \sum_{r=2}^{R=20} w_{r|\tau} \tau, \quad (31)$$

where $w_{r|\tau}$ are the AIC weights for each candidate model given the change point τ , τ is all possible change points which is from 1 to 500 and T is the total 500 time point in this time series.

The use of AIC weight induces the ‘‘K-L (Kullback-Leibler) prior’’ which makes the AIC weights approximate to the posterior model probability distribution (Burnham & Anderson, 2004). Therefore, the posterior model probability distribution of AIC weight is:

$$w_i = P(M_i | \mathbf{y}) \approx \frac{\exp\left(-\frac{1}{2} AIC_i\right)}{\sum_{r=1}^R \exp\left(-\frac{1}{2} AIC_r\right)} \quad (32)$$

However, the K-L prior has strong effect on the AIC weight, and the AIC weight tends to favor models with more parameters (Link & Barker, 2006). Due to this fact, the use of AIC weights for model averaging would tend to choose models with larger values of the autoregressive model order p which may cause bias. Thus, we present an alternative method to account for the uncertainty.

2.3.2 Bayesian Model Averaging

Bayesian model averaging (BMA) is an approach that accounts for the uncertainty in model choice (Hoeting et al., 1999). BMA averages the posterior distribution over the model set by weighting posterior model probability for each model under consideration. Madigan and Raftery (1994) also noted that the BMA prediction will often be better than using a single model.

Our goal is to estimate τ . Due to the uncertainty of the value of the autoregressive model order p in our model, we perform a grid search over set of all possible p from 2 to $R = 20$ instead of choosing the optimum model with fixed τ . Therefore, we used AR(2) to AR(20) models to fit the data and used BMA to account for model uncertainty, reweighting models based on how well they fit with the data.

For estimating arrival time τ with uncertainty, we need to compute the posterior distribution of the estimate arrival time τ . We fit a model for each possible value of τ , then perform model averaging over the set of all possible choices of change point. Thus, the posterior distribution we are interested in is:

$$[\tau | \mathbf{y}] = \frac{\sum_{p=2}^{20} [\mathbf{y} | M_p, \tau] [M_p | \tau] [\tau]}{\sum_{\tau=1}^T \sum_{p=2}^{20} [\mathbf{y} | M_p, \tau] [M_p | \tau] [\tau]} \quad (33)$$

In equation (32), M_2 to M_{20} are the set of possible models with M_p representing the AR(p) models. $[M_p | \tau]$ is $\frac{1}{19}$ because we assume each model is equally likely, *a priori*. The prior distribution over arrival times $[\tau]$ is $\frac{1}{T}$ because we assume before any data are collected that any time point in the time series is uniformly likely to be the change point. $[\mathbf{y} | M_p, \tau]$, the likelihood of the data under model M_p times the prior distributions and given change point τ , is defined as:

$$[\mathbf{y} | M_p, \tau] = \int [\mathbf{y} | M_p, \tau, \boldsymbol{\theta}_p][\boldsymbol{\theta}_p | M_p, \tau] d\boldsymbol{\theta}_p, \quad (34)$$

where $\boldsymbol{\theta}_p$ is a set of parameters $\{\phi_1, \phi_2 \dots, \phi_p, \sigma_\eta^2\}$ for model M_p . The integral of equation (33) is calculated by MCMC sampling, accounting for parameter uncertainty.

3 Simulation Study

3.1 Simulation Method

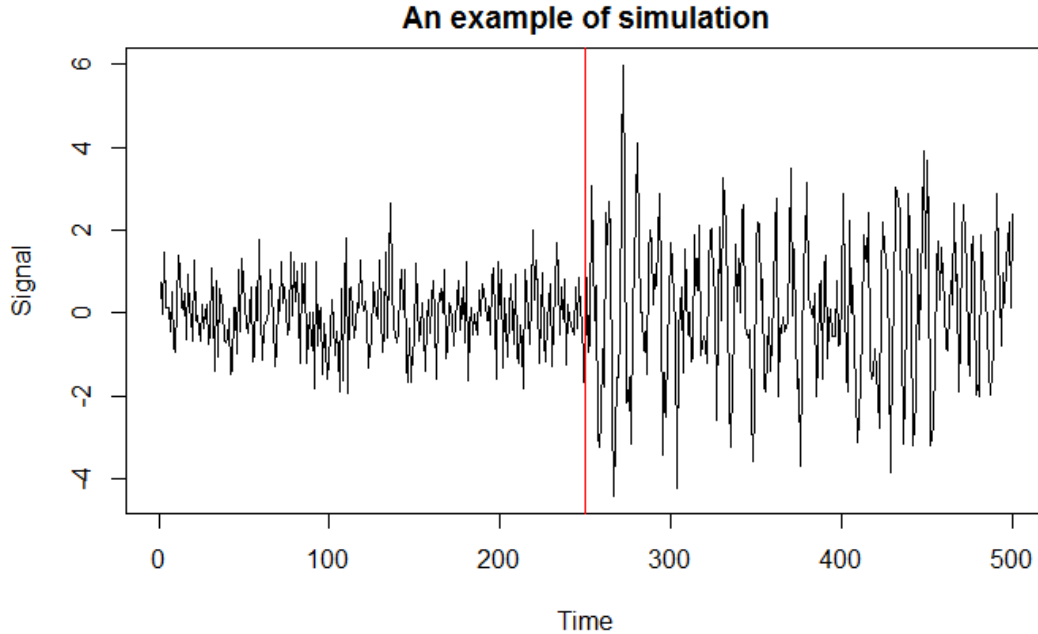


Fig. 2. Simulated data from the change point model (11) with $p = 4$. The redline at time 250 indicates the arrival time.

In order to examine how well the model framework performs, we conducted a simulation study.

We simulated 54 independent datasets from change point model (11) with $\sigma_\varepsilon^2 = 0.9$, $\sigma_\eta^2 = 1$, $p = 4$, $\{\phi_1, \phi_2, \phi_3, \phi_4\} = \{0.5, 0.3, -0.5, -0.2\}$, $\tau = 250$, and $T = 500$. An example of one of the simulated datasets is shown in Fig. 2. We noted our simulated process is similar to the real seismic waves on Fig. 1 because the pattern in the data is noticeably different after the change point which is the arrival time. Thus, our change point model could describe real seismic signals well.

We fit 19 different models given 500 different change points for each of 54 independent simulated datasets resulting in total 513,000 model fits. Because the study required fitting many models, we used the high performance “Razor computing cluster” at University of Arkansas that has 4,328 cores with a peak performance of 76 TF (Trillion Floating Point Operations per Second), and supports various statistical software including Matlab, Python, and R, among others (“Arkansas High Performance Computing Center (AHPCC)”, n.d.). The use of the computing resource allowed for efficiently fitting many models for each of the 54 simulated datasets.

We used Stan (Stan Development Team, 2014) to fit the 19 models for each of the 54 datasets by sampling 700 samples per chain, keeping 200 samples after 500 warmup iterations and fitting 4 chains giving 800 posterior samples per model fit. To generate posterior distributions for model averaging, we fit a grid of all values of p and τ for each simulated dataset. Next, we implemented BMA using the equation (32) and obtained the model averaged posterior for the simulated arrival time weights $[\tau \mid \mathbf{y}]$ for each simulated dataset.

3.2 Results

To check the estimates of arrival time τ , we plot the posterior density of the arrival time τ with the simulated true value of $\tau = 250$ as a vertical red line for 6 of the 54

Estimate of Arrival Time by 6 simulations

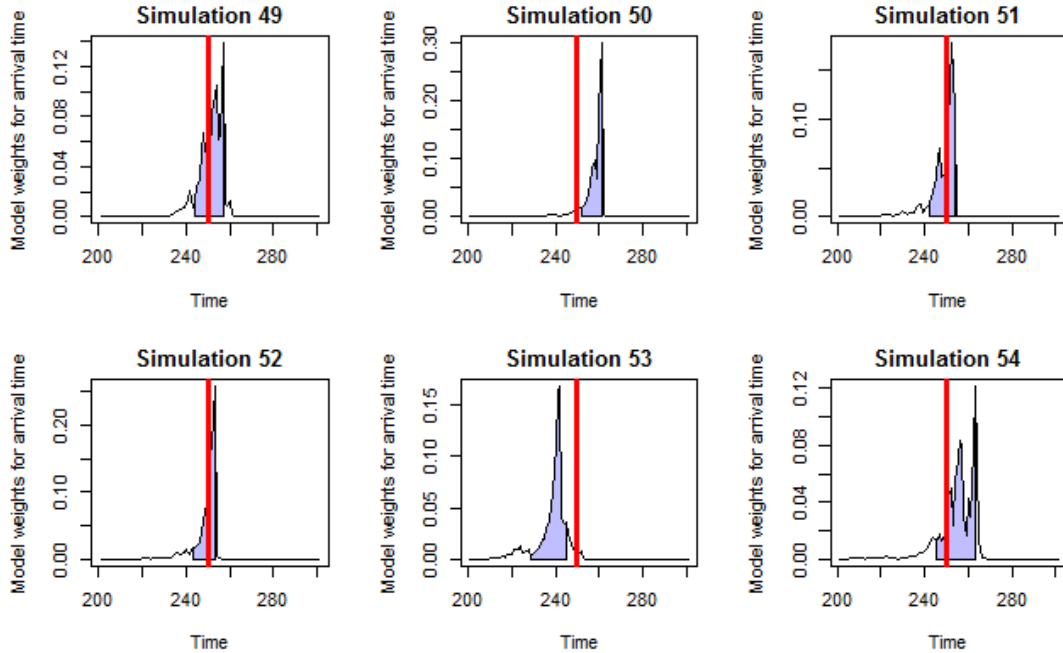


Fig. 3. Posterior densities for estimation of arrival time averaging over all possible model orders. (zoomed in to the time period t from 200 to 300). The red line at $t = 250$ indicates that the true arrival time we set for the simulated data. The blue shaded region shows the central 80% credible interval for arrival time. The heights of the curves represent the BMA posterior density.

simulated datasets (Fig. 3). The central 80% credible interval for estimating the arrival time τ is shown in the shading. Simulation 49, 51, 52, and 54 have change point predictions that contain the simulated arrival time $\tau = 250$ within the 80% central credible interval. However, some of the predictions are failed to cover the arrival time, such as simulation 50 and 53, whose credible intervals do not contain the simulated arrival time $\tau = 250$. The empirical coverage for τ based on a central 80% credible interval in these 6 simulations is 67%. However, the empirical coverage for τ in all 54 simulations is 83.33% which is close to its corresponding theoretical coverage (80%); thus, the central 80% Bayesian credible interval estimate of arrival time in all 54 simulations appears to be well calibrated.

Calibration Plot of Arrival Time

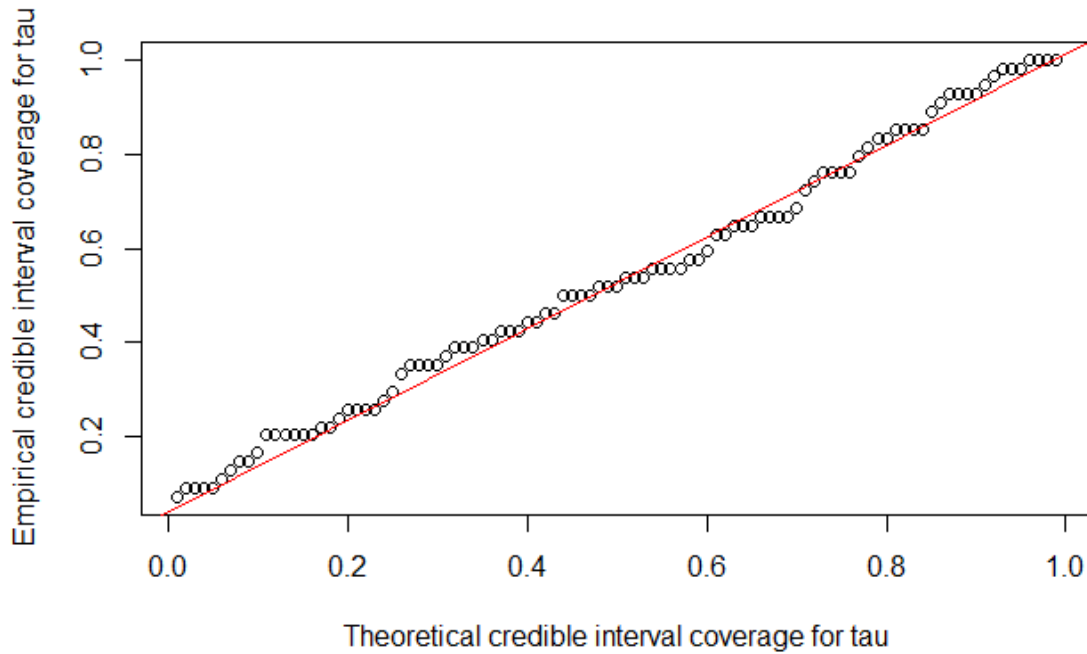


Fig. 4. Plot for theoretical credible interval coverage of arrival time and empirical credible interval coverage of arrival time. The empirical coverage for arrival time is the proportion of simulated credible intervals that contain the simulated true arrival time $\tau = 250$. The red line is the one-to-one line that represents a well calibrated prediction.

To check the general performance of the model in estimating the arrival time, we calculated the empirical coverage for τ , which is the proportion of simulated datasets where the estimated $(1-\alpha) * 100\%$ credible interval estimate τ contains the simulated arrival time $\tau = 250$. To check for any potential issues in model fit, we set α at values 0.01 to 0.99 in 0.01 increments to estimate empirical coverage. Then, we plotted the theoretical coverage $(1 - \alpha)$ and versus its corresponding empirical coverage for τ (Fig. 4). We note that the points in Fig. 4 approximately coincide with the one-to-one line, suggesting that the theoretical coverage and the corresponding empirical coverage for τ are approximately equal. Therefore, the BMA estimate of arrival time is well calibrated for estimation of τ .

Average Estimate of Arrival Time

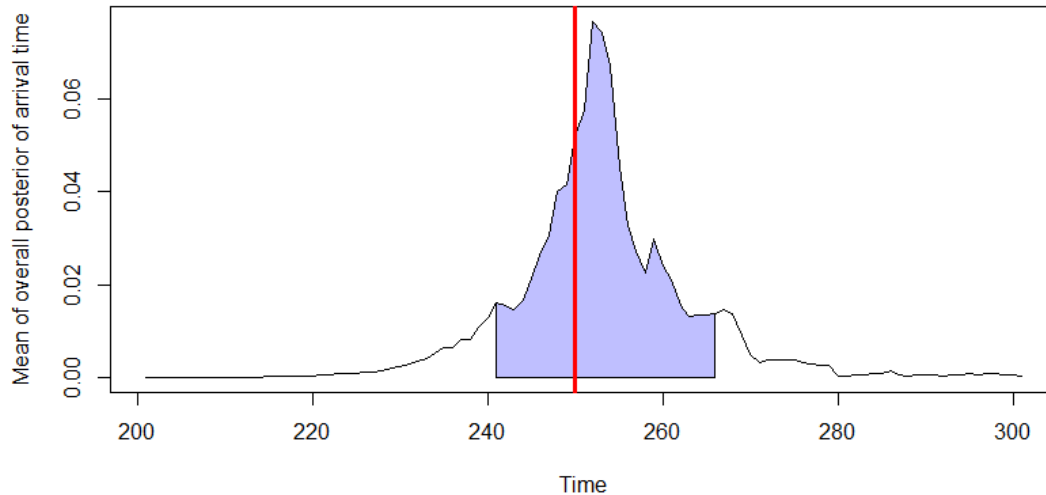


Fig. 5. Averaged posterior density for estimate of arrival time of 54 simulations (zoomed in to the time period t from 200 to 300). The blue shaded region shows the central 80% credible interval for the arrival time and the red line at $t = 250$ shows the true arrival time for the simulated data. The height of the curves represents the averaged BMA posterior density.

Another way to check the general performance of the arrival time estimate is to average all the posterior densities for the 54 simulations and plot the averaged posterior density of τ with the simulated true value of $\tau = 250$ as a vertical red line (Fig. 5). The shaded region shows central 80% credible interval for the arrival time τ . We see the average highest posterior density of arrival time over all 54 simulations is close to the simulated arrival time $\tau = 250$ and is contained within the 80% central credible interval. This provides evidence that, on average, the BMA approach to estimating arrival time accurately.

Estimate of Model's Order by 6 Simulations

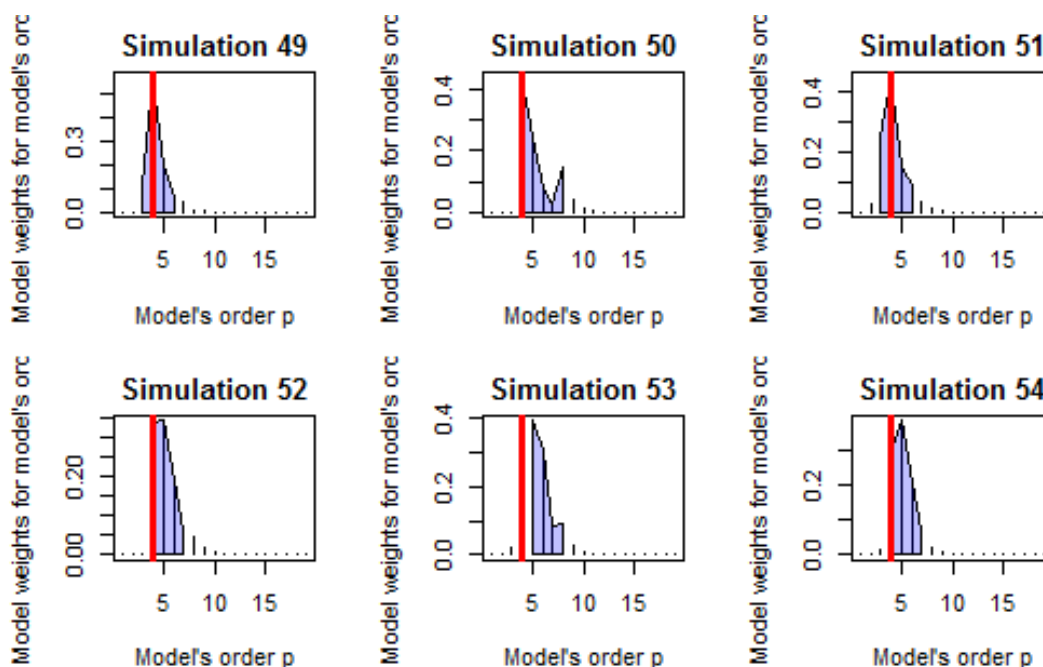


Fig. 6. Posterior densities for estimation of model order averaging over all possible arrival times. The blue shaded region shows the central 80% credible interval for the model order and the red line at $p = 4$ shows the true model order for the simulated data. The heights of the histograms represent the BMA posterior density.

We plotted the posterior density for the autoregressive model parameter order p and selected the same 6 simulated datasets in Fig 6. The red line shows the true simulated model order of $p = 4$ we shade the 80% central Bayesian credible interval. Fig. 6 shows that the predictions on simulations 49, 50, 51, 52 and 54 contain the simulated model order $p = 4$ within the 80% central credible interval. Especially on simulation 49, 50 and 51, the simulated model order $p = 4$ is at the highest BMA posterior density of estimation of model order p ; thus, simulation 49, 50 and 51 perform very well on estimating of model order p . Because only the estimate of model order p for simulation 53 does not contain the simulated model order $p = 4$, the empirical coverage for p is 83% for the six example plots, which is close to its

theoretical coverage (80%) in these 6 simulations. The overall empirical coverage for p in all 54 simulations is 75.93%, and it is also close to its corresponding theoretical coverage (80%).

However, if we increase the theoretical coverage to 90%, the general empirical coverage in all 54 simulations for p is 77.78%, and it is not close to its corresponding theoretical coverage.

Therefore, the general estimate of model order p in all 54 simulations may not always perform well depending on different theoretical coverage values.

Ideally, because all datasets were simulated from an AR(4) model, the estimates of the autoregressive model parameter order p should perform well. However, Gonzalez and Foy (1997), and Bedossa, Dargère and Paradis (2003) have stressed that due to the sampling error, sampling variability exists in their estimates. We found similar results where the sampling variability influences our sampling. If we repeat the same procedure many times, the average estimate of autoregressive model parameter order p will be close to the true model order we set. Thus, it is important to check the general performance of the model in estimating the autoregressive model parameter order p and the average estimate of autoregressive model parameter order p in all 54 simulations in next steps.

Calibration Plot of Model's Order

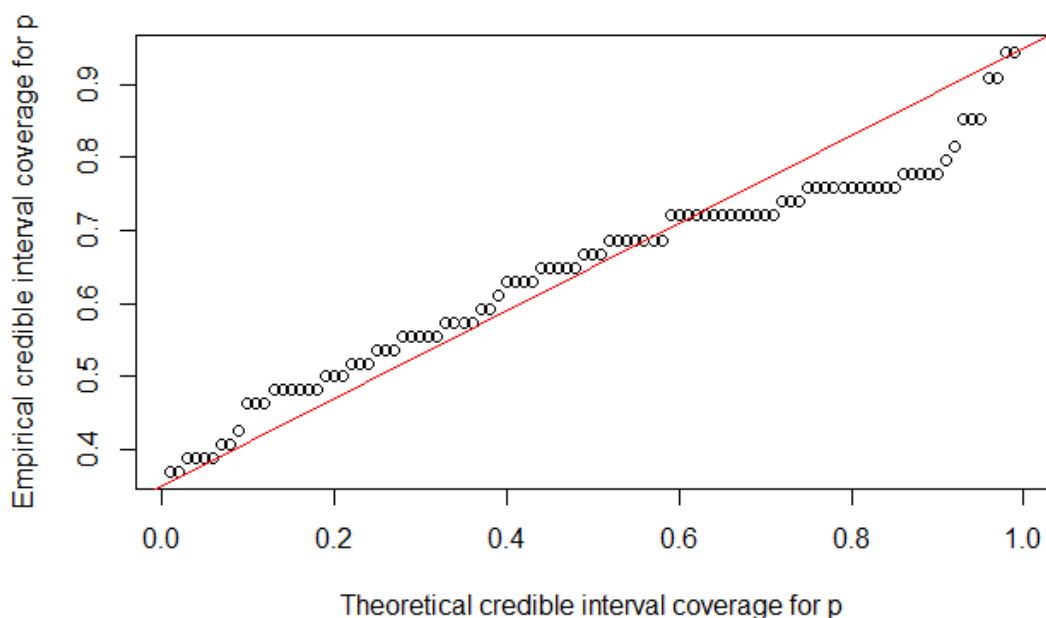


Fig. 7. Plot for theoretical credible interval coverage of arrival time and empirical credible interval coverage of BMA model order estimate. The empirical coverage for the model order p is the proportion of the simulated credible intervals that contain the simulated true model order $p = 4$. The red line is the one-to-one line that represents a well calibrated prediction.

We repeated the same procedure as Fig. 4 for the autoregressive order parameter p (Fig. 7) to check the general performance of the model in estimating the autoregressive model parameter order p . The plot is not as tight with the one-to-one line, especially as the theoretical coverage increases. For larger values of α , the empirical coverage is far away from the one-to-one line. Because of that, the theoretical coverage and its each corresponding empirical coverage for p are not approximately equal. Therefore, the BMA might not be calibrated for estimation of p ; however, this might simply due to the impact of the model space being a discrete parameter and the sample size being relatively small. A more computationally intensive simulation study could be used to further explore this relationship.

Average Estimate of Model's Order

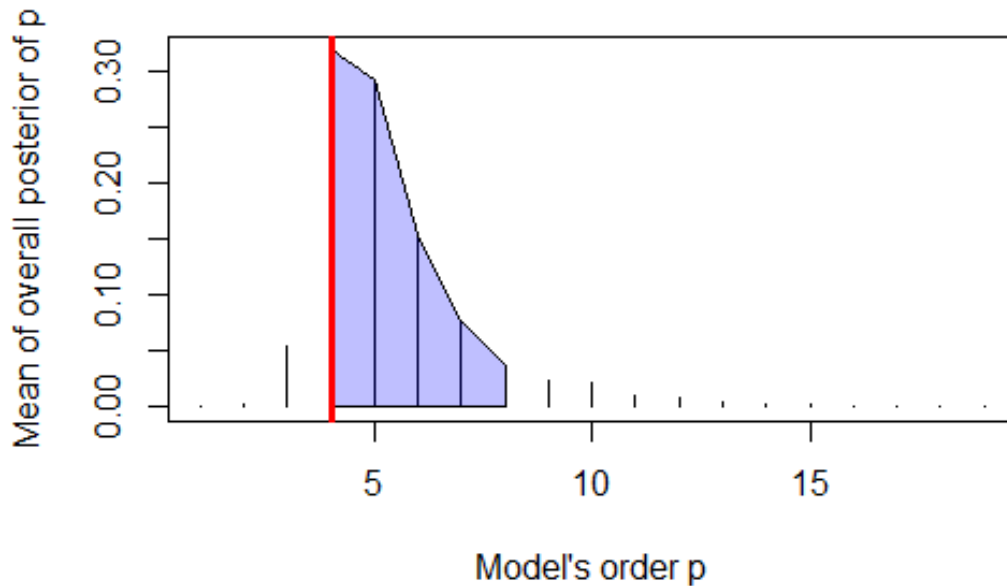


Fig. 8. Averaged posterior density for estimation of model order of 54 simulations. The blue shaded region shows the central 80% credible interval for the model order and the red line at $p = 4$ shows the true model order for the simulated data. The height of the histogram represents the averaged BMA posterior density.

Following the same procedure as Fig. 5, we plotted the averaged posterior density of autoregressive model parameter order p of 54 simulations (Fig. 8). The red line shows the true simulated model order of $p = 4$ we set and the shade regions show the central 80% credible interval for model order. We see the true model order of $p = 4$ we set is at left boundary of central 80% credible interval; however, the simulated value of $p = 4$ is at the highest average posterior density. Therefore, the estimate of autoregressive model parameter seems to be performing well even though the simulated parameter $p = 4$ is at the boundary of the central 80% credible interval. Thus, there is need for more investigation into the performance of the BMA framework in estimating the model order.

For readers' convenience, we used our results to build shiny app and deployed it to the cloud to share with readers (<https://hareluyaboy.shinyapps.io/thesisapp/>). A shiny app is a self-service platform that is convenient for users to visualize and share their projects on website (Shinyapps.io team, 2018).

4 Discussion

In this thesis, we demonstrated that the change point model framework (11) can estimate the seismic wave arrival time accurately while the use of BMA allows for calibrated estimation of uncertainty about arrival time. One concern about using BMA is whether the true model is in the model set; although we only used AR(2) to AR(20) models in our model set, the final BMA model represents a model that is outside the model set which can help reduce this issue. However, the 19 models we used may be not enough to account for model uncertainty in future analyses, and if this is the case, more models can be introduced into the model set.

In this simulation study, data were simulated from an AR(4) model only. Because the AR(4) model is in the model set, the BMA should be able to fit the data well. As such, the estimation of the quality of prediction of seismic arrival time on real data is probably over optimistic. One way to test this issue would be to explore different data generating models and perform the simulation study over these new classes of simulated data.

Moreover, although the result shows the BMA may not calibrate the estimation of autoregressive model parameter p , the average estimate of the model parameter p performs well. There is need to investigate this reason in future study.

In reality, researchers desire to pick the arrival time of a seismic wave as quickly as they can. In order to increase the calculation speed, parallel computing is needed. Parallel computing is a method that breaks a problem into many parts, and uses many computing tools to solve each separate part at the same time (Grama et al., 2003). For example, in this study, we used 19

models to fit the data; if we use 19 computing cores for each model to implement the model fitting at the same time, the calculation speed will theoretically increase nearly 19 times relative to using one computing core. Therefore, although we increase the candidate models in practice, it is possible to optimize the calculation speed for nearly real-time estimation if we properly use parallel computing resources.

To fully account for the uncertainty in future studies, there should be as many models to account for uncertainty as possible. We have already seen that our Bayesian framework performs very well on estimating the change point in our simulation study. In order to check our model's practicability, it is necessary to fit the data from real seismic wave and check its performance in future work.

References

- Arkansas High Performance Computing Center. (n.d.). Retrieved from <https://hpc.uark.edu/>.
- Bedossa, P., Dargère, D., & Paradis, V. (2003). Sampling Variability of Liver Fibrosis in Chronic Hepatitis C. *Hepatology*, 2003; 38(6):1449-1457.
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods and Research* 33:261–304.
- Banks, H.T., & Joyner, M.L. (2017). AIC Under the Framework of Least Squares Estimation. *Appl. Math. Lett.*, 74, 33-45.
- Burke, N. (2018). *Metropolis, Metropolis-Hastings and Gibbs Sampling Algorithms*. Lakehead University Thunder Bay, Ontario.
- Chatfield, C. (2003). *The Analysis of Time Series: An Introduction*. ISBN 9781584883173 - CAT# C3170.
- Cryer, J. D., & Chan, K. S. (2008). *Time Series Analysis with Applications in R*. Springer-Verlag New York. DOI: 10.1007/978-0-387-75959-3.
- Colombelli, S. (2014). *Early Warning for Large Earthquakes: Observations, Models and Real-Time Data Analysis*. Alma Mater Studiorum Università Di Bologna.
- Earle, S (2016). *Physical Geology*. Create Space Independent Publishing Platform 11, 325. ISBN-13: 978-1537068824.
- Gonzalez, E. J., & Foy, P. (1997). *Third International Mathematics and Science Study Technical Report, Volume II: Implementation and Analysis – Primary and Middle School Years*, 81-86. ISBN: 1-889938-06-8.
- Grama, A., Karypis, G., Kumar, V., & Gupta, A. (2003). *Introduction to Parallel Computing*. ISBN-13: 978-0201648652.
- Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57, 97--109.
- Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging: a tutorial. *Statistical Sciences*, 14(4), 382–417.

- Hayah, N. A., & Kim, D. (2013). Seismic Fragility Estimation of Base-Isolated NPP USING ARMA Synthesized Long Period Ground Motions. Transactions, SMiRT-22 San Francisco, California, USA - August 18-23, 2013 Division VII.
- Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* 15 (2014) 1593-1623.
- Link, W. A., & Barker, R. J. (2006). Model Weights and The Foundations of Multimodel Inference. *Ecology*, 87(10), 2626-2635. Retrieved from <http://www.jstor.org/stable/20069272>.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21, 1087-1092. doi: 10.1063/1.1699114.
- Monahan, J. F. (1984). A Note on Enforcing Stationarity in Autoregressive-moving Average Models. *Biometrika* 71 (2) (August 1): 403-404.
- Madigan, D., & Raftery, A. (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. doi:10.2307/2291017.
- Ozaki, T., & Tong, H. (1975). On The Fitting of Non-stationary Autoregressive Models in Time Series Analysis. Proceedings of the 8th Hawaii International Conference on System Science, Western Periodical, Hawaii.
- Ravenzwaaij, D. V., Cassey, P., & Brown, S.D. (2016). A Simple Introduction to Markov Chain Monte-Carlo Sampling. <https://doi.org/10.3758/s13423-016-1015-8>.
- Robert, C., & Casella, G. (2010). *Introducing Monte Carlo Methods with R*. ISBN:978-4419-1575-7.
- Shearer, P. (2003). Introduction to Seismology. *Journal of Seismology*, 7: 137. DOI: <https://doi.org/10.1023/A:1021219818569>.
- Stan Development Team. (2014). *Stan Modeling Language User's Guide and Reference Manual*. Retrieved from <https://mc-stan.org/>.
- Shinyapps.io Team. (2018). *shinyapps.io User Guide*. Retrieved from <https://docs.rstudio.com/shinyapps.io/index.html>.

Takanami, T., & Kitagawa, G. (1991). Estimation of The arrival Times of Seismic Waves by Multivariate Time Series Models. *Ann. Inst. Statist. Math.* Vol. 43, No. 3, 407-433.