Forecasting Surabaya Rainfall in 2017 Using First Order Autoregressive Model through INLA Bayesian Approach

Hery Tri Sutanto¹
¹Mathematics Department, Universitas Negeri Surabaya, C8 Building, Ketintang Surabaya, Indonesia
E-mail: herysutanto@unesa.ac.id

Abstract. Water is one of the natural factors which affects the life of organisms on earth. If in one area, there is a sufficient amount of rain, then it will support the life of organisms in that area. However, if an area has an excessive amounts of rain, flooding will occur and it will damage the infrastructure, the agricultural land and some existing ecosystems. One way to reduce unwanted floods is by predicting accurate flood discharge. It can be used not only to design water buildings but also to manage water flow. Flood discharge can be predicted using the time series method, namely the First Order Autoregressive model, also known as the AR model (1). Estimating parameters in the AR model (1) can be done using the Moment, least Square and Maximum Likelihood methods. However, because the sample size used in this study is small then the Bayesian method can be used to estimate the AR model parameters (1). It has some advantages such as: 1. It can produce logical interpretations to conclude statistically in time series analysis. 2. The output can always be updated based on the latest information. The Markov Chain Monte Carlo method is one method that can be used for Bayesian inference calculations, specifically the Gibbs sampler and Metropolis-Hastings algorithms. Both of these algorithms are more flexible and easily adapted to estimate complex models with many parameters or characterized by posterior nonstandard distributions. The MCMC theory states that the convergence distribution of the simulation value to the posterior distribution has a number of iterations to infinity, the posterior distribution will not be feasible. So that the processing of data consume a lot of time to reach convergence, then if it is not stopped then it will give suggestions about how some MCMC iterations become convergent. If the model is more complex then the MCMC algorithm is slower and the calculation becomes not feasible. Therefore, Rue et al. (2009) proposed a fast and accurate algorithm called the Integrated Nested Laplace Approximation Algorithm (INLA).

1. Introduction
Water is one of the factors that influence the survival of living things on the earth. If an area has a sufficient amount of rain, then it can supports all life in the area. Conversely, if an area has an excessive amounts of rain that will exceed the carrying capacity of the environment, then the presence of water will harm agriculture, plantations, housing and other living things. It will cause damage to infrastructure that has been built by the government, damage to agricultural land, and damage to some existing ecosystems [1].

Regard to the flood, efforts are needed to overcome it. Rainfall data recorded by BMKG (Meteorology and Geophysics Agency) can be used to predict the amount of rainfall that will occur in the future. This allows the government to be able to plan, improve development and anticipate future
floods. In order to reduce an unwanted flooding, we can just predict the incidence of it. Furthermore, it can be used to design water buildings, rivers, gutters, irrigation, water seepage, dams and water management [2].

Geographically, the city of Surabaya has an area of 326.81 km$^2$ with an average rainfall of 1,321 mm per year. From May to October in Surabaya, there is a dry season. In November to April the rainy season occurs. If the rainfall is too many and the carrying capacity of the environment (the place of seepage, rivers, ditches, rice fields, lakes) is inadequate then there will be a big flood in Surabaya [3].

There are several studies on rainfall in Indonesia using univariate time series. One of them is an autoregressive (AR1) model which is assumed to be a stationary rain condition while the rain is not stationary. Then the AR (1) model has developed using the ARMA (Autoregressive Moving -Average) model which is assumed to be a stationary rain process and lognormal distribution. To estimate parameters in the AR (1) model: 1). Moment Method, 2). Least Square Method, and 3). Maximum Likelihood Method [4,5].

It is because the sample size used in this study is small then to estimate the parameters of the AR model (1), we use Bayesian method. The reason for involving Bayesian analysis in time series analysis are, 1). In this approach, it can produce a logical interpretation to conclude statistically in time series analysis, 2). The output can always be updated based on the latest information [5].

The Monte Carlo Markov Chain method is a supporter of calculations for Bayesian inference. In particular the Gibbs sampler and Metropolis-Hastings algorithm. They are more flexible and easily adapted to estimate complex models that use many parameters or characteristics by a nonstandard posterior distribution. The MCMC theory states that the distribution of the convergence of the simulation value to the target density (posterior distribution) if the number of iterations is infinite then the posterior distribution is not feasible to run an infinite Markov chain. Hence, in order to process the data, it takes too long to reach convergence, if it is not stopped, it will give a suggestion how some MCMC iterations need convergence [6,5].

The main key in drawing conclusions with MCMC is greater caution (a lot of time) to decide and monitor the conference stage in determining the best setting (parameterization, prior distribution, initial value and distribution of Martin-Hasting proposals) which results in a reliable and accurate MCMC output (Brooks et al., 2011).

Several methods for producing existing posterior distributions have their advantages and disadvantages. Rue et al. (2009) have developed a method for estimating parameters that is better and more efficient than the methods described earlier. This method is called Integrated Nested Laplace Approximation (INLA) [7].

2. Study Literature

2.1. Time Series

The main purpose of time series analysis is to produce predictions of future values based on current data. Time series analysis is a method used to study theoretically and application. According to Cryer and Chan (2008), time series data is a series of observational data arranged according to time sequence. Time series forecasting is a model to predict future time values based on data (events) that have occurred [4,6].

2.2. Stationary Data

If a time series data is plotted then there is no evidence of change in the mean value from time to time then it can be said that the data series is stationary at its mean. If a time series data is plotted then it does not show any change in variance from time to time then it is said that the data series is stationary in its variants. If a time series data is plotted then it shows the change in mean (some trend-cycle) over time then it is said the data series is not stationary at its mean. If a time series data is plotted, then it does not show any change in mean and variance from time to time then it is said that the data series is not stationary at the mean and its variants [4,8].
2.3. Stationary Average Data

According to (Hanke J E, Ritse A., and Witchen, 1999) the autocorrelation values on stationary data to the average will not be significant after the second or third lag time, while the non-stationary data has a significant autocorrelation value after the third lag time or several periods. Stationary to the average can be known graphically by looking at the ACF plot, the data is said to be stationary to the average if at most the first 3 lags on the ACF plot do not exceed the $\pm \frac{2}{\sqrt{n}}$. If the data is not stationary to the average, it is necessary to do a differentiation process until the data is stationary to the average [4,8]. The first form of differentiation can be stated as follows.

$$\nabla Y_t = Y_t - Y_{t-1}$$

Statistic used

$$t_{hit} = \frac{\hat{\phi}^*}{Se(\hat{\phi}^*)}$$

where,

$\hat{\phi}^*$: Estimated value of the Autoregressive (AR) parameter

$Se(\hat{\phi}^*)$: Error standar of $\hat{\phi}^*$

2.4. Forecasting First Order Autoregressive Model or AR Model (1)

The simplest model of AR (1) with zero on its mean [9]

$$Y_t = C + \rho Y_{t-1} + e_t$$

where $-1 < \rho < 1$

$$t = 1, 2, ..., n$$

It is a stationary AR (1) process (Box et al, 1970)

$Y_t$ is $t^{th}$ observation time series

$\{ e_t \}$ is a process

C is an arbitrary constant

$\rho$ is a parameter autoregressive

2.5. Laplace Approximation

The INLA approach was first proposed by Rue et al. It is to determine a Bayesian inference algorithm based on integrated nested Laplace approximations (INLA). INLA was designed specifically for Gaussian latent models that are quite flexible which are used for different applications. See Rue et al. [10] who review some of the examples used in the R-INLA package program.

Suppose $f(x)$ is probability density function for random variable $x$, then

$$\int f(x) \, dx = \int \exp(\log f(x)) \, dx$$

The value of $\log f(x)$ can be obtained by using Taylor series and take $x = x_0$ as follows

$$\log f(x_0) \approx \log f(x_0) + (x - x_0) \frac{\partial \log f(x)}{\partial x} \bigg|_{x=x_0} + \frac{(x-x_0)^2}{2} \frac{\partial^2 \log f(x)}{\partial x^2} \bigg|_{x=x_0} = 0$$

If $x_0 = x^* = \text{argmax} \log f(x)$ then

$$\frac{\partial \log f(x)}{\partial x} \bigg|_{x=x_0} = 0$$

Equation (2) will be

$$\log f(x_0) \approx \log f(x^*) + \frac{(x-x^*)^2}{2} \frac{\partial^2 \log f(x)}{\partial x^2} \bigg|_{x=x_0} = x^*$$

Substitute (3) into (1) then

$$\int f(x) \, dx \approx \int \exp \left( \log f(x^*) + \frac{(x-x^*)^2}{2} \frac{\partial^2 \log f(x)}{\partial x^2} \bigg|_{x=x_0} \right) \, dx$$

$$= \exp(\log f(x^*)) \int \exp \left( \frac{(x-x^*)^2}{2} \frac{\partial^2 \log f(x)}{\partial x^2} \bigg|_{x=x_0} \right) \, dx$$

Where the integrand relates to the density of normal distribution.

Choose

$$\sigma^{2*} = -\frac{1}{\frac{\partial^2 \log f(x)}{\partial x^2} \bigg|_{x=x^*}}$$

Such that
\[
\int f(x)dx \approx \exp(\log f(x^*)) \int \exp\left(\frac{-(x-x^*)^2}{2\sigma^2}\right) \, dx
\]  
(5)
where \(\exp\left(\frac{(x-x^*)^2}{2\sigma^2}\log f(x)\right)\) is a kernel of normal distribution \((x^*, \sigma^2)\).

If the integral above has boundary at domain \((\alpha, \beta)\) then (5) becomes

\[
f_\alpha^\beta f(x) \, dx \approx f(x^*) \sqrt{2\pi \sigma^2} \left(\phi(\beta) - \phi(\alpha)\right)
\]
(6)
where \(\phi(.)\) is cumulative density function of normal distribution.

2.6. Latent Gaussian Model

INLA is designed with the Gaussian latent model, where the observed (response) variable \(y_i\) is assumed to belong to a distribution family (not necessarily from an exponential family) [7]. Some parameters of the \(\Phi_i\) family are links from the additive predictor structure of \(\eta_i\) through a link function \(g(.)\), such that \(g(\Phi_i) = \eta_i\).

Additive linear predictor \(\eta_i\) is defined as follows

\[
\eta_i = \beta_0 + \sum_{m=1}^{m=M} \beta_m x_{mi} + \sum_{l=1}^{L} f_l(Z_{il})
\]

where:
- \(\beta_0\) is a scalar for intercept
- \(\beta = \{\beta_1, \beta_2, ..., \beta_M\}\) is a coefficient of linear effect quantity from several covariates \(x = (x_1, x_2, ..., x_M)\) in response
- \(f = \{f_1(\cdot), ..., f_L(\cdot)\}\) is a collection of function which is defined from a covariate \(Z = (z_1, ..., z_L)\). In \(f_l(\cdot)\), it can be assumed some different form such that smoothness, non-linear effect covariate, time series, seasonal effect, intercept and random slope (temporal random effect or spatial random effect).

A group of latent gaussian model is flexible and able to accommodate a wide range of models ranging from generalize and dynamic linear models to spatial and spatio-temporal (Martins et al., 2013).

A group of spatio-temporal model is usually constructed by hierarchical bayesian with three step. First step is construct an observation mode \(\pi(y|x)\) where \(\pi(\cdot | \cdot)\) is density and \(y\) is an observation vector. If it is assumed \(y_i\) independent, latent component (cannot be observed) \(x\) and hyperparameter vector \(\theta\) then the distribution of \(N\) observation is likelihood:

\[
\pi(y|x, \theta) = \prod_{i=1}^{i=N} \pi(y_i|x_i, \theta)
\]

Second step is latent gaussian field \(\pi(x|\theta)\) where a prior of multivariate gaussian with mean is equal to zero and precision matrix \(q\) for \(x\). It depends on hyperparameter \(\theta\) (third step) which is no need to be Gaussian. Hence, \(x \sim N(0, q^{-1}(\theta))\) with density function as follows:

\[
\pi(x|\theta) = (2\pi)^{-N/2} |q(\theta)|^{1/2} \exp\left(-\frac{1}{2} x^T q(\theta) x\right)
\]

Suppose, a components of latent gaussian field \(x\) is conditionally independent such that \(q(\theta)\) is sparse precision matrix [7].

3. Discussion

3.1. AR Model (1) Identification

The rainfall data is identify as follows. Find the scatter plot of the data from day one to last day in the same year.
The results of the plot turned out to be not stationary data. Therefore data needs to be transformed with first differencing. Unfortunately, because on the first differencing, the data remained not stationary on the 210th day up to the 366th day, then it requires to differencing in the second order to obtain stationary data.

3.2. AR Model (1) and test parameters

Regard to the identification above, there are some possibilities model such as:

i. Model AR(1) atau model ARMA(1,0) atau Model ARIMA(1,0,0)

ii. Model Arma (0,1) atau Arima(0,0,1)

iii. Model Arma(1,1) atau Arima(1,0,1)

iv. Model Arima(1,2,0)

v. Model Arima(0,2,1)

By using R software and forecast library, we get

```r
> library(forecast)
> auto.arima(sbydat)
```

Series: sbydat
ARIMA(1,0,1) with non-zero mean

Coefficients:
```
            ar1      ma1    mean
           0.8633  -0.7966   6.5592
```

s.e. 0.1347 0.1591 1.2537

sigma^2 estimated as 260.5:  log likelihood=-1535.84
AIC=3079.67   AICC=3079.78   BIC=3095.28

So, Arma (1,1) or Arima (1,0,1) had been choosen then it can be used to predict the rainfall for next two months or next 60 days based on the data with program as follow.

```r
> sby_arma_c<-forecast(arma_c,h=60,level=c(99.5))
> sby_arma_c
```

Therefore, the result can be plotted like follows
3.3. **AR Model (1) and Bayesian INLA**

library(INLA)

```r
temp = read.csv("D:/BAYESIANINLA2019/Datainla/Tugas AR1 INLA/Olahdatahujansby2016.csv", header=F)
n = nrow(temp)
data = data.frame(y = temp[,1], t=1:n)
dates <- data[,2]
plot(dates, data$y, lwd=2, xlab='harian', ylab='Data Curah Hujan')
lines(dates,data$y)

It gave this result
```

By observing the likelihood and AR(1) model, we can get this result

```r
family <- "gaussian"
formula1 <- y ~ f(t,model='ar1')

Calculate the posterior distribution \( \pi(\theta|y) \)
where \( \theta = (\theta_1, \theta_2) \)

- \( \theta_1 \) is hyperparameter with \( \theta_1 = \log(K) \)
- \( \theta_2 \) is hyperparameter with \( \theta_2 = \log(1 + \rho_1 - \rho) \)

By using transformation, it will give \( \pi(K|y) \) and \( \pi(\rho|y) \)
where \( K = \tau(1 - \rho_2) \)

By using this program

```r
res1 <- inla(formula=formula1, data=data, family=family)
summary(res1)
```
Call:
  "inla(formula = formula1, family = family, data = data)"

Time used:
  Pre = 2.83, Running = 1.96, Post = 0.183, Total = 4.97

Fixed effects:
  mean  sd  0.025quant  0.5quant  0.975quant  mode  kld
(intercept)  6.393 0.847  4.729  6.393  8.056  6.393  0

Random effects:
  Name  Model
  t  AR1 model

Model hyperparameters:
  mean  sd  0.025quant
Precision for the Gaussian observations 4.00e-03 0.00e+00  0.003
Precision for t  1.91e+04  1.86e+04  1270.650
Rho for t  4.00e-03  9.88e-01  -0.997

Expected number of effective parameters(stdev): 1.00(0.00)
Number of equivalent replicates: 365.94
Marginal log-Likelihood: -1555.17

4. Conclusion
Before processing, it takes 2.83 and then it runs (running) = 1.96. Furthermore, it takes 0.183 (Post), hence it takes 4.97 in total. In summary, the output above shows the distribution of $\beta_0$ (intercept), distribution $\tau$ (precision for Gaussian observation), $K$ distribution (precision for t) and distribution $\rho$ (Rho for t) which have 95% of the shortest Highest Posterior Density Region (HPDR) is $\tau$ (precision for Gaussian observations). From the results of the fitting, we can also see two types of points that almost overlap everywhere, which means that the results of the model and real data are almost the same.

References
[1] Tim D 2007 Fundamentals of Hydrology, Routledge Fundamentals of Physical geography (New Zealand)
[2] Hidayah E, Anwar N E, and Iriawan N 2010 Evaluating Error of Temporal Disaggregation from Daily into hourly Rainfall using Heytos Model at sampean catchments Area IPTEK The Journal of Technology and Science 21(1)
[3] Kristianta F and Fithrisani K 2016 Predict rainfall in East Surabaya with Vector Autoregressive Neural Network Journal Sains and Art ITS 5(2)
[4] Agung N G I 2009 Time Series Data Analysis Using EViews (Asia: John Wiley & Sons Pte Ltd.)
[5] Maoumi H, Niekerk J, and Staden P 2010 Bayesian Analysis of AR(1) model (South Africa: University of Pretoria)
[6] Detting M 2016 Applied Time Series Analysis, Institute for Data Analysis and Process Design (Winterthur: Zurich University of Applied Sciences)
[7] Blangiardo M and Cameletti M 2015 Spatial and Spatio-temporal Bayesian Models with R_INLA (United Kingdom: John Wiley & Sons, Ltd)
[8] Rosadi D 2011 Ekonometrika & Analisis Runtun Waktu Terapan dengan Eviews Aplikasi untuk bidang ekonomi, bisnis dan keuangan (Yogyakarta)
[9] Zheng R and Bakka H 2018 Analysing unemployment data with the AR1 Model
[10] Rue H and Martino 2009 Approximation Bayesian Inference for Latent Gaussian Models Using INLA
[11] Chib S 1991 Bayes regression with autoregressive errors A Gibbs sampling approach (USA: Washington University)
[12] Gomes-Rubio V and Rue H 2017 Markov Chain Monte Carlo with the Integrated Nested Laplace Approximation (Spain)
[13] Otieno E 2013 *Bayesian Spatial and Spatiotemporal Modeling (Applied to precipitation data set).*