Coro-coro Dinggo Ngeramal Data: Catetan Cekak lan Pepak –Tapi Durung Rampung–

Canggih Puspo Wibowo
`canggihpw@gmail.com`

Terakhir penulis ngedit: May 29, 2020 15:07:00

Abstract

[illegible]

Isi Catetan

1. Awalan

Tulisan cekak iki digawe tujuane ben iso dinggo ngeling-eling coro opo wae neng babagan *forecasting*. Dadi isine yo mung catetan wae sing dikumpulake penulis dinggo sinau lan ngajar neng kelas kuliah. Mergo mung catetan, dadi yo mungkin wae akeh tulisan sing kleru utawa ora jelas. Jaluk ngapura lah.

Kanggo sing durung ngerti, *forecasting* kui gampange ilmu dinggo nebak masa depan, yo koyo ramalan ngono. Ananging iki dudu ramalan koyo mbah dukun lo; iki ramalan sing dasare ilmu-ilmu matematika lan statistika. Nah, coro-coro sing ono neng catetan iki kui yo sakjane coro ngeramal kui mau, nanging ditulis luwih *detil* babagan rumus matematika lan statistikane. Nek misal sampeyan isih kangelan moco rumus-rumus, yo sing penting iso paham konsepe koyo piye. Penulis bakal usaha ben konsepe yo tetep jelas. Tapi nek ora jelas yo sampeyan sinau matematika dasar dhisik yo ora popo. Sarane penulis, sampeyan iso sinau babagan *time-series* sadurunge moco iki, tapi yo ora wajib.

Masalah sing meh dirampungke nganggo coro *forecasting* kui simpel. Misale awake dewe duwe data *time-series* sebut wae y_t sing isine y_1 tekan y_T. Total datane ono T. Seko data kui, awake dewe nggoleki $\hat{y}_{T+h}$, ramalan pendekatan seko y_{T+h}. Neng kene, h kui dijenengke *forecast horizon* yokui rentang wektu neng masa depan sing meh diramal. Nek ditulis nganggo bahasa matematika, tujuane *forecasting* koyo ngene

$$\text{tujuane } \textit{forecasting} : \hat{y}_{T+h|T} \tag{1}$$

Dadi, awake dewe goleki ramalan neng $T + h$ dasare nganggo data neng T. Uwis kui tok sakjane. Neng catetan iki, penulis sengojo mbagi coro-coro neng 3 bagian: coro simpel, coro ES (*exponential smoothing*), lan coro ARIMA.

2. Coro Simpel

Yo iso dikiro-kiro seko jenenge, nek bagian iki sing dibahas ki coro sing sederhana banget. Nek dinggo ngramal tenanan sakjane yo ora bakal akurat. Ananging coro-coro iki iso dadi dasar dinggo sinau coro sing luwih ampuh.

2.1 Naive Method

Coro iki mungkin coro sing paling simpel. Intine yokui nilai sing pengen awake dewe ramal kui dianggep podo karo nilai terakhir. Rumuse koyo ngene

$$\hat{y}_{T+h|T} = y_T \tag{2}$$

Nek sampeyan wis tahu sinau babagan ngisi nilai kosong (*missing value imputation*) neng data *time-series*, mesti wis tahu krungu coro *LOCF*[1] karo *NOCB*[2]. Nah coro Naive iki intine podo karo kui kan yo. Ananging wis jelas nek cara iki ora cocok nek dinggo ngeramal akeh utawa adoh neng ngarep (nilai h gede).

2.2 Simple Average

Selanjute ono coro *Simple Average*. Iki yo coro sing gampang. Intine yokui ngetung rerata seko data sing wis ono, trus hasile dinggo ramalan. Rumuse koyo ngene

$$\hat{y}_{T+h|T} = \frac{1}{T}\sum_{i=1}^{T} y_i \tag{3}$$

Nek dipikir-pikir, coro iki isone mung nek dinggo ngeramal data *time-series* sing variasi datane ora gede, alias kabeh data cedhak karo rerata. Nek datane jungkir walik, munggah mudhun, pethakilan, yo rodo ora cocok nganggo iki.

2.3 Drift Method

Coro selanjute yoiku *Drift Method*. Nek coro Naive ki mung nganggo data terakhir y_T, nek coro iki luwih cerdas sithik, yokui nganggo data pertama y_1

[1]Last Observation Carried Forward
[2]Next Observation Carried Backward

lan terakhir y_T. Etungane koyo ngene

$$\begin{aligned} \hat{y}_{T+h|T} &= y_T + \frac{h}{T-1}\sum_{t=2}^{T}(y_t - y_{t-1}) & (4) \\ &= y_T + h\left(\frac{y_T - y_1}{T-1}\right) & (5) \end{aligned}$$

Sampeyan iso delok neng rumus nek coro ngeramale nganggo ilmu *ekstrapolasi linier*. Opo kui? Maksude antara y_1 lan y_T digawe siji garis lurus, trus data ramalane yo dijupuk seko titik neng garis kui nganggo nilai h sing tertentu. Wis jelas, coro iki ora mentingke data sing neng tengah, mung asal comot sing pertama lan terakhir.

2.4 Moving Average Smoothing

Nah, selanjute, coro iki terkenal banget neng donya *time-series*. Terkenal amargo coro sing paling simpel dinggo gawe data *time-series* dadi alus (*smoothing*). Biasane awake dewe nganggo iki dinggo keperluan kui. Nek data wis alus, ora akeh variasi, iso didelok gambaran umum seko datane, misale ono *tren* opo ora, lan sapanunggalane. Neng kene, coro iki yo iso dinggo ngeramal. Corone yoiku jupuk k data sedurunge terus dietung reratane, hasile dinggo prakiraan ramalan data. Rumuse yoiku

$$\hat{y}_t = \frac{1}{k}(y_{t-1} + y_{t-2} + \cdots + y_{t-k}) \tag{6}$$

Coro iki rodo mirip karo *Simple Average*. Bedane nek mau nganggo kabeh rerata, nek iki mung nganggo data cacahe k wae.

2.5 Weighted Moving Average

Coro iki sakjane bentuk umum seko *Moving Average* biasa. Bedane mung neng kene ki tiap data dikei bobot w. Nah misale kabeh bobot w kui nilaine 1, yo coro iki podo persis karo sadurunge. Rumus etungane koyo ngene

$$\hat{y}_t = \frac{1}{k}(w_1 y_{t-1} + w_2 y_{t-2} + \cdots + w_k y_{t-k}) \tag{7}$$

Nilai bobot sing diwenehi kui sakjane bebas wae isine. Tapi, nek manut konsepe *time-series* kudune bobot seko data sing paling cedhak yo diwenehi nilai sing luwih gede dibanding sing adoh, soale porsine luwih gede. Dadi, mengko $w_1 \geq w_2 \geq \cdots \geq w_k$.

3. Exponential Smoothing Methods

Saiki awake dewe mlebu babagan etungan sing mulai rodo serius, jenenge *exponential smoothing*. Coro ngeramal nganggo metode iki lumayan terkenal. Neng kene bakal dibahas pie carane ngeramal data *time-series* sing ono bagian *tren* lan *musiman*.

3.1 Time-Series Decomposition

Sadurunge bahas babagan ramalan, ben iso paham coro-coro neng kene, awake dewe kudu bahas *dekomposisi* data *time-series* dhisik. Intine *dekomposisi* yoiku data *time-series* kui sakjane iso dipecah dadi beberapa bagian. Sakjane ono akeh coro-corone, misale sing terkenal banget kui *Fourier series*. Ananging dinggo bahas peramalan neng kene, awake dewe bakal memecah data dadi tren, musiman, lan level. Dadi misal iso digambarke

$$\text{time-series} = \begin{cases} \text{level} \\ \text{tren} \\ \text{musiman} \end{cases} \tag{8}$$

Nah, coro-coro neng bagian *Exponential Smoothing* iki bakal memanfaatke *dekomposisi* kui.

3.2 Simple Exponential Smoothing

Ono sing nyebut jeneng liyane *Single Exponential Smoothing* (SES). Kenopo kok *single*? Soale neng SES, komponen dekomposisi (Rumus 8) sing digunakke mung 1, yoiku *level*. Dadi, rumus ngeramale dadi

Rumus *Forecast*

$$\hat{y}_{t+h|t} = l_t \tag{9}$$

nilai l_t neng kene adalah komponen *level* sing rumuse

Rumus *Level*

$$l_t = \alpha y_t + (1 - \alpha) l_{t-1} \tag{10}$$

Ono α neng dhuwur kui jenenge *smoothing parameter* sing nilaine $0 \le \alpha \le 1$. Nah nek meh nganggo rumus SES dinggo ngeramal, awake dewe kudu menetapke dhisik nilai α sing bakal dinggo. Nilai iki bakal duweni pengaruh dinggo aluse *time-series*.

Durung di-*proofread* opo meneh di-*peer-review*

3.3 Holt's Linear Trend

Jeneng liyane yoiku *Double Exponential Smoothing*. Wis sesuai prakiraan awake dewe, ono single, ono double, berarti sakjane iki ono tambahan dibandingke sing single. *Holt's Linear Trend* sakjane yoiku SES sing ditambahi komponen *tren*. Tujuane ditambahi komponen *tren* yoiku ben iso digunakke dinggo data *time-series* sing ono tren-e. Rumus ngeramale dadi

Rumus *Forecast*

$$\hat{y}_{t+h|t} = l_t + hb_t \tag{11}$$

Mergo ono pengaruh seko *tren*, makane etungan *level/smoothing* neng Rumus 10 yo berubah dadi

Rumus *Level*

$$l_t = \alpha y_t + (1 - \alpha)(l_{t-1} + b_{t-1}) \tag{12}$$

Nek didelok meneh, Rumus 12 dasare podo karo Rumus 10 mung ditambahi b_{t-1} sing etungane entuk seko

Rumus *Trend*

$$b_t = \beta^*(l_t - l_{t-1}) + (1 - \beta^*)b_{t-1} \tag{13}$$

Neng kene ono β^* sing dinggo dadi *trend parameter*. Konsep β^* iki mirip karo gunane α neng level.

3.4 Holt-Winters

Nah selanjute iki jeneng liyane *triple exponential smoothing*. Bedane karo sing sedurunge yoiku, Holt-Winters iki nambahi komponen musiman (*seasonal*) s_t. Babagan coro iki, ono 2 pendekatan sing iso dinggo, yoiku *additive* lan *multiplicative*.

3.4.1 Additive Model

Coro *additive* iki luwih cocok digunakke nek variasi neng musimane rodo konstan. Rumus ngeramal nganggo coro iki yoiku

Rumus *Forecast*

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)} \tag{14}$$

Nek didelok seko Rumus 14, awake dewe iso ngerti yen ono tambahan komponen s sing dietung seko informasi musiman. Nilai m neng komponen musim kui maksude cacahe musiman. Misale nek datane wulanan, brarti nilai $m = 12$. Trus nilai k kui dijupuk seko etungan $(h-1)/m$. Tambahan komponen iki marakke rumus level yo berubah dadi

Rumus *Level*

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{15}$$

Nah wis ketok seko rumus neng dhuwur nek ono bagian pengurangan nilai y_t nganggo komponen musiman s. Iki dinggo ngilangi pengaruh musiman neng persamaan level. Selanjute neng bagian tren, tetep wae ora berubah seko double smoothing, soale tren ki efeke adoh, ora koyo musiman.

Rumus *Trend*

$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1} \tag{16}$$

Nah, persamaan musiman iki sing bedo karo itungan sadurunge. Ono perhitungan komponen musiman koyo ing ngisor iki:

Rumus *Seasonal*

$$s_t = \gamma^*(y_t - l_{t-1} - b_{t-1}) + (1-\gamma^*)s_{t-m} \tag{17}$$

Neng persamaan 17 kui, γ^* yoiku *smoothing parameter* neng musiman.

3.4.2 Multiplicative Model

Sadurunge wis dijabarke itungan Holt-Winters nganggo *additive model*, nah saiki coro liyo yoiku nganggo *multiplicative model*. Coro iki luwih cocok digunakake yen variasi musiman kui berubah-ubah.

Rumus *Forecast*

Ono bedo sithik antara itungan rumus *forecast* iki dibanding bagian *additive* sadurunge. Neng kene komponen musiman dadi pengali neng komponen tren lan level, makane rumuse dadi:

$$\hat{y}_{t+h|t} = (l_t + hb_t)s_{t+h-m(k+1)} \tag{18}$$

Rumus *Level*

Neng rumus level juga komponen musiman ora dinggo pengurang, tapi saiki berubah dadi pembagi.

$$l_t = \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{19}$$

Rumus *Trend*

$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1} \tag{20}$$

Rumus *Seasonal*

$$s_t = \gamma^* \frac{y_t}{(l_{t-1} + b_{t-1})} + (1-\gamma^*)s_{t-m} \tag{21}$$

Intine rumus-rumuse berubah seko penjumlahan lan pengurangan dadi perkalian lan pembagian, kanggo komponen musiman. Itungan liyane ora berubah.

Bagian iki mengisor durung tak ganti seko enggres neng jowo. Durung sempat wae

4. ARIMA

While exponential smoothing models are based on a description of the trend and seasonality in the data, ARIMA models aim to describe the autocorrelations in the data.

4.1 Stationarity

4.2 Komponen penyusun

4.2.1 Differencing

Differencing is one method to produce a stationary time series from the non-stationary one.

$$y'_t = y_t - y_{t-1} \tag{22}$$

The equation above can be rewritten as

$$y_t = c + y_{t-1} + \epsilon_t \tag{23}$$

where c is the average of the changes between consecutive observations and ϵ_t is a white noise. Eq 23 will be use intensively further in this section.

4.2.2 Backshift Notation

backshift notation B operating on y_t means shifting the data back one period, for example

$$By_t = y_{t-1} \tag{24}$$
$$B(By_t) = B^2 y_t = y_{t-2} \tag{25}$$

Some references use L notation instead of B. This notation is convenient for describing the process of differencing.

$$y'_t = y_t - y_{t-1} = y_t - By_t = (1 - B)y_t \tag{26}$$

Thus, a dth-order difference can be rewritten as

$$(1 - B)^d y_t \tag{27}$$

4.2.3 Auto Regressive Model

In an autoregression model, we forecast the variable of interest using a linear combination of past values of the variable.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t \tag{28}$$

where ϵ_t is white noise. This is like regression with lagged values of y_t as the predictors. With the maximum lag p, this is called an autoregressive model of order p, **AR(p) model**.

4.2.4 Moving Average Model

Moving average models should not be confused with the moving average smoothing we discussed in Section 1. A moving average model is used for forecasting future values, while moving average smoothing is used for estimating the trend-cycle of past values.

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \tag{29}$$

We call this as an **MA(q) model**, a moving averae model of order q.

4.3 Non-Seasonal ARIMA Model

A non-seasonal ARIMA (AutoRegressive Integrated Moving Average) is the combination of differencing, autoregression, and moving average model.

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \tag{30}$$

This is the full model of arima, which can be called **ARIMA(p, d, q)**, where

$$
\begin{array}{lcl}
p & = & \text{order of the autoregressive part} \\
d & = & \text{degree of first differencing involved} \\
q & = & \text{order of the moving average part}
\end{array}
$$

With this notation, we can consider our aforementioned models as special cases of ARIMA

White noise	ARIMA (0,0,0)
Random walk	ARIMA (0,1,0)
Autoregression	ARIMA (p,0,0)
Moving average	ARIMA (0,0,q)

Eq 30 can be rewritten using *backshift notation* B to simplify the equation as follows

$$(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - B)^d y_t = c + (1 + \theta_1 B + \cdots + \theta_q B^q)\epsilon_t \tag{31}$$

4.4 Seasonal ARIMA Model (SARIMA)

ARIMA models are capable of modelling seasonal data. The basic idea is the same with the non-seasonal model, but with the additional seasonal terms. The seasonal ARIMA model can be writtern as

$$\text{ARIMA}(p, d, q)(P, D, Q)_m$$

where m is the number of observations per year. The uppercase notations are for the seasonal part, while the lowercase notations are for non-seasonal part. The seasonal part consists a similar model to the non-seasonal, but involve backshift of the seasonal period. The complete seasonal ARIMA model is shown as follows

$$
\begin{aligned}
&(1 - \phi_1 B - \cdots - \phi_p B)(1 - \Phi_1 B - \cdots - \Phi_P B^m)(1 - B)^d (1 - B^m)^D y_t = \\
&c + (1 + \theta_1 B + \cdots + \theta_q B^q)(1 + \Theta_1 B + \cdots + \Theta_Q B^m)\epsilon_t
\end{aligned} \tag{32}
$$

Based on the model, suppose we want to calculate ARIMA(1,1,1)(1,1,1)$_4$ model without a constant, then the equation will become

$$(1 - \phi_1 B)(1 - \Phi_1 B^4)(1 - B)(1 - B^4) y_t = (1 + \theta_1 B)(1 + \Theta_1 B^4)\epsilon_t \tag{33}$$

4.5 Seasonal ARIMA Exogeneous Model (SARIMAX)

SARIMAX is the SARIMA with additional external variables to be use as the predictors of the autoregression part. The seasonal SARIMAX model can be writtern as

$$\mathrm{ARIMA}(p, d, q)(P, D, Q)_m(X)$$

SARIMAX general model is defined as

$$y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_k X_{k,t} + \omega_t \tag{34}$$

where $(X_{1,t}, X_{1,t}, \ldots, X_{k,t})$ are observations of k number of external variables corresponding to the dependent variable y_t; $(\beta_0, \beta_1, \ldots, \beta_k)$ are regression coefficients of external variables; ω_t is a stochastic residual which can be represented in the form of ARIMA model as follows

$$\omega_t = \frac{\theta_q(B)\Theta_Q(B^m)}{\phi_p(B)\Phi_P(B^m)(1-B)^d(1-B^m)^D}\epsilon_t \tag{35}$$

Notice that Eq 35 is similar to Eq 32.

4.6 Fractional ARIMA (ARFIMA)

ARFIMA (autoregressive fractionally integrated moving average) model is the same as ARIMA, however the value of differencing parameter d is allowed to be non-integer values. Therefore, the calculation of differencing part is redefined to accommodate it using binomial series expansion as follows

$$\begin{aligned}
(1-B)^d &= \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k && (36)\\
&= \sum_{k=0}^{\infty} \frac{\prod_{a=0}^{k-1}(d-a)(-B)^k}{k!} && (37)\\
&= 1 - dB + \frac{d(d-1)}{2!}B^2 - \ldots && (38)
\end{aligned}$$

These models are useful in modeling time series with long memory—that is, in which deviations from the long-run mean decay more slowly than an exponential decay.

Bahan Woconan

[1] Brown, R. G. (1957, January). Exponential smoothing for predicting demand. In Operations Research (Vol. 5, No. 1, pp. 145-145). 901 ELKRIDGE LANDING RD, STE 400, LINTHICUM HTS, MD 21090-2909: INST OPERATIONS RESEARCH MANAGEMENT SCIENCES.

[2] Holt, C. C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages, Office of Naval Research. Research memorandum, 52.

[3] Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on 29 February 2020

[4] Cools, M., Moons, E., & Wets, G. (2009). Investigating the variability in daily traffic counts through use of ARIMAX and SARIMAX models: assessing the effect of holidays on two site locations. Transportation research record, 2136(1), 57-66.