<MULT1>

Elastic-Net Regularization in Statistical Downscaling to Estimate Rainfall

Dindha Fadhilah Dinati [a] Akbar Rizki [b] Aji Hamim Wigena [c]

Department of Statistics, Faculty of Mathematics and Natural Science, Bogor Agricultural University, Bogor, Indonesia

Corresponding authors:
[a]dindhafdinati@gmail.com
[b]akbar.ritzki@gmail.com
[c]ajiwigena@ymail.com

Abstract. The impact of climate change and the importance of rainfall as the climate elements so that estimation of rainfall is important to be implement. The estimation of rainfall used statistical downscaling (SD) technique which utilize functional relationship approach between local scale data (rainfall) and global scale data (Global Circulation Model output-GCM). In general, GCM output data are multicollinear, so it is required a technique to overcome multicollinearity such as elastic-net regularization. The data used are GCM output above Indramayu regency as explanatory variables and the monthly rainfall data ZOM 79 as response variables. The result shows that the rainfall estimations with elastic-net at ZOM 79 from 2010 to 2013 were consistent.

Keywords: Elastic-net, multicolinearity, statistical downscaling.

INTRODUCTION

Rainfalls are important things to be observed because of the climate change that has occurred today. Extreme rainfall will give negative effect on various fields, especially agriculture. Therefore, rainfall estimation is necessary to avoid damage.

Statistical Downscaling represents transfer function between a local scale variable and global scale variables. GCM is a basic tool used for modeling climate changes [1]. Due to GCM is one of global information models, an estimation technique of local scale climate variables is needed to generate highly accuracy results [2]. One of the techniques that obtain local scale information from GCM output is Statistical Downscaling.

GCM output has large dimensions and generally consists of correlated variables or in condition of multicollinearity which is problem in multiple regression modeling [3]. This problem usually overcomes by Principal Component Regression (PCR), Ridge regularization, and Least Absolute Shrinkage and Selection Operator (LASSO) regularization. Pusporini applied ridge and lasso regularization to overcome multicollinearity [4].

Ridge regression as continuous shrinkage method achieves its better estimation performance through a bias-variance trade-off. However, ridge regression cannot produce a parsimonious model, for it always keeps all the estimators in the model [5]. Lasso regularization arises to overcome ridge problem. Lasso does both continuous shrinkage and automatic variable selection simultaneously [5]. Although the lasso has shown success in many situations, it has some limitations. One limitation in lasso is if there is a group of variables among which the pair wise correlations are very high, then the lasso tend to select only one variable from the group and does not care which one is selected [5]. Therefore, Zou and Hastie propose a new regularization technique called the elastic-net.

Elastic-net is new regularization that combines the penalties of ridge and lasso. Similar to the lasso, the elasticnet simultaneously does automatic variable selection and continuous shrinkage, and is able to select groups of correlated variables [5]. The elastic-net selects variables like the lasso, and shrinks together the coefficients of correlated predictors like ridge [6].

LITERATURE REVIEW

Statistical Downscaling and General Circulation Model

GCM is important data in understanding the climate system because it provides information about the climate change in the future [7]. GCM data in the form of grids indicate that the GCM is a form of spatial data. GCM is mathematical representation of the interaction of physics, chemistry, and dynamics of the Earth's atmosphere. Research using GCM is often constrained in the broadest selection, methods, and domain; it can certainly make it difficult to continue the analysis. The approach can be adopted to solve the problem is to determine the domain using statistical downscaling [8].

Downscaling is a technique to make future projections of local climate data using the coarse resolution data GCM outputs. There are several kinds of downscaling approaches; one of them is statistical downscaling. Statistical downscaling is an empirical approach to the statistical relation between the atmospheric circulation and rainfall. Generally, any successful statistical downscaling should satisfy three main conditions: (i) the link between predictands and predictors has to be strong in order to explain satisfactorily the local climate variability; (ii) the predictor variable should be well simulated by the GCM; and (iii) the relationship between predictands and predictors should not change within change in time, and should remain the same in a changed future climate [9]. A common form of SD model is:

$$y_{(tx1)} = f(X_{(txp)})$$

$y_{(tx1)}$ = vector local climate variable (rainfall)

$X_{(txp)})$ = matrix GCM output variable (precipitation)

t = amount of time (monthly)

p = amount of domains GCM grid

Ridge Regularization

Gulud regularization, introduced by Arthur E Hoerl and Robert W Kennard in 1970, became one of the solutions for multicollinearity problems [10]. Coefficient estimation of ridge regression carried out by adding a penalty in the minimization of sum squared error in linear regression (ordinary least square):

$$\sum_{j=1}^{p} \beta_j^2 \leq t$$

Coefficient estimation can be written in equation:

$$\hat{\beta} = \underset{\beta \in R}{argmin}\left[\sum_{i=1}^{t} (y_i - x_i{}'\beta) + \lambda \sum_{j=1}^{p} \beta_j^2\right]$$

With the penalty can be done by coefficient shrink of ridge regression. The amount of shrinkage is controlled by ridge parameter (λ). The larger the value of λ, the greater the amount of shrinkage. The coefficients are shrunk toward zero.

LASSO Regularization

Least absolute shrinkage and selection operator was introduced by Robert Tibshirani in 1995[11]. Lasso is shrinkage method like ridge, with subtle but important differences. The lasso estimate is defined by:

$$\hat{\beta} = \underset{\beta \in R}{argmin}\left[\sum_{i=1}^{t}(y_i - x_i{'}\beta) + \lambda \sum_{j=1}^{p}|\beta_j|\right]$$

With lasso penalty:

$$\sum_{j=1}^{p}|\beta_j| \leq t$$

That penalty causes the equations non linear in y, therefore it cannot be obtained solution in closed-form like ridge. In this case the quadratic programming required. Unlike the ridge, lasso regularization make the selection of correlated estimator variables because the small value of t can causes the coefficients is zero [6]. Therefore, lasso can do selection model.

Elastic-net Regularization

Lasso regularization has some limitations, there are [5]:

1. When $p > n$, then lasso only choose n variables included in the model
2. If there is a set of variables with high correlation, then lasso only pick one variable randomly
3. When $p < n$, lasso performance is dominated by ridge [11].

Therefore, Hui Zou and Trevor Hastie in 2005 introduced elastic-net penalty as follows:

$$\lambda \sum_{j=1}^{p}\left[(1-\alpha)\beta_j^2 + \alpha\,|\beta_j|\right]$$

If $\alpha = 0$ then elastic-net become ridge penalty, and if $\alpha = 1$ then elastic-net become lasso penalty. Elastic-net penalty give parameter estimation as the solution of the following equation:

$$\hat{\beta} = \underset{\beta \in R}{argmin}\left[\sum_{i=1}^{t}(y_i - x_i{'}\beta) + \lambda \sum_{j=1}^{p}\left[(1-\alpha)\beta_j^2 + \alpha\,|\beta_j|\right]\right]$$

METHODOLOGY

Data

This study uses the rainfall data in Indramayu district as dependent variable and GCM precipitation from climate model inter comparison project (CMIP5) with time lag as independent variables. Each of these data is a monthly data from January 1981 to December 2013. GCM data was obtained from website http://climexp.knmi.nl/ issued by the Netherlands KNMI (Koninklijk Nederlands Meteorologisch Instituut). The size of domain 8x8 grids over the area of Indramayu showed estimator more stable or consistent and not overly sensitive to outlier [1].

Methods

The steps of the analysis in this study were:

1. Determine the time lag of GCM using CCF (Cross Correlation Function). Time lag is used before modeling because estimation of rainfall using the GCM precipitation data with time lag was more accurate than using GCM precipitation without time lag [12]. If $r_{xy}(l)$ is cross-correlation between the x and y series at the time lag-l, $C_{xy}(l)$ is covariance between x and y at time lag-l, S_x and S_y is the standard deviation of x and y series respectively, the CCF can be formulated as equation:

$$r_{xy}(l) = C_{xy}(l)/S_x S_y$$

2. Modeling SD based on elastic-net regularization

   a. Divide data into training and testing data

      -period 1981-2009 as training data, period 2010 as testing data

      -period 1981-2010 as training data, period 2011 as testing data

      -period 1981-2011 as training data, period 2012 as testing data

      -period 1981-2012 as training data, period 2013 as testing data

   b. Perform cross validation with training data to determine optimum lambda from alpha 0.1-0.9

   c. Make statistical downscaling model with elastic-net regularization

3. Choose the best model based on criterion of Root Mean Square Error Estimation (RMSEP) and correlation between estimation and actual data. Best model uses to estimate rainfall.

4. See the consistency of rainfall estimation in 2010, 2011, 2012 and 2013 with observed mean and standard deviation value of RMSEP and correlation each year.

RESULTS AND DISCUSSION

Exploration

Highest average rainfall in ZOM 79 is at January with the intensity 263.36 mm/month. While, the lowest average rainfall intensity is 14.05 mm/month that happen in August. Highest rainfall in ZOM 79 occurred at January 2006 with the intensity 498 mm/month. Lowest rainfall occurs at July, August, September and October with the intensity 0 mm/month.

According to Oldeman Climate Classification, January, February and December are included as the wet months. Meanwhile, June, July, August, September and October are included as dry months. Then, March, April, May and November are included as humid months. It means the pattern of Oldeman Classification resembles monsoon pattern. Figure 1 shows the pattern of rainfall in ZOM 79 from 1981-2013 monsoon patterns. Figure shows that March, May, June, September and December there are outlier. It means that rainfall intensity is higher than normal condition in these months.

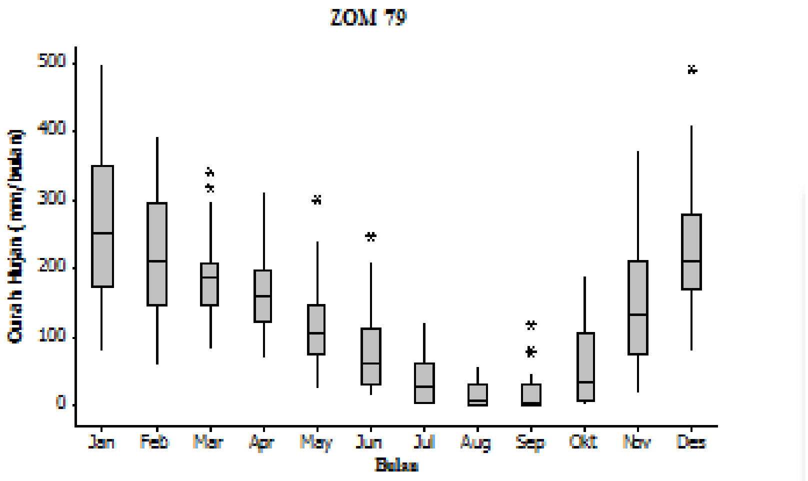

FIGURE 1. Boxplot rainfall in ZOM 79

Statistical Downscaling with Elastic-net Regularization

Cross validation use to find lambda optimum. Table 1 display the result of cross validation process. The value of alpha used to modeling is between 0.1 and 0.8. Smallest RMSEP are at 2011 and 2012, so this data used to estimate rainfall.

TABLE 1. Optimum parameter value in ZOM 79

Tahun	Alpha (α)	Lambda (λ)	CVE	Lasso parameter ($\lambda\alpha$)	Ridge parameter ($\lambda[1-\alpha]$)	RMSEP
2010	0.8	1.45	4978.22	1.16	0.29	71.41
2011	0.1	20.01	4963.32	2.00	18.01	66.35
2012	0.2	7.51	4934.05	1.50	6.01	48.07
2013	0.1	14.98	4867.35	1.50	13.48	76.35

Furthermore, Table 2 shows actual and estimation rainfall for 2011 and 2012. Rainfall estimation at 2011 has RMSEP 66.35. While at 2012, better value of RMSEP that is 48.07. Correlation value is used to show how close the pattern between actual and estimation rainfall. Closer to |1| then more better estimation result. Then Figure 2 shows observed pattern of actual and estimation rainfall.

TABLE 2. Actual and estimation rainfall at 2011 and 2012

	Estimation		Actual	
Month	2011	2012	2011	2012
Jan	258	247	110.13	205.88
Feb	240	219	129.25	175.63
Mar	201	179	202.38	170.25
Apr	156	148	267.5	104.75

Mei	116	121	80.85	60.13
Jun	76	77	60.5	24.75
Jul	40	34	21	0
Agu	19	4	0	0
Sep	30	18	0	0
Okt	63	63	106.63	22.25
Nov	156	141	122.73	40
Des	234	229	248.5	179.25
rmsep	-	-	66.35	48.07
r	-	-	0.72	0.95

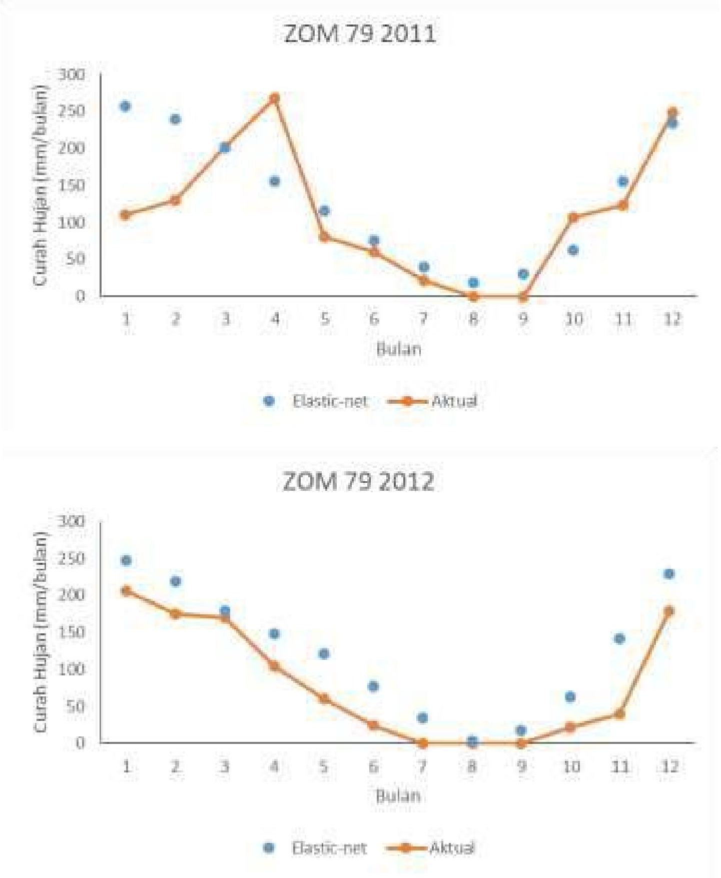

FIGURE 2. Actual and estimations rainfall patterns at 2011 and 2012

Consistency

Consistency of estimation result is necessary to determine the best model. SD elastic-net model is consistent in rainfall estimation from 2010 to 2013. Table 3 shows value of RMSEP and correlation in each year. The standard deviation shows the distance between rainfall estimation each year. Standard deviation of RMSEP value is 12.34 that indicate, the distance of estimation each year was small enough and the rainfall estimation is consistent. It also applied on correlation value. RMSEP and correlation are small determiner that the rainfall estimation is consistent. Consistent means that the estimation did not change within changes in year.

TABLE 3. Data consistency ZOM 79

Criteria	2010	2011	2012	2013	Standard Deviation
RMSEP	76.35	66.35	48.07	71.41	12.34
Correlation	0.67	0.72	0.95	0.75	0.12

CONCLUSION

The rainfall estimation using elastic-net regularization had good result especially in 2011 and 2012. The best model of ZOM 79 produced in 2012 with value of RMSEP 48.07 and correlation 0.95. This model is also consistent for annual estimation of rainfalls.

REFERENCES

1. A.H. Wigena, Modeling of Statistical Downscaling using Projection Pursuit Regression for Forecasting Monthly Rainfall, [Dissertation]. Bogor Agricultural University (in Indonesian), Indonesia, 2006.
2. E. Zorita, H. V. Storch, The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. Journal of Climate, 12 (1999) 2474-2489. http://dx.doi.org/10.1175/1520-0442(1999)012<2474:tamaas>2.0.co;2
3. A.H. Wigena, A. Djuraidah, A. Rizki, Semi parametric Modeling in Statistical Downscaling to Predict Rainfall. Applied Mathematical Sciences, Vol. 9 (2015), No. 88:4371-4382. http://dx/doi.org/10/12988/ams.2015.54362
4. Pusporini, Penerapan regresi gulud dan least absolute shrinkage and selection operator (lasso) dalam penyusutan koefisien regresi, [Skripsi]. Bogor Agricultural University, Indonesia, 2012.
5. H. Zou, T. Hastie, Regularization and variable selection via the elastic-net. J.R. Statist. Soc. B 67, Part 2:301-320 (2005).
6. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. Data Mining, Inference, and Estimation. Second edition. New York (US): Springer, 2008.
7. L. Handayani, A. H. Wigena, A. Djuraidah, Statistical downscaling with generalized additive model for extreme rainfall estimation. IOSR Journal of Mathematics. Vol.3, Issue 3:21-25.
8. U. Haryoko, Pendekatan Reduksi Dimensi Luaran GCM untuk Penyusunan Model SD, Bogor Agricultural University, Indonesia, 2014.
9. Busuioc, D. Chen, C. Hellstrom, Performance of statistical downscaling models in GCM validation and regional climate change estimates: Application for Swedish precipitation. Int. J. Climatol, 21 (2001) 557-578. http://dx.doi.org/10.1002/joc.624
10. A.E. Hoerl, R. W. Kennard, Gulud regression. biased estimation for nonorthogonal problems. Technometrics, Vol. 12, No. 1:55-67 (1970).
11. R. Tibshirani, Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 58, Issue 1:267-288 (1996).
12. S. Sahriman, A. Djuraidah, A. H. Wigena, Application of principal component regression with dummy variable in statistical downscaling to forecast rainfall, Open Journal of Statistics, 4 (2014) 678-686. http://dx/doi/org/10/4236/ojs.2014/19063