Prediction of Extremal Precipitation: Use of Quantile Regression Forests

Seoncheol Park[1], Junhyeon Kwon[2], Joonpyo Kim[3] and Hee-Seok Oh[4]

Department of Statistics

Seoul National University

Seoul 08826, Korea

November 20, 2017

[1]Graduate Student, Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea. pscstat@gmail.com

[2]Graduate Student, Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea. junhyeonkwon@gmail.com

[3]Graduate Student, Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea. joonpyokim@snu.ac.kr

[4]Professor, Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea. heeseok.oh@gmail.com

Abstract

This paper suggests a random forest method for spatio-temporal extreme value prediction of precipitation data. The proposed method is based on a quantile regression forests method, a kind of data mining method, improves other methods, tree and bagging. To represent monthly correlation of the data, we adopt circular transformed predictors. Throughout real data analysis, the empirical prediction performance of the proposed method is better than other methods and the prediction score of the proposed method is competitable with other team's method.

Keywords: Quantile regression forests, Circular transform, Ensemble, Spatio-temporal extremes.

1 Introduction

앙상블 기법은 그동안 많은 발전이 있었다. 그 중에서도 랜덤 포레스트는 성공적인 앙상블 예측 기법 중 하나다. 한편, (Meinshausen, 2006)은 quantile regression forests를 개발해 평균 예측만 사용되는 random forests를 분위수 예측에 사용할 수 있게 하였다. 분위수 회귀 포레스트를 가지고 높은 quantile을 예측하려는 시도는 여럿 있었다. Quantile regression forests는 기상 자료 분석에 많이 사용되어왔다. (Taillardat et al., 2016)은 quantile regression forest 기반 예측 방법을 프랑스 surface temperature와 wind speed 예측에 사용하고 기존 방법보다 좋은 예측 결과를 얻었다. (Aznarte, 2017)는 스페인 마드리드에서 기상요소들이 extreme NO_2 농도에 미치는 영향을 알아보기 위해 quantile regression forests를 사용했다.

한편, Circular-linear or circular-circular regression methods have been used in various area, especially in the environmental study. There are some previous studies using those kind of regressions. (Johnson and Wehrly, 1978) explained special joint distributions when linear and circular variables are in explanatory variables and they have special marginal distributions. After then they applied it to the regression problem with air pollution as a response variable, temperature as a linear predictor and wind direction as a circular predictor.. (Jammalamadaka and Lund, 2006)는 sine, cosine 변환을 한 month, wind direction을 predictor로써 포함하여 이들이 ozone level에 미치는 영향을 circular-circular regression으로 분석하였다. 본 연구에서는 이러한 사전연구들에 힌트를 얻어 네덜란드의 각 관측장소별 월 별 20-year return level에 해당하는 극단 강수량 예측을 QRF에 기반한 방법들로 예측해 보고 그 결과를 비교해 보았다. 월(month) 간 correlation을 반영하기 위해 circular-transformed predictor variable을 고려하였고 그 결과 예측 성능을 높일 수 있었다. 2장에서는 자료에 대한 소개 및 결측치 제거 방법에 대해 말하였다. 3장에서는 본 논문에서 사용한 방법론들을 소개하였다. 4장에서는 모형 구축 및 제안한 모형의 성능을 제시하였고 그 결과를 비교하였다. 마지막으로 5장에서는 본 연구의 한계점과 앞으로 나아가야 할 점들을 제시하였다.

2 Data Description and Missing Data Treatment

In this paper, we can consider five explanatory variables candidates, `year`, `month`, `date`, `longitude` and `latitude`, etc. Figure 1 shows scatter plots between predictor variable candidates and response variables.

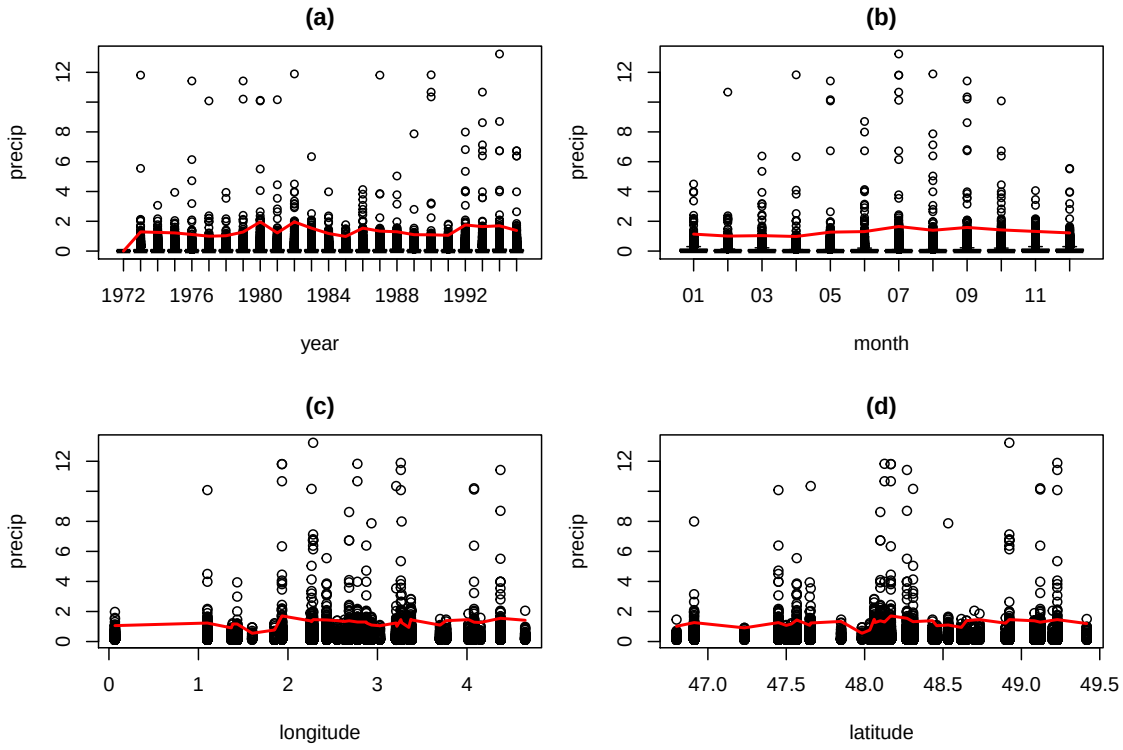

Figure 1: Scatter plot between precipitation and year (top left), month (top right), longitude (bottom left) and latitude (bottom right). Red lines mean empircal 0.998 conditional quantile.

From the explanatory analysis, we know some characteristics of the data. First, we have no strong linear trend between year and precipitation. That means we can' find significant increasing or decreasing trend of extreme precipitation across year. In this paper, we assume that year doesn't effect for the extreme precipitation in future 20 years. Second, we also find the relationship between longitue or latitude and extreme precipitation is nonlinear. In this paper, we assume that there exists some local increasing or decreasing behavior of extreme precipitation through longitude or latitude. Finally, there is a little seasonal trend of extreme precipitation. Extreme precipitaion is high in summer, low in spring.

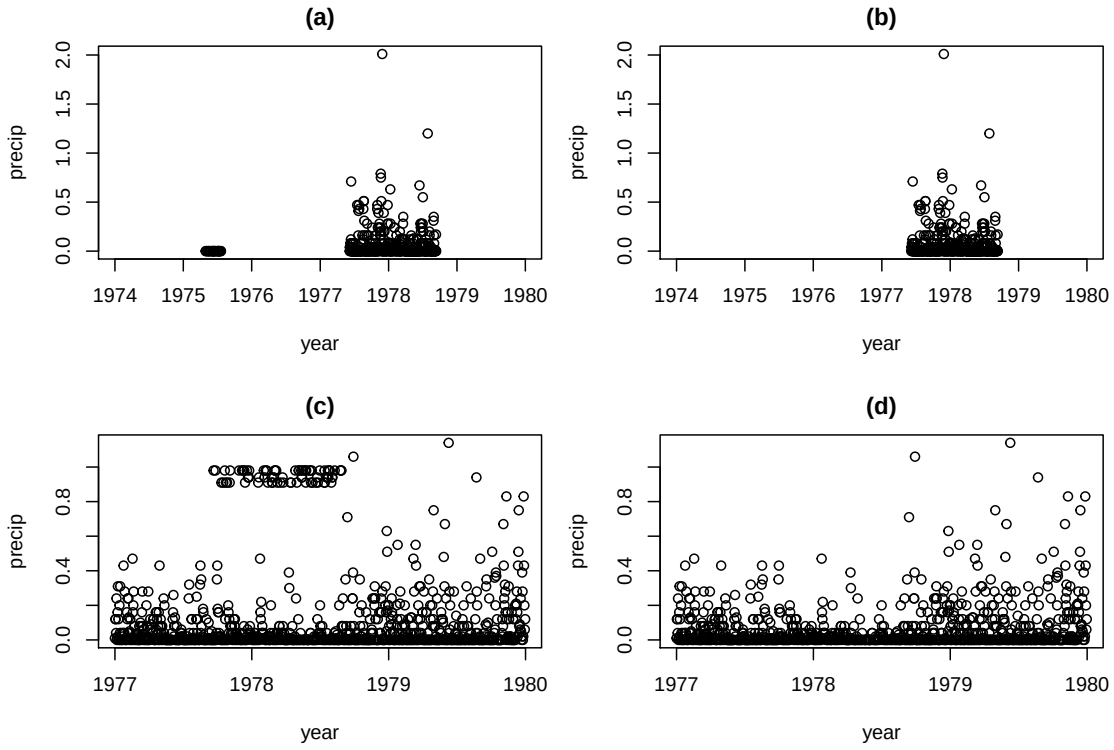

Figure 2: Illusration xample of data treatment. Since (a) original time series at station 31 have some zero values in 1975, (b) we delete those values. (c) In station 4, time series have some strange values from 1977 to 1979, therefore (d) we delete those values.

On the other hand, most of the time series have many missing values, as in Figure 2. That means, it is hard to apply time-series based approach to this data. As a result of explanatory data analysis, we could observe that some stations have too many missing values. Therefore,

traditional time-series based apporoach is hard to use in this data. A circular transform is used to express the monthly correlation.

3 Methods

3.1 Regression Tree

Decision tree is a kind of data mining methods and first suggested by (Breiman et al., 1984). When response variable is numeric, then it is called 'regression tree'. In this paper, we focus on regression tree.

Suppose there are n observations, let the set of data $\mathbf{Z}_i$, $\mathcal{D}_n = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ be a learning set. The data $\mathbf{Z}_i$ is $\mathbf{Z}_i = (Y_i, \mathbf{X}_i)$, where $Y_i, i = 1, \ldots, n$ be dependent variables and $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,p})$ be p-dimensional independent variables. In this paper, we assume that there is a non-linear, complex relationship between dependent variables and independent varibles. That means, we assume there is a non-linear, complex function f such that

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \qquad i = 1, \ldots, n.$$

Let $\mathcal{X}$ be the covariate space. The main idea of regression tree is making adequate partitions of $\mathcal{X}$ by diving $\mathcal{X}$ into L disjoint sets $\mathcal{R}_1, \ldots, \mathcal{R}_L$ to have homogeneous Y_i values in each partition. After then, regression tree fits a piecewise-constant prediction at each partition.

For further explanation, we introduce some notations in (Breiman, 2001). The leaf of a tree $l(\mathbf{x})$ is a rectangular subspace of $\mathcal{X}$. Then, the prediction of a single tree for a new data point $\mathbf{X} = \mathbf{x}$ is obtained by averaging over the observed values in leaf $l(x)$. Let the weight vector $w_i(\mathbf{x})$ represents whether the observation $\mathbf{X}_i$ is part of leaf $l(\mathbf{x})$ or not,

$$w_i(\mathbf{x}) = \frac{1_{\{\mathbf{X}_i \in R_{l(\mathbf{x})}\}}}{\#\{j : \mathbf{X}_j \in R_{l(\mathbf{x})}\}}. \tag{1}$$

The prediction of a single tree, given covariance $\mathbf{X} = \mathbf{x}$, is then the weighted average of the original observations $Y_i, i = 1, \ldots, n$,

$$\hat{f}(\mathbf{X}) = \hat{E}(Y|\mathbf{X} = \mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) Y_i. \tag{2}$$

Regression tree has some benefits. It is easy to interpret, can do explanatory variable selection implicitly. It also needs only a few tuning parameters, so the computation consideration is simple. It doesn't need any formal distributional assumptions, so it can also model the nonlinear relationship between explanatory variables and response variable. However, tree method also has some disadvantages. First, trees usually have high variance due to its greedy split process. That means small change in training data can give very different splits. Second, since the tree estimation is not smooth, it is not good when underlying function is smooth.

3.2 Random Forests

To overcome the disadvantages of regression tree, ensemble methods are suggested. (Breiman, 1996) suggested a new method called Bagging. The main idea of bagging is averaging many trees to bootstrap-resampled versions of the training data to reduce the variance of regression trees.

Furthermore, (Breiman, 2001) suggested random forests. In addition to bagging, random forests uses a random subset of predictor variables for covariate space split selection to make more decorrelated trees. It makes an improvement of the prediction performace of random forest compared to bagging.

Based on the notations of (Meinshausen, 2006), we summarize random forest algorithms. Suppose we generate k single trees for model construction. Then, for $t = 1$ to k, we repeat following three steps:

1. Draw a bootstrap sample from the training data.
2. Select m ($m < p$) variables at random and call them $\boldsymbol{\theta}_t$.
3. Grow a single tree $T(\boldsymbol{\theta}_t)$.

Then, the prediction of conditional mean $E(Y|\mathbf{X} = \mathbf{x})$ is given by an average prediction of k single trees,

$$\hat{f}(\mathbf{X}) = \hat{E}(Y|\mathbf{X} = \mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) Y_i, \tag{3}$$

where

$$w_i(\mathbf{x}) = \frac{\sum_{t=1}^{k} w_i(\mathbf{x}, \boldsymbol{\theta}_t)}{k}, \qquad w_i(\mathbf{x}, \boldsymbol{\theta}) = \frac{1_{\{\mathbf{X}_i \in R_{l(\mathbf{x},\boldsymbol{\theta})}\}}}{\#\{j : \mathbf{X}_j \in R_{l(\mathbf{x},\boldsymbol{\theta})}\}}. \tag{4}$$

For each tree grown on a bootstrap sample, the error rate for observations left out of the bootstrap sample is monitored. (Breiman, 2001) called it as "out-of-bag (oob)" error rate. In the notation of (Gregorutti et al., 2017), we consider an estimator based on the observation of a out-of-bag sample $\bar{\mathcal{D}}$

$$\hat{R}(\hat{f}, \bar{\mathcal{D}}) = \frac{1}{|\bar{\mathcal{D}}|} \sum_{i:(\mathbf{X}_i, Y_i) \in \bar{\mathcal{D}}}^{|\bar{\mathcal{D}}|} (Y_i - \hat{f}(\mathbf{X}_i))^2 \tag{5}$$

Then we can get the empirical permutation performance of the variable X_j,

$$\hat{I}(X_j) = \frac{1}{k} \sum_{t=1}^{k} [\hat{R}(\hat{f}_t, \bar{\mathcal{D}}_n^{tj}) - \hat{R}(\hat{f}_t, \bar{\mathcal{D}}_n^{t})], \qquad j = 1, \ldots, p \tag{6}$$

which can be used to measure the importance of each predictor variable.

3.3 Quantile Regression Forests

Sometimes we have an interest about conditional quantile, not conditional mean. (Meinshausen, 2006) tried to do random forest for conditional quantiles, names as quantile regression forests (QRF). The conditional distribution function of Y given $\mathbf{X} = \mathbf{x}$ is given by

$$F(y|\mathbf{X} = \mathbf{x}) = P(Y \leq y|\mathbf{X} = \mathbf{x}) = E(1_{\{Y \leq y\}}|\mathbf{X} = \mathbf{x}). \tag{7}$$

(Meinshausen, 2006) showed that these w_i in (4) can also be used to estimate the conditional distribution function, $\hat{F}(y|\mathbf{X} = \mathbf{x})$ by plugging in $1_{\{Y \leq y\}}$ instead of Y in equation (3):

$$\hat{F}(y|\mathbf{X} = \mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x}) 1_{\{Y_i \leq y\}}. \tag{8}$$

In QRF, they uses same weight $w_i(\mathbf{x})$ as in random forests. Therefore, QRF only changes the equation (3) to (8). The corresponding conditional quantile of level τ is

$$\hat{Q}_\tau(y|\mathbf{x}) = \inf\{y : \hat{F}_{Y|\mathbf{X}}(y|\mathbf{x}) \geq \tau\}. \tag{9}$$

3.4 Generalized Quantile Regression Forests

(Athey et al., 2016) suggest a new method called generalized random forests (GRF). The main differences between quantile regression forest and generalized quantile random forests are twofolds. First, they used esimating equation to solve various loss minimization problem. Second, they changed splitting scheme according to score function.

Suppose that we want to estimate a quantity, $\theta(\mathbf{x})$. In this challenge, $\theta(\mathbf{x})$ is the τ-th conditional quantile function, i.e., $\theta(\mathbf{x}) = Q_\tau(y|\mathbf{x})$. GRF solve the following local estimating equation:

$$E[\psi_{Q_\tau(\mathbf{x})}(Y_i)|\mathbf{X}_i = \mathbf{x}] = 0, \qquad \forall \mathbf{x} \in \mathcal{X}, \tag{10}$$

where $\psi(\cdot)$ is a appropriate score function. In quantile regression forests, $\psi_{Q_\tau(\mathbf{x})}(Y_i) = \tau \mathbf{1}(\{Y_i > Q_\tau(\mathbf{x})\}) - (1-\tau)\mathbf{1}(\{Y_i \leq Q_\tau(\mathbf{x})\})$. In this paper, we called it generalized quantile regression forests (GQRF).

(Athey et al., 2016) suggest a new spliting method based on above estimation equation using gradient-based approximation. Let the parent node be $P \subset \mathcal{X}$. Suppose $\hat{Q}_{\tau,P}(\mathbf{x})$ is computed by an empirical version of (10),

$$\hat{Q}_{\tau,P}(\mathcal{D}_n) \in \arg\min_{Q_\tau(\mathbf{x})} \{\| \sum_{\{i:(\mathbf{X}_i,Y_i)\in\mathcal{D}_n, X_i \in P\}} \psi_{Q_\tau}(Y_i)\|_2\} \tag{11}$$

The main ingredient of tree is dividing a parent leaf P into two child leaves, H_1 and H_2 with an suitable criteria.

1. First, GQRF gets pseudo-outcomes $\rho_i = \mathbf{1}(\{Y_i > \hat{Q}_{\tau,P}\})$ using the q-the quantile of the parent P, $\hat{Q}_{\tau,P}$ to .

2. Next, GQRF runs a regression split on the pseudo-outcomes ρ_i. Specially, GQRF splits P into two axis-aligned children H_1 and H_2 to maximize the criterion

$$\tilde{\Delta}(H_1, H_2) = \sum_{j=1}^{2} \frac{-1}{|\{i : X_i \in H_j\}|} \Big(\sum_{\{i\, X_i \in H_j\}} \rho_i \Big)^2. \tag{12}$$

3. Then, GQRF computes a gradient-based approximation of q-th quantile of the child nodes $\hat{Q}_{\tau,H}$ by

$$\tilde{Q}_{\tau,H} = \hat{Q}_{\tau,P} + \frac{1}{|\{i : X_i \in C\}|} \sum_{\{i:X_i\in H\}} \rho_i, \qquad H \in \{H_1, H_2\}. \tag{13}$$

4. GQRF recursively continues above steps.

Note that (Athey et al., 2016) don't claim their GQRF is alway better than quantile regression forests. Empirically, GQRF would produce smoother sample paths than QRF. Splitting rule of GRF is specially sensitive to quantile shifts in a way that regression splits are not. However, traditional mean-based regression trees are sensive to shifts of $E(Y|\mathbf{X})$.

3.5 Circular Transform

In this paper, we consider three types of circular transform of month. One of the natural way is that in Figure 3, we places a month in the clock. 12월을 (1,0) 즉 시계로 얘기하면 3시 방향에 놓고 counter-clockwise 방향으로 달들을 배치한 다음 이들의 `cosine`, `sine` 값을 설명변수로 사용하였고 이들 두 개의 변수들을 `cosmonth`, `sinmonth`로 부르기로 한다. We called circular-transformed month variable as `cosmonth` and `sinmonth`.

We consider monthly-circular transformed data (Figure 3). In monthly-circular transformed data, we put 12 (cosine, sine) value at each month.

We also consider finer-scale circular transforms. In Figure 4는 Figure 3를 좀 더 세분화하여 한 달을 early, late part로 나눈 후 circular transform을 취한 것이다. 이렇게 생성된 변수들을 `cosmonth*`, `sinmonth*`라고 부르기로 한다. Figure 5는 한 달을 early, middle, late part 셋으로 세분화 한 후 circular transform을 취한 것이다. 이렇게 생성된 변수들을 `cosmonth**`, `sinmonth**`라고 부르기로 한다.

원래 날은 연속된 형태의 자료이나 부득이하게 12개의 class로 나누어 계산되었는데, 이것을 좀 더 continuous하게 만들고자 하는 의도가 있었다. 그리고 여러 모형의 예측값을 가중평균으로 나타내 안정적인 예측 결과를 얻고자 하는 목표가 있었다. 그러나 완전한 continuous한 변수로 보기에는 자료의 수가 너무 적었기에 상순 중순 하순 보다 더 세밀하게 나누는 작업은 고려하지 않았다.

Further, when there is an irrgularity of maximum within the month, it may be useful to adopt finer division. (Figure 4 and Figure 5). It may also solve the discrete problem of original monthly-circular transform. 2번째 모형에서는 한 달의 상순, 하순의 0.998 quantile을 예측 한 후 이들의 평균을 취했으며, 3번째 모형에서는 한 달의 상순, 중순, 하순의 0.998 quantile을 예측한 후 이들의 평균을 취하였다.

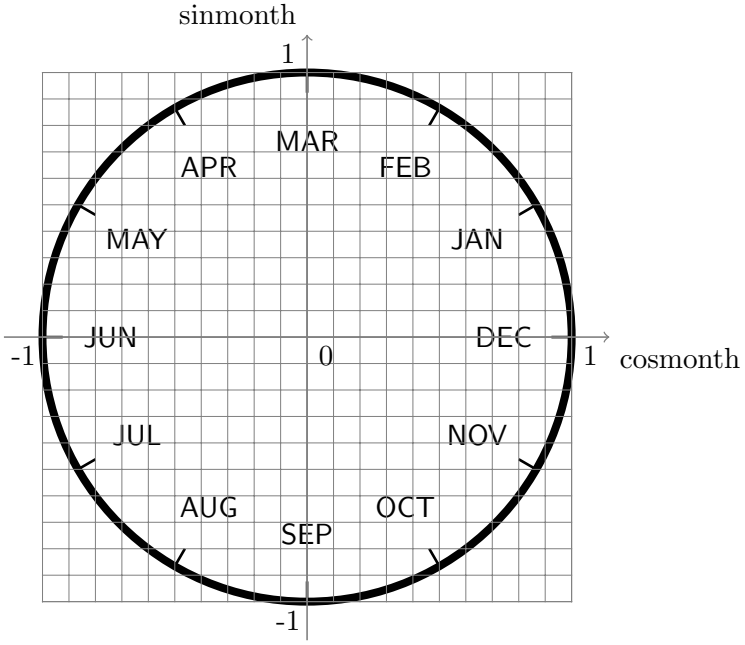

Figure 3: Circular transform of type 1.

We divide a month into two parts: early and late (cosmonth* and sinmonth*), or three parts: early, middle and late (cosmonth** and sinmonth**).

Model 2 and 3 gives a better prediction result when there is a strong relationship (달이 바뀔 시.) .

4 Models and Results

4.1 Models

We consider several models:

1. For preliminary results of challenge, we adopted circular-transformed quantile regression forests (CQRF) with precipitation(`precip`) as a response variable, `longitudes`, `latitudes`, (`cosmonth`) and (`sinmonth`) are predictor variables.

$$\text{CQRF: } \texttt{precip} \sim \texttt{lon + lat + cosmonth + sinmonth.} \qquad \text{(CQRF)}$$

2. For final results of challenge, we used an emsemble of circular-transformed quantile regression forests (ECQRF) with data-adaptive weight using inverse of losses. The

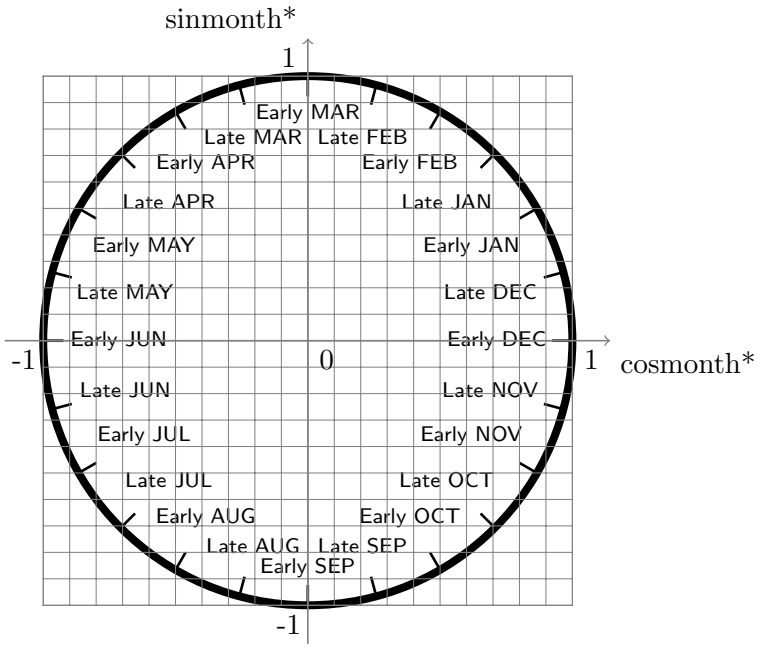

Figure 4: Circular transform of type 2.

idea is based on using more detailed division of the circular transform to consider more time information.

제시한 예측 모형은 3개 모형의 앙상블이다. 첫 번째 모형은 CQRF를 그대로 사용한 것이다. 두 번째 모형과 세 번째 모형은 `cosmonth`, `sinmonth`를 좀 더 세분화한 예측변수를 사용한 CQRF 모형을 사용하였다.

Then, we perform circular transform and quantile regression forests:

$$\text{Model 1: } \texttt{precip} \sim \texttt{lon + lat + cosmonth + sinmonth.} \qquad \text{(CQRF)}$$

$$\text{Model 2: } \texttt{precip} \sim \texttt{lon + lat + cosmonth* + sinmonth*.}$$

$$\text{Model 3: } \texttt{precip} \sim \texttt{lon + lat + cosmonth** + sinmonth**.}$$

ECQRF is the weighted average of three models. To decide the weights, we compute the inverse of losses. To do this, we divide the given data into training (80%) and validation (20%) set and compute the empirical losses on validation set. Given station $j, j = 1, \ldots, 40$ and month $k, k = 1, \ldots, 12$ sum of empirical losses on validation set is

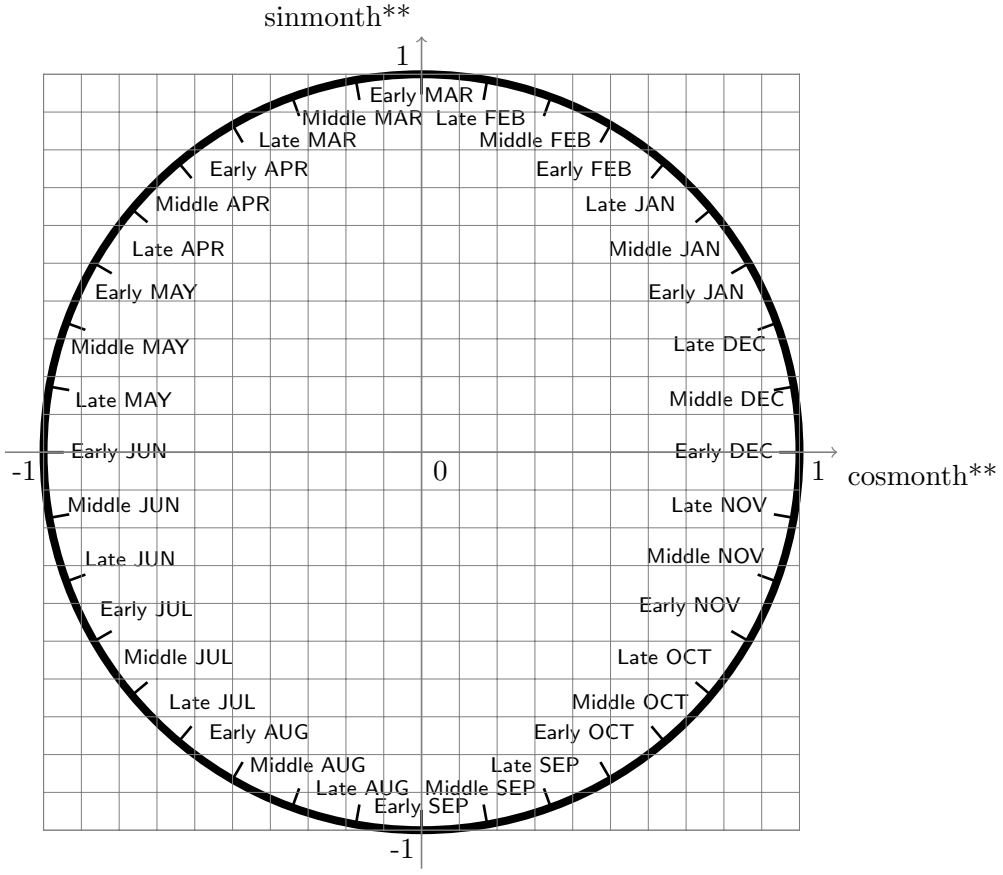

Figure 5: Circular transform of type 3.

$$S_{j,k}(\hat{Q}_{\text{Model i},j,k}) = \sum_{\text{day t of the validation period in month k}} l(P_{j,t}, \hat{Q}_{\text{Model i},j,k}), \qquad i = 1, 2, 3.$$

Then the weight matrix is

$$W_{\text{Model i}} = \left[w^i_{j,k} \right], \qquad w^i_{j,k} = \frac{1/S_{j,k}(\hat{Q}_{\text{Model i},j,k})}{\sum_{i=1}^{3} 1/S_{j,k}(\hat{Q}_{\text{Model i},j,k})}.$$

If there is no available training data for some (j, k) we use the average weight of whole losses, i.e., $w^i_{j,k} = \frac{1/\sum_{j,k} S_{j,k}(\hat{Q}_{\text{Model i},j,k})}{\sum_{i=1}^{3} 1/\sum_{j,k} S_{j,k}(\hat{Q}_{\text{Model i},j,k})}$.

Final ECQRF prediction is

$$\text{ECQRF: } W_{\text{Model 1}} \circ \hat{Q}_{\text{Model 1}} + W_{\text{Model 2}} \circ \hat{Q}_{\text{Model 2}} + W_{\text{Model 3}} \circ \hat{Q}_{\text{Model 3}},$$

where $\hat{Q}_{\text{Model}}$ is the average of predicted values from ten quantile regression forests using whole data, $\circ$ is the Hadamard product (entrywise product).

3. 비교군으로 circular transformed predictor 대신 linear (month) predictor를 사용한 것이 있다. (LQRF) Circular variable의 예측 중요성을 확인해보기 위해 linear numeric 변수로 취급하고 넣은 LQRF 방법 또한 예측력을 계산하였다.

4.2 Tuning parameter selection

Tuninig parameter selection is an important topic in random forests. By choosing appropriate tuning parameters, prediction model get better prediction performance. In this paper, we consider five tuning parameters. First three tuning parameters are fixed in our research:

1. number of trees to grow (`ntree` in `R`): we found that the number of trees has no significant effect to the prediction performance. In this paper, we fix the number of trees to grow at each iteration to 500.

2. Number of variables randomly sampled as candidates at each split (`mtry` in `R`): the default number of original random forest is $\frac{\lfloor p \rfloor}{3}$. Therfore, in this data, the default `mtry` is 1. We changed `mtry` to 2 and 4.

3. Number of samples in each bootstrap (`samplesize` in `R`): 0.632 * nrow(x), it is a default value of `randomForest` function in `R`. (Efron and Tibshirani, 1997)

4. Maximum size of terminal nodes (`nodesize` in `R`): default number is 5. We changed node size to 10 and 20.

To get more robust result, we compute each model at 100 times and computed their mean (and standard deviation).

4.3 Results

Table 1 and 2 show the prediction score results of various prediction methods. Among all method, CQRF with `ntrees=500`, 1 predictor and `minnode=5` is the best in both case, challenge 1 and 2. The selection of adequate `minnode` is important. However, there is no universal rule. When `mtry=1`, small `minnode` works well among all prediction methods. On the other hand, when `mtry` is equal to the number of predictor variables, i.e., bagging, performance is better when `minnode` is bigger.

In GCQRF, there is a small change across different `mtry` and `minnode`. That means GCQRF is robust method. However, the effect of decorrelated tree on the prediction performance is low in GCQRF.

bagging의 결과들은 왜 많이 차이 났는가?) `mtry`가 커질수록 tree들 사이의 correlation이 커진다. (Hastie et al., 2009) 따라서 `mtry`가 적절히 낮을 때, 본 연구에서는 `randomForest` 패키지의 default value인 1일 때 예측력이 제일 좋았다.

(Prediction result가 가장 높은 지역의 분석 결과 넣기)

(Variable importance plot) We explore variable importance plot of the best prediction model, CQRF with `mtry=1` and `minnode=5` in Figure 6.

4.4 Discussion

5 Limitations and Further Works

The limitations of our alorithm are follows:

`mtry`	1			2			bagging		
`minnode`	5	10	20	5	10	20	5	10	20
LQRF	0.5840	0.5835	0.5826	0.4686	0.4693	0.4681	0.2044	0.2061	0.2143
CQRF	**0.5977**	0.5967	0.5975	0.5685	0.5683	0.5723	0.2069	0.2041	0.2150
ECQRF	0.5936	0.5945	0.5770	0.5582	0.5598	0.5613	0.1023	0.1727	0.3114
GCQRF	0.5928	0.5927	0.5923	0.5777	0.5777	0.5777	0.5585	0.5620	0.5660

Table 1: Prediction score result with `ntree=500` for challenge 1 (after data treatment).

`mtry`	1			2			bagging		
`minnode`	5	10	20	5	10	20	5	10	20
LQRF	0.5650	0.5637	0.5636	0.4478	0.4487	0.4470	0.1924	0.1942	0.2109
CQRF	**0.5780**	0.5769	0.5775	0.5501	0.5494	0.5540	0.1931	0.1916	0.2020
ECQRF	0.5744	0.5751	0.5557	0.5401	0.5419	0.5428	0.0903	0.1589	0.2953
GCQRF	0.5730	0.5729	0.5725	0.5585	0.5585	0.5583	0.5397	0.5430	0.5470

Table 2: Prediction score result with `ntree=500` for challenge 2 (after data treatment).

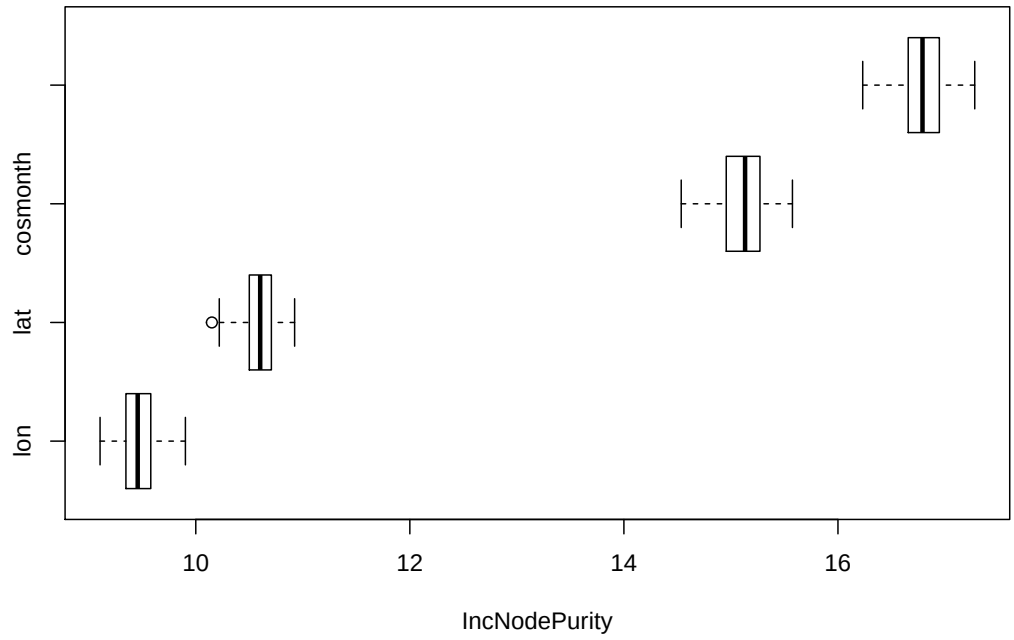

Figure 6: Example of variable importance plot of CQRF with `mtry=1` and `minnode=5`.

First, circular transform is unique when we consider cosine and sine transformed value simultaneously. However, in this paper, traditional random forest algorithm do prediction parameter selection separately. We need to improve the algorithm.

Second, it is known that the performance of random forest is better when the number of prediction variables is large. However, in this case, the number of predictor variables is small. 따라서 예측결과가 좋지 않을 수 있다.

We need to develop quantile regression forests for an extreme quantile (Generalized random forest 및 extreme conditional quantile updgrade 버전 참고).

References

Athey, S., Tibshirani, J., and Wager, S. (2016). Generalized random forests.

Aznarte, J. L. (2017). Probabilistic forecasting for extreme NO_2 pollution episodes. *Environmental Pollution*, 229:321–328.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. (1984). *Classification and Regression Trees*. Taylor and Francis.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.

Gregorutti, B., Michel, B., and Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, 27(3):659–678.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,*. Springer Science and Business Media, 2nd edition.

Jammalamadaka, S. R. and Lund, U. J. (2006). The effect of wind direction on ozone levels: a case study. *Environmental and Ecological Statistics*, 13(3):287–298.

Johnson, R. A. and Wehrly, T. E. (1978). Some angular-linear distributions and related regression models. *Journal of the American Statistical Association*, 73(363):602–606.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999.

Taillardat, M., Mestre, O., Zamo, M., and Naveau, P. (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Journal of Machine Learning Research*, 144(6):2375–2393.