Write Paper: Mongolia's Rainfall Forecast using Rstudio

Morais, Manoela N.

Abstract—One of the essential sectors of Mongolia's economy is agriculture, which is sensitive to climate variation. The most important climatic element which impacts agriculture productivity is the rainfall. Therefore, rainfall prediction becomes an important issue in Mongolia, especially for small farmers that lack technological resources. In this paper, I propose to model the daily rainfall prediction over Dornod Province to help farmers to forecast rain over 1 year. I applied several modeling methodologies, which include time series regression and Machine learning. The method that presented the lower MAE and RMSE was the Arima Model, followed by the default Simple Times series.

Keywords—Time Series Regression, Rainfall Forecasting, Statistical forecasting, Mongolia weather

I. INTRODUCTION

Rainfall information is important for crop selection and water resource management in agriculture. The occurrence of prolonged dry periods or heavy rain at the critical stages of crop growth and development may lead to a significant reduction of the farmer's crop yield. Align with that, climate changes that have been occurring in the last years make it difficult for the farmers to guess/predict the weather by just looking for past results. Nearly 90% can be used for agricultural or pastoral pursuits, 9.6% is forest and 0.9% is covered by water (XXXX). Therefore, rainfall prediction becomes a significant factor in agriculture to reach the best crop performance. I select to study for this paper, Dornod, an agricultural province in Mongolia, which its economy is largely based upon agriculture.

1.2 Research Goal

The main goal of this research is to propose a solution to help farmers selecting their crop and its start farming period by giving them useful rainfall prediction. Therefore, I am going to test several simples' models to identify one that can predict with lower errors the rainfall forecast for Dornod Province.

DATA

I am using 5 years of meteorological data between January 2015 and May 2020 from the Nasa website from Dornod Province (see all Soums in Figure 01). From the Nasa available indices, a set of 14 indices of climate variables were selected for this study (Figure 2). For long-range forecast (LFR), which is more than 2 weeks and up to 2 years, generally, the rain forecast is consistent with fundamental variables such as previously temperature and precipitation (Doblas-Reyes & all, pp. 12). However, we are using other

	Dornod Province	
	Soum (unit/district)	**Climate**
1	Chuluun khoroot (Ereentsav)	wet
2	Dashbalbar	wet
3	Bayandun	wet
4	Bayanuul (Javartkhoshuu)	wet
5	Gurvan zagal	wet
6	Tsagaan Ovoo	dry
7	Sergelen	dry
8	Choibalsan	dry
9	Bayantumen (Tsagaanders)	dry
10	Kherlen	dry
11	Bulgan	dry
12	Hulunbuir	moderately mild
13	Matad	moderately mild
14	Khalkhgol	wet

References: North, Central, South

Figure 1. Dornord Provinces used for weather model study

variables to identify the interference and importance in the rainfall prediction results.

Of the selected variables, seven indices refer to temperature, one to precipitation, and 4 to other variables. The temperature indices describe cold extremes as well as warm extremes. The precipitation indices describe wet extremes.

Parameter(s):	Units	Description
T2M	C(Celsius)	Temperature at 2 Meters
TS	C(Celsius)	Earth Skin Temperature
PS	kPa	Surface Pressure
T2M_MIN	C(Celsius)	Minimum Temperature at 2 Meters
RH2M	%	Relative Humidity at 2 Meters
WS2M	m/s	Wind Speed at 2 Meters
ALLSKY_TOA_SW_DWN SRB	(MJ/m^2/day)	Top-of-atmosphere Insolation
ALLSKY_SFC_SW_DWN SRB	(MJ/m^2/day)	All Sky Insolation Incident on a Horizontal Surface
PRECTOT	(mm day-1)	Precipitation
T2M_RANGE	C(Celsius)	Temperature Range at 2 Meters
T2MWET	C(Celsius)	Wet Bulb Temperature at 2 Meters
T2MDEW	C(Celsius)	Dew/Frost Point at 2 Meters
ALLSKY_SFC_LW_DWN	(MJ/m^2/day)	Downward Thermal Infrared (Longwave) Radiative Flux
T2M_MAX	C(Celsius)	Temperature at 2 Meters

Figure 2. 14 Indices selected from the Nasa climatology website available indices.

II. Modeling

The methodology used will be first, get historical data and random from the Nasa website. If the model is not times series related, I will divide it in 2 categories in 80% for training (build the models) and 20% for test the models (Figure 04). If the model contains time-series I will dived for the training 2015-2019 data and 2020 results will be used for testing. Then, I am going to run several regression models for rain forecast for autoregressive and multivariable regressions. Finally compare RMSE and MAE results and select the one, that present lower results. For illustration of the models I going to represent initially Dashbalbar results, for the others soum please vide table at attachment.

A quantitative forecast of rainfall is extremely difficult and realizable. Generally, it is done only a couple hours of their occurrence with a Doppler. However, for agriculture operations, a quantitative forecast of rain is not as important as a forecast of the (i) non-occurrence of rains (dry spells) and (ii) type of rain spell that can be expected. Therefore, after modeling and get the results in "millimeters of rainfall", I am going to classify the rainfall based on the USG definition of raining (Figure X). USG defines that Absent of raining <=0 mm/hh, Slight rain <=0.5mm/hh, Moderate<=4mm/hh, and Heavy rain (High)>8mm/hh rainfall. Although the information is given by USG in mm/hour, I am going to consider the same value for the whole day (mm/day) to classify. The main goal of doing this classification is to give farmers the information they need and decrease forecast errors. The main goal of doing this classification is to give to farmers the information they need and decrease the forecast errors.

Classification from website		
No rain	0 - 0.001	mm/hour
Slight rain:	0.001 - 0.5	mm/hour
Moderate rain:	0.5 - 4	mm/hour
Heavy rain	4 - 8	mm/hour

Figure 3. Rain data classification extracted from the USG. Gov Resource: https://www.usgs.gov/mission-areas/water-resources

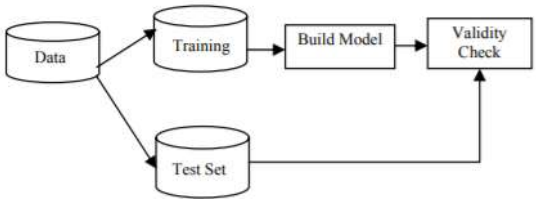


Figure 4. Overview of the forecasting Planning methodology

III. Single Regression Models

The autoregressive process is a regression on itself. Therefore, Yt is a linear combination of the p most recent past values of itself plus the term "et" that incorporates the error (Equation 01). We are going to start this stage by establishing a Baseline model,

which is basically the mean of the historical data, to be used as a reference and compared with the linear Times series model, Times series *Seasonal naïve Method, and Arima model. The Y(t) will be rainfall (PRECTOT).*

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t$$

Baseline Model

Assuming that all variables are independent and there is no rainfall prediction (no time series), I constructed a dependent guess (not random) based on the mean of the value itself, using the training data. The Results obtained were RMSE= 2.187456 and MAE= 1.190749, these results will be used as a reference to compare with other models.

```
> Base.Model <- mean(train$PRECTOT)
```

Times Series linear model (TS)

The rainfall observations collected through daily data is sequential over time. Therefore, we are going to use time-series to model the stochastic mechanism that gives rise to an observed series. I am going to predict and compare it with the test set. If one of those models has the best fit, I will forecast it. We will start by analyzing the data extracted from NASA, if the variables' time series are (non) stationarity and possess a unit root.

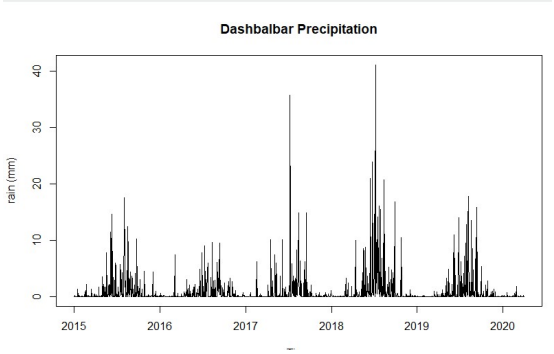


Figure 5.Precipitation for Dashbalbar through the last 5 years. Extracted from RStudio. We can see times series and a seasonal trend.

Checking UnitRoot

To make statistical inferences about the structure of a stochastic process on the basis of an observed record of that process, we need to verify the assumption that the data is stationarity and therefore, the probability laws that govern the behavior of the process do not change over time. In a sense we formulate the null and alternative hypotheses for the unit root. Considering the equation for AR(p):

$$\Delta Y_t = \alpha + \delta t + \rho Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \cdots + \gamma_{p-1} \Delta Y_{t-p+1} + \varepsilon_t$$

H0: ρ=0 (non-stationary, has UNIT ROOT)

Ha: $\rho < 0$ (stationarity)

When Y has no unit root and stationarity condition $2<\rho<0$ holds after incorporating a time trend, we call these series trend stationary. We tested the unit roots with Augmented Dickey-Fuller Test (ADF).

```
#Unit root test
library(fUnitRoots)

#new data
adfTest(da$WS2M,type="c") # type="nc" no cons
adfTest(da$TS,type="c") # type="nc" no consta
adfTest(da$PRECTOT,type="c") # type="nc" no c
adfTest(da$RH2M,type="c") # type="nc" no cons
adfTest(da$T2MDEW,type="c") # type="nc" no co
adfTest(da$T2M_MAX,type="c") # type="nc" no c
adfTest(da$T2M_MIN,type="c") # type="nc" no c
adfTest(da$T2M,type="c") # type="nc" no const
adfTest(da$T2M_RANGE,type="c") # type="nc" no
adfTest(da$PS,type="c") # type="nc" no consta
adfTest(da$T2MWET,type="c") # type="nc" no co
adfTest(da$ALLSKY_TOA_SW_DWN,type="c") # type
adfTest(da$ALLSKY_SFC_SW_DWN,type="c") # type
adfTest(da$ALLSKY_SFC_LW_DWN ,type="c") # typ
```

The Augmented Dickey-Fuller (ADF) test uses the t-statistic from the regression and compares it to a DF-t critical value that can be found in statistical packages. If the unit root hypothesis ρ=0 has not been rejected, the conclusion is that at least one-unit root exists in the process and we need to test for possible second unit root for the differenced series ΔY until we get a stationary process. From results of all 14 indices we see that for H0: unit root hypothesis is rejected since p-value <0.01 <significance level of 10% (or 5%). And therefore, we can assume that the indices are stationary and can be used as it is, without the need of differentiation.

After creating a simple linear time series with a simple seasonality and an increasing trend and forecast for a year value we obtained a the RMSE=1.784812 and a MAE =0.9783169, both lower than the base model.

```
######################## Times Series -linear model ########

ts.train<-ts(train$PRECTOT,frequency=365.25,start=c(2015))
ts.reg<-ts(x$PRECTOT,frequency=365.25,start=c(2015))

#TEST PERFOMANCE
test.pred.ts<- predict(ts.train,test$PRECTOT, h=365)
RMSE.ts <- sqrt(mean((test.pred.ts$mean)^2))
RMSE.ts
MAE.ts <- mean(abs(test.pred.ts$mean))
MAE.ts
```

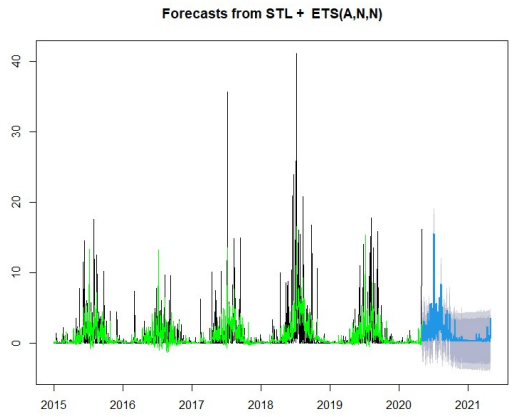


Figure 6. Forecast from Datalkar. PRECTOT for 365 days using default Times series method. Where we can see in green the forecast model in top of the past historical values.

After some days I tested again this model adding more recently data. The plot gave me Figure 07. Although, RMSE and MAE slight changed, we can see that the mean in non-Zero (there is no period without raining!). Which shows that for long periods (365 days) we might have a problem forecasting values.

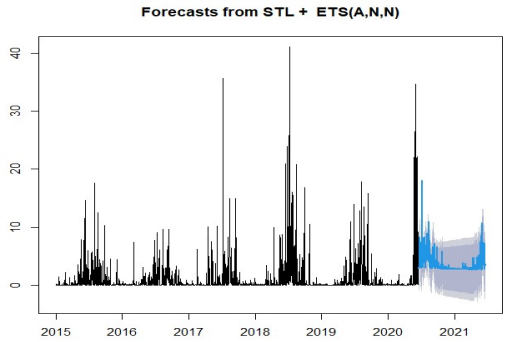


Figure 7.Forecast from Datalkar. PRECTOT for 365 days using default Times series method. After recently shower in the place.

TS + Seasonal naïve Method

The seasonal Naïve (snaive) model verify if rainfall in the place are highly seasonal. The seasonal naive model makes the forecast using the most recently observation from the same season (Equation 2). The Z[t] is normal error. The Exponential smoothing generally is good for short term forecast and refer to error, trend and seasonality. The model uses the exponentially weighted moving average (EWMA) to "smooth" a time series and trying to eliminate the random effect (RPubs).

$$Y[t]=Y[t-m]+Z[t] \quad \text{Equation 2}$$

```
###################### Method 02: TS+ Naive (Random Walk Forecasts)

fit_SN.train <- snaive(ts.train,frequency=12*30.4375)
fit_SN <- snaive(ts.reg,frequency=12*30.4375)


#preformance
test.pred.sn<- predict(fit_SN.train,test$PRECTOT, h=365.25)
RMSE.sn <- sqrt(mean((test.pred.sn$mean)^2,na.rm=TRUE))
RMSE.sn
MAE.sn <- mean(abs(test.pred.sn$mean),na.rm=TRUE)
MAE.sn
```

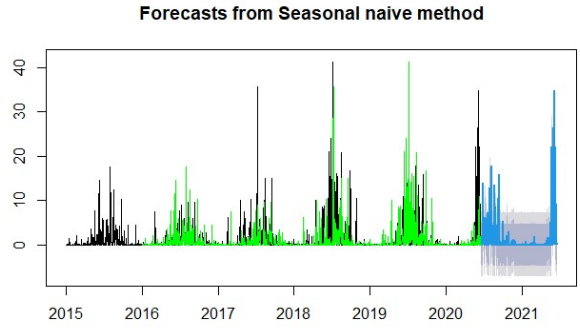


Figure 8. Forecast PRECTOT for 365 days using Times series with exponential smoothing method.

I got for this model a RMSE = 3.373169, which increased compared with than the base model, however the MAE= 1.110575 was lower. Which means the ET+ Snaive model fit better, however the error when happens is bigger the base model.

Arima Method

Arima also known as AutoRegressive Integrated Moving Average models is widely used approaches for time series weather forecasting. Arima Model is a combination of three mathematical models, using autoregressive(p), integrated(d), moving-average (q) (ARIMA) models for time series data. An ARIMA (p, d, q) model can account for temporal dependence in several ways. Firstly, the time series is d-differenced to render it stationary. If d = 0, the observations are modelled directly, and if d = 1, the differences between consecutive observations are modelled.

For this paper I started using the auto. arima function in Rstudio, the result was Arima (3,1,2) with non-Zero mean. This shows that Rstudio could not identify the seasonal part of the zero. This happened because Daily data is challenging as it often involves multiple seasonal patterns, and so we need to use a method that handles such complex seasonality. Therefore, I also runned the arima function with Fourier varying K from 1 to 25 and and the result was the same. The forecast can be seen in Figure 07.

$$\text{ARIMA}\ \underbrace{(p,d,q)}_{\text{Non-seasonal part of the model}}\ \underbrace{(P,D,Q)_m}_{\text{Seasonal part of of the model}}$$

```
#PERFOMANCE
  #train data
    y<- ts(x$PRECTOT, frequency=365, start =c(2015), end=c(2019))
    y.test<- ts(x$PRECTOT,frequency=365, start =c(2019), end=c(2020))
    head(y.test)
    plot(y.test)

  #fit_ARIMA <- auto.arima(y,seasonal = TRUE,approximation = FALSE,trace = FALSE)
    bestfit <- list(aicc=Inf)
    for(K in seq(25)) {
      fit <- auto.arima(y, xreg=fourier(y, K=K),
                        seasonal=FALSE)
      if(fit[["aicc"]] < bestfit[["aicc"]]) {
        bestfit <- fit
        bestK <- K
      }
    }
  summary (bestfit )

fit_ARIMA.predict <- forecast(bestfit,xreg=fourier(y, K=bestK, h=365))
fit_ARIMA.predict1<- predict(fit_ARIMA.predict, y.test )
```

Both the RMSE =1.497 and MAE=1.098 decreased compared with the the base model. The autocorrelation plot -

ACF graph in figure 08 shows shows that for the first500 lags, almost all sample autocorrelations fall inside the 95 % confidence bounds indicating the residuals appear to be random.

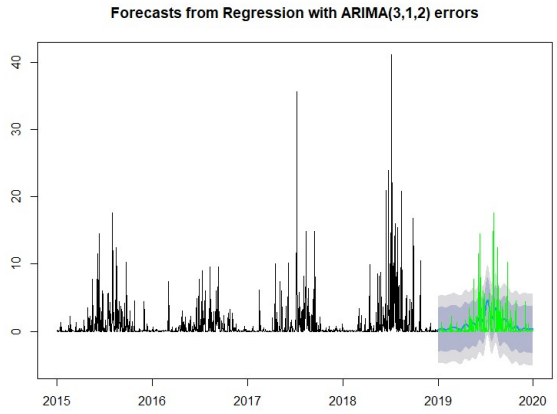


Figure 9.Forecast PRECTOT for 365 days using Times series with auto.arima method. Green the real value and in blue the forecast.

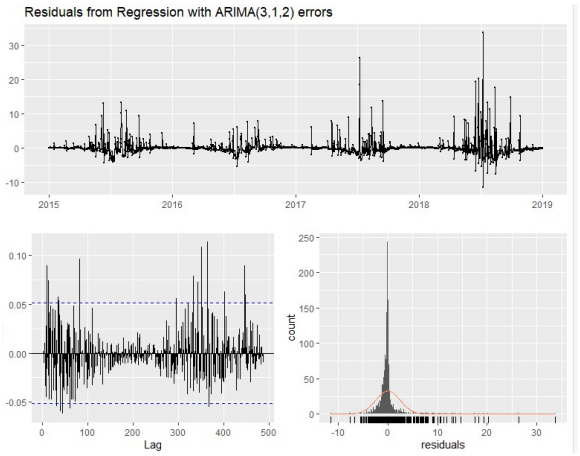


Figure 10. Residual from the forecast of PRECTOT for 365 days using Times series with auto.arima method.

IV. Multivariable linear Regression

For a Multiple linear regression, the dependent variable is assumed to be a linear function of several independent variables (predictors), where each of them has a weight (regression coefficient) that is expected to be statistically significant in the final model.

Multiple Log- linear regression

Using a log- linear regression model to the dependent variable (y_i =PRECTOT) in Equation x. We started with the all potential variables (Figure 06) and then eliminated from the model, those that were not statistically significant $p < 0.05$ (Figure 10). The adjusted R-squared for the log-linear model with all 12 variables is 0.59 which means that 59% of the variance in our dependent variable mm in rainfall (PRECTOT) can be explained by the set of predictors in the model; Although not all variables are significant in this first model, after we try to used just the significant values (figure 12), we can observe that the results of the Adjusted R-square for all variables > Adjusted R-square for significant ones. This happens, probably because the variables might have some degree of dependent between each other.

$$yi=\beta 0+\beta 1xi1+\ldots+\beta nxin+\epsilon i, \quad \text{Equation 01}$$

$$i=1,2,\ldots,n, \text{ where, } \varepsilon i \sim i.i.d.N(0,\sigma 2)$$

```
#(log)linear regression model using the training data

lin.reg <- lm(log(da$PRECTOT+1) ~ da$RH2M + da$T2M_MAX + da$RH2M  + da$WS2M, data = train)
summary(lin.reg)# Inspect the model
accuracy(lin.reg)
```

The result obtained from the log linear model with all the variables was a RMSE= 2.791962 that have slight increases when compared with the base model, however the MAE= 1.186184 was lower. This shows that the model fits better than the base model. This model makes some assumptions which includes linearity, constant variance, normality and independence between the parameters. We can see on figure 13 the normal QQ plot of the residuals, the values are not normal after x-axis grether than 1.5 where points are away from the line. In figure 14, we can see the plot of the forecast of the variable PRECTOT against the past value, using linear regression.

```
Call:
lm(formula = log(da$PRECTOT + 1) ~ +da$RH2M + da$T2M + da$T2M_MAX +
    da$WS2M + da$PS + da$TS + da$T2M_MAX + da$T2M_RANGE + da$ALLSKY_TOA_SW_DWN +
    da$ALLSKY_SFC_SW_DWN + da$ALLSKY_SFC_LW_DWN, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-2.03310 -0.21987 -0.03082  0.16271  2.22492

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.087496   1.578767   1.322 0.186247
da$RH2M              0.009889   0.001027   9.630  < 2e-16 ***
da$T2M              -0.042697   0.024636  -1.733 0.083237 .
da$T2M_MAX           0.017019   0.015060   1.130 0.258604
da$WS2M              0.027632   0.007134   3.873 0.000111 ***
da$PS               -0.048621   0.016499  -2.947 0.003248 **
da$TS                0.001602   0.015931   0.101 0.919911
da$T2M_RANGE        -0.031729   0.008296  -3.825 0.000135 ***
da$ALLSKY_TOA_SW_DWN  0.021880   0.001866  11.728  < 2e-16 ***
da$ALLSKY_SFC_SW_DWN -0.021240   0.001911 -11.114  < 2e-16 ***
da$ALLSKY_SFC_LW_DWN  0.095360   0.007770  12.272  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3829 on 1936 degrees of freedom
Multiple R-squared:  0.5925,    Adjusted R-squared:  0.5904
F-statistic: 281.5 on 10 and 1936 DF,  p-value: < 2.2e-16
```

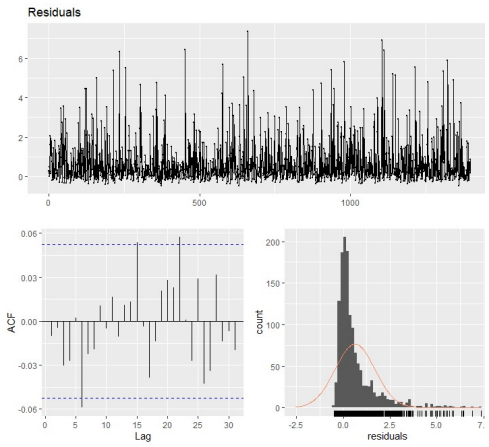


Figure 11. Results(summary) and residuals from the log linear regression for the variable PRECTOT against all others 12 variables.

```
Call:
lm(formula = log(da$PRECTOT + 1) ~ da$RH2M + da$WS2M + da$PS +
    da$T2M_RANGE + da$ALLSKY_TOA_SW_DWN + da$ALLSKY_SFC_SW_DWN +
    da$ALLSKY_SFC_LW_DWN, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-2.31184 -0.22521 -0.03196  0.16242  2.29516

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.4622050  1.5787914   2.193  0.02843 *
da$RH2M              0.0135607  0.0007869  17.232  < 2e-16 ***
da$WS2M              0.0320494  0.0070887   4.521 6.52e-06 ***
da$PS               -0.0532391  0.0166325  -3.201  0.00139 **
da$T2M_RANGE        -0.0318915  0.0029151 -10.940  < 2e-16 ***
da$ALLSKY_TOA_SW_DWN  0.0239658  0.0018510  12.947  < 2e-16 ***
da$ALLSKY_SFC_SW_DWN -0.0242215  0.0018773 -12.902  < 2e-16 ***
da$ALLSKY_SFC_LW_DWN  0.0442631  0.0027572  16.054  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3876 on 1939 degrees of freedom
Multiple R-squared:  0.5818,    Adjusted R-squared:  0.5803
F-statistic: 385.4 on 7 and 1939 DF,  p-value: < 2.2e-16
```

Figure 12. Results(summary) and residuals from the log linear regression for the variable PRECTOT against significant variables.

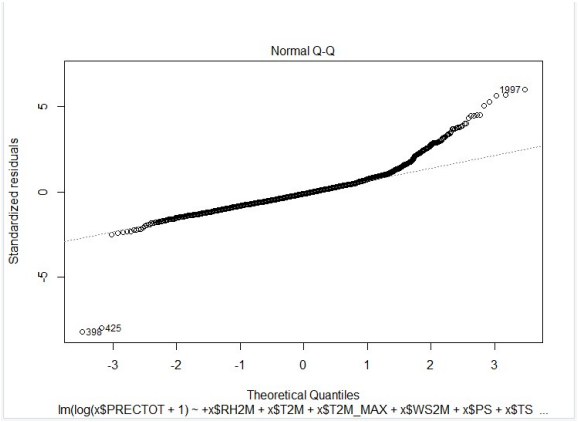


Figure 13. QQ plot of the residuals of the log-linear models for y(PRECTOT) against the 14 variables.

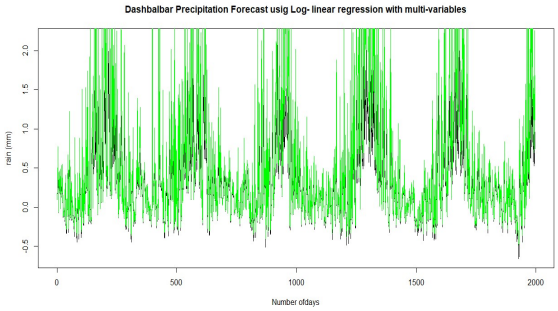


Figure 14. Graph pf PRECTOT (Y) Log- linear regression against the 14 multivariable. In green, the forecast and in black past historical rainfall measures.

```
Call:
lm(formula = log(x$PRECTOT + 1) ~ +x$RH2M + x$T2M_MAX + x$PRECTOT:x$T2M_MAX +
    x$WS2M + x$PS + x$TS + x$T2M_MAX + x$ALLSKY_TOA_SW_DWN +
    x$ALLSKY_SFC_SW_DWN, data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.61156 -0.15608 -0.03484  0.10990  1.78097

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.8005933  1.1011824   5.268 1.53e-07 ***
x$RH2M               0.0095309  0.0005910  16.125  < 2e-16 ***
x$T2M_MAX           -0.0449861  0.0033383 -13.476  < 2e-16 ***
x$WS2M               0.0175076  0.0051368   3.408 0.000667 ***
x$PS                -0.0660196  0.0117874  -5.601 2.43e-08 ***
x$TS                 0.0515676  0.0034551  14.925  < 2e-16 ***
x$ALLSKY_TOA_SW_DWN  0.0164018  0.0012586  13.032  < 2e-16 ***
x$ALLSKY_SFC_SW_DWN -0.0189389  0.0013047 -14.516  < 2e-16 ***
x$T2M_MAX:x$PRECTOT  0.0061400  0.0001269  48.376  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2833 on 1994 degrees of freedom
Multiple R-squared:  0.796,     Adjusted R-squared:  0.7952
F-statistic: 972.8 on 8 and 1994 DF,  p-value: < 2.2e-16
```

Figure 15.Result of the Log- linear regression after include an interaction term between precipitation and include only significant variable.

By incorporating an interaction term into our regression model and just including significant variables. Where β3xi1xi2 is the interaction between Precipitation and maximum temperature, we found a R-adjust of 0.795 (see figure 15). However, we found

a RME=16.29 and MAE=2.02, which means that the errors when happens are far much greater than the previous models.

Vector Autoregression (VAR)

VAR is an AR(p) in vector form where *Y*t " is a vector of several variables describing the dynamic system with variables for Rainfall attributes. A vector autoregression is a system of equations, where each equation for each variable contains lagged values of itself and lagged values of all other variables in the system. Below we present a VAR model which generated several equations estimated separately using OLS regression method. For this study was selected only the precipitation variable (PRECTOT). For computing the model I used the library (vars) in Rstudio and select the number of lags by using Schwarz information criterion (here it is called SC(n)):

```
$selection
AIC(n)  HQ(n)  SC(n) FPE(n)
     5      2      2      5

$criteria
                1           2           3           4
AIC(n)   8.873722    8.354423    8.276219    8.280325
HQ(n)    9.068972    8.730976    8.834076    9.019485
SC(n)    9.404172    9.377434    9.791791   10.288457
FPE(n) 7141.838373 4249.037184 3929.626722 3946.212891
                5           6           7           8
AIC(n)   8.273273    8.323168    8.355269    8.416600
HQ(n)    9.193737    9.424935    9.638340    9.880974
SC(n)   10.773966   11.316421   11.841083   12.394974
FPE(n) 3919.163537 4120.733557 4256.688260 4528.080432
                9          10          11          12
AIC(n)   8.480407    8.530198    8.618485    8.707522
HQ(n)   10.126085   10.357180   10.626770   10.897111
SC(n)   12.951343   13.493694   14.074542   14.656140
FPE(n) 4829.371837 5079.786319 5553.820833 6077.773638
```

Since Schwarz criterion selected a VAR model with only two lag we can proceed estimating VAR(1) model with p=2

```
Estimation results for equation PRECTOT:
========================================
PRECTOT = WS2M.l1 + TS.l1 + PRECTOT.l1 + RH2
M.l1 + T2MDEW.l1 + T2M_MAX.l1 + T2M_MIN.l1 +
T2M.l1 + PS.l1 + T2MWET.l1 + ALLSKY_TOA_SW_D
WN.l1 + ALLSKY_SFC_SW_DWN.l1 + ALLSKY_SFC_LW
_DWN.l1 + WS2M.l2 + TS.l2 + PRECTOT.l2 + RH2
M.l2 + T2MDEW.l2 + T2M_MAX.l2 + T2M_MIN.l2 +
T2M.l2 + PS.l2 + T2MWET.l2 + ALLSKY_TOA_SW_D
WN.l2 + ALLSKY_SFC_SW_DWN.l2 + ALLSKY_SFC_LW
_DWN.l2 + const + trend
```

```
                      Estimate Std. Error t value Pr(>|t|)
WS2M.l1             -0.1527719  0.0491256  -3.110  0.00190 **
TS.l1                0.1425533  0.1114537   1.279  0.20104
PRECTOT.l1           0.2752689  0.0283964   9.694  < 2e-16 ***
RH2M.l1             -0.0562982  0.0228753  -2.461  0.01394 *
T2MDEW.l1           -0.0317577  0.1980003  -0.160  0.87259
T2M_MAX.l1           0.1244869  0.0536593   2.320  0.02045 *
T2M_MIN.l1           0.0182663  0.0547874   0.333  0.73887
T2M.l1              -0.5145864  0.1809029  -2.845  0.00450 **
PS.l1                0.1716499  0.1496709   1.147  0.25159
T2MWET.l1            0.2332984  0.1854624   1.258  0.20857
ALLSKY_TOA_SW_DWN.l1 -0.0115907  0.0131763  -0.880  0.37915
ALLSKY_SFC_SW_DWN.l1  0.0159977  0.0130829   1.223  0.22156
ALLSKY_SFC_LW_DWN.l1  0.2872944  0.0557677   5.152 2.85e-07 ***
WS2M.l2              0.0466316  0.0495522   0.941  0.34679
TS.l2               -0.1635510  0.1132864  -1.444  0.14899
PRECTOT.l2          -0.0792970  0.0275322  -2.880  0.00402 **
RH2M.l2              0.0034075  0.0222369   0.153  0.87823
T2MDEW.l2            0.1160982  0.1980357   0.586  0.55778
T2M_MAX.l2          -0.0192834  0.0528793  -0.365  0.71540
T2M_MIN.l2          -0.0139423  0.0539395  -0.258  0.79606
T2M.l2               0.1489908  0.1808523   0.824  0.41014
PS.l2               -0.1881564  0.1513578  -1.243  0.21398
T2MWET.l2           -0.0976223  0.1825558  -0.535  0.59288
ALLSKY_TOA_SW_DWN.l2 -0.0089699  0.0131550  -0.682  0.49541
ALLSKY_SFC_SW_DWN.l2  0.0109489  0.0130296   0.840  0.40084
ALLSKY_SFC_LW_DWN.l2  0.0171055  0.0552715   0.309  0.75699
trend                0.0002081  0.0001182   1.760  0.07853 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 2.353 on 1888 degrees of freedom
Multiple R-Squared: 0.2863,     Adjusted R-squared: 0.2761
F-statistic: 28.05 on 27 and 1888 DF,  p-value: < 2.2e-16
```

Considering for the model has “trend and constant” the R-squared is 0.20, which means that with this equation just 20% of the precipitation is explained by this equation. Considering that the model has only “trend” it increased for 27%. Considering just the significant values it reduces for 26% for rainfall

```
# Check Perfomance
#Model for forecast
train1<- cbind(x, frequency=365.25, start =c(2015), end=c(2019))
train.test<- cbind(x,frequency=365, start =c(2019), end=c(2020))

PRECTOT1= train1$PRECTOT
trainvar=cbind(train1$WS2M1,train1$TS1,PRECTOT1,train1$RH2M1,train1$T2MDEW1,'
var.TB1=VAR(trainvar, p=4, type="trend")
pred.VAR <- predict(var.TB1,train.test, na.rm = TRUE)
str(pred.VAR$fcst$PRECTOT1)
RMSE.VAR <- sqrt(mean((pred.VAR$fcst$PRECTOT1)^2,na.rm=FALSE))
RMSE.VAR
MAE.VAR<- mean(abs(pred.VAR$fcst$PRECTOT1),na.rm=TRUE)
MAE.VAR
```

```
Estimation results for equation T2M:
====================================
T2M = PRECTOT.l1 + RH2M.l1 + T2M_MAX.l1 + T2M.l1 + PRECTOT.l2 + RH2M.l2 + T2M_MAX.l2 + T2M.l2 + trend

              Estimate Std. Error t value Pr(>|t|)
PRECTOT.l1 -9.381e-02  2.957e-02  -3.172  0.00154 **
RH2M.l1     3.905e-02  7.783e-03   5.018 5.72e-07 ***
T2M_MAX.l1  2.240e-01  3.595e-02   6.232 5.67e-10 ***
T2M.l1      8.071e-01  5.289e-02  15.260  < 2e-16 ***
PRECTOT.l2  1.140e-01  2.785e-02   4.092 4.45e-05 ***
RH2M.l2    -3.661e-02  7.472e-03  -4.899 1.04e-06 ***
T2M_MAX.l2 -2.710e-01  3.528e-02  -7.680 2.52e-14 ***
T2M.l2      2.257e-01  4.952e-02   4.558 5.49e-06 ***
trend       8.189e-05  1.061e-04   0.772  0.44015
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 2.593 on 1906 degrees of freedom
Multiple R-Squared: 0.9697,     Adjusted R-squared: 0.9696
F-statistic:  6780 on 9 and 1906 DF,  p-value: < 2.2e-16
```

Granger Causality

Clive Granger developed the notion of Granger causality: A variable X is said to Granger-cause Y if past values of X can help explain current Y. Thus, Granger causality is

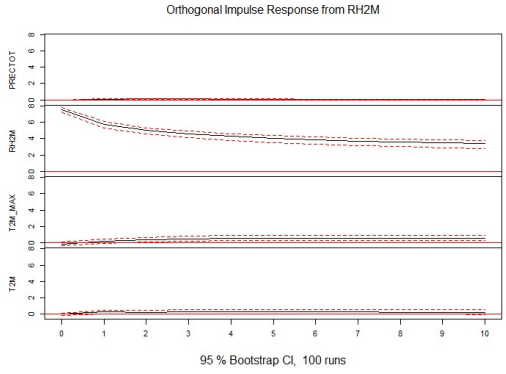
only relevant for time series as this definition is implemented by regressing Y on lagged values of itself and lagged values of X (it is important that error is white noise. The resultsfrom all Granger casuality tests performed in Rstudio for null hypotheses are rejected: H0: One variable does not Granger-cause X. We also note that there is no significant instantaneous causality in the full output of the code.

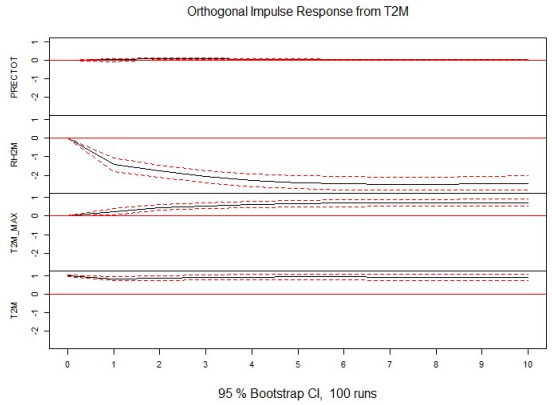
Impulse Response Function

Impulse response function shows the dynamic effects of the error processes (e) on each variable in the VAR system . We can see that the 95% confidence intervals include zero horizontal line making the effect not statistically significant. As we can see on the figures below the variables to not affect PRECTOT. However, we can say that PRECTOT influences the temperature for several days (Figure 13).

Performance Evaluation

In Conclusion VAR method not seems to be a very good method for Mongolia Rainfall Forecast. Both the RMSE= 5.10147 and MAE= 4.632723, for the precipitation increased related with all others models.





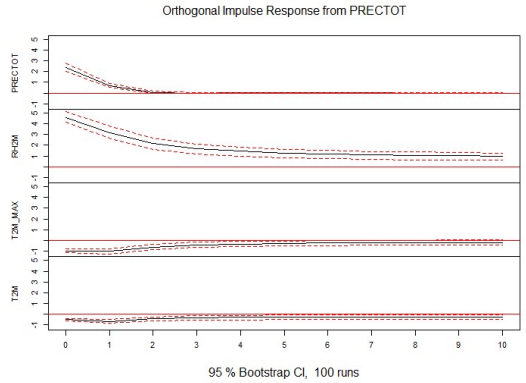


Figure 16. Impulse response for VAR Model

V. Machine Learning Models

Neural Network Model – library (nnetar)

The Artificial Neural Networks (ANN) are a set of algorithms, designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. rtificial ANNs have become very popular, and prediction using ANN is one of the most widely used techniques for rainfall forecasting.

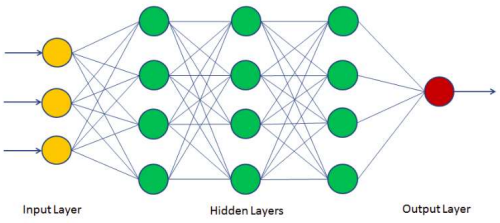


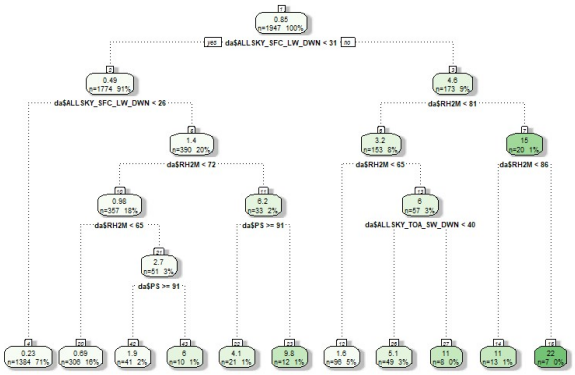
Figure 17. Artificial Neural Network structure Source: https://www.datacamp.com/community/tutorials/neural-network-models-r

To perform the Neural Network forecast I used the nnetar function in the **forecast** package for R, that fits a neural network model to a non-linear time series. Used with big data sets. The NN model is organized in multiples layers, the simplest networks contain no hidden layers and are equivalent to linear regressions. The coefficients attached to these predictors are called "weights". The forecasts are obtained by a linear combination of the inputs.

```
Forecast method: NNAR(17,1,10)[365]

Model Information:

Average of 20 networks, each of which is
a 18-10-1 network with 201 weights
options were - linear output units

Error measures:
                      ME     RMSE       MAE MPE MAPE      MASE       ACF1
Training set -0.01211589 1.422581 0.6547206 NaN  Inf 0.5161361 0.06305773
```

Regression trees

Regression tree for continuous outcome variables, is a simple and popular machine learning algorithm. In contrast with previous linear models it makes no assumptions about the relation between the outcome and predictors. It is the basis of a very powerful method that we will also use in this tutorial, called random forest



We can interpret that as when the downward Thermal Infrared (longwave) is lower than 26 we have lower than 0.23mm of precipitation (71% of the days). In the same way when the downward Thermal Infrared is between 26-31 and the humidity is lower than 65% (16% of the days), predict a wet day of 0.69mm.

```
Regression tree:
rpart(formula = da$PRECTOT ~ da$RH2M + da$T2M + da$T2M_MAX +
    da$WS2M + da$PS + da$TS + da$T2M_MAX + da$T2M_RANGE + da$ALLSKY_TOA_SW_DWN +
    da$ALLSKY_SFC_SW_DWN + da$ALLSKY_SFC_LW_DWN, data = train)

Variables actually used in tree construction:
[1] da$ALLSKY_SFC_LW_DWN da$ALLSKY_TOA_SW_DWN da$PS                da$RH2M

Root node error: 13491/1947 = 6.9291

n= 1947

         CP nsplit rel error  xerror     xstd
1  0.196050      0   1.00000 1.00127 0.179083
2  0.179909      1   0.80395 0.87971 0.169585
3  0.052839      2   0.62404 0.73262 0.107885
4  0.046187      3   0.57120 0.70195 0.112547
5  0.037965      5   0.47883 0.66476 0.110786
6  0.020276      6   0.44086 0.64293 0.109442
7  0.018138      7   0.42059 0.64668 0.109806
8  0.012971      8   0.40245 0.62434 0.095690
9  0.010267      9   0.38948 0.59774 0.093644
10 0.010000     10   0.37921 0.58842 0.093367
```

As you can see, we were able to prune our tree, from the initial 10 splits on 11 variables, to only 4 splits on three variables, gaining simplicity without losing performance (RMSE and MAE are about equivalent in both cases). However, in this model we observed that again both RMSE= 3.511225 and MAE= 1.410056 have increased significantly when compared with the baseline model.

Random Forecast

The Random forecast is a technique of machine learning, that ensemble a learning model for classification, regression and other

tasks. It combines in the same distribution, multiples dependents regression trees dependent of a random vector. However, each tree grown with its own version of the training data.

```
library(randomForest)
set.seed(123)

rf <- randomForest(da$PRECTOT ~ da$RH2M + da$T2M + da$T2M_MAX + da$WS2M +
                   da$PS + da$TS  + da$T2M_MAX  + da$T2M_RANGE + da$ALLSKY_TOA_SW_DWN +
                   da$ALLSKY_SFC_SW_DWN+ da$ALLSKY_SFC_LW_DWN,
                   data = train, importance = TRUE, ntree=1000)
which.min(rf$mse)
imp <- as.data.frame(sort(importance(rf)[,1],decreasing = TRUE),optional = T)
mp <- as.data.frame(sort(importance(rf)[,1],decreasing = TRUE),optional = T)
names(imp) <- "% Inc MSE"
imp
```

The model used 443 trees (Figure 07) and each variables importance could be seen in Figure 08. For this method we can see that 36% of the prediction depends of the relative humidity and 29% in the downward Thermal Infrared. Differently from the single regression tree model, where the downward Thermal Infrared was alone responsible for 85% of the data**.**

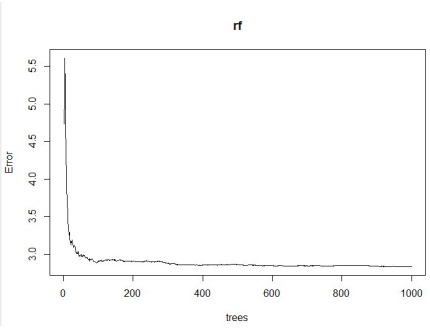


Figure 18. Error per quantity of trees runned in the Random forecast model

	% Inc MSE
da$RH2M	36.272208
da$ALLSKY_SFC_LW_DWN	29.313339
da$ALLSKY_SFC_SW_DWN	18.791845
da$ALLSKY_TOA_SW_DWN	16.685015
da$T2M	16.200720
da$T2M_RANGE	15.516982
da$TS	14.522884
da$T2M_MAX	12.528049
da$WS2M	4.756857
da$PS	3.510637

Figure 19. Variables importance in Random forecast Method

We can see that RMSE= 3.305167 and MAE= 1.414745 slight improved when compared to the decision tree model, However, still not better from the performance of the linear regression model nor base model.

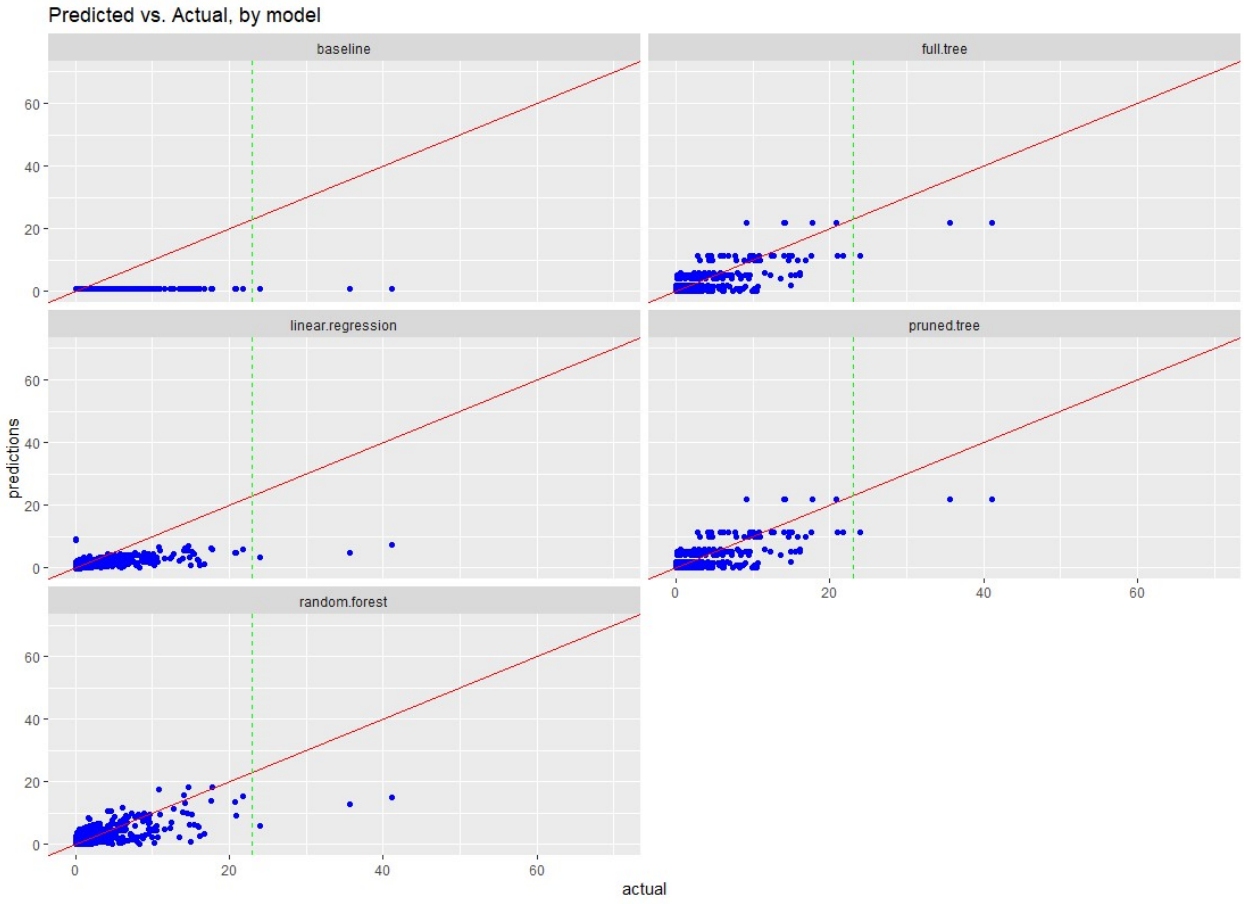
I. Model Evaluation

For evaluate the models the main error metric we will use is the RMSE (root mean squared error) and the MAE (mean absolute error). As a combination of metrics, are often required to assess model performance Chai, T. and Draxler1, R. R. (2014). The RMSE measure gives more weight to larger residuals than smaller ones (a residual is the difference between the predicted and the observed value). This means that it penalizes more the greater the error. The MAE interpret the model performance, by analyzing the magnitude of the errors when compared the model prediction with the actual observed values.

WE can see a summary of RMSE and MAE in the Attachment 02 Arima and Times Series presented the best results.

ATTACH 01

Results from Model related versus actual.



ATTACH

RMSE and MAE Results from the models for Darbarr.

Type of Model	Model	*RMSE*	Rank	*MAE* % of erro	Rank
Autoregressive Regression Models	Base.Model	2.729586	4	1.302	6
Autoregressive Regression Models	Time Series	1.784812	2	0.9783169	2
Autoregressive Regression Models	Times Series with exponential smoothing method	3.373169	7	1.110575	4
Autoregressive Regression Models	Auto ARIMA	1.496647	1	1.098219	3
Multivariate linear Regression	Log-Linear Regression	2.791962	5	1.186184	5
Multivariate linear Regression	Vector autoregression (VAR)	5.10147	9	4.632723	9
Machine Learning	Neural network models	2.22929	3	0.9534483	1
Machine Learning	Tree Model	3.511225	8	1.410056	7
Machine Learning	Random Forecast	3.305167	6	1.414745	8

Entry

Nasa Data .Retrieved from: https://power.larc.nasa.gov/data-access-viewer/

Alexy, V. (2018). USA advanced retail sales 24 month forecasting analysis. *RPubs*. Retrieved from: https://www.rpubs.com/alev2301/421553

Das, H. P, Doblas-Reyes F. J., Garcia, A., Hansen, J., Mariani, L., Nain, A., Ramesh, K., Rathore L. S. & Venkataraman, R. Weather and Climate Forecasts for Agriculture. Agrometeorology. Retrieved from: http://www.agrometeorology.org/files-folder/repository/gamp_chapt4.pdf

Pedro M. (2015). Modelling – predicting the amount of rain. *R-bloggers*. Retrieved from: https://www.r-bloggers.com/part-4a-modelling-predicting-the-amount-of-rain/

Chai, T. and Draxler1, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*. Retrieved from: https://www.geosci-model-dev.net/7/1247/2014/gmd-7-1247-2014.pdf

FAO. *Mongolia.* retrieved from : http://www.fao.org/3/y2722e/y2722e0y.htm