2020 - STUDY

MONGOLIA'S RAINFALL FORECAST USING RSTUDIO

I study to help small farmers in emerging countries fight climate change

Mongolia's Rainfall Forecast using Rstudio

Morais, Manoela N. Moraismn@hotmail.com

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, Farming Technology

I. INTRODUCTION

The agricultural industry is one of the most vulnerable to climate change as it directly depends on rainfall and temperature. Rainfall information is normally used by farmers 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. Although climate change affects differently different crops and regions, it is expected that a decrease in farming productivity (Lobell, 2007). In fact, some decline can already be seen. In the last years, developed countries with a good farm infrastructure had more than 17% yield decrease because of climate changes (Nelson et al, 2014).

Climate variability is a major source of risk for agriculture and food systems and its changes are making it harder for the farmers to guess/predict the weather patterns by just looking for last year's results. The climate change impact is even more accentuated to small-low-income farmers in emerging countries, given their small scale, the burden of economic exclusion, and lack of access to critical resources. Around 500 million smallholder farms in the developing world are supporting almost 2 billion people, and 70 % depending partly or completely on agriculture for their livelihoods (IFAD, 2011). Those farmers do not have adequate networking or tools from which they can learn and easily exchange know-how. Therefore, temperature and rainfall prediction become a significant factor in agriculture to reach the best crop performance.

To this study, I selected the Dornod province, in Mongolia, and an emerging country, which its economy is largely based upon agriculture. The focus of this study is to provide to the small-house-hold-farmers the rainfall forecast with a good prediction and low resources.

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.

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

References: North, Central, South

Figure 2. Dornod Provinces used for weather model study

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 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	kW/m^2/day	Top-of-atmosphere Insolation
ALLSKY_SFC_SW_DWN SRB	kW/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	kW/m^2/day	Downward Thermal Infrared (Longwave) Radiative Flux
T2M_MAX	C (Celsius)	Temperature at 2 Meters

Figure 1. 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 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 03). 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

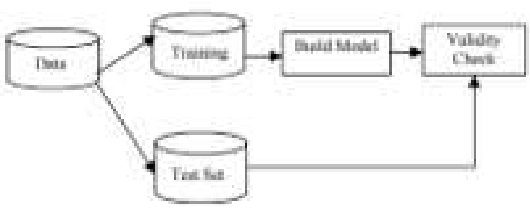


Figure 4. Overview of the forecasting Planning methodology

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.

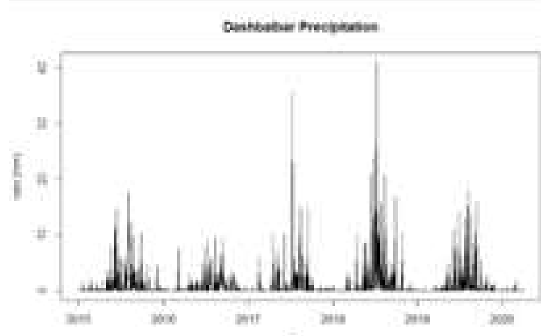


Figure 5.Precipitation for Dashbalbar through the last 5 years. Extracted from RStudio. We can see times series and a seasonal trend.

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: ρ < 0 (stationarity)

When Y has no unit root and stationarity condition 2<ρ<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. <u>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.</u>

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

#TEST PERFORMANCE
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
```

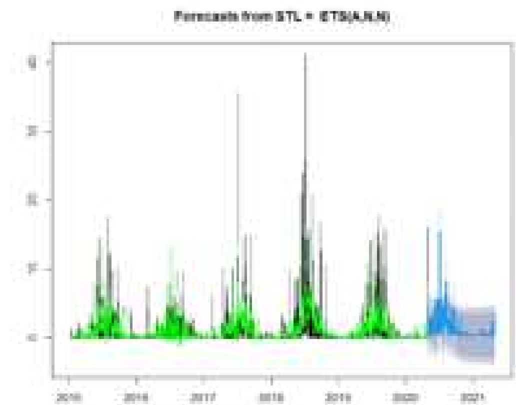


Figure 6. Forecast from Datalkar. PRECTOT for 365 days using default Times series method. Where we can see in green the forecast model in top of the past historical values.

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.

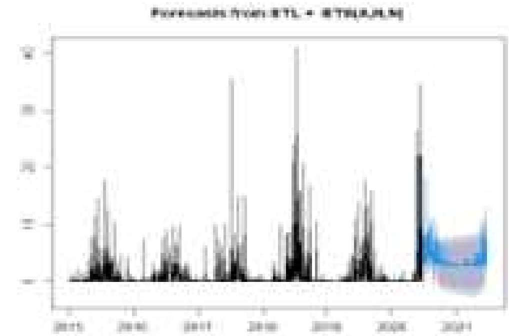
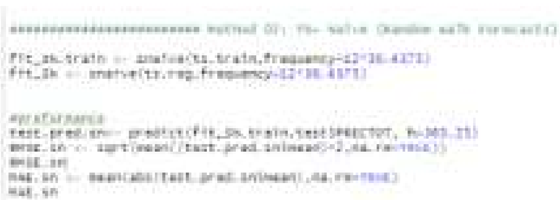


Figure 7.Forecast from Datalkar. PRECTOT for 365 days using default Times series method. After recently shower in the place.

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).

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



```
######################## Method 03: The naive (random walk forecasts)

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

#performance
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
```

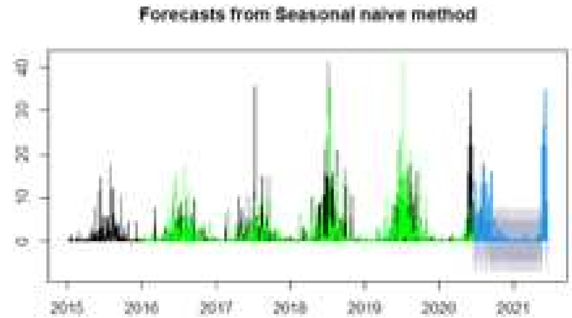


Figure 8. Forecast PRECTOT for 365 days using Times series with exponential smoothing method.

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.

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

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.

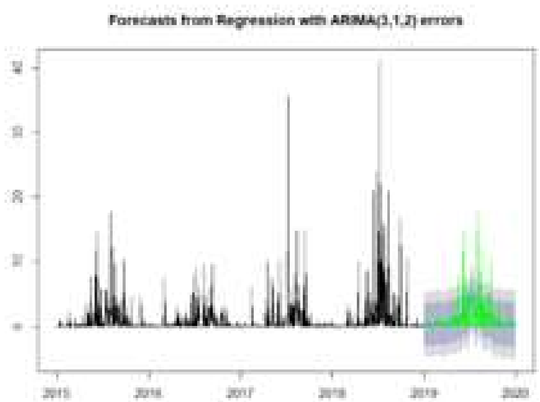


Figure 9.Forecast PRECTOT for 365 days using Times series with auto.arima method. Green the real value and in blue the forecast.

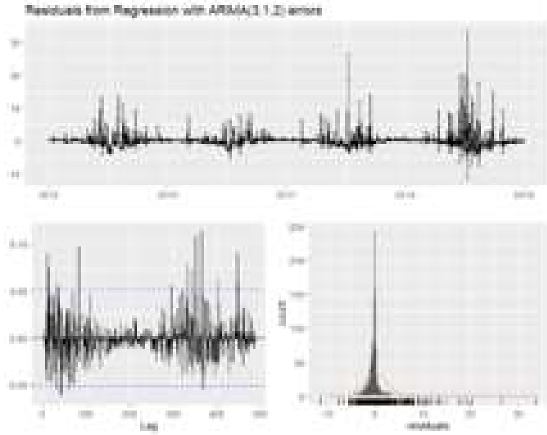


Figure 10. Residual from the forecast of PRECTOT for 365 days using Times series with auto.arima method.

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.

$yi=\beta0+\beta1xi1+\ldots+\beta nxin+\epsilon i$ Equation X

$i=1,2,\ldots,n$, where, $\epsilon i\sim i.i.d.N(0,\sigma 2)$

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.

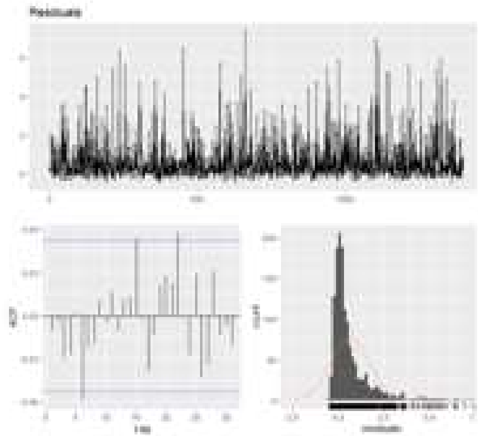


Figure 11. Results(summary) and residuals from the log linear regression for the variable PRECTOT against all others 12 variables.

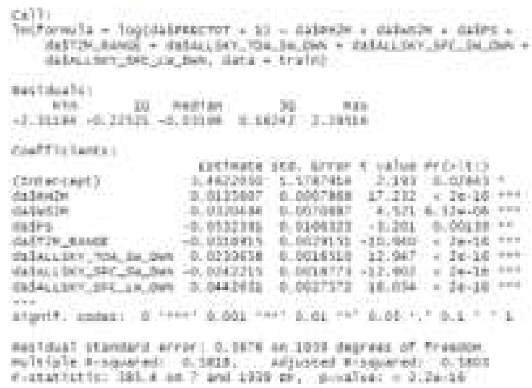


Figure 12. Results(summary) and residuals from the log linear regression for the variable PRECTOT against significant variables.

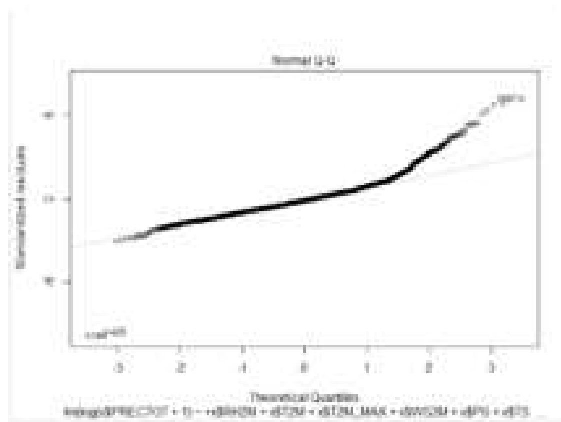


Figure 13. QQ plot of the residuals of the log-linear models for y(PRECTOT) against the 14 variables.

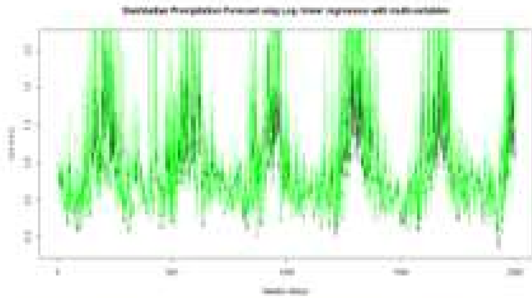


Figure 14. Graph pf PRECTOT (Y) Log- linear regression against the 14 multivariable. In green, the forecast and in black past historical rainfall measures.

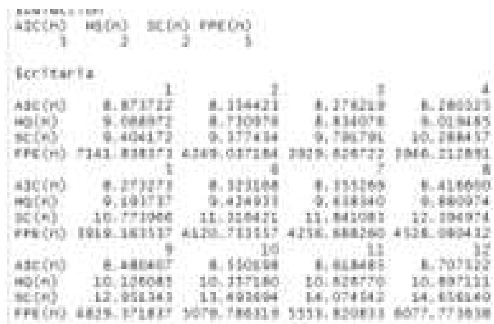
Figure 15.Result of the Log- linear regression after include an interaction term between precipitation and include only significant variable.

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)

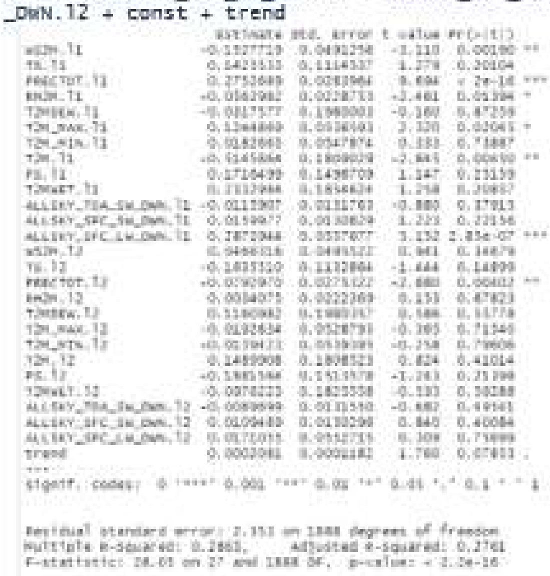
VAR is an AR(p) in vector form where *Yt* " is a vector of several variables describing the dynamic system with variables for Rainfall attributes. A vector autoregression is a system of equations, where each equation for each variable contains lagged values of itself and lagged values of all other variables 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)):



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
```



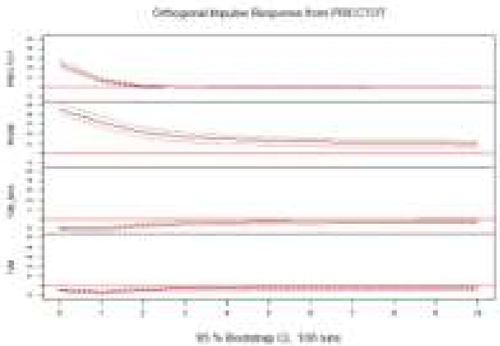


Figure 16. Impulse response for VAR Model

V. Machine Learning Models

<u>Neural Network Model – library (nnetar)</u>

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.

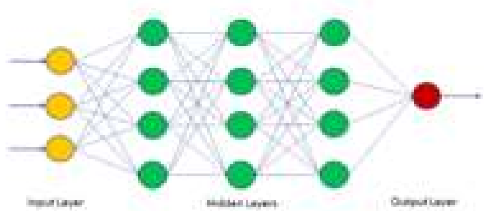
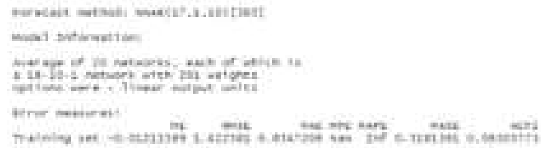


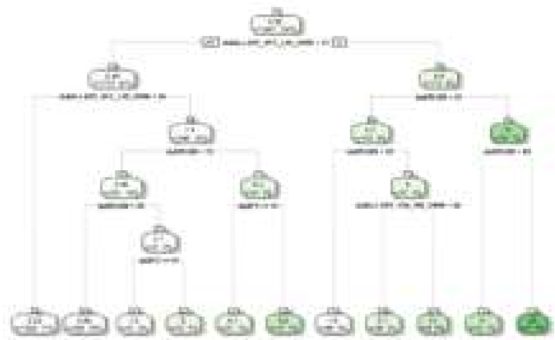
Figure 17. Artificial Neural Network structure Source: https://www.datacamp.com/community/tutorials/neural-network-models-r

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.



<u>Regression trees</u>

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.

…. For more please e-mail moraismn@hotmail.com....

ATTACH 02

RMSE and MAE Results from the models for Darbarr. *(need to add for other places)*

Type of Model	Model	RMSE	Rank	MAE % of err	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	3.335167	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. retrieved from: http://www.fao.org/3/y2722e/y2722e0y.htm

Lobell D.B., Field C.B. Global scale climate–crop yield relationships and the impacts of recent warming. Environ. Res. Lett. 2007;2:014002. doi: 10.1088/1748-9326/2/1/014002. Retrieved from: https://iopscience.iop.org/article/10.1088/1748-9326/2/1/014002

IFAD. 2011. Rural poverty report 2011. New realities, new challenges: new opportunities for tomorrow's génération (available at http://www.ifad.org/rpr2011/report/e/rpr2011.pdf).

Nelson, G., van der Mensbrugghe, D., Ahammad, H., Blanc, E., Calvin, K., Hasegawa, T., Havlik, P., Heyhoe, E., Kyle, P., Lotze-Campen, H., von Lampe, M., Mason d'Croz, D., van Meijl, H., Müller, C., Reilly, J., Robertson, R., Sands, R., Schmitz, C., Tabeau, A., Takahashi, K., Valin, H. & Willenbockel, D. 2014b. Agriculture and climate change in global scenarios: why don't the models agree? Agricultural Economics. 45(1): 85–101.