**MACHINE LEARNING FOR DATA SCIENCE**

**Instructions:**

Build a HMM model using any one of the pollutant measurements as response variable. Based on your result what can you say about the air quality of the period? Provide a brief explanation.

**Solution:**

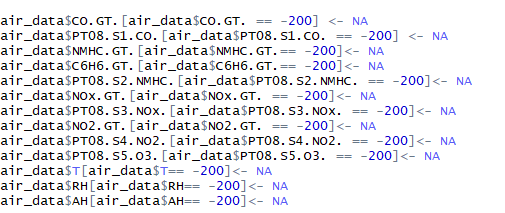
Steps to be followed:

1. Load the dataset into the R studio

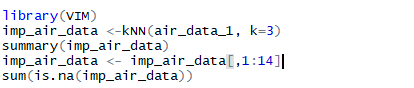
2. After looking into the dataset, we can notice that there are missing values.

3. The attribute “NMHC.GT.” has 90% of missing values., so this column can be dropped.

4. Reassign all the value -200 to NA and handle the missing values.



5. We can impute the missing values using the kNN missing value imputation techniques.



6. Now that we have handled the missing values, we can merge the date and time columns to create the HMM model and convert it into the DATE format.

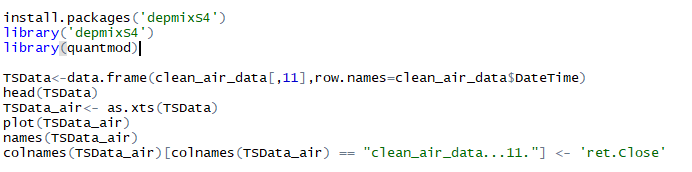


7. The above code merges 2 columns into one DateTime column with the Date (POSIXIt format)

8. There are only two missing value in the DateTime column, so they can be removed using the complete.cases()

9. Now that we have our cleaned dataset, let us create the Hidden Markov’s model.

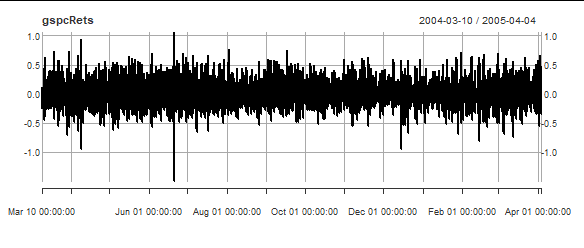
10. In order to do this, we need to create our data into the time series dataset.



11. Once the dataset has been converted into time series object, we will pick the response

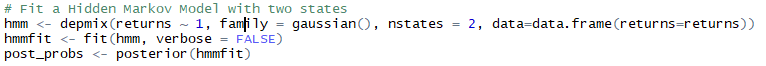
Variable to be analyzed over the time

12. The response variable here is the Relative Humidity “RH” to analyse the air quality over the time period.



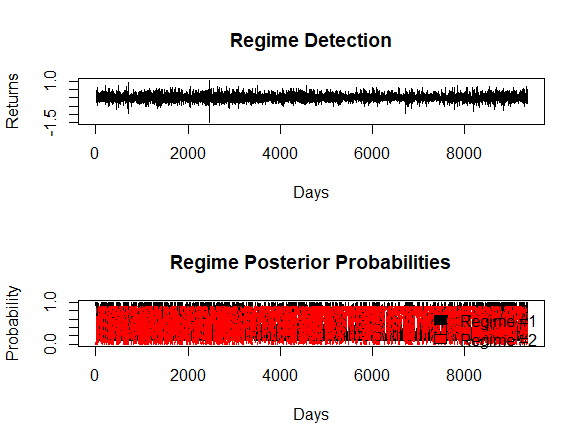
13.The above plot describes the relative humidity over the time period from March 2004 to April 2005

Lets us now fit the Hidden Markov’s model with 2 states



14. After getting the posterior states of the model, we can plot the return stream and the

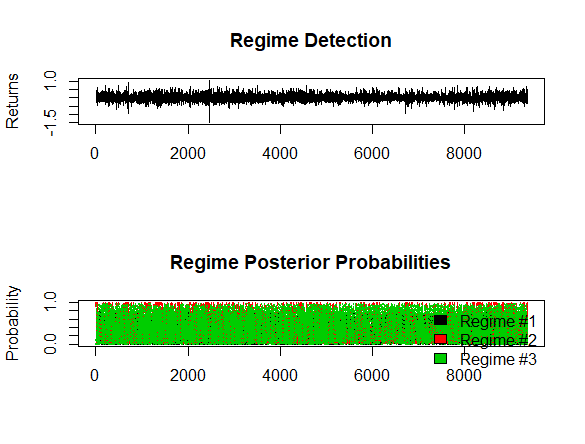
posterior probabilities of the separate regimens with respect to their days.



15. We can see from the above plot that there are clear fluctuations in the probabilities and

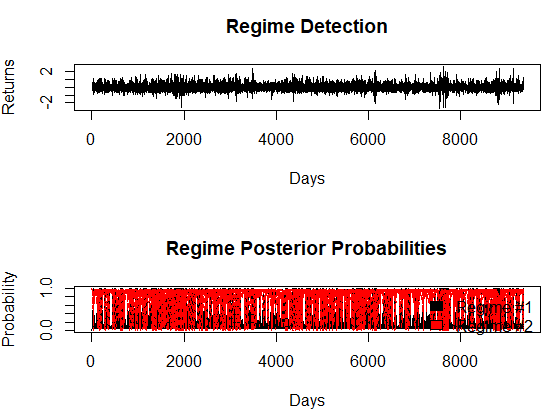
Returns

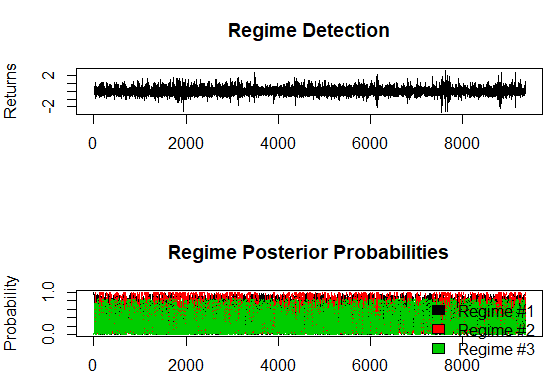
16. Let us try to create the model with 3 states and we can see the plot for 3 regimes



17. The same type of analysis has been followed with the 2 more attributes. The attribute “CO.GT”

and “C6H6.GT.”





The exposure to pollution sources is strongly significant only when the pollution episodes Occur. The model predicts a serious pollution episode only when exceedances are observed in locations that are exposed to low pollution sources. The estimated transition probabilities of the latent chain, varying with weather conditions with the respect to the relative humidity.

We can observe that a regime of acceptable air quality (state 2) is persistent during the whole year, as

indicated by the large probabilities of remaining in state 2. As a result, the probabilities of moving from state 2 to a different state are generally low. The high variability of the probabilities of remaining in state 1 and in state 3 confirm that pollution episodes are not persistent during the period of interest.