VECTOR AUTO-REGRESSION (VAR)

- **Autoregression** is a time series model that uses observations from previous time steps (called as the variable's lags) as input to a regression equation to predict the value at the next time step. It is a very simple idea that can result in accurate forecasts on a range of time series problems.
- A univariate autoregression is a single-equation, single-variable linear model in which the current value of a variable is explained by its own lagged values.
- A **VAR** is a K-equation, K-variable linear model in which each variable is in turn explained by its own lagged values, plus current and past values of the remaining K - 1 variables. This simple framework provides a systematic way to capture rich dynamics in multiple time series.
- The equation can be expressed as:

$$y_t = \nu + A_1 y_{t-1} + \ldots + A_p y_{t-p} + u_t,$$

where *v* is a constant K–vector for the intercept and all A_j , j = 1, . . . , p are K ×K–matrices, while u_t denotes a multivariate **white noise process**. In this scheme, any of the component variables y_t^k, k =

1. . . K depends on p lags of itself and of the other K − 1 component variables.

- A white noise process is one with a mean zero and no correlation between its values at different times. This correlation between a value and it's lags is called **autocorrelation.** A white noise process should also have an identical variance matrix Σ(u) which doesn't vary with time.
- In data description and forecasting, VARs have proven to be powerful and reliable tools that are now, rightly, in everyday use.
- To use the VAR model, the data usually has to be first pre-processed to an appropriate form and should be carefully analysed. The subsequent sections talk briefly about the theory behind these steps. Then, the EXPERIMENTS section displays how these steps are put into practice.

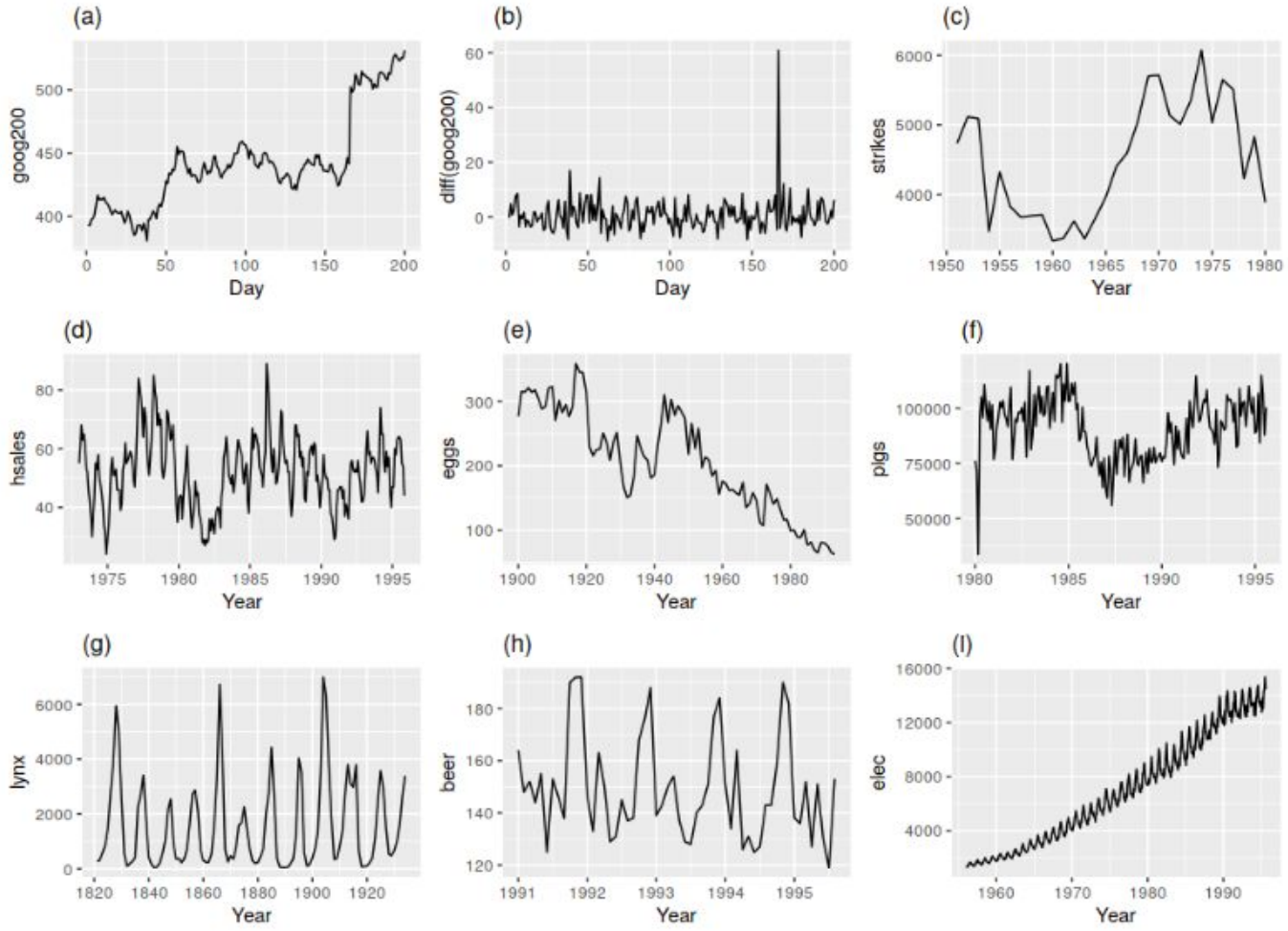
STATIONARITY

- Let X be the matrix denoting the dataset. Normally while modelling the time series' using VAR, or any other **OLS (Ordinary Least Square) Estimator,** the following assumptions are made about the model:
    - All parameters are linear in the modelled equations

- All dependent variables are non-stochastic in nature
- The mean of all error terms are 0. That is, E(u(t) | X) = 0 for all t from 0,1…….T.
- The variance of error terms are constant. They don't change with time. var(u(t)| X) = $\boldsymbol{\sigma * \sigma}$
- No auto-correlation among the error terms
- No multicollinearity (no linear correlation among the dependent variables)
- No specification bias (No superfluous variables, no core variables excluded from the model, no measurement errors, no outliers etc.)

- **GAUSS-MARKOV THEOREM**: It states that given the above assumptions, the ordinary least squares (OLS) estimator is **BLUE(Best Linear Unbiased Estimator)**.
  - Best means it has the lowest sampling variance within the class of linear unbiased estimators.
  - Unbiased means as no. of times sampling is done reaches infinity, the average of the error approaches the true error.

- **STATIONARY TIME SERIES:** Intuitively, a stationary time series is one whose properties (like mean, variance, autocorrelation etc.) do not depend on the time at which the series is observed. Thus, time series with trends, or with seasonality, are not stationary — the trend and seasonality will affect the value of the time series at different times. On the other hand, a white noise series is stationary — it does not matter when you observe it, it should look much the same at any point in time.



The above figure illustrates the concept of stationarity. Which of the above figures are stationary?

Obvious seasonality rules out series (d), (h) and (i). Trends and changing levels rules out series (a), (c), (e), (f) and (i). Increasing variance also rules out (i). That leaves only (b) and (g) as stationary series. At first glance, the strong cycles in series (g) might appear to make it non-stationary. But these cycles are aperiodic. In the long-term, the timing of these cycles is not predictable. Hence the series is stationary.

- **FORMAL DEFINITION OF STATIONARITY:**
  - **Strong stationarity:** A stochastic process whose unconditional joint probability distribution does not change when shifted in time.

$$F_X(x_{t_1+\tau}, \ldots, x_{t_n+\tau}) = F_X(x_{t_1}, \ldots, x_{t_n}) \quad \text{for all } \tau, t_1, \ldots, t_n \in \mathbb{R} \text{ and for all } n \in \mathbb{N} \qquad \text{(Eq.1)}$$

    Here, x_t represents a random variable generating the stochastic process at time t and F_X denotes the cumulative probability distribution of the joint probability.

    Consequently, parameters such as mean and variance also do not change over time.

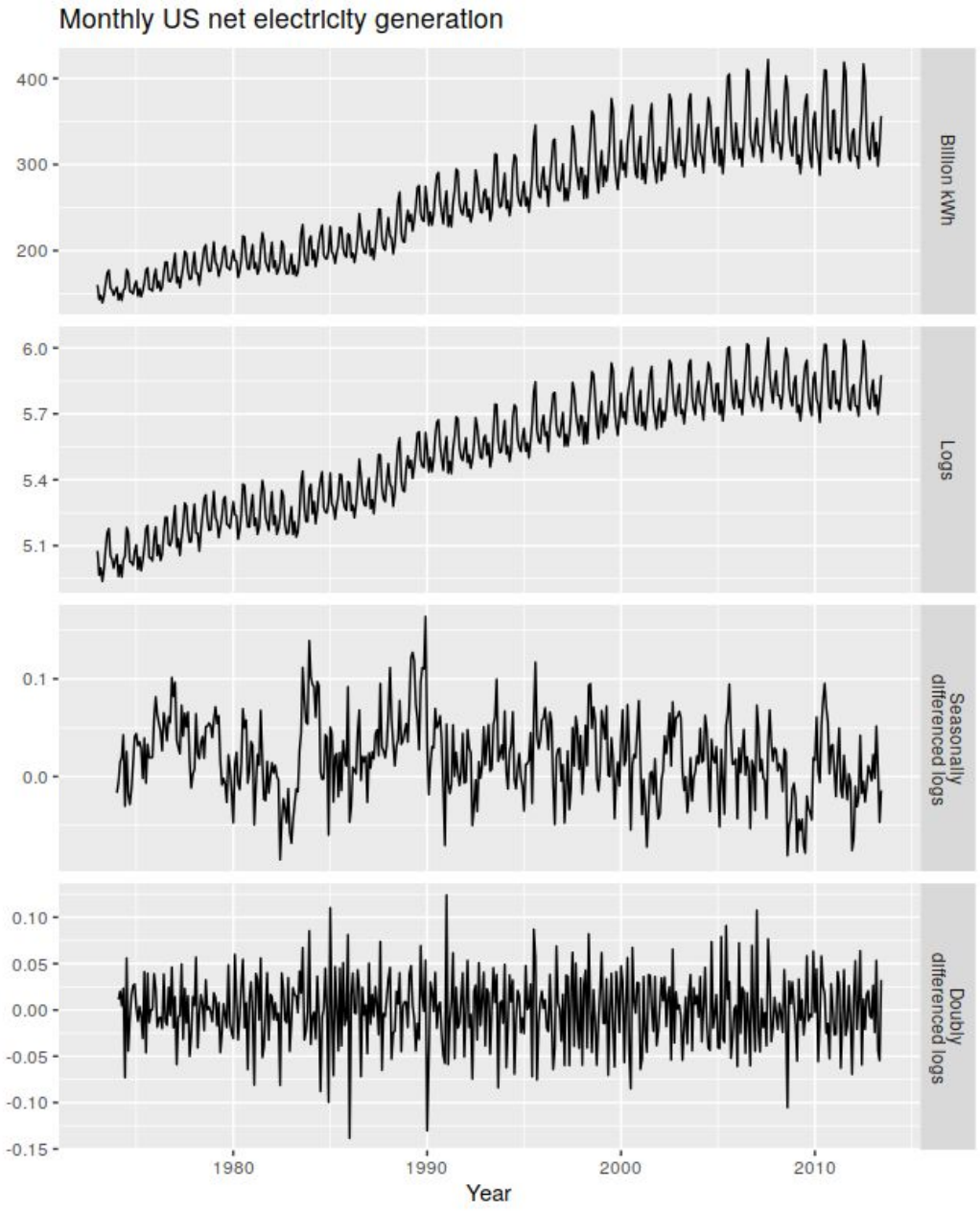
  - **Weak stationarity:** Weak stationarity only requires the shift-invariance (in time) of the first moment and the cross moment (the auto-covariance). This means the process has the same mean at all time points, and that the covariance between

the values at any two time points, t and t−k, depend only on k, the difference between the two times, and not on the location of the points along the time axis. Note that this directly implies that the variance of the process is also constant (by taking k = 0).

- In order for Gauss- Markov Theorem to apply, ensuring weak stationarity of the time series data is sufficient. Also, weak stationary processes are easier to analyze, model and investigate. Due to these properties, weak stationarity has become a common assumption for many practices and tools in time series analysis, including VAR.
- From now on, we will refer to a weak stationary process as just a stationary process.
- **HOW TO DEAL WITH NON-STATIONARY DATA:** If the data has:
    - **Linear Trends:** Compute the differences between consecutive observations. This is known as **differencing**. This would ensure that mean remains constant as in a linear graph, the rate of change in value always remains constant.
    - **Unstable variance:** Transformations such as logarithms or square root can help to stabilise the variance of a time series.
    - **Non-Linear Trends:** The differenced data will not appear to be stationary and it may be necessary to difference the data multiple times to obtain a stationary series.

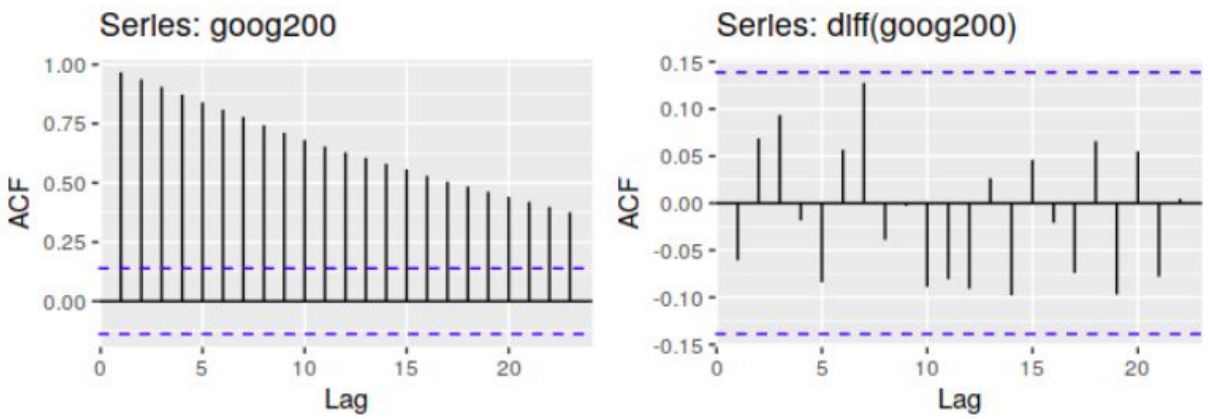
- **Seasonal Trends:** A seasonal difference is the difference between an observation and the previous observation from the same season. We can do a seasonal difference:

$y'_t = y_t - y_{t-m},$ where m denotes the no. of seasons.



- The above picture is an example of converting a non-stationary time series to a stationary one. The original data (1st picture) is first log transformed (2nd picture), seasonally differenced (3rd picture) and then normally differenced (4th picture).

- As well as looking at the time plot of the data, the **Auto-Correlation Function (ACF)** plot (A plot of autocorrelation vs. the no. of lags) is also useful for identifying non-stationary time series. For a stationary time series, the ACF will drop to zero relatively quickly, while the ACF of non-stationary data decreases slowly as shown in the following figure.



The left figure is the ACF plot of a non-stationary data, while the right one is that of a stationary data.

The ACF plot which doesn't decay to zero has more chances of being non-stationary, because if there is an increase of a value at a

particular time t, the effect would propagate further down the series and have a cumulative effect on later values in the time series, thus leading to a change in the mean values.

- **UNIT ROOT TESTS FOR STATIONARITY:** Though graphical visualization is a good way to check for stationarity, it is subjective in nature. One way to determine more objectively whether a time series is stationary is to use a unit root test. These are statistical hypothesis tests of stationarity.
    - **DICKEY-FULLER TEST:** Let a simple model be

      $y_t = \rho y_{t-1} + u_t$ where u_t is the error term. A unit root is present if ρ=1. The model would be non-stationary in this case, as in that case magnitude of y_t will continue to rise with t, thus violating the condition that the mean will remain constant with time.

      The regression model can be written as

      $$\Delta y_t = (\rho - 1) y_{t-1} + u_t = \delta y_{t-1} + u_t$$

      .

      Taking $\delta \equiv \rho - 1$, the test has three versions that differ in the model of unit root process they test for.

        - Test for a unit root ($\delta = 0$): $\Delta y_i = \delta y_{i-1} + u_i$

- Test for a unit root(δ) with drift: $\Delta y_i = a_0 + \delta y_{i-1} + u_i$
- Test for a unit root(δ) with drift and deterministic time trend:

  $\Delta y_i = a_0 + a_1 * t + \delta y_{i-1} + u_i$

- **AUGMENTED DICKEY-FULLER (ADF) TEST:** The ADF test expands the Dickey-Fuller test equation to include high order regressive process in the model.

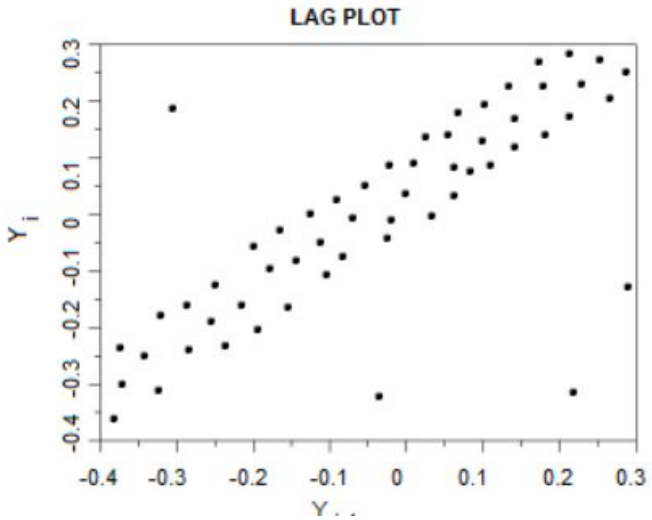
$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t$$

  where α is a constant, β is the coefficient on a time trend and p the lag order of the autoregressive process. We can create the 3 situations just as in the normal Dickey Fuller Test shown above, by adjusting values of α and β

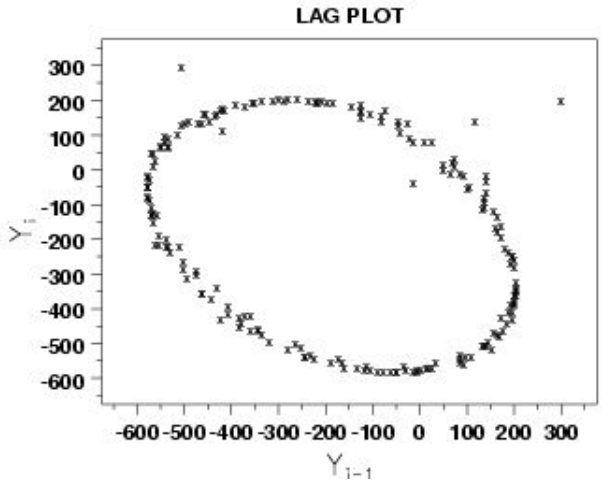
- The above 2 tests are fundamentally statistical significance tests. That means, there is a hypothesis testing involved with a null hypothesis (that unit root exists) and an alternate hypothesis. A test statistic is computed and p-values get reported. Smaller the p-value, the more confident we are that a particular time series is stationary.

LAG PLOTS

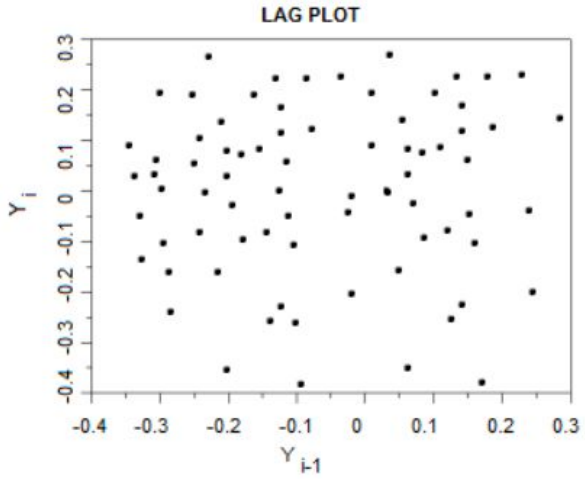
- A lag plot is a special type of scatter plot with a variable and it's lag, typically a lag of 1. They are very useful in making some design choices about what model to use for the time series data.
- The shape of the lag plot can provide clues about the underlying structure of the data. For example:
    - A linear shape to the plot suggests that an autoregressive model like VAR is probably a better choice as a linear shape implies good autocorrelation among the data. (See the below figure).



- An elliptical plot suggests that the data comes from a single-cycle sinusoidal model. (See the below figure).



- Creating a lag plot also enables us to check for randomness. Random data will spread fairly evenly both horizontally and vertically. If you cannot see a pattern in the graph, the data is most probably random and hence not suitable for any time series modelling. (See the below figure).



MULTICOLLINEARITY

- One of the assumptions in the Gauss-Markov theorem is multicollinearity: that all variables in the model are not correlated to each other.
- Though perfect multicollinearity is not possible in practice, it should still be ensured that it is not high. An acceptable collinearity between 2 variables is generally 0.7 or less.
- Collinearity in the dataset can be checked by calculating the correlation matrix of the dataset. If multicollinearity is present, it is desirable that feature selection or feature extraction (like PCA) is done.
- If this is not done, the model that VAR fits would still have high accuracy. But, the model may not be reliable enough when dealing with new data, because in the regression equations, the coefficients calculated for the redundant variables may be small and insignificant.
- Also, if multicollinearity exists, we may not be able to make a sound analysis about the way the variables affect each other.
- Also, feature extraction is anyhow desirable, especially for high dimensional data, to account for the curse of dimensionality.

CHOOSING THE LAG ORDER

- To choose the lag orders of the VAR model, estimators are used which estimate out-of-sample prediction error and thereby relative quality of statistical models for a given set of data. They provide means for model selection.
- These estimators have to deal with both overfitting and underfitting of the model. Some estimators used in common are:
    - **AKAIKE INFORMATION CRITERIA (AIC):** AIC estimates the relative amount of information lost by a given model. It's value is given by: $\mathrm{AIC} = 2k - 2\ln(\hat{L})$ where k is the number of estimated parameters in the model and L^ is the maximum value of the likelihood function for the model.
    - **BAYESIAN INFORMATION CRITERION (BIC):** It is similar to the formula for AIC, but with a different penalty. With AIC the penalty is 2k, whereas with BIC the penalty is ln(n) k, where n is the no. of data points.
    - **HANNAN–QUINN INFORMATION CRITERION (HQC):**

      $$\mathrm{HQC} = -2L_{max} + 2k\ln(\ln(n))$$

    - **FINAL PREDICTION ERROR(FPE):** Final Prediction Error (FPE) criterion provides a measure of model quality by simulating the situation where the models being compared are tested on a

different data set. It is given by:

$$FPE = \det\left(\frac{1}{N}\sum_{1}^{N} e(t,\hat{\theta}_N)\left(e(t,\hat{\theta}_N)\right)^T\right)\left(\frac{1+d/N}{1-d/N}\right)$$ where:

N is the number of values in the estimation data set.

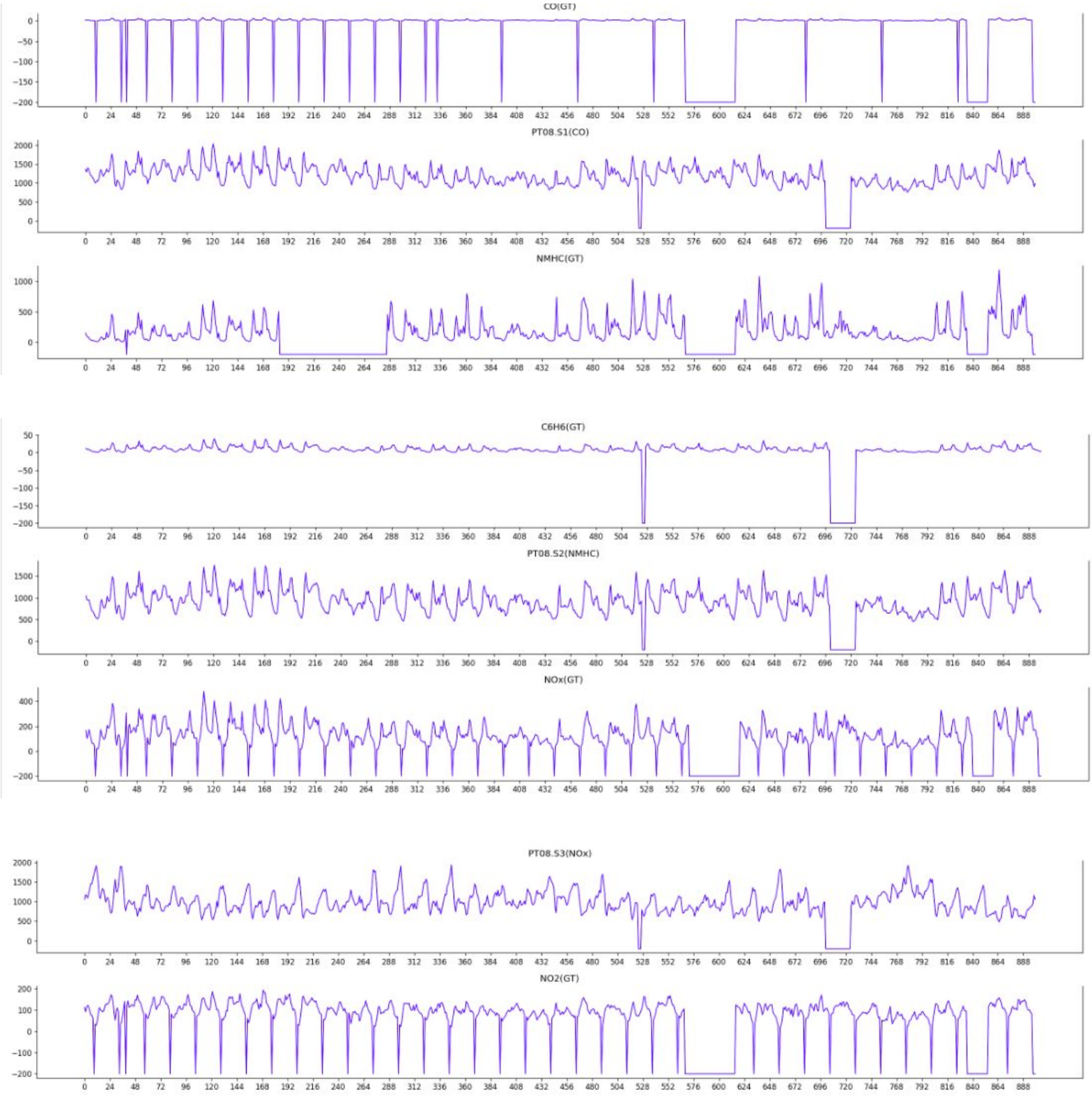
e(*t*) is a *ny*-by-1 vector of prediction errors.

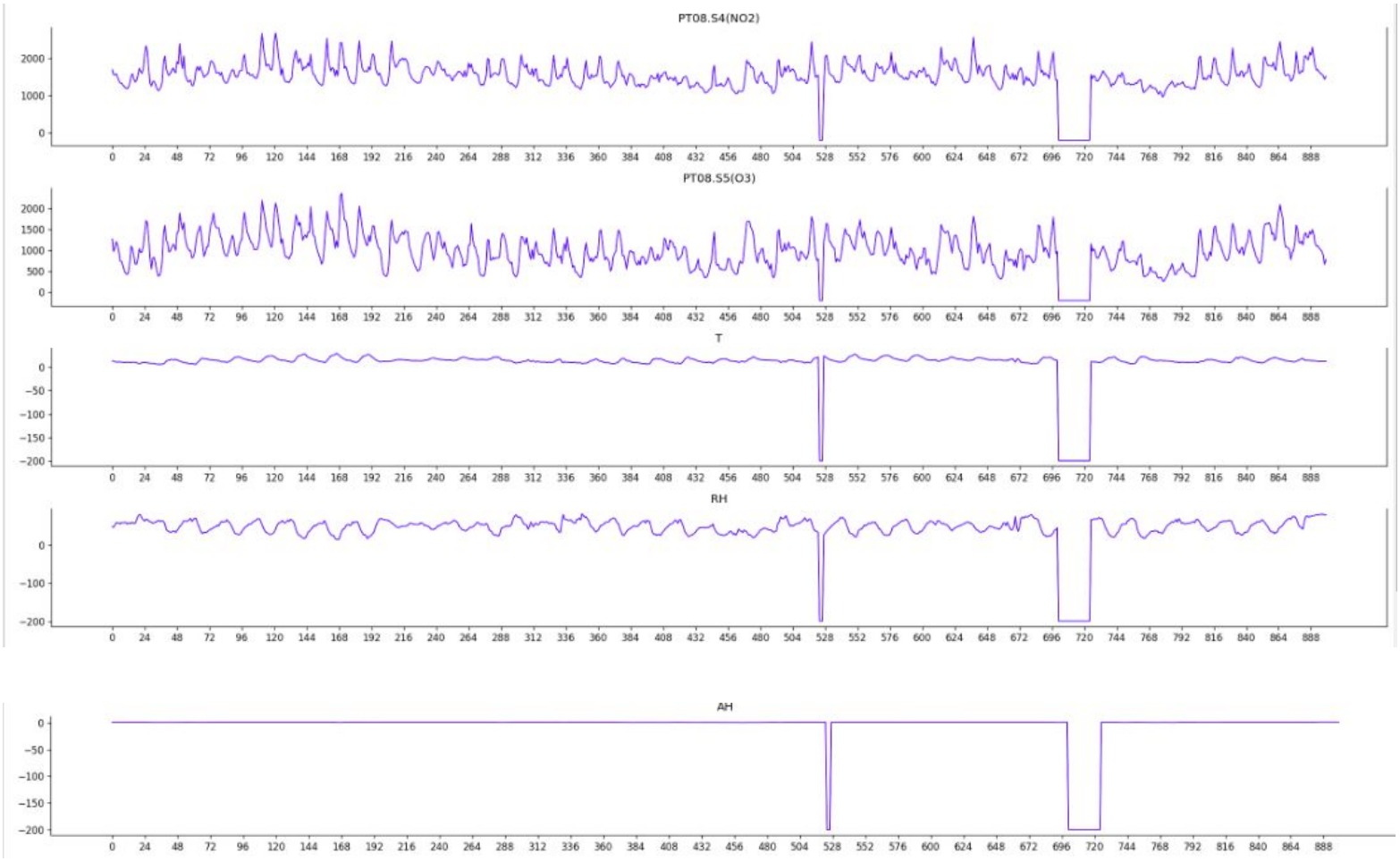
θ_N represents the estimated parameters.

d is the number of estimated parameters.

EXPERIMENTS

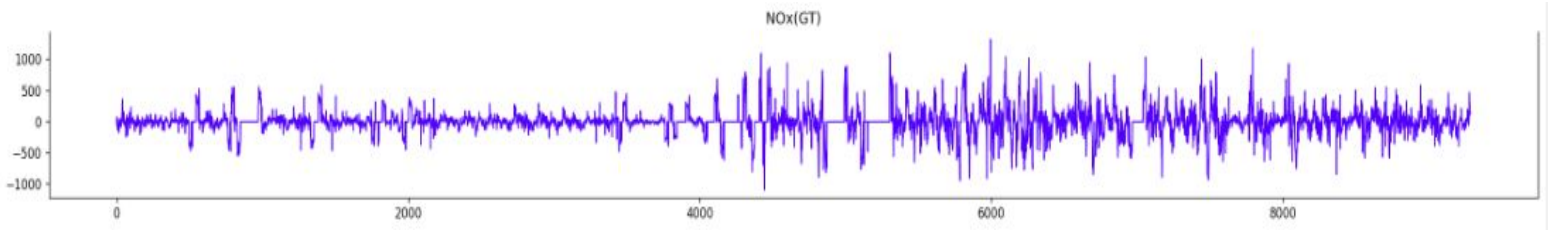
- We have tried to use a VAR model on a dataset of Air Quality to model the data and to forecast future values . The dataset has hourly averaged responses of 13 atmospheric metrics. The dataset is available in the same folder as this report, by the name AirQualityUCI.csv.
- The experiments described in this section can be run interactively on the python notebook VAR.ipynb, which is also present in the same folder.
- First, the graphs of each time series are visualized as shown below.


CO(GT)
PT08.S1(CO)
NMHC(GT)
C6H6(GT)
PT08.S2(NMHC)
NOx(GT)
PT08.S3(NOx)
NO2(GT)

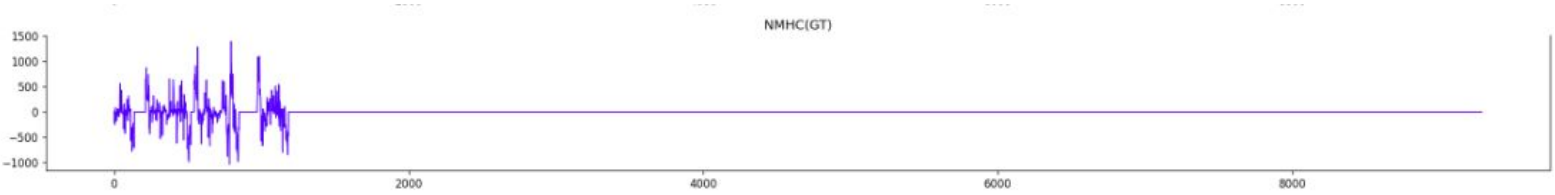


- It can be observed that almost all time series are seasonal in nature and have a time period of 24 hours. Also, it can be observed that the time period of CO(GT)'s graph (the first graph) initially was 24, but then after a point of time becomes 72 hours.
- Therefore accordingly, CO(GT)'s graph was seasonally differenced with a period of 72 hours, while the rest of the graphs were differenced with a period of 24 hours.

- The resultant plots were almost stationary. The only exceptions were NMHC(GT) and NOx(GT) where it is clear that the variance of the data is not constant as shown below:

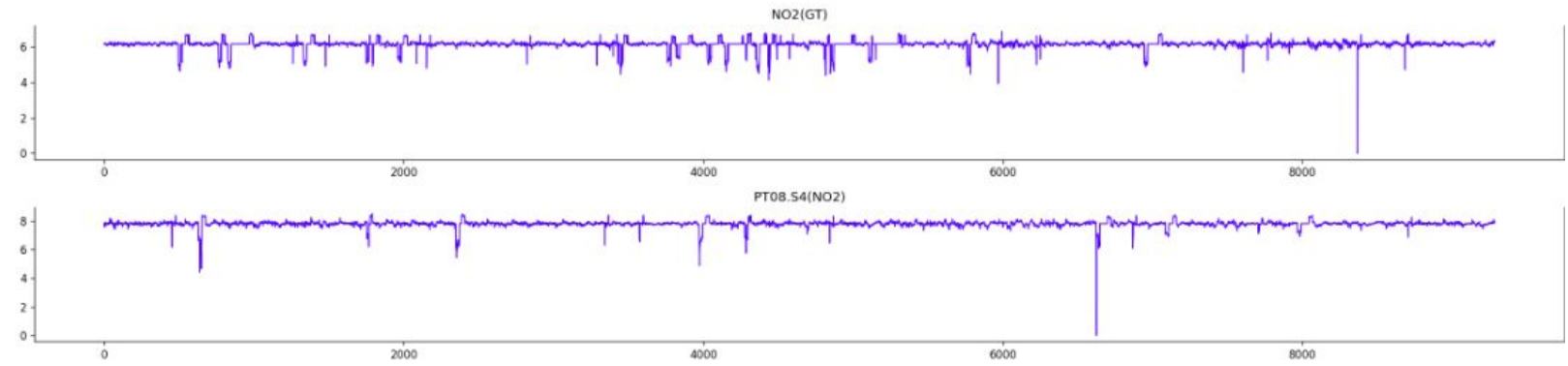


(The variance increases from small to big)

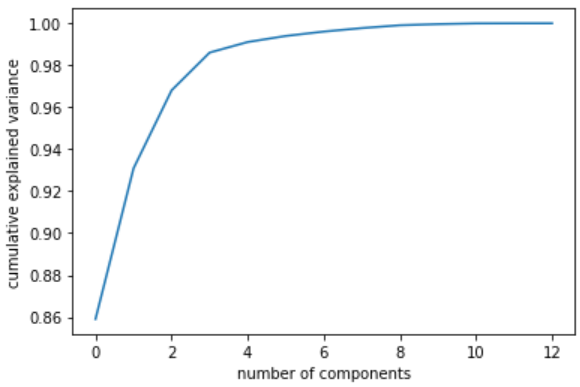


(The variance decreases from big to small)

- So, each time series was offset by a certain amount so that all the readings were greater than 1 and then were made to undergo log transformations, to stabilize the variance.
- The resulting time series were all almost stationary. But, some time series had outliers like the ones in the below figure. These outliers were removed by filtering out from each time series all the data points which lied outside [μ - 3σ, μ + 3σ]. (Assuming a normal distribution, we would retain 99% of the data after this filtering)



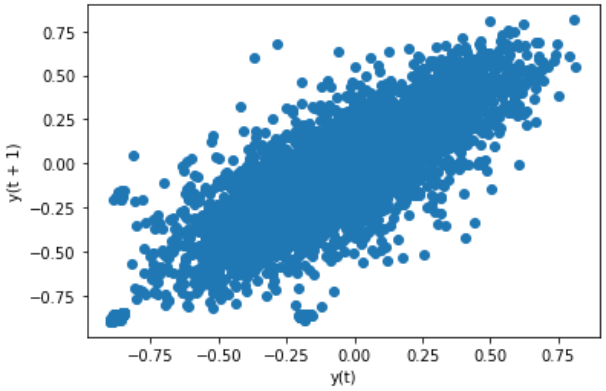
- Then the correlation matrix was calculated of all the time series. We found that PT08.S1(CO), PT08.S2(NMHC), PT08.S4(NO2) and PT08.S5(O3) were highly correlated with each other. The same for C6H6(GT), AH, RH and T as well. So, there was a need for feature extraction to make our model more reliable.
- PCA was used for the feature extraction. Given below is the variance ratio plot of PCA:



It is clear from the above figure that the plot saturates after 3 dimensions. So, the whole dataset was reduced to 3 dimensions.

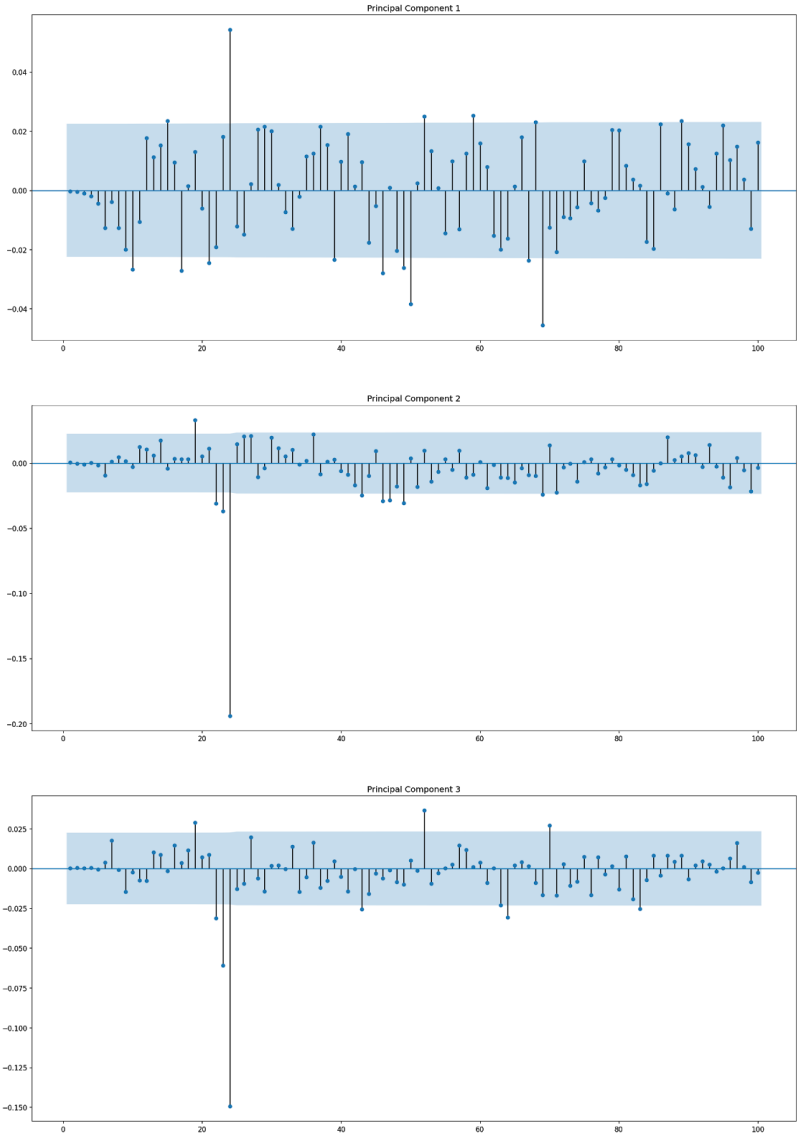
- The outliers were removed again as described above and then the lag plot for 1 lag was plotted. The plot clearly had a linear trend and was

not random. This gave the confidence that VAR would be a suitable choice to model the data.



- Also, ACF's were plotted for each time series. Though, they did not decay as fast as would have been expected, a compensation was that the autocorrelations were not insignificant giving confidence that a reliable VAR model could be built.
- We apply the ADF test to each time series. The null hypothesis was rejected for all the time series, which meant none of the time series had any unit roots, and thus could be confidently declared to be stationary.
- The lag order was selected by calculating the AIC, BIC, FPE and HQIC for the model. According to AIC, the lag order turned out to be 6, according to BIC, it turned out to be 2, according to FPE 5, and according to HQIC 2. We experimented on all the time lags and obtained best results on lag order 6.

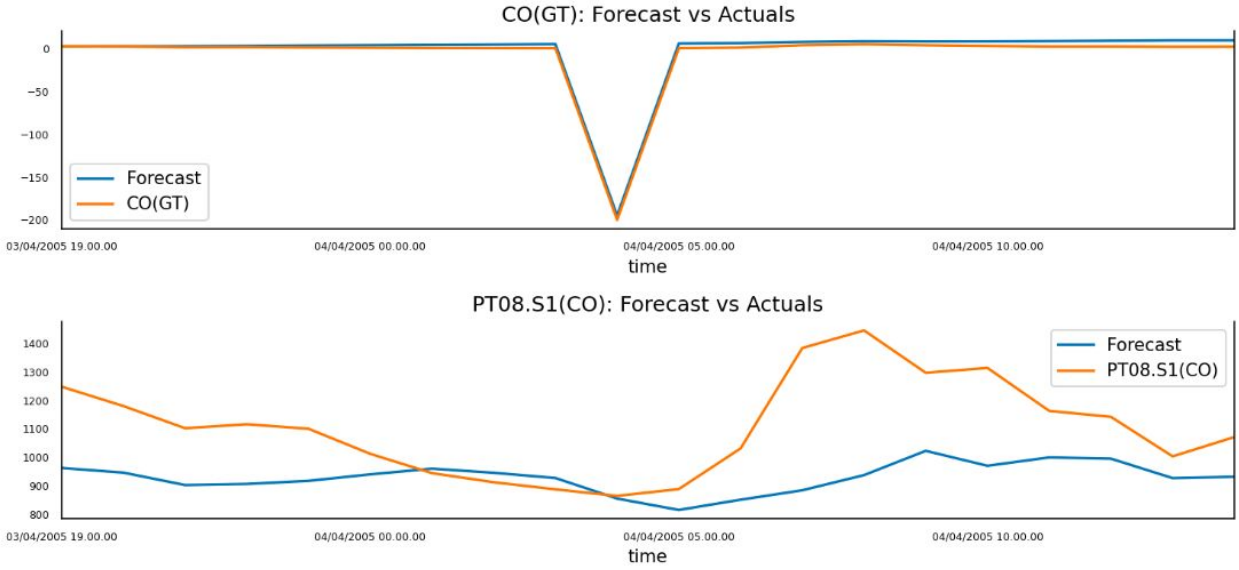
- The model summary showed that none of the coefficients were insignificant. This gave the confidence that the model was reliable.
- The ACF of the residuals were plotted (See the figure below). Ideally, the autocorrelations should have been insignificant, but spikes at lag numbers 24, 48 and 72 were observed. This was not entirely surprising given that the time period of the original time series was 24 or 72.

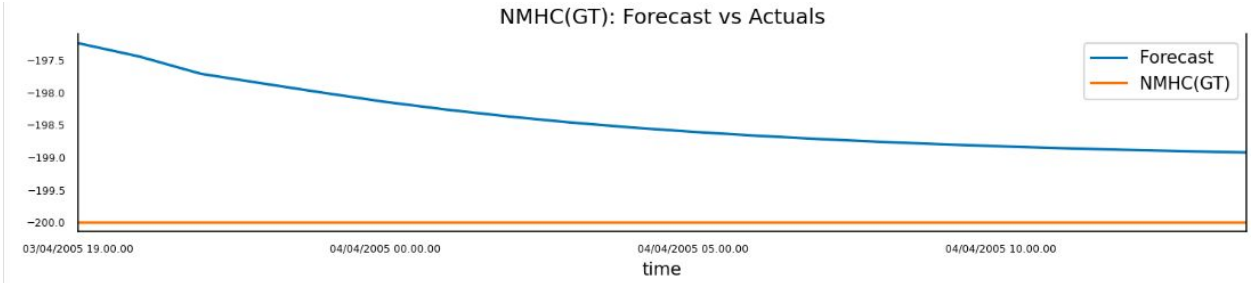


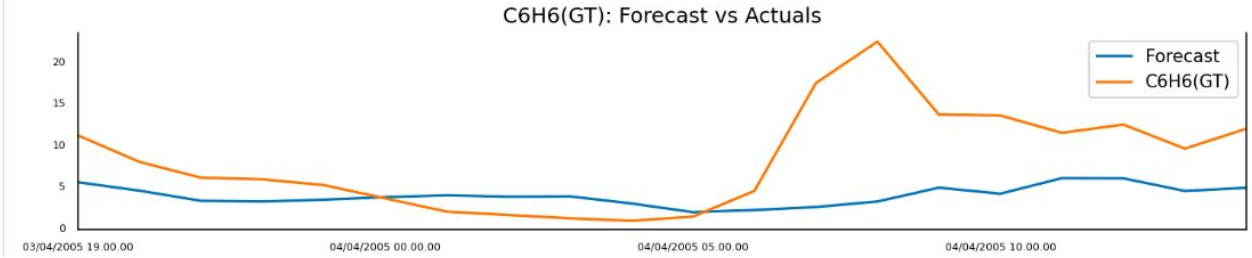
- Such spikes demonstrate the deficiencies of a VAR model, which may not account for all the autocorrelations in the data. A more dynamic model is necessary to deal with such issues, which take the error terms also as additional parameters in the regression.

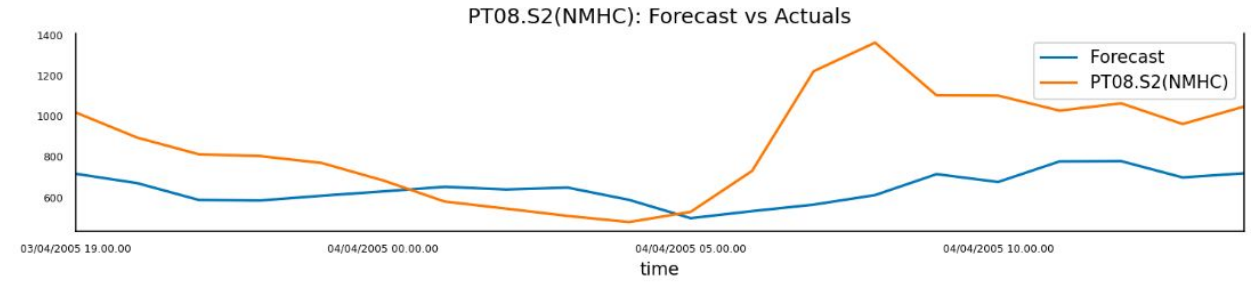
RESULTS

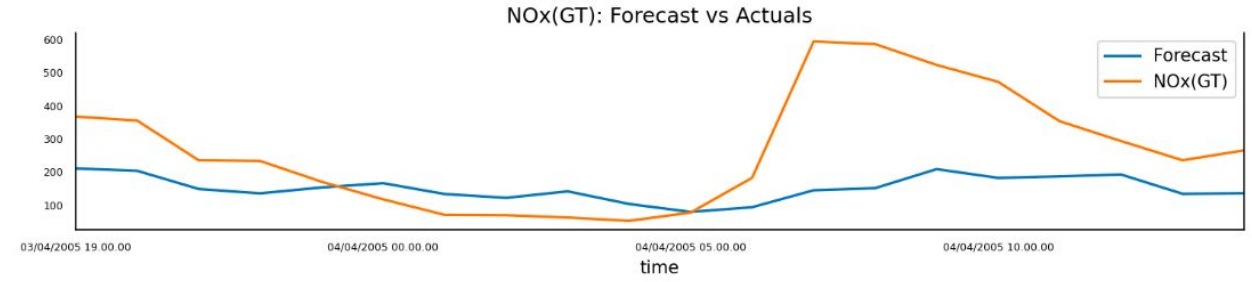
- The model is used to forecast all the time series for 20 hours.
- The forecasted values were transformed back to the original space (reversing the PCA feature extraction, log transformation and seasonal differencing) and then compared with the ground truth. The results were plotted, as shown in the below figures:

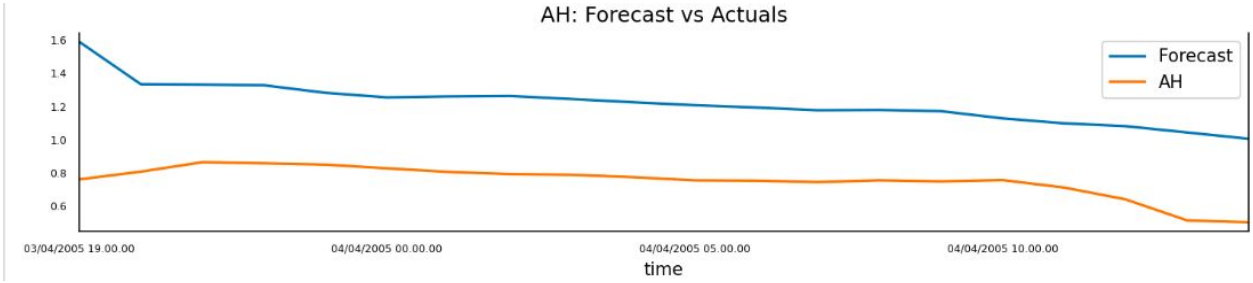


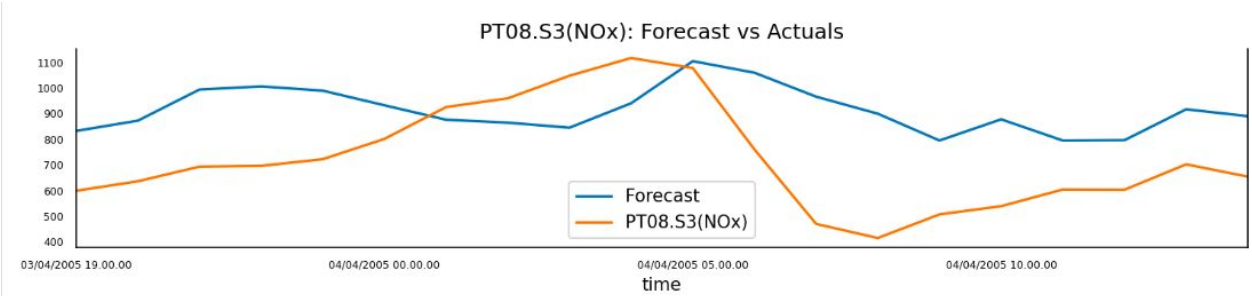

NMHC(GT): Forecast vs Actuals
Forecast
NMHC(GT)
−197.5
−198.0
−198.5
−199.0
−199.5
−200.0
03/04/2005 19.00.00
04/04/2005 00.00.00
04/04/2005 05.00.00
04/04/2005 10.00.00
time

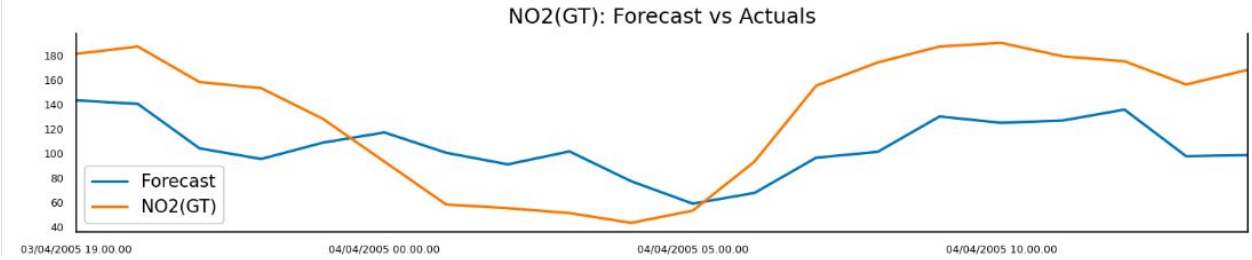

C6H6(GT): Forecast vs Actuals
Forecast
C6H6(GT)
20
15
10
5
0
03/04/2005 19.00.00
04/04/2005 00.00.00
04/04/2005 05.00.00
04/04/2005 10.00.00

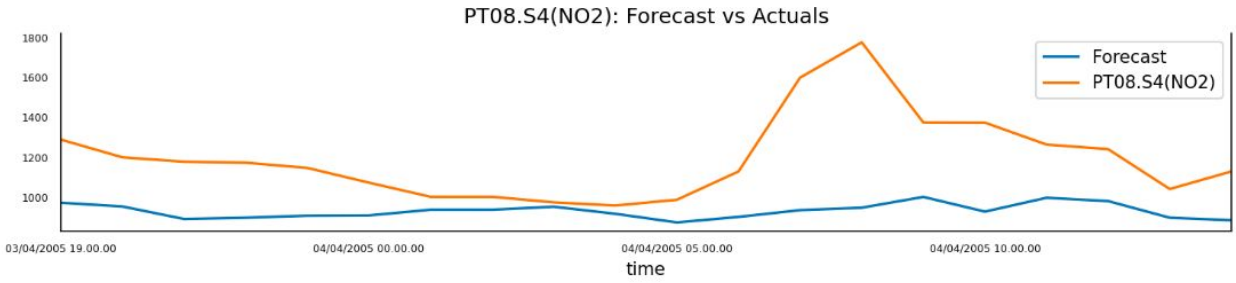

PT08.S2(NMHC): Forecast vs Actuals
Forecast
PT08.S2(NMHC)
1400
1200
1000
800
600
03/04/2005 19.00.00
04/04/2005 00.00.00
04/04/2005 05.00.00
04/04/2005 10.00.00
time

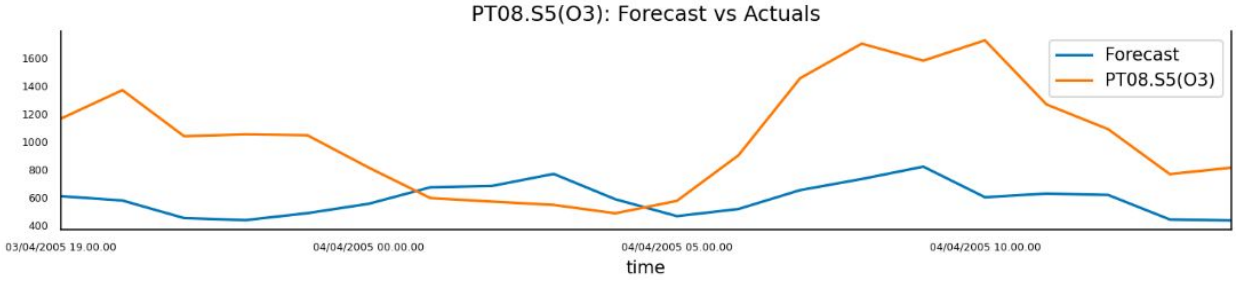

NOx(GT): Forecast vs Actuals
Forecast
NOx(GT)
600
500
400
300
200
100
03/04/2005 19.00.00
04/04/2005 00.00.00
04/04/2005 05.00.00
04/04/2005 10.00.00
time

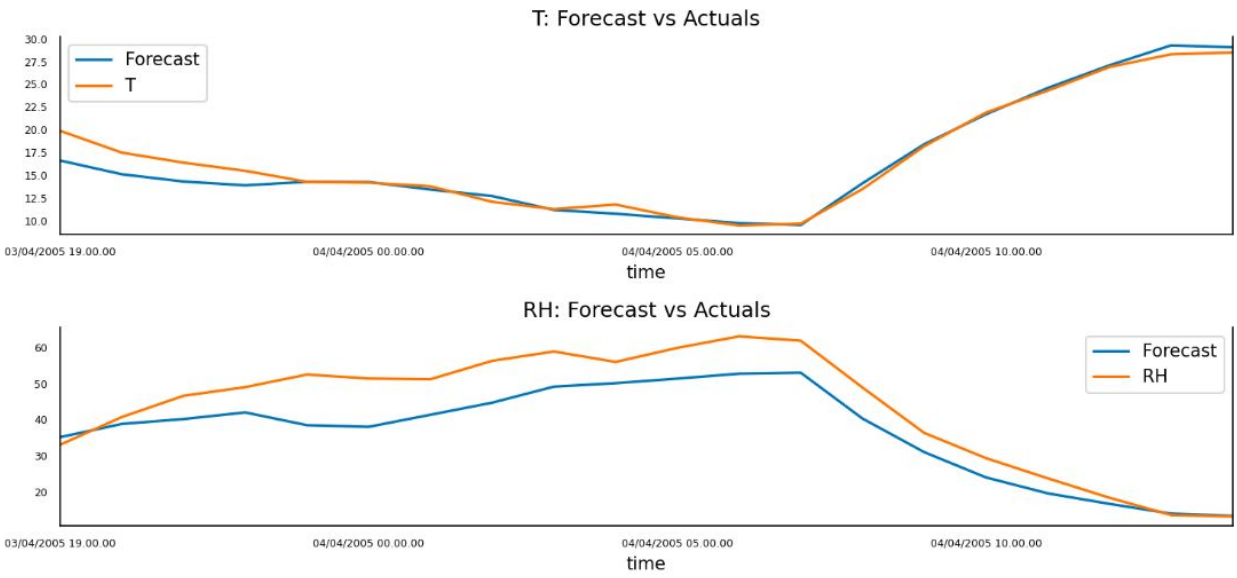

AH: Forecast vs Actuals
Forecast
AH
1.6
1.4
1.2
1.0
0.8
0.6
03/04/2005 19.00.00
04/04/2005 00.00.00
04/04/2005 05.00.00
04/04/2005 10.00.00
time


PT08.S3(NOx): Forecast vs Actuals
1100
1000
900
800
700
600
500
400
Forecast
PT08.S3(NOx)
03/04/2005 19.00.00
04/04/2005 00.00.00
04/04/2005 05.00.00
04/04/2005 10.00.00
time


NO2(GT): Forecast vs Actuals
180
160
140
120
100
80
60
40
Forecast
NO2(GT)
03/04/2005 19.00.00
04/04/2005 00.00.00
04/04/2005 05.00.00
04/04/2005 10.00.00


PT08.S4(NO2): Forecast vs Actuals
1800
1600
1400
1200
1000
Forecast
PT08.S4(NO2)
03/04/2005 19.00.00
04/04/2005 00.00.00
04/04/2005 05.00.00
04/04/2005 10.00.00
time


PT08.S5(O3): Forecast vs Actuals
1600
1400
1200
1000
800
600
400
Forecast
PT08.S5(O3)
03/04/2005 19.00.00
04/04/2005 00.00.00
04/04/2005 05.00.00
04/04/2005 10.00.00
time



- The plots do seem satisfactory. Do remember that this forecast is under the assumption that we don't get any data for the 20 hours forecasted. But in real life, we continue to get data hourly, thus we can keep updating our model accordingly with the extra information, and can thus produce even better results than above.
- The corresponding result metrics are as follows:

```
Forecast Accuracy of: CO(GT)

mape   :  3.3287

me     :  4.0911

mae    :  4.1116

mpe    :  3.3185

rmse   :  4.6294

corr   :  0.9988
```

minmax : 0.6005

Forecast Accuracy of: PT08.S1(CO)

mape   : 0.1532

me   : -175.7069

mae   : 184.6075

mpe   : -0.1434

rmse   : 232.0626

corr   : 0.4151

minmax : 0.1531

Forecast Accuracy of: NMHC(GT)

mape   : 0.008

me   : 1.6003

mae   : 1.6003

mpe   : -0.008

rmse   : 1.6755

corr   : nan

minmax : -0.0081

Forecast Accuracy of: C6H6(GT)

mape   : 0.7965

me   : -4.2835

mae   : 5.2443

mpe   : 0.0019

rmse   : 7.0633

corr   : 0.3168

minmax : 0.5408

Forecast Accuracy of: PT08.S2(NMHC)

mape   : 0.2724

me   : -217.4309

mae   : 259.0307

mpe   : -0.1921

rmse   : 316.5663

corr   : 0.4187

minmax : 0.2653

Forecast Accuracy of: NOx(GT)

mape   : 0.5523

me   : -114.5714

mae   : 144.2068

mpe   : -0.1164

rmse   : 191.1628

corr   : 0.5803

minmax : 0.4484

Forecast Accuracy of: PT08.S3(NOx)

mape   : 0.3905

me   : 185.936

mae   : 238.181

mpe   : 0.3401

rmse   : 265.4882

corr   : 0.3144

minmax : 0.2581

Forecast Accuracy of: NO2(GT)

mape   : 0.3943

me   : -26.2104

mae   : 45.4034

mpe   : -0.0422

rmse   : 48.5786

corr   : 0.6913

minmax : 0.321

Forecast Accuracy of: PT08.S4(NO2)

mape   : 0.2031

me   : -264.8875

mae   : 264.8875

mpe   : -0.2031

rmse   : 329.3925

corr   : 0.3852

minmax : 0.2031

Forecast Accuracy of: PT08.S5(O3)

mape   : 0.4336

me   : -440.9066

mae   : 492.199

mpe   : -0.3399

rmse   : 573.6261

corr   : 0.2839

minmax : 0.4236

Forecast Accuracy of: T

mape   : 0.0451

me   : -0.3683

mae   : 0.7531

mpe   : -0.0233

rmse : 1.1543

corr : 0.9846

minmax : 0.0447

Forecast Accuracy of: RH

mape : 0.1429

me : -6.515

mae : 6.7817

mpe : -0.1324

rmse : 7.906

corr : 0.9764

minmax : 0.1427

Forecast Accuracy of: AH

mape : 0.6407

me : 0.4693

mae : 0.4693

mpe : 0.6407

rmse : 0.478

corr : 0.687

minmax : 0.3847

CONCLUSION

- This report has given a theoretical introduction to VAR modelling, and has demonstrated its practical application on a 13 variable time series dataset. The results demonstrate the power of VAR modelling.
- But one of its shortcomings has also been highlighted where all the auto-correlation of the variables could not be accounted for by the VAR model demonstrating the need for a more dynamic version of VAR, where the past error terms would also be considered in the modelling.
- Another major criticism of VAR is that it does not say anything about the underlying structure of the variables. We would sometimes not only want to forecast our predictions, but also want to find causal relations among the variables. An extension of the VAR model, called Structural VAR (SVAR) would be necessary for such cases.