Time Series Analysis of Output of Continuous Flow Process

Name : Aditi Ganesh Joshi

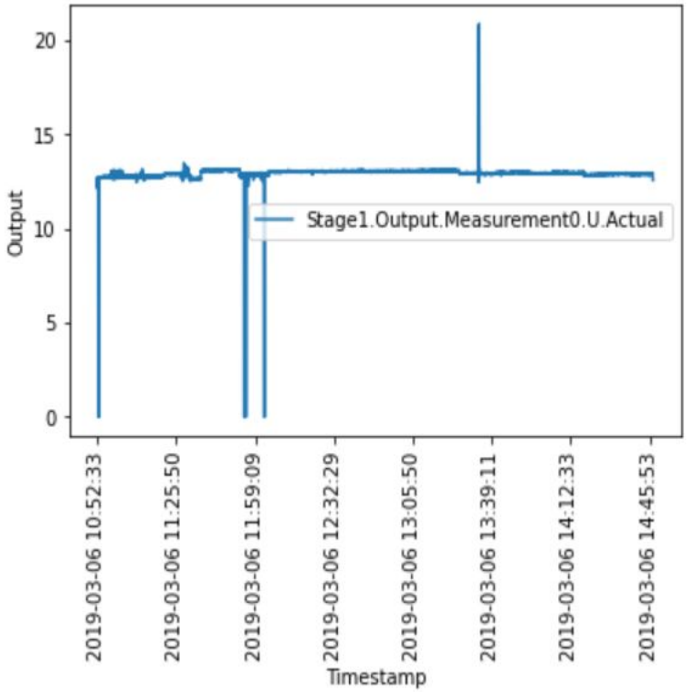
Roll No. : 180020010

Problem Statement

- The aim is to predict output of a continuous flow process from a live plant in Detroit Michigan.
- Three machines operate in parallel and their outputs are combined to give the output of stage 1. This is the dependent variable.
- The output is a **univariate time series**.
- There are 41 input/independent variables that influence this output. These are the **exogenous variables**. They consist of different process variables and system parameters.
- The size of the dataset is **14008 x 43**
- Hence, this is a **multivariate regression problem**.

Data Preparation and Visualization

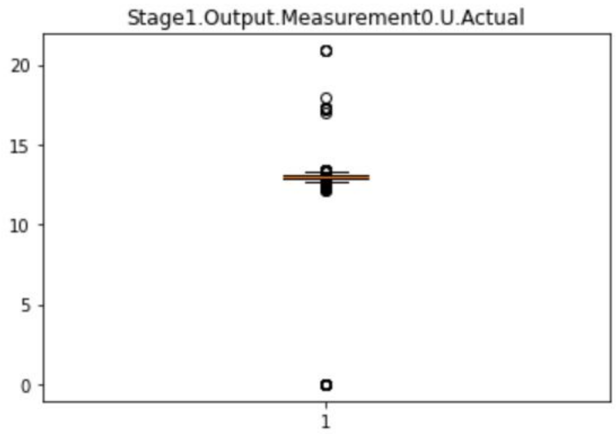
The time-series is mean-centered.



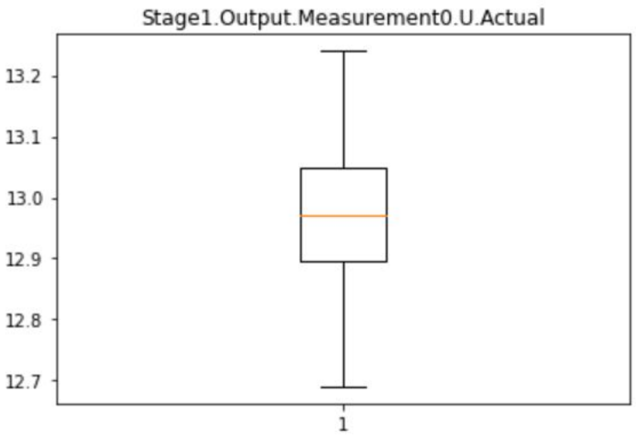
Removal of Outliers

Points beyond 1.3*(interquartile range) are replaced with the mean.

Before

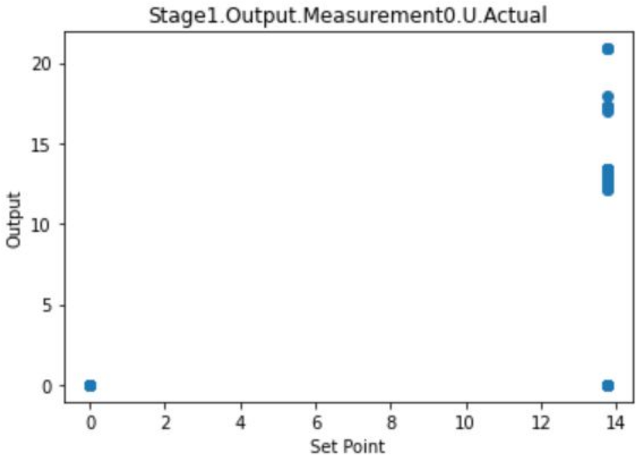


After



Identification of Feedback Loop

The output doesn’t depend on the setpoint. The points where setpoint and output both are zero are exceptions and very less in number. Hence, no feedback loop is present.



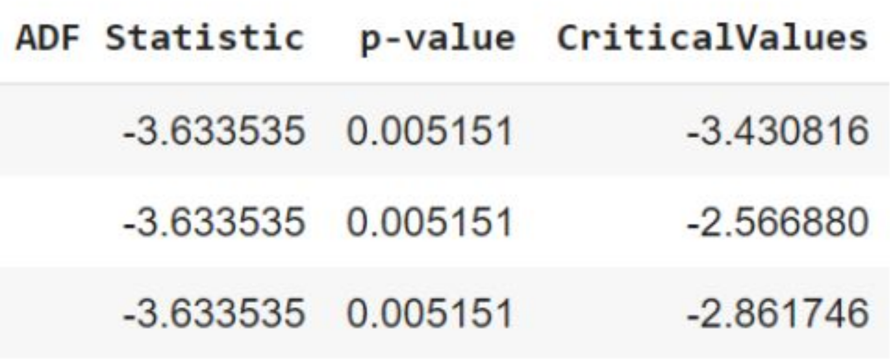
Unit Root Test for Stationarity

Augmented Dickey-Fuller Test

ADF Statistic > Critical value(at 1%)

P-value < 0.05

Hence, hypothesis is rejected. The data is **stationary.**

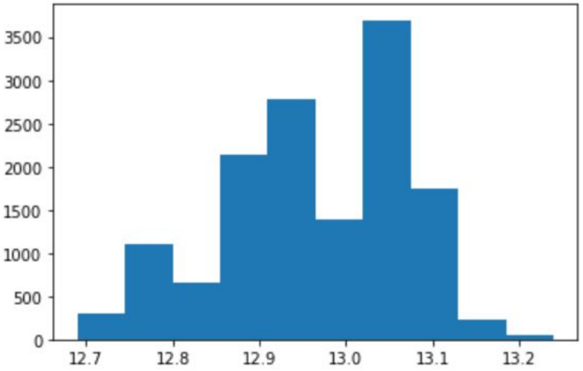


ADF Statistic	p-value	CriticalValues
-3.633535	0.005151	-3.430816
-3.633535	0.005151	-2.566880
-3.633535	0.005151	-2.861746

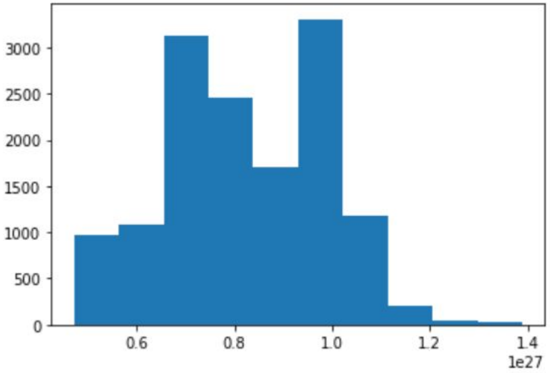
Box Cox Transformation

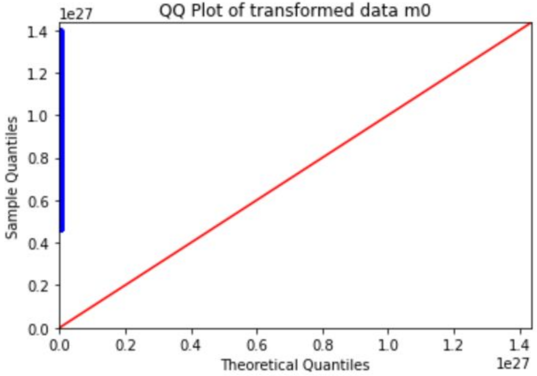
Box cox transforms are not as effective in this case.

Before



After

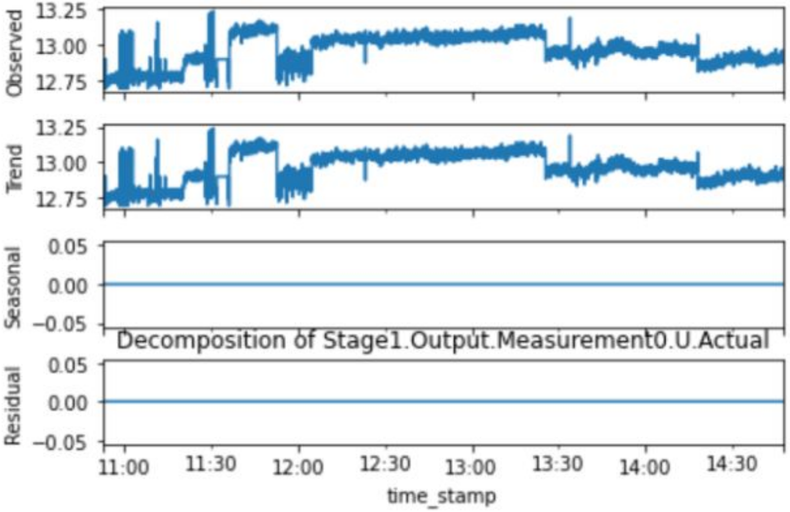




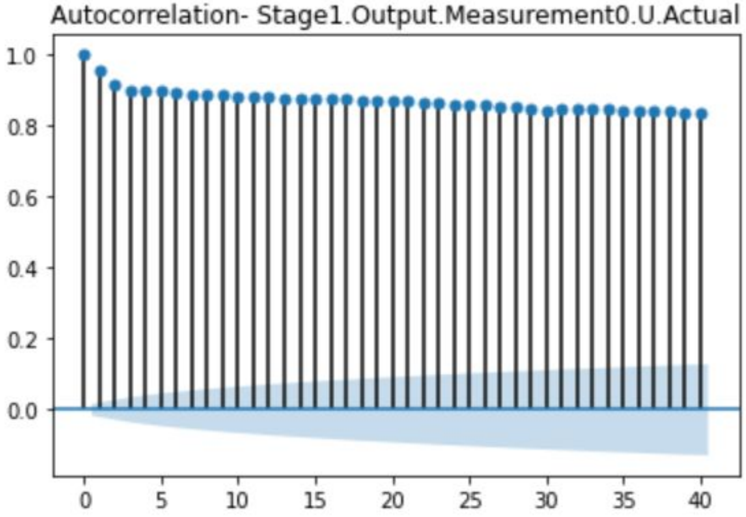
Decomposition of Time Series

y(t) = Level + Trend + Seasonality + Residual

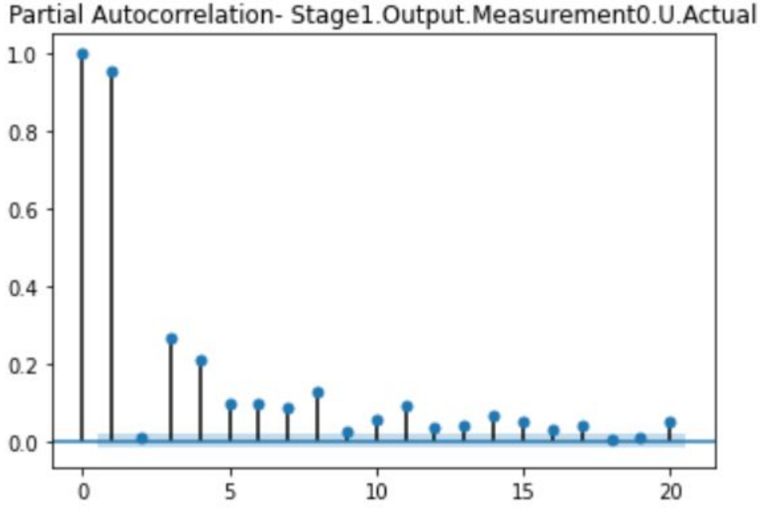
No significant trend or seasonality is observed.



Autocorrelation and Partial Autocorrelation



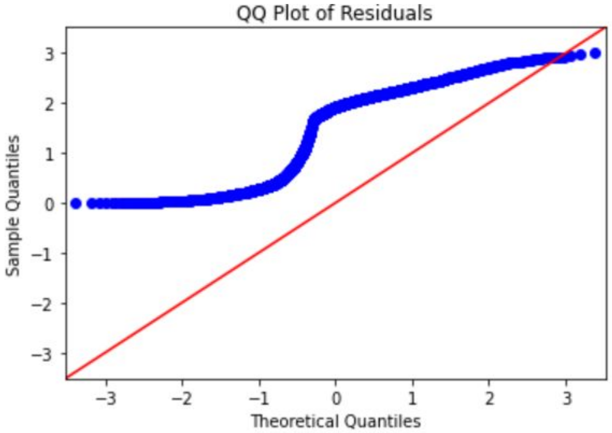
Autocorrelation- Stage1.Output.Measurement0.U.Actual

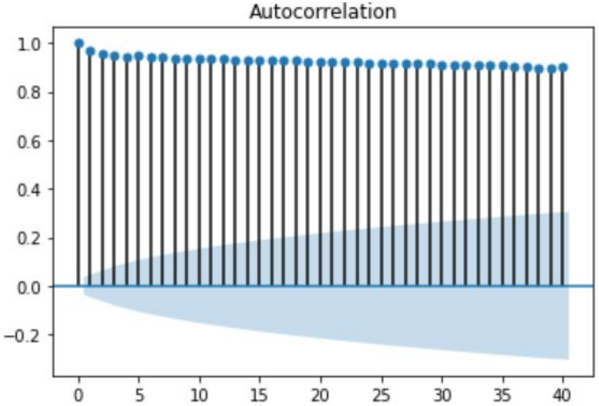


Partial Autocorrelation- Stage1.Output.Measurement0.U.Actual

AR_X Model with p = 1

- The acf curve trails off.
- The pacf curve shows hard cut-off at p=2
- This suggests an **AR_x model with p=1.**
- Model equation: $y(t) + a_1y(t-1) = e(t) + u(t)$
- Mean squared error for p = 1 is **3.0653**
- Normalized Root Mean Squared error for p=1 is **4.1988**
- The QQ plots and autocorrelation of residuals shows that the residuals are not white noise. Hence, the model is not a good fit.





LSTM

LSTMs have a long range memory. They are sequence models and consider time dependence between input samples.

Features:

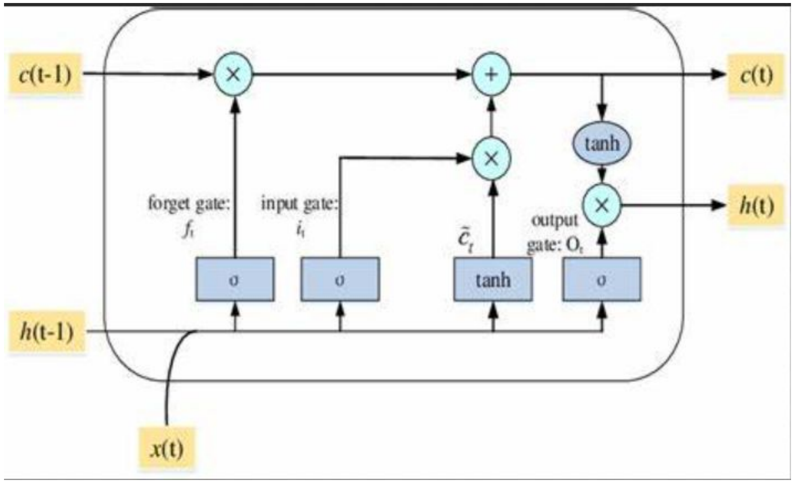
1. Cell State- ensured memory
2. Forget Gate- throws unnecessary information
3. Input Gate- adds new information
4. Tanh Layer- Creates new candidate for cell state
5. Cell state update

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

6. Output Evaluation

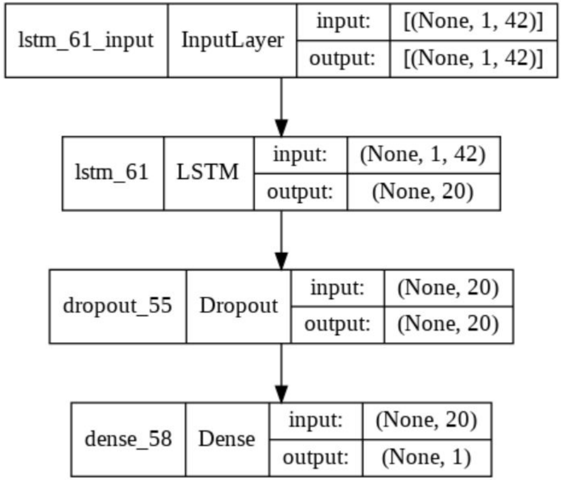
$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

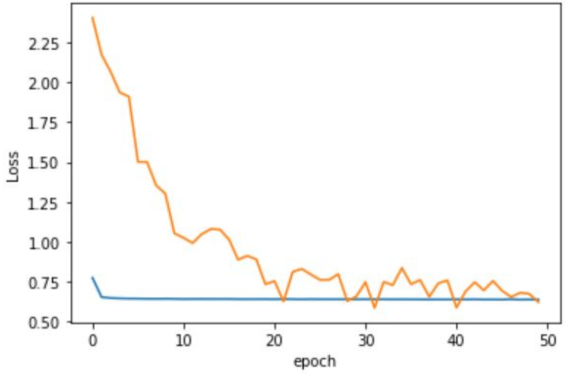


LSTM Model with 20 units

- The LSTM layer uses a **tan(h)** activation function.
- The Dense layer utilizes **ReLU** as its activation function, since this is a regression problem.
- The optimizer used to solve for minimum loss(mean squared error) is **Adam** optimizer.
- A **Dropout** layer with a probability of 0.2 is employed to avoid overfitting.
- Root Mean Squared error is **0.7897**
- Normalized Root Mean Squared error is **1.8938**

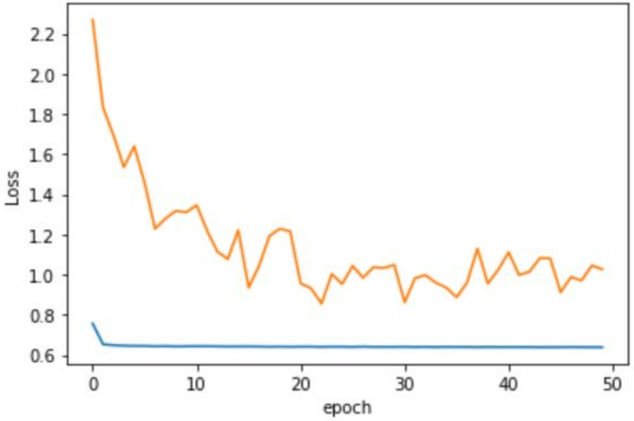


lstm_61_input	InputLayer	input:	[(None, 1, 42)]
		output:	[(None, 1, 42)]
lstm_61	LSTM	input:	(None, 1, 42)
		output:	(None, 20)
dropout_55	Dropout	input:	(None, 20)
		output:	(None, 20)
dense_58	Dense	input:	(None, 20)
		output:	(None, 1)



LSTM Model with 26 units

- The model complexity increases.
- Model becomes prone to overfitting.
- Root Mean Squared error is **1.0132**
- Normalized Root Mean Squared error is **2.4299**



Conclusion

- No feedback loop is present in the given system and hence, this is a **regularization problem.**
- An ADF Statistic lower than a critical value, with 1% probability (in this case), indicates a **stationary** time series.
- Box cox transforms are not always effective while converting a time series to a normally distributed series.
- The seasonal decomposition of the data indicates no significant trend or seasonality.
- Regularization and reduction in complexity of models help in preventing the overfitting of models.
- **LSTMs** provide better models for time series data over traditional statistical models like **AR**. In this case, it reduced the normalized root mean squared error by almost **4** times.