Sequential Analysis Forecasting

Yevhen (Jake) Horban

Problem and assumptions

- Predicting the amount of roses produced
- Project assumes that if the environmental factors were controlled the production would follow a sinusoidal pattern with no noise
- Let’s build a model

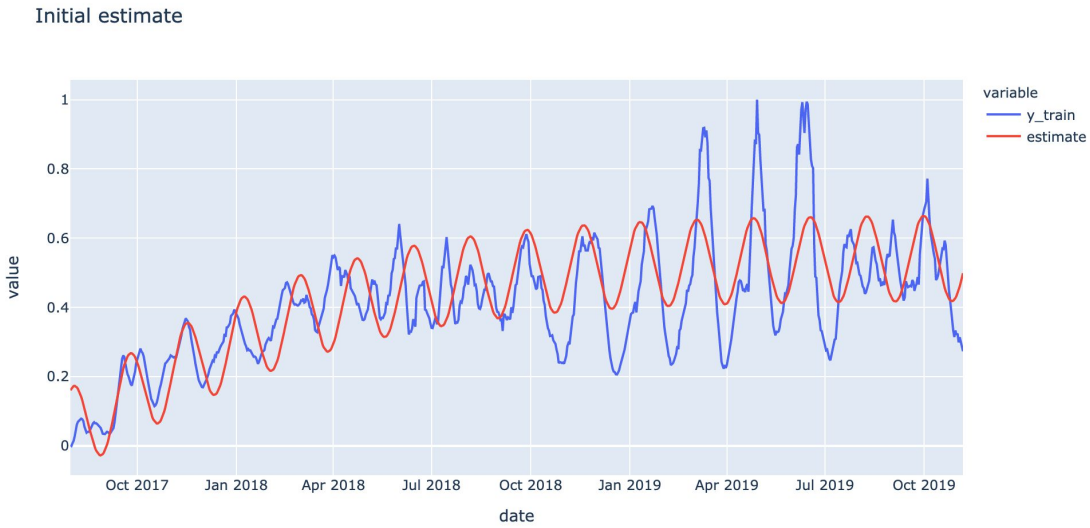
Method: Step 1. Initial estimate

$$a(x) = \frac{A}{1 + e^{Bx}} + D$$

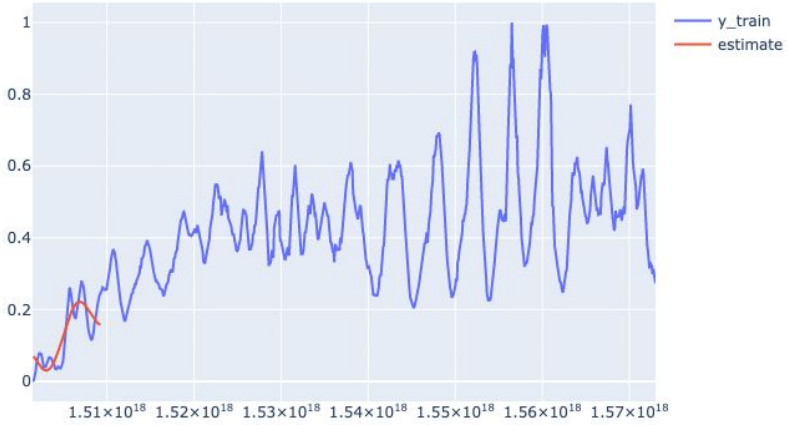
$$b(x) = Asin(Bx + C) + D$$

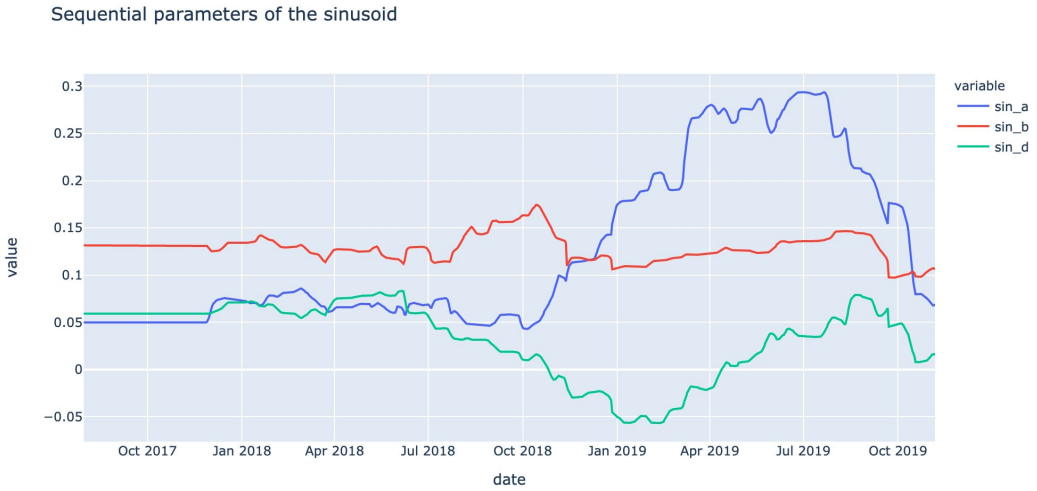
$$f(x) = a(x) + b(x)$$

```
log_param: [-0.00729785]
sin_params: [1.23348336e-01 1.19482305e-01 5.47880525e+02 4.40281044e-02]
R2 = 0.5080295152765394
```

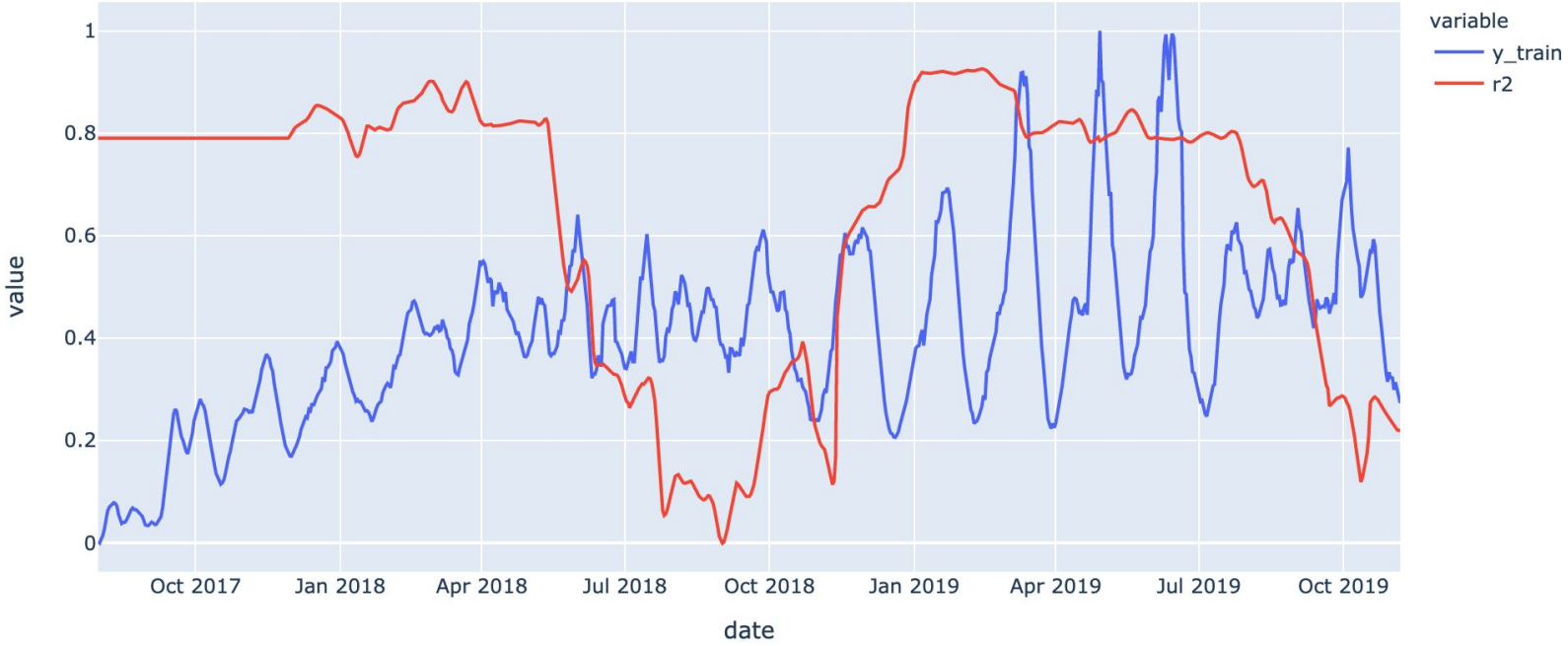


Method: Step 2. Sequential scanning



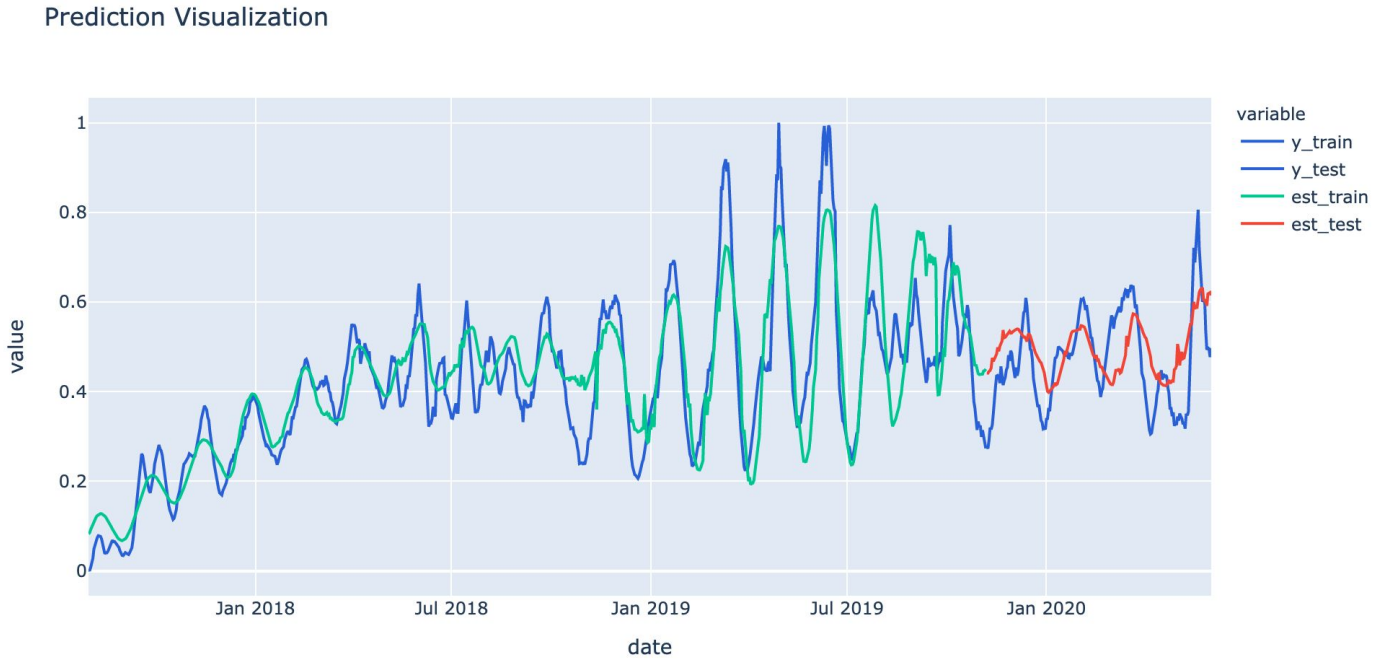


R2 of the sequential fitting



Method: Step 3. Prediction

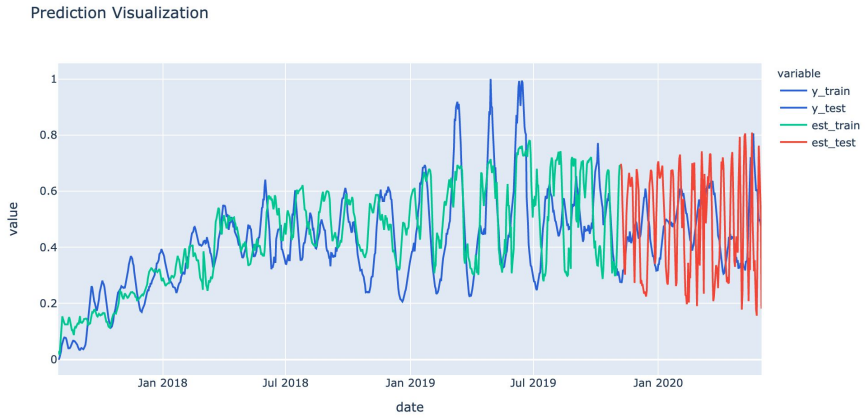
```
sin_params    0.999
train         0.755
test          0.259
Name: R2 of RandomForestRegressor(random_state=0), dtype: float64
```



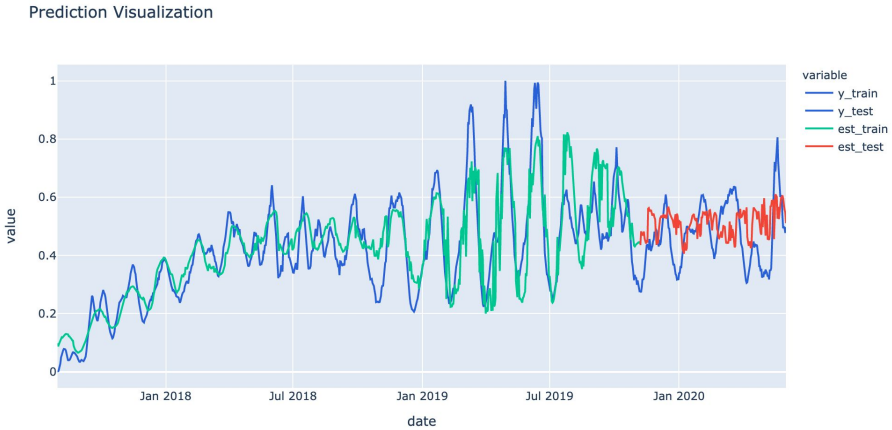
Modifications

- Using R2 values as weights for fitting the model
- Not recording the vertical shift in the sin curve
- Adding an index column to the training and testing data

```
sin_params    0.432
train         0.417
test         -3.455
Name: R2 of LinearRegression(), dtype: float64
```




```
sin_params    0.996
train         0.713
test         -0.204
Name: R2 of GradientBoostingRegressor(random_state=0), dtype: float64
```



Future work

- Figure out why it works, without messing up the result
- Tune hyperparameters to see if the model can do better
- Compare performance to other methods like ARIMA, LSTM and Prophet
- Suggestions are welcome