Midterm Project

CUNY DATA 624 | 2023 Summer I

Group 3: Daniel Sullivan, Jeff Parks, Lwin Shwe, Matthew Katz

Contents

# Overview

For this project, we were presented with six batches of blinded time series data, containing historical performance on twelve separate variables. Our goal is to provide you with statistically rigorous, but actionable advice on expected performance for each of these twelve variables - whether they correspond to product sales, web traffic, or other important business metrics.

In this report, we will illustrate for you the overall performance of each of the variables, as well as any relationships we suspect, might exist between them. We will also briefly cover any adjustments that were made to the data to aid in clean and accurate forecast models. Finally, we will provide you with our best estimates for future performance along with guidance on measures of uncertainty in the models.

# Series One

## Variables & Analysis

For the first series we examined data for **Variable 01** and **Variable 02**.

It seems that for both there was little if any seasonality but where the data seemed to have 5-day intervals with scattered gaps. A work days per year pattern was used however I do not believe that there is a major correlation with these series since most of the decomposition at various frequencies showed large runs of wave motion remainders showing there may be some other pattern to the data. each variable had a few missing variables which were imputed where necessary being replaced with the most recent non-NA value before it.

**Variable 01** shows an upward trend through the time series with what seems to be plateaus at either end with the midsection appearing linear and two almost camel humps at either end of the linear phase.

|  |  |
| --- | --- |
|  |  |

**Variable 02** shows a very slight downward trend but generally appears to be near white noise.

|  |  |
| --- | --- |
|  |  |

Both Var01 and Var02 showed a very high amount of correlation as seen with the lag plots with Var01 having a near 1.0 correlation almost 40 points out and Var02 with many values hovering around 0.5.

|  |  |
| --- | --- |
|  |  |

Because of the non-seasonal pattern of the data we moved forward using the forecast function which implemented an STL+ETS method and SES with a tuned alpha variable. After comparing these forecasts with an 80/20 training/testing split with the last 20 on the last half we tested the forecast optimizing the MAPE. We found that in all cases the tuned alpha performed better so used this for the Forecast. we found the alpha values optimized at a=0.03 for Var01 and a=0.27.

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

## Model Selection & Forecasts

Since the SES models performed slightly better, these were chosen for the forecast.

|  |  |
| --- | --- |
|  |  |

# Series Two

## Variables & Analysis

For the second series we examined data for **Variable 02** and **Variable 03**.

**Variable 02** seems to represent a product or activity with a high volume, approximately 40 million units per day. Overall there seems to be a gradually decreasing trend for this item, with neither seasonality nor long-term cyclic patterns detected. The presence of many extreme values for this variable could interfere with the accuracy of our forecasts. To correct these conditions while preserving as much information as possible, we chose to identify and smooth those 66 outliers with linear interpolation.

|  |  |
| --- | --- |
|  |  |

**Variable 03** seems to represent a product or activity with low volume, approximately 14 units per day. Overall there seems to be a steady trend for this item, with no seasonality detected. This dataset has only two missing values, and one notable extreme value toward the end of the series, smoothed here with linear interpolation.

|  |  |
| --- | --- |
|  |  |

## Model Selection & Forecasts

To evaluate and select models, we created an 80/20 train/test split and score performance using the Mean Absolute Percentage Error (MAPE) statistic.

The three models evaluated for Variable 02 were ARIMA(2,1,3) with drift, Holt (Damped), and Simple Exponential Smoothing (SES). We selected the SES method for this forecast since it produced the smallest MAPE score (30.22).

Our resulting forecast for the next 140 days has a stable mean of approximately 20.7M units daily but with a fairly large predictive interval.

The three models evaluated for Variable 03 were ARIMA(2,1,0), Holt (Damped), and Simple Exponential Smoothing (SES). We selected the ARIMA(2,1,0) method for this forecast since it produced the smallest MAPE score (16.19).

Our resulting forecast for the next 140 days has a stable mean of approximately 16.2 units daily but with a fairly large predictive interval.

|  |  |
| --- | --- |
|  |  |

# Series Three

## Variables & Analysis

For the third series we examined data for **Variable 05** and **Variable 07**.

**Variable 05** seems to represent a product or activity with low volume, approximately 76 units per day. Overall there seems to be a gradually increasing trend, but this does not appear to be a stationary dataset. By examining the differenced time series were able to confirm the lack of autocorrelations and seasonality. We also identified and smoothed several outliers and missing values with linear interpolation before modeling.

|  |  |
| --- | --- |
|  |  |

**Variable 07** seems to represent an almost-identical time series to Variable 05 - in fact, the two datasets are 98% correlated.

|  |
| --- |
|  |

## Model Selection & Forecasts

To evaluate and select models, we created an 80/20 train/test split and score performance using the Mean Absolute Percentage Error (MAPE) statistic.

The three models evaluated for Variable 05 were: ARIMA(1,1,1) with drift, ETS, and Naive. We selected the Naive method for this forecast since it produced the smallest MAPE score (15.91).

Our resulting forecast for the next 140 days has a stable mean of approximately 130 units daily but with a fairly large predictive interval.

The three models evaluated for Variable 07 were ARIMA(3,1,2) with drift, ETS, and Naive. We selected the Naive method for this forecast since it produced the smallest MAPE score (14.79).

Our resulting forecast for the next 140 days has a stable mean of approximately 128.5 units daily but with a fairly large predictive interval.

|  |  |
| --- | --- |
|  |  |

# Series Four

## Variables & Analysis

For the fourth series we examined data for **Variable 01** and **Variable 02**.

Variable 01 has two main phases with one early plateau and a hump of rapid growth towards the end. Variable 02 looks mostly like white noise with a hump towards the beginning with what may be a steep hump towards the end and a more middle hump towards the middle. However these humps are difficult to pick out because of the noise.

|  |  |
| --- | --- |
|  |  |

We chose to approach these variables with a dual forecasting approach using an ARIMA and Naive model for both variables. The time series were split into an 80/20 train/test split to test the model for accuracy.

## Model Selection & Forecasts

Comparing the MAPE output we found that the Naive model was best at predicting both variables, and was used to forecast both series.

|  |  |
| --- | --- |
|  |  |

# Series Five

## Variables & Analysis

For the fifth series we examined data for **Variable 02** and **Variable 03**.

Both contained small amounts of missing values and where applicable were replaced based on the previously held value. Additionally, since all of these data sets appear to be on a 5-day cycle the work year(260) days was used for establishing frequency.

In the **Variable 2** plots we see a gradual small decrease in value with an almost white noise-looking fluctuation. It does not seem to show any seasonality as in the decomposition plots we found in most cases there were runs of remainders that formed an almost Wave-like pattern.

**Variable 3** has what seems to be three patterns in the center of the plot there seems to be slow steady growth that is near linear with a fair amount of fluctuation. Additionally, at the start of the data set, there is a steep rise and at the end a steep fall. The seasonal patterns seem to make these more gradual however there is a good amount of wave motion in the remainder so there is an aspect of the data this is not accounted for despite the seasonal graph matching a few of the data motifs.

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

Both variable three and variable five show high levels of correlation within the lag plot with var02 hovering near 0.5-0.4 and var03 hovering around 1-0.8 for over 40 values.

|  |  |
| --- | --- |
|  |  |

For this analysis, we compared how the forecast () function performed using a STL+ETS model and how an SES(simply exponential smoothing) model with a tuned alpha variable performed. We tuned the SES by using an algorithm to reduce MAPE to determine the alpha value. We found using the highest alpha=0.99 in both cases gave the most accurate forecast when testing on the last 20% of the data with an 80/20 training testing split and when we compared this to the forecast function performed slightly better.

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

## Model Selection & Forecasts

Since the SES models performed slightly better these were chosen for the forecast.

|  |  |
| --- | --- |
|  |  |

# Series Six

## Variables & Analysis

For the sixth series we examined data for **Variable 06** and **Variable 07**.

These strikingly-similar datasets show mostly linear patterns with a few humps with larger increases and dramatic dips with one drastically larger point than any other.

The autocorrelation shows that between time points there is a very high level of correlation hovering around one for both of these time points as seen in the ACF plot. To help make clean predictions for these variables, a single high value was omitted from the dataset and all five empty values were imputed using linear imputation. While the graphs look extraordinarily similar, these datasets are not identical.

|  |  |
| --- | --- |
|  |  |

[,1]  
holt\_pval5 0.9021570  
ses\_pval5 0.9284836

[,1]  
holt\_pval7 0.9606495  
ses\_pval7 0.9168281

It seems after testing all of these models most of them perform relatively similarly with ARIMA being the worst model in both variable cases, SES being the best for Variable 06 and ETS being the best for Variable 07. However, these are such small differences that these may be performing the same and just one is affected more because of the randomness in data partitions. altogether we chose to move forward to forecast with the SES for variable 05 and ETS for Variable 07

MAPE results by model - S06\_Var05

ses holt ets arima  
data\_S06\_Var05 35.21299 35.21073 35.21299 35.47032

MAPE results by model - S06\_Var07

ses holt ets arima  
data\_S06\_Var07 35.98002 35.9767 35.98035 36.36105

## Model Selection & Forecasts

|  |  |
| --- | --- |
|  |  |

# Appendix

## R Code

# Available upon request.