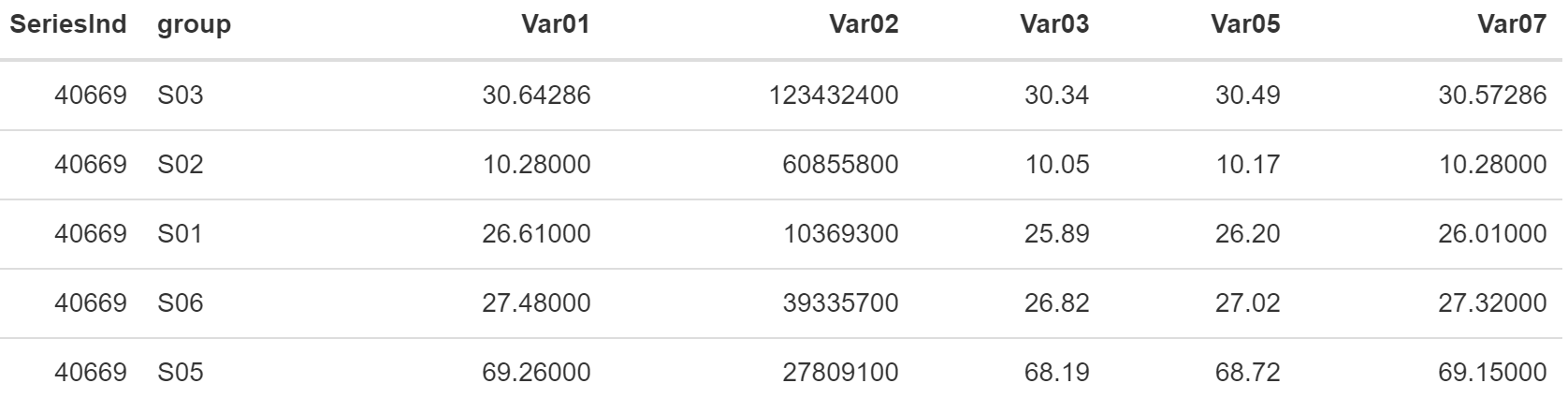
Project 1: Time Series Prediction

DATA 624-01 Group 3

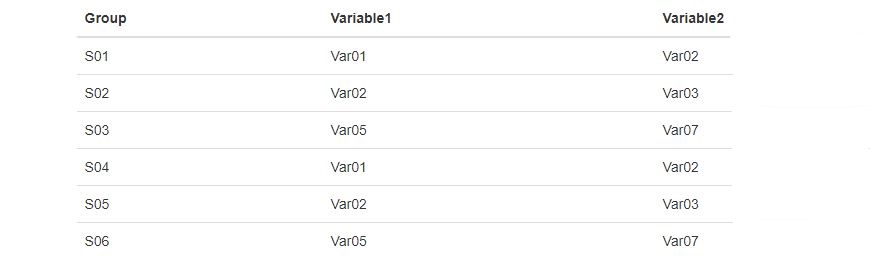
Z. Palmore, K. Popkin, K. Potter, C. Nan, J. Ramalingam

# Introduction

Given an unknown data source with several groups, we attempt to predict the next 140 values of a times series data set based on 1622 entries provided on multiple events. Our predictions will be fine-tuned to reduce the mean absolute percentage error (MAPE) as much as possible. The packages we will be using and all associated code to produce the models can be found in the attached markdown file. The data with its first five rows, are shown below.



We create forecasts for two preselected variables within each of six predetermined groups. These groups are denoted S01, S02, S03, S04, S05, and S06 respectively. There are five variables within each group that we have to work with. They are Var01, Var02, Var03, Var05, and Var07 respectively. Our date variable ‘SeriesInd,’ is displayed in its numeric serial number form calculated with Excel. Although we do not know what the variables stand for, we can develop models to try and forecast their behavior. This chart contains a breakdown of which variables are forecast in each group.



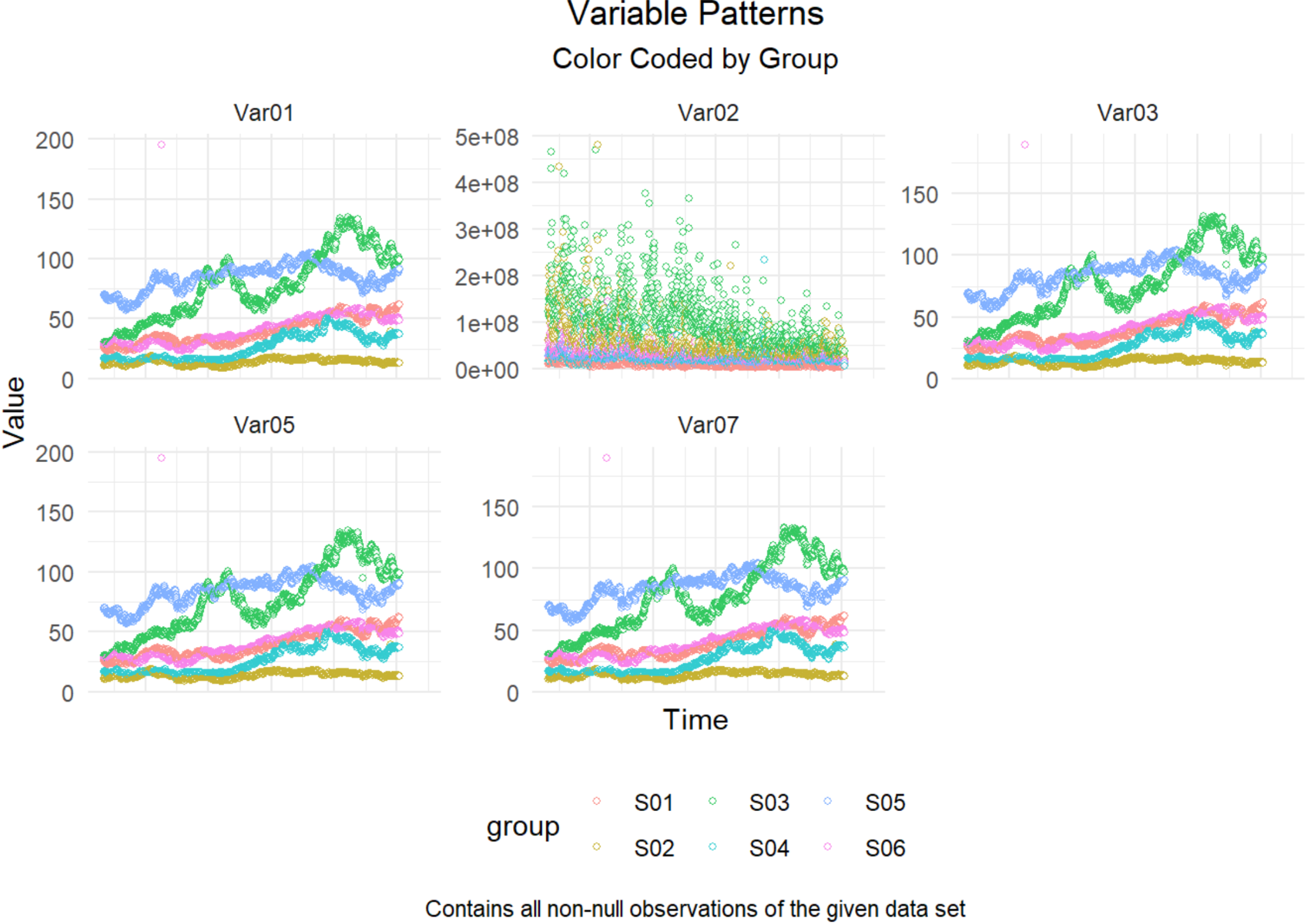
Before we begin, the data is filtered to extract each time series by group. This isolates the Var01, Var02, Var03, Var05, and Var07 variables associated with groups S01, S02, and so on. Then, with each group and its respective variables’ behavior isolated, we clean and adjust the data to make use of it in the analysis. Once we determine the most appropriate models to forecast the proper variable in each group, we evaluate the results of our predictions. Our final forecasts are captured in the excel spreadsheet attached.

# Analysis

We began by addressing missing values. Given 10,572 observations, about 8% of each variable was missing. Several methods were tried to address this but the best were Kalman smoothing and simple imputation by the median of each ‘Var0X’ variable to fill in where appropriate. The ‘SeriesInd’ numeric date was also converted from its serial number form to a common date-time series. We then examined each group’s variables separately.

Statistical summaries, box plots, and histograms were run on each group to evaluate where the average value of each variable was, if its distribution was skewed, determine whether outliers were present, and provide other descriptors of the data. These informed us that the average value (mean) of the variables are similar but their range varies widely with Var05 at 186.01 while Var02 covers a range of 479 million. Our analysis solves this potential problem by focusing on variables of the same scales as the intended target.

Additionally, all but group S03 of the histograms exhibited right skewness, and Var02 and Var03 had outliers. These were replaced using Friedman’s super smoothing method. Due to the randomness of these variables, determining outliers was difficult and there is a presence of additional overly influential points as determined using Cook's distance formula. We acknowledge the presence of these points but are unable to alter them as they are likely intentional based on the patterns in the data. For reference, the observations are shown in the scatter plot with color coding by each group.

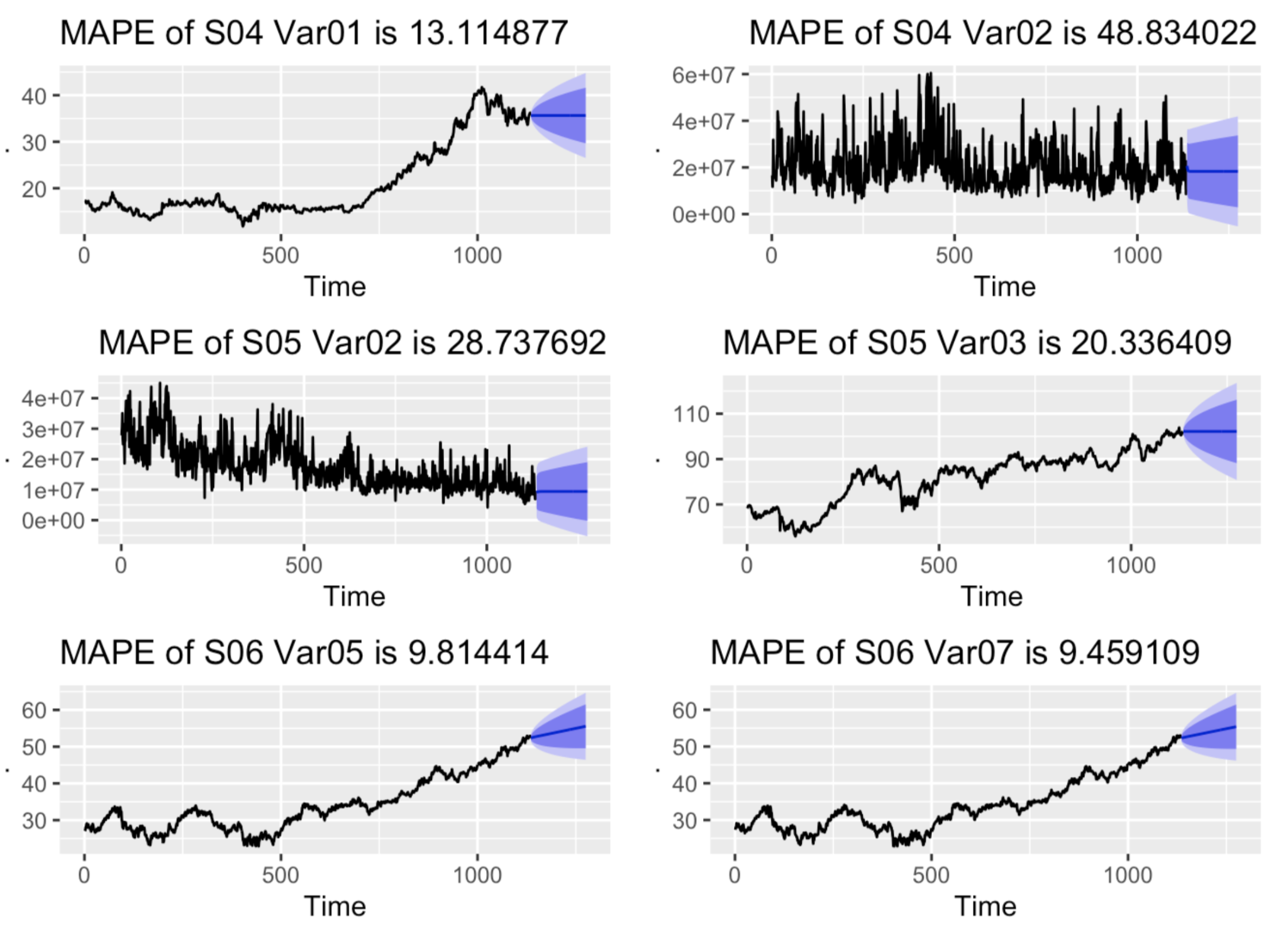


Seasonality was also considered. It is possible this data follows a weak seasonal trend that increases during summer months but there is not a lot of evidence to support regular fluctuations. Regular gaps were noticed in the time series on a weekly basis and several methods were used in attempts to fix this. However, the data appears randomly distributed and as such, acts randomly. For this reason, we left the gaps alone and any further adjustments made were minimal to avoid disturbing any existing patterns in the data.

We determined that the best model type was an Auto Regressive Integrated Moving Average (ARIMA) with drift. Unfortunately, all variables required differencing to achieve stationarity. This indicates that any predictions made with these variables may be unrealistic because of inherent random changes in statistics like the mean and variance of these variables over time. We transform the data in our attempts to achieve stationarity, but it should be noted that our review of stationarity is only a rough estimate using the aforementioned summary statistics so that we may apply this ARIMA method. Otherwise, we would have to conclude this data is inherently unpredictable and as such, render model forecasts useless. Rather, we focus on forecasting each variable individually and try to keep it simple.

# Prediction

To evaluate our models, we split the data into training and testing data sets with 70% reserved for training and 30% set aside for testing. We use the training data to build our models to forecast the proper variables in each group. We repeat the cleaning performed in the analysis and forecast the values with our ARIMA model method. These predicted values are then compared with the testing data to see if they agree using mean absolute percentage error (MAPE). Predicted values that differ from expectations increase the amount of error. Three groups are selected to visualize major trends.



Variables ‘Var05’ and ‘Var07’ in group S06 performed best based on this selection; however, it is easy to see none of them are predictable. The blue areas mark a cone of possibilities with darker shades having higher probabilities of occurrence. The straight black line shown in the center of these cones is clearly misinterpreting the data. The observed variation in each variable produces a zig-zagging black line which our model does not estimate. This is to be expected given the stationarity of data and the inherently unpredictable nature of this data set. Overall, the data appears to drift randomly from its historical observations. All remaining variables follow the same pattern.