A Sequence Analysis Approach to Improve New Product Forecasts

Shubham Gupta, Thuy Nguyen, Muthuraja Palaniappan, Sudarshana Singh, Matthew A. Lanham

Purdue University Krannert School of Management

gupta515@purdue.edu; nguye441@purdue.edu; palaniam@purdue.edu; singh554@purdue.edu; lanhamm@purdue.edu

Abstract

This study investigates a novel application of sequence analysis to forecast new product demand, which is a modified version of the approach that has been used in the bioinformatics field for protein sequencing. For retailers, the problem of forecasting demand for products which have never existed previously is challenging. Knowledge of how demand will fluctuate, especially for new product in the portfolio, will enable stores and distribution centers to manage inventory and utilize their shelf space more efficiently. We collaborated with a U.S. national retailer to develop a sequence analysis model to identify the likely sequence of purchasing future spares for various types of products on the basis of similar historical SKU sales data.

Introduction

- As new products are introduced to the market, new replacement spares are also added to stores to serve them.
- There are many possible SKUs to choose from to stock in stores and DCs, but space and purchasing budgets are limited. So, the company has to decide which SKUs should have stocking precedence and the number of SKUs.
- For most SKUs they rely on past sales history, stocking information from other locations, market data, and lifecycle curves to adequately stock stores. Since for new SKUs this information does not exist, demand profiles from similar vehicles and spares are used as a proxy, which is often highly inaccurate.

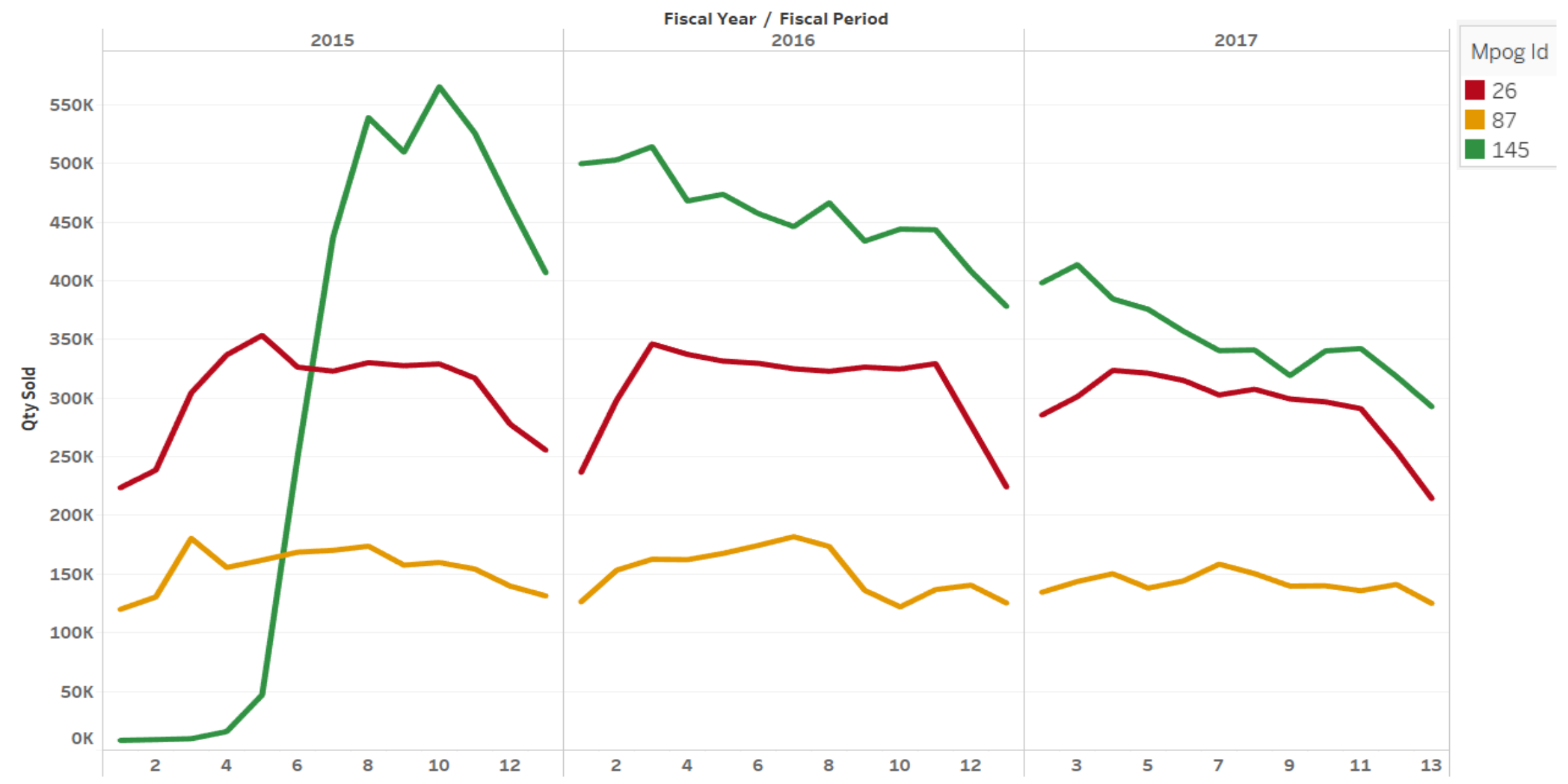


Figure 1. Different Sale Trends of various spares of a vehicle model

- Sequence analysis has not been performed in a new spare purchase context in past research so this approach has explored a new method to identify the spare replacement order and make inventory management more efficient.
- Time series prediction models using the sequence generated in first part to get higher accuracy than the traditional method not utilizing the sequence.

Literature Review

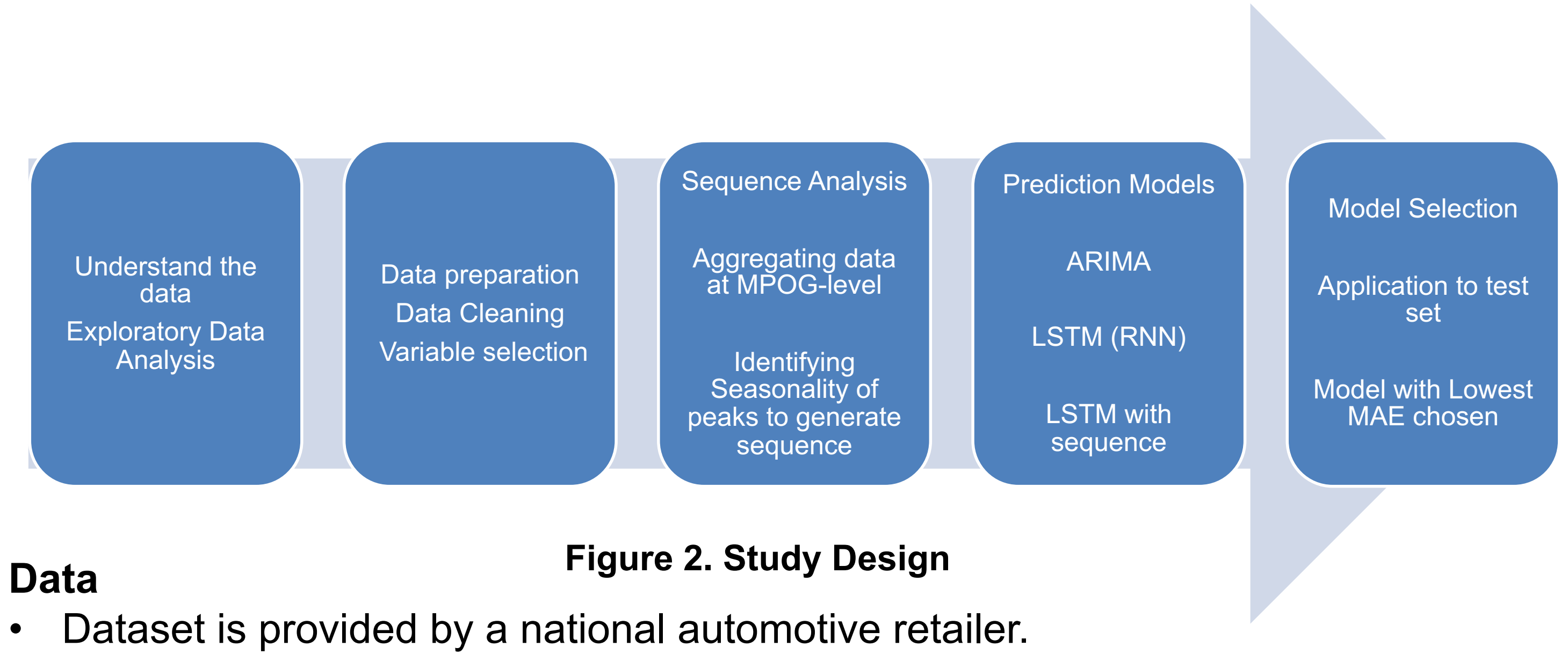
- Previous studies about sequence generation were mainly in the field of bio-informatics or mining for product data to get customer purchasing trends.
- In this research many such algorithms were studied and a new algorithm was created to make a model that supports our business problem.

Study	Sequence Generation	Demand Forecasting
Natasha A. & John B.	-	Demand forecasting with decision trees
Yun, U.	Analyzing Sequential Patterns in Retail Databases	-
Ghassen Chniti et al	-	Using LSTM and SVM
Sima & Akbar Namin	-	Comparison of LSTM and ARIMA
Press, W., & Teukolsky.	Savitzky-Golay Smoothing Filters	-
Chowdhury & Garai	Multiple sequence alignment review from a genetic algorithm perspective	-
Our Study	Sequence analysis using Savitzky-Golay Filter and Minimum Edit Distance Dynamic Programming	Forecasting using ARIMA Time Series and LSTM model

Table 1. Literature review summary by method used

- In our study, we are not dealing with individual users' patterns but rather the pattern of the entire company's sales data. Hence, we identify the sequence of the peaks of sales data.
- This study is novel because we adapt various sequence generation techniques used in protein sequencing for the demand prediction problem to identify the peak sales.
- We then apply prediction models with the sequence as one of the inputs and check accuracy is increasing or not.

Methodology



Understand the data
Exploratory Data Analysis

Data preparation
Data Cleaning
Variable selection

Sequence Analysis
Aggregating data at MPOG-level
Identifying Seasonality of peaks to generate sequence

Prediction Models
ARIMA
LSTM (RNN)
LSTM with sequence

Model Selection
Application to test set
Model with Lowest MAE chosen

Figure 2. Study Design

Data

- Dataset is provided by a national automotive retailer.
- SKU-level sales data for 5 similar vehicles that are different in brand and type of spares that they require is included.
- Spares/SKUs are organized into MPOGs (master plan-o-grams) which are large collections of similar spares that are organized into category lines based on the SKU's function.
- Data ranges from 2015 to 2017 about each SKU and more information about the product's lifecycle.
- 8,662 unique SKUs and 50 unique MPOGs are in the database.

Data Cleaning & Pre-Processing

- For sequence building, SKU- & period-level data is aggregated at MPOG level.
- The SKU sales for every vehicle and corresponding model years were grouped for the MPOG it belonged to in every period and then peaks were identified.

Methodology

Sequence Generation

- Using the detect_peak function developed by Marcos Duarte to detects peaks of the sale of each MPOG for a specific model-year vehicle. The identified peaks assist the arrangement of MPOGs in the sequence.
- In the below graph, there are 2 peaks for each year (13 periods), except the first year having 3 peaks. To be careful and not identify small blips in the sales data as peaks the Savitzky-Golay filter was used, it smooths the data by increasing the signal-to-noise ratio without distorting the signal greatly. Using convolution, this filter fits successive sub-sets of adjacent data points with a low-degree polynomial by the method of linear least squares.

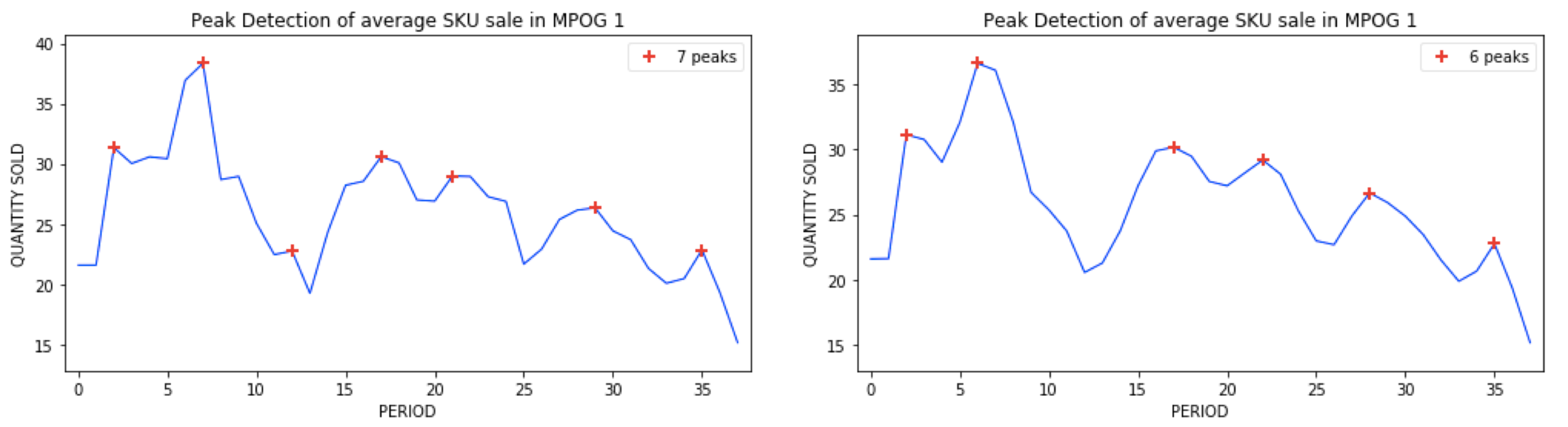
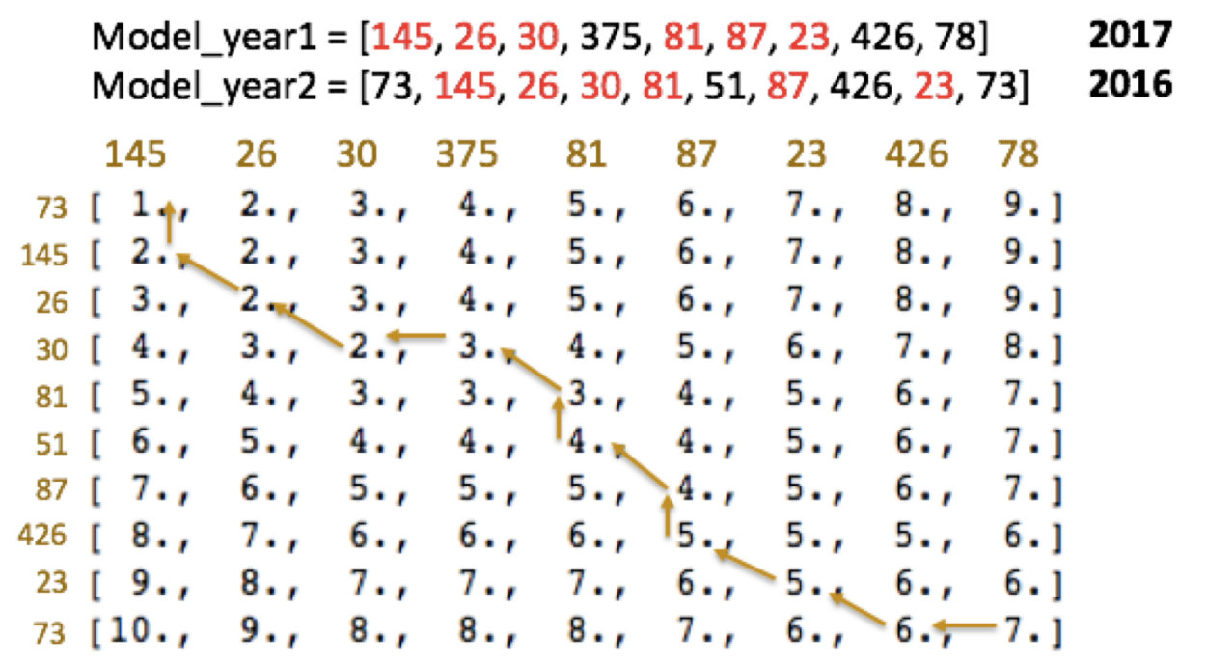


Figure 3. Before and after peaks using Savitzky-Golay Filter for smoothing peaks

- For predicting this year's new SKU sales, the sales data of each period of the previous year of the MPOG that the new SKU fits into along with the sales data of the immediately preceding MPOG in the sequence were used as inputs. Also the number of retailers and distribution centers having the MPOG into which the new SKU fits was given as input.
- The detected peaks were used to calculate seasonality which in-turn helped form sequence of MPOGs demanded by vehicles. Sequence alignment technique using minimum edit distance dynamic programming was applied to find common demand patterns among various model vehicles with newer vehicle model having more relevance (Figure 4).



```
Model_year1 = [145, 26, 30, 375, 81, 87, 23, 426, 78]     2017
Model_year2 = [73, 145, 26, 30, 81, 51, 87, 426, 23, 73]  2016

       145   26   30   375   81   87   23   426   78
 73  [ 1.,   2.,  3.,  4.,   5.,  6.,  7.,  8.,   9.]
145  [ 2.,   2.,  3.,  4.,   5.,  6.,  7.,  8.,   9.]
 26  [ 3.,   2.,  3.,  4.,   5.,  6.,  7.,  8.,   9.]
 30  [ 4.,   3.,  2.,  3.,   4.,  5.,  6.,  7.,   8.]
 81  [ 5.,   4.,  3.,  3.,   3.,  4.,  5.,  6.,   7.]
 51  [ 6.,   5.,  4.,  4.,   4.,  4.,  5.,  6.,   7.]
 87  [ 7.,   6.,  5.,  5.,   5.,  4.,  5.,  6.,   7.]
426  [ 8.,   7.,  6.,  6.,   6.,  5.,  5.,  5.,   6.]
 23  [ 9.,   8.,  7.,  7.,   7.,  6.,  5.,  6.,   6.]
 73  [10.,   9.,  8.,  8.,   8.,  7.,  6.,  6.,   7.]
```

Pattern Found:

[145, 26, 30, 81, 87, 23]

Figure 4. Minimum Edit Distance Dynamic Programming for sequence alignment

- With patterns found for each pair of model-year sequences, a heatmap scoring the pairwise alignment is created (Figure 5) identifying clusters of vehicle model years with similar sequences. To predict the sale of new SKU, the sequence pattern from the rightmost cluster is chosen as input for demand forecasting.

Results

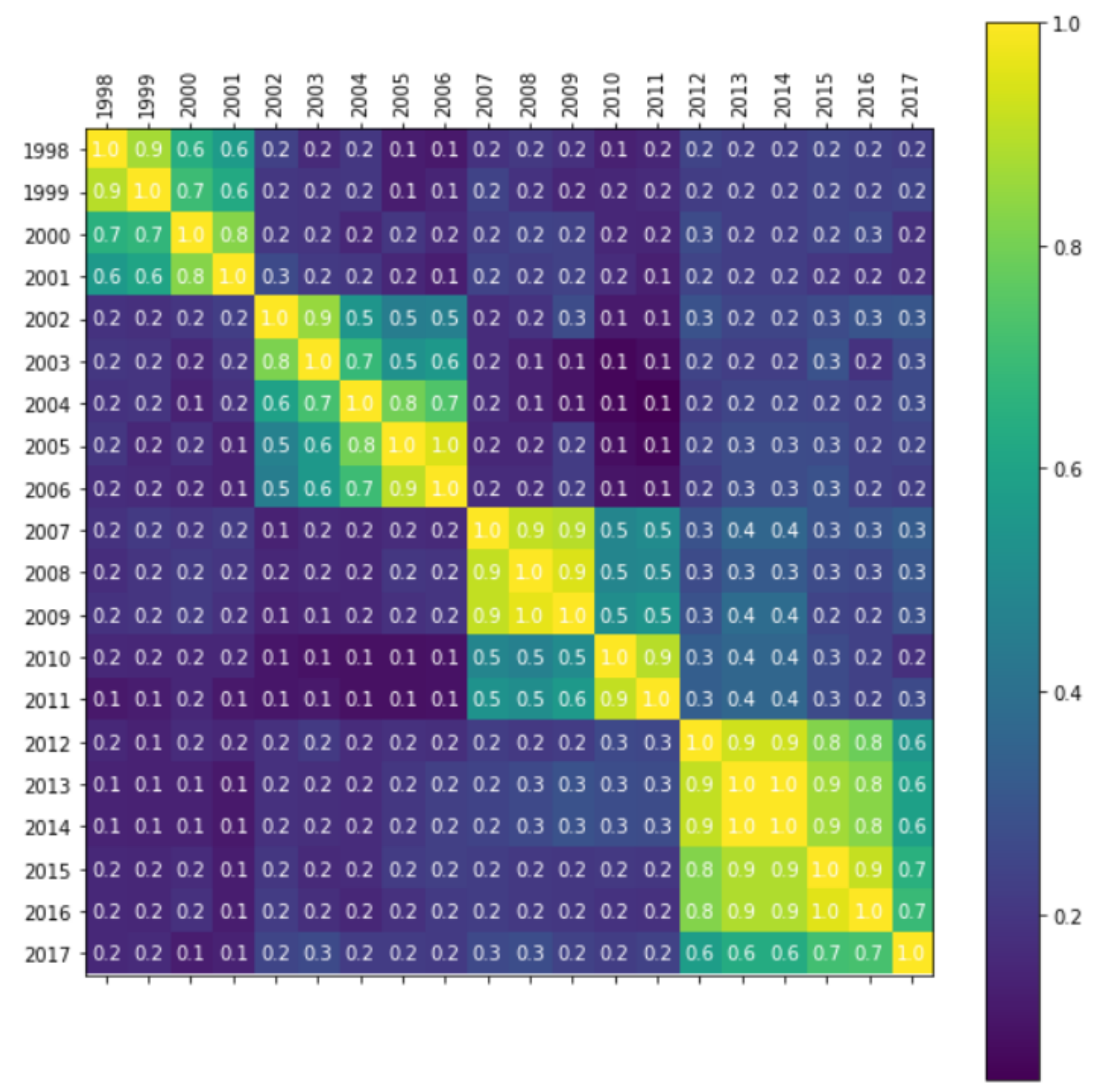


Figure 5. Heatmap table showing 4 different clusters of vehicle model-year sequences

Demand Forecasting

- With time series data, ARIMA is the common algorithm for prediction. Therefore, ARIMA model was built with only the preceding year's sales data of the MPOG that the new SKU fits in.
- For ARIMA, Augmented Dickey Fuller test was used to check stationarity. Best combination of different AR, Differencing and MA picked using grid search with lowest AIC as criteria.
- LSTM (Long-Short Term Memory) model (Recurrent Neural Network) with and without the sequence data were built. The models without the sequence are the base models and the efficiency of the sequence was compared to the base models to check the prediction accuracy.
- For LSTM model **without sequence**, the input includes the sale of the MPOG that the new SKU fits in the previous periods, number of retailers and distribution centers having stocks of that SKU were given additionally.
- For LSTM model **with sequence**, from the common sequence for various model years, we obtain the immediately preceding MPOG in the sequence. Therefore, addition to the inputs same as the LSTM model without sequence, another input for prediction is the sales of the preceding MPOG in previous periods.
- Predictions are done to check whether the MAE has improved. We chose MAE as it does not penalize the larger errors more as MSE does. MAE improved significantly when the sequence data was also used.
- Data points were less for the ARIMA to perform optimally. LSTM gave good results. With the Sequence data, LSTM gave forecasts that were closer to the actual sale results of the new SKUs with 20% improvement in MAE.

Models	MAE
ARIMA time-series model for all MPOGs (average)	16,832,435
LSTM model without sequence	63.166
LSTM model with sequence	51.009 (20% IMPROVEMENT)

Table 2. MAE for new SKU sales forecast with and without utilizing sequence data

Conclusions

The improvement of MAE from the LSTM model without sequence to the one with sequence data as an input shows that sequence analysis is an important addition to the existing approach of demand forecasting. Knowing the order in which parts get replaced by customers, stores can reduce lost sales and increase customer satisfaction. Therefore, any brick-and-mortar or online retailer having such similar SKU range and having varying seasonal demand across your product line can utilize this sequence approach to save significant financial inventory costs as a result of the reduction in inaccuracies of your demand forecasting.

Acknowledgements

We thank Professor Matthew Lanham for constant guidance on this project, and our industry partners for giving insights about their business and their data.