RC25279 (W1204-007) April 12, 2012
Mathematics

IBM Research Report

Modeling and Forecasting of Enterprise-level Retail Time Series Data with Implementation in SPSS

Ramesh Natarajan, Xiaoxuan Zhang
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Modeling and Forecasting of Enterprise-level Retail Time Series Data with Implementation in SPSS

Ramesh Natarajan, Xiaoxuan Zhang

IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Abstract

This report describes a forecasting methodology for retail chains and consumer products manufacturers, which is based on the data management tools and the time-series modeling techniques in the SPSS Statistics. The proof-of-concept of this methodology was carried out using three years price and weekly-aggregated unit sales data for the products in a specific category (Bread) that was obtained from individual grocery stores of a retail chain in a certain market geography (Kroger retail chain in the metropolitan Denver area). This methodology can incorporate additional data on exogenous demand factors which invariably improve the forecasting accuracy; these exogenous factors include for example, holiday and seasonal effects, product delivery and inventory data, promotional and marketing information, and product drag and competition effects. The implementation of this methodology with the SPSS Statistics engine called within Python programs flow provides retail chains and consumer product manufacturers with a cost-effective and scalable enterprise-level forecasting solution across the full range of their individual products and retail outlets.

1 Introduction

The various entities in a retail supply chain, such as the retail chains and stores, the intermediate distribution depots, and the consumer product manufacturers, all share a common and critical need for accurate demand forecasts for all their products across all their end-point retail outlets. These end-point forecasts at the level of individual stores and sku's are required for

many important retail decision-support applications; for instance, to plan replenishment levels and fix stocking lead times in stores and depots, to ensure on-shelf product availability and to reduce the levels of returns and unsold inventory in stores, to plan distribution capacity and order fulfillment policies in depots, and to evaluate and improve the effectiveness of pricing decisions and product promotion campaigns.

The ability of these retail supply-chain entities to perform the required end-point demand forecasts is enhanced by the availability of historic time-series demand data from one or more data sources; these data sources include for example, scanner-level sales data (for VMI products), consolidated sales data sets provided by third-party information collectors (such as Nielsen and IRI), and demand signal repositories (DSR) that are used by retail supply-chain participants for mutual visibility to shared data for monitoring and diagnostics.

Although there are a number of specialized packages available for retail demand forecasting, the use of a general-purpose statistics package such as SPSS Statistics for this forecasting application provides certain advantages over these specialized packages including for example, rigorous and tested statistical algorithm implementations, extensive capabilities for project and data management, ancillary statistical analysis, data pre-processing and results post-processing, and formatted report generation. Furthermore, these wide-ranging capabilities are easily accessed using intuitive graphical user interfaces and wizards, and likewise, easily automated and deployed using customized command scripts. Therefore, in this report, we review certain aspects of the implementation, such as missing value estimation, model specification and selection, and the use of the Python-based interface for application integration, which emphasize the particular relevance of SPSS for the retail forecasting application.

The outline of this report is as follows. Section 2 investigates the missing value issues in the typical retail time series data, and introduces an EM (Expectation Maximization) based algorithm to improve the quality of data. Section 3 reviews the analytical methodology used for the forecasting application based on customized time-series modeling. In section 5, we investigate using linear mixed model to extract common features of the same product across all the stores, which is useful in finding fixed effects in forecasting and examine the predicted sales for a particular store. Section 6 covers the use of SPSS including the Python programming interface for implementing this analysis and forecasting procedure, along with details on the forecasting performance. Section 7 provides the summary discussion of the methodology, and some of the directions of our future work.

2 Missing Values

The original data set contains many missing values, which correspond either to the case when the product was not available for sale in the given store, or to the case when some process error in the data collection lead to the omission of the relevant data. Figure 1 plots the time series data of sales versus unit price in retail store 80477 for different bread products. Usually the sales and price time series are negatively correlated. The imputation of the missing data in these two cases are quite different.

2.1 Missing Values for Unit Price

For the same product sku, the unit price is approximately equal across different stores, or at least has the same pattern of increasing/decreasing over time. This can be observed from the time series plot of unit price for all the stores in Figure 2.

We can find from Figure 2 that the missing unit price value for some store can be estimated by exploring the values of the same product sold in other stores in the same week. We use the Expectation-Maximization Algorithm (EM) with the assumption that the underlying distribution for the unit price of all the stores in the same week is Normal.

2.2 Missing Values for Sales Quantity

Since the sales quantity has a strong correlation with the unit price, it is natural to make the assumption that the quantity values also have a strong correlation for the same product in the same week from different stores. A first check of the time series plot of sales quantity in Figure 3 for all the stores helps to confirm this assumption.

The time series of different stores are not quite identical as that in Figure 2. Many factors, such as size and location of the stores, lead to different customer responses and thus sales quantities, even though the unit prices are quite similar. To check the underlying sales pattern, we scale the quantity by dividing it with the annual average sales of that store and get the time series plot in Figure 4. The underlying scaled sales quantity across all the stores in the same week shows more similar pattern and stronger correlations.

We use the EM method to estimate the missing values in sales quantity scaled by yearly mean. Let q_i denote the sales quantity, where the observations for $i = 1, 2, \ldots, n$ are available and for $i = n+1, \ldots, N$ are missing. Let μ denote the mean value of sales averaged by all the observations. The

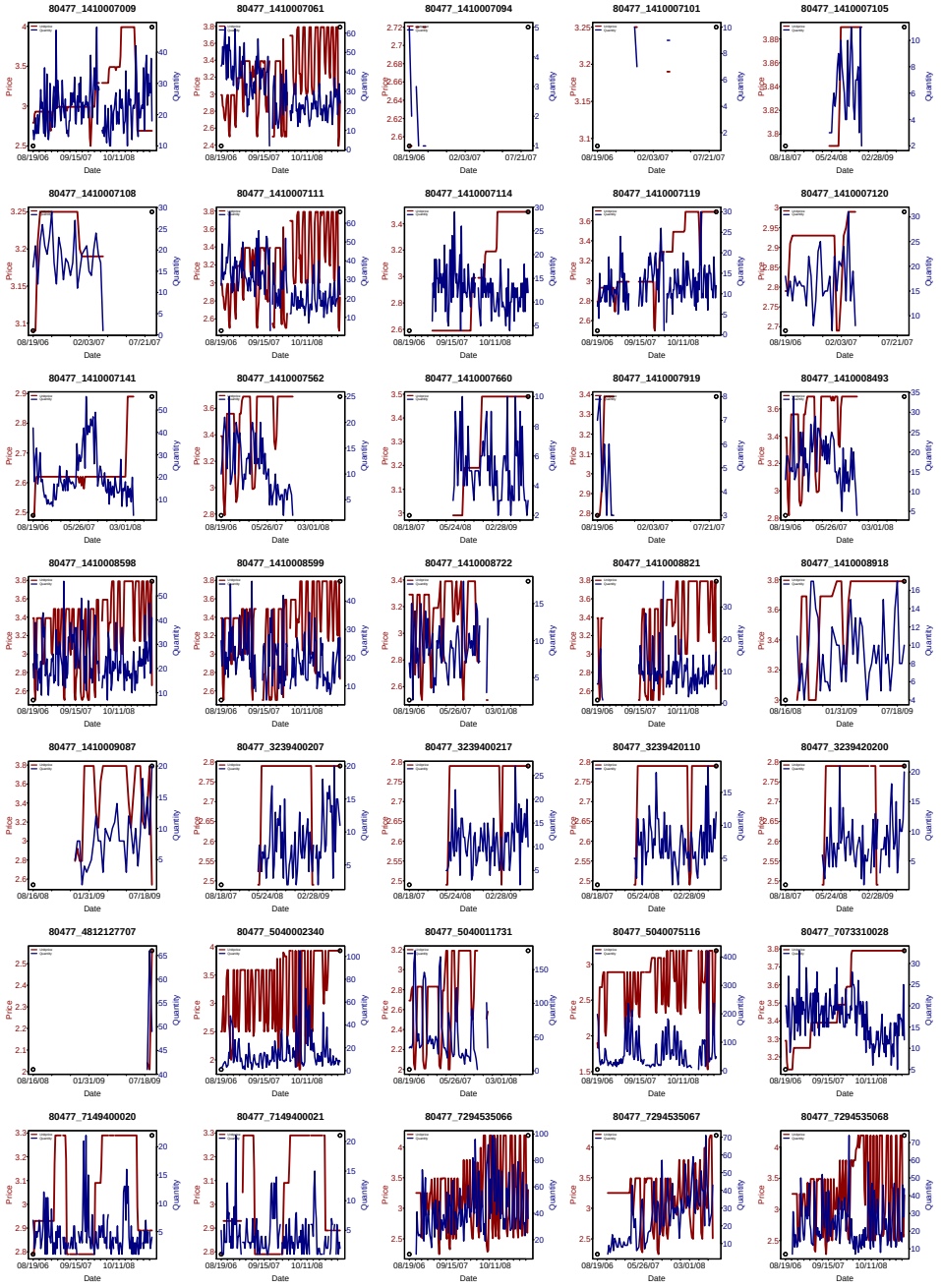

Figure 1: Sales Vs. Price of Retail Store 80477

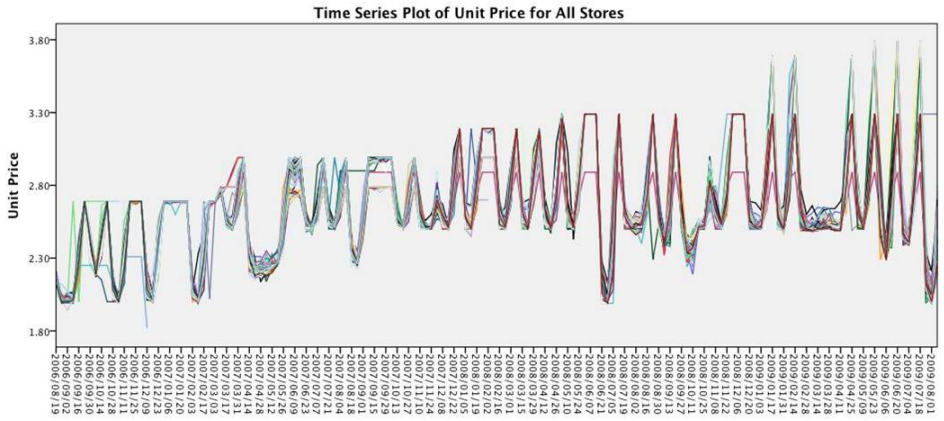

Figure 2: Time Series Plot of Unit Price for SKU 7294570544 Across All Stores

Figure 3: Time Series Plot of Sales for SKU 7294570544 Across All Stores

scaled quantity is calculated as $Q_i = q_i/N\mu$. The algorithm is as follows:

1. Estimate $Q_{n+1}, \ldots, Q_N$ by EM method based on the assumed distribution of $\{Q_i\}$.
2. Update $q_{n+1}, \ldots, q_N$ by $Q_i * N\mu$ for $i = n+1, , N$.
3. Update the mean value μ with all the updated q_i, for $i = 1, \ldots, N$.
4. Update $Q_i = q_i/\mu$ for $i = 1, \ldots, N$, where μ is from Step 3.

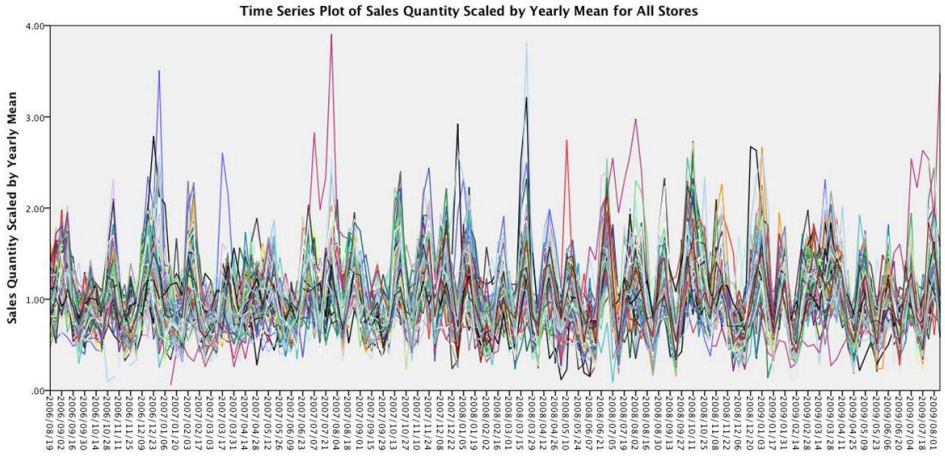

Figure 4: Time Series Plot of Annual-Revenue-Scaled Sales for SKU 7294570544 Across All Stores

5. Repeat this procedure from Step 1 until convergence of $q_{n+1}, \ldots, q_N$.

Our missing data imputation using scaled sales data provides better predicted scores comparing to the standard multiple imputation procedure.

3 Sales Forecasting

In this section we describe the methodology that is used for time series modeling and forecasting which is primarily based on the nonseasonal and seasonal ARIMAX models. The essential steps in the methodology are described in [1] , which consist of model identification, model estimation and model checking are illustrated here for the time series corresponding to a particular product and store. Similar procedures can be applied to all the product and store combinations.

3.1 Sales without Seasonal Trend

In this subsection, we would like to build the time series model for sales quantity with no obvious seasonal trend. The objective is to make the demand forecasting given future unit price. Figure 5 shows the time series plot of sales quantity and unit price for product ID 7294570544 and store 80477. The sales quantity and unit price both show an increasing trend.

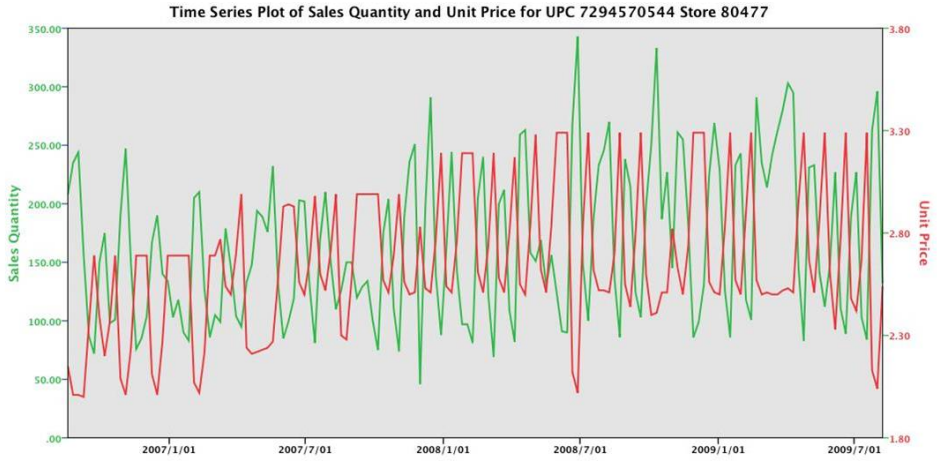

Figure 5: Sku 7294570544 and Store 80477

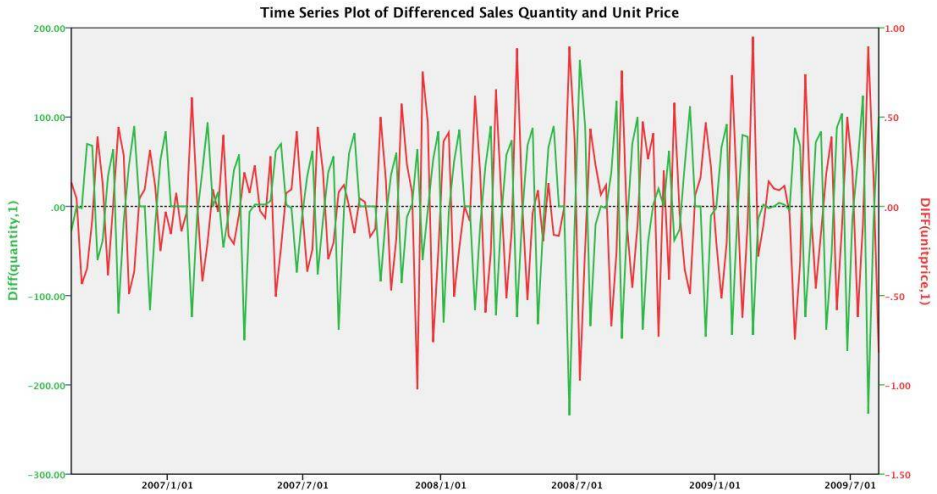

Figure 6: Sku 7294570544 and Store 80477: Differenced Series

Thus we check the time series plot of quantity and unit price with difference order 1 in Figure 6.

In addition, the plot of cross correlation (CCF) between sales quantity and unit price indicates a time series model of quantity with unit price as a predictor variable 7.

To find (p, q) for ARIMAX(p,1,q), we check the auto correlation (ACF)

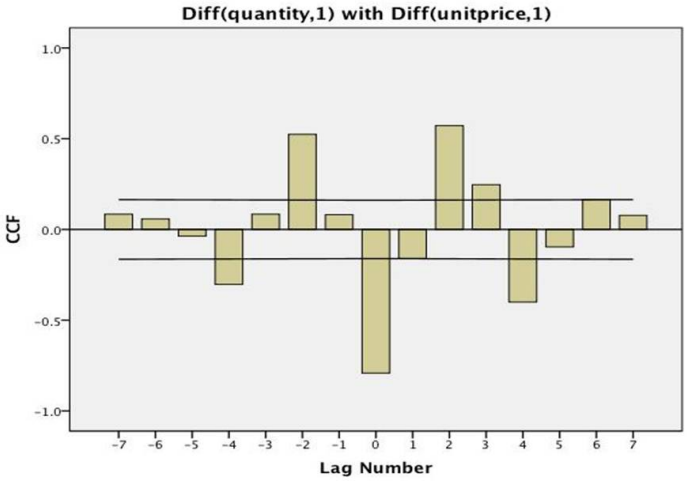

Figure 7: Cross Correlation of Sales and Price

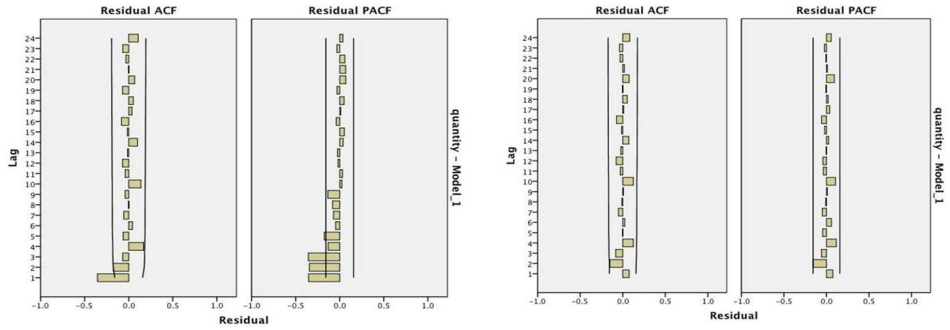

Figure 8: Auto Correlations

and partial auto correlation (PACF) for the residuals of the model ARIMAX(0,1,0) in left of Figure 8.

The ACF has a clear cutoff after lag 1, while PACF has several significant lags before decaying, thus we choose to add an MA(1) to the model. There is no more significant lags in both ACF and PACF for ARIMAX(0,1,1) in right of Figure 8.

Let q_t denote the quantity and p_t denote the unit price, then

$$\Delta q_t = b\Delta p_t + \theta\epsilon_{t-1} + \epsilon_t, \tag{1}$$

where ϵ_t is the white noise at time t. The meaning of (1) is that the change

Model Statistics					
Model	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
	Stationary R-squared	Statistics	DF	Sig.	
ARIMAX(0,1,1)	.756	19.991	17	.275	0

Table 1

ARIMA Model Parameters							
				Estimate	SE	t	Sig.
quantity	No Transform	Difference		1			
		MA	Lag 1	.870	.045	19.487	.000
unitprice	No Transfrom	Numerator	Lag 0	-166.387	9.976	-16.680	.000
		Difference		1			

Table 2

of sales quantity mainly depends on the change of the price and the average of the recent sales.Table 1 and Table 2 are the model statistics and model parameters. Figure 9 shows the fitted values and forecasting results comparing to true observations. The stationary R-squared is 0.756, implying 75.6% of variance has been explained by the specified model. The Ljung-Box significance value is 0.275, which is not significant at a 0.05 level. We can say that the residual of the fitted model is random and the model is correctly specified. In Table 2, neither of the two t-statistics is significant at 0.05 level, implying the model parameters are robust.

3.2 Sales with Seasonal Trend

In this section, we are exploring the seasonal models for sales quantity forecasting. A direct way to check the existence of seasonality is by autocorrelation plot of sales quantity. Figure 9 is the plot of ACF for upc 7294570544 and store 110044. Instead of decaying quickly with increasing lag number, we can find the spikes at lag 52, 53. Recalling our data is recorded on a weekly basis, and 52 is the number of weeks in a calendar year in Figure 10. Thus we consume there exists an annual seasonality, or a cycle of 52 periods. To further confirm this consumption, we use the spectral analysis in SPSS. The plot of the periodogram from Figure 11 shows a series of peaks

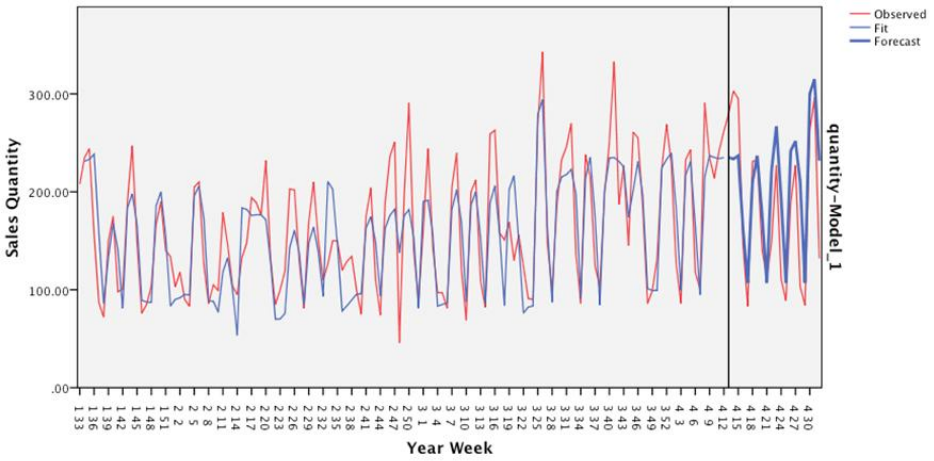

Figure 9: Predicted vs. Actual

excluding the background white noise. We can see that the first peak is at frequency close to 0.02. Since period and frequency are reciprocals of each other and1/52 is approximately 0.193, so a 52-week period corresponds to a frequency of 1/52 (0.0193). So a 52-week period implied a peak in the periodogram at around 0.019 to 0.02. In addition, between 0 to 0.1, there exists 5 almost equally spaced peaks. These facts are consistent with the existence of a 52-week period.

A preliminary model would be ARIMAX(p,1,q)(P,1,Q), with unit price as a predictor variable differencing at both lag 1 and seasonal lag 1 (or regular lag 52). We first implement ARIMAX(0,1,0)(0,1,0) and check the ACF and PACF of the residuals.Similarly, the clear cutoff of ACF at lag 1 and the slow decaying of PACF indicates an MA(1) to be added into the model. Next we implement ARIMAX(0,1,1)(0,1,0). The model statistics is shown in Table 3, and the ARIMA parameters are in Table 4. Ljung-Box Q test shows the model is not significant at 0.05 level (0.853), indicating the residuals of the fitted model are random. This can also be confirmed from Figure 9, the plot of ACF and PACF of residuals. The t-statistics shows the parameters are robust on 0.05 level.

A good explanation is that the sales can be decomposed into three parts, one of which is explained by the most recent change of the price, the second part is explained by the seasonal change of the price, and the last part is explained by the average of recent sales quantities. This is usually the case for some seasonal product, or when the store has seasonal promotions due

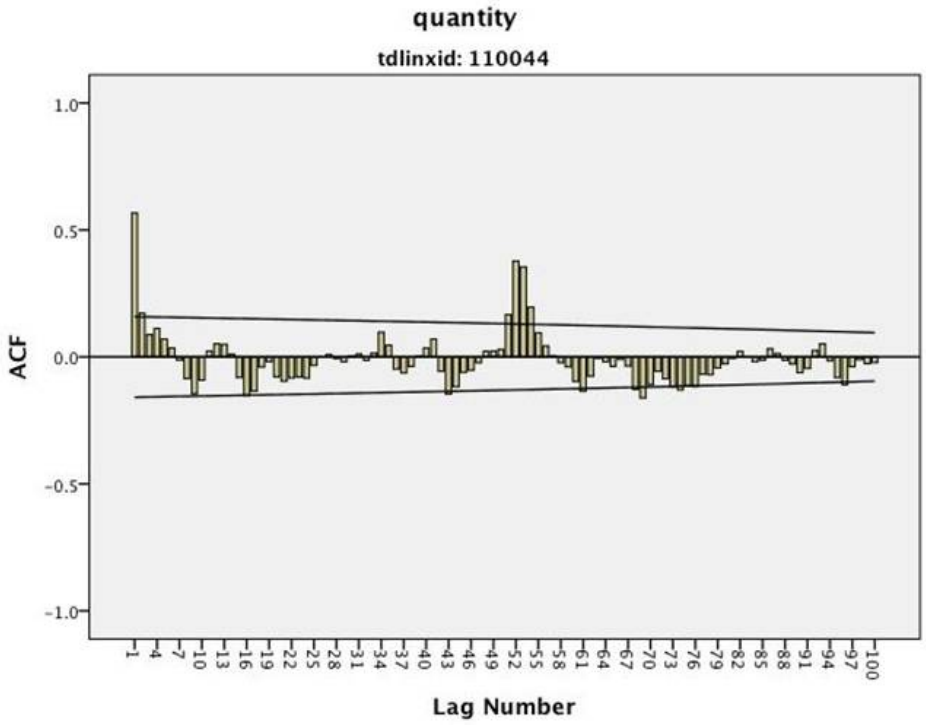

Figure 10: Autocorrelation of Seasonal Product

to holidays or events, that influence the sales.

4 Unit Price Forecasting

Usually the future price information is available to forecast demand. However, the future price information of the competitive retail brand is not known. Such price information is important since the total demand of similar kinds of products are usually stable, thus if other retail brands decide to have promotional price next week and we don't, our sales will decrease almost for sure.We can also apply time series modeling approach for unit price forecasting with no predictor variables. The performance is not quite good. Notice that the unit price fluctuates between promotional price level and non-promotional price level most of the time, and stays in each level for some weeks before next jump. This observation suggests that we can use Croston's method [2] to forecast the unit price.

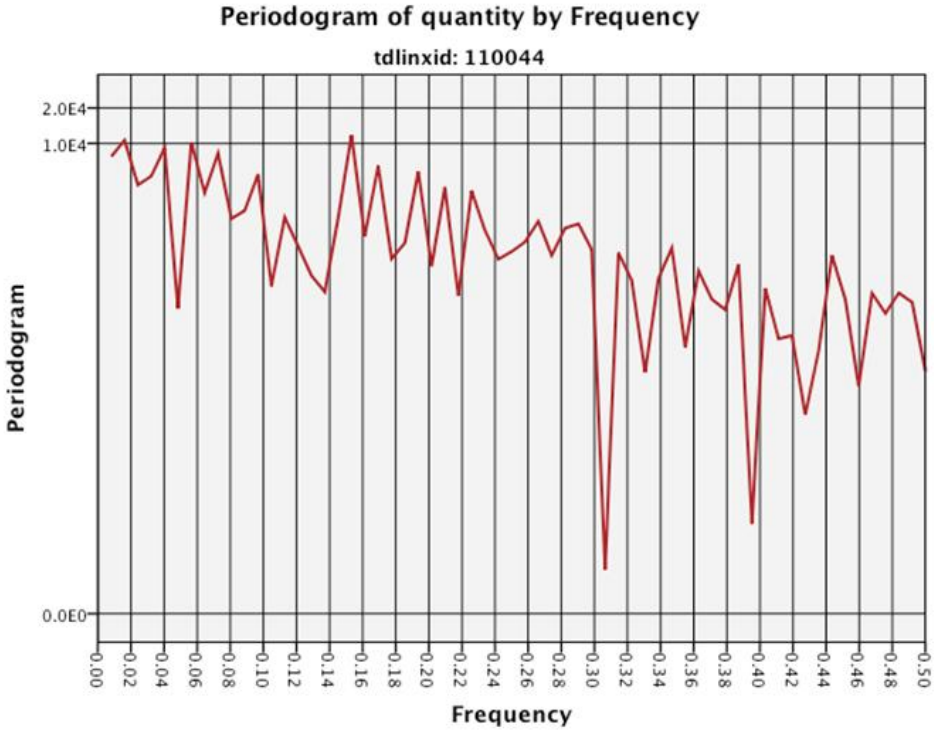

Figure 11: Periodogram

The basic idea of Croston's method is to forecast the time series of price levels and durations between price changes independently, and then combine the two forecasts. Smoothing methods such as exponential smoothing are often used in the prediction of the two series.

The original Croston's method is for zero and non-zero series forecasts. Our case here is a different in this manor. The price series consist of non-zero values only, but with promotional and non-promotional values. Such two-state feature is similar to the original Croston's method.

5 Linear Mixed Modeling Across Stores

We use the linear mixed model to find a common model for each product across all the stores, and also to further test our modeling approach or examine predicted sales. We want to find the relationship between change of sales and change of price, as well as store effects. In SPSS Mixed Linear

ARIMA Model Parameters						
				Estimate	SE	t	Sig.
	quantity	Difference		1			
		MA	Lag 1	.737	.083	8.836	.000
		Seasonal Difference		1			
	unitprice	Numerator	Lag 0	-43.241	7.217	-5.991	.000
		Difference		1			
		Seasonal Difference		1			

Table 3

Model Statistics					
		Model Fit statistics	Ljung-Box Q(18)		
Model	Number of Predictors	Stationary R-squared	Statistics	DF	Sig.
quantity-Model_1	1	.453	11.059	17	.853

Table 4

module, we specify the change of price and the store as the fixed effects, and the repeated covariance type as AR(1). The result for sku 7294570544 across 144 stores are listed below. Table 5 is the estimate parameters for fixed effects. We can see that the change of price fixed effect is negatively correlated to the change of sales, and is significant. The store effects are not significant, which is consistent with our modeling approach. Table 6 is the estimate for covariance parameters. The AR1 rho is negative, which means the increase in sales now may lead to decrease in sales next.

Figure 12 and Figure 13 are the plots of predicted sales for two stores analyzed in section 2.1. We can see that the model pooled from all the stores mainly captures the average behavior for a particular store. In addition, the seasonal component is not captured in Figure 13, which is the increasing in sales during the holiday seasons each year. This also confirms that a seasonal modeling approach is more accurate if the sales show a seasonal

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
unitpr_1	1	20285.360	14186.817	.000
tdlinxid	144	9855.445	.023	1.000

a. Dependent Variable: DIFF(quantity,1).

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
unitpr_1	-38.783138	.325612	20285.360	-119.108	.000	-39.421364	-38.144912
[tdlinxid=80477]	-.226869	1.150638	9853.490	-.197	.844	-2.482356	2.028618
[tdlinxid=80495]	.015068	1.150638	9853.490	.013	.990	-2.240419	2.270554
[tdlinxid=80514]	.014506	1.150638	9853.489	.013	.990	-2.240981	2.269993
[tdlinxid=92974]	.438769	1.150639	9853.488	.381	.703	-1.816719	2.694256
[tdlinxid=92980]	.099227	1.150638	9853.489	.086	.931	-2.156259	2.354714
[tdlinxid=93001]	.292482	1.150638	9853.490	.254	.799	-1.963005	2.547968
[tdlinxid=93010]	.122091	1.150639	9853.489	.106	.915	-2.133396	2.377578
[tdlinxid=109076]	.137954	1.150638	9853.490	.120	.905	-2.117533	2.393440
[tdlinxid=109096]	.203115	1.150638	9853.489	.177	.860	-2.052372	2.458602
[tdlinxid=109160]	.039545	1.150638	9853.489	.034	.973	-2.215942	2.295032

Covariance Parameters

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error
Repeated Measures	AR1 diagonal	352.535748	3.616095
	AR1 rho	-.265393	.006655

a. Dependent Variable: DIFF(quantity,1).

pattern. In addition, recall that the dependent variable is the change in sales, so the pooled model predicts more accurate for non-seasonal sales rather than seasonal ones. This is because for non-seasonal sales pattern, the change of sales is consistent through the year, while the seasonal ones have spikes occasionally and such spikes are not captured by the pooled parameters across all stores.

6 SPSS

The analytical and forecasting engine can be implemented with Python SPSS module, which calls SPSS engine with spss.Submit command. The commands within spss.Submit are the SPSS syntax that can run with SPSS Statistics independently. In addition, the split feature in SPSS allows analysis and forecasting on different product and/or store combinations levels

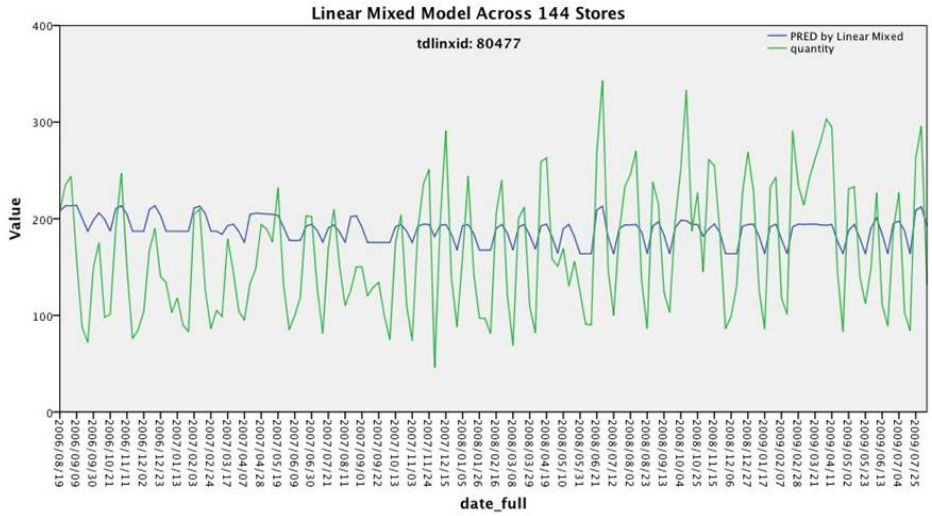

Figure 12: Linear Mixed Model Forecasts for Store with Weak Seasonal Sales

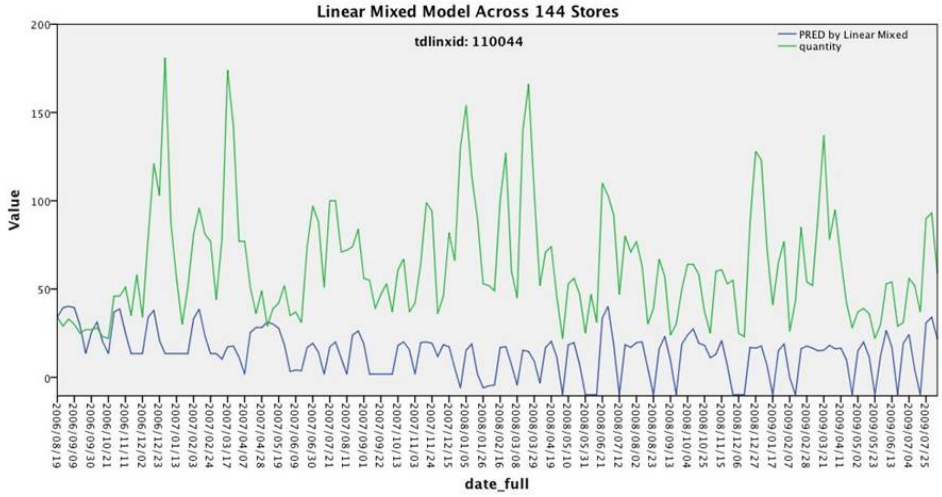

Figure 13: Linear Mixed Model Forecasts for Store with Strong Seasonal Sales

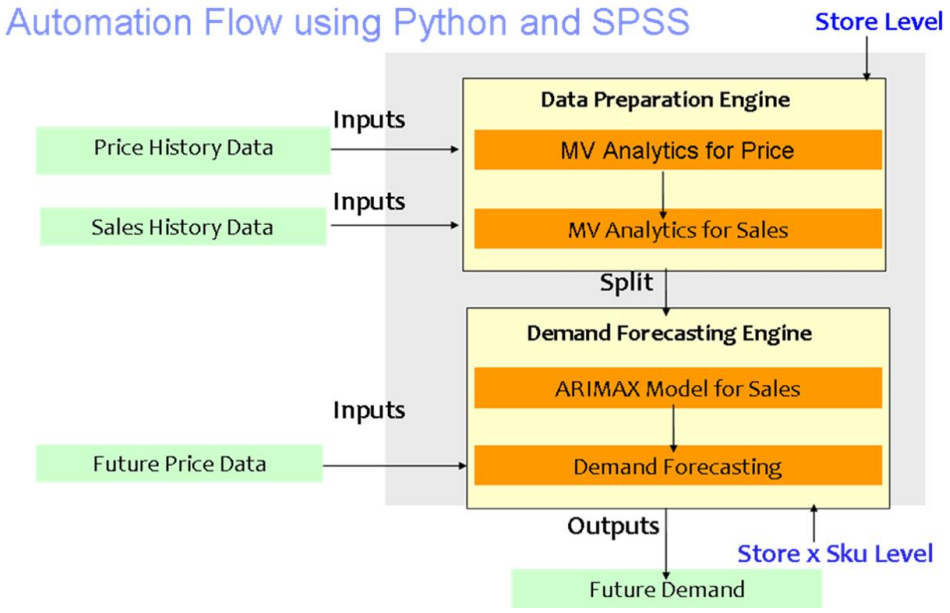

Figure 14: SPSS and Python Automation Flow

using the same analytical procedure. The outline of the engine flow is in Figure 14.

7 Conclusion and Future Work

The use of enterprise-level demand forecasting in a retail enterprise is likely to entail significant storage and computational requirements; even more so when the required forecasting cycle is more frequent (e.g., days instead of weeks), or when the forecasts must cover a wider range of products (e.g., multiple categories instead of a single category) or when a wider range of retail outlets must be covered (e.g., multiple retail chains and multiple market geographies instead of a single retail chain and geography). These potential and evolving resource-intensive requirements can be met in a cost-effective way by using cloud computing platforms for this application, and therefore, a prospective future work is to implement the forecasting methodology with a parallel computing framework.

References

[1] G.E.P. Box and G.M. Jenkins, *Time series analysis: forecasting and control*, Prentice Hall PTR, 1994.

[2] JD Croston, *Forecasting and stock control for intermittent demands*, Operational Research Quarterly (1972), 289–303.