Industrial Engineering
& Management Systems
Vol 14, No 2, June 2015, pp.196-209
ISSN 1598-7248 | EISSN 2234-6473 |
http://dx.doi.org/10.7232/iems.2015.14.2.196

Lessons Learned and Challenges Encountered in Retail Sales Forecast

Qiang Song*
Atlanta Forecasting Systems, Atlanta, GA, U.S.A.

(Received: December 17, 2014 / Revised: March 28, 2015; May 18, 2015 / Accepted: May 26, 2015)

ABSTRACT

Retail sales forecast is a special area of forecasting. Its unique characteristics call for unique data models and treatment, and unique forecasting processes. In this paper, we will address lessons learned and challenges encountered in retail sales forecast from a practical and technical perspective. In particular, starting with the data models of retail sales data, we proceed to address issues existing in estimating and processing each component in the data model. We will discuss how to estimate the multi-seasonal cycles in retail sales data, and the limitations of the existing methodologies. In addition, we will talk about the distinction between business events and forecast events, the methodologies used in event detection and event effect estimation, and the difficulties in compound event detection and effect estimation. For each of the issues and challenges, we will present our solution strategy. Some of the solution strategies can be generalized and could be helpful in solving similar forecast problems in different areas.

Keywords: Multiple Seasonalities, Event Handling, Modeling, Noise Suppression, Best Forecasting Practice

* Corresponding Author, E-mail: qsong3@hotmail.com

1. INTRODUCTION

In retail industry, it is important to forecast sales volumes, customer counts, items sold, and transaction counts at each store for the next few weeks in order to prepare for proper labor schedules. Collectively, we call sales volumes, customer counts, items sold, and transaction counts as retail sales data. Various models, methods and even forecasting systems can be found in the literature of retail sales forecasting (Chen and Ou, 2011; Lundholm *et al*., 2010; Chu and Zhang, 2003; Ni and Fan, 2010; Guo *et al*., 2013).

Accurate forecasts will provide reliable input to the labor scheduling system, reduce operational cost and increase service level. What are the major factors that affect retail sales forecasting accuracy? One would answer without hesitation that the major factors are the forecasting models. Although this statement is true to a certain degree, this is not a complete and satisfactory answer. There is no doubt that models play a very important role in retail sales forecasting. As we know, retail sales data have strong seasonalities. Usually sales are high during the weekends and low during the weekdays. Depending on the business locations, different stores may exhibit completely different seasonal patterns. For this reason, not all models work equally well for retail sales forecasting. For example, the simple exponential smoothing model is not a good one because it cannot model seasonal time series equally as well as seasonal models. Holt's model is not a good one either for the same reason. Autoregressive time series models are not good ones unless they could handle the seasonal component well. However, seasonal time series models such as SAR (Seasonal Autoregressive) and Winters models are good candidates. It is well-known that retail sales data are strongly affected by events, such as promotions. Unfortunately, none of the aforementioned models can model and forecast well sales data influenced by promotion events even if the historical data do contain the information about these events. In this sense, models are not critical if events cannot be processed properly. To compensate the weakness of the aforementioned forecasting models, data must

be preprocessed before they can be used in forecasting. Our experience is that to have good retail sales forecast, we need to follow the best-forecast practice. As a practitioner of forecast who had the opportunity to design and implement the forecasting engine for a leading retail software company in the world, and more importantly had the opportunities to access a huge amount of real life retail sales data and hence conducted data and forecasting analysis, I would like to share the lessons and issues that have been learned and encountered in performing my jobs. This paper consists of 5 parts. In Section 2, data models of retail sales data are presented. In Section 3, we will describe in detail the lessons learned in retail sales forecast. Section 4 details the challenges that we have encountered in retail sales forecast. Those challenges will have significant influence to retail sales forecast once solved. Concluding remarks are found in Section 5.

2. RETAIL SALES DATA MODEL

Let us first introduce the data model that we have been using for retail sales forecast.

At any given time, the observation data x_t can be expressed as the sum of three different components:

$$x_t = s_t + e_f + \varepsilon_t \tag{1}$$

where s_t is the seasonal component which is deemed to be deterministic, e_t is the event effect which is stochastic in nature, and ε_t is the noise component. This is an additive model. With such a model, our major task is to estimate the seasonal component s_t, model and forecast using the seasonal component as accurately as we can, and estimate the event effect e_t. ε_t is a random variable of N(0, σ^2) distribution in nature and is assumed to be independent at different times. Note that the event effect e_t exists only within a time window of an event. This window could be symmetric or asymmetric about the time of the event occurrence. Outside this time window, e_t will be out of the equation and we will obtain the following model:

$$x_t = s_t + \varepsilon_t \tag{2}$$

We say that equation (2) is a model of the normal business. That is, when there are no events, the observations are simply the summation of the seasonal component and the noise that explains the random deviation of the actual business from the normal business. Both (1) and (2) model a mature and stable business.

If we take the expected values of (1) and (2) respectively, we would obtain

$$\bar{x}_t = s_t + \bar{e}_t \tag{1'}$$

and

$$\bar{x}_t = s_t \tag{2'}$$

Those are the data models of forecasts. (1') indicates that the forecast is the superposition of the seasonal component and the event effect when events exist over the forecast horizon whereas (2') indicates that when no events exist the forecast is simply the seasonal component.

It is necessary to note that multiplicative data models are also possible. However, we are only interested in and focused on the additive models in this paper.

3. LESSONS LEARNED IN RETAIL SALES FORECAST

What is the first thing I want to share with readers in this paper? It is the lessons that I have learned in the past in retail sales forecast. Those lessons have become precious experience to me, and help me have a better understanding of retail sales forecasting.

3.1 Understanding Multiple Seasonalities in the Data Is Critical

Retail sales data are seasonal. For the majority of retail sales data, they are not only seasonal, but also they are seasonal with multiple seasonal cycles. Without knowing this, it will be hard to achieve good forecasts. Usually, retail sales data possess strong weekly and annual seasonalities. Some sales data may also possess monthly or even quarterly seasonality, depending on the nature of the business and business locations. For this reason, models used in forecasting must be able to handle multiple seasonal patterns.

What will happen if the model can only handle a single seasonal cycle length? For example, what will happen if the model can only handle weekly seasonality but the data contain annual seasonality? Then, the forecast will trail the seasonal pattern due to seasonal changes in the data. Let's use the data part of which is presented in Figure 1 as an example. The data used to produce Figure 1 indicates the strongest annual seasonality, in addition to a weekly seasonality. However, in forecasting, if we use this seasonal model $\hat{x}_t = \sum_k a_k x_{t-7k}$ which models only weekly seasonality, we will get a forecast that trails the actual data due to seasonal changes. As shown in Figure 1, the time series picks up due to seasonal changes at the beginning of July 2005. But, the forecast slowly picks up after about three weeks. When the time series moves down during August after the peak in July, the forecast cannot capture the change until about 3 weeks later. Therefore, if the model cannot handle multiple seasonal cycles, or if the seasonal cycle lengths are incorrect, unfavorable forecasts could be obtained. It has been reported in the forecasting literature (Armstrong,

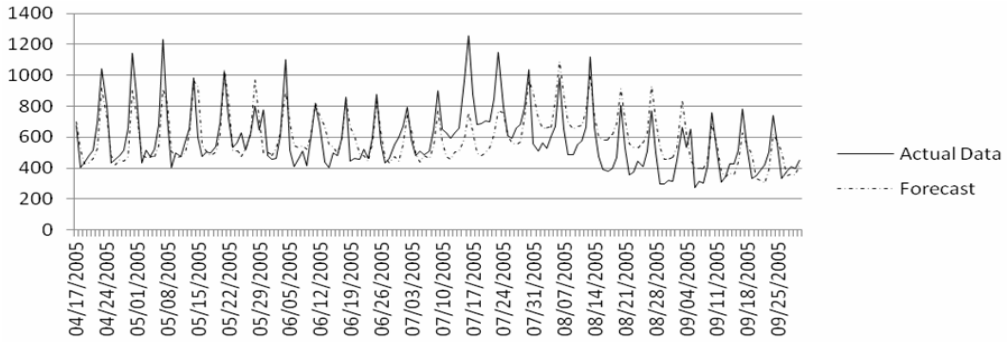

Figure 1. Actual and Forecasts when only weekly seasonality is used in forecasting.

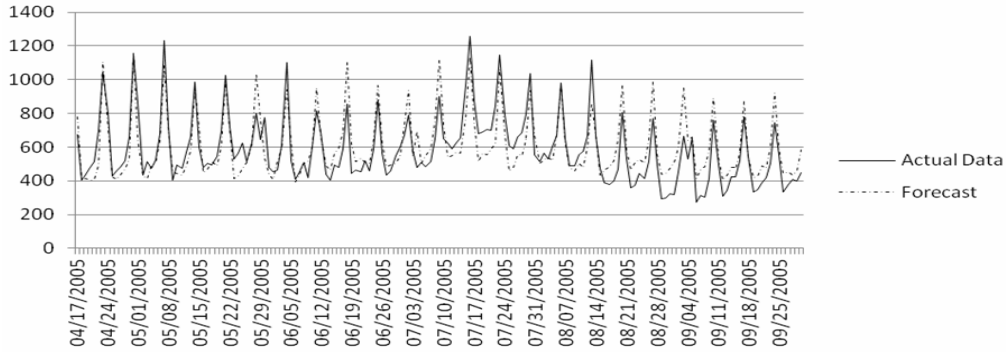

Figure 2. Actual and Forecast when weekly and annual seasonalities are used in forecasting.

2001; 231) that simple models could sometimes outperform complex models. It should not be hard to explain this phenomenon: If the time series has an annual seasonality while a complex model handles only weekly and/or monthly seasonality, the forecasts produced by the latter can be outperformed by simply using a naïve model where last year's data is used as the forecast. When the annual seasonality is incorporated in the model, forecasts could be improved and this can be seen in Figure 2. In Figure 2, the forecast can adapt itself quickly as the time series changes, although the forecasting errors are still less favorable. Therefore, it is important to understand the seasonality of the data, and incorporate the seasonality into the model. Especially, if the data possess multiple seasonal cycles, it is crucial to incorporate all the seasonal cycles in the model.

3.2 Ineffectiveness of Auto-Correlation Function and Power Spectrum Plots in Detecting Multiple Seasonal Cycles

Retail sales data are of multiple seasonal cycles. Therefore, we need not only to detect these seasonal cycles, but also we need to know which seasonal cycle is dominating in a seasonal time series. In other words, if a seasonal model with only one single seasonal cycle is used, which seasonal cycle should be used in the model which produces the smallest forecasting error?

In the literature, empirical autocorrelation function (ACF) plots and empirical power spectrum plots have been used as the major instruments in detecting seasonal cycles in a time series (Box *et al*., 1994; Taylor, 2008). The autocorrelation function plot has a peak at the lag which coincides with the seasonal cycle length. The power spectrum plot also has a peak at the time of the seasonal cycle (Stoica and Moses, 2005). In Figure 3, a seasonal time series of sold items is presented. Figure 4 is the empirical autocorrelation function plot of this time series and Figure 5 is the empirical power spectrum plot of this time series.

The ACF plot indicates that this time series has a weekly seasonalty. As the ACF plot has a peak at lag of 364, it has an annual seasonality as well. The power spectrum reveals about the same information. However, which seasonality is stronger, the weekly or the annual? Both the ACF plot and the power spectrum plots indicate the weekly seasonality is stronger. Is this true? To answer this question, let us take a totally different approach-Let us conduct an experiment using the simplest forecast model $\hat{x} = x_{t-p}$ where p is the seasonal cycle length. With the actual data in the figure, it turns out that when p = 7, the mean absolute percent error (MAPE) is 53.33%, and when p = 364 the forecasting error is 47.92% (forecasting horizon starts from 08/15/2010 to 05/28/2011). This is equivalent to about 10% of reduction in forecasting errors. Therefore, this simple experiment indicates that the stronger seasonality is the annual, not the weekly. Hence, both ACF and the power spectrum plots fail to capture this. What is worse is that, if we must pick a few of the strongest seasonal cycle lengths to be

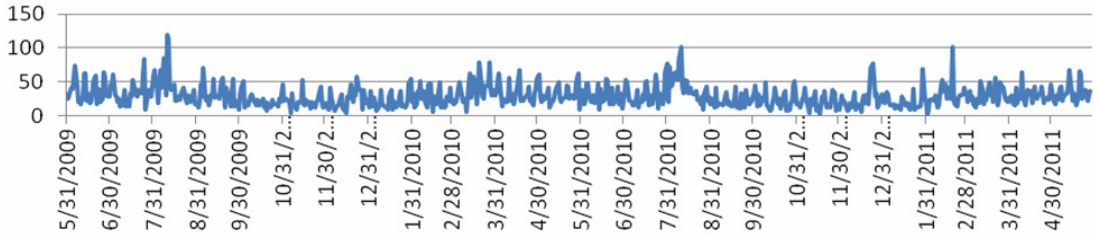

Figure 3. Sold items data which indicate strong seasonality.

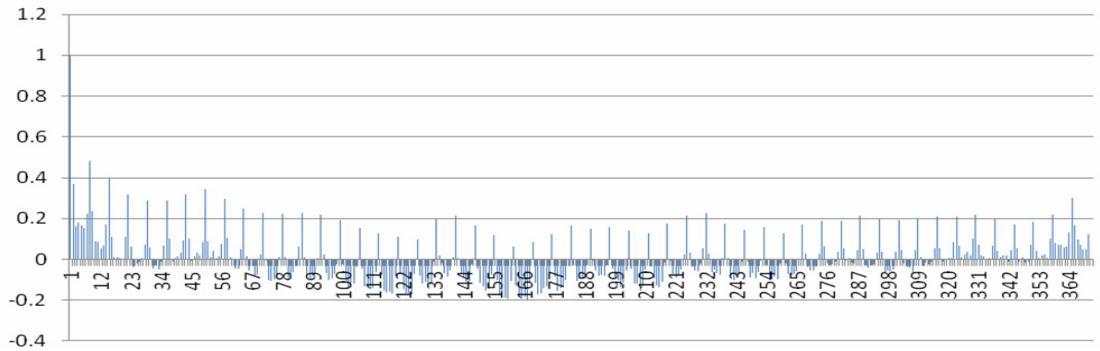

Figure 4. Autocorrelation Plot of sold items data.

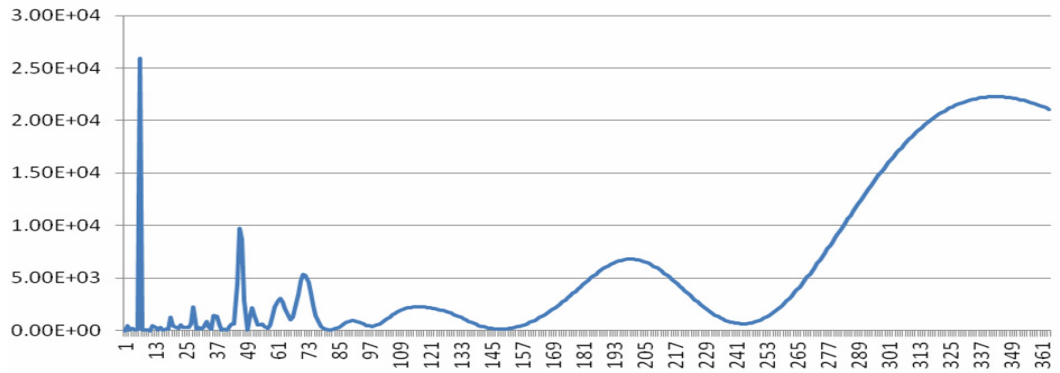

Figure 5. Power spectrum of sold items data.

used in the model, say three, both plots will produce different results. With the ACF plot, if we pick the lags corresponding to the highest peaks as the seasonal cycle lengths, cycle lengths of 7, 14 and 49 will be picked. With the power spectrum plot, cycle lengths of 7, 343, and 46 will be picked.

In the past, the author spent nearly two years in designing various automated algorithms in cooperating the ACF plot and power spectrum plot to detect the strongest seasonality in seasonal time series. The main idea of those algorithms was to detect the peaks of the ACF and the power spectrum plots as the seasonal cycle lengths. The original time series or the differenced time series were used in the research. In spite of a significant amount of time and effort invested, the author never achieved satisfactory results, or succeeded in obtaining what was expected. For some time series, the ACF plot and the power spectrum plot couldn't even pick the right seasonal cycle length that could be easily detected with our naked eyes. In despair, the author has turned to a different road map and has therefore developed a new instrument called the Average Power Function of Noise (APFN) to detect multiple seasonalities. The following is the formula used in calculating APFN for a stochastic process (Song, 2011):

$$APFN(\tau) = \lim_{T\to\infty} E\left[\int_{-T}^{T} \frac{\left(x(t+\tau)-x(t)\right)^2}{2T} dt\right]$$

or

$$APFN(p) = \lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N}\left(x(t+p)-x(t)\right)^2$$

for a time series.

From Song (2011), we know that on the APFN plot, a seasonal cycle of the time series will create repeating local minima. The seasonal cycle that has the global minimum on the APFN plot will indicate the dominating seasonality in the data. The APFN plot for the time series in Figure 3 is given in Figure 6. It can be seen that the APFN plot has multiple local minima, located at time lags of 364, 7 and 371. This indicates that 364, 7, and 371 are the strongest seasonal cycle lengths. Obviously, only APFN has found the strongest seasonality in the data. Since late 2008, the author has been using APFN exclusively in detecting multiple seasonalities, and feels very confident that APFN is the right apparatus in detecting multiple seasonalities in a seasonal time series.

Figure 7 illustrates a different time series. The time series indicates a very strong seasonality. The AFC plot in Figure 8 indicates that the strongest seasonality is weekly, and the power spectrum in Figure 9 also indicates so. To verify this finding, we use this simple forecast model $\hat{x} = x_{t-p}$ to produce forecasts and calculate the forecast errors where p is the seasonal cycle length. When p = 7, the forecasting error (MAPE) is 45.25%, and when p = 364 the forecasting error is 37.10%. This is about 18% of reduction in forecasting errors. By com-

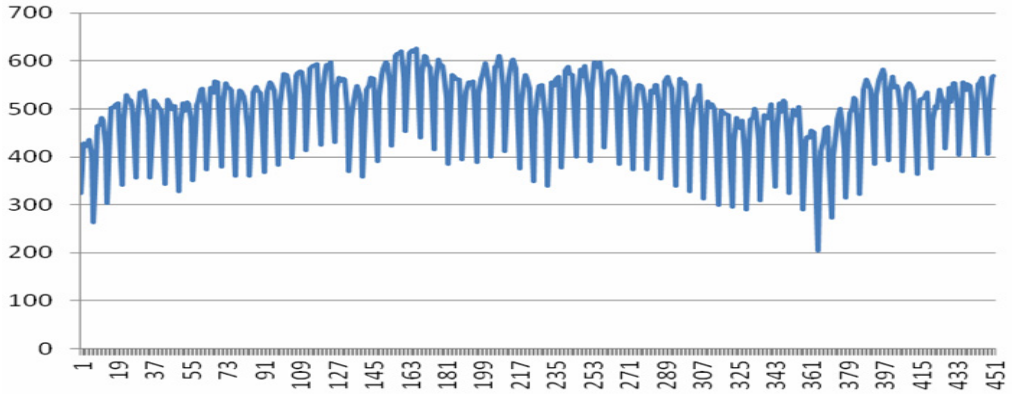

Figure 6. Average Power Function of Noise of sold items data.

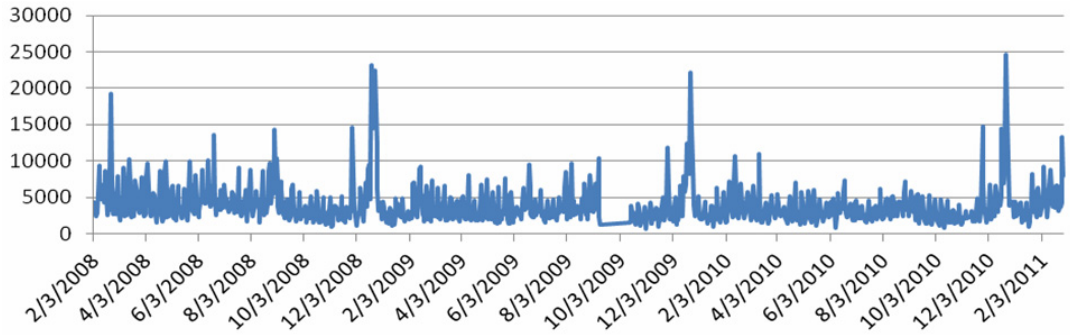

Figure 7. Total store sales indicating strong seasonality.

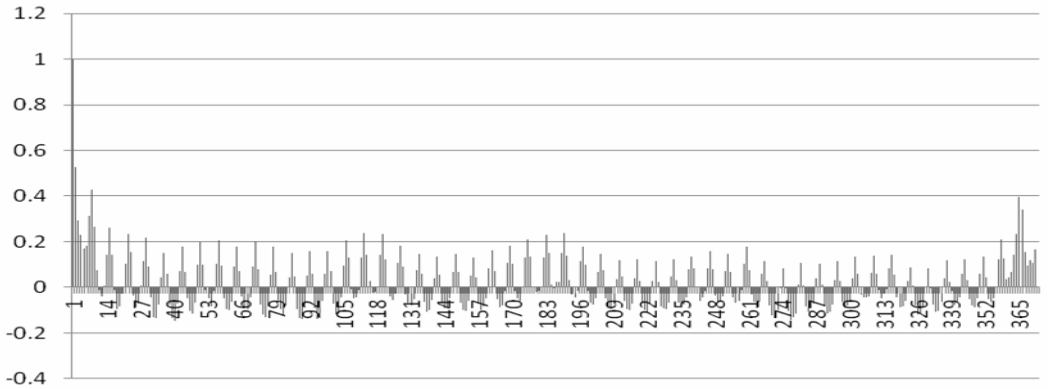

Figure 8. Autocorrelation function plot indicating strong weekly and annual seasonality.

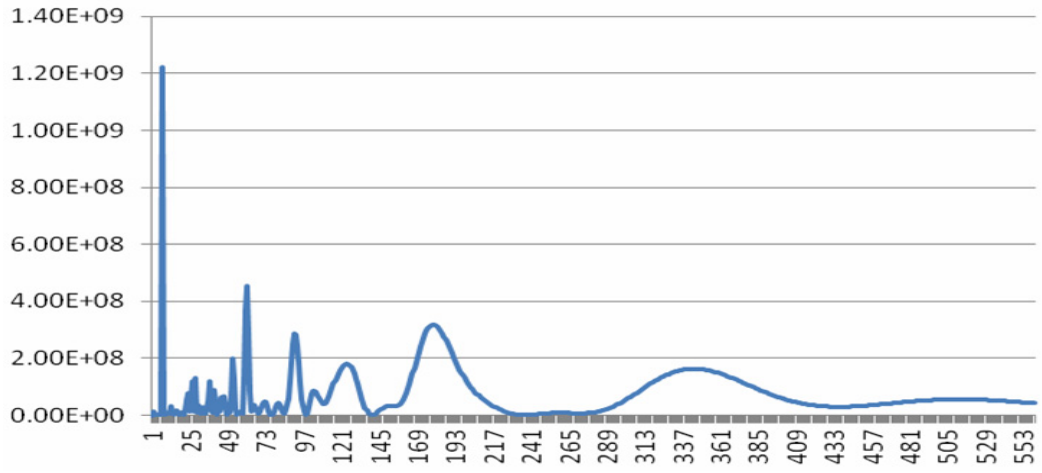

Figure 9. Power spectrum plot indicating strong weekly seasonality.

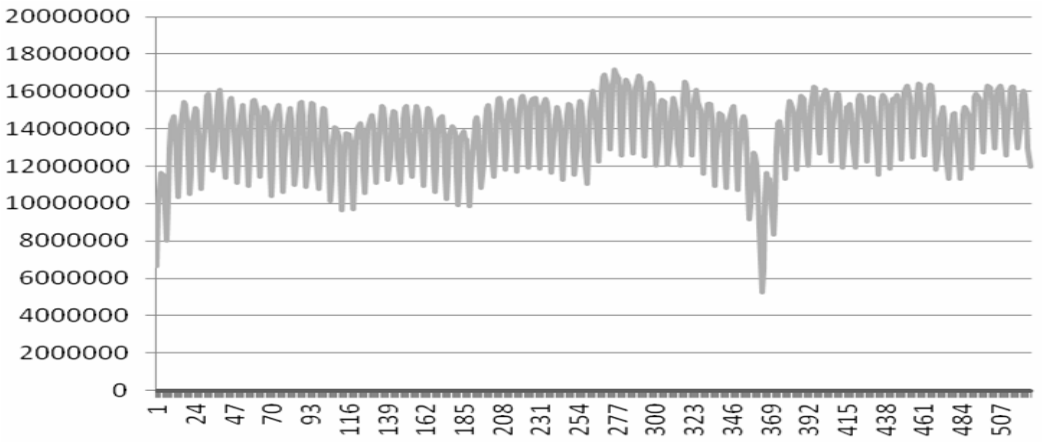

Figure 10. APFN indicating strong annual seasonality.

paring the forecast errors, we see that the strongest seasonality is annual. However, both the ACF plot and the power spectrum plot indicate weekly as the strongest. In addition, if we must pick 3 seasonal cycle lengths by sorting the values of the peaks, on the power spectrum the annual seasonality is missed. However, the APFN plot in Figure 10 indicates that the annual seasonality is the strongest.

Both examples indicate that as an instrument of detecting multiple seasonal cycle lengths, APFN is superior to both the ACF and the power spectrum plots.

3.3 Roles of Modeling Optimization Are Conditional

Modeling optimization is the process of finding the optimal model parameters once the structure of the model is determined. It is used in almost all the commercial products of forecasting software. How important is modeling optimization in achieving good forecasts? How much contribution does this process make to achieving good forecasts?

My conclusion is that modeling optimization is not a critical step in forecasting and is less important than finding the proper structure of the model. Its merit is conditional simply because if the data are not processed properly, the model produced is a model of the noisy data, and this type of model will be very sensitive to the change in data due to the noise. In forecasting, there are numerous cases where a simple naïve model can produce better forecast produced by an optimized model. There are two possible causes for this. First, the model structure might have been inappropriately chosen. For example, for a seasonal time series the seasonal cycle lengths might have been incorrectly determined in the model. Second, the data are very noisy and the model is determined based on the noisy data.

To demonstrate that optimization is not critical in forecasting, we will compare the forecasts using one seasonal model, $\hat{x}_t = a_1 x_{t-7} + a_2 x_{t-14}$, for 106 retail sales time series which show very strong weekly seasonality in two different scenarios and in two different versions. The forecasts will be created for the period from April 17, 2008 to Feb. 24, 2010. In the first version, both the model parameters a_1 and a_2 are chosen to be 0.5, and in

Table 1. Comparison between two versions of model and two scenarios of data of 106 time series

	Version 1 model parameters = 0.5	Version 2 model parameters determined algorithmically
Scenario 1 Noise not suppressed	10.31%	10.38%
Scenario 2 Noise suppressed	8.23%	8.23%

the second version, both parameters are optimized dynamically to minimize the modeling error. In the first scenario, the original data without any preprocessing are used while in the second scenario noise, event effects and any seasonal components other than weekly are suppressed by means of a simple algorithm. Table 1 lists the overall forecasting errors (MAPE) for these two versions and two scenarios. Surprisingly, the results in the first scenario indicate that no advantages can be obtained from purely optimizing the forecasting model. The reason is that the data contain other information such as event effects and annual seasonality that are not incorporated in the model. In this case, optimizing model will make the model more sensitive to the event information and other unused seasonality, and will make the forecasting results inferior to the non-optimization based model. In both versions, no advantages can be seen from the forecasting errors achieved. Only after the data has been preprocessed are the forecasting errors reduced. Therefore, **the roles of modeling optimization are conditional on properly preprocessing of the data. Modeling optimization becomes helpful only when all the information is incorporated in the model.**

3.4 It Is Critical to Suppress Noise in the Data

Noise exists ubiquitously in retail sales data. We need to reduce the level of noise in the data before using them in modeling and forecasting. This might be a controversial practice. However, it helps significantly in forecasting. What is noise in our real life data? If not in the mathematical context, we can interpret noise as any changes of patterns in the data that cannot be explained with a good cause or reason. For example, if there is a trend in the data for which we cannot find a good cause or explanation, it should be regarded as noise, and the data should be modified properly. If the data are not modified, then a model will capture this "trend," and would extrapolate this trend to generate forecast and cause large forecasting errors. The same can occur to levels of data. Missing data should be regarded as noise because it changes the seasonal patterns. Suppressing noise in the data will help forecasting significantly.

The illustration in the last section has demonstrated the significance of noise suppression in forecasting. In the illustration there, with 106 time series of total store sales which show very strong weekly seasonality, an optimized additive seasonal average model, $\hat{x}_t = a_1 x_{t-7} + a_2 x_{t-14}$, was used in forecasting for a horizon from April 17, 2008 to Feb. 24, 2010. When the original data without any noise suppression were used in modeling and forecasting, the overall forecasting error for all the time series was 10.38%. When a simple algorithm was used to reduce the noise in the data and then the processed data were used in forecasting, the overall forecasting MAPE errors for all the time series became 8.23%. In this example, with noise suppressed in the data, we could achieve about 20% of improvement. See Table 1 for a comparison. For the same data sets, with the simple average of the last two weeks data as the forecast, the improvement could be also 20% due to noise suppression. Although the improvement can vary from data to data, and from model to model, in our experience **properly suppressing noise in data is always helpful to improve forecast accuracy assuming the model could capture all the seasonality information of the data**.

Figures 11 and 12 illustrate a different example where the time series of store sales contains both sea-

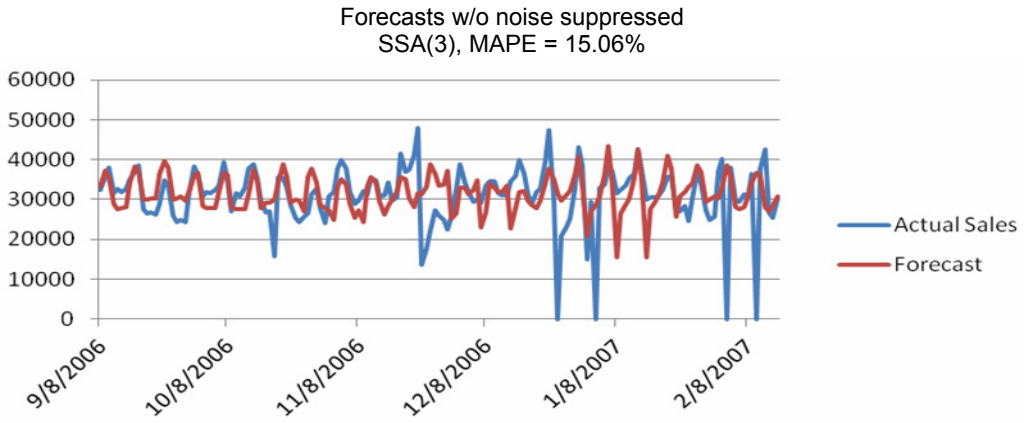

Figure 11. Forecasts without reducing noise level in the data.

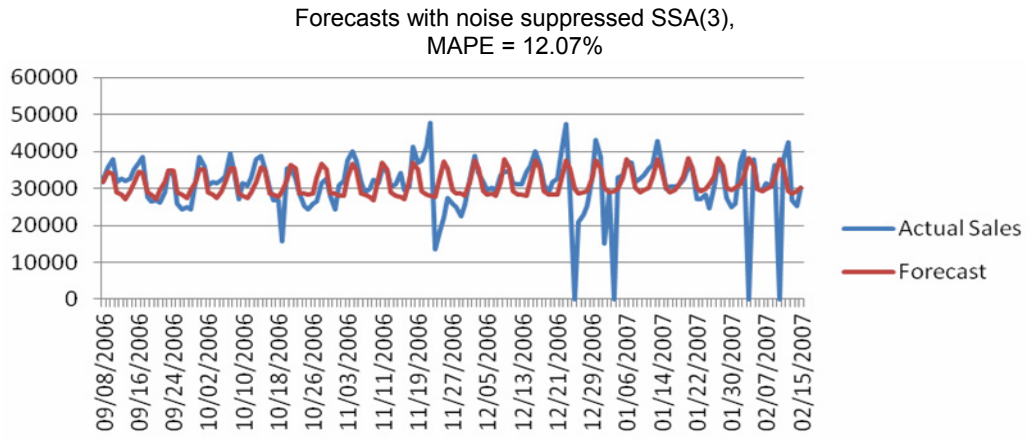

Figure 12. Forecasts with simple algorithm to reduce noise in data.

sonal components and noise. First, we generate the forecast using the average of the same day sales in the last 3 weeks for the current week with no attention to reduce the noise level in the data. The achieved forecasting error (MAPE) is 15.06%. Figure 11 illustrates the actual data and the forecasts obtained. Next, we use a simple algorithm to reduce the noise level in the data, and use the same forecasting algorithm to generate forecast for the current week. Figure 12 displays the actual data and the forecast. This time, the forecasting error (MAPE) becomes 12.07%, about 19.8% of improvement over 15.06% simply due to the suppression of noise in the data. Therefore, the importance of noise suppression in forecasting cannot be overemphasized.

3.5 What Is Benchmark in Retail Forecasting?

I have been often asked by customers about benchmarking in retail forecast accuracy. Customers want to know what kind of forecasts accuracy can be obtained for their data. They simply want to have a number to be pleased with. I doubt there is such a thing as benchmark of accuracy in retail sales forecast. There has been a debate on benchmarking of forecasting accuracy in the literature (Kolassa, 2008; Hoover, 2008; McCarthy *et al*., 2008). Some authors published benchmark results in forecasting which means that to evaluate your own forecast, their benchmark results should be used as a reference. To me, this is quite misleading. What I have found is that forecasting accuracy depends on many different factors. For example, it depends on the level of noise in the data, it depends on how much you understand your data, it depends on the type of the business where the data are from, it depends on the events that affect the business, it depends on the modeling techniques, it depends on the noise suppression techniques applied, and it depends on the event handling techniques. Just name a few. In retail sales forecast, what I have observed is that even for the same metric, for example customer counts, forecasting errors could differ significantly from stores to stores within the same chain, and the same occurs from metrics to metrics of even the same store. Not to mention different metrics from different chain stores. It is impossible to use a single number or a few numbers as the references for all forecasts in the retail industry. What is most important is to adapt the best forecasting process and the best forecasting practice that will be discussed later.

3.6 Aggregating Data to a Higher Level Could Do More Harm than Good

Aggregating data to a higher level (either spatial or temporal) is a common practice in forecasting (Hubrich, 2005; Chu and Zhang, 2003). It is hoped that when forecasting at a higher level in the data hierarchy, a better forecast will be obtained. For example, to forecast weekly sales people may aggregate the daily data to the weekly and then model and forecast the weekly data. My lesson is that if more seasonal information is gained from aggregating data to a higher level, it might be beneficial to do so. If more seasonal information exists at a lower level than at a higher level, it might be beneficial to forecast at the lower level and aggregate the forecast to a higher level.

Figure 13 displays a weekly time series of sales data that is aggregated from the daily time series data. For comparison purposes, we will generate the weekly sales forecast with two different approaches. In the first approach, we generate daily sales forecast and then aggregate the daily forecasts to obtain weekly sales forecast. In the second approach, we aggregate daily sales data to obtain weekly sales data, and then model the weekly sales data and generate weekly sales forecast with an annual seasonal model. In both approaches, models are optimized using history data dynamically. With the first approach, we can obtain a forecast error (MAPE) of 4.53%, and with the second approach we can obtain a forecast error (MAPE) of 5.06%. From the plots in Figure 13, the weekly data exhibits only annual seasonality although

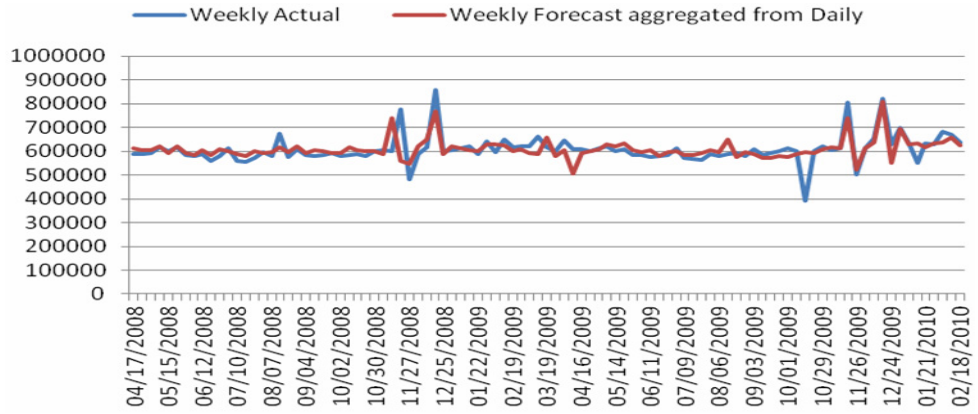

Figure 13. Weekly sales and weekly forecasts aggregated from daily forecast with MAPE of 4.53%.

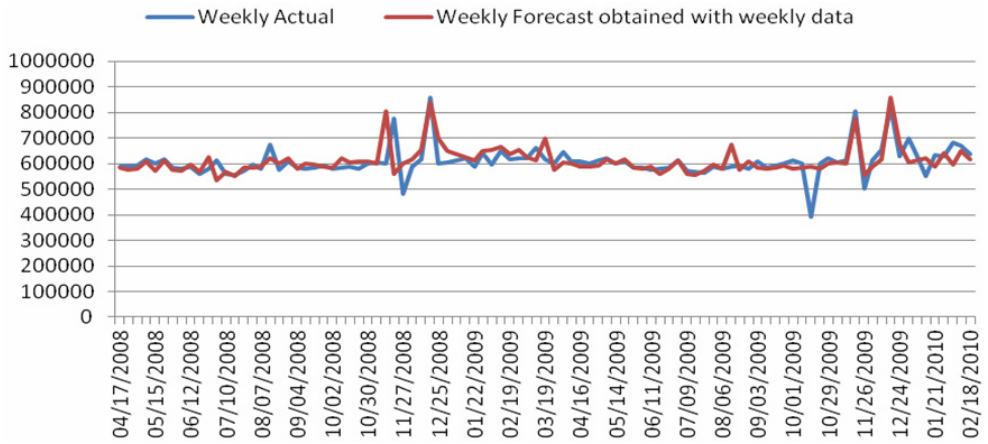

Figure 14. Weekly actual and weekly forecast obtained using weekly data with MAPE of 5.06%, Annual Seasonal Model.

the annual seasonality is not strong. However, the daily sales data exhibits a very good weekly seasonality and an annual seasonality. When forecasting at the daily level, we can model the time series using both the weekly and the annual seasonality whereas when forecasting at the weekly level, only the annual seasonality is used. This example demonstrates that **when forecasting, it is beneficial to model and forecast at the level of data where stronger and more seasonal information can be collected.**

To further verify this finding, we repeat this process using the daily traffic data from 306 stores from a different sector of retail industry with two different approaches. The range of the history data is from Feb. 01, 2004 to the end of 2008 so that the forecasts could cover a 3-year period. Again in both approaches models are optimized using history data dynamically. In the first approach, we forecast the daily traffic data and then aggregate the daily forecasts to get the weekly forecasts, and an overall MAPE of 10.29% is obtained. In the second approach, we forecast the weekly traffic using the aggregated weekly traffic data. An overall MAPE of 14.52% was obtained using an annual seasonal model and an overall MAPE of 11.69% using an autoregressive model. The forecasts produced by the annual seasonal model are 41% worse than the forecasts produced by forecasting at the daily level and then aggregating the daily forecasting to get the weekly forecast. The forecasts produced by the autoregressive model at the weekly level are about 13% worse than the forecasts produced by the daily approach. The results here are consistent with some literature (Dunn *et al*., 1976) and tell us again that **when forecasting, try to forecast at the level where maximum amount of information can be gathered from the data**. For example, at the daily level, we have weekly seasonality, monthly seasonality and even annual seasonality. In addition, we could collect information on events and holidays. However, at the weekly level, we will lose the weekly seasonality information that is usually very strong; we could lose holiday events and other event information. All these factors lead to inferior forecasts at the aggregated weekly level.

3.7 Not All Events Are Helpful in Improving Forecasts

I often heard complaints from customers that events they incorporated in their forecast did not help improve

their forecasts. To understand the cause, we need to know that there are two types of events: business events and forecasting events. Business events are those associated with any planned business activities, such as promotions, advertisements, holidays, etc., and forecasting events are those by incorporating which forecasting accuracy may be improved. Not all business events are forecasting events, and at the same time not all forecasting events are business events. This is something not everyone is aware of. To be qualified as a forecasting event, a business event should satisfy two conditions: (1). Event effects must be stable and repeatable over time, and (2). The timing of its occurrence is deterministic or predictable. Unfortunately, not many people comprehend these two properties. To some people, any business events, when incorporated in forecasting, should improve their forecast accuracy. But, this does not happen all the time! Indeed, most business events will boost business when measured at an aggregated level. But, boosting business is different from enhancing forecast accuracy. Before incorporating an event in forecasting, we need to check if these two conditions can be satisfied. If not, it is very likely that the event will not help improve the forecast. Often the times, incorporating such an event may worsen the forecast accuracy.

4. CHALLENGES ENCOUNTERED IN RETAIL FORECAST

During the past years, I have encountered numerous challenges in retail sales forecast. Some of them have been solved satisfactorily and others are still in work-in-process. In this section, I want to share some of the challenges in a hope to inspire some thoughts from readers.

4.1 Estimation of Event Effects

With the data model introduced in Section 2,

$$x_t = s_t + e_f + \varepsilon_t$$

if we take the expected value of both sides of the equation, we would get this equation:

$$\bar{x}_t = s_t + \bar{e}_t \tag{3}$$

where we assume the noise process is white and s_t is deterministic. Hence, the mean of the event effect is the difference between the mean value of the observation and the deterministic seasonal component, *i.e.*, $\bar{e}_t = \bar{x}_t - s_t$.

To obtain the event effects $\bar{e}_t$, we need to estimate the seasonal component s_t. This can be achieved by solving the following maximum likelihood function problem (Schervish, 1997):

$$\max_{S_1, S_2, \cdots, S_P} f_{X|S}(X \mid S) = f_{X|S}(X \mid \hat{S}) \tag{4}$$

where $f_{X|S}(X \mid S)$ is the conditional density function of the data given any seasonal component estimates $\hat{S}$ and X contains no event information. When the noise is Gaussian, (4) can reduce to the least square estimator obtained by solving the following minimization problem:

$$\min_{S_1, S_2, \cdots, S_P} \sum_{k=1}^{P} \sum_{t} \left(x_{t-k} - s_k\right)^2 \tag{5}$$

where P is the largest seasonal cycle length of the time series, and x_{t-k} does not contain any event effect information. Both (4) and (5) are easy to solve.

Once the seasonal components are obtained, within the window of an event, the event effects can be computed using the following formula:

$$\bar{e}_t = \bar{x}_t - s_t \tag{6}$$

The real challenge here is how to estimate event effects when there are multiple concurrent (or compound) events occurring over the same time period or having overlaps in time. When multiple events occur at the same time, it is observed that the overall effect is not simply a superposition of each individual event's effect when occurring separately. For example, Promotion A brings in 100 extra sold items when executed alone, and Promotion B brings in 200 extra sold items when run separately. However, the number of extra sold items brought in by executing both Promotions A and B at the same time is not simply 300. There are interactions (cannibalism) between different concurrent events. Therefore, in the case of multiple events, we should estimate each individual event effects, and then estimate the interactions between them.

For simplicity, we will consider only two events. Suppose event A and event B occur on the same day. Let e_A be the event effect when event A occurs alone, and e_B be the event effect when event B occurs alone, and e_{AB} be the interactive event effect between A and B. Then, the following formula models the overall effects when A and B occur at the same time:

$$e = e_A + e_B - e_{AB} \tag{7}$$

Thus, the data model of (1) can be written as

$$x = s + e_A + e_B - e_{AB} + \varepsilon \tag{8}$$

where the time index t is omitted.

Both e_A and e_B can be estimated using (6). To estimate the interactive effect e_{AB}, we may use the following formula:

$$\bar{e}_{AB} = s + \bar{e}_A + \bar{e}_B - \bar{x} \tag{9}$$

When there are more than two concurrent events, their interactions can be estimated with a similar method.

Nevertheless, another challenge in handling multiple concurrent event effects is the data collection. In general, concurrent multiple events tend to occur much less frequently than a single event does. For this reason, the sample size of the data containing the event information might be too small to be useful. To overcome this difficult, an alternative approach is to define a new event and use the event effect of one of the events as an approximation to the newly defined event. Note that when events occur, the level of noise tends to be higher than when there are no events at all. Therefore, it is acceptable to use a single event as an approximation to the concurrent event.

4.2 Event Detection

How do we detect an event? Events are not simply outliers defined in statistics. If events were simply the outliers as discussed in statistics, then many forecasting events could not be detected because they may not satisfy the definitions statistically. However, they are significant events in business and in forecast.

To detect an event, people may check the ratio of the value of a particular data point to an average value. If the percentage is beyond a threshold, the data is treated as an event. One of the drawbacks of this appraoach is that the seasonal compoments are not filtered out from the event effects. In addition, using percentage as the critieron makes the detecting algorithm too sensitive to changes in data, thus too sensitive to noise, and causes frequent false alarms. According to our data model, to detect an event, we need to estimate the seasonal components first, and then calculate the mean and the standard devition of the residuals. If the value of a particular residual is outside of a confidence interval, then we should treat it as an event. Specifically, let $\hat{e}_t = x_t - s_t$ be the residual between the observation data and the seasonal component. Let $\hat{\bar{e}}_t$ be the sample mean of the residuals, and $\hat{\sigma}_t$ be the sample standard deviation of $\hat{e}_t$. Then, when $\hat{e}_t$ is outside the confidence interval defined by $[\hat{\bar{e}}_t - k\hat{\sigma}_t, \hat{\bar{e}}_t + k\hat{\sigma}_t]$, an event is detected. In this paradigm, the value of k should be determined and tested by checking against some known events such as holidays to ensure that all well-known events can be detected and to ensure that at the same time the number of false alarms is minimized.

It's interesting to note that the value of k when detecting events at the store level is different from when detecting events at the chain level due to the change in noise levels.

4.3 Event Analysis

We know that business events are not the same as forecasting events, and not all business events are forecasting events. Before incorporating an event in forecasting, we need to conduct quantitative analysis of events. One of the major tasks of event analysis is to understand if events have stable and repeating patterns.

4.3.1 Similarity Analysis

A good forecasting event should have a stable and repeating temporal pattern, and the timing of occurance is deterministic. Thus, we can measure the similiarties between different occurences of an event in the history. Similarity can be measured in terms of correlation coefficient of the event effects of two different occurences. It has been found that if the correlation coefficient of the event effects between the two different occurences is greater than 0.97, then usually the event effects of the last occurence could be close to the next one. If the correlation coefficient is less than 0.97, in general the event effect of the last occurrence may not be close to the next one. Therefore, if such events are incorporated in the forecast, the forecast accuracy is hard to improve.

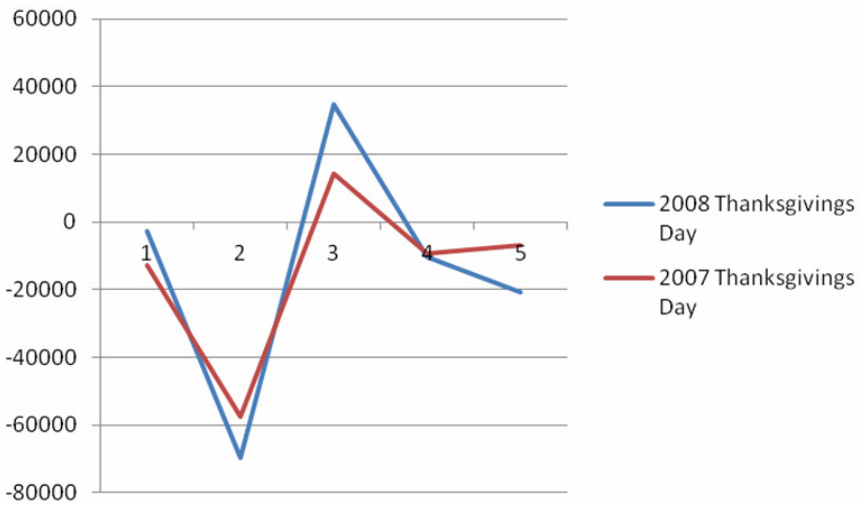

Figure 15. Event effects of Thanksgivings Day in 2007 and 2008 for total store sales.

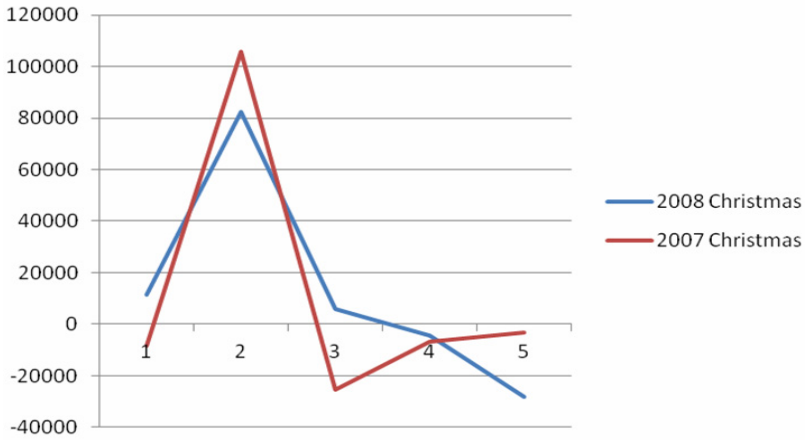

Figure 16. Event effects of Christmas in 2007 and 2008 for total store sales.

Figures 15 and 16 illustrate the event effects of Thanksgivings Day and Christmas for the total sales of a store in 2007 and 2008, respectively. Both events have a deterministic timing of occurances. But, the shapes are different in 2007 and in 2008. The shapes of the Thanksgivings Day are more similar to each other (0.956) than those of Christmas (0.88). Therefore, the Thanksgivings Day event is a better forecasting event than the Christmas event.

Other measures of similarity are also possible. For example, e^{-MAPE} is a good measure of similarities between two occurrences where MAPE is calculated with one occurrence as the forecast and the other as the actual. When two instances of an event are identical, the MAPE value calculated as such will be equal to zero and therefore the value of e^{-MAPE} will be equal to 1. When the two instances are quite different, the MAPE value will be large and e^{-MAPE} will be small. Hence, e^{-MAPE} can be used as a measure of similarity between two instances of an event.

When working with events, it is always beneficial to analyse their effects at different levels of data aggregations, for example, at the store level and also at the chain level. Customers usually think that business events should help in forecast. This is true often at a higher level of data aggregation. For example, a promotion may not be a good forecasting event for a specific store. But, if this promotion is run for the entire chain, when aggregated at the chain level, this event could be a good forecating event. This is usually because when aggreated to the chain level, data has a lower level of noise than at the store level. This will in turn help to make this event possess a stable and repeatable temporal pattern at the chain level. Once we know the different behaviors of an event at both the chain level and at the store level, by calculating its similarities, we can have a better understanding of the event and its roles in improving forecast accuracy.

More often than not, customers would provide a list of events that could be used as forecasting events. A good practice is for each event to calculate the similarities between two consective occurences in the history, plot the similarities over time to spot visually any changes in the similarities over time. This plot will reveal whether a customer event is a good forecasting event, and if there are any changes in similarities we can discover when those changes occured.

4.3.2 Event Impact Analysis

Customers are used to measuring the effect of an event in terms of percentage. For example, they could tell us that an event boosts their sales by 30% on a particular day. We call this percentage the event impact, in order to distinguish it from event effects. Therefore, event impact analysis is also a progmatic aspect of event analysis.

The most important work in event impact analysis is to estimate the seasonal component of the observation data, s_t, and use it as the basis in calculating the event impact percentage given by $\frac{\bar{x}_t}{s_t}$ 100% where $\bar{x}_t$ is the mean value of the observation data and $\bar{x}_t$ is used to suppress noise in the data. Event impact analysis should be performed at different levels of aggregations. For example, it should be performed at the chain level and at the store level. By comparing the impacts of the same event at different levels, customers can collect information regarding the effectiveness of the event and the distribution of the effectiveness cross their organizations.

4.4 What Is the Best Forecast?

"What is the best forecast for my data?" This is the question our customers often ask and they want to know the lowest forecasting error to be achieved. Unfortunately, this is a question that has no definite answers for.

I can tell customers what the best of forecast I can get for their data. But, I cannot tell them what the theoretically minimal forecast errors in MAPE are for their

data, given the state of art of forecast techiques, albeit there are some statistical procedures that may help to assess if the forecast error achieved is the theoretical minimum. For instance, we can calculate the forecasting error and calculate the empirical autocorrelation function and the empirical partial correlation function. If the forecasting errors behave like a white noise, then we know that we have done our best and the forecasts obtained might be the best. However in my experience, I have seen cases where there were two competetive forecasts, one having a larger MAPE and a white noise forecasting error process, and the other having a smaller MAPE but a forecasting error process that is not white noise. It seems that the whiteness of the forecasting error process doesn't garranttee that the forecasting error measured in MAPE will be the smallest. Often the times, it is the events not captured and utilized that contribute the most to forecasting errors.

How do we know if the forecasts are the best achieved given the state of art of forecasting techniques? From my personal experience, I have found that as long as all the seasonality information has been utilized in modeling and forecasting, as long as the noise in the data has been suppressed properly, and as long as all the event information has been incorporated in forecasting, the possibility of improving the obtained forecast is very small, if not impossible. Hence, I would like to call it **the best forecasting practice to utilize all the seasonality information, suppress noise, extract all event information and use proper modeling techniques.** Therefore, to answer that question regarding the best forecasts, it is better to know whether the best forecasting practice has been implemented in our forecast. If the best forecasting practice has been exercized, then we have done our best, and the possibility of improving the obtained forecast is slim, and if it is not impossible the cost associated will be very high.

4.5 Estimating Forecasting Errors Prior To Running Forecasts

Very often, customers would like to know, before runnning the actual forecast, what the forecasting error could be like. Without seeing the actual data and running the forecasts, we cannot give them an exact answer. However, by checking the variation coefficient (VC) of the data, which is the ratio of the standard deviation to the mean of the data, we could provide an estimate of the achieved forecasting MAPE without running the actual forecast. I have found that the achieved forecasting error in terms of MAPE is positively correlated to the variation coefficient (VC) of the data. Usually, larger VC will lead to larger MAPE. The value of the correlation coefficient between MAPE and VC depends on the types of the data.

VC is a good indicator of the achieved forecasting MAPE. The following is my observation from extensive real life data and forecasting analysis. Usually, when VC is less than 0.5, the achieved forecating MAPE could range from 1/2 to 1/3 of the VC, and if the estimated MAPE values are achieved after running the actual forecast, usually the forecasts are very satifactory. When VC is larger than 0.5, the achieved forecasting MAPE tends to be larger and close to VC. This is just an approximated relation between the achieved MAPE and the VC of the data. Figure 17 displays a scatterplot of VC and MAPE for 106 time series. The scatter plot exhibits a possitive correlation between VC and MAPE with $r = 0.347$.

5. CONCLUDING REMARKS

In this paper, I have presented lessons that I have learned in retail sales forecast as well as challenges that I have encountered, which will be summarized below:

(1) To obtain good retail sales forecasts, it is critical to understand the data and the business that the data are from.
(2) Retail sales data are seasonal with multiple seasonalities. A model with a single seasonal cycle length may not be satisfactory.
(3) Conventional methodology may not work well in de-

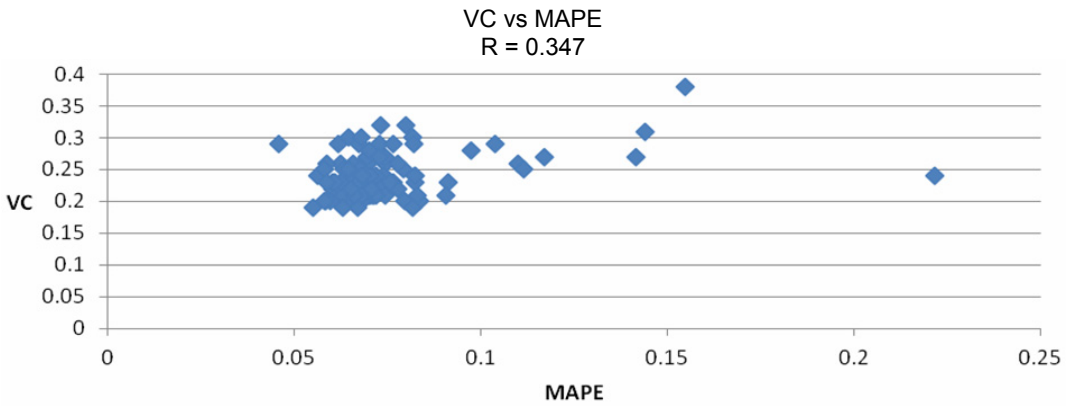

Figure 17. Scatter Plot of VC vs. MAPE.

tecting multiple seasonalities. New methodology is needed and needs refinement.
(4) Suppressing noise in data is important and helpful.
(5) Modeling optimization seems to have a marginal and limited role.
(6) The lower bound of forecasting error is still unknown. However, we can estimate the achieved forecasting error using VC.
(7) Estimating event effects is important in obtaining good forecasts. More efforts should be given to this area.
(8) It is critical to follow the best-forecast practice to obtain the best achievable forecast.

ACKNOWLEDGEMENTS

This paper is based on a plenary speech delivered at The International Conference on Computational Intelligence and Software Engineering (CiSE 2011), Dec. 9-11, 2011, Wuhan, China. The author wants to express his sincere thanks to the anonymous referees whose comments help improve the writing of the paper.

REFERENCES

Armstrong, J. S. (2001), Principles of Forecasting: A Handbook for Researchers and Practitioners, Springer.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994), Time Series Analysis: Forecasting and Control, Third Edition, Prentice-Hall, Inc. New Jersey.

Bunn, D. W. and Taylor, J. W. (2001), Setting accuracy targets for short-term judgmental sales forecasting, *International Journal of Forecasting*, **17**, 159-169.

Chen, F. L. and Ou, T. Y. (2011), Sales forecasting system based on Gray extreme learning machine with Taguchi method in retail industry, *Expert Systems with Applications*, **38**(3), 1336-1345.

Chu, C. W. and Zhang, P. G. (2003), A comparative study of linear and nonlinear models for aggregate retail sales forecasting, *International J. of Production Economics*, **86**(3), 217-231.

Dunn, D. M., William, W. H., and Dechaine, T. L. (1976), Aggregate versus subaggregate models in local area forecasting, *Journal of the American Statistical Association*, **71**(353), 68-71.

Hoover, J. (2008), Commentary on benchmarking, Foresight, **11**.

Hubrich, K. (2005), Forecasting euro area inflation: Does aggregating forecasts by HICP component improve forecast accuracy?, *International Journal of Forecasting*, **21**(1), 119-136.

Guo, Z. X., Wong, W. K., and Li, M. (2013), A multivariate intelligent decision-making model for retail sales forecasting, *Decision Support Systems*, **55**(1), 247-255.

Kolassa, S. (2008), Can we obtain valid benchmarks from published surveys of forecast accuracy? Foresight, **11**.

Lundholm, R., McVay, S. and Randall, T. (2010), Forecasting sales: A model and some evidence from the retail industry, *Unpublished working paper. University of British Columbia and University of Washington.*

McCarthy, T., Davis, D., Golicic, S., and Mentzer, J. (2008), Commentary on benchmarking, Foresight, **11**.

Ni, Y. and Fan, F. (2010), A two-stage dynamic sales forecasting model for the fashion retail, *Expert Systems with Applications*, **38**(3), 1529-136.

Schervish, M. J. (1997), Theory of Statistics, Springer-Verlag, New York.

Song, Q. (2011), Average Power Function of Noise and Its Applications in Seasonal Time Series Modeling and Forecasting, *American Journal of Operations Research*, **1**, 293-304.

Stoica, P. and Moses, R. (2005), Spectral Analysis of Signals, Pearson Prentice Hall, New Jersey.

Taylor, J. W. (2008), A comparison of univariate time series methods for forecasting intraday arrivals at a call center, *Management Science*, **54**, 253-265.