**Forecasting Retailer Product Sales at the SKU level in the presence of structural break**

Tao Huang, Robert Fildes, Didier Soopramanien

Abstract

Conventional forecasting methods for retailer product sales at the SKU level presume no change in the effectiveness of the price and promotional activities, which cause the model to be subject to structural break and potential forecast bias. In this study, we propose more effective methods to generate accurate forecasts for retailers by taking into this issue. Our proposed models outperform the industrial practice model and as well as the conventional models. We found that our models are especially effective for products which have less deep price cut or low coefficient of variation in product sales.

Selling point:

Easy to implement

Keywords:

Sales Forecasting, Marketing Analytics, Promotion

1. **Introduction**

Grocery retailers rely on accurate sales forecasts for their inventory management, scheduling, planning and strategical management {Petropoulos, 2014 #3}. Poor forecasts of product sales lead to out-of-stock conditions or overstock conditions. When the product is out-of-stock, retailers lose profit not only lose profits but also dissatisfy customers. In the long term, retailers may see customers switching to other retail chains and never return ([Corsten and Gruen 2003](#_ENREF_21)). Retailers may intentionally overstock, which however significantly raises inventory costs (e.g., capital cost, warehousing, and deterioration etc.) and reduces profits ([Cooper, Baron et al. 1999](#_ENREF_19)). In the year of 2014, Retailers in North American lost $634.1 billion due to out-of-stocks and $471.9 billion due to overstocks ([OrderDynamics 2015](#_ENREF_60)).

In practice, many retailers forecast their product sales at the SKU level using a two-stage ‘base-lift’ approach. The products are forecasted separately depending on whether or not the focal product is being promoted. The sales for the time period when the product is being promoted will be adjusted by brand/category managers based on their experience. In the literature, some studies focus on developing a procedure for managers to improve their judgments (e.g., [Goodwin 2002](#_ENREF_30), [Fildes, Nikolopoulos et al. 2008](#_ENREF_27), [Nikolopoulos 2010](#_ENREF_59)) or proposed models to determine the optimal judgment based on the data ([Cooper, Baron et al. 1999](#_ENREF_19), [Cooper and Giuffrida 2000](#_ENREF_20), [Trusov, Bodapati et al. 2006](#_ENREF_66)). Other studies directly generate the final forecasts of the product sales by proposing models with sophisticated structures and with additional information. For example, [Gür Ali, et al. (2009](#_ENREF_30)) proposed the regression tree model with a range of variables constructed from the sales, price, and promotion of the focal product in the previous time periods. [Huang, Fildes et al. (2014)](#_ENREF_35) proposed general-to-specific Autoregressive Distributed Lag models which incorporate the promotional information of not only the focal product but also of the competitive products within the same product category. [Ma, Fildes et al. (2016)](#_ENREF_48) further integrated the promotional information from the products not only within the same product categories but also across other related product categories.

One of the limitations of these studies is that they assume constant effects of the marketing activities (e.g., price reductions and promotions). In practice, evidence shows that the effect of price reductions and promotions tend to change overtime due to many influencing factors including the change of economic conditions, the change in consumer tastes, and media habits, and new competitor entry etc. which are normally not observable or measurable ([Wildt 1976](#_ENREF_72), [Wildt and Winer 1983](#_ENREF_73)). For example, customers may become more price/deal sensitive during an economic crunch. When a new competitor enters the market, it becomes more difficult to attract customers using the same budget for promotions and advertising. In reality, the German low-price retail chain Aldi has opened more than 400 stores in the United States just in the year of 2014, which leaves great pressures to other existing retail chains ([Loeb 2015](#_ENREF_47)).

Under such circumstance, conventional models which assume no change of the effect of the marketing variables may potentially be subject to structural break. A structural break is defined as a large change in the parameter coefficients of the model ([Allen and Fildes 2001](#_ENREF_3), [Armstrong 2001](#_ENREF_8)). The model which is subject to structural break may generate biased and thus less accurate forecasts. The issue of structural break have been historically addressed in the macroeconomics literature ([see Clements and Hendry 1994](#_ENREF_16)). In this study, we aim to propose more effective forecasting models which generate more accurate forecasts by mitigating the forecast bias due to the structural break. The research problem is challenging for the following reasons: 1) the product sales data at the disaggregated SKU level contains more variations compared to macroeconomic data. It is possible for the improved forecasting accuracy to submerged in the noise of the data. 2) the methods we propose in the study try to mitigate the forecast bias at a cost of increased forecasting error variance which also affects the forecasting accuracy. Therefore, in the retailing context, whether or not the mitigation of potential forecast bias due to structural break could lead to higher forecasting accuracy becomes an empirical question.

Our research is significant for the following contribution: 1) unlike any earlier study which contributes higher forecasting accuracy by incorporating additional information (e.g., the promotional information of other products from the same product category or other related product categories), our methods focus on how the information can be effectively utilized by taking into account the issue of structural break and forecast bias potentially due to the change of the effectiveness of the marketing activities. 2) Our methods have superior forecasting performance compared to conventional models which assume no change in the effectiveness of the marketing activities including prices and promotions. 3) Methodologically our study provides an evaluation of various methods which offers operational guidance to not only retailers as to how to produce more accurate forecasts but also manufacturers when competitive promotional information become not accessible. 4) we evaluate the forecasting performance of the models for 1834 SKUs from 30 product categories in 30 retail stores, which not only provide robust results but also allows us to further explore the relationship between the improved forecasting accuracy and the characteristics of the data series for each SKU. Our results suggest that it is more likely to obtained higher forecasting accuracy by implementing our methods for products with fewer deep price cuts or low coefficient of variation in product sales. The finding of the research can be used as a guidance for practitioners to select which forecasting method to use based on ex-ante analysis {Petropoulos, 2014 #3}. For example, they may decide whether or not implement the new forecasting method based on the potential benefit.

The remainder of the paper is arranged as follows: Section 2 summarizes previous research findings. Section 3 explains the issue of structural break and the subsequent forecast bias when conventional models overlook the change in the effectiveness of marketing activities. In section 4, we propose our models which may potentially improve the forecasting accuracy by mitigating the forecast bias due to structural breaks. Section 5 and section 6 explore the data and introduce the candidate models. Section 7 describes the design of the model evaluation. Section 8 summarizes and discusses the evaluation results. In Section 9, we draw conclusions. We make recommendations for both manufacturers and retailers, address research limitations, and highlight directions for future research.

1. **Literature review**

2.1 Forecasting retailer product sales at the SKU level

In practice, many retailers produce forecasts for their product sales at the SKU level using a two stage ‘base-lift’ approach. They initially generate the ‘baseline’ forecasts using simple univariate methods with the data excluding the time periods when the focal product is being promoted. Then they make adjustments to the baseline forecasts if there is an incoming promotional event in the future ([Fildes, Nikolopoulos et al. 2008](#_ENREF_27), [Fildes, Goodwin et al. 2009](#_ENREF_26)). The adjustments are usually made by brand/category managers and therefore subject to human cognitive bias ([Fildes, Goodwin et al. 2009](#_ENREF_26)). A stream of studies has been devoted to helping managers with their adjustment procedure ([Fildes and Goodwin 2007](#_ENREF_25), [Arenas, Pedregal et al. 2013](#_ENREF_7)). Some other studies try to improve the adjustment with model-based forecasting systems. For example, they may estimate the ‘lift’ effect by the promotional event based on historical information related to previous promotions, store/category features, and manufacturers etc. ([Cooper, Baron et al. 1999](#_ENREF_19), [Cooper and Giuffrida 2000](#_ENREF_20), [Trusov, Bodapati et al. 2006](#_ENREF_66)). One of the common limitation for these methods of two stages is that they generate forecasts separately depending on whether or not the focal product is being promoted. Therefore, the information when the focal product is being promoted are inevitably overlooked when forecasting the sales of the product when the product is not being promoted, and vice versa.

Previous studies have also proposed holistic methods to forecast the grocery product sales at the same time. [Gür Ali, SayIn et al. (2009)](#_ENREF_31) evaluated the forecasting performance of the variants of support vector machine models and regression tree models. Their models incorporated a range of constructed variables based on the promotional information of the focal product. Divakar et al. (2005) proposed the CHAN4CAST system with models of a dynamic regression structure to forecast brand sales for manufacturers/channels. [Huang, Fildes et al. (2014)](#_ENREF_35) proposed to forecast retailer product sales using the general-to-specific Autoregressive Distributed Lag (ADL) model with selected competitive promotional information within the same product category. The competitive promotional information was selected with variable selection methods (e.g., the stepwise selection and the LASSO algorithm) or constructed using principal component analysis. [Ma, Fildes et al. (2016)](#_ENREF_48) further integrated the promotional information not only from the same category of the focal product but also from other related categories. They resorted to Granger causality test to indicate the relevant product categories and then relied on the LASSO algorithm not only as a variable selection procedure but also as a model simplification strategy.

2.2 The changing effectiveness of marketing activities

Price reductions and promotions have a significant impact on product sales. For example, price reductions and promotions significantly increase short-term sales of the focal product ([Blattberg, Briesch et al. 1995](#_ENREF_10)). Price reductions and promotions have a positive (negative) impact on complementary (competitive) products ([Wittink, Addona et al. 1988](#_ENREF_75), [Dekimpe, Hanssens et al. 1999](#_ENREF_23), [Andrews, Currim et al. 2008](#_ENREF_6)). The impact of price reductions and promotions can be asymmetrical regarding different brands ([Wedel and Zhang 2004](#_ENREF_70)). Price reductions and promotions may either accelerate customers’ consumption or postpone their purchases if customers anticipate future promotional events ([Van Heerde, Gupta et al. 2003](#_ENREF_67), [Mace and Neslin 2004](#_ENREF_49)). The findings of these studies have been addressed by the most recent forecasting models (e.g., Gur Ali et al, 2009; Huang et al, 2014; Ma et al, 2016).

However, all the forecasting methods introduced in section 2.1 all presume invariant effectiveness of the marketing activities (e.g., price reductions, display promotions, and feature advertising). The potential change of the effectiveness of the marketing activities has been intensively explored in the literature (e.g. [Little 1966](#_ENREF_46), [Morrison 1966](#_ENREF_54), [Myers and Nicosia 1970](#_ENREF_57), [Myers 1971](#_ENREF_56), [Houston and Weiss 1975](#_ENREF_34), [Monroe and Guiltinan 1975](#_ENREF_53), [Moinpour, McCullough et al. 1976](#_ENREF_52), [Wildt 1976](#_ENREF_72), [Wichern and Jones 1977](#_ENREF_71), [Winer 1979](#_ENREF_74), [Mahajan, Bretschneider et al. 1980](#_ENREF_50)). The effectiveness of the marketing activities may change because of various reasons including the change in economic condition, legislation, consumer tastes, media habits, and advertising etc. ([Wildt 1976](#_ENREF_72), [Wildt and Winer 1983](#_ENREF_73)). The effectiveness of promotions may change during the different stages of the product lifecycle ([Mahajan, Bretschneider et al. 1980](#_ENREF_50)). Marketing theory suggests that the elasticity of the marketing activities will tend to vary at each stage of the product lifecycle ([Kotler 1997](#_ENREF_41)). The effectiveness of the marketing activities may change due to new competition. The introduction of new products (especially the store-owned brand) decrease promotional elasticities of premium national brands and increase promotional elasticities of the second tier national brands ([Nijs, Dekimpe et al. 2001](#_ENREF_58), [Van Heerde, Srinivasan et al. 2008](#_ENREF_68)). Evidence also suggests that intensive price reductions and promotions may decrease consumers responsiveness to these marketing activities by reducing their reference price ([Lattin and Bucklin 1989](#_ENREF_42), [Lichtenstein and Bearden 1989](#_ENREF_45), [Kalwani, Yim et al. 1990](#_ENREF_38), [Kalwani and Yim 1992](#_ENREF_37), [Foekens, Leeflang et al. 1999](#_ENREF_29), [Kopalle, Mela et al. 1999](#_ENREF_39), [Levy, Grewal et al. 2004](#_ENREF_44)). [Verhoef, Neslin et al. (2007)](#_ENREF_69) found the effectiveness of the marketing activities may be (negatively) affected by the introduction of a new distribution channel (e.g., online website).

Evidence also shows that consumers’ response to price reductions and promotions by competitive products may be affected by the introduction and termination of a loyalty program as consumers included in the loyalty program may find the promotions of alternative brands less or more attractive. ([Melnyk and Bijmolt 2007](#_ENREF_51)). [Leenheer, van Heerde et al. (2007)](#_ENREF_43).

Some studies tried to model the change of the effectiveness of the marketing activities. Foekens, S.H. Leeflang et al. ([1999](#_ENREF_29)) proposed an extended SCAN\*PRO model with time-varying parameters for the marketing activities including prices and promotions. In their model, the parameters of the marketing activities are modeled as functions of previous promotional information of the focal brand and other competitive brands. The model tried to capture how the effect of the marketing activities change over time, which may benefit managers with more effective budget allocation. Kopalle, Mela et al. ([1999](#_ENREF_40)) also proposed extensions of the SCAN\*PRO model to investigate the changing effectiveness of promotions on the baseline sales. These models, however, are all descriptive models and are not used in forecasting retailer product sales.

1. **The issue of structural break and potential forecast bias**

When the effectiveness of marketing activities on product sales change, as explained in the previous section, conventional models will be subject to structural break which is defined as large changes in the model’s parameters ([Allen and Fildes 2001](#_ENREF_3), [Armstrong 2001](#_ENREF_8)). The parameter estimates of the models then become the weighted average of the true parameters before and after the structural break. The forecasts generated by the model will potentially be biased and less accurate[[1]](#footnote-1). The impact of the structural break on the model’s forecasting performance has been addressed by many studies in the macroeconomics literature (e.g. [Cooper and Nelson 1975](#_ENREF_18), [Muellbauer 1994](#_ENREF_55), [Hendry 1995](#_ENREF_32), [Clements and Hendry 1999](#_ENREF_17), [Pesaran and Timmermann 2007](#_ENREF_61), [Castle, Doornik et al. 2008](#_ENREF_12)). [Pesaran and Timmermann (2005)](#_ENREF_63) provided an example using a simple regression model with a single structural break to indicate how structural break leads to forecast bias. In the retailing context, suppose that we have the sales and price information of the focal product from week 1 to week *T,* i.e.,, and we presume that the sales are driven by prices. We may assume that there is a structural break at week (where ), and the parameter of the price variable change from to after . In practice, this may be caused by the impact of a new brand entry, a new advertisement by other existing brands, or even the change of the temperature (e.g., for ice cream products) etc. which we cannot measure or do not have available data. Thus the unobservable real demand as follows:

where, is an indicator which equals to 1 before week and 0 otherwise. and are respectively the sales and the price of the product at week *t*. We assume that retailers do not change product price based on their short-term sales, thus we consider to be strictly exogenous[[2]](#footnote-2). and are the true parameters before and after the structural break at week . is the error term. when and when .

We may estimate a model with a functional form which is congruent with the demand (e.g., ) where the estimation window starts before the structural break, e.g., at week *m* . Thus the OLS estimate for the model based on the data [*m*,T] becomes:

where and are the matrices for the sales and price respectively for the time period from week *m* to week T. We assume no structural break after week T, and the true demand after week T remains as . Therefore, the *h*-step ahead forecast error at week *T*+*h* (with *m* as the starting observation of the estimation window) can be represented as:

where , is the matrix for the error term at week .

Therefore, the forecast bias at week can be represented as , which is unequal to zero as .

The structural break can further lead to less accurate forecasts. We illustrate this using a simulation example. We construct an artificial variable to represent price. The values of are 2.99 for most of the observations but occasionally reduced to 2.29 or 1.99[[3]](#footnote-3). i.e., or 2.29 or 1.99. We assume that we have the data of 100 weeks and we need to generate the forecast for the last 25 weeks based on the data of the first 75 weeks. The product sales to be determined by the price with two structural breaks at week 25 and week 50 respectively. Thus the unobserved real demand can be represented as follows:

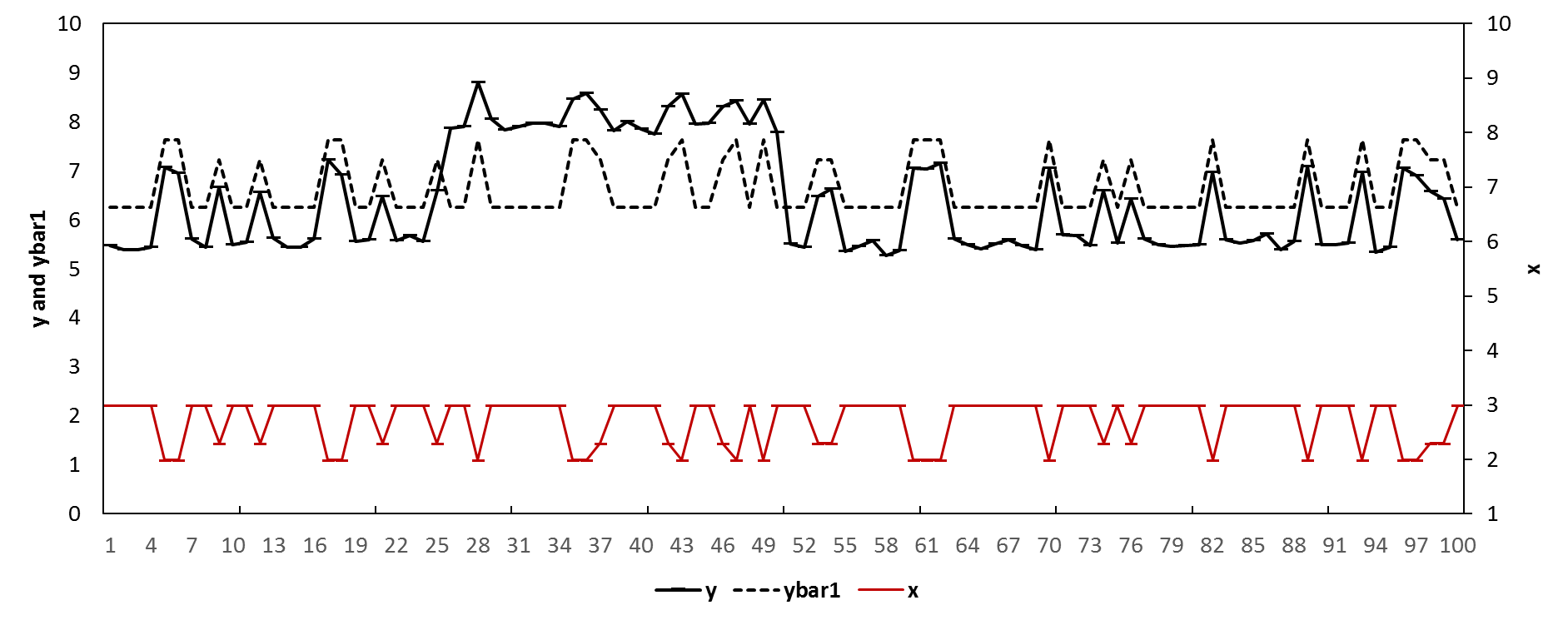
, , when

, , when

, , when

where represents the sales at week t, represents the price at week t. is the error term. Therefore, the real demand may indicate that product sales increase after week 25 but also become less responsive to temporary price reductions after the break. This is reasonable especially for those products in the growth to mature stage in their product life cycle. After week 50, product sales decrease but become more responsive to temporary price reductions, which may simulate the situation of an economic crunch. The sales and price data from week 1 to week 100 are depicted in Figure 1 by the solid black line and the solid red line respectively.

Figure 1. Simulated sales with a structural break: model with full data



In Figure 1, the blue area represents the time period before the first structural break (e.g., week [1,25]), the yellow area represents the time period after the second structural break until the forecast origin (e.g., week [51, 75]), the green area represents the period between the two structural breaks (e.g., [26, 50]), and the red area represents the forecast period (e.g., week [76, 100]). Suppose we have the data from week 1 to week 75 and we want to forecast the product sales from week 76 to week 100. We may estimate the model with the function form as using the data from week 1 to week 75 while overlooking the changes of the effectiveness of the price at week 25 and week 50. Under such circumstance, we will have estimates as the weighted average of the true parameters in the three periods (e.g., blue, green, and yellow). As suggested by the graph, this would over-predict the product sales for the time period from week 1 to week 25, under-predict the product sales for the time period from week 26 to week 50, over-predict the product sales for the time period from week 51 to week 70, and finally would produce downwards-biased out-of-sample forecasts for the time period from week 76 to week 100. The predictions/forecasts are represented by the black dashed line (as *ybar1*) in Figure 1. Table 1 shows the forecasting performance of this model regarding various error measures (e.g., with MAE= 0.715, MSE= 0.520, MAPE= 12.2%, and SMAPE= 11.5%).

Alternatively, we may estimate the model exclusively using the data from week 51 to week 75 and generate unbiased forecasts, if we know there are structural breaks at week 25 and week 50. These unbiased forecasts are represented by the black dashed (as *ybar2*) line in Figure 2[[4]](#footnote-4). Table 1 also shows the forecasting performance of the model which is estimated exclusively with post-break data (e.g., with MAE= 0.294, MSE= 0.184, MAPE= 5.0%, and SMAPE= 4.3%). The model exclusively estimated with post-break data outperforms the model estimated with all the data in the estimation sample. This indicates that the latter generates less accurate forecasts because of the structural break and forecast bias. However, in the retailing context. There are so many factors which may change the effectiveness of the price, as mentioned earlier in section 2.2. As a result, we are unable to observe the structural break at week 25 and week 50. For some occasions where structural breaks occur at the locations which are very close to the forecast origin, there will not be enough post-break observations to estimate the model.

Figure 2. Simulated sales with a structural break: model with post-break data

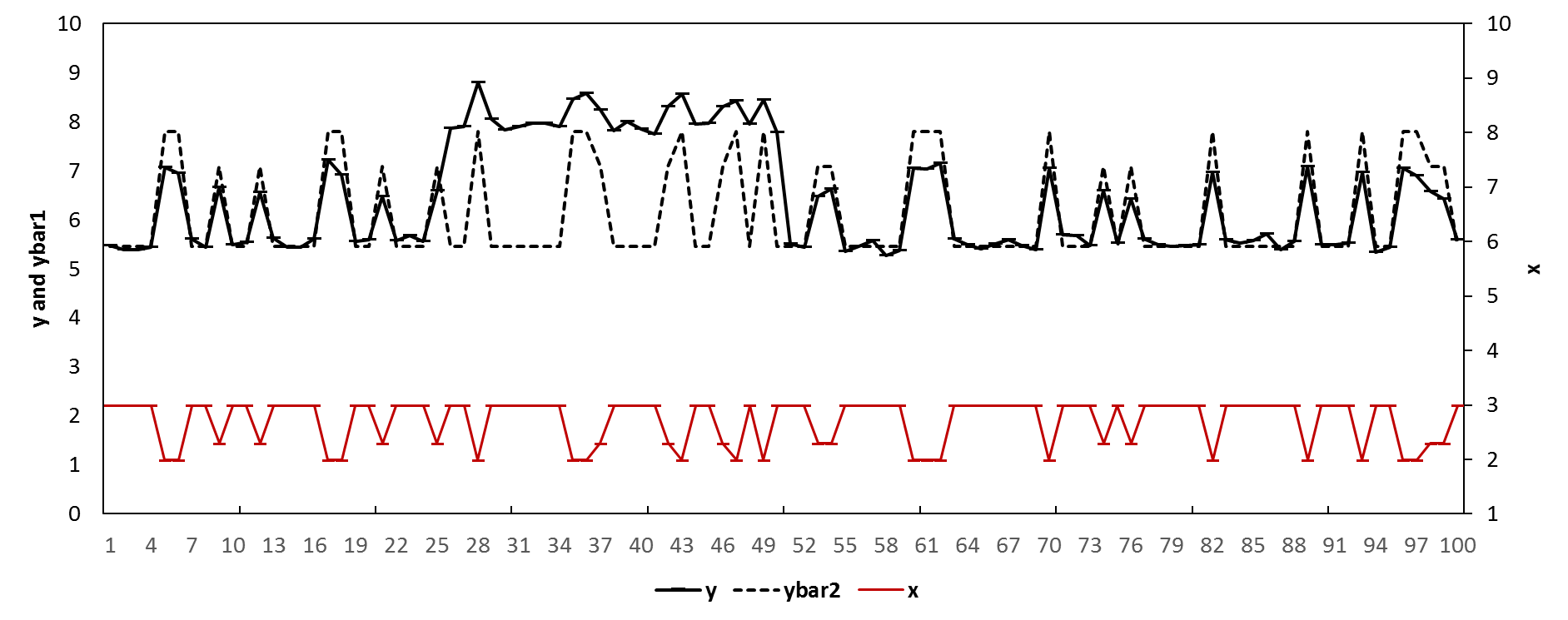


Table 1. The forecasting performance of different models in the simulation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | MAE | MSE | MAPE | SMAPE |
| In Figure 1: Model with full data | 0.715 | 0.520 | 12.2% | 11.5% |
| In Figure 2: Model with Post-break data only | 0.294 | 0.184 | 5.0% | 4.3% |
| In Figure 3: Model with intercept correction | 0.101 | 0.015 | 1.7% | 1.8% |
| In Figure 4: Model with estimation window combining | 0.647 | 0.425 | 11.0% | 10.5% |

1. **Dealing with structural breaks**

4.1 The Intercept Correction method

One of the methods to deal with the forecast bias caused by the structural break is the intercept correction method. The intercept correction method estimates the forecast bias and then offset the forecast bias (e.g., regime shifts) by specifying non-zero values for the model’s errors in the forecasting period ([Clements and Hendry 1994](#_ENREF_16), [Clements and Hendry 1999](#_ENREF_17), [Clark and McCracken 2007](#_ENREF_14)). The method may potentially improve the forecasting accuracy by mitigating the forecast bias through specify non-zero error terms but at the cost of inflated forecasting error variance ([Clements and Hendry 1999](#_ENREF_17)).

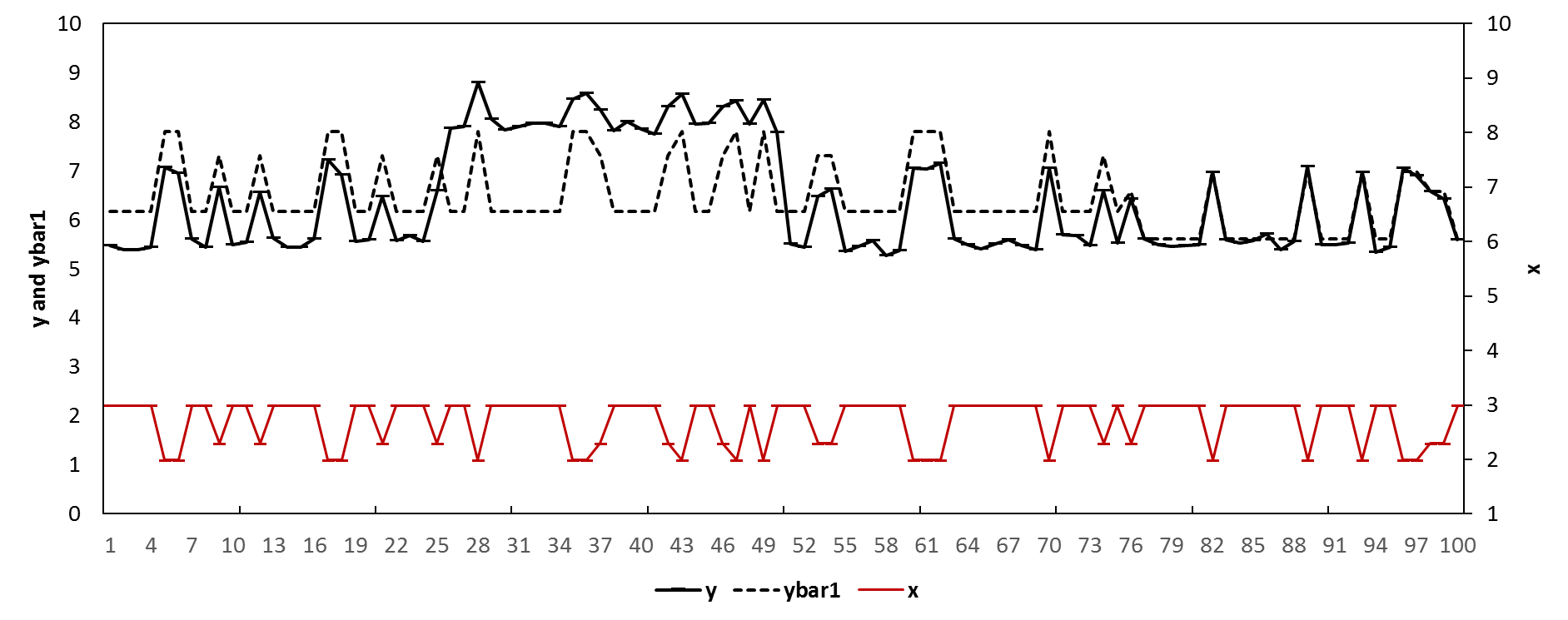
The IC method can be illustrated based on the same simulation example which is described in section 3. We assume that we have the congruent model as but with no prior knowledge related to the existence or the location of the structural break. A sequential [Chow (1960)](#_ENREF_13) test can be conducted for every observation in the whole estimation period[[5]](#footnote-5). The rejection of the null hypothesis of no structural break for any of the observation in the estimation sample would indicate that the model is subject to structural break (though it does not indicate how many structural breaks and their locations). Figure 3 shows the *p*-values of the sequential Chow test assuming there is one single structural break occurring at each week. The chow test rejects the null hypothesis of no structural break for some weeks (e.g., week 20) but fails to do so for some other weeks (e.g., week 35). Therefore, the results indicate that the model is subject to structural break with the estimation sample from week 1 to week 75, though do not indicate how many structural breaks and their locations[[6]](#footnote-6). More advanced statistic tests have been proposed to detect the locations of the structural breaks but they all need to assume additional priori knowledge such as the number of potential structural breaks ([Andrews 1993](#_ENREF_4), [Andrews and Ploberger 1994](#_ENREF_5), [Bai and Perron 2003](#_ENREF_9)).

Figure 3 P-values of the sequential Chow test



Therefore, we consider the model to be subject to structural break and we consider the forecasts as biased. We may estimate the forecast bias with different schemes. For example, as the predictive error at the forecast origin (i.e., , where *T* =75) or as the average value of an ad hoc number of predictive errors before the forecast origin (e.g. , where *i* can be arbitrarily chosen). In this example, we estimate the forecast bias as the average of the predictive errors for the most recent four observations in the estimation period. e.g., . We add the bias estimate back to the forecasts. e.g., , where represent the final ‘intercept corrected’ forecasts and are illustrated by the black dashed line (as *ybar3*) in Figure 4[[7]](#footnote-7). The results in Table 1 suggest that the intercept corrected model outperforms the original model (e.g., with MAE= 0.101, MSE= 0.015, MAPE= 1.7%, and SMAPE= 1.8%).

Figure 4. Simulated sales with a structural break: model with intercept correction



However, one of the limitations of the intercept correction method is that it heavily relies on the detection and the estimation of the forecasts bias. In the retailing context, the product sales at the SKU level have large variations, which may cause challenges to the estimation of the forecast bias. Also, the IC method mitigates the forecast bias by adding the estimated bias back to the forecasts but at a cost of inflated error variance of the forecasts (Clements and Hendry ([1999](#_ENREF_17)). Whether we can generate more accurate forecasts by implementing the IC method to conventional models for retailer product sales at the SKU level is an empirical question.

4.2 The Estimation Window Combining method

An alternative method to deal with the forecast bias due to structural break is the estimation window combining method ([Pesaran and Timmermann, 2005)](#_ENREF_63). The method does not estimate the forecast bias. It aims to take an effective trade-off between the forecast bias and the forecast error variance by combining the forecasts generated by the same model but with different estimation windows. In the simulation example in section 3, if we know the location of the structural break, we could estimate the model exclusively with the post-break data (e.g., the data from week 51 to week 75) and generate unbiased forecasts. In reality, we neither know whether structural breaks exist nor the location of the potential structural breaks. We may estimate the model with the most recent observations close to the forecast origin. It is less likely for the model to be subject to structural break as we keep *m* as large as possible (so that we discard more old data). When *m* by chance becomes larger than , the model will be estimated exclusively with post-break data and generate unbiased forecasts.

However, the reduction of the forecast bias comes with the cost of inflated forecasting error variance as we estimate the model with less information (e.g., the estimation sample is smaller). In the same example in section 3, the forecast error is:

The corresponding Mean Square Error (MSE), as one of the measures for the forecasting accuracy, at week can be represented as:

where

is interpreted as the squared forecast bias, and is interpreted as the efficiency term ( is the forecasting error variance). The change of the MSE for week when we include one more observation in the model estimation is:

where is the MSE for the model which is estimated with the data from week m-1 to week T. [Pesaran and Timmermann (2005)](#_ENREF_63) show that the bias term () (i.e., the change of the squared forecast bias) is always larger than or equal to zero (i.e., with one more observation before the structural break, the forecast will get more biased), but the sign for the efficiency term depends on the change of the error variance before and after the structural break (i.e., )). If (e.g., there are more post-break variations in the product sales which cannot be explained by the price variable), will be larger than or equal to , and the MSE may increase as both terms are non-negative. If (e.g., there are fewer variations in the product sales which cannot be explained by the price variable), may be smaller than or equal to . Under this condition, the MSE may either increase or decrease depending on how the non-negative squared bias term compares to the non-positive efficiency term. Therefore, when we exclude pre-break data in the model estimation, we may have either better or worse forecasting performance depending on the trade-off between the reduced forecast bias and the potentially inflated forecasting error variance.

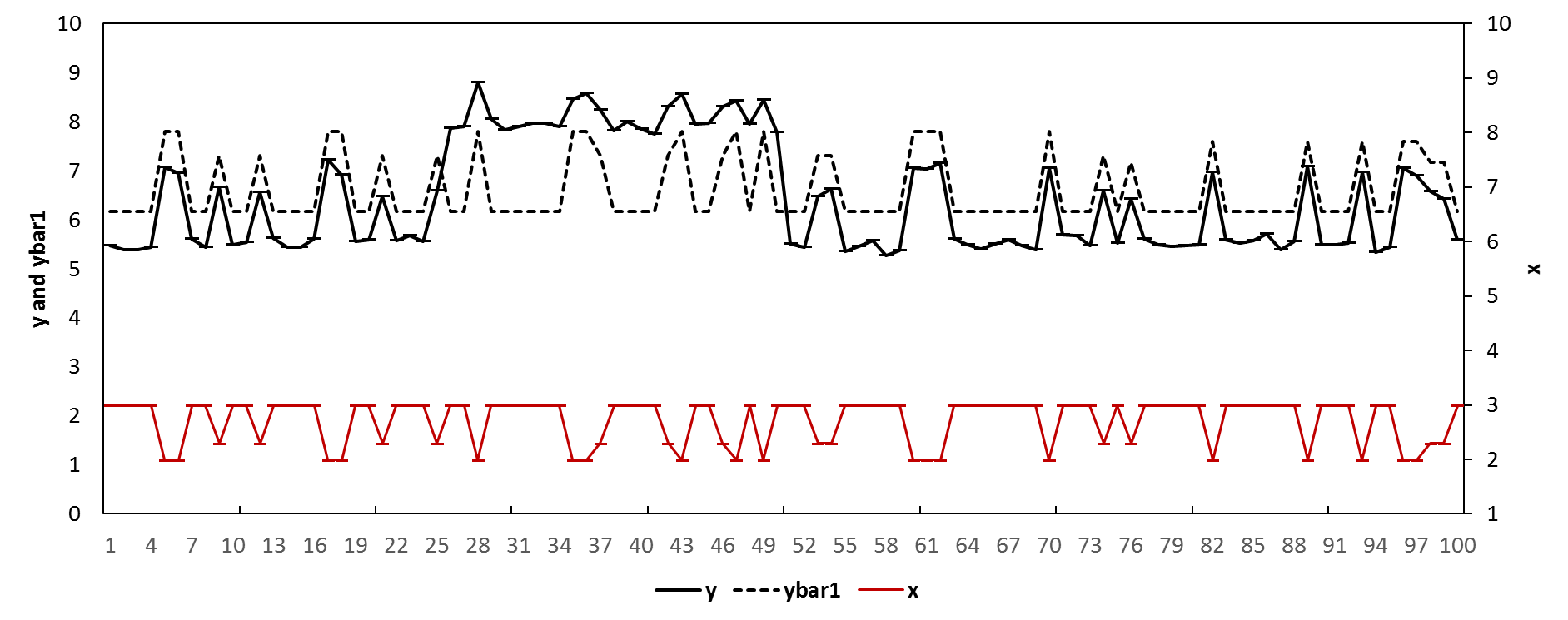
[Pesaran and Timmermann (2005)](#_ENREF_63) suggest combining the forecasts generated by the model of the same specification but estimated with different sample windows to potentially achieve an effective trade-off between the forecast bias the forecasting error variance. Evidence also show that combining forecasts leads to higher forecasting accuracy {Clemen, 1989 #745}{Jose, 2008 #746}. In this study, we combine these forecasts following a scheme of equal weights as combining forecasts with equal weights has been proved to be effective and easy to implement.([Clements and Hendry 1998](#_ENREF_15), [Fildes and Stekler 2002](#_ENREF_28), [Dekker, Van Donselaar et al. 2004](#_ENREF_24), [Pesaran, Schuermann et al. 2009](#_ENREF_62)). Specifically, we estimate the model using the most recent observations to generate the 1st set of the *h*-step-ahead forecast, e.g., , where represents the parameters estimated with the sample window . The value of is arbitrarily chosen given there are enough observations to estimate the model and there are enough variations for the explanatory variables. We may then add one more observation to the estimation window and generate the 2nd set of the *h*-step-ahead forecast, e.g., and so forth. We have the set of the *h*-step-ahead forecasts, e.g., . Finally, we combine these () sets of *h*-step-ahead forecasts with equal weights:

where represents the final forecasts.

The method can be illustrated with the same simulation example in section 3. Suppose there is an unknown structural break within the estimation period at week 31. We may estimate the model with different lengths of estimation windows and combine the corresponding forecasts. For example, we first estimate the model using the data from week 1 to week 75, and generate the forecasts for the period after week 75. We denote this set of forecasts as which are subject to the full bias. We then estimate the same model but using the data from week 2 to week 75, and generate forecasts for the period after week 75 and denote them as , and so forth. The forecasts such as will be less biased compared to but associated with inflated forecasting error variance because they were generated by models with less information. We choose to be 16 and thus we could combine sets of forecasts with equal weights. e.g., which are the final forecasts. The final forecasts are illustrated by the black dashed line in Figure 5. Table 1 shows that the forecasts are more accurate compared to the forecasts by the original model (e.g., 0.647 for MAE, 0.425 for MSE, 11.0% for MAPE, and 10.5% for SMAPE).

In this study, we explore the empirical question that whether we can generate more accurate forecasts by implementing the estimation window combining method to conventional models for retailer product sales at the SKU level.

Figure 5. Simulated sales with a structural break: model with estimation window combing



1. **The data**

In this study, we evaluate our models using the retail dataset made available by the Information Resources, Inc. (IRI) company. A description of the dataset can be found in [Bronnenberg, Kruger et al. (2008)](#_ENREF_11). The dataset contains weekly data at the SKU level including unit sales, price, features, and displays etc. over 30 product categories on a weekly basis. We conduct our evaluation based on 1834 SKU’s with positive movements for at least 90% of the time for 30 product categories from 30 stores. Table 2 shows the basic statistics for the selected SKU’s for each of the categories. The table indicates that some product categories (e.g., Carbonated Beverages and Hotdog) have much higher promotional intensity compared to others (e.g., Margarine/Butter and Mayonnaise). Figure 6 depicts the sales data for a typical SKU in the Beer category. The product has occasional price reductions and feature/display events where the product sales exhibit spikes accordingly. In Figure 6, the calendar events include Halloween, Thanksgiving, Christmas, New Year’s Day, President’s Day, Easter, Memorial Day, the 4th of July, and Labour Day. The promotional events include feature and display.

Table 2. Statistical description for the product in the categories

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Category | Price mean | Sales mean | Display percentage | Feature percentage | Number of SKU's | Category | Price mean | Sales mean | Display percentage | Feature percentage | Number of SKU's |
| Beer | 8.34 | 20.61 | 13.90% | 4.00% | 169 | Milk | 2.45 | 222.26 | 2.10% | 1.80% | 30 |
| Blades | 8.13 | 14.59 | 7.40% | 2.20% | 22 | Mustard & Ketchup | 2.06 | 64.51 | 5.30% | 0.90% | 22 |
| Carbonated Beverages | 2.1 | 113.59 | 26.80% | 15.60% | 82 | Paper towels | 3.66 | 68.07 | 4.00% | 3.60% | 3 |
| Cigarette | 22.28 | 22.22 | 0.00% | 0.80% | 202 | Peanut butter | 3.67 | 34.23 | 3.20% | 0.60% | 15 |
| Coffee | 5.19 | 14.5 | 5.20% | 2.90% | 86 | Photo | 7.18 | 9.19 | 4.60% | 5.10% | 13 |
| Cold cereal | 3.45 | 70.7 | 4.00% | 18.10% | 125 | Razors | 5.6 | 7.99 | 22.60% | 2.10% | 4 |
| Deodorant | 2.66 | 6.94 | 4.10% | 5.20% | 126 | Salty snacks | 2.28 | 50.89 | 6.70% | 5.00% | 100 |
| Face tissue | 2.12 | 75.82 | 0.30% | 11.70% | 6 | Shampoo | 3.51 | 9.89 | 12.80% | 7.10% | 70 |
| Frozen Dinner | 2.04 | 43.79 | 5.30% | 23.70% | 87 | Soup | 1.54 | 61.59 | 1.20% | 9.70% | 139 |
| Frozen pizza | 3.44 | 31.17 | 8.90% | 9.10% | 147 | Spaghetti sauce | 2.43 | 39.14 | 1.60% | 6.50% | 51 |
| Household Cleaner | 2.48 | 29.92 | 0.30% | 3.60% | 18 | Sugar substitutes | 2.76 | 14.49 | 0.10% | 1.40% | 20 |
| Hotdog | 3.99 | 68.63 | 13.20% | 15.60% | 35 | Toilet Tissue | 5.42 | 89.13 | 4.30% | 8.30% | 20 |
| Laundry Detergent | 8.78 | 28.94 | 2.30% | 8.80% | 57 | Toothbrush | 2.56 | 8.69 | 3.10% | 6.30% | 27 |
| Margarine/Butter | 1.95 | 71.36 | 0.10% | 6.30% | 36 | Toothpaste | 2.77 | 35.49 | 11.00% | 12.50% | 25 |
| Mayonnaise | 2.97 | 79.74 | 3.00% | 0.40% | 22 | Yogurt | 1.13 | 115.07 | 0.70% | 6.30% | 75 |

Figure 6. Unit sales, price (in USD), and promotional events (feature and display) for an SKU in the Beer category at the store level.



1. **Models**

In this study, we include the Base-lift method as the benchmark model. This method has been widely used by retailers to forecast product sales at the SKU level (Cooper et al., 1999; Huang et al., 2014). The method generates baseline forecasts using simple exponential smoothing method with data when there is no promotion for the focal product. It then makes adjustments for any incoming promotional event based on the lift effect by the most recent promotional event. The method can be represented as follows:

where is the final forecast for week *t* by the base-lift method, is the baseline forecast for week , is the actual sales for the previous week when the focal product is not being promoted, is the parameter for the simple exponential smoothing model. The adjustment is calculated as the increased sales by the most recent promotional event for the focal product.

We also include the autoregressive distributed lag (ADL) models which were introduced by Huang et al. (2014). The ADL model captures the dynamic effects of price reductions and promotional events with parsimonious specifications. We firstly construct the ADL model with the dynamic terms of the price and promotional information of the focal product (we refer this model as the ADL-own model thereafter). We initially construct the following model:

where:

is the log sales of the focal product at week

is the term for the determinist trend which captures any potential steady change during the estimation period {Song, 2003 #1216}.

is the log price of the focal product at week

is the promotional index of the focal product at week

is the four-week-dummy variable  
 is the dummy variable for the calendar event at week . The dummy variable represents the week of the calendar event when , , and the week before the event if . takes the values from 1 to 9 representing all the calendar events *[[8]](#footnote-8)*

are the parameters  
 is the error term and we assume

is the order of the lags and is set to as 2.

The initial model is then reduced by the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm following Ma et al. (2016). The LASSO algorithm is a regularization algorithm which put a constraint to the sum of the absolute values of all the parameter coefficients of the initial ADL model ([Tibshirani 1996](#_ENREF_65)):

where

is the vector of observations on the dependent variable  
 represents explanatory variables included in the initial ADL model

*u* is the identically distributed random error

is the vector of the parameter coefficients  
*N* is the number of parameters  
 is the shrinkage factor which equals to the sum of all the parameter coefficients.

Therefore, some of the explanatory variables will be dropped if their coefficients are pushed towards zeros by the constraint. The model reduction process is controlled by a shrinkage factor based on 10-fold cross-validation following Ma et al. (2016). The ADL-own model can be illustrated by Figure 7a. For example, an ADL model is initially constructed with the dynamics of the price and the promotional variables of the focal product (i.e., own predictors) and then reduced by the LASSO algorithm. The resulted model will be used to generate the final forecasts.

We also include the ADL model with promotional information not only of the focal products but also of other competitive products within the same product category following Huang et al. (2014). We refer the model as the ADL-intra model in this study. The model is initially specified as follows:

where

is the log price of competitive product at week .

is the promotional index of competitive product at week .

is the number of competitive price variables in the product category.

is the number of competitive promotional variables in the product category.

The initial model is then reduced by the LASSO procedure and the retained explanatory variables are combined with the explanatory variables retained by the ADL-own model to generate the final forecasts for the same SKU. It would be less likely for the final combined model to miss important and relevant variables though at a potential cost of efficiency. The ADL-intra model can be illustrated by Figure 7b. For example, we conduct the LASSO selection procedure respectively for the models with only own predictors and the models with both own predictors and competitive predictors. The final model will have all the retained explanatory variables.

Figure 7c. An illustration for the ADL-intra-IC models

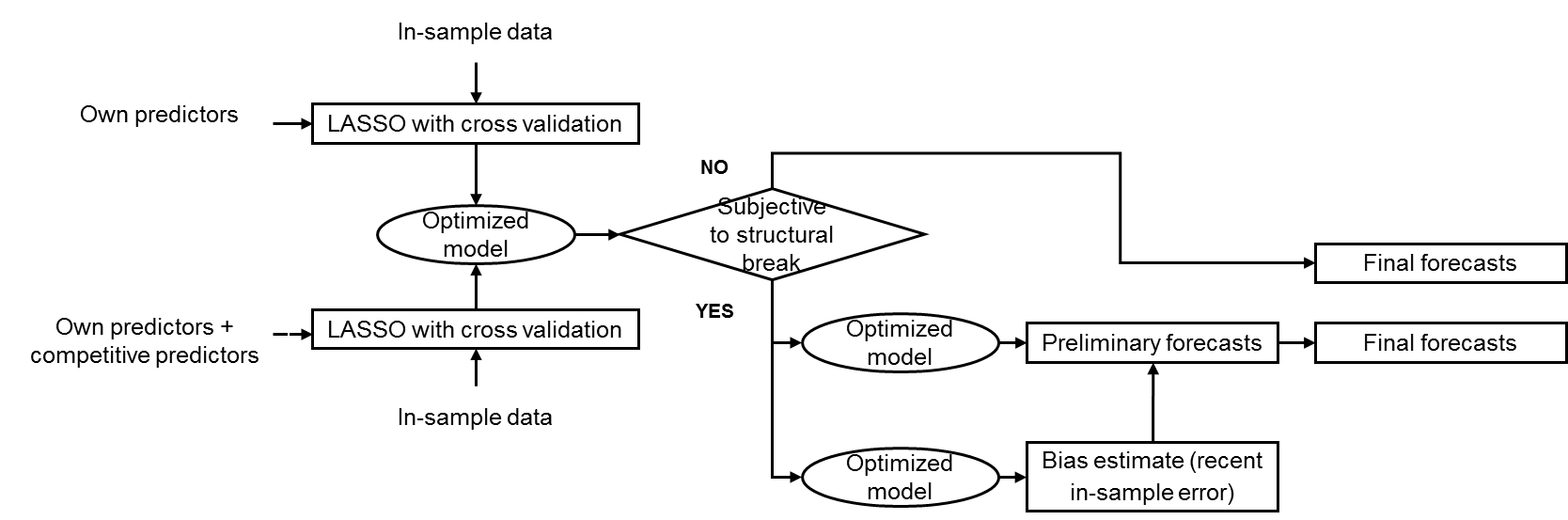
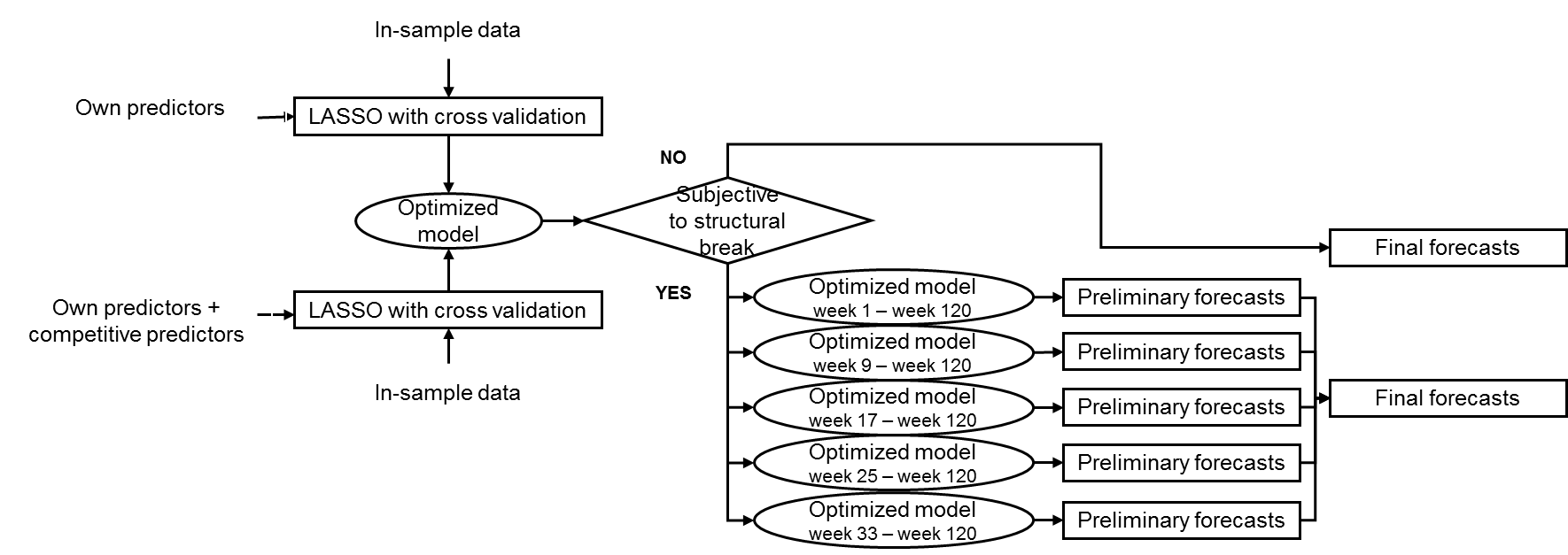


Figure 7d. An illustration for the ADL-intra-EWC models



The ADL-own model and the ADL-intra model both ignore the potential change in the effectiveness of the marketing activities. As a result, they may potentially be subject to structural break and generate biased and less accurate forecasts. In this study, we propose the following models which may potentially generate more accurate forecasts by taking into account this issue: 1) the ADL-intra-IC method; 2) the ADL-intra-EWC method; 3) the ADL-own-IC method; 4) the ADL-own-EWC method. The modeling procedure for the models are illustrated in Figure 7c, 7d, 7e, and 7f respectively. For the ADL-intra-IC model, we first construct the ADL-intra model as illustrated in Figure 8b and then conduct the sequential Chow test based on all the observations in the estimation sample. If the test fails the reject the null hypothesis of no structural break, the forecasts by the ADL-intra model will be the final forecasts. Otherwise, we estimate the forecast bias as the average value of the four most recent error terms. For the ADL-intra-EWC model, we also construct the ADL-intra model and then conduct the sequential Chow test. If the test results indicate no structural break, the forecasts by the ADL-intra model will be the final forecasts. Otherwise we re-estimate the ADL-intra model with five different estimation windows (e.g., if our initial estimation window is week 1 to week 160, then we re-estimate the model with the time period from week 1 to week 160, week 5 to week 160, week 9 to week 160, and so forth, until week 41 to week 160) and generate five sets of forecasts. The final forecasts will be the equal weighted average of these five sets of forecasts. The ADL-own-IC model and the ADL-own-EWC model are also built in the same way expect that the two models do not contain the retained competitive price and promotional information. Compared to Huang et al. (2014) where the general-to-specific models were specified manually, all the models we propose in this study are specified automatically using the LASSO procedure in SAS 9.4. The automation of the statistical forecasting procedure becomes essential as typically grocery retailers have more than 30,000 SKUs (Cooper et al. 1999; {Petropoulos, 2014 #3}).

1. **The experimental design**

In this study, we evaluate the forecasting performance of the models with rolling origins ([Tashman 2000](#_ENREF_64)). The models are re-specified and re-estimated for each rolling event. For example, we specify the models with the time period of [1,160], [9, 168], [17, 176], [25,184], and [33,192]. For each rolling event, we generate one to week-ahead forecasts, where is 1, 4, and 12, to approximate the situation retailers face in practice. We use the actual values of the exogenous variables (e.g., price, promotion, or calendar events etc.) and the forecasts of the lagged dependent variables when they are not available.

The models are evaluated using four error measures: the Mean Absolute Percentage Error (MAPE), the symmetric Mean Absolute Percentage Error (sMAPE), the Mean Absolute Scaled Error (MASE) proposed by [Hyndman and Koehler (2006)](#_ENREF_36), and the Relative Average Mean Absolute Error (RelAvgMAE) proposed by [Davydenko and Fildes (2013)](#_ENREF_22). These error measures approximate the loss function of the retailer from different aspects. The error measures for SKUs and rolling events based on forecast horizon of 1 to (i.e. , , and =1, 4 and 12) are as follows:

where and are respectively the actual value and forecast value of the forecast period for data series based on the rolling event. We add one-half mean squared error to the final forecasts before we transform the log values to levels (Cooper et al.,1999). We apply is the total number of observations in the full estimation window. and are the Mean Absolute Errors for the candidate model and the benchmark model for data series *s*, with forecast horizon of *H*, for the rolling event.

1. **Results and discussion**

8.1 Overall forecasting results

Table 3 shows the forecasting performance of the candidate models. The Base-lift model generates the least accurate forecasts for almost all the scenarios. The ADL-own model gets outperformed by the ADL-intra model for all the scenarios, which highlights the value of competitive promotional information as suggested by Huang et al. (2014). In the results, the ADL-own-EWC model outperforms the ADL-own model for all the scenarios. The ADL-own-IC model outperforms the ADL-own model for short and moderate forecast horizons (e.g., when *h*=1 and *h*=4) but was losing advantages for long forecast horizon (e.g., when *h*=12)[[9]](#footnote-9). In practice, competitive promotional information may not always be available especially for manufacturers ([Ali and Boylan 2011](#_ENREF_2)). Under such circumstance, the ADL-own-EWC model and the ADL-own-IC model can be implemented with the information of the focal product by taking into account the issue of structural break and potential forecast bias. The ADL-intra-EWC model outperforms the ADL-intra model for all the scenarios. The ADL-intra-IC model has superior forecasting performance for 1-week-ahead and 1 to 4-week-ahead forecast horizons but gets outperformed for 1 to 12-week-ahead forecast horizon. Overall, the ADL-intra-EWC model and the ADL-intra-IC model generate the most accurate forecasts.

Table 3. The forecasting performance of candidate models for all forecast period for different forecast horizons

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| All forecast period, Forecast horizon= 12 | | | | | | | | |
| Model/measure | MAPE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank |
| Base-lift | 71.17% | 7 | 47.55% | 7 | 0.783 | 7 | 1.1458 | 7 |
| ADL-own | 67.99% | 5 | 41.16% | 5 | 0.700 | 5 | 1.0000 | 5 |
| ADL-intra | 67.00% | 2 | 40.84% | 2 | 0.696 | 3 | 0.9925 | 2 |
| ADL-own-EWC | 67.86% | 4 | 40.95% | 3 | 0.696 | 2 | 0.9932 | 3 |
| ADL-intra-EWC | 66.87% | 1 | 40.64% | 1 | 0.691 | 1 | 0.9856 | 1 |
| ADL-own-IC | 68.34% | 6 | 41.39% | 6 | 0.704 | 6 | 1.0031 | 6 |
| ADL-intra-IC | 67.27% | 3 | 41.05% | 4 | 0.700 | 4 | 0.9958 | 4 |
| All forecast period, Forecast horizon= 4 | | | | | | | | |
| Model/measure | MAPE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank |
| Base-lift | 66.41% | 7 | 45.94% | 7 | 0.751 | 7 | 1.0944 | 7 |
| ADL-own | 65.73% | 6 | 40.53% | 6 | 0.688 | 6 | 1.0000 | 6 |
| ADL-intra | 64.88% | 3 | 40.18% | 3 | 0.680 | 3 | 0.9907 | 3 |
| ADL-own-EWC | 65.54% | 5 | 40.31% | 4 | 0.683 | 5 | 0.9925 | 4 |
| ADL-intra-EWC | 64.70% | 2 | 39.96% | 1 | 0.676 | 1 | 0.9834 | 1 |
| ADL-own-IC | 65.15% | 4 | 40.47% | 5 | 0.682 | 4 | 0.9955 | 5 |
| ADL-intra-IC | 64.46% | 1 | 40.12% | 2 | 0.677 | 2 | 0.9866 | 2 |
| All forecast period, Forecast horizon= 1 | | | | | | | | |
| Model/measure | MAPE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank |
| Base-lift | 61.82% | 6 | 44.06% | 7 | 0.726 | 7 | 1.0062 | 7 |
| ADL-own | 62.04% | 7 | 39.53% | 6 | 0.666 | 6 | 1.0000 | 6 |
| ADL-intra | 61.44% | 4 | 39.22% | 4 | 0.657 | 4 | 0.9974 | 5 |
| ADL-own-EWC | 61.71% | 5 | 39.33% | 5 | 0.663 | 5 | 0.9954 | 4 |
| ADL-intra-EWC | 60.93% | 3 | 39.02% | 2 | 0.654 | 2 | 0.9836 | 3 |
| ADL-own-IC | 60.46% | 2 | 39.09% | 3 | 0.656 | 3 | 0.9707 | 2 |
| ADL-intra-IC | 60.15% | 1 | 38.82% | 1 | 0.650 | 1 | 0.9677 | 1 |

We also conduct the Wilcoxon Sign Rank (WSR) test for the statistical significance of the difference between the models’ forecasting performance. Table 4 shows the p-values of the results comparison and it has the following indications: 1) the ADL-own model significantly outperforms the Base-lift model for all the scenarios except for the MAPE when the forecast horizon is one week (e.g., there forecasting performance is not significantly different). 2) the ADL-intra model significantly outperforms the ADL-own model for all the scenarios. 3) the ADL-own-IC model significantly outperforms the ADL-own model for h=1 and h=4, but the improvements are only statistically significantly for h=4. The ADL-own-IC model even gets significantly outperformed by the ADL-own model when h=12. The results between the ADL-intra-IC model and the ADL-intra model are similar. The ADL-intra-IC model significantly outperforms the ADL- intra model for h=1 and h=4, but the improvements are only statistically significantly for h=4. The ADL-intra-IC model are significantly outperformed by the ADL-intra model when h=12; 4) the ADL-own-EWC model significantly outperforms the ADL-own model for most scenarios, while the ADL-intra-EWC model significantly outperforms the ADL-intra model. The only exception is for the MAPE when h=12, where the performance of these models and their counterparts are not significantly different (e.g., p-value= 0.184 and 0.414 respectively).

Table 4. P-values for the Pairwise Wilcoxon Sign Rank Test

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Benchmark | Candidate model | MAPE | | | SMAPE | | | MASE | | |
| h=1 | h=4 | h=12 | h=1 | h=4 | h=12 | h=1 | h=4 | h=12 |
| ADL-intra | ADL-intra-EWC | 0.002 | 0.000 | 0.414 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ADL-intra | ADL-intra-IC | 0.144 | 0.002 | 0.000 | 0.118 | 0.017 | 0.000 | 0.247 | 0.003 | 0.000 |
| ADL-own | ADL-own-EWC | 0.013 | 0.000 | 0.184 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ADL-own | ADL-own-IC | 0.110 | 0.003 | 0.000 | 0.134 | 0.028 | 0.000 | 0.172 | 0.005 | 0.000 |
| ADL-own | ADL-intra | 0.003 | 0.000 | 0.000 | 0.010 | 0.000 | 0.000 | 0.010 | 0.000 | 0.030 |
| ADL-own | Base-lift | 0.268 | 0.039 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

The strategical decision of the forecast time horizon will have an impact on which method to implement.

8.2 Overall results for the promoted period and the non-promoted period

We also investigate the models’ forecasting performance for the period when the focal product is being promoted and when the focal product is not being promoted[[10]](#footnote-10). Table 5 shows the forecasting performance of candidate models for the promoted forecast period. The results are in line with the results for all the forecast period described in section 8.1. For the promoted period, the Base-lift model has the least accurate forecasts but have competitive performance for the MAPE. This is because the results for a small number of SKU’s with very low sales has dominated the ranking for the MAPE.[[11]](#footnote-11) This result is different from the result in Huang et al. (2014) because in the latter the forecasts are conducted based on aggregated data across multiple stores.

The ADL-own model is outperformed by the ADL-intra model for all the scenarios. The ADL-own-EWC model outperforms the ADL-own model for most of the scenarios. The ADL-own-IC model has mixed forecasting performance compared to the ADL-own model: it outperforms the ADL-own model for short forecast horizons (e.g., when *h*=1). The ADL-intra-EWC model outperforms the ADL-intra model for most of the scenarios. The ADL-intra-IC model has superior forecasting performance for 1-week-ahead forecast horizon and mixed forecasting performance compared to the ADL-intra model, and gets outperformed by the ADL-intra model for the 12-week-ahead forecast horizon. The results for the non-promoted period are in line with the results in Table 3 and we do not show it here for simplicity.

Table 5. The forecasting performance of candidate models for the promoted forecast period for different forecast horizons

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| promoted period, Forecast horizon= 12 | | | | | | | | |
| Model/measure | MAPE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank |
| Base-lift | 66.36% | 1 | 81.78% | 7 | 2.162 | 7 | 1.4518 | 7 |
| ADL-own | 74.93% | 4 | 49.89% | 5 | 1.656 | 4 | 1.0000 | 5 |
| ADL-intra | 73.89% | 3 | 48.82% | 2 | 1.633 | 2 | 0.9769 | 2 |
| ADL-own-EWC | 74.97% | 5 | 49.56% | 4 | 1.645 | 3 | 0.9906 | 3 |
| ADL-intra-EWC | 73.58% | 2 | 48.42% | 1 | 1.616 | 1 | 0.9659 | 1 |
| ADL-own-IC | 76.98% | 7 | 50.56% | 6 | 1.696 | 6 | 1.0265 | 6 |
| ADL-intra-IC | 75.35% | 6 | 49.28% | 3 | 1.668 | 5 | 0.9962 | 4 |
| promoted period, Forecast horizon= 4 | | | | | | | | |
| Model/measure | MAPE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank |
| Base-lift | 64.18% | 1 | 81.35% | 7 | 2.107 | 7 | 1.5123 | 7 |
| ADL-own | 73.50% | 6 | 50.05% | 5 | 1.617 | 6 | 1.0000 | 5 |
| ADL-intra | 72.02% | 3 | 48.46% | 2 | 1.560 | 2 | 0.9676 | 2 |
| ADL-own-EWC | 74.84% | 7 | 49.82% | 4 | 1.605 | 4 | 0.9897 | 4 |
| ADL-intra-EWC | 73.26% | 5 | 48.30% | 1 | 1.548 | 1 | 0.9593 | 1 |
| ADL-own-IC | 73.18% | 4 | 50.41% | 6 | 1.612 | 5 | 1.0131 | 6 |
| ADL-intra-IC | 71.92% | 2 | 48.65% | 3 | 1.563 | 3 | 0.9753 | 3 |
| promoted period, Forecast horizon= 1 | | | | | | | | |
| Model/measure | MAPE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank |
| Base-lift | 66.42% | 4 | 87.12% | 7 | 2.220 | 7 | 1.4461 | 7 |
| ADL-own | 67.11% | 7 | 50.14% | 5 | 1.603 | 5 | 1.0000 | 5 |
| ADL-intra | 66.63% | 5 | 48.56% | 3 | 1.510 | 1 | 0.9792 | 3 |
| ADL-own-EWC | 66.79% | 6 | 50.00% | 4 | 1.605 | 6 | 1.0044 | 6 |
| ADL-intra-EWC | 65.85% | 2 | 48.47% | 2 | 1.516 | 3 | 0.9715 | 2 |
| ADL-own-IC | 66.39% | 3 | 50.20% | 6 | 1.587 | 4 | 0.9902 | 4 |
| ADL-intra-IC | 65.71% | 1 | 48.30% | 1 | 1.511 | 2 | 0.9523 | 1 |

8.3 Results for each product category

We explore models’ forecasting performance at the category level. Figure 8a to Figure 8d show the improved forecasting accuracy for the various forecast error measures by our proposed models. The improved forecasting accuracy is calculated as the percentage reduction of the error measure by applying the proposed model compared to the ADL-intra model. For example, in Figure 8a, The percentage reduction of the MAPE by using the ADL-intra-EWC model in Figure 8a is calculated as

and the value of the MAPE could be reduced by 5.61% using the ADL-intra-EWC model and by 18.51% using the ADL-intra-IC model. Figure 8a indicates that our proposed models outperform the ADL-intra model especially for most of the categories including Paper Towel, Coffee, Face Tissue, Frozen Dinner, Mayonnaise, and Soup etc. The proposed models get outperformed by the ADL-intra model for categories including Carbonated Beverage, Frozen Pizza, and Hotdogs etc. Figure 8b, 8c, and 8d suggest a consistent finding that the proposed models outperform the ADL-intra model for most of the product categories. We also investigate the results across different forecast horizons and we find the results are consistent except that the improvement of the ADL-own-IC model and the ADL-intra-IC model over their counterparts (e.g., the ADL-own model and the ADL-intra model) are getting marginal for long forecast horizon (e.g., h=4 and h=12). Results also show similar improvements by the ADL-own-EWC model and the ADL-own-IC model at the category level. The Figures are not shown here for simplicity but will be available upon request.

Figure 8a. The percentage reduction of the MAPE at the category level

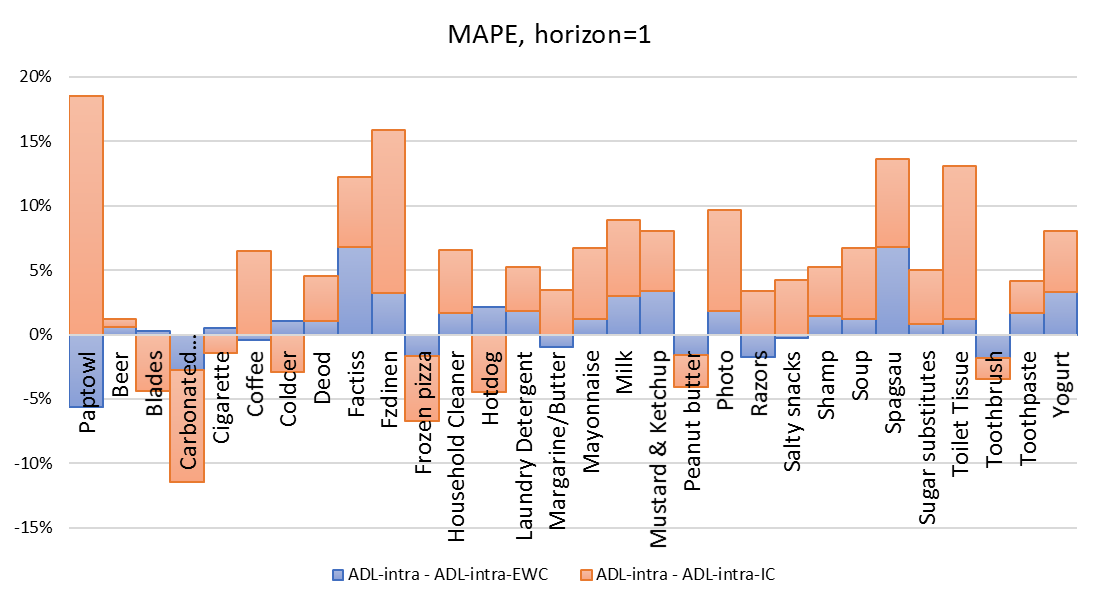


Figure 8b. The percentage reduction of the MASE at the category level

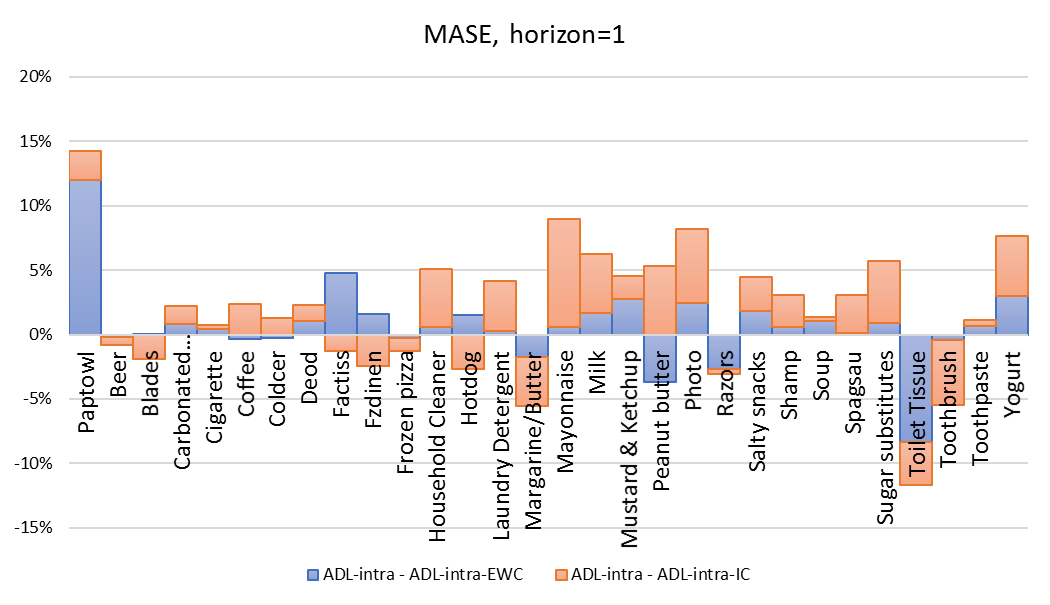


Figure 8c. The percentage reduction of the SMAPE at the category level

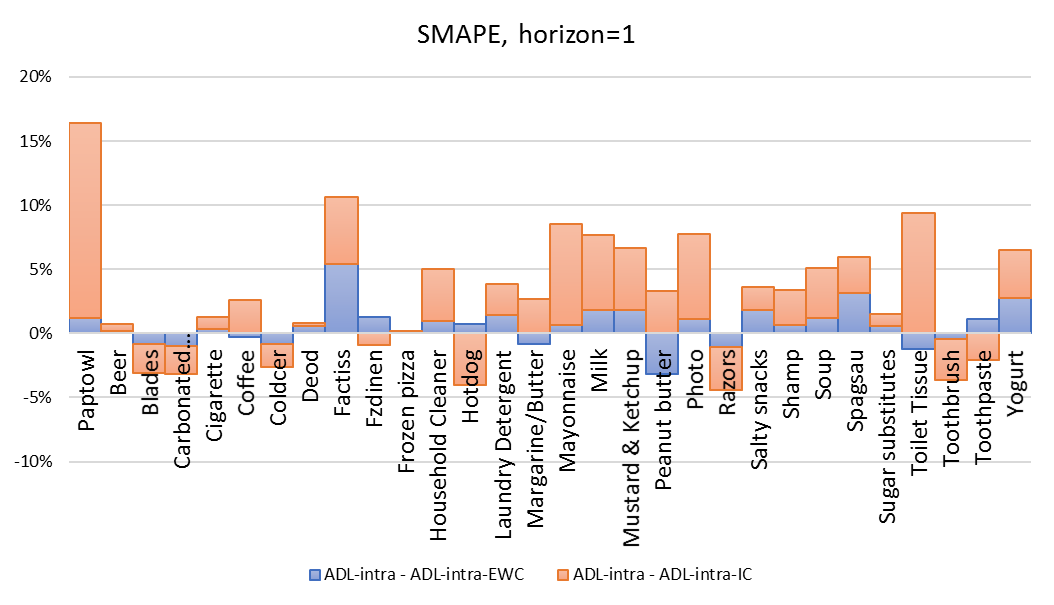
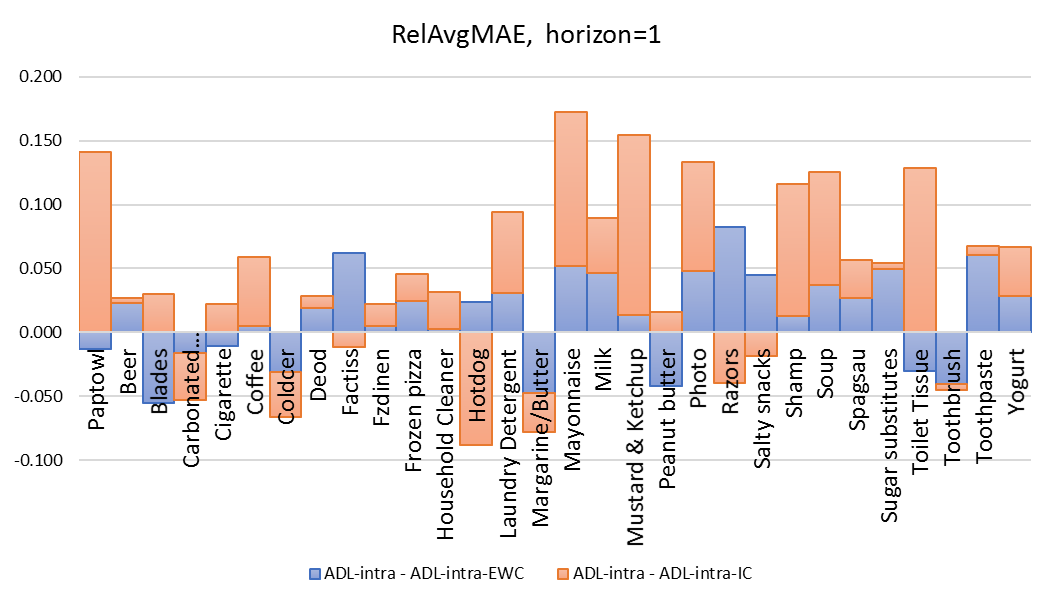


Figure 8d. The percentage reduction of the RelAvgMAE at the category level



1. **Conclusions, limitations and future research**

Grocery retailers have been struggling with producing accurate sales forecasts to effectively manage their inventory planning and customer satisfaction. In practice, many retailers use simple univariate models with adjustments for incoming promotional events. Some recent studies focused on taking advantage of the impact of promotional activities. For example, Gur Ali et al. (2009) proposed models with sophisticated function forms (e.g., the regression tree model) with the price and promotional information of the focal product. Huang et al. (2014) incorporated the competitive promotional information within the same product category by resorting to variable selection methods and the principle component analysis which mitigated the problem of high dimensionality. Ma et al. (2016) integrated the promotional information both within the same product category and across difficult product categories.

These conventional forecasting models all presume invariant effectiveness of marketing activities such as price reductions and feature and display promotions which may actually change over time because of the impact of many influencing factors including the change of economic condition, the change of the consumer taste, and new competition entry etc. However, the data for these influencing factors may not be available. As a result, the conventional models will be subject to structural break and potentially generate biased and less accurate forecasts.

In this study, we propose the ADL-intra-EWC model and the ADL-intra-IC model which take into account the potential forecast bias caused by the structural break. The ADL-intra-EWC model generates forecasts which are the combination of various sets of forecasts by the ADL-intra model with different estimation windows under the condition where structural breaks are detected. The ADL-intra-EWC model tries to obtain an effective trade-off between the forecast bias and the forecast error variance. In our experiment, the ADL-intra-EWC model generates the most accurate forecasts overall across all 30 product categories for various scenarios (e.g., forecast horizons and error measures). Table 7 shows the percentage of reductions by the model compared to other models for all these scenarios. For example, the ADL-intra-EWC model reduces the MAPE of the ADL-intra model by 0.20% and reduces the MAPE of the Base-lift model by 6.04% based on 1 to 12-week forecasting horizon. The ADL-intra-IC model tries to offset the potential forecast bias by adding the estimate of the forecast bias back to the error term at a cost of inflated forecast error variance when structural breaks are detected. The ADL-intra-IC model also has superior overall forecasting performance across all the product categories, though its advantages are getting marginal for longer forecast horizons (e.g., when *h*=12). At the category level, our proposed models have superior forecasting performance for most of the product categories.

We also evaluate the forecasting performance of the ADL-own-EWC model and the ADL-own-IC model. The models are especially valuable for manufacturers when competitive promotional information cannot be accessed ([Ali and Boylan 2011](#_ENREF_2)). In our experiment, the ADL-own -EWC model outperforms the ADL-own model across all the product categories for various scenarios, while the ADL-own -IC model outperforms the ADL-own model across all product category for short forecast horizons. Table 7 shows that the ADL-own-EWC model reduces the MAPE of the ADL-own model by 0.20% based on the 1 to 12 week ahead forecast horizon and 0.53% based on the 1 week ahead forecast horizon. The models also have superior forecasting performance for most product categories.

Table 7. Forecasting performance regarding percentage reductions in various error measures

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Forecast horizon | Proposed model | Benchmark | percentage of increase/decrease | | | |
| MAPE | SMAPE | MASE | AvgRelMAE |
| h=12 | ADL-intra-EWC | ADL-intra | -0.20% | -0.48% | -0.69% | -0.70% |
| ADL-intra-IC | ADL-intra | 0.40% | 0.51% | 0.55% | 0.33% |
| ADL-own-EWC | ADL-own | -0.20% | -0.52% | -0.65% | -0.68% |
| ADL-own-IC | ADL-own | 0.51% | 0.56% | 0.56% | 0.31% |
| ADL-intra | ADL-own | -1.46% | -0.78% | -0.61% | -0.75% |
| ADL-intra-EWC | Base-lift | -6.04% | -14.52% | -11.76% | -13.98% |
| ADL-intra-IC | Base-lift | -5.48% | -13.66% | -10.66% | -13.09% |
| h=4 | ADL-intra-EWC | ADL-intra | -0.28% | -0.53% | -0.61% | -0.74% |
| ADL-intra-IC | ADL-intra | -0.66% | -0.15% | -0.41% | -0.41% |
| ADL-own-EWC | ADL-own | -0.30% | -0.55% | -0.72% | -0.75% |
| ADL-own-IC | ADL-own | -0.89% | -0.16% | -0.76% | -0.45% |
| ADL-intra | ADL-own | -1.29% | -0.88% | -1.08% | -0.93% |
| ADL-intra-EWC | Base-lift | -2.58% | -13.00% | -10.03% | -10.14% |
| ADL-intra-IC | Base-lift | -2.95% | -12.67% | -9.85% | -9.85% |
| h=1 | ADL-intra-EWC | ADL-intra | -0.83% | -0.53% | -0.51% | -1.38% |
| ADL-intra-IC | ADL-intra | -2.11% | -1.02% | -1.01% | -2.98% |
| ADL-own-EWC | ADL-own | -0.53% | -0.52% | -0.56% | -0.46% |
| ADL-own-IC | ADL-own | -2.54% | -1.12% | -1.53% | -2.93% |
| ADL-intra | ADL-own | -0.96% | -0.78% | -1.43% | -0.26% |
| ADL-intra-EWC | Base-lift | -1.44% | -11.44% | -9.99% | -2.25% |
| ADL-intra-IC | Base-lift | -2.70% | -11.88% | -10.44% | -3.83% |

We also explored the determinants of the improved forecasting accuracy by the EWC method and the IC method. We find that in general, the EWC method is especially effective for products with fewer deep price cuts and the IC method is especially effective for products with a low coefficient of variations in product sales.

There are potentials to further improve the forecasting accuracy and we leave it to future research. For example, 1) for the EWC method, we combine five sets of forecasts based on five different estimation windows using equal weights. The forecasting performance may potentially be improved by changing the number of the estimation windows, by changing the length of the estimation windows, and by exploring alternative forecasting combination schemes (e.g., based on k-fold evaluation). For the IC method, Clements and Hendry (1999) is a summary of various correction schemes each of which may have different effects on the trade-off between the bias and the error variance[[12]](#footnote-12). 2) Ma et al. (2016) proposed models which integrate both the intra and the inter-category promotional information. We may investigate if we can further improve the forecasting performance of the ADL-intra-EWC model and the ADL-intra-IC model with inter-category information. 3) A method alternative to the EWC method and the IC method is to directly incorporate the changing process of the effectiveness of the marketing activities into the model so that the structural break may potentially be eliminated even when the influencing factors are not observed. For example, the change of the effectiveness of the marketing activities may be modeled by an autoregressive process of the marketing activities themselves. [Foekens et al. (1999)](#_ENREF_29) modeled the effectiveness of the price variables using the level of previous prices and the recency and the frequency of previous promotional events. The models are for descriptive purposes and not evaluated for forecasting. However, one of the challenges for this type of model to generate accurate forecasts is that it is sophisticated and lacks parsimony. 4) Another alternative method is the impulse saturation technique introduced by [Hendry and Krolzig (2001)](#_ENREF_33) and [Castle, Doornik et al. (2008)](#_ENREF_12). They proposed to specify the ADL model with dummy variables for each of the observations and then recursively simplify the model with the *Autometrics* algorithm based on a general-to-specific modeling strategy. The final model usually retains a large number of dummy variables to prevent the model from structural breaks and the potential forecast bias. However, the method comes with a cost of losing information by retaining these dummy variables, which makes the forecasting performance of the method an empirical question. We leave all these potential opportunities to the next stage of our research.

**Acknowledgement**

We thank the IRI company for making the data available. All the analysis and findings in this paper based on the IRI dataset are by the author and not by IRI company.

**Reference:**

Aburto, L. and R. Weber (2007). "Improved supply chain management based on hybrid demand forecasts." Applied Soft Computing **7**: 136-144.

Ali, M. and J. Boylan (2011). "Feasibility principles for Downstream Demand Inference in supply chains." Journal of the Operational Research Society **62**.

Allen, P. G. and R. Fildes (2001). Econometric forecasting. Principles of Forecasting: A Handbook for Researchers and Practitioners. J. S. Armstrong. Boston, Kluwer Academic Publishers.

Andrews, D. W. K. (1993). "Tests for Parameter Instability and Structural Change with Unknown Change Point." Econometrica **61**: 825-851.

Andrews, D. W. K. and W. Ploberger (1994). "Optimal tests when a nuisance parameter is present only under the alternative." Econometrica **62**: 1383-1414.

Andrews, R. L., et al. (2008). "Estimating the SCAN\*PRO model of store sales: HB, FM or just OLS?" International Journal of research in marketing **25**(1): 22-33.

Arenas, T., et al. (2013). "Analysis of judgmental adjustments in the presence of promotions." International Journal of Forecasting **29**(2).

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

Bai, J. and P. Perron (2003). "Computation and Analysis of Multiple Structural-Change Models." Journal of Applied Econometrics **18**: 1-22.

Blattberg, R. C., et al. (1995). "How promotions work?" Marketing Science **14**(3).

Bronnenberg, B. J., et al. (2008). "The IRI Marketing Data Set." Marketing Science **27**(4): pp. 745–748.

Castle, J. L., et al. (2008). "Model Selection when there are Multiple Breaks." Working paper No. 407, Economics Department, University of Oxford.

Chow, G. C. (1960). "Tests of Equality Between Sets of Coefficients in Two Linear Regressions." Econometrica **28**(3).

Clark, T. E. and M. W. McCracken (2007). Forecasting with Small Macroeconomic VARs in the Presence of Instabilities. Finance and Economics Discussion Series

Divisions of Research & Statistics and Monetary Affairs

Federal Reserve Board, Washington, D.C.

Clements, M. and D. Hendry (1998). Forecasting Economic Time Series, Cambridge University Press.

Clements, M. B. and D. F. Hendry (1994). Towards a theory of economic forecasting. Nonstationary Time Series Analysis and Cointegration. C. P. Hargreaves, Oxford University Press.

Clements, M. P. and D. F. Hendry (1999). Forecasting non-stationary economic time series. London, The MIT Press.

Cooper, J. P. and C. R. Nelson (1975). "The Ex Ante Prediction Performance of the St. Louis and FRB-MIT-PENN Econometric Models and Some Results on Composite Predictors." Journal of Money, Credit and Banking **7**(1).

Cooper, L. G., et al. (1999). "Promocast": a New Forecasting Method for Promotion Planning." Marketing Science **18**(3): 301-316.

Cooper, L. G. and G. Giuffrida (2000). "Turning Datamining into a Management Science Tool: New Algorithms and Empirical Results." Management Science **46**(2): 249.

Corsten, D. and T. Gruen (2003). "Desperately seeking shelf availability: an examination of the extent, the causes, and the efforts to address retail out-of-stocks." International Journal of Retail & Distribution Management **31**(12).

Davydenko, A. and R. Fildes (2013). "Measuring forecasting accuracy: the case of judgmental adjustments to SKU-level demand forecasts." International Journal of Forecasting **29**(3).

Dekimpe, M., et al. (1999). "Long-run effects of price promotions in scanner markets." Journal of Econometrics **89**(1/2): 261-291.

Dekker, M., et al. (2004). "How to use aggregation and combined forecasting to improve seasonal demand forecasts." International Journal of Production Economics **90**(2): 151-167.

Fildes, R. and P. Goodwin (2007). "Fine judgements: do organizations follow best practice when applying management judgement to forecasting?" Interfaces **37**: 570-576.

Fildes, R., et al. (2009). "Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning." International Journal of Forecasting **25**(1): 3-23.

Fildes, R., et al. (2008). "Forecasting and operational research: A review." Journal of the Operational Research Society **59**.

Fildes, R. and H. Stekler (2002). "The state of macroeconomic forecasting." Journal of Macroeconomics **24**(4): 435-468.

Foekens, E. W., et al. (1999). "Varying parameter models to accommodate dynamic promotion effects." Journal of Econometrics **89**(1-2): 249-268.

Goodwin, P. (2002). "Integrating management judgment and statistical methods to improve short-term forecasts." Omega **30**(2): 127-135.

Gür Ali, Ö., et al. (2009). "SKU demand forecasting in the presence of promotions." Expert Systems with Applications **36**(10).

Hendry, D. F. (1995). Dynamic Econometrics: Advanced Texts in Econometrics. Oxford, UK, Oxford University Press.

Hendry, D. F. and H.-M. Krolzig (2001). Automatic Econometric Model Selection using PcGets. London, Timberlake Consultants Press.

Houston, F. S. and D. L. Weiss (1975). "CUMULATIVE ADVERTISING EFFECTS: THE ROLE OF SERIAL CORRELATION\*." Decision Sciences **6**(3): 471-481.

Huang, T., et al. (2014). "The value of competitive information in forecasting FMCG retail product sales and the variable selection problem." European Journal of Operational Research **237**(2): 738-748.

Hyndman, R. J. and A. B. Koehler (2006). "Another look at measures of forecast accuracy." International Journal of Forecasting **22**: 679-688.

Kalwani, M. U. and C. K. Yim (1992). "Consumer Price and Promotion Expectations: an Experimental Study." Journal of Marketing Research **29**(February): 90-100.

Kalwani, M. U., et al. (1990). "A Price Expectations Model of Customer Brand Choice." Journal of marketing research **27**(3): 251-262.

Kopalle, P. K., et al. (1999). "The Dynamic Effect of Discounting on Sales: Empirical Analysis and Normative Pricing Implications." Marketing Science **18**(3): 317-332.

Kopalle, P. K., et al. (1999). "The Dynamic Effect of Discounting on Sales: Empirical Analysis and Normative Pricing Implications." Marketing Science **18**(3): 317.

Kotler, P. (1997). Marketing management, analysis, planning, implementation and control, Prentice Hall.

Lattin, J. M. and R. E. Bucklin (1989). "Reference effects of price and promotion on brand choice behavior." Journal of Marketing Research **26**: 299-310.

Leenheer, J., et al. (2007). "Do loyalty programs really enhance behavioral loyalty? An empirical analysis accounting for self-selecting members." international Journal of research in marketing **24**(1): 31-47.

Levy, M., et al. (2004). "Emerging trends in retail pricing practice: implications for research." Journal of Retailing **80**.

Lichtenstein, D. R. and W. O. Bearden (1989). "Contextual Influences on Perceptions of Merchant-Supplied Reference Prices." Journal of Consumer Research: An Interdisciplinary Quarterly **16**(1): 55-66.

Little, J. D. C. (1966). "A Model of Adaptive Control of Promotional Spending." Operations research **14**(6).

Loeb, W. (2015). "Unrelenting Competition: The Biggest Retail Story of 2015." 2016.

Ma, S., et al. (2016). "Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information." European Journal of Operational Research **249**(1): 245-257.

Mace, S. and S. A. Neslin (2004). "The determinants of pre- and postpromotion dips in sales of frequently purchased goods." Journal of Marketing Research **XLI**: 339-350.

Mahajan, V., et al. (1980). "Feedback Approaches to Modeling Structural Shifts in Market Response." Journal of Marketing **44**: 71-80.

Melnyk, V. and T. H. A. Bijmolt (2007). "Effects of having and terminating a loyalty program: the role of monetary and non-monetary rewards." Working paper New Zealand: University of Waikato.

Moinpour, R., et al. (1976). "Time Changes in Perception: A Longitudinal Application of Multidimensional Scaling." Journal of marketing research **13**(3): 245-253.

Monroe, K. B. and J. P. Guiltinan (1975). "A Path-Analytic Exploration of Retail Patronage Influences." The Journal of Consumer Research **2**(1): 19-28.

Morrison, D. G. (1966). "Interpurchase Time and Brand Loyalty." Journal of Marketing Research **3**.

Muellbauer, J. (1994). "The Assessment: Consumer Expenditure." Oxford Review of Economic Policy **10**(2): 1-41.

Myers, J. G. (1971). "The Sensitivity of Dynamic Time-Path Typologies." Journal of marketing research **8**(4): 472-479.

Myers, J. G. and F. M. Nicosia (1970). "Time-Path Types: From Static to Dynamic Typologies." Management Science **16**(10): B584-B596.

Nijs, V. R., et al. (2001). "The Category-Demand Effects of Price Promotions." Marketing Science **20**(1): 1-22.

Nikolopoulos, K. (2010). "Forecasting with quantitative methods: the impact of special events in time series." Applied Economics **42**: 947-955.

OrderDynamics (2015). Retailers and the Ghost Economy: The Haunting of Returns. <http://engage.dynamicaction.com/WS-2015-06-IHL-Ghost-Economy-Haunting-of-Returns-AR_LP.html>.

Pesaran, H. M. and A. Timmermann (2007). "Selection of estimation window in the presence of breaks." Journal of Econometrics **137**: 134-161.

Pesaran, M. H., et al. (2009). "Forecasting Economic and Financial Variables with Global VARs." International Journal of Forecasting **25**: 642-675.

Pesaran, M. H. and A. Timmermann (2005). "Small sample properties of forecasts from autoregressive models under structural breaks." Journal of econometrics **129**(1-2): 183-217.

Tashman, L. J. (2000). "Out-of-sample tests of forecasting accuracy: an analysis and review " International Journal of Forecasting **16**(4).

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." Journal of the Royal Statistical Society. Series B (Methodological) **58**(1): 267-288.

Trusov, M., et al. (2006). "Retailer Promotion Planning: Improving Forecasting Accuracy And Interpretability." Journal of Interactive Marketing **20**(3-4): 71-81.

Van Heerde, H. J., et al. (2003). "Is 75% of the Sales Promotion Bump Due to Brand Switching? No, Only 33% Is." Journal of Marketing Research **XL**: 481-491.

Van Heerde, H. J., et al. (2008). "Decomposing the Demand for a Pioneering Innovation." Working paer, University of Waikato, Department of Marketing.

Verhoef, P. C., et al. (2007). "Multichannel customer management: Understanding the research-shopper phenomenon." international Journal of research in marketing **24**(2): 129-148.

Wedel, M. and J. Zhang (2004). "Analyzing brand competition across subcategories." Journal of Marketing Research **41**(4): 448-456.

Wichern, D. W. and R. H. Jones (1977). "Assessing the Impact of Market Disturbances Using Intervention Analysis." Management Science **24**(3): 329-337.

Wildt, A. R. (1976). The empirical investigation of time dependent parameter variation in marketing models. American Marketing Association. E. proceedings**:** 466-472.

Wildt, A. R. and R. S. Winer (1983). "Modeling and Estimation in Changing Market Environments." The Journal of Business **56**(3).

Winer, R. S. (1979). "An Analysis of the Time-varying Effects of Advertising: The Case of Lydia Pinkham." The Journal of Business **52**(4).

Wittink, D., et al. (1988). SCAN\*PRO: the estimation, validation and use of promotional effects based on scanner data. Internal paper, Cornell University.

1. More specifically, the forecast bias comes from the change of the deterministic mean of the model due to the change of the model parameters. There is a possibility that the deterministic mean could retain unchanged even if the parameters change. Under such circumstance, there will be no forecast bias even when the model is subject to structural break. However, in this study we do not explain this situation as it only happens theoretically when very restrictive conditions are met. Details based on an example of a VAR model can be found in Clements and Hendry (1999). [↑](#footnote-ref-1)
2. Analytical evidence for the models with endogenous explanatory variables can be found in Clements and Hendry (1999) and Pesaran and Timmerman (2005, 2007). [↑](#footnote-ref-2)
3. This setting is very common in the retailer context. In this example we artificially make up the data series but we keep the data series to be stationary. [↑](#footnote-ref-3)
4. The values from week 1 to week 50 are predicted by the model estimated with the data from week 51 to week 75. [↑](#footnote-ref-4)
5. The Chow test is a variant of F-test which compares the fitting of the model before and after the structural break. It assumes the locations of one structural break known a priori and also invariant error variations before and after the break. For example, we conduct the Chow test assuming the break occurs at a specific week (e.g., week 30). A small p-value would reject the null hypothesis of no structural at week 30. [↑](#footnote-ref-5)
6. To mitigate the multiple comparison problem, we may adopt very small threshold (e.g., 0.0001) for the p-value of the sequential test. [↑](#footnote-ref-6)
7. In Figure 5, the black dashed line for the estimation period (e.g., week 1 to week 75) represents the predicted value of the original regression model which is estimated using the full sample. [↑](#footnote-ref-7)
8. We include the following US calendar events including *Halloween*, *Thanksgiving*, *Christmas*, *New Year’s Day*, *President’s Day*, *Easter*, *Memorial Day*, the *4th of July*, and *Labour Day*. [↑](#footnote-ref-8)
9. In this study, when h=4 and h=12, the error measures are calculated based on 1 to 4 step ahead forecasts and 1 to 12 step ahead forecasts respectively. [↑](#footnote-ref-9)
10. The values of the error measures we report in Table 4a, 4b, and 4c are all aggregate values. The values of the error measures for all forecast period may not necessarily drop in the range between the values for the non-promoted forecast period and the values for the promoted forecast period. This is because that the lengths of the promoted time periods are different across the SKU’s. As a result, the weights we use to combine the various data series for all forecast period, for the promoted forecast period, and for the non-promoted forecast period can be different. [↑](#footnote-ref-10)
11. These SKU’s have very low baseline sales and very high sales spikes for most of the promotional events except the most recent promotions in the estimation period and the forecast period (they do not generate big sales spikes possibly due to other unobserved factors). As a result, the ADL family models substantially over-predict the product sales in the forecast period while the Base-lift method may generate good forecasts (because it estimates the promotional effect based on the effect by the previous promotion). These results will dominate the overall MAPE ranking because the denominator in the MAPE is very small (due to low baseline sales). [↑](#footnote-ref-11)
12. For example, in this study we generate the forecasts first and then add the estimated bias to all the forecasts. One of the alternative schemes is to first make adjustments to the one-step-ahead forecast, and then calculate the two-step-ahead forecast based on the value of the one-step-ahead forecast which has already adjusted, and so forth. [↑](#footnote-ref-12)