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

Tao Huang, Robert Fildes, Didier Soopramanien

Abstract

Retailers need accurate forecasts at SKU level for their strategical and tactic decisions. Previous studies have developed forecasting models which incorporate the variables of price and promotions. These models overlook the potential change of the effect of the price and promotions. As a result, they may potentially be subject to structural break and generate biased and less accurate forecasts. In this study, we integrate recently developed techniques to takes into account the structural break. Our proposed three-stage methods generate more accurate forecasts compared to industrial practice and conventional econometric models.

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, Makridakis et al. 2014](#_ENREF_65)). Poor forecasts of product sales cause out-of-stock conditions and overstock conditions. When the product is out-of-stock, retailers not only directly 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_25)). Retailers may intentionally overstock to maintain a high customer satisfaction level, which however significantly raises inventory costs (e.g., capital cost, warehousing, and deterioration etc.) and reduces profits ([Cooper, Baron et al. 1999](#_ENREF_23)). In the year of 2014, Retailers in North American had a loss of $634.1 billion due to stock-outs and spent $471.9 billion to overstock ([OrderDynamics 2015](#_ENREF_58)). Retailers need more accurate forecasts for their product sales at SKU level.

In practice, many retailers generate their forecasts using a two-stage ‘base-lift’ approach. The forecasts are generated separately depending on whether or not the focal product is being promoted. The ‘base’ forecasts are usually generated using simple univariate models, while the ‘lift’ effect, which is the effect of the promotion, is estimated by the brand/category manager based on his/her experience. Some previous studies propose a procedure to help managers improve their judgments (e.g., [Goodwin 2002](#_ENREF_35), [Fildes, Nikolopoulos et al. 2008](#_ENREF_32), [Nikolopoulos 2010](#_ENREF_57)). Others develop models to estimate the ‘lift’ effect based on the data ([Cooper, Baron et al. 1999](#_ENREF_23), [Cooper and Giuffrida 2000](#_ENREF_24), [Trusov, Bodapati et al. 2006](#_ENREF_71)). Some recent studies directly generate the final forecast of the product sales. 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. [Huang, Fildes et al. (2014)](#_ENREF_40) proposed a two stage general-to-specific Autoregressive Distributed Lag (ADL) 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_45) further integrated the promotional information from the products across other related product categories.

These studies all assume constant effects of the price and promotions. In practice, the effect of prices and promotions may change due to many influencing factors including the change of economic conditions, the change in consumer tastes, and new competitor entry etc. which are usually not observable or measurable ([Wildt 1976](#_ENREF_76), [Wildt and Winer 1983](#_ENREF_77)). Customers may become more price/deal sensitive during an economic crunch. Customers may change their preference to the products due to their cognitive bias, product familiarity, change of their lifestyle and social status ([Meeran, Jahanbin et al. 2017](#_ENREF_49)). When a new competitor enters the market, the effect of prices and promotions of the focal product may be reduced not only because the new competitor launches their marketing activities but also because customers seek variety. In the year of 2014, the German low-price retail chain Aldi has opened more than 400 stores in the United States, which leaves great pressures to other existing retail chains ([Loeb 2015](#_ENREF_44)).

Under such circumstance, conventional models which assume no change of the effect of the prices and promotions 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_2), [Armstrong 2001](#_ENREF_8)). The model which is subject to structural break may generate biased and less accurate forecasts. The issue of structural break have been historically addressed in the economics literature (see [Clements and Hendry 1994](#_ENREF_20), [Pesaran and Timmermann 2005](#_ENREF_64)). In this study, we propose effective forecasting methods which generate more accurate forecasts by mitigating the forecast bias due to the structural break. Our methods contain three stages: we first select the most important predictors (e.g., competitive prices and competitive promotions) ([e.g., Huang, Fildes et al. 2014](#_ENREF_40)). We then incorporate these predictors into a general ADL model which is simplified afterwards using the LASSO algorithm ([e.g., Ma, Fildes et al. 2016](#_ENREF_45)). Finally, we implement the estimation window combining technique and the intercept correct technique to the obtained simplified model based on the results a sequential structural break test.

Our research is significant for the following contribution: 1) our methods have superior forecasting performance compared to conventional models which assume no change in the effect of product prices and promotions; 2) unlike any earlier study which rely on incorporating additional information or construct models of sophisticated structure, our methods rely on how promotional information could be effectively utilized. Our methods match managerial intuition that the effect of the prices and promotions do change in practice; 3) Our study provides an evaluation of various forecasting methods which offers operational guidance to not only retailers but also manufacturers when competitive promotional information become not accessible. 4) the method we propose is fully automatic compared to [Huang, Fildes et al. (2014)](#_ENREF_40) and easy to implement; 5) we conduct the evaluation for 1834 SKUs across 30 product categories in 30 retail stores, which provides robust results.

The remainder of the paper is arranged as follows: Section 2 summarizes previous studies. Section 3 explains the issue of structural break and the subsequent forecast bias when conventional models overlook the change in the effect 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**

In practice, many retailers forecast their product sales at SKU level using a two stage ‘base-lift’ method. The method splits the data into promoted and non-promoted periods based on whether or not the focal product is being promoted. The method is a combine of simple univariate methods (for the non-promoted period) and human judgments by brand/category managers (for the promoted period) ([Fildes, Nikolopoulos et al. 2008](#_ENREF_32), [Fildes, Goodwin et al. 2009](#_ENREF_31)). A stream of studies has been devoted to helping managers with better adjustment procedure ([Fildes and Goodwin 2007](#_ENREF_30), [Arenas, Pedregal et al. 2013](#_ENREF_7)). Other studies try to improve the adjustment with model-based forecasting systems. e.g., 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_23), [Cooper and Giuffrida 2000](#_ENREF_24), [Trusov, Bodapati et al. 2006](#_ENREF_71)). One of the limitations for these methods is that, as they split the data into two periods, they tend to overlook the information in the promoted period when forecasting the product sales in the non-promoted period, and vice versa.

Previous studies have also proposed holistic methods to conduct the forecast for the promoted and non-promoted periods. [Gür Ali, SayIn et al. (2009)](#_ENREF_36) evaluated the forecasting performance of the support vector machine (SVM) models and regression tree models. Their methods incorporate a range of variables constructed based on the promotional information of the focal product. Divakar et al. (2005) developed the CHAN4CAST system with models of a dynamic regression structure to forecast brand volume sales for the manufacturer/channel. [Huang, Fildes et al. (2014)](#_ENREF_40) proposed a two-stage general-to-specific ADL model with competitive promotional information within the same product category. The competitive promotional information are selected with variable selection methods or are constructed using principal component analysis. [Ma, Fildes et al. (2016)](#_ENREF_45) further integrated the promotional information not only from the same category of the focal product but also from other related product categories. They resort to Granger causality test to indicate the relevant product categories and then rely on the Least Absolute Shrinkage and Selection Operator Algorithm (LASSO) not only as a variable selection procedure but also as a model simplification strategy.

These studies incorporated price and promotional information to forecast retailer product sales. This is because price and promotion have strong impact on product sales. A large number of studies has been devoted to exploring the mechanism of price and promotions. Some studies find that price and promotions significantly increase short-term sales ([Blattberg, Briesch et al. 1995](#_ENREF_11)), exhibit (asymmetrical) competitive effect ([Wittink, Addona et al. 1988](#_ENREF_79), [Dekimpe, Hanssens et al. 1999](#_ENREF_27), [Wedel and Zhang 2004](#_ENREF_74), [Andrews, Currim et al. 2008](#_ENREF_5)), lead to purchase acceleration and anticipation ([Van Heerde, Gupta et al. 2003](#_ENREF_72), [Mace and Neslin 2004](#_ENREF_46)).

Other studies reveal that the effect of prices and promotions may change over time (e.g. [Little 1966](#_ENREF_43), [Morrison 1966](#_ENREF_52), [Myers and Nicosia 1970](#_ENREF_55), [Myers 1971](#_ENREF_54), [Houston and Weiss 1975](#_ENREF_39), [Monroe and Guiltinan 1975](#_ENREF_51), [Moinpour, McCullough et al. 1976](#_ENREF_50), [Wildt 1976](#_ENREF_76), [Wichern and Jones 1977](#_ENREF_75), [Winer 1979](#_ENREF_78), [Mahajan, Bretschneider et al. 1980](#_ENREF_47)). [Wildt (1976)](#_ENREF_76) and [Wildt and Winer (1983)](#_ENREF_77) attribute the change of the effect of the marketing activities to the change of economic conditions, consumer tastes, and competition situations etc. Customers may price reductions and sales promotions more attractive when there is an economic crunch. Customers may change their taste and preference as they accumulate knowledge of the product and when they try to seek varieties, and they may also change their preference when they reach a different social status or decide to adopt a different lifestyle ([Meeran, Jahanbin et al. 2017](#_ENREF_49)). Research at the store level also find that introductions of new products (e.g., 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_56), [Van Heerde, Srinivasan et al. 2008](#_ENREF_73)). Lastly, the effect of prices and promotions may change during the different stages of the product lifecycle ([Mahajan, Bretschneider et al. 1980](#_ENREF_47)). The change of the effect of prices and promotions however has been overlooked by all the existing methods for retailer product sales.

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

The forecasting models which overlooks the change of the effect of prices and promotions will be subject to structural break which is defined as large changes in the model’s parameters ([Allen and Fildes 2001](#_ENREF_2), [Armstrong 2001](#_ENREF_8)). The changes may occur at the intercept and/or the parameters of the explanatory variables, and lead to a shift of the deterministic mean ([Clements and Hendry 1999](#_ENREF_21)). The estimated deterministic mean then becomes the weighted average of the true deterministic means before and after the structural break. As a result, the forecasts generated by the model will be biased and less accurate. The negative impact of the structural break on the model’s forecasting performance has been addressed in the macroeconomics literature (e.g. [Cooper and Nelson 1975](#_ENREF_22), [Muellbauer 1994](#_ENREF_53), [Hendry 1995](#_ENREF_37), [Stock and Watson 1996](#_ENREF_68), [Clements and Hendry 1999](#_ENREF_21), [Pesaran and Timmermann 2007](#_ENREF_60), [Castle, Doornik et al. 2008](#_ENREF_14), [Pesaran and Pick 2011](#_ENREF_61)). These studies suggest that the parameters of their forecasting models may change due to influencing factors including the shift of the market sentiments, the regulation, and debt management etc. A large number of studies have been devoted into take into account the impact by the change of the parameters in order to achieve higher forecasting accuracy in financial interest rate and stock market return (e.g., [Perez-Quiros and Timmermann 2000](#_ENREF_59), [Ang and Bekaert 2002](#_ENREF_6), [Pesaran and Timmermann 2002](#_ENREF_63)). In this study, we aim to take into account the impact by the change of the effect of prices and promotions in order to generate more accurate forecasts for retailer product sales.

[Pesaran and Timmermann (2005)](#_ENREF_64) illustrates analytically how a structural break leads to forecast bias using a simple regression model. In the retailer context, suppose 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 but there is a structural break at week (where ). The parameter of the price variable changes from to after . In reality, 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 are unknown to the forecaster. Therefore, the unobservable real demand can be represented as follows:

where, is an indicator which equals to 1 before week and 0 afterwards. 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, and we consider to be strictly exogenous[[1]](#footnote-1). 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 estimates for the model based on the data from week *m* to week *T* become:

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

Thus the forecast at week is obviously biased because , which is unequal to zero.

We illustrate the impact of structural break on forecasting accuracy with an example using simulation. For example, we construct a price variable with its values being 2.99 for most of the observations (say, weeks) but occasionally reduced to 2.29 or 1.99[[2]](#footnote-2). We assume that, during a period of 100 weeks, product sales are driven by product prices and subject to two structural breaks[[3]](#footnote-3):

, , when

, , when

, , when

where and represent the sales and the price at week *t*. is the error term. In practice, the structural breaks may occur because of new product introduction (which reduces the price elasticity of the focal product) and a credit crunch (so that customers become more price sensitive). The sales and price are represented in Figure 1 by the solid black line and the solid red line respectively.

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



Suppose that we need to develop models to forecast the product sales for the period from week 76 to week 100. If we know the existence and the locations of the breaks, we may develop a congruent model (i.e., ) exclusively based on the post break data (i.e., data from week 51 to week 75) and generate unbiased forecasts. Figure 1 represents the predictions/forecasts using the black dashed line (e.g., *ybar\_post breaks*). Table 1 shows the forecasting performance of the model with post break data (e.g., with MAE= 0.3, MSE= 0.18, MAPE= 5.0%, and SMAPE= 4.3%).

However, the existence and the locations of the structural breaks are usually unknown. As mentioned in section 2, the effect of the price may change due to influencing factors which are unobservable and/or measurable to the retailer. Therefore, if we develop the model using all the data (i.e., from week 1 to week 75) but without taking into account the two structural breaks, we obtain estimates of the parameters as the weighted average of the true parameters before and after the breaks and generate biased forecasts. We tend to over-predict the sales from week 1 to week 25, under-predict the sales from week 26 to week 50, then again over-predict the sales from week 51 to week 70, and finally generate downwards-biased out-of-sample forecasts from week 76 to week 100. Figure 2 shows the biased predictions/forecasts with the black dashed line (as *ybar\_all data*). Table 1 shows the forecasting performance of the model with the full data (e.g., with MAE= 0.7, MSE= 0.52, MAPE= 12.2%, and SMAPE= 11.5%). The forecasts are substantially inferior compared to the model with post break data.

Figure 2. Simulated sales with a structural break: model with full data[[4]](#footnote-4)



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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | MAE | MSE | MAPE | SMAPE |
| Model with post-break data | 0.3 | 0.18 | 5.0% | 4.3% |
| Model with full data | 0.7 | 0.52 | 12.2% | 11.5% |
| Model with full data, with IC | 0.1 | 0.01 | 1.7% | 1.8% |
| Model with full data, with EWC | 0.6 | 0.43 | 11.0% | 10.5% |

1. **Dealing with structural breaks**

In this study, we propose effective forecasting models for retailer product sales at SKU level by taking into account the impact of structural breaks. In this section, we introduce the Intercept Correction (IC) method and the Estimation Window Combining (EWC) method which are incorporated into one of the stages in our proposed method.

4.1 The Intercept Correction method

The IC method suggests using conventional models to generate forecasts but specify non-zero values for the model’s errors in the forecasting period ([Clements and Hendry 1994](#_ENREF_20), [Clements and Hendry 1999](#_ENREF_21), [Clark and McCracken 2007](#_ENREF_17)). Based on the example in section 3, if we believe the model is subject to structural break and forecast bias, we may estimate the bias with the average value of the most recent residuals, i.e., , where is the number of residuals being used to estimate the forecast bias. When , the estimate reduces to , which is the residual at the forecast origin ([e.g., Chevillon 2016](#_ENREF_15)). The estimated bias are added back to the out-of-sample forecasts, which may potentially improve the forecasting accuracy though at a cost of inflated forecasting error variance ([Clements and Hendry 1999](#_ENREF_21)).

We illustrate how the IC method improves the forecasting accuracy using the same example in section 3. We construct the congruent model as but presuming no priori knowledge for the locations of the breaks. We conduct a sequential [Chow (1960)](#_ENREF_16) test based on every observation in the estimation period[[5]](#footnote-5). The rejection of the null hypothesis of no structural break for any of the observation would suggest that the model is subject to structural break though without indicating 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 for all the weeks one at each time. The results reject the null hypothesis of no structural break for some of the weeks (e.g., week 20) but fail to do so for some other weeks (e.g., week 35). Based on the results we consider the model being subject to structural breaks[[6]](#footnote-6). Alternative advanced tests considering multiple breaks, heteroskedasticity, and unit roots are available but require additional priori knowledge such as the number of potential structural breaks ([Andrews 1993](#_ENREF_3), [Andrews and Ploberger 1994](#_ENREF_4), [Bai and Perron 1998](#_ENREF_9), [Bai and Perron 2003](#_ENREF_10)).

Figure 3 *P*-values of the sequential Chow test



In the retailer context, sales data at SKU level exhibit large variations, unexpected outliers, and missing values, which makes the bias estimation exclusively based on the residual at the forecast origin unreliable. In this example, we estimate the forecast bias as the average value of the four residuals close to the forecast origin. e.g., . We obtain the final corrected forecasts as . Figure 4 represents the predictions/forecasts using the black dashed line (e.g., *ybar\_IC*). Table 1 shows the forecasting performance of the intercept corrected model (e.g., with MAE= 0.1, MSE= 0.01, MAPE= 1.7%, and SMAPE= 1.8%). The intercept corrected model substantially outperforms the model with the full data.

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



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_64). Unlike the IC method, the EWC method does not estimate the forecast bias. It takes a 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 ([Pesaran and Pick 2011](#_ENREF_61)). In the example in section 3, if we know the location of the structural break, we could estimate the model with the post-break data and generate unbiased forecasts. Alternatively, if we suspect the break but do not know the location, we may use the most recent observations close to the forecast origin as long as there are enough observations to estimate the model. It becomes less likely for the model to be subject to structural break if we keep *m* as large as possible (so that we discard more old data). If *m* by chance becomes larger than , the model will be estimated with post-break data and generate unbiased forecasts.

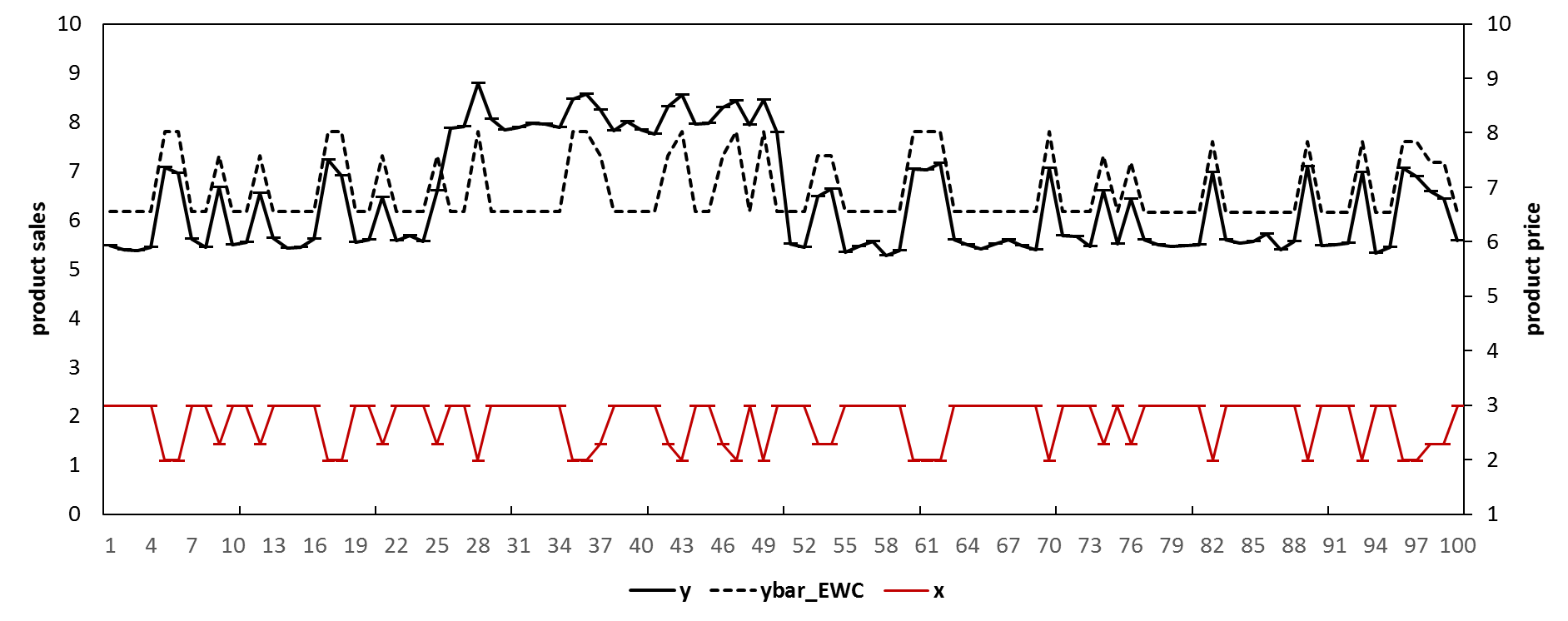
However, this does not suggest that when we have suspicion for structural breaks we must adopt estimation windows as small as possible. This is because that the reduction of the forecast bias comes with the cost of inflated forecasting error variance as we will be using less information (e.g., the estimation sample becomes smaller). In the example in section 3, the Mean Square Error (MSE) at week can be represented as , where , interpreted as the squared forecast bias, , is interpreted as the efficiency term ( is the forecasting error variance), , , and . The change of the MSE for week when we estimate the model with one additional observation (i.e., week m-1) can be represented as:

where is the MSE at week based on the estimation window [m-1, T]. [Pesaran and Timmermann (2005)](#_ENREF_64) show that, when one additional observation is included in the estimation, the change of the squared bias term (e.g., ) is always non-negative (i.e., the bias will increase), but the change of the efficiency term (e.g., ) depends on the error variance before and after the structural break. If (e.g., there are more pre-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 pre-break 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 and adopt a smaller estimation window, 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, and vice versa.

[Pesaran and Timmermann (2005)](#_ENREF_64) suggest combining the forecasts generated by the model of the same specification but estimated with different sample windows to achieve an effective trade-off between forecast bias and forecasting error variance, as forecast combination usually leads to higher accuracy ([Clemen 1989](#_ENREF_18), [Jose and Winkler 2008](#_ENREF_42)). In this study, we combine the forecasts with equal weights as the equal weight combining scheme has been found effective and easy to implement.([Clements and Hendry 1998](#_ENREF_19), [Fildes and Stekler 2002](#_ENREF_33), [Dekker, van Donselaar et al. 2004](#_ENREF_28), [Pesaran, Schuermann et al. 2009](#_ENREF_62)). For example, 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 then add more observations (e.g., one) 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 calculate the final forecasts as the average value of the () sets of *h*-step-ahead forecasts:

We illustrate how the EWC method improves the forecasting accuracy using the same example in section 3. For example, we estimate the model using the data from week 1 to week 75, and generate the forecasts which are subject to the full bias (referred as ). We then estimate the same model but use the data with one less observation (e.g., from week 2 to week 75) and generate the forecasts (referred as ), and so forth. The forecasts including are less biased compared to but associated with inflated forecasting error variance because they were generated by models with less information. We arbitrarily choose to be 16 and thus we calculate the final forecasts as the average of sets of forecasts. e.g.,. Figure 5 represents the predictions/forecasts using the black dashed line (e.g., *ybar\_EWC*). Table 1 shows the forecasting performance of the EWC model (e.g., 0.6 for MAE, 0.43 for MSE, 11.0% for MAPE, and 10.5% for SMAPE). The EWC method outperforms the conventional model with the full data.

Figure 5. Simulated sales with a structural break: model with EWC



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_12). The dataset contains weekly data at SKU level with variables including product unit sales, price, features, and displays etc. We conduct our evaluation based on 1834 SKU’s for 30 product categories from 30 stores[[7]](#footnote-7). Table 2 shows the basic statistics for the selected SKU’s for each of the categories. Some categories (e.g., Carbonated Beverages and Hotdog) exhibit much higher promotional intensity compared to others (e.g., Margarine/Butter and Mayonnaise). Figure 6 illustrates the data for a typical SKU in the Beer category. For example, the product has occasional price reductions and feature/display events which are associated with sales spikes.

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. Store level data for an SKU in the Beer category[[8]](#footnote-8)



1. **The models**

We include the base-lift method as one of the benchmark models. This method has been widely used by retailers to forecast product sales at SKU level ([Cooper, Baron et al. 1999](#_ENREF_23)). It generates baseline forecasts using simple exponential smoothing method and makes adjustments for any incoming promotional event based on the lift effect by the most recent promotional event. Details of the method can be found in [Huang, Fildes et al. (2014)](#_ENREF_40).

In this study, we propose new effective models which generate more accurate forecasts by taking into account the issue of structural break and forecast bias. Our method contains three stages. At the first stage, we identify the most informative competitive explanatory variables for the focal product. Grocery retailers typically sell hundreds of SKU’s and this leads to hundreds of potential competitive explanatory variables for the focal product. Incorporating all the variables into the model would easily overfit the model and even make the estimation infeasible ([Martin and Kolassa 2009](#_ENREF_48)). Therefore, we select the variables using the Least Absolute Shrinkage and Selection Operator (LASSO) ([Tibshirani 1996](#_ENREF_70)). For example, we have the following model:

where

represents log product sales of the focal product at week *t*  
 represents the matrix for the explanatory variables including the product price, feature, and display of all the products in the same product category

*u* represesnt the identically distributed error term

is the vector of the parameter coefficients  
*N* is the number of parameters which is the total number of SKUs for the category  
 is the shrinkage factor

The LASSO procedure puts a constraint to the sum of the absolute values of all the parameter coefficients of the model. Therefore, the selection procedure removes some of the explanatory variables by pushing their parameters towards zero. The model simplification process is controlled by a shrinkage factor based on 10-fold cross-validation following Ma et al. (2016).

At the second stage, we construct the ADL model based on the retained variables by the LASSO procedure with their dynamic terms following Huang et al. (2014). We have 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 and Witt 2003](#_ENREF_67)).

and represent the log price of the focal product and the competitive product at week

and represents the Feature dummy for 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 *[[9]](#footnote-9)*

are the parameters  
 is the error term and we assume

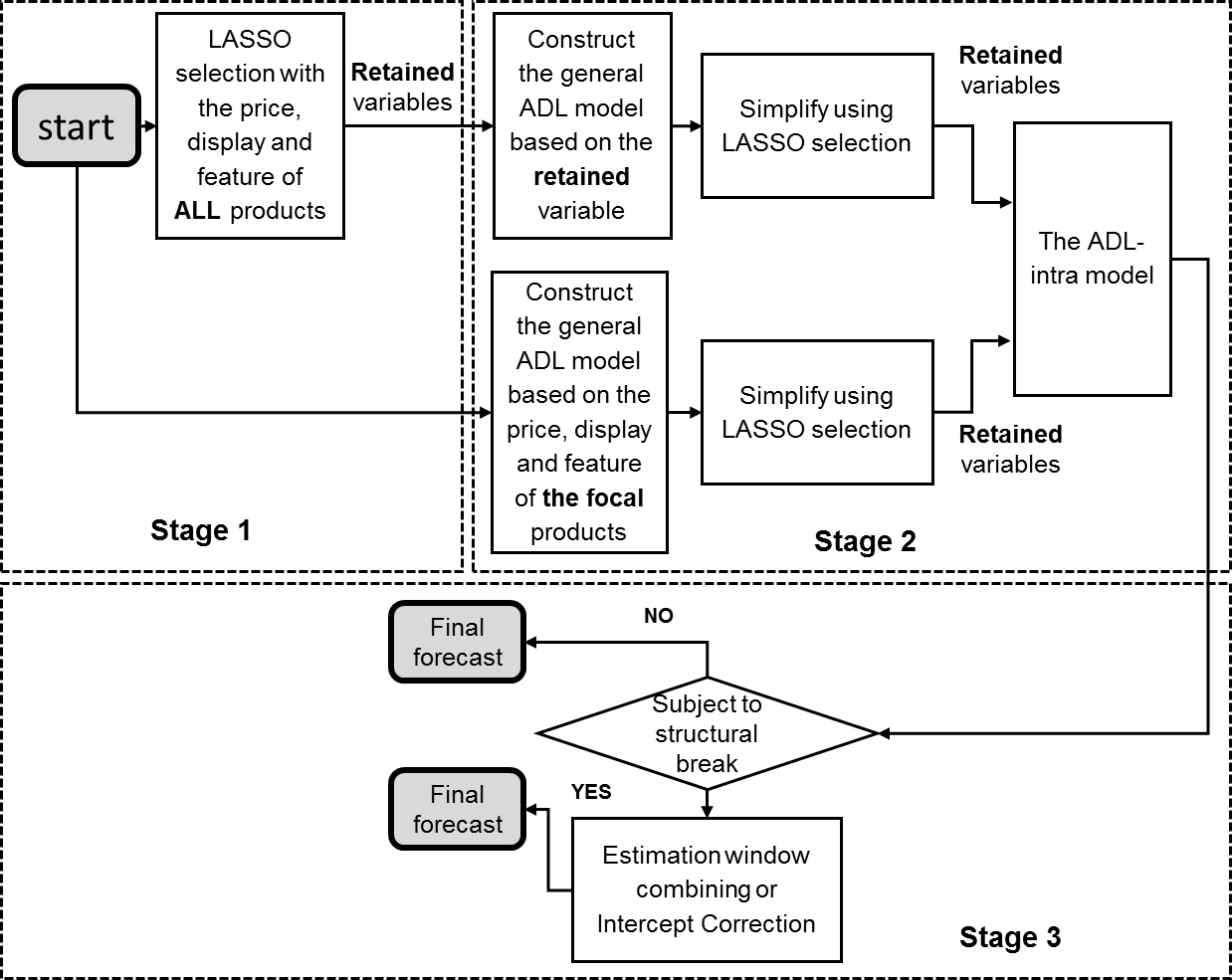
is the order of the lags and is set to as 2.

*, ,* and are the numbers of selected competitive price, Feature, and Display variables for the product category.

The model is then simplified using the LASSO procedure. However, the LASSO procedure is subject to the limitation of potentially missing important variables under the condition multicollinearity ([Fan and Lv 2008](#_ENREF_29), [Ma, Fildes et al. 2016](#_ENREF_45)). In practice, retailers tend to promote relevant products at the same time, which may even exaggerate the challenge. As a result, the LASSO procedure may get exposed to the risk of picking up less important variables. Therefore, we construct a parallel model which only include the price and promotion variables of the focal product. For example,

We then simplify this parallel model using the LASSO procedure (we refer this simplified parallel model as the ADL-own model thereafter). The final model at the second stage will contain the retained variables by the LASSO procedure in these two ADL models in combine. We incorporate the retained variables in the ADL-own model because previous studies suggest that own promotional variables are more important compared to variables of other products ([Bucklin, Gupta et al. 1998](#_ENREF_13)). We therefore increase the probability to include the own promotional variables at a potential cost of efficiency (we refer the final model at the second stage as the ADL-intra model throughout the study).

Figure 7. An illustration for the three-stage ADL-intra-EWC and the three-stage ADL-intra-IC models



At the final stage, we integrate the ADL-intra model with the EWC method and the IC method to take into account the issue of structural breaks. We implement the EWC method and the IC method to the ADL-intra model based on the ADL-intra model if the sequential Chow test rejects the null hypothesis of no structural break and otherwise keep the forecasts of the ADL-intra model as the final forecasts. To implement the EWC method, we re-estimate the ADL-intra model with five estimation windows with different lengths (e.g., [1, 160], [5, 160], [9, 160], [13, 160], and [17, 160], given an initial estimation window of 160 weeks, for example), and generate five sets of forecasts. We then combine the five sets of forecasts with equal weights. To implement the IC method, we estimate the forecast bias as the average value of the four most recent residuals and add the value equally to the forecasts for each forecast horizon. We refer the models at this final stage as the ADL-intra-EWC model and the ADL-intra-IC model. Figure 7 illustrates the steps for the two methods.

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, Makridakis et al. 2014](#_ENREF_65))). In this study, we also include two additional models which are the ADL models without competitive promotional information but integrated with the EWC method and the IC method accordingly, i.e., the ADL-own-EWC method and the ADL-own-IC method.

1. **The experimental design**

In this study, we evaluate the forecasting performance of the models with rolling origins ([Tashman 2000](#_ENREF_69)). 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_41), and the Relative Average Mean Absolute Error (RelAvgMAE) proposed by [Davydenko and Fildes (2013)](#_ENREF_26). 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**

Table 3 shows the forecasting performance of the models for the various error measures and forecast horizons based on the overall results across all the 30 product categories. Table 4 shows the p-values of the Wilcoxon Sign Rank (WSR) test for the statistical significance. The results have the following indications: 1) the base-lift model generates the least accurate forecasts. 2) The ADL-intra model significantly outperform the ADL-own model, which highlights the value of competitive promotional information as suggested by Huang et al. (2014). 3) The ADL-own-EWC model outperforms the ADL-own model. 4) The ADL-intra-EWC model significantly outperforms the ADL-intra model. 5) The ADL-own-IC model outperforms the ADL-own model for short and moderate forecast horizons (e.g., for *h*=1 and *h*=4) but gets outperformed the ADL-own model for long forecast horizon (e.g., for *h*=12)[[10]](#footnote-10). 6) The ADL-intra-IC model outperforms the ADL-intra model for h=1 and h=4, but gets outperformed by the ADL-intra model for long forecast horizon (e.g., for *h*=12. The results suggest that the IC method may be more effective for short forecast horizons, which is consistent with the findings in other disciplines (e.g., [Pettenuzzo and Timmermann 2011](#_ENREF_66), [Chevillon 2016](#_ENREF_15)). Overall, The ADL-intra-EWC model and the ADL-intra-IC model generate the most accurate forecasts.

Table 3. The forecasting performance of the models

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |

Table 4. p-values for the Wilcoxon Sign Rank Test

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Benchmark | Candidates | 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 |

We also investigate the models’ forecasting performance for the time period depending on whether or not the focal product is being promoted. Table 5 shows the forecasting performance of the models for the promoted forecast period. The results are in line with those in Table 3. The only exception is that the base-lift method has surprisingly superior forecasting performance for the MAPE. This is because that a small number of SKU’s with very low sales volume has dominated the ranking of the MAPE. The result is different from that in Huang et al. (2014) where the forecasts are generated based on aggregated data across multiple stores and the MAPE may not easily be distorted. The results for the non-promoted period are also consistent with the results in Table 3 and we do not show it here for simplicity.

Table 5. The forecasting performance of the models (for the time period when the focal product is being promoted)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |

In Table 6, we compare the forecasting performance of the ADL-intra model, the ADL-intra-EWC model and the ADL-inter-IC model, for each individual product category. We select the three models because the ADL-intra-EWC model and the ADL-inter-IC model are the models with best overall forecasting performance and the ADL-intra model is the one which overlooks the issue of structural break. The forecasts are based on one to four weeks forecast horizon. Figure 8a and 8b show further details using boxplot for the MAPE for one to four week forecast horizon[[11]](#footnote-11). In the boxplot, positive values indicate (percentage) improvements by the ADL-intra-EWC model or the ADL-intra-IC model compared to the ADL-intra model.

Generally, both the ADL-intra-EWC model and the ADL-inter-IC models outperform the ADL-intra model for most of the 30 product categories. However, we do not observe improve forecasting accuracy for some product categories. For example, the ADL-intra-EWC model is outperformed by the ADL-intra model for 6 categories including *Cold Cereal, Hot dog, Margarine/Butter, Mayonnaise, Razors, Toilet Tissue*, and *Toothbrush,* while the ADL-intra-IC model is outperformed for 7 categories including *Carbonated Beverages, Cold Cereal, Face Tissue, Frozen Pizza, Hot Dog, Toothbrush,* and *Toothpaste.* One explanation is that the change of the effect of the price and promotions, the structural break, and the forecast bias may vary for different categories. As a result, the potential improvement may be submerged in the variations of the data.

Table 6. Comparing forecasting performance for each product category for one to four week forecast horizon

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Category | ADL-intra | | |  | ADL-intra-EWC | | |  | ADL-intra-IC | | |  |
| MAPE | MASE | SMAPE | RelAvgMAE | MAPE | MASE | SMAPE | RelAvgMAE | MAPE | MASE | SMAPE | RelAvgMAE |
| Beer | 53.8% | 0.717 | 38.4% | 0.988 | 53.2% | 0.712 | 38.2% | 0.932 | 53.3% | 0.719 | 38.2% | 1.101 |
| Blades | 50.6% | 0.780 | 35.0% | 0.991 | 49.9% | 0.772 | 34.9% | 0.983 | 50.1% | 0.771 | 35.1% | 0.983 |
| Carbonated Beverages | 57.5% | 0.505 | 37.4% | 0.987 | 59.0% | 0.504 | 37.7% | 0.983 | 62.3% | 0.506 | 38.5% | 0.966 |
| Cigarette | 52.2% | 0.900 | 37.4% | 0.987 | 52.1% | 0.890 | 37.1% | 1.003 | 53.1% | 0.893 | 37.2% | 1.016 |
| Coffee | 57.3% | 0.814 | 38.5% | 1.001 | 57.8% | 0.814 | 38.6% | 0.989 | 55.6% | 0.819 | 38.4% | 0.994 |
| Cold Cereal | 78.2% | 0.394 | 44.3% | 0.993 | 79.4% | 0.395 | 44.6% | 0.997 | 83.6% | 0.408 | 45.3% | 0.987 |
| Deodorant | 78.3% | 0.765 | 48.3% | 1.002 | 77.7% | 0.760 | 48.1% | 1.013 | 75.6% | 0.758 | 48.1% | 1.031 |
| Face Tissue | 39.8% | 0.647 | 27.5% | 0.992 | 38.3% | 0.634 | 26.7% | 0.983 | 41.7% | 0.671 | 28.0% | 0.976 |
| Frozen Dinner | 94.8% | 0.449 | 45.9% | 1.007 | 92.2% | 0.449 | 45.5% | 0.980 | 84.9% | 0.464 | 46.5% | 1.006 |
| Frozen pizza | 56.6% | 0.707 | 38.9% | 0.972 | 58.5% | 0.717 | 38.9% | 0.957 | 59.4% | 0.712 | 39.4% | 1.000 |
| Household Cleaner | 32.2% | 0.882 | 28.1% | 0.995 | 31.3% | 0.869 | 27.6% | 0.998 | 31.7% | 0.851 | 27.3% | 1.024 |
| Hotdog | 81.6% | 0.606 | 48.9% | 1.000 | 82.2% | 0.627 | 49.0% | 0.974 | 89.3% | 0.634 | 51.0% | 0.962 |
| Laundry Detergent | 75.3% | 0.538 | 45.6% | 1.004 | 73.7% | 0.530 | 45.1% | 1.006 | 73.3% | 0.526 | 44.6% | 1.073 |
| Margarine/Butter | 56.7% | 0.569 | 28.7% | 0.980 | 57.0% | 0.578 | 28.9% | 0.975 | 56.5% | 0.576 | 28.1% | 0.940 |
| Mayonnaise | 32.5% | 0.894 | 27.2% | 1.025 | 32.4% | 0.896 | 27.2% | 1.044 | 33.3% | 0.878 | 27.0% | 1.016 |
| Milk | 29.8% | 0.916 | 22.2% | 0.999 | 28.8% | 0.902 | 21.9% | 0.999 | 28.1% | 0.873 | 21.4% | 0.996 |
| Mustard & Ketchup | 46.1% | 0.730 | 33.2% | 0.992 | 45.0% | 0.724 | 33.1% | 0.975 | 45.0% | 0.736 | 32.7% | 0.948 |
| Paper Towels | 61.6% | 0.465 | 47.3% | 1.012 | 62.7% | 0.441 | 47.3% | 1.004 | 61.1% | 0.444 | 46.7% | 0.989 |
| Peanut butter | 32.4% | 1.144 | 28.4% | 1.009 | 31.4% | 1.142 | 28.1% | 0.993 | 33.1% | 1.077 | 27.0% | 0.984 |
| Photo | 71.1% | 0.681 | 48.2% | 0.998 | 69.4% | 0.667 | 47.7% | 0.983 | 65.0% | 0.654 | 46.0% | 0.940 |
| Razors | 69.8% | 0.680 | 43.4% | 1.000 | 72.1% | 0.693 | 44.1% | 1.034 | 67.5% | 0.676 | 44.3% | 1.003 |
| Salty snacks | 69.6% | 0.602 | 40.1% | 0.977 | 70.3% | 0.599 | 39.8% | 0.966 | 69.6% | 0.608 | 40.1% | 0.982 |
| Shampoo | 79.6% | 0.697 | 49.3% | 0.995 | 77.5% | 0.691 | 48.9% | 0.985 | 76.0% | 0.689 | 48.7% | 0.972 |
| Soup | 74.3% | 0.554 | 42.8% | 0.964 | 74.1% | 0.546 | 42.3% | 0.954 | 72.1% | 0.550 | 42.1% | 0.948 |
| Spaghetti sauce | 48.7% | 0.708 | 34.7% | 0.990 | 45.3% | 0.672 | 33.2% | 0.920 | 47.0% | 0.706 | 34.6% | 0.975 |
| Sugar substitutes | 53.3% | 0.787 | 36.8% | 0.986 | 52.9% | 0.779 | 36.5% | 0.977 | 50.5% | 0.734 | 35.4% | 0.944 |
| Toilet Tissue | 102.8% | 0.515 | 48.5% | 0.991 | 101.6% | 0.533 | 48.7% | 1.017 | 94.8% | 0.496 | 45.6% | 0.922 |
| Toothbrush | 95.4% | 0.747 | 52.1% | 0.993 | 97.2% | 0.755 | 52.6% | 1.009 | 97.7% | 0.780 | 53.5% | 1.037 |
| Toothpaste | 197.0% | 0.582 | 69.7% | 0.981 | 197.1% | 0.590 | 69.2% | 0.970 | 192.1% | 0.583 | 70.3% | 1.011 |
| Yogurt | 26.9% | 0.783 | 23.5% | 1.000 | 26.1% | 0.756 | 22.9% | 0.967 | 25.7% | 0.740 | 22.6% | 0.950 |

Figure 8a. The ADL-intra-EWC model versus the ADL-intra model, for the MAPE, for h=4

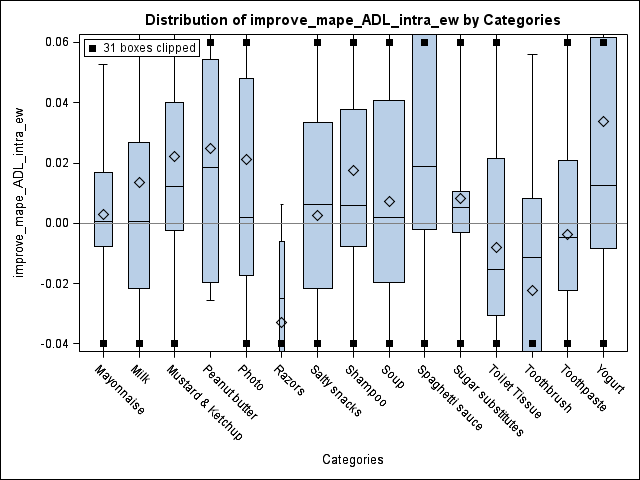
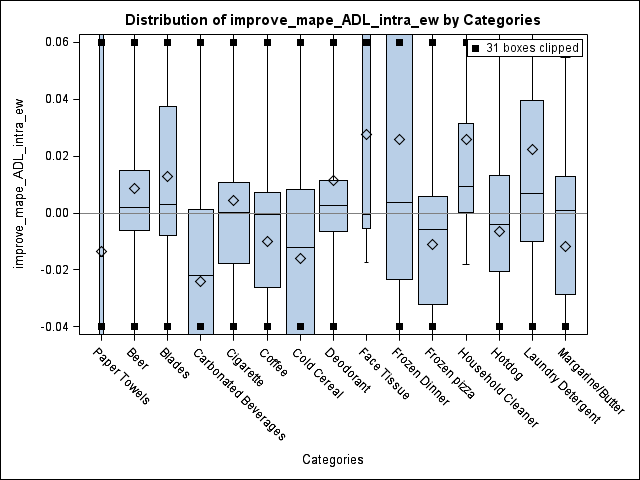
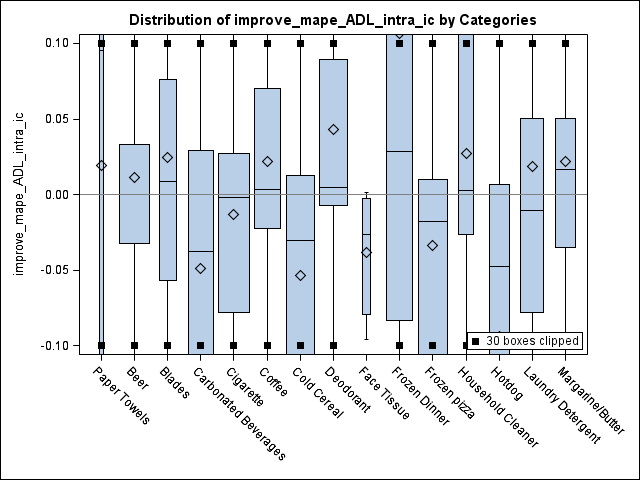
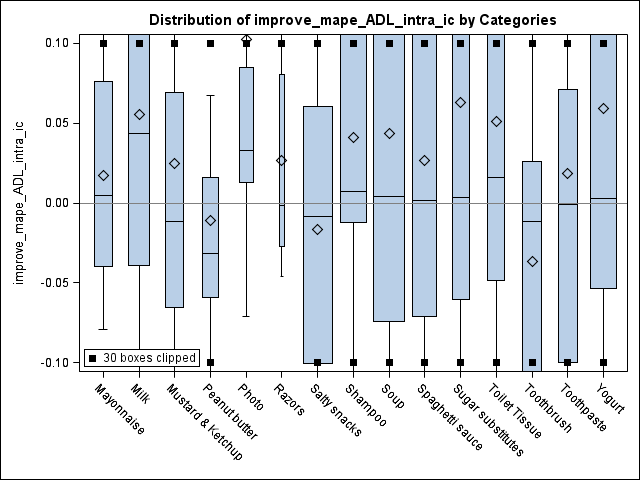


Figure 8b. The ADL-intra-IC model versus the ADL-intra model, for the MAPE, for h=4





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 effect 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_1)). 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 effect 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 effect of the marketing activities may be modeled by an autoregressive process of the marketing activities themselves. [Foekens, Leeflang et al. (1999)](#_ENREF_34) modeled the effect 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_38) and [Castle, Doornik et al. (2008)](#_ENREF_14). 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:**

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.

Ang, A. and G. Bekaert (2002). "Regime Switches in Interest Rates." Journal of Business & Economic Statistics **20**(2): 163-182.

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 (1998). "Estimating and Testing Linear Models with Multiple Structural Changes." Econometrica **66**: 47- 78.

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.

Bucklin, R. E., et al. (1998). "Determining Segmentation in Sales Response across Consumer Purchase Behaviors." Journal of Marketing Research **35**(2): 189-197.

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

Chevillon, G. (2016). "Multistep forecasting in the presence of location shifts." International Journal of Forecasting **32**(1): 121-137.

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.

Clemen, R. T. (1989). "Combining forecasts: A review and annotated bibliography." International Journal of Forecasting **5**(4): 559-583.

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.

Fan, J. and J. Lv (2008). "Sure independence screening for ultrahigh dimensional feature space (with discussion)." Journal of Royal Statistical Society **70**(Series B): 849–911.

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.

Jose, V. R. R. and R. L. Winkler (2008). "Simple robust averages of forecasts: Some empirical results." International Journal of Forecasting **24**(1): 163-169.

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.

Martin, R. and S. Kolassa (2009). Challenges of Automated Forecasting in Retail. International Symposium on Forecasting, Hong Kong.

Meeran, S., et al. (2017). "When do changes in consumer preferences make forecasts from choice-based conjoint models unreliable?" European Journal of Operational Research **258**(2): 512-524.

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>.

Perez-Quiros, G. and A. Timmermann (2000). "Firm Size and Cyclical Variations in Stock Returns." The Journal of Finance **55**(3): 1229-1262.

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. and A. Pick (2011). "Forecast Combination Across Estimation Windows." Journal of Business & Economic Statistics **29**(2): 307-318.

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 (2002). "Market timing and return prediction under model instability." Journal of Empirical Finance **9**(5): 495-510.

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.

Petropoulos, F., et al. (2014). "‘Horses for Courses’ in demand forecasting." European Journal of Operational Research **237**(1): 152-163.

Pettenuzzo, D. and A. Timmermann (2011). "Predictability of stock returns and asset allocation under structural breaks." Journal of Econometrics **164**(1): 60-78.

Song, H. and S. F. Witt (2003). "Tourism Forecasting: The General-to-Specific Approach." Journal of Travel Research **42**: 65-74.

Stock, J. H. and M. W. Watson (1996). "Evidence on Structural Instability in Macroeconomic Time Series Relations." Journal of Business and Economic Statistics **14**.

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.

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. 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-1)
2. This setting is very typical in a retailer context. In this example, we artificially make up the data series (but we keep the data series to be stationary). [↑](#footnote-ref-2)
3. In this example, for simplicity, we choose to illustrate the impact of structural breaks on forecasting accuracy using two structural breaks and also by holding the error variance to be constant before and after the breaks. Alternative settings (e.g., with different number of structural breaks and with changing error variance before and after the structural breaks) would provide the same indication. [↑](#footnote-ref-3)
4. In Figure 1, we use the blue area to represent the period before the first structural break (e.g., week [1,25]), use the yellow area to represent the period after the second structural break until the forecast origin (e.g., week [51, 75]), use the green area to represent the period between the two structural breaks (e.g., [26, 50]), and we use the red area to represent the forecast period (e.g., week [76, 100]). [↑](#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 a sequential Chow test, we conduct the Chow test assuming the break occurs at each individual week. [↑](#footnote-ref-5)
6. We would consider the model not being subject to structural breaks only when all the p-values are above the threshold. To mitigate the multiple comparison problem, we adopt very small threshold (e.g., 0.0001 rather than the usual 0.05) for the p-values. [↑](#footnote-ref-6)
7. We select the SKU’s which exhibit positive movements for at least 90% of the time. [↑](#footnote-ref-7)
8. 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. [↑](#footnote-ref-8)
9. 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-9)
10. 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-10)
11. For illustration purpose the boxplots are clipped to highlight the average values. For simplicity reason, we do not include the boxplots for other error measures and for alternative forecast horizons and they are consistent with the results in Table 6. [↑](#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)