**Forecasting Retailer Product Sales in The Presence of Structural Change**

Tao Huang[[1]](#footnote-1)

Surrey Business School, University of Surrey, GU2 7XH, UK

Robert Fildes

Centre for Marketing Analytics and Forecasting, Lancaster University, LA1 4YX, UK

Didier Soopramanien

School of Business and Economics, Loughborough University, Loughborough LE11 3TU

Abstract

Grocery retailers need accurate forecasts at Stock Keeping Unit (SKU) level to effectively manage their inventory. Previous studies have developed forecasting methods which incorporate the effect of various marketing activities including prices and promotions. These methods, however, have overlooked whether the effects of marketing activities on product sales may change over time. These methods may potentially be subject to the structural change problem as they are unable to capture any varying effect of the marketing activities. As a result, they could generate biased and less accurate forecasts. In this study, we propose effective forecasting methods for retailer product sales which take into account the problem of structural change. Our methods outperform conventional models based on a sample from a popular dataset for US retailers.

Keywords:

Analytics, Forecasting, OR in marketing, Retailing

1. **Introduction**

Grocery retailers rely on accurate sales forecasts to coordinate their supply chains (Syntetos, Babai, Boylan, Kolassa, & Nikolopoulos, 2016). Inaccurate forecasts of product sales lead to out-of-stock conditions or inflated costs due to overstocking. When a specific item is out-of-stock, retailers directly lose the profit from the sale of the item. If out of stocks situations happen on a regular basis, it can further lead to consumer dissatisfaction which, in the long term, can lead customers permanently switching to other retail chains (Corsten & Gruen, 2003). To avoid such situations, retailers may intentionally overstock to maintain a high customer satisfaction level. However, this significantly raises inventory costs (e.g., capital cost, warehousing, and deterioration, etc.) and reduces profits (L. Cooper, Baron, Levy, Swisher, & Gogos, 1999). In 2014, retailers in North America had a loss of $634.1 billion due to out-of-stock and spent $471.9 billion on overstock (OrderDynamics, 2015). One of the solutions to mitigate this dilemma is to generate more accurate sales forecasts at Stock Keeping Unit (SKU) level which improves the effectiveness of the supply chain management by reducing the bullwhip effect and enabling the Just-In-Time delivery (Ouyang, 2007; Sodhi & Tang, 2011).

Some recent studies have proposed effective methods to forecast retailer product sales at SKU level. For example, Gür Ali, SayIn, van Woensel, and Fransoo (2009) proposed the regression tree method with a range of variables constructed from sales, price, and promotion of the focal product. Huang, Fildes, and Soopramanien (2014) proposed a two-stage general-to-specific Autoregressive Distributed Lag (ADL) models. Their models incorporate the promotional information of not only the focal product but also of the promotional effect of competing products within the same product category. Ma, Fildes, and Huang (2016) further proposed a three-stage forecasting model which integrates the promotional information of the products from related product categories. The various models in the literature have recently been surveyed by Fildes, Ma, and Kolassa (2018).

These studies assume that the impact of marketing activities such as the price and promotions on product sales remains constant over time. In practice, the effect of prices and promotions on sales may change because of external non-controllable factors which may include, for instance, changing economic conditions, changes in consumer tastes and the entry of new competitors, introductions of new products, and terminations of existing products etc. Some of these effects are also neither observable or measurable (Wildt, 1976; Wildt & Winer, 1983). For example, customers can become more sensitive to prices and promotions during an economic crunch. Customers may also change their tastes due to their familiarity with the product and their changing lifestyles and social status (Meeran, Jahanbin, Goodwin, & Quariguasi Frota Neto, 2017). When a new competitor enters the market, the effect of prices and promotions of the focal product may decrease not only because of the marketing activities launched by the new competitor but also because customers seek variety. In the year of 2014, the German discounting retail chain Aldi opened more than 400 stores in the United States, leading to changes in customer grocery purchasing habits which exerted severe competitive pressure on other retail chains (Loeb, 2014).

Under any of the circumstances described above, forecasting models assuming constant effects of the price and promotions may potentially be subject to the problem of structural changes (Allen & Fildes, 2001). As a result, the forecasts generated by these models could be biased, which may potentially lead to lower forecast accuracy. The structural change problem has been addressed by previous studies (see a summary in Clements & Hendry, 1999) but overlooked in the domain of forecasting retailer product sales. In this study, we propose novel methods by taking into account the problem of structural change. Specifically, we examine the Autoregressive Distributed Lag (ADL) models with the Estimation Window Combining method and the ADL model with the Intercept Correction method. The former combines different sets of forecasts generated by the same ADL model but with different estimation windows. The latter makes corrections to the final forecasts based on the estimate of the forecast bias.

Our research falls in the domain of retail forecasting and makes the following contributions. First, our research is, as far as we are aware, the first to investigate the problem of structural change in forecasting retailer product sales. The empirical results based on the data suggest that our proposed methods have superior forecasting performance compared to conventional models which do not account for the problem of structural change. Second, our proposed methods focus on effectively utilizing available promotional information and thus do not incur additional costs for data collection. Third, our research provides an evaluation of various forecasting methods, the results of which offers operational guidance to not only retailers but also to manufacturers when competitive promotional information is unavailable. Fourth, the methods we propose are fully automatic and easy to implement. Finally, the focus on structural change in the retailer context offers exploratory insights into those situations where our proposed methods work more effectively compared to models with similar specifications but overlook the problem of structural change.

The remainder of the paper is organized as follows: section 2 summarizes previous studies which forecast retailer product sales at SKU level and also summarize the effect of marketing activities including price and promotions. Section 3 explains the structural change problem and the rationale of the methods which may potentially mitigate the problem. Section 4 explores the data. In section 5, we describe our new three-stage forecasting methods. Section 6 introduces the design of the model evaluation. Section 7 summarizes and discusses the evaluation results to provide a convincing demonstration of their performance. In Section 8, we explore the characteristics of the situations where the proposed methods garner the greatest improvements compared to models with similar specifications but overlook the problem of structural change. In the last section, we make recommendations for both manufacturers and retailers, address research limitations, and highlight directions for future research.

## Literature review

2.1 Forecasting retailer product sales at SKU level

In practice, many retailers still forecast their product sales at the SKU level using a two-stage ‘Base-lift’ method. The method (and the software) entails dividing the data into promoted and non-promoted periods based on whether the focal SKU is being promoted. Specifically, they use simple univariate methods to generate the ‘baseline’ forecasts for the non-promoted period and then make adjustments for the ‘lift’ effect of any incoming promotional events. They estimate the ‘lift’ effect of the promotional events relying on the experience of the brand/category managers (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009; Fildes, Nikolopoulos, Crone, & Syntetos, 2008). A stream of studies have been devoted to helping managers to effectively tackle their own biases typically reflecting their understanding of the market conditions (Lee, Goodwin, Fildes, Nikolopoulos, & Lawrence, 2007; Petropoulos, Fildes, & Goodwin, 2016). Other studies try to estimate the ‘lift’ effect with model-based forecasting systems. For example, the PromoCast™ system relates the ‘lift’ effect to previous promotions of the focal product, the characteristics of product categories and stores, and manufacturer information etc. (L. Cooper et al., 1999; L. G. Cooper & Giuffrida, 2000; Trusov, Bodapati, & Cooper, 2006). Aburto and Weber (2007) used neural network models to estimate the ‘lift’ effect for the product sales for a Chilean supermarket. One limitation of these two-stage 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. Other studies have proposed integrated methods to directly generate the final forecasts. Kuo (2001) used neural network models to forecast product sales of daily milk in convenience stores. Gür Ali et al. (2009) proposed the regression tree method and the support vector regression (SVR) method to forecast retailer product sales for the non-perishable food categories at SKU level. Their models incorporate variables constructed based on statistical measures of past information (e.g., the sales, prices, and promotions of the focal product). Their regression tree method has the best forecasting performance. However, it gets beaten by the Base-lift method for the time periods when the focal product is not being promoted. One of the limitations for the model is that it overlooks the effect of competitive promotions on the sales of the focal product. Divakar et al. (2005) proposed the CHAN4CAST method to forecast product volume sales for beverage manufacturers. Their method incorporates the promotional information of the main competitors of the focal product. However, their method is only applicable when there are a small number of competitors (e.g., just like Coca *versus* Pepsi). Huang et al. (2014) proposed two-stage Autoregressive Distributed Lag (ADL) methods to forecast retailer product sales at SKU level. Their methods were the first to account for the competitive promotional information for the whole product category. They initially implemented a variable selection procedure to identify the most important variables for the competitive activities within the product category. They then specified general-to-specific ADL models based on these selected variables. Their methods has been found the best forecasting performance for five grocery categories such as *Bottled Juice*, *Soft Drinks*, and *Bath Soap* etc. Ma et al. (2016) proposed three-stage ADL methods which further integrate the promotional information not only from the same product category but also from other related product categories. Their methods are extensions of those in Huang et al. (2014) and also benefit from an automatic model specification procedure. Their methods outperform the Base-lift benchmark model for 15 food product categories. These studies suggest that promotional information are valuable in forecasting retailer product sales, and evidence shows that modern commercial software has also started to integrate promotional information (Fildes et al., 2018). However, all the studies described here assume constant effects of the marketing activities.

2.2 The effect of marketing activities including price and promotions

Previous studies have summarized the effect of marketing activities on product sales. For example, early studies have found that product sales can be increased in the short term by price reductions and promotions (e.g., Blattberg, Briesch, & Fox, 1995; Christen, Gupta, Porter, Staelin, & Wittink, 1997; L. Cooper et al., 1999; Gupta, 1988; Gür Ali et al., 2009; Lattin & Bucklin, 1989; Mulhern & Leone, 1991). Product sales after the price reduction and promotions may decrease because customers may stockpile their purchases (Mace & Neslin, 2004; H. J. Van Heerde, Gupta, & Wittink, 2003). Product sales may be negatively affected by the marketing activities of competitive products (Demirag, Keskinocak, & Swann, 2011; Rudolph W. Struse, 1987; Walters, 1991; Walters & Rinne, 1986). The effect of competitive marketing activities may not come from products within the same category but also from related categories. (R. L. Andrews, Currim, Leeflang, & Lim, 2008; Wedel & Zhang, 2004).

Further evidence also shows that the effect of marketing activities such as prices and promotions may change over time. For example, Wildt (1976) and Wildt and Winer (1983) suggest that the effect of the marketing activities may change due to the change in economic conditions, consumer tastes, and the competition environment, etc. Customers may find price reductions and promotions more attractive during the period of an economic crunch compared to other time periods. Mahajan, Bretschneider, and Bradford (1980) found that the effect of prices and promotions change during the different stages of the product lifecycle. Meeran et al. (2017) found that customers have different tastes and preferences when they accumulate more knowledge of the product, when they seek variety, and when they reach a different social status and then decide to adopt a different lifestyle. These individual changes lead to substantial aggregate effects on product sales. Other studies find that the introduction of store-own brands in a product category decreases the promotional elasticities of premium national brands and increase promotional elasticities of the second tier national brands (e.g., Nijs, Dekimpe, Steenkamps, & Hanssens, 2001; H. J. Van Heerde, Srinivasan, & Dekimpe, 2008). The effect of the marketing activities can also change depending on how retailers communicate their marketing events. For example, they may promote the products through mobile applications and adopt new prominent promotion shelf tags, which could makes the promotions more effective (H. van Heerde, M. Dinner, & Neslin, 2015). In practice, retailers may record their marketing activities using aggregate terms such as Features and Displays (e.g., Bronnenberg, Kruger, & Mela, 2008). However, these terms may have various forms such as Buy One Get One free (BOGO), store flyers, billboard advertising, and temporary price reduction (TPR) for shopper card holders only etc. Under such conditions, the effect of the marketing activities may differentiate.

## The problem of structural change

The problem of structural change has been addressed by previous forecasting studies[[2]](#footnote-2) (e.g., Castle, Doornik, & Hendry, 2008; Hendry, 2018; Pesaran & Timmermann, 2007). Pesaran and Timmermann (2005) demonstrated analytically how a structural change leads to biased forecasts using a simple regression model without an intercept. For example, suppose that for the time periods of , the unobserved data generating process is:

(1)

where, and are the vectors of the dependent variable and independent variable respectively. is the vector of the error term. (where *i*=1,2) are the vectors of the parameter coefficients. is an indicator which equals to 1 before week (where ) and 0 afterwards. Therefore, we have a structural change where the true parameter of the independent variable changes from to after . We can estimate a model with a functional form congruent with the data generating process (e.g., ) based on the data before and after the structural change, e.g., ,. Thus, the OLS estimate of the parameter is:

(2)

where and are respectively the vectors of the dependent variable and independent variable for the time periods from week *m* to week *T*. We assume that there is no structural change after week *T*. e.g., . Suppose that we are interested in the one-step ahead forecast, the one-step ahead forecast error is:

(3)

where is the vector of the independent variable for the time periods from week *m* to . is the vector of error term for the time periods from week *m* to *T*. is the error term at week . Therefore, the expected value of the equation (3) is:

. (4)

Equation (4) is unequal to zero as is unequal to . This indicates that the forecast at week is biased. For more general cases where the model has an intercept term and endogenous explanatory variables, the forecast bias can be demonstrated using Monte Carlo simulation (see Clements & Hendry, 1999; Pesaran & Timmermann, 2005, 2007)[[3]](#footnote-3).

In this study, we implement two methods to mitigate the problem of structural change. The first method is the Intercept Correction (IC) method which specifies non-zero values for the model’s errors in the forecast period (Clark & McCracken, 2007; Clements & Hendry, 1994, 1999). If we identify that the model is subject to structural changes, we may try to estimate the forecast bias, e.g., by taking the average value of those most recent residuals, e.g., , where is the number of residuals. When , the estimated bias reduces to , which is the residual at the forecast origin (e.g., Chevillon, 2016). Ideally, we can simply add the estimated bias back to the out-of-sample forecasts. However, the IC method has a limitation. In practice, sales at SKU level sometimes exhibit large variations and unexpected outliers, which renders the task of estimating the forecast bias challenging. For example, the bias can be submerged by high variations in the product sales. As a result, it is possible that the average value of the most recent residuals may mostly represent random variations. Also, by adding the estimated bias back to the out-of-sample forecasts, we inevitably incur the cost of inflated forecast error variance (see the analytical evidence in Clements & Hendry, 1999). The second method is the Estimation Window Combining (EWC) method which combines the forecasts generated by the same model but with different estimation windows (e.g., Pesaran & Pick, 2011; Pesaran, Schuermann, & Smith, 2009; Pesaran & Timmermann, 2005). More specifically, we can combine those forecasts with equal weights as it has been found effective and easy to implement.(Clements & Hendry, 1998; Dekker, van Donselaar, & Ouwehand, 2004; Fildes & Stekler, 2002; Pesaran et al., 2009). In the example demonstrated in equation (1), we may estimate the model using the most recent observations to generate the first set of forecasts, e.g., , where represents the parameters estimated based on the observation window . The value of can be arbitrarily chosen given there are enough observations to estimate the model and enough variations in the explanatory variable. We then add more observations (e.g., one) to the estimation window and generate the second set of forecasts, e.g., , and so forth, until we generate the set of forecasts based on the estimation window . Thus, we may equally combine those forecasts to generate the final forecasts:

(5)

Pesaran and Timmermann (2007) show analytic evidence that, for the example in equation (1), the forecasts generated by the models with smaller estimation windows tend to be less biased (e.g., the models will utilize fewer observations before the structural change). However, these forecasts inevitably bear a cost of inflated forecast error variance. This is because they are generated when the model is estimated with less data especially if the data before the structural change are more informative. The EWC method thus tries to generate more accurate forecasts by accepting a trade-off between the reduced forecast bias and the inflated forecast error variance. Compared to the IC method, the EWC method does not estimate the size of the bias.

The two methods described above have been found good forecasting performance by previous studies. For example, the IC method has good performance in forecasting wage, unemployment, and CPI inflation etc. (e.g., Clark & McCracken, 2007; Clements & Hendry, 1996), and the EWC method has good forecasting performance for exchange rate, inflation, and equity index futures etc. (e.g., Pesaran & Pick, 2011; Pesaran et al., 2009; Rapach & Strauss, 2008). For retailer product sales, whether accounting for structural change and which of the two methods, the IC method and the EWC method, could generates more accurate forecasts becomes empirical questions.

## The data

We evaluate the forecasting performance of various models using the retail dataset made available by the Information Resources, Inc. (IRI) company. A more comprehensive description of the dataset can be found in Bronnenberg et al. (2008). The dataset contains weekly data at SKU level with variables including product unit sales, price, features, and displays, etc. We initially conduct our evaluation based on 1831 SKU’s for 28 product categories from 28 different stores. We select the SKU’s for the same category from the same store, and we select the SKUs with positive movements for at least 90% of the time. Table 1 shows the basic statistics for the selected SKU’s during a period of 202 weeks for each product category, which suggests a wide variety in the marketing activities across the different categories. Figure 1 shows the data series for a typical SKU in the Beer category. e.g., the product sales spikes are usually associated with the price reductions and feature/display promotions of the focal product, as well as calendar events (e.g., Halloween, Thanksgiving, and Christmas, etc.).

Table 1. Statistical descriptions for each product category

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Category | Price mean | Sales mean\* | Display percentage\*\* | Feature percentage\*\*\* | Number of SKU's |
| Beer | 8.3 | 20.6 | 13.9% | 4.0% | 169 |
| Blades | 8.1 | 14.6 | 7.4% | 2.2% | 22 |
| Carbonated Beverages | 2.1 | 113.6 | 26.8% | 15.6% | 82 |
| Cigarette | 22.3 | 22.2 | 0.0% | 0.8% | 203 |
| Coffee | 5.2 | 14.5 | 5.2% | 2.9% | 86 |
| Cold cereal | 3.5 | 70.7 | 4.0% | 18.1% | 125 |
| Deodorant | 2.7 | 6.9 | 4.1% | 5.2% | 126 |
| Face Tissue | 2.1 | 75.8 | 0.3% | 11.7% | 6 |
| Frozen Dinner | 2 | 43.8 | 5.3% | 23.7% | 87 |
| Frozen pizza | 3.4 | 31.2 | 8.9% | 9.1% | 147 |
| Household Cleaner | 2.5 | 29.9 | 0.3% | 3.6% | 18 |
| Hotdog | 4 | 68.6 | 13.2% | 15.6% | 35 |
| Laundry Detergent | 8.8 | 28.9 | 2.3% | 8.8% | 57 |
| Margarine/Butter | 2 | 71.4 | 0.1% | 6.3% | 36 |
| Mayonnaise | 3 | 79.7 | 3.0% | 0.4% | 22 |
| Milk | 2.5 | 222.3 | 2.1% | 1.8% | 30 |
| Mustard & Ketchup | 2.1 | 64.5 | 5.3% | 0.9% | 22 |
| Peanut butter | 3.7 | 34.2 | 3.2% | 0.6% | 15 |
| Photo | 7.2 | 9.2 | 4.6% | 5.1% | 13 |
| Salty snacks | 2.3 | 50.9 | 6.7% | 5.0% | 101 |
| Shampoo | 3.5 | 9.9 | 12.8% | 7.1% | 70 |
| Soup | 1.5 | 61.6 | 1.2% | 9.7% | 139 |
| Spaghetti sauce | 2.4 | 39.1 | 1.6% | 6.5% | 52 |
| Sugar substitutes | 2.8 | 14.5 | 0.1% | 1.4% | 20 |
| Toilet Tissue | 5.4 | 89.1 | 4.3% | 8.3% | 20 |
| Toothbrush | 2.6 | 8.7 | 3.1% | 6.3% | 28 |
| Toothpaste | 2.8 | 35.5 | 11.0% | 12.5% | 25 |
| Yogurt | 1.1 | 115.1 | 0.7% | 6.3% | 75 |

\* Sales mean represents the average unit sales for all the SKU’s for the category for the specific store.

\*\* \*\*\*Display percentage and Feature percentage indicate the percentage of weeks during the 202-week time periods when the focal product is being promoted for Display and Feature.

Figure 1. Store level data for an SKU in the Beer category



In Figure 1, week 1 indicates the first week in the year of 2001. 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.

## Methodology

In this study, we propose two novel forecasting methods for retailer product sales at SKU level. Our methods consider the problem of structural change. The methods consist of three stages. During the first stage, we identify the most relevant competitive explanatory variables for the focal product within the product category. Grocery retailers typically sell hundreds of SKU’s in a typical product category, 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 render the estimation task infeasible (Martin & Kolassa, 2009). Therefore, we initially select the most relevant competitive explanatory variables using the Least Absolute Shrinkage and Selection Operator (LASSO) procedure (Tibshirani, 1996). That is, we construct the following model for each SKU:

(6)

where represents log product sales of the focal product for the store at week *t.* is the matrix for the explanatory variables including prices, features, and displays of all the products in the same product category. *u* represents the identically distributed error term. represents the vector of the parameter coefficients. *N* is the total number of SKUs for the category. is the shrinkage factor. The LASSO procedure imposes a constraint to the sum of the absolute values of the models’ parameter coefficients. It removes the less relevant explanatory variables by pushing their parameter coefficients towards zero. We control the model simplification process using the shrinkage factor based on 10-fold cross validation (Ma & Fildes, 2017; Ma et al., 2016)[[4]](#footnote-4).

During the second stage, we construct the General Autoregressive Distributive Lag (ADL) model following Huang et al. (2014) by incorporating the variables retained by the LASSO procedure during the first stage. The LASSO procedure has a limitation that it may potentially miss important variables especially under the condition of high multicollinearity (Fan & Lv, 2008; Ma et al., 2016). Previous studies suggest the product sales are usually mostly influenced by the prices and promotions of the products themselves (Bucklin, Gupta, & Siddarth, 1998). Thus, we intentionally incorporate the prices and promotions of the focal product in the general ADL model even these variables were not retained by the LASSO procedure during the first stage. We also incorporate the dynamic effects of these explanatory variables as well as a time variable to capture the potential trend, twelve deterministic four-week dummy variables to capture seasonality, and other dummy variables to capture calendar events. The constructed general ADL model for each product in a specific store can be demonstrated as follows:

where is the log sales of the focal product at week . is the term which captures any potential trend during the estimation period (Song & Witt, 2003). and represent the log price of the focal product and a competitive product, *m*, at week . and represents the feature and the display dummy variables 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*[[5]](#footnote-5)*. are the parameters.  
 is the error term and is assumed that . is the order of the lags and is set as 2. *, ,* and are the numbers of selected competitive price, Feature, and Display variables for the product category.

The general ADL model, as shown in equation (7), could have too many explanatory variables and lack parsimony. Thus, we simplify the model using the LASSO procedure following Ma et al. (2016) (we refer to the resulted model as the ADL-raw model thereafter). During this stage, we use the LASSO procedure as a model specification strategy rather than a variable selection method as previous studies indicate that models simplified by the LASSO procedure could have good forecasting performance and outperform traditional models specified based on statistical significance (Epprecht, Guegan, & Veiga, 2013; Ma et al., 2016). Also, the LASSO procedure enables the automation of the statistical forecasting task which becomes essential as typically grocery retailers stock a tremendous number of SKUs (L. Cooper et al., 1999). To prevent the LASSO procedure missing important variables, we initially construct a supplementary parallel ADL model which has a similar specification compared to the general ADL model but only includes the price and promotion variables of the focal product:

(8)

We simplify the supplementary parallel ADL model shown in equation (8) using the LASSO procedure (we refer to the resulted model as the ADL-own model thereafter). We then incorporate the marketing variables retained in the ADL-own model into the ADL-raw model (we refer to the resulted model as the ADL-intra model). This enables us to selectively retain the potentially important variables in the ADL-intra model, such as the price and promotions of the focal product and their dynamic terms. If these variables get removed by the ADL-raw model, they will be added back to the ADL-intra model if they are retained by the ADL-own model. That is, we try to prevent the ADL-intra model from missing important variables at the cost of reduced efficiency. The supplementary parallel ADL model, i.e., in equation (8), by definition, has fewer explanatory variables compared to the general ADL model, i.e., in equation (7), and is less likely to suffer from multicollinearity compared to the latter. Thus, if the price and promotions of the focal product truly have effects on the product sales, it is less likely that they will be removed by both the ADL-raw model and the ADL-own model[[6]](#footnote-6).

Figure 2. An illustration for the three-stages of our proposed methods



During the final stage, we integrate the ADL-intra model with the EWC method and the IC method respectively to account for the structural change problem. We implement the EWC method and the IC method to the ADL-intra model only if the presence of the structural change is confirmed. If this is not the case, we keep the forecasts generated by the ADL-intra model as the final forecasts. In this study, we conduct a sequential Chow test for up to 95% of the weeks in the estimation period. That is, if we have an estimation period of 160 weeks, we conduct the Chow test for each of the 152 weeks. For example, we initially conduct the Chow test assuming a structural change occurring at week 5 and we obtain the corresponding p-value. We then conduct the Chow test for week 6, 7, and so forth until week 156 and each time we obtain the p-value accordingly. We keep at least 5% of the weeks for the estimation of the test. Thus, we may obtain up to 152 p-values in total. The null hypothesis of no structural change will be rejected if any of these p-values is below a threshold. To mitigate the multiple comparison problem, we adopt a very small threshold, i.e., 0.001[[7]](#footnote-7).Previous studies have proposed alternative tests which focus on estimating multiple structural changes and their locations and are usually associated with very stringent assumptions (e.g., Donald W K Andrews, 1993; Donald W. K. Andrews & Ploberger, 1994; Bai & Perron, 1998, 2003; Brown, Durbin, & Evans, 1975). In our study, we only need to know if structural change is present in our data. Thus, we conduct the sequential Chow test which is appropriate for that purpose and is also simple to implement. We refer to the final resulting models as the ADL-intra-EWC model and the ADL-intra-IC model respectively. Figure 2 provides a summary guide for the implementation of the ADL-intra-EWC model and the ADL-intra-IC model.

## The experimental design

In this study, we consider the Base-lift method as the benchmark model. The method is widely used in practice, and its forecasting performance has been evaluated in previous studies (e.g., L. Cooper et al., 1999; Gür Ali et al., 2009; Huang et al., 2014; Ma et al., 2016). The forecasts for week *t* by this method can be described as follows:

(9)

where represents the baseline forecast for week by the simple exponential smoothing (SES) model. The SES model is estimated exclusively based on the data when the focal product is not being promoted. Thus, represents the sales of the focal product for the previous time when the focal product was not promoted. is the smoothing parameter of the simple exponential smoothing model, and is estimated by minimizing the in-sample mean squared errors. The adjustment for the ‘lift’ effect is calculated as the increased sales of the focal product during its most recent promotion compared to the corresponding baseline sales. We also have the following candidate models:

1. The ADL-own model, i.e., the model in equation (8) simplified by the LASSO procedure
2. The ADL-intra model; i.e., the model in equation (7) simplified by the LASSO procedure and then include the marketing variables retained in the ADL-own model.
3. The ADL-own-EWC model: the ADL-own model implemented with the EWC method
4. The ADL-own-IC model: the ADL-own model implemented with the IC method
5. The ADL-intra-EWC model: the ADL-intra model implemented with the EWC method
6. The ADL-intra-IC model: the ADL-intra model implemented with the IC method

We specify the models with an estimation window of 160 weeks, and we evaluate their forecasting performance using 18 rolling origins for robustness (Tashman, 2000). For each rolling event, we move the estimation window two weeks forward and re-specify the model. We assume that the value of the price and any promotional information to be known as it is part of the retailer’s inventory plan. We use the forecast value of product sales when the forecast horizon is beyond one week. We generate one to week-ahead forecasts, where is 1, 4, and 8, to approximate the situation retailers face in practice. For the EWC method, we generate the final forecasts by equally combining the forecasts by the same model with ten estimation windows (e.g., for the estimation period, e.g., [1,160], we estimate the model with ten estimation windows including [1, 160], [3, 160], and so forth, until [19, 160]). This generates ten sets of forecasts). For the IC methods, we estimate the forecast bias as the average value of the sixteen most recent residuals and add the value to the forecasts of all the forecast horizons. We implement the models using the MODEL procedure with macros in SAS 9.4. The model parameters are estimated using the OLS estimator.

We evaluate the models with different error measures which approximate the unknown loss function of the retailer from different aspects. We include traditional error measures including the Mean Absolute Error (MAE), the symmetric Mean Absolute Percentage Error (sMAPE) and the scaled Mean Squared Error (scaled MSE)[[8]](#footnote-8). We also include more recently developed error measures including the Mean Absolute Scaled Error (MASE) and the Relative Average Mean Absolute Error (RelAvgMAE) respectively developed by Hyndman and Koehler (2006) and Davydenko and Fildes (2013). Such relative measures have more desirable properties, e.g., equally penalize positive and negative errors, more robust to outliers while the latter is readily interpretable as the percentage improvement (or worsening) of the focal method compared to a benchmark. The two latter error measures can be demonstrated as follows:

(10)

, where ,

(11)

where and are the MASE and the AvgRelMAE based on one to *H* forecast horizon (=1, 4 and 8) across SKUs (e.g., *S*= 1831) for *K* rolling events (e.g., *K*=18). and are respectively the *h*-step ahead actual value and forecast value for data series based on the rolling event. is the total number of observations in the estimation window (i.e., ). Before we transform the log values to levels for evaluation, we adjust the final forecasts by adding one-half mean squared error, which mitigate the bias caused by the logarithm transformation (e.g., L. Cooper et al., 1999; Ma et al., 2016).

## Results and discussion

In Table 2, we summarize the forecasting performance of the models across all the products. Table 3 shows the results of the Diebold-Mariana (DM) test for the statistical significance of the difference between the models’ forecasting performance. (Diebold & Mariano, 1995; Harvey, Leybourne, & Newbold, 1997)[[9]](#footnote-9). The following findings emerge from the analysis:

1. The Base-lift model generates the least accurate forecasts across all the error measures.
2. The ADL-intra model outperforms the ADL-own model across all the error measures, which is consistent with the findings in Huang et al. (2014).
3. The ADL-own-EWC model outperforms the ADL-own model for all the error measures.
4. The ADL-own-IC model generally outperforms the ADL-own model except for the MAE.
5. The ADL-intra-EWC model outperforms the ADL-intra model for all the error measures.
6. The ADL-intra-IC model generally outperforms the ADL-intra model except for the MAE and the scaled MSE for longer forecast horizons (e.g., Forecast horizon is one to four weeks ahead and one to eight weeks ahead).
7. Overall, The ADL-intra-EWC model and the ADL-intra-IC model generate the most accurate forecasts.

Table 2. The forecasting performance of the models for all forecast period

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Forecast horizon is one to eight weeks ahead | | | | | |
| Model/measure | MAE | sMAPE | MASE | AvgRelMAE | scaled MSE |
| Base-lift | 22.92 | 46.98% | 0.775 | 1.1444 | 0.2234 |
| ADL-own | 15.75 | 40.81% | 0.697 | 1.0000 | 0.1575 |
| ADL-intra | 15.44 | 40.51% | 0.695 | 0.9941 | 0.1553 |
| ADL-own-EWC | 15.67 | 40.68% | 0.696 | 0.9956 | 0.1570 |
| ADL-own-IC | 16.23 | 40.76% | 0.694 | 0.9992 | 0.1596 |
| ADL-intra-EWC | **15.35** | **40.41%** | 0.694 | **0.9905** | **0.1548** |
| ADL-intra-IC | 15.59 | 40.46% | **0.693** | 0.9936 | 0.1568 |
| Forecast horizon is one to four weeks ahead | | | | | |
| Model/measure | MAE | sMAPE | MASE | AvgRelMAE | scaled MSE |
| Base-lift | 22.67 | 46.24% | 0.762 | 1.1365 | 0.2186 |
| ADL-own | 15.63 | 40.45% | 0.69 | 1.0000 | 0.1548 |
| ADL-intra | 15.16 | 40.12% | 0.686 | 0.9913 | 0.1514 |
| ADL-own-EWC | 15.55 | 40.31% | 0.688 | 0.9950 | 0.1540 |
| ADL-own-IC | 15.94 | 40.25% | 0.684 | 0.9948 | 0.1553 |
| ADL-intra-EWC | **15.09** | 40.01% | 0.685 | 0.9876 | **0.1509** |
| ADL-intra-IC | 15.21 | **39.93%** | **0.681** | **0.9871** | 0.1517 |
| Forecast horizon is one week ahead | | | | | |
| Model/measure | MAE | sMAPE | MASE | AvgRelMAE | scaled MSE |
| Base-lift | 24.99 | 45.42% | 0.762 | 1.1279 | 0.2261 |
| ADL-own | 16.66 | 39.87% | 0.689 | 1.0000 | 0.1561 |
| ADL-intra | 15.66 | 39.43% | 0.686 | 0.9883 | 0.1529 |
| ADL-own-EWC | 16.59 | 39.72% | 0.686 | 0.9955 | 0.1549 |
| ADL-own-IC | 17.02 | 39.52% | 0.68 | 0.9902 | 0.1552 |
| ADL-intra-EWC | **15.59** | 39.33% | 0.684 | 0.9850 | 0.1523 |
| ADL-intra-IC | 15.65 | **39.15%** | **0.679** | **0.9804** | **0.1520** |

We also investigate the models’ forecasting performance for the time periods depending on whether the focal product is being promoted. This is because that retailer product sales tend to exhibit very high levels of variations when the focal product is being promoted, and tend to be comparably stable otherwise (Gür Ali et al., 2009). We refer these two periods as the promoted period and non-promoted period respectively afterward. Table 4 shows the forecasting performance of the models for the promoted forecast period and the non-promoted forecast period respectively for one to eight-week forecast horizon[[10]](#footnote-10). The following are particularly important. The ADL-intra-IC model has the best forecasting performance for the non-promoted period but only has moderate performance for the promoted period. A possible explanation is that the estimated bias added back to the error term in the forecast period may get submerged by the high variations of the product sales when the focal product is being promoted. In contrast, the ADL-intra-EWC model has the best performance for the promoted period. Therefore, we develop an exploratory combined method between these two methods, named as the ADL-EWC-IC model. The ADL-EWC-IC model is identical to the ADL-intra-EWC model for the promoted period and identical to the ADL-intra-IC model for the non-promoted period. To allow for a fair comparison, we evaluate the performance of the ADL-EWC-IC model based on previously unseen data (e.g., the data are based on 1605 SKU’s for the same 28 product categories but from a new set of 28 stores). Table 5 shows the forecasting performance of the ADL-EWC-IC model compared to other three models[[11]](#footnote-11). The exploratory results indicate that the ADL-EWC-IC model generally generates the most accurate forecasts across all the models even when we consider previously unseen data.

We further explore the percentage reduction of the MASE by the ADL-intra-EWC method and the ADL-intra-IC method compared to the ADL-intra model for each product category. The comparison highlights the value for taking consideration of the structural change problem as the ADL-intra model has a similar specification compared to the two proposed methods but overlooks the problem of structural change. We calculate the percentage reductions of the MASE by the ADL-intra-EWC method and by the ADL-intra-IC method for product as follows:

(12)

(13)

We then take the average value of and respectively across all the SKU’s for each product category. Table 6 shows the results for each product category for one to eight weeks forecast horizon. The ADL-intra-EWC method and the ADL-intra-IC method outperform the ADL-intra model for most of the product categories (e.g., 18 and 16 respectively, out of 28 categories). They do not outperform the ADL-intra model for all product categories due to the heterogeneity of the data characteristics across different product categories (Ma et al., 2016). The comparison results for other error measures and horizons are similar and we omit them for simplicity. For each proposed method, we highlight the product categories where the method has its highest advantages compared to the ADL-intra model. For example, the ADL-intra-EWC method has its highest advantages compared to the ADL-intra model for six product categories including *Spaghetti sauce*, *Face Tissue*, and *Toothpaste* etc. Figure 3(a) shows the distributions of the percentage reduction of the MASE by the ADL-intra-EWC method compared to the ADL-intra model for these categories. The ADL-intra-IC method has its highest advantages compared to the ADL-intra model for another six product categories including *Peanut butter, Milk, Yogurt,*and *Toilet Tissue* etc. Figure 3(b) shows the distributions of the percentage reduction of the MASE by the ADL-intra-IC method compared to the ADL-intra model for these categories.

Table 3. The results of the Diebold-Mariana (DM) test

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model 1 | Model 2 | MAE | | | sMAPE | | | MASE | | | scaled MSE | | |
| *H*=1 | *H*=1 to 4 | *H*=1 to 8 | *H*=1 | *H*=1 to 4 | *H*=1 to 8 | *H*=1 | *H*=1 to 4 | *H*=1 to 8 | *H*=1 | *H*=1 to 4 | *H*=1 to 8 |
| ADL-own | Base-lift | 0.000\* | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| ADL-own | ADL-intra | 0.000 | 0.001 | 0.015 | 0.000 | 0.000 | 0.000 | 0.233 | 0.026 | 0.157 | 0.443 | 0.380 | 0.453 |
| ADL-own | ADL-own-EWC | 0.106 | 0.005 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.104 | 0.294 | 0.148 | 0.335 | 0.258 |
| ADL-own | ADL-own-IC | 0.064 | 0.008 | 0.000 | 0.000 | 0.000 | 0.259 | 0.000 | 0.000 | 0.009 | 0.388 | 0.138 | 0.001 |
| ADL-intra | ADL-intra-EWC | 0.138 | 0.013 | 0.002 | 0.000 | 0.000 | 0.000 | 0.005 | 0.124 | 0.100 | 0.652 | 0.259 | 0.308 |
| ADL-intra | ADL-intra-IC | 0.946 | 0.469 | 0.021 | 0.000 | 0.000 | 0.277 | 0.000 | 0.000 | 0.030 | 0.169 | 0.011 | 0.001 |

\*0.000 indicates that the p-value is smaller than 0.001.

Table 4. The forecasting performance of the models for the promoted and non-promoted forecast period

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Forecast horizon is one to eight weeks ahead, for the promoted period | | | | | |
| Model/measure | MAE | sMAPE | MASE | AvgRelMAE | scaled MSE |
| Base-lift | 119.33 | 87.26% | 1.915 | 1.3705 | 2.4742 |
| ADL-own | 65.27 | 47.56% | 1.329 | 1.0000 | 1.0719 |
| ADL-intra | 63.10 | 46.04% | 1.307 | 0.9795 | 1.0265 |
| ADL-own-EWC | 65.01 | 47.43% | 1.325 | 0.9955 | 1.0662 |
| ADL-own-IC | 69.68 | 47.95% | 1.354 | 1.0208 | 1.1299 |
| ADL-intra-EWC | **62.74** | **45.91%** | **1.303** | **0.9756** | **1.0196** |
| ADL-intra-IC | 65.01 | 46.30% | 1.327 | 1.0035 | 1.0651 |
| Forecast horizon is one to eight weeks ahead, for the non-promoted period | | | | | |
| Model/measure | MAE | sMAPE | MASE | AvgRelMAE | scaled MSE |
| Base-lift | 8.84 | 41.10% | 0.609 | 1.0083 | 0.0973 |
| ADL-own | 8.52 | 39.83% | 0.605 | 1.0000 | 0.0921 |
| ADL-intra | 8.47 | 39.70% | 0.606 | 0.9986 | 0.0922 |
| ADL-own-EWC | 8.47 | 39.70% | 0.604 | **0.9963** | 0.0920 |
| ADL-own-IC | 8.43 | 39.71% | **0.598** | 0.9995 | **0.0916** |
| ADL-intra-EWC | 8.43 | 39.61% | 0.605 | 0.9964 | 0.0921 |
| ADL-intra-IC | **8.38** | **39.61%** | 0.600 | 0.9976 | 0.0918 |

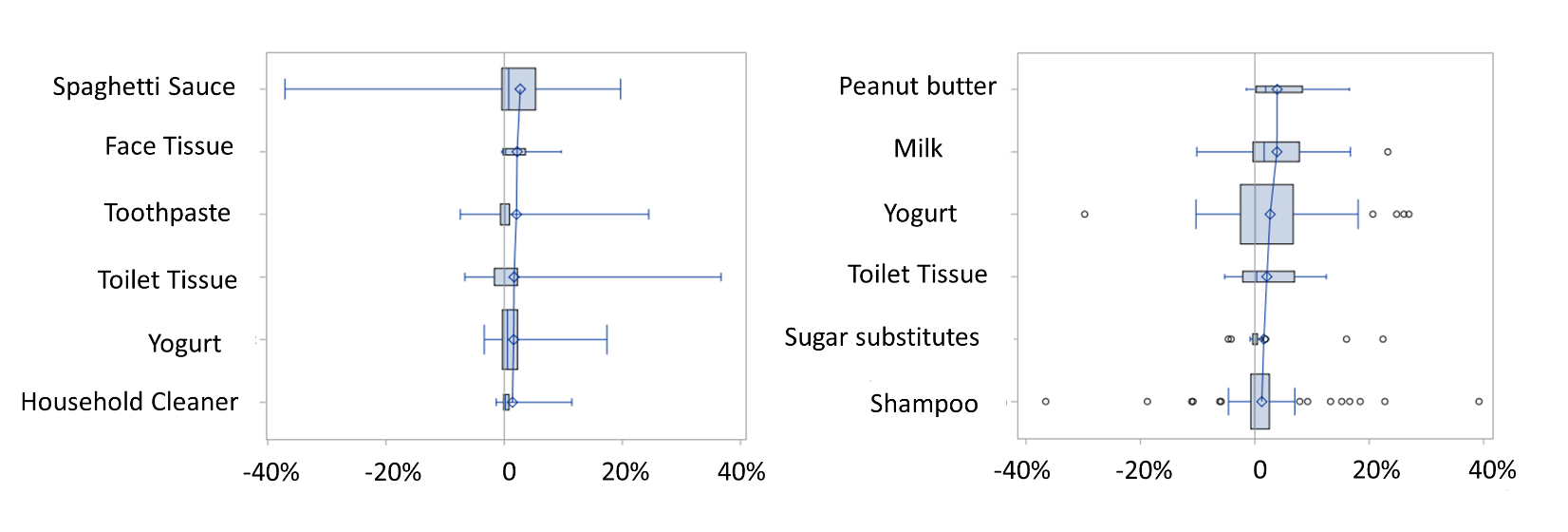
Table 5. The forecasting performance of the models based on previously unseen data for one to eight-week forecast horizon for 1605 SKU’s for the same 28 product categories from a different set of 28 stores

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Forecast horizon is one to eight weeks ahead, for all forecast period | | | | | |
| Model/measure | MAE | sMAPE | MASE | AvgRelMAE | scaled MSE |
| ADL-intra | 13.44 | 40.01% | 0.770 | 1.0000 | 0.1689 |
| ADL-intra-EWC | 13.47 | 39.89% | 0.769 | 0.9964 | 0.169 |
| ADL-intra-IC | **13.34** | 39.60% | 0.76239 | 0.9885 | **0.1674** |
| ADL-EWC-IC | 13.39 | **39.59%** | **0.76236** | **0.9876** | 0.1677 |
| Forecast horizon is one to eight weeks ahead, for the promoted forecast period | | | | | |
| Model/measure | MAE | sMAPE | MASE | AvgRelMAE | scaled MSE |
| ADL-intra | **55.11** | 45.96% | 1.5694 | 1.0000 | 1.2509 |
| ADL-intra-EWC | 55.55 | **45.90%** | **1.5689** | **0.9960** | 1.2549 |
| ADL-intra-IC | 55.112 | 45.99% | 1.5691 | 1.0090 | **1.2477** |
| ADL-EWC-IC | 55.55 | **45.90%** | **1.5689** | **0.9960** | 1.2549 |
| Forecast horizon is one to eight weeks ahead, for the non-promoted forecast period | | | | | |
| Model/measure | MAE | sMAPE | MASE | AvgRelMAE | scaled MSE |
| ADL-intra | 8.3 | 39.27% | 0.671 | 1.0000 | 0.10473 |
| ADL-intra-EWC | 8.28 | 39.15% | 0.67 | 0.9963 | 0.10472 |
| ADL-intra-IC | **8.18** | **38.81%** | **0.663** | **0.9871** | **0.1036** |
| ADL-EWC-IC | **8.18** | **38.81%** | **0.663** | **0.9871** | **0.1036** |

Table 6. The percentage reduction of the MASE by the ADL-intra-EWC model and the ADL-intra-IC model compared to the ADL-intra model for one to eight-week forecast horizon for each product category

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Category/MASE | ADL-intra-EWC | ADL-intra-IC | Category/MASE | ADL-intra-EWC | ADL-intra-IC |
| Beer | 0.13% | -0.31% | Mayonnaise | -0.07% | 0.04% |
| Blades | 0.37% | 1.30% | Milk | 0.82% | **4.84%** |
| Carbonated Beverages | -0.33% | -2.13% | Mustard & Ketchup | 0.45% | -0.84% |
| Cigarette | 0.07% | 0.72% | Peanut butter | -0.19% | **4.88%** |
| Coffee | -0.22% | 0.89% | Photo | 1.09% | 1.05% |
| Cold Cereal | 0.44% | -1.92% | Salty snacks | 0.00% | 0.56% |
| Deodorant | -0.04% | 1.26% | Shampoo | 0.19% | **1.52%** |
| Face Tissue | **2.19%** | -0.60% | Soup | 0.97% | -4.04% |
| Frozen Dinner | -0.67% | -1.74% | Spaghetti sauce | **2.72%** | 1.38% |
| Frozen pizza | -0.44% | -2.02% | Sugar substitutes | 0.18% | **1.97%** |
| Hotdog | 0.27% | -3.46% | Toilet Tissue | **1.67%** | **2.62%** |
| Household Cleaner | **1.40%** | 0.89% | Toothbrush | -0.18% | -1.12% |
| Laundry Detergent | 0.71% | -0.34% | Toothpaste | **2.09%** | 0.55% |
| Margarine/Butter | -0.76% | -0.89% | Yogurt | **1.61%** | **3.35%** |

Figure 3. The boxplots for the percentage reduction of the MASE by the ADL-intra-EWC method and the ADL-intra-IC method compared to the ADL-intra model for one to eight weeks forecast horizon for selected product categories (e.g., those with their relative performance highlighted in Table 6)



1. the ADL-intra-EWC method (b) the ADL-intra-IC method

The box widths are proportionate to the number of SKU’s for the category. The square symbols, which are joined by lines for illustration, indicate the group means for the category.

## Exploring the determinants of the improvement in the forecasts

The results in Table 6 show that our proposed methods generate more accurate forecasts especially for some product categories (e.g., Yogurt, Toilet Tissue, and Spaghetti sauce etc.) compared to the models with similar specifications but overlook the problem of structural change (e.g., the ADL-intra model). This is due to the unique characteristics of the data for the products in those categories. Thus, we may further explore the potential determinants of the improved forecasting performance by our proposed methods at SKU level. This may potentially provide exploratory insights into what types of SKUs may most benefit from using the proposed methods. We consider the following data characteristics as potential determinants: 1) the average and standard deviation of both the prices and sales variables; 2) the frequency of the feature and display promotions for each of the focal products; 3) more advanced statistical measures suggested by Fildes (1992). For example, we include the proportion of outliers for the sales of each SKU. The value of the sales for product *i* will be identified as an outlier if or , where is the differenced value of the sales for product *i*. and are the first and third quantiles of . For retailer product sales, these outliers are usually due to promotional activities. We also include the randomness measure by regressing on , where is the sales value for product *i* at week *t* given that the outliers are removed and *T* is the time trend. The fitness of this autoregressive model (e.g., the R square) represents the systematic variation in the sales data which could be captured by simple models. Lastly, we include the linear trend of product sales measured as the absolute value of the correlation between and the time trend. We then construct five orthogonal factors to represent the information originally contained in the nine explanatory variables described above, which mitigates the issue of multicollinearity[[12]](#footnote-12). Table 7 shows the correlation between the original fourteen explanatory variables and the constructed factors. We may interpret factor 1 as “Price level and variation”, factor 2 as “Sales level and variation”, factor 3 as “Randomness and trend”, factor 4 as “Outliers and Feature intensity”, and factor 5 as “Display intensity”.

Table 7. The pattern of the factors (Small values are omitted for simplicity)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Factor1 | Factor2 | Factor3 | Factor4 | Factor5 |
| Standard deviation of price | 0.956 |  |  |  |  |
| Average price | 0.930 |  |  |  |  |
| Average sales |  | 0.940 |  |  |  |
| Standard deviation of sales |  | 0.898 |  |  |  |
| Proportion of outliers |  |  | 0.921 |  |  |
| Frequency of Feature |  |  | 0.849 |  |  |
| Trend |  |  |  | 0.922 |  |
| Randomness |  |  |  | 0.913 |  |
| Frequency of Display |  |  |  |  | 0.961 |

We then explore the relationship between these five factors and the forecasting improvement by the proposed methods using regression models. We consider the dependent variables as the percentage reductions of the MASE by the ADL-intra-EWC model and the ADL-intra-IC model compared to the ADL-intra model for one to eight weeks forecast horizon for each SKU, as demonstrated in equation (12) and (13). Table 8 reports the estimate of the parameters. For example, the estimates of “Randomness and trend” are positive (e.g., 0.21 and 0.58) and statistically significant (e.g., with p-values of 0.01 for both parameters) for the models with the dependent variables of and [[13]](#footnote-13). This suggests that, adopting the ADL-intra-EWC method or the ADL-intra-EWC method leads to higher percentage reductions of the MASE for the SKU’s which are associated with higher randomness and trend (e.g., those which are more difficult to forecast and tend to exhibit a trend in product sales). This is possibly because that the SKU’s of this type are more heavily associated with the structural change problem and forecast bias. The results also show that the ADL-intra-IC model tends to have less of an advantage compared to the ADL-intra model for the SKU’s with higher proportions of outliers and higher intensities for Feature promotions (e.g., the parameter is -1.13 with a p-value smaller than 0.01). This is possibly because of the fact that the ‘intercept correction’ for the bias can be submerged by high sales spikes which are usually ‘outliers’ and caused by the Feature promotional activities. This is consistent with the moderate forecasting performance of the ADL-intra-IC method for the promoted forecast period. We conduct the analysis for other error measures and forecast horizons and we have consistent findings. Overall, we attempt to provide exploratory insights on the situations where our proposed methods may gain most benefits compared to the ADL-intra model.

Table 8 The determinants of the percentage reduction of the MASE by the ADL-intra-EWC method and the ADL-intra-IC method compared to the ADL-intra model for one to eight weeks forecast horizon

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameters/ Estimates and p-values/ Dependent variables |  | |  | |
| Estimate | P-value | Estimate | P-value |
| Price level and variation | -0.11\* | 0.21 | 0.11 | 0.61 |
| Sales level and variation | 0.16 | 0.06 | -0.17 | 0.43 |
| Outliers and Feature intensity | -0.01 | 0.93 | -1.13 | 0.00 |
| Randomness and trend | 0.21 | 0.01 | 0.58 | 0.01 |
| Display intensity | -0.11 | 0.19 | -0.06 | 0.77 |
| Intercept | 0.30 | 0.00 | -0.38 | 0.08 |

\*The estimates are all multiplied by 100.

## Conclusions, limitations and future research

Grocery retailers need to effectively manage their supply chain and, to achieve that, they rely on effective forecasting models and welcome new approaches that will enable them to improve their current inventory management practices. Previous studies focus on incorporating additional information (e.g., Gür Ali et al., 2009; Huang et al., 2014; Ma et al., 2016). However, they assume the effect of marketing activities such as price and promotions (e.g., Feature and Display) to be constant over time. This assumption may not hold because of the impact of external factors such as the change in economic conditions, and the change due to consumers’ taste and the entry of new competitors. The data on these external factors are typically not always available. Conventional models assuming constant effects of the marketing activities may be subject to the problem of structural change and generate biased and less accurate forecasts.

Table 9. The percentage reductions of different error measures compared to the Base-lift method

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | MAE | SMAPE | MASE | AvgRelMAE | Scaled MSE |
| ADL-own-EWC | -31.6% | -13.4% | -10.2% | -13.0% | -29.7% |
| ADL-own-IC | -29.2% | -13.3% | -10.5% | -12.7% | -28.6% |
| ADL-intra-EWC | -33.0% | -14.0% | -10.5% | -13.4% | -30.7% |
| ADL-intra-IC | -32.0% | -13.9% | -10.6% | -13.2% | -29.8% |

In this study, we propose effective methods to forecast retailer product sales by taking into account the problem of structural change. We propose the ADL-intra-EWC method which combines various sets of forecasts by the ADL-intra method with different estimation windows under the condition when structural changes are detected. The method tries to achieve an effective trade-off between the reduced forecast bias and the inflated forecast error variance. We also propose the ADL-intra-IC method which attempts to offset the potential forecast bias. The method adds the estimate of the forecast bias back to the error term at the cost of inflated forecast error variance when structural changes are detected. Our models significantly outperform the industrial practice method. Table 9 shows the percentage reductions of various error measures by the ADL-intra-EWC method and the ADL-intra-IC model compared to the Base-lift method for one to eight-week forecast horizon. Specifically, by using the ADL-intra-EWC method and the ADL-intra-IC method, we can reduce the MASE by 10.5% and 10.6% respectively compared to the current practice of using the Base-lift method. Therefore, our study may provide retailers more effective forecasting methods. We have also evaluated the forecasting performance of the ADL-own-EWC method and the ADL-own-IC method . These methods are particularly valuable to manufacturers when competitive promotional information is not available (e.g., Mohammad M. Ali, Babai, Boylan, & Syntetos, 2017; M. M. Ali & Boylan, 2011). Table 9 also shows the percentage reductions of various error measures by the ADL-own-EWC method and the ADL-own-IC method compared to the Base-lift method for one to eight-week forecast horizon. Specifically, by using the ADL-own-EWC method and the ADL-own-IC method, we can reduce the MASE by 10.2% and 10.5% respectively compared to the current practice of using the Base-lift method. The improvements are consistent across different forecast horizons and such improvements in accuracy are estimated to translate into a similar improvement in profits (Kremer, 2015).

In this study, we evaluate the models’ forecasting performance separately depending on if the focal product is being promoted. We find that the ADL-intra-EWC model has the best performance for the promoted forecast period and the ADL-intra-IC model dominates the non-promoted forecast period. We, therefore, forge an exploratory ADL-EWC-IC model which is a combination of the ADL-intra-EWC model and the ADL-intra-IC model based on whenever the focal product is being promoted. We evaluate the forecasting performance of the ADL-EWC-IC model based on previously unseen data for 1605 SKU’s from a different set of 28 stores, and we find that the ADL-EWC-IC model generates the most accurate forecasts overall. These results on the relative strengths of the two approaches to structural change are new.

We also explore the relationship between the improved forecasting performance of the proposed methods (compared to the methods with similar model specifications but overlook the structural break problem) and the data characteristics of the product SKU. We find that the ADL-intra-EWC model tends to have better forecasting performances compared to the ADL-intra model for the SKUs with higher levels of randomness and trend. This suggests that our methods are especially beneficial for the products which are more difficult to forecast and with a trend in their sales. We also find that the ADL-intra-IC model tends to accrue greater advantages compared to the ADL-intra model for the SKU’s with a lower proportion of outliers and lower Feature promotion frequency. This may be due to the fact that the estimated bias can get submerged in the high sales variations caused by promotions. However, we note that the findings are still exploratory as we may capture the characteristics of the data by include more variables (e.g., those for seasonality). However, we leave this problem for future research.

The methods we propose in this study is new to the area of forecasting retailer product sales at SKU level, but we have also identified areas where we feel further improvements in forecasting performance could be found. For example, we may use alternative methods to capture the seasonality. Nagbe, Cugliari, and Jacques (2018) used the splines smoothing method to model the seasonality for electricity demand. For the EWC method, we equally combine the forecasts generated by the ADL-intra model with ten different estimation windows. We may further explore the model’s forecasting performance with a different number of the estimation windows, and with different forecasting combination schemes (e.g., based on *k*-fold evaluation). For the IC method, we may explore the model’s forecasting performance when using different correction schemes (Clements & Hendry, 1999). For example, one alternative correction scheme 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 adjusted, and so forth. Ma et al. (2016) have proposed models which integrate both the intra- and the inter-category promotional information. Thus, it is possible that the forecasting performance might improve with both the intra- and the inter-category promotional information considering the structural change problem which we have brought to attention in this paper. Also, an alternative to the ADL-intra-EWC method and the ADL-intra-IC method is to directly model the change in the effect of the marketing activities, such as the time-varying parameter model. However, a disadvantage of this method is that we need to make strong assumptions of how the effects of the marketing activities change. For example, Foekens, Leeflang, and Wittink (1999) modeled the effect of marketing activities as a linear function of previous promotional activities. Their models were not developed for forecasting purposes. In summary, the methods we have proposed in this study produce consistently accurate forecasts. They also satisfy the practical requirements of retail forecasting in that they are intuitive, they can be developed and operated automatically and also use readily available data on marketing activities.

**Acknowledgments**

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 authors and not by the IRI company.

**References**

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

Ali, M. M., Babai, M. Z., Boylan, J. E., & Syntetos, A. A. (2017). Supply chain forecasting when information is not shared. *European Journal of Operational Research, 260*(3), 984-994. doi: <https://doi.org/10.1016/j.ejor.2016.11.046>

Ali, M. M., & Boylan, J. E. (2011). Feasibility principles for Downstream Demand Inference in supply chains. *Journal of the Operational Research Society, 62*(3), 474-482.

Allen, P. G., & Fildes, R. (2001). Econometric forecasting. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. 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., & Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica, 62*, 1383-1414.

Andrews, R. L., Currim, I. S., Leeflang, P., & Lim, J. (2008). Estimating the SCAN\*PRO model of store sales: HB, FM or just OLS? *international Journal of research in marketing, 25*(1), 22-33.

Bai, J., & Perron, P. (1998). Estimating and Testing Linear Models with Multiple Structural Changes. *Econometrica, 66*, 47- 78.

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

Blattberg, R. C., Briesch, R., & Fox, E. J. (1995). How promotions work? *Marketing Science, 14*(3), G122-G132.

Bronnenberg, B. J., Kruger, M. W., & Mela, C. F. (2008). The IRI Marketing Data Set. *Marketing Science, 27*(4), pp. 745–748.

Brown, R. L., Durbin, J., & Evans, J. M. (1975). Techniques for Testing the Constancy of Regression Relationships over Time. *Journal of the Royal Statistical Society. Series B (Methodological), 37*(2), 149-192.

Bucklin, R. E., Gupta, S., & Siddarth, S. (1998). Determining Segmentation in Sales Response across Consumer Purchase Behaviors. *Journal of Marketing Research, 35*(2), 189-197. doi: 10.2307/3151847

Castle, J. L., Doornik, J. A., & Hendry, D. F. (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.

Christen, M., Gupta, S., Porter, J. C., Staelin, R., & Wittink, D. R. (1997). Using market level data to understand promotion effects in a nonlinear model. *Journal of Marketing Research, 34*(3), 322-334.

Clark, T. E., & McCracken, M. W. (2007). Forecasting with Small Macroeconomic VARs in the Presence of Instabilities *Finance and Economics Discussion Series*: Federal Reserve Board, Washington, D.C.

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

Clements, M. P., & Hendry, D. F. (1996). Intercept Corrections and Structural Change. *Journal of Applied Econometrics, 11*(5), 475-494.

Clements, M. P., & Hendry, D. F. (1998). *Forecasting Economic Time Series*: Cambridge University Press.

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

Cooper, L., Baron, P., Levy, W., Swisher, M., & Gogos, P. (1999). Promocast": a New Forecasting Method for Promotion Planning. *Marketing Science, 18*(3), 301-316.

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

Corsten, D., & Gruen, T. (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), 605-617.

Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: the case of judgmental adjustments to SKU-level demand forecasts. *International Journal of Forecasting, 29*(3), 510-522.

Dekker, M., van Donselaar, K., & Ouwehand, P. (2004). How to use aggregation and combined forecasting to improve seasonal demand forecasts. *International Journal of Production Economics, 90*(2), 151-167.

Demirag, O. C., Keskinocak, P., & Swann, J. (2011). Customer rebates and retailer incentives in the presence of competition and price discrimination. *European Journal of Operational Research, 215*(1), 268-280. doi: <http://dx.doi.org/10.1016/j.ejor.2011.04.006>

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics, 13*, 253-263.

Epprecht, C., Guegan, D., & Veiga, Á. (2013). Comparing variable selection techniques for linear regression: LASSO and Autometrics: Université Panthéon-Sorbonne (Paris 1), Centre d'Economie de la Sorbonne.

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

Fildes, R. (1992). The evaluation of extrapolative forecasting methods. *International Journal of Forecasting, 8*, 81-98.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (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., Ma, S., & Kolassa, S. (2018). *Retail forecasting: research and practice*. Working paper. Lancaster University Management School. Lancaster University.

Fildes, R., Nikolopoulos, K., Crone, S., & Syntetos, A. (2008). Forecasting and operational research: a review. *Journal of the Operational Research Society, 59*(9), 1150-1172.

Fildes, R., & Stekler, H. (2002). The state of macroeconomic forecasting. *Journal of Macroeconomics, 24*(4), 435-468. doi: Pii S0164-0704(02)00055-1

Foekens, E. W., Leeflang, P., & Wittink, D. R. (1999). Varying parameter models to accommodate dynamic promotion effects. *Journal of Econometrics, 89*(1-2), 249-268.

Gupta, S. (1988). Impact of sales promotions on when, what, and how much to buy. *Journal of Marketing Research, 25*, 322-355.

Gür Ali, Ö., SayIn, S., van Woensel, T., & Fransoo, J. (2009). SKU demand forecasting in the presence of promotions. *Expert Systems with Applications, 36*(10), 12340-12348.

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting, 13*(2), 281-291.

Hendry, D. F. (2018). Deciding between alternative approaches in macroeconomics. *International Journal of Forecasting, 34*(1), 119-135. doi: <https://doi.org/10.1016/j.ijforecast.2017.09.003>

Huang, T., Fildes, R., & Soopramanien, D. (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., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*, 679-688.

Kremer, M. S., Enno & Thomas, Doug. (2015). The Sum and Its Parts: Judgmental Hierarchical Forecasting. *Management Science, 62*(10), 1287.

Kuo, R. J. (2001). Sales forecasting system based on fuzzy neural network with initial weights generated by genetic algorithm. *European Journal of Operational Research, 129*(3), 496-517.

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

Lee, W. Y., Goodwin, P., Fildes, R., Nikolopoulos, K., & Lawrence, M. (2007). Providing support for the use of analogies in demand forecasting tasks. *International Journal of Forecasting, 23*(3), 377-390. doi: <https://doi.org/10.1016/j.ijforecast.2007.02.006>

Loeb, W. (2014). Unrelenting Competition: The Biggest Retail Story of 2015, from <https://www.forbes.com/sites/walterloeb/2014/12/16/unrelenting-competition-the-retail-story-of-2015/#4893092419f1>

Ma, S., & Fildes, R. (2017). A retail store SKU promotions optimization model for category multi-period profit maximization. *European Journal of Operational Research, 260*(2), 680-692.

Ma, S., Fildes, R., & Huang, T. (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., & Neslin, S. A. (2004). The determinants of pre- and postpromotion dips in sales of frequently purchased goods. *Journal of Marketing Research, XLI*, 339-350.

Mahajan, V., Bretschneider, S. I., & Bradford, J. W. (1980). Feedback Approaches to Modeling Structural Shifts in Market Response. *Journal of Marketing, 44*, 71-80.

Martin, R., & Kolassa, S. (2009). *Challenges of Automated Forecasting in Retail.* Paper presented at the International Symposium on Forecasting, Hong Kong.

Meeran, S., Jahanbin, S., Goodwin, P., & Quariguasi Frota Neto, J. (2017). When do changes in consumer preferences make forecasts from choice-based conjoint models unreliable? *European Journal of Operational Research, 258*(2), 512-524. doi: <https://doi.org/10.1016/j.ejor.2016.08.047>

Mulhern, F. J., & Leone, R. P. (1991). Implicit price bundling of retail products: A multiproduct approach to maximizing store profitability. *Journal of Marketing, 55*, 63-76.

Nagbe, K., Cugliari, J., & Jacques, J. (2018). Short-Term Electricity Demand Forecasting Using a Functional State Space Model. *Energies, 11*, 1120. doi: doi:10.3390/en11051120

Nijs, V. R., Dekimpe, M. G., Steenkamps, J.-B. E. M., & Hanssens, D. M. (2001). The Category-Demand Effects of Price Promotions. *Marketing Science, 20*(1), 1-22.

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>.

Ouyang, Y. (2007). The effect of information sharing on supply chain stability and the bullwhip effect. *European Journal of Operational Research, 182*, 1107-1121.

Pesaran, M. H., & Pick, A. (2011). Forecast Combination Across Estimation Windows. *Journal of Business & Economic Statistics, 29*(2), 307-318. doi: 10.1198/jbes.2010.09018

Pesaran, M. H., Schuermann, T., & Smith, V. (2009). Forecasting Economic and Financial Variables with Global VARs. *International Journal of Forecasting, 25*, 642-675.

Pesaran, M. H., & Timmermann, A. (2005). Small sample properties of forecasts from autoregressive models under structural breaks. *Journal of Econometrics, 129*(1-2), 183-217. doi: DOI: 10.1016/j.jeconom.2004.09.007

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

Petropoulos, F., Fildes, R., & Goodwin, P. (2016). Do ‘big losses’ in judgmental adjustments to statistical forecasts affect experts’ behaviour? *European Journal of Operational Research, 249*(3), 842-852. doi: <https://doi.org/10.1016/j.ejor.2015.06.002>

Rapach, D. E., & Strauss, J. K. (2008). Structural Breaks and Garch Models of Exchange Rate Volatility. *Journal of Applied Econometrics, 23*(1), 65-90.

Rudolph W. Struse, III. (1987). Commentary—Approaches to Promotion Evaluation: A Practitioner's Viewpoint. *Marketing Science, 6*(2), 150-151.

Sodhi, M. S., & Tang, C. S. (2011). The incremental bullwhip effect of operational deviations in an arborescent supply chain with requirements planning. *European Journal of Operational Research, 215*(2), 374-382.

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

Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research, 252*(1), 1-26. doi: <https://doi.org/10.1016/j.ejor.2015.11.010>

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review *International Journal of Forecasting, 16*(4), 437-450.

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., Bodapati, A. V., & Cooper, L. G. (2006). Retailer Promotion Planning: Improving Forecasting Accuracy And Interpretability. *Journal of Interactive Marketing, 20*(3-4), 71-81.

van Heerde, H., M. Dinner, I., & Neslin, S. (2015). Creating Customer Engagement Via Mobile Apps: How App Usage Drives Purchase Behavior. *Working paper, 10.2139/ssrn.2669817*.

Van Heerde, H. J., Gupta, S., & Wittink, D. R. (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., Srinivasan, S., & Dekimpe, M. G. (2008). *Decomposing the Demand for a Pioneering Innovation*. Working paper. Department of Marketing. University of Waikato.

Walters, R. G. (1991). Assessing the impact of retail price promotions on product substitution, complementary purchase, and interstore sales displacement. *Journal of Marketing, 55*, 17-28.

Walters, R. G., & Rinne, H. J. (1986). An empirical investigation into the impact of price promotions on retail store performance. *Journal of Retailing, 62*(3), 237-266.

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

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

Wildt, A. R., & Winer, R. S. (1983). Modeling and Estimation in Changing Market Environments. *The Journal of Business, 56*(3), 365-388.

1. Corresponding author at Surrey Business School, University of Surrey, GU2 7XH, UK. Tel: 01483 68 6359, email: [t.huang@surrey.ac.uk](mailto:t.huang@surrey.ac.uk); [r.fildes@lancaster.ac.uk](mailto:r.fildes@lancaster.ac.uk) (r.Fildes); [D.G.Soopramanien@lboro.ac.uk](mailto:D.G.Soopramanien@lboro.ac.uk)

   (d.soopramanein) [↑](#footnote-ref-1)
2. The term of ‘structural change’ is used interchangeably with the term of ‘structural break’ in the literature. In this study, we use the term “structural change” as in the retailer context we expect the effect of the marketing activities to change sometimes gradually rather than but sometimes in a sudden and abrupt way. We thank one of the anonymous reviewers for to pointing this out. [↑](#footnote-ref-2)
3. We demonstrate the impact of the structural change on the forecasting performance using a simulation example where the model has an intercept term. We include this in the supplementary material. [↑](#footnote-ref-3)
4. Huang et al. (2014) used alternative schemes such as the Akaike’s Information Criterion. In this study, we find little difference in the results between these different schemes. [↑](#footnote-ref-4)
5. 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-5)
6. We do not further reduce the ADL-intra models using the LASSO procedure as further simplification using the LASSO procedure will potentially remove important variables. [↑](#footnote-ref-6)
7. The results in our study suggest that for most scenarios (e.g., 99.8%) the ADL-intra models are subject to structural change if we conduct the Chow test for 95% of the observations. For robustness, we have conducted the whole evaluation by implementing the sequential Chow test for less observations (e.g., 70% of weeks). We find little difference in the final results. [↑](#footnote-ref-7)
8. Compared to the Mean Absolute Percentage Error (MAPE) which do not have an upper bound, the sMAPE is more robust to outliers. [↑](#footnote-ref-8)
9. We conduct the DM test based on all the error measures except for the AvgRelMAE which does not fit into the framework of the DM test. [↑](#footnote-ref-9)
10. The results for other forecasting horizons are similar and are omitted for simplicity. [↑](#footnote-ref-10)
11. Other models including the Base-lift method, the ADL-own model, the ADL-own-EWC model, and the ADL-own-IC model are outperformed by the four models in Table 5 and are not shown here for simplicity. [↑](#footnote-ref-11)
12. We choose to retain five factors based on the Scree plot and 90.2% of the original information have been retained. [↑](#footnote-ref-12)
13. For robustness, we have developed an alternative regression model which also include dummy variables to capture potentially unobserved category effects, and we find the parameter estimate for the five factors to be consistent with those shown in Table 8. [↑](#footnote-ref-13)