**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 sales forecasts at Stock Keeping Unit (SKU) level to effectively manage their inventory. Previous studies have proposed forecasting methods which incorporate the effect of various marketing activities including prices and promotions. However, their methods have overlooked that the effects of the marketing activities on product sales may change over time. Therefore, these methods may potentially be subject to the structural change problem and generate biased and less accurate forecasts. In this study, we propose more effective methods to forecast retailer product sales which take into account the problem of structural change. Our methods outperform conventional forecasting models based on the data from a popular US retailer.

Keywords:

Analytics, Forecasting, OR in marketing, Retailing

1. **Introduction**

Grocery retailers rely on accurate sales forecasts to coordinate their supply chains (Syntetos, Babai et al. 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 missed 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 and 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 (Cooper, Baron et al. 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 and Tang 2011).

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

These studies assume that the impact of marketing activities such as the price and promotions on product sales remains constant over time. However, in practice, the effect of prices and promotions may change because of many external non-controllable factors. For example, customers may become more sensitive to prices and promotions during an economic crunch period (Wildt 1976, Wildt and Winer 1983). Customers may change their tastes due to the change of their familiarity with the product, the change of their lifestyles, and the change of their social status (Meeran, Jahanbin et al. 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 (Meeran, Jahanbin et al. 2017). 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 and Fildes 2001). As a result, the forecasts generated by these models could be biased and potentially of lower forecast accuracy. The structural change problem has been addressed by previous studies (see a summary in Clements and Hendry 1999) but overlooked in the domain of forecasting retailer product sales. In this study, we propose novel methods to forecast retailer product sales by taking into account the problem of structural change. Specifically, we examine the forecasting performance of the Autoregressive Distributed Lag (ADL) models with the Estimation Window Combining (EWC) method and the ADL model with the Intercept Correction (IC) 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 into 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 the area of forecasting retailer product sales. The empirical results based on the data suggest that our methods have superior forecasting performance compared to conventional models which do not account for the problem of structural change. Second, our methods focus on effectively utilizing available promotional information and thus do not incur the costs of collecting additional data (also, in reality, collecting additional data may not even be possible). 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 for which competitive promotional information becomes unavailable. Fourth, our methods are fully automatic (e.g., the specification of the model for each product does not require human to intervene) and easy to implement, which meets the requirement by retailers who nowadays sell tens of thousands of products. Finally, we shed light upon the characteristics of the products for which our proposed methods have even more superior forecasting performance compared to the conventional model which overlook the problem of structural change.

The remainder of the paper is organized as follows: section 2 initially summarizes previous studies which forecast retailer product sales at SKU level and then highlights the findings by previous studies which explains why the effect of marketing activities including price and promotions may change over time. Section 3 explains the structural change problem and the methods which may potentially mitigate the problem. Section 4 explores the data. In section 5, we introduce our proposed three-stage forecasting methods. Section 6 describes the design of the model evaluation. Section 7 summarizes and discusses the evaluation results to provide a convincing demonstration of the methods’ performance. In Section 8, we highlight the characteristics of the product for which our proposed methods garner the greatest improvements compared to conventional models which overlooks 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 SKU level using a two-stage ‘Base-lift’ method (Cooper, Baron et al. 1999, Fildes, Ma et al. 2018). The method 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. The adjustment is usually estimated relying on the experience of brand/category managers or based on the lift effect by the previous promotional event (Fildes, Nikolopoulos et al. 2008, Fildes, Goodwin et al. 2009). A stream of studies have been devoted to helping retail managers effectively tackle their own cognitive biases typically reflecting their understanding of the market conditions (Lee, Goodwin et al. 2007, Petropoulos, Fildes et al. 2016, Fildes, Goodwin et al. 2018). Some other studies also divide the data into promoted and non-promoted periods but estimate the ‘lift’ effect with model-based forecasting approaches. For example, the PromoCast™ system relates the ‘lift’ effect to various driving factors including previous promotions of the focal product, the characteristics of product categories and stores, and manufacturer information etc. (Cooper, Baron et al. 1999, Cooper and Giuffrida 2000, Trusov, Bodapati et al. 2006). Aburto and Weber (2007) used Neural Network models to estimate the ‘lift’ effect for the product sales for a Chilean supermarket though their evaluation is only based on very limited numbers of products. A limitation for all 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.

Some other studies have proposed wholistic methods which directly generate the final forecasts. Kuo (2001) used Fuzzy Neural Network models to forecast product sales of daily milk in convenience stores. However, their models have been evaluated based on a very limited number of products. Gür Ali, SayIn 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 methods incorporate variables constructed based on statistical measures of past information (e.g., the sales, prices, and promotions) of the focal product and have overall superior forecasting performance. Their methods get beaten by the Base-lift method for the time periods when the focal product is not being promoted. One of the limitations for their methods is that they overlook 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 a small number of known competitors of the focal product (e.g., the main competitors, Coca *versus* Pepsi). Their method however is not applicable for retailers where there are hundreds of competitive products. Huang, Fildes et al. (2014) proposed two-stage Autoregressive Distributed Lag (ADL) methods to forecast retailer product sales at SKU level, which is the first to account for the competitive promotional information for the whole product category where there is a large number of competitive products. They initially implemented a variable selection procedure to identify the most important variables for the competitive activities within the product category. Then they specified the ADL models following a general-to-specific modelling strategy based on these selected variables. Their methods has superior forecasting performance for five grocery categories such as *Bottled Juice*, *Soft Drinks*, and *Bath Soap* etc. However, their methods specify the models relying human expertise and thus do not directly meet the need of automatic modelling which is essential by today’s retailers. Ma, Fildes 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, Fildes 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, Ma et al. 2018). However, all the studies described here assume constant effects of the marketing activities.

2.2 The effect of marketing activities may change over time

Previous studies suggest that the effect of marketing activities may change over time. Wildt (1976) and Wildt and Winer (1983) find 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 et al. (1980) found that the effect of prices and promotions changes during different stages of the product lifecycle. Meeran, Jahanbin et al. (2017) find 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. The change in the behaviour of individual customers may eventually lead to substantial change in the aggregate effect by the marketing activities on product sales. Pauwels and Srinivasan (2004) find that the introduction of store-own brands in a product category reduces the price elasticities of premium national brands and increase price elasticities of the second-tier national brands. The effect of the marketing activities can also change depending on how retailers communicate their marketing events. For example, retailers may promote the products through mobile applications and adopt new prominent promotion shelf tags, which can make the promotions more effective (van Heerde, M. Dinner et al. 2015). The effect of the marketing activities can also change because how retailers manage/store their data. For example, retailers may record their marketing activities using aggregate terms such as feature or displays (e.g., Bronnenberg, Kruger et al. 2008). However, even the same type of event may have various forms such as Buy One Get One free (BOGO), store flyers, billboard advertising, and temporary price reduction (TPR), or TPR for shopper card holders only etc. Therefore, the effect of these events could differentiate as the events initially have different forms.

## The problem of structural change

The problem of structural change has been addressed by previous studies in the forecasting literature[[2]](#footnote-2) (e.g., Pesaran and Timmermann 2007, Castle, Doornik et al. 2008, Hendry 2018). Pesaran and Timmermann (2007) demonstrated analytically how a structural change could lead to forecast bias using a simple regression model without an intercept. For example, suppose that for the time period of , the unobserved data generating process (DGP) is:

(1)

where, and are respectively the vectors of the dependent variable at week *t*+1 and independent variable at week *t*. is the vector of the error term at week *t*+1. (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, the DGP has 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 DGP (e.g., ) based on the data before and after the structural change, e.g., ,. Thus, the OLS estimate of the parameter is:

(2)

where is the vectors of the dependent variable for the time period from week *m* to week *T*, and is the vector of the independent variable for the time period from week *m* to week *T*. We assume that there is no structural change after week *T*. e.g., . Thus, the one-step ahead forecast error is:

(3)

where is the vector of the independent variable for the time period from week *m* to . is the vector of error term for the time period from week *m* to *T*. is the error term at week . The conditional mean of equation (3) is:

(4)

Equation (4) is unequal to zero as is unequal to , which indicates that the forecast at week is biased. The forecast bias may subsequently lead to lower forecast frequency (Clements and Hendry 1999). Previous studies also demonstrated the bias for more general cases (e.g., models with an intercept term and endogenous explanatory variables) using Monte Carlo simulation (see Clements and Hendry 1999, Pesaran and Timmermann 2005, Pesaran and Timmermann 2007)[[3]](#footnote-3).

In this study, we implement two methods to mitigate the problem of structural change. The first is the Intercept Correction (IC) method which specifies non-zero values for the model’s errors in the forecast period given that the model is subject to structural change (Clements and Hendry 1994, Clements and Hendry 1999, Clark and McCracken 2007). If the model is subject to structural changes, we can estimate the forecast bias, e.g., by taking the average value of the most recent residuals, e.g., , where is the number of residuals. When , the bias will be estimated to be the residual at the forecast origin, i.e., , (e.g., Chevillon 2016). We then add the estimated bias back to the out-of-sample forecasts. The final forecasts would be less biased and may potentially be more accurate. However, the IC method comes with limitations. For example, 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 and Hendry 1999). Also, in practice, product sales at SKU level often exhibit large variations and unexpected outliers caused by marketing activities, which renders the task of estimating the forecast bias challenging. The bias could be submerged by high variations in the product sales. Under this circumstance, it is possible that the average value of the most recent residuals may mostly represent random variations rather than the bias caused by the structural change. 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 and Timmermann 2005, Pesaran, Schuermann et al. 2009, Pesaran and Pick 2011). The forecasts can be combined based on equal weights, which has been found effective and easy to implement (Clements and Hendry 1998, Fildes and Stekler 2002, Dekker, van Donselaar et al. 2004, Pesaran, Schuermann et al. 2009). For 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 parameter estimated based on the observation window . The value of can be arbitrarily chosen provided that there are enough observations to estimate the model and enough variations in the explanatory variable. We may 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 obtain the final forecast by equally combining all the sets of forecasts:

(5)

Pesaran and Timmermann (2007) show analytically 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 may bear a cost of inflated forecast error variance. This is because that the models with smaller estimation windows tend to ignore some of the data before the structural change (which may potentailly be more informative compared to the data after the structrual change). The EWC method thus tries to generate more accurate forecasts by taking a trade-off between the reduced forecast bias and the potentially 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 effective in some previous studies. For example, the IC method has been applied to forecast wage, unemployment, and CPI inflation etc. (e.g., Clements and Hendry 1996, Clark and McCracken 2007), and the EWC method has superior forecasting performance for exchange rate, inflation, and equity index futures etc. (e.g., Rapach and Strauss 2008, Pesaran, Schuermann et al. 2009, Pesaran and Pick 2011). However, 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

In this study, we use the retail dataset made available by the Information Resources, Inc. (IRI) company. A more comprehensive description of the dataset can be found in Bronnenberg, Kruger et al. (2008). The dataset contains weekly data at SKU level with variables including product unit sales, price, features, and displays, etc. We initially evaluate the forecasting performance of various models 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 SKU’s with positive movements for at least 90% of the time. Table 1 shows basic statistical measures 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 different product 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, feature, or display 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 period when the focal product is being promoted for display and feature respectively.

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 methods to forecast retailer product sales at SKU level by taking into account the problem of structural change. Both 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. In practice, grocery retailers typically sell hundreds of SKU’s in a product category. This leads to hundreds of potential competitive explanatory variables (e.g., competitive price and competitive promotions) for the focal product. Incorporating all the variables into the model can easily overfit the model and render the estimation task infeasible (Martin and Kolassa 2009). Therefore, we 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 sales of the focal product for a 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 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 thus 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 a 10-fold cross validation (Ma, Fildes et al. 2016, Ma and Fildes 2017)[[4]](#footnote-4).

During the second stage, we construct the General Autoregressive Distributive Lag (ADL) model following Huang, Fildes et al. (2014) based on 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 and Lv 2008, Ma, Fildes et al. 2016). Previous studies suggest that product sales are usually mostly influenced by the prices and promotions of the products themselves (Bucklin, Gupta et al. 1998). Thus, we intentionally incorporate the prices and promotions of the focal product in the general ADL model even if 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, four trigonometric variables to capture the seasonal effect, and other dummy variables to capture the calendar effect. 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 . We include the time as a variable to capture any potential trend during the estimation period (Song and Witt 2003). and respectively represent the log price of the focal product and the log price of a competitive product, *m*, at week . and represents the feature and the display dummy variables for the focal product at week . The first two trigonometric variables, e.g., and captures the month of the year effect, and the other two trigonometric variables, e.g., , and captures the week of the month effect[[5]](#footnote-5). is the dummy variable for the calendar event at week . The dummy variable represents the week of the calendar event if , and the week before the event if . takes the values from 1 to 9 representing all the calendar events*[[6]](#footnote-6)*. are the parameters. is the error term and we 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), can have too many explanatory variables and lack parsimony. Therefore, we simplify the model using the LASSO procedure following Ma, Fildes 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 show that models simplified by the LASSO procedure could have good forecasting performance and outperform traditional models specified based on statistical significance (Epprecht, Guegan et al. 2013, Ma, Fildes 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 (Cooper, Baron et al. 1999). To mitigate the limitation of the LASSO procedure that it may potentially miss important variables, we 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 using the LASSO procedure (we refer to the resulted model as the ADL-own model thereafter). We incorporate the explanatory 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 potentially important variables such as the price and promotions of the focal product and their dynamic terms only at a cost of efficiency. The supplementary parallel ADL model, by definition, has fewer explanatory variables compared to the general ADL model and thus 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 would be less likely for these variables to be removed from both the ADL-raw model and the ADL-own model[[7]](#footnote-7).

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 only when the ADL-intra model is subjected to structural changes, and we keep the forecasts generated by the ADL-intra model as the final forecasts otherwise. In this study, we conduct a sequential Chow test for up to 95% of the weeks in the estimation period[[8]](#footnote-8). Suppose we have an estimation period of 160 weeks, we conduct the Chow test for each of the central 152 weeks. For example, we conduct the Chow test each time assuming a structural change occurring at a specific week from week 5 to week 156 and we obtain the p-values. 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[[9]](#footnote-9). Previous studies have proposed alternative tests which focus on estimating multiple structural changes and their locations but are usually associated with stringent assumptions (e.g., Brown, Durbin et al. 1975, Andrews 1993, Andrews and Ploberger 1994, Bai and Perron 1998, Bai and Perron 2003). In our study, we only need to identify the presence of structure changes. Thus, we conduct the sequential Chow test which meets the requirement and also benefits from simple implementation. We refer to these two three-stage methods as the ADL-intra-EWC method and the ADL-intra-IC method respectively. Figure 2 provides a guide for the implementation of the two methods.

## The experimental design

In this study, we consider the Base-lift method as the benchmark model. The method is widely used in practice and has been used as benchmark models in previous studies (e.g., Cooper, Baron et al. 1999, Gür Ali, SayIn et al. 2009, Huang, Fildes et al. 2014, Ma, Fildes 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 SES 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. In this study, we 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 explanatory variables retained in the ADL-own model.
3. The ADL-own-EWC model: the ADL-own model with the EWC method
4. The ADL-own-IC model: the ADL-own model with the IC method
5. The ADL-intra-EWC model: the ADL-intra model with the EWC method
6. The ADL-intra-IC model: the ADL-intra model 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]). Thus, we have ten sets of forecasts in total. For the IC methods, we estimate the forecast bias as the average value of the sixteen most recent residuals and add the value directly 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’ forecasting performance using 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)[[10]](#footnote-10). We also include relative measures including the Mean Absolute Scaled Error (MASE) proposed by Hyndman and Koehler (2006) and the Relative Average Mean Absolute Error (RelAvgMAE) proposed by Davydenko and Fildes (2013). These measures have more desirable properties, e.g., equally penalizing positive and negative errors and being more robust to outliers. Also, the RelAvgMAE is readily interpretable as the percentage improvement (or worsening) of the focal method compared to a benchmark. The MASE and the RelAvgMAE 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., ). The measures the forecasting performance of one model relative to another and the corresponding and are the MAE by these two models based on one to *H* forecast horizon across SKUs for *K* rolling events. In this study, we use the to measure the forecasting performance of each model relative to the ADL-own model. Thus the is the MAE by the candidate model and the is the MAE by the ADL-own model. 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., Cooper, Baron et al. 1999, Ma, Fildes et al. 2016, Ma and Fildes 2017).

## Results and discussion

In Table 2, we summarize the forecasting performance of the models across all the products with respect to different forecast horizons. Table 3 shows the results of the Diebold-Mariano (DM) test for the statistical significance of the difference between the models’ forecasting performance (Diebold and Mariano 1995, Harvey, Leybourne et al. 1997)[[11]](#footnote-11). 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, Fildes 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 | 1.1508 | 0.2234 |
| ADL-own | 15.70 | 40.74% | 0.693 | 1.0000 | 0.1552 |
| ADL-intra | 15.36 | 40.39% | 0.692 | 0.9934 | 0.1530 |
| ADL-own-EWC | 15.61 | 40.61% | 0.691 | 0.9954 | 0.1542 |
| ADL-own-IC | 16.14 | 40.67% | 0.690 | 0.9986 | 0.1570 |
| ADL-intra-EWC | **15.27** | **40.29%** | 0.690 | **0.9893** | **0.1525** |
| ADL-intra-IC | 15.54 | 40.37% | **0.690** | 0.9935 | 0.1545 |
| Forecast horizon is one to four weeks ahead | | | | | |
| Model/measure | MAE | SMAPE | MASE | AvgRelMAE | scaled MSE |
| Base-lift | 22.67 | 46.24% | 0.762 | 1.1413 | 0.2186 |
| ADL-own | 15.62 | 40.39% | 0.687 | 1.0000 | 0.1530 |
| ADL-intra | 15.11 | 40.02% | 0.684 | 0.9908 | 0.1498 |
| ADL-own-EWC | 15.53 | 40.25% | 0.684 | 0.9948 | 0.1519 |
| ADL-own-IC | 15.88 | 40.19% | 0.681 | 0.9941 | 0.1533 |
| ADL-intra-EWC | **15.02** | 39.91% | 0.682 | **0.9865** | **0.1492** |
| ADL-intra-IC | 15.19 | **39.87%** | **0.679** | 0.9877 | 0.1502 |
| Forecast horizon is one week ahead | | | | | |
| Model/measure | MAE | SMAPE | MASE | AvgRelMAE | scaled MSE |
| Base-lift | 24.99 | 45.42% | 0.762 | 1.1294 | 0.2261 |
| ADL-own | 16.67 | 39.86% | 0.687 | 1.0000 | 0.1551 |
| ADL-intra | 15.65 | 39.40% | 0.685 | 0.9892 | 0.1525 |
| ADL-own-EWC | 16.60 | 39.72% | 0.684 | 0.9952 | 0.1540 |
| ADL-own-IC | 16.97 | 39.49% | 0.678 | 0.9895 | 0.1539 |
| ADL-intra-EWC | **15.58** | 39.29% | 0.683 | 0.9849 | 0.1515 |
| ADL-intra-IC | 15.62 | **39.12%** | **0.678** | **0.9810** | **0.1514** |

We also investigate the models’ forecasting performance for the time periods depending on whether the focal product is being promoted. In practice, retailer product sales tend to exhibit high levels of variations when the focal product is being promoted and tend to become comparably stable otherwise (Gür Ali, SayIn 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[[12]](#footnote-12). The following findings 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 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 across these two methods and we refer this model as the ADL-EWC-IC model. The ADL-EWC-IC model is identical to the ADL-intra-EWC model for the promoted period and 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 for 1605 SKU’s for the same 28 product categories but from a different set of 28 stores). Table 5 shows the forecasting performance of the models[[13]](#footnote-13). 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 benefit of taking account for the problem of structural change by focusing on 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 ADL-intra model has a similar specification compared to the ADL-intra-EWC method and the ADL-intra-IC method but overlooks the problem of structural change. The percentage reductions of the MASE by the ADL-intra-EWC method and by the ADL-intra-IC method for product can be demonstrated as follows[[14]](#footnote-14):

(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[[15]](#footnote-15). 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, Fildes et al. 2016). Figure 3(a) and Figure 3(b) show the boxplots for the percentage reduction of the MASE for selective product categories where the ADL-intra-EWC method and the ADL-intra-IC method respectively have the greatest improvement in forecasting performance compared to the ADL-intra model.

Table 3. The results of the Diebold-Mariano (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 | 0.000 | 0.007 | 0.000 | 0.000 | 0.000 | 0.555 | 0.100 | 0.294 | 0.352 | 0.973 | 0.304 |
| ADL-own | ADL-own-EWC | 0.092 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.669 | 0.604 | 0.388 |
| ADL-own | ADL-own-IC | 0.106 | 0.022 | 0.000 | 0.000 | 0.000 | 0.175 | 0.000 | 0.000 | 0.007 | 0.554 | 0.469 | 0.019 |
| ADL-intra | ADL-intra-EWC | 0.165 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.061 | 0.048 | 0.488 | 0.368 | 0.301 |
| ADL-intra | ADL-intra-IC | 0.791 | 0.296 | 0.009 | 0.000 | 0.002 | 0.532 | 0.000 | 0.000 | 0.078 | 0.590 | 0.059 | 0.006 |

\*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 | 1.381 | 2.474 |
| ADL-own | 64.80 | 47.49% | 1.319 | 1.000 | 1.048 |
| ADL-intra | 62.57 | 45.95% | 1.294 | 0.981 | 0.999 |
| ADL-own-EWC | 64.58 | 47.36% | 1.315 | 0.996 | 1.043 |
| ADL-own-IC | 68.95 | 47.94% | 1.344 | 1.022 | 1.104 |
| ADL-intra-EWC | **62.16** | **45.79%** | **1.289** | **0.975** | **0.992** |
| ADL-intra-IC | 64.62 | 46.32% | 1.316 | 1.009 | 1.040 |
| 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 | 1.0120 | 0.0973 |
| ADL-own | 8.53 | 39.76% | 0.602 | 1.0000 | 0.0912 |
| ADL-intra | 8.47 | 39.58% | 0.604 | 0.9977 | 0.0914 |
| ADL-own-EWC | 8.46 | 39.62% | 0.599 | **0.9957** | 0.0905 |
| ADL-own-IC | 8.43 | 39.61% | **0.594** | 0.9984 | **0.0904** |
| ADL-intra-EWC | 8.42 | 39.49% | 0.602 | 0.9950 | 0.0912 |
| ADL-intra-IC | **8.37** | **39.50%** | 0.598 | 0.9961 | 0.0909 |

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

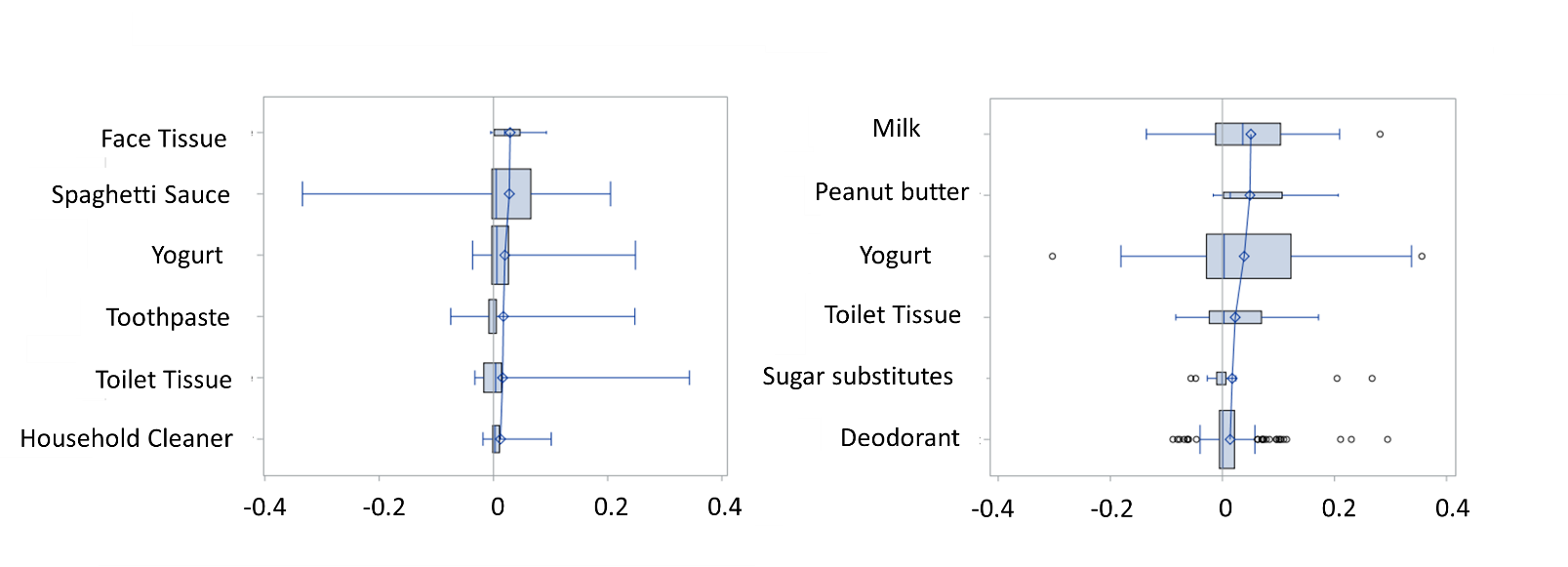
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| All forecast period, for 1 to 8 weeks ahead | | | | | |
| Model/measure | MAE | SMAPE | MASE | AvgRelMAE | scaled MSE |
| ADL-intra | 13.46 | 39.91% | 0.7669 | 0.997 | 0.1674 |
| ADL-intra-EWC | 13.47 | 39.79% | 0.7650 | 0.993 | 0.1674 |
| ADL-intra-IC | **13.39** | 39.50% | 0.7592 | 0.986 | **0.1660** |
| ADL-EWC-IC | 13.41 | **39.49%** | **0.7588** | **0.985** | 0.1661 |
| promoted period, for 1 to 8 weeks ahead | | | | | |
| Model/measure | MAE | SMAPE | MASE | AvgRelMAE | scaled MSE |
| ADL-intra | **55.02** | 45.88% | 1.566 | 0.988 | 1.2459 |
| ADL-intra-EWC | 55.36 | **45.83%** | **1.564** | **0.982** | 1.2482 |
| ADL-intra-IC | 55.23 | 45.93% | 1.567 | 0.993 | **1.2451** |
| ADL-EWC-IC | 55.36 | **45.83%** | **1.564** | **0.982** | 1.2482 |
| non-promoted period, for 1 to 8 weeks ahead | | | | | |
| Model/measure | MAE | SMAPE | MASE | AvgRelMAE | scaled MSE |
| ADL-intra | 7.692 | 38.28% | 0.622 | 0.989 | 0.0904 |
| ADL-intra-EWC | 7.644 | 38.13% | 0.618 | 0.985 | 0.0897 |
| ADL-intra-IC | **7.451** | **37.46%** | **0.605** | **0.967** | **0.0869** |
| ADL-EWC-IC | **7.451** | **37.46%** | **0.605** | **0.967** | **0.0869** |

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 | 0.18% | -0.53% | Mayonnaise | 0.00% | -0.11% |
| Blades | 0.32% | 1.08% | Milk | 1.06% | 5.09% |
| Carbonated Beverages | -0.30% | -2.44% | Mustard & Ketchup | 0.31% | -0.62% |
| Cigarette | 0.11% | 0.80% | Peanut butter | -0.18% | 4.90% |
| Coffee | -0.22% | 0.13% | Photo | 1.00% | -0.98% |
| Cold Cereal | 0.61% | -1.88% | Salty snacks | 0.10% | 1.12% |
| Deodorant | 0.11% | 1.39% | Shampoo | 0.31% | 1.34% |
| Face Tissue | 2.93% | -1.31% | Soup | 0.97% | -4.39% |
| Frozen Dinner | -0.39% | -2.15% | Spaghetti sauce | 2.79% | 0.70% |
| Frozen pizza | -0.46% | -2.16% | Sugar substitutes | 0.09% | 1.75% |
| Hotdog | -0.45% | -4.88% | Toilet Tissue | 1.61% | 2.29% |
| Household Cleaner | 1.24% | 0.66% | Toothbrush | -0.14% | -1.11% |
| Laundry Detergent | 1.14% | -0.17% | Toothpaste | 1.75% | -0.83% |
| Margarine/Butter | -0.84% | -2.70% | Yogurt | 2.01% | 3.89% |

\* positive numbers refer to forecast improvements by our proposed methods with respect to the ADL-intra model.

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.



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. Positive numbers refer to forecast improvements by our proposed methods with respect to the ADL-intra model.

## Exploring the determinants of the improvement in the forecasts

In this section, we try to shed lights on the characteristics of the product for which the ADL-intra-EWC method and the ADL-intra-IC method tend to achieve higher improvements compared to the ADL-intra model which overlooks the problem of structural change. We consider the following data characteristics for each of the SKU’s : 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[[16]](#footnote-16). Table 7 shows the correlation between the original nine explanatory variables and the five constructed factors, and 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 the forecasting improvement by our proposed methods and these five factors 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 respectively compared to the ADL-intra model for one to eight weeks forecast horizon for each SKU, e.g., as defined in equation (12) and equation (13), and we consider the five factors as the independent variables. Table 8 shows the estimation results. We find that “Randomness and trend” has positive and statistically significant effect on the dependent variable. For example, for the models with the dependent variables of , the parameter coefficient for “Randomness and trend” is 0.26 with a p-value of smaller than 0.01. For the models with the dependent variables of, the parameter coefficient is 0.57 with a p-value of 0.01[[17]](#footnote-17). This suggests that, the ADL-intra-EWC method and the ADL-intra-EWC method garner greater improvements compared to the ADL-intra models especially for the SKU’s associated with higher levels of 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 the SKU’s of this type are more heavily associated with the structural change problem and forecast bias. We also find that the ADL-intra-IC method tends to have less improvement compared to the ADL-intra model for the SKU’s with higher proportions of outliers and higher levels of feature intensity (e.g., the parameter is -1.08 with a p-value smaller than 0.01). The possible explanation is that the adjustment for the forecast bias can get submerged by high sales variations which are usually ‘outliers’ and caused by the feature promotional activities. This is also consistent with the moderate forecasting performance of the ADL-intra-IC method for the promoted forecast period. We also conduct the analysis for other error measures and forecast horizons and our findings are consistent.

Table 8 The estimation results for the regression model with five factors as independent variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameters/ Estimates and p-values/ Dependent variables |  | |  | |
| Estimate | P-value | Estimate | P-value |
| Price level and variation | -0.15\* | 0.07 | 0.07 | 0.73 |
| Sales level and variation | 0.19 | 0.02 | -0.19 | 0.36 |
| Outliers and Feature intensity | 0.03 | 0.74 | -1.08 | 0.00 |
| Randomness and trend | 0.26 | 0.00 | 0.57 | 0.01 |
| Display intensity | -0.15 | 0.07 | -0.22 | 0.31 |
| Intercept | 0.35 | 0.00 | -0.43 | 0.05 |

\*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 welcome new approaches that will improve their forecasting accuracy. Previous studies focus on incorporating additional information (e.g., Gür Ali, SayIn et al. 2009, Huang, Fildes et al. 2014, Ma, Fildes 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 etc. The data on these external factors are typically not available. Thus, conventional models assuming constant effects of the marketing activities may be subject to the problem of structural change. As a result, these models may generate biased and potentially 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 | -31.9% | -13.6% | -10.9% | -13.5% | -31.0% |
| ADL-own-IC | -29.6% | -13.4% | -11.0% | -13.2% | -29.7% |
| ADL-intra-EWC | -33.4% | -14.2% | -11.0% | -14.0% | -31.7% |
| ADL-intra-IC | -32.2% | -14.1% | -11.1% | -13.7% | -30.8% |

In this study, we propose novel 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 when structural changes are present. 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 11.0% and 11.1% respectively compared to the Base-lift method. 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., Ali and Boylan 2011, Ali, Babai et al. 2017). 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 10.9% and 11.0% respectively compared to 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 depending on if the focal product is being promoted. We find that the ADL-intra-EWC method has the best performance for the promoted forecast period and the ADL-intra-IC method dominates the non-promoted forecast period. We, therefore, forge an exploratory ADL-EWC-IC model which is a combination of the ADL-intra-EWC method and the ADL-intra-IC method 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.

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 intensity. This may be because that the estimated bias we add back to the forecasts gets submerged in the high sales variations caused by promotions. However, we note that the findings are still exploratory as the improvement by the methods can also be driven by other unknown factors, and we leave the exploration for future research.

The methods we propose in this study is new to 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 the ADL-intra-EWC method, we equally combine the forecasts generated by the ADL-intra model with ten different estimation windows. It is possible to further explore the model’s forecasting performance with different numbers of the estimation windows, and with different forecasting combination schemes (e.g., based on *k*-fold evaluation). For the ADL-intra-IC method, it is possible to explore the model’s forecasting performance when using different correction schemes (Clements and Hendry 1999). One of the alternative correction schemes is to 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 been adjusted, and so forth. In this study, we have brought the problem of structural change to attention. An alternative method to account for this problem is to directly model the change in the effect of the marketing activities, such as using time-varying parameter models. However, a disadvantage of this type of models is that we need to make strong assumptions of how the effects of the marketing activities change. For example, Foekens, Leeflang et al. (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. and R. Weber (2007). "Improved supply chain management based on hybrid demand forecasts." Applied Soft Computing **7**: 136-144.

Ali, M. M., et al. (2017). "Supply chain forecasting when information is not shared." European Journal of Operational Research **260**(3): 984-994.

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

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.

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.

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

Brown, R. L., et al. (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., 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.

Clark, T. E. and M. W. McCracken (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. 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 (1996). "Intercept Corrections and Structural Change." Journal of Applied Econometrics **11**(5): 475-494.

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

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

Cooper, L., 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): 605-617.

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): 510-522.

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.

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

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

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. (1992). "The evaluation of extrapolative forecasting methods." International Journal of Forecasting **8**: 81-98.

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. (2018). "Use and misuse of information in supply chain forecasting of promotion effects. International Journal of Forecasting." International Journal of Forecasting **35**(1): 144-156.

Fildes, R., et al. (2018). Retail forecasting: research and practice, Lancaster University.

Fildes, R., et al. (2008). "Forecasting and operational research: a review." Journal of the Operational Research Society **59**(9): 1150-1172.

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.

Gür Ali, Ö., et al. (2009). "SKU demand forecasting in the presence of promotions." Expert Systems with Applications **36**(10): 12340-12348.

Harvey, D., et al. (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.

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.

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.

Lee, W. Y., et al. (2007). "Providing support for the use of analogies in demand forecasting tasks." International Journal of Forecasting **23**(3): 377-390.

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. and R. Fildes (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., 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.

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.

Nagbe, K., et al. (2018). "Short-Term Electricity Demand Forecasting Using a Functional State Space Model." Energies **11**: 1120.

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.

Pauwels, K. and S. Srinivasan (2004). "Who benefits from store brand entry?" Marketing Science **23**(3): 364-390.

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 (2005). "Small sample properties of forecasts from autoregressive models under structural breaks." Journal of Econometrics **129**(1-2): 183-217.

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

Petropoulos, F., et al. (2016). "Do ‘big losses’ in judgmental adjustments to statistical forecasts affect experts’ behaviour?" European Journal of Operational Research **249**(3): 842-852.

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

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

Syntetos, A. A., et al. (2016). "Supply chain forecasting: Theory, practice, their gap and the future." European Journal of Operational Research **252**(1): 1-26.

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

van Heerde, H., et al. (2015). "Creating Customer Engagement Via Mobile Apps: How App Usage Drives Purchase Behavior." Working paper **10.2139/ssrn.2669817**.

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

Wildt, A. R. and R. S. Winer (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 gradually rather than but in a sudden and abrupt way. We thank one of the anonymous reviewers for 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 rare difference in the results between these different schemes. [↑](#footnote-ref-4)
5. We thank one of the anonymous reviewers for this suggestion to capture the seasonal effect using trigonometric variables. We find that models with trigonometric variable generally have higher forecasting accuracy compared to models which capture the seasonal effect using four-week dummy variables. [↑](#footnote-ref-5)
6. 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-6)
7. 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-7)
8. We keep at least 5% of the weeks for the estimation of the test. [↑](#footnote-ref-8)
9. The results in our study suggest that for most scenarios (e.g., above 99%) 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 the final results consistent. [↑](#footnote-ref-9)
10. The sMAPE is more robust to outliers compared to the Mean Absolute Percentage Error (MAPE) as the latter does not have an upper bound. We have also conducted the analysis for the MAPE and the results are consistent with the results based on the sMAPE. We do not report the results for the MAPE for simplicity. [↑](#footnote-ref-10)
11. 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-11)
12. The results for other forecasting horizons are similar and are omitted for simplicity. [↑](#footnote-ref-12)
13. The results based on the unseen data for the 1605 SKU’s are consistent with the results based on the previous 1831 SKU’s. In Table 5, we do not show the forecasting performance for the Base-lift method, the ADL-own model, the ADL-own-EWC model, and the ADL-own-IC model for simplicity. [↑](#footnote-ref-13)
14. In Equation (12) and (13), all the MASE’s have the same denominator, thus the percentage reductions of the MASE is equivalent to the percentage reductions of the MAE. [↑](#footnote-ref-14)
15. The comparison results for other error measures and horizons are similar and thus omitted for simplicity. [↑](#footnote-ref-15)
16. We choose to retain five factors based on the Scree plot. With five factors we are able to retain 90.2% of information contained in the original nine variables. [↑](#footnote-ref-16)
17. 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-17)