**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 the 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 that the effect of these marketing activities on product sales may change over time. These methods may potentially be subject to a structural change problem as they are unable to capture the time varying effects of marketing activities on sales. Consequently, such approaches generate biased and less accurate forecasts which, importantly, from an operational decision-making perspective, impacts the inventory management of retailers. In this study, we propose new forecasting methods for product sales which considers the problem of structural change. Our methods generate more accurate forecasts compared to the conventional models that do not account for structural changes.

Keywords:

Forecasting, OR in marketing, Analytics, Retailing

1. **Introduction**

Grocery retailers rely on accurate sales forecasts for their inventory management (Petropoulos, Makridakis, Assimakopoulos, & Nikolopoulos, 2014). Inaccurate forecasts of product sales lead to poor service arising from out-of-stock conditions or, alternatively, inflated costs due to overstocking. When a specific item is out-of-stock, retailers directly lose the profit from the sale of the item. Out of stocks situations happen on a regular basis, this 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 the 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).

In practice, many retailers generate sales forecasts at the SKU level using a two-stage ‘base-lift’ approach. They generate the ‘base’ forecasts for the periods when the focal product is not being promoted, using simple univariate models. They add the ‘lift’ effect to the ‘base’ forecasts when the focal product is being promoted. The ‘lift’ is effectively to account for the impact of marketing activities such as promotions, and usually estimated by the brand/category managers based on their experience. In the retailer context, previous studies have proposed various procedures to help managers improve the accuracy of their judgments (e.g., Fildes, Nikolopoulos, Crone, & Syntetos, 2008; Goodwin, 2002; Nikolopoulos, 2010). Some studies have developed models to estimate the ‘lift’ effect based on data (L. Cooper et al., 1999; L. G. Cooper & Giuffrida, 2000; Trusov, Bodapati, & Cooper, 2006). Other studies develop methods that directly generate the final forecasts of product sales. 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. These 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) proposed a three-stage forecasting model which further integrates the promotional information of the products from related product categories.

Generally, all 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 non-controllable factors which may include, for instance, changing economic conditions, changes in consumer tastes and the entry of new competitors etc. Some of these effects are also neither observable or measurable (Wildt, 1976; Wildt & Winer, 1983). For example, customers may become more sensitive to prices and promotions during an economic crisis. Customers can also change their tastes due to factors including their familiarity with the product and their changing lifestyle 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 the new competitor launches their marketing activities 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 their competitors (Loeb, 2014).

Under any of the circumstances that have been described above, forecasting models the parameters of the effects of the price and promotions can potentially be subject to a structural change (Allen & Fildes, 2001; Armstrong, 2001). As a result, the forecasts generated by models that do not account for such changes are likely to be less accurate. The structural change problem has been addressed by previous studies (see a summary in M. P. Clements & Hendry, 1999) but has been overlooked in the domain of forecasting retailer product sales. In this study, we propose new methods to generate more accurate forecasts by taking into account the problem of structural change. Specifically, we propose the Autoregressive Distributed Lag (ADL) models with the Intercept Correction method and the ADL model with the Estimation Window Combining method. The former estimates and offsets the bias and the latter provides combined forecast in order to achieve an effective trade-off between the reduced forecast bias and the inflated forecasting error variance.

Our research falls in the domain of retail forecasting makes the following contributions. First, our research is, as far as we are aware, the first to investigate the problem of structural change for forecasting retailer product sales. Our proposed models have superior forecasting performance compared to conventional models which do not account for the presence 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 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.

The remainder of the paper is organised as follows: section 2 summarizes previous studies in the literature related to forecasting retailer product sales and the change of the effect of marketing activities. Section 3 explains the structural change problem and the two methods we propose to apply. Section 4 explores the data. In section 5, we propose our new three-stage forecasting methods. Section 6 describes the design of the model evaluation. Section 7 summarizes and discusses the evaluation results in order to provide a convincing demonstration of their performance. In Section 8, we explore the characteristics of the situations where the proposed models garner the greatest improvements. In the last section, we make recommendations for both manufacturers and retailers, address research limitations, and highlight directions for future research.

## Literature review

In practice, many retailers forecast their product sales at the SKU level using a two-stage ‘base-lift’ method. The method entails dividing the data into promoted and non-promoted periods based on whether the focal SKU is being promoted. The method is a combination of simple univariate methods (e.g., the simple exponential smoothing method) and judgmental adjustments by brand/category managers (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009; Fildes et al., 2008). These adjustments by managers can be biased and a number of studies have been devoted to helping managers with that task that effectively tackles their own biases typically reflecting their own understanding of the market conditions (Lee, Goodwin, Fildes, Nikolopoulos, & Lawrence, 2007; Petropoulos, Fildes, & Goodwin, 2016). Other studies try to improve these adjustments with model-based forecasting systems. e.g., they estimate the ‘lift’ effect by the promotional event based on information related to previous promotions, store/category features, and manufacturers etc. (L. Cooper et al., 1999; L. G. Cooper & Giuffrida, 2000; Trusov et al., 2006). One limitation of 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. Other studies have proposed more integrated methods to directly generate the final forecasts. Divakar et al. (2005) developed the CHAN4CAST system with models of dynamic regression structures to forecast brand volume sales for the manufacturer/channel. Gür Ali et al. (2009) evaluated the forecasting performance of support vector machine (SVM) models and regression tree models for grocery sales. Huang et al. (2014) proposed two-stage general-to-specific ADL models to forecast grocery sales for five categories such as *Bottled Juice*, *Soft Drinks*, and *Bath Soap* etc. The models incorporate competitive promotional information within the same product category of the focal product. Ma et al. (2016) further integrated the promotional information not only from the same category but also from other related categories.

These studies generate forecasts by capturing various effects of the marketing activities including prices and promotions. For example, previous studies suggest that price reductions and promotions increase the short term sales of the focal product (Blattberg, Briesch, & Fox, 1995). The price reductions and promotions not only increase the product sales at the focal period but also may potentially reduce the sales before and after the focal period as customers may delay or stockpile their purchases (Mace & Neslin, 2004; Van Heerde, Gupta, & Wittink, 2003). The price and promotions also have competitive effects on the sales of other products within and across product categories. (R. L. Andrews, Currim, Leeflang, & Lim, 2008; Wedel & Zhang, 2004).

Evidence also shows that the effect of prices and promotions may change over time. For example, Wildt (1976) and Wildt and Winer (1983) suggest the effect of the marketing activities may change due to the change in economic conditions, consumer tastes, and the competition environment. 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) also 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 the product sales. Previous studies found 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 (Nijs, Dekimpe, Steenkamps, & Hanssens, 2001; Van Heerde, Srinivasan, & Dekimpe, 2008). However, previous studies which forecast retailer product sales at SKU level all assume constant effect of the marketing activities. As a result, their methods may potetnially be subject to the problem of structural change, which we will demonstrate in the next section.

## The problem of structural change

Previous forecasting methods for retailer product sales assume constant parameters and overlook the change in the effect of the marketing activities. As a result, their generated forecasts may potentially be biased and less accurate (Allen & Fildes, 2001; Armstrong, 2001). This has been addressed by previous studies and is referred to as the problem of structural change[[2]](#footnote-2) (e.g., Castle, Doornik, & Hendry, 2008; Hendry, 2018; H. M. Pesaran & Timmermann, 2007). H. M. Pesaran and Timmermann (2005) demonstrated analytically how a structural change leads to biased forecasts using a simple regression model without an intercept. For example, we describe the time periods of and 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 independent variable and dependent variable for the time periods from week *m* to week *T*. We assume no structural change after week *T*. e.g., . Suppose that we are interested in the one-step ahead forecast. Thus, the one-step ahead 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 forecast at week is biased as the expected value of the equation (3) is unequal to zero. e.g.,

. (4)

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 M. P. Clements & Hendry, 1999; H. M. Pesaran & Timmermann, 2005, 2007)[[3]](#footnote-3).

In this study, we implement two methods to address the problem of structural change. The first one is the Intercept Correction (IC) method which specifies non-zero values for the model’s errors in the forecasting period (Clark & McCracken, 2007; M. B. Clements & Hendry, 1994; M. P. Clements & Hendry, 1999). For example, if we identify that the model is subject to structural changes, we can estimate the forecast bias by taking the average value of 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). We can then add the estimated bias back to the out-of-sample forecasts. In the context of our research of retail forecast, sales at SKU level often exhibit large variations, unexpected outliers and missing values which renders the task of estimating the forecast bias difficult. Also, by adding the estimated bias back to the out-of-sample forecasts, we inevitably incur the cost of inflated forecasting error variance (see the analytical evidence in M. P. Clements & Hendry, 1999). The second approach is the Estimation Window Combining (EWC) method which combines the forecasts generated by the same model but with different estimation windows (H. M. Pesaran & Timmermann, 2005; M. H. Pesaran & Pick, 2011; M. H. Pesaran, Schuermann, & Smith, 2009). More specifically, we can combine the forecasts with equal weights as it has been found effective and easy to implement.(M. Clements & Hendry, 1998; Dekker, van Donselaar, & Ouwehand, 2004; Fildes & Stekler, 2002; M. H. Pesaran et al., 2009). In the example, as shown 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 the forecast, e.g., and so forth. We have sets of the forecasts. Thus, the final forecast is computed is effectively an average forecast as follows:

(5)

H. M. Pesaran and Timmermann (2007) show analytically that the forecasts generated by a model with smaller estimation windows tend to be less biased but could also be associated with inflated forecasting error variance. The EWC method does not estimate the size of the bias (compared to the IC method) but take a trade-off between the reduced forecast bias and the inflated forecasting error variance. For the retailer product sales, whether the IC method and the EWC method could generate more accurate forecasts becomes an empirical question. [can we say that which of the two is better ultimately depends on the data or the type of data]

## The data

In this study, we evaluate the forecasting performance of various models using the retail dataset which is made open access by the Information Resources, Inc. (IRI) company. A more comprehensive description of the dataset can be found in Bronnenberg, Kruger, and Mela (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. Table 1 shows the basic statistics of the selected SKU’s during a period of 202 weeks for each product category[[4]](#footnote-4). 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 description of each product category

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Category | Price mean | Sales mean | Display percentage\* | Feature percentage\*\* | Number of SKU's |
| Beer | 8.3 | 20.6 | 13.90% | 4.00% | 169 |
| Blades | 8.1 | 14.6 | 7.40% | 2.20% | 22 |
| Carbonated Beverages | 2.1 | 113.6 | 26.80% | 15.60% | 82 |
| Cigarette | 22.3 | 22.2 | 0.00% | 0.80% | 203 |
| Coffee | 5.2 | 14.5 | 5.20% | 2.90% | 86 |
| Cold cereal | 3.5 | 70.7 | 4.00% | 18.10% | 125 |
| Deodorant | 2.7 | 6.9 | 4.10% | 5.20% | 126 |
| Face Tissue | 2.1 | 75.8 | 0.30% | 11.70% | 6 |
| Frozen Dinner | 2 | 43.8 | 5.30% | 23.70% | 87 |
| Frozen pizza | 3.4 | 31.2 | 8.90% | 9.10% | 147 |
| Household Cleaner | 2.5 | 29.9 | 0.30% | 3.60% | 18 |
| Hotdog | 4 | 68.6 | 13.20% | 15.60% | 35 |
| Laundry Detergent | 8.8 | 28.9 | 2.30% | 8.80% | 57 |
| Margarine/Butter | 2 | 71.4 | 0.10% | 6.30% | 36 |
| Mayonnaise | 3 | 79.7 | 3.00% | 0.40% | 22 |
| Milk | 2.5 | 222.3 | 2.10% | 1.80% | 30 |
| Mustard & Ketchup | 2.1 | 64.5 | 5.30% | 0.90% | 22 |
| Peanut butter | 3.7 | 34.2 | 3.20% | 0.60% | 15 |
| Photo | 7.2 | 9.2 | 4.60% | 5.10% | 13 |
| Salty snacks | 2.3 | 50.9 | 6.70% | 5.00% | 101 |
| Shampoo | 3.5 | 9.9 | 12.80% | 7.10% | 70 |
| Soup | 1.5 | 61.6 | 1.20% | 9.70% | 139 |
| Spaghetti sauce | 2.4 | 39.1 | 1.60% | 6.50% | 52 |
| Sugar substitutes | 2.8 | 14.5 | 0.10% | 1.40% | 20 |
| Toilet Tissue | 5.4 | 89.1 | 4.30% | 8.30% | 20 |
| Toothbrush | 2.6 | 8.7 | 3.10% | 6.30% | 28 |
| Toothpaste | 2.8 | 35.5 | 11.00% | 12.50% | 25 |
| Yogurt | 1.1 | 115.1 | 0.70% | 6.30% | 75 |

\* \*\*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, 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 new forecasting methods which consider the problem of structural change. Our methods consist of three stages. In 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 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 at week *t.* is the matrix for the explanatory variables including product prices, features, and displays of all the products in the same product category. *u* represents the identically distributed error term. represents the vector for 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)[[5]](#footnote-5).

In 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 at the first stage. The LASSO procedure has a limitation that it may potentially misses important variables especially under the condition of high multicollinearity (Fan & Lv, 2008; Ma et al., 2016). Previous studies suggest that sales of a product are usually mostly influenced by the product’s own prices and promotions (Bucklin, Gupta, & Siddarth, 1998). Thus, we intentionally incorporate the prices and promotions of the focal product in the general ADL model even they are not retained by the LASSO procedure. We also incorporate the dynamic effects of these marketing 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. We refer this model as the general ADL model:

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 dummy for the focal product at week . is the four-week-dummy variable. is the dummy variable for the calendar event at week . The dummy variable represents the week of the calendar event when , and the week before the event if . takes the values from 1 to 9 representing all the calendar events*[[6]](#footnote-6)*. are the parameters.  
 is the error term and we assume . 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 specified general ADL model, i.e., model (7), could have many explanatory variables and thus generate poor forecasts due to lack of parsimony. Thus, we simplify the general ADL model using the LASSO procedure following Ma et al. (2016) (we refer to the resulted model as the ADL-raw model thereafter). At 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). In addition, 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). However, we need to mitigate the possibility that the LASSO procedure could miss important variables if we have multicollinearity even at the cost of reduced efficiency. Thus, we construct a supplementary parallel ADL model which has a similar specification compared to the ADL-raw model but only includes the price and promotion variables of the focal product:

(8)

We simplify this model using the LASSO procedure (we refer to this simplified model as the ADL-own model thereafter), and then incorporate the marketing variables retained in the ADL-own model to the ADL-raw model (we refer to the resulted model as the ADL-intra model). In our model developing process at this stage, we initially simplify the general ADL model using the LASSO procedure. We then try to bring the potentially important variables (e.g., the price and promotions of the focal product and their dynamic terms) back to the model in a selective way (e.g., those retained in the ADL-own model will be included in the ADL-intra model). This supplementary parallel ADL model, i.e., model (7), by definition, has fewer explanatory variables compared to the corresponding general ADL model, model (6), and becomes 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 unlikely that they will be removed out of both the ADL-own model and the ADL-raw model[[7]](#footnote-7).

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



In 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 if the existence of any 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, for an estimation period of 160 weeks, we can conduct the Chow test for each of the 152 weeks. We initially conduct the Chow test assuming a structural change occurring at week 5 and we obtain the 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[[8]](#footnote-8). Thus, we may obtain up to 152 p-values in total. The null hypothesis of no structural change will be rejected only if none of these p-value is below a threshold. To avoid the problem of different comparisons using different thresholds, we adopt a very small threshold, i.e., 0.001. Previous studies have proposed alternative tests which focus on estimating a number of 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 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. In Figure 2, the ADL-intra-IC model can be implemented analogously by replacing the EWC method with the IC method once we confirm that the model is subject to structrual change.

## 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* from this approach can be described as follows:

(9)

where represents the initial baseline forecast for week by the simple exponential smoothing model. represents the actual sales of the focal product during the previous week given that the it was not promoted. is the parameter of the simple exponential smoothing model. It is estimated by minimizing the in-sample mean squared errors. The adjustment is calculated as the increased sales of the focal product by its most recent promotion compared to the corresponding initial baseline. In this study, we have the following candidate models:

1. The ADL-own model, i.e., model (6) simplified by the LASSO procedure
2. The ADL-intra model; i.e., model (5) simplified by the LASSO procedure plus the marketing variables retained by 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
6. The ADL-intra-IC model

In this study, we evaluate the forecasting performance of these models using 18 rolling origins for robustness (Tashman, 2000). We specify the model with an estimation window of 160 weeks. 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 is known as it is part of the retailer’s inventory plan. We then use the forecast value of the 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 engage ten estimation windows with different lengths (e.g., for the initial estimation period [1,160], we estimate the model with ten estimation windows including [1, 160], [3, 160], and so forth, until [19, 160]), and generate ten sets of forecasts accordingly). We combine the ten sets of forecasts with equal weights. For the IC methods, we estimate the forecast bias as the average value of the sixteen most recent residuals and add the value equally to the forecasts of all the forecast horizons. We implement the models using MODEL procedure with macros in SAS 9.4. The model parameters are estimated using the OLS estimator.

We evaluate the models with various error measures which approximate the 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). We also include error measures that have recently been developed including the Mean Absolute Scaled Error (MASE) and the Relative Average Mean Absolute Error (RelAvgMAE) developed by Hyndman and Koehler (2006) and Davydenko and Fildes (2013) respectively. The two latter error measures for SKUs based on a forecast horizon of 1 to (e.g., and =1, 4 and 8) are as follows:

, where ,

(10)

Where and are the MASE and the AvgRelMAE based on one to *H* forecast horizon (=1, 4 and 8). and are respectively the *h*-step ahead actual value and forecast value for data series based on the rolling event. There are *S* data series and *K* rolling events (*S*= 1831 and *K*=18). 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 product categories. 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 these analyses:

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 which is scale dependent.
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 MSE for longer forecast horizons (e.g., *h*=4 and 8).
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 1 to 8 weeks ahead, for all forecast period | | | | | | | | | | |
| Model/measure | MAE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank | Scaled MSE | Rank |
| Base-lift | 22.919 | 7 | 46.98% | 7 | 0.775311 | 7 | 1.1444 | 7 | 0.2234 | 7 |
| ADL-own | 15.755 | 5 | 40.81% | 6 | 0.697303 | 6 | 1.0000 | 6 | 0.1575 | 5 |
| ADL-intra | 15.436 | 2 | 40.51% | 3 | 0.695222 | 4 | 0.9941 | 3 | 0.1553 | 2 |
| ADL-own-EWC | 15.673 | 4 | 40.68% | 4 | 0.695964 | 5 | 0.9956 | 4 | 0.1570 | 4 |
| ADL-own-IC | 16.233 | 6 | 40.76% | 5 | 0.694034 | 3 | 0.9992 | 5 | 0.1596 | 6 |
| ADL-intra-EWC | 15.354 | 1 | 40.41% | 1 | 0.693915 | 2 | 0.9905 | 1 | 0.1548 | 1 |
| ADL-intra-IC | 15.595 | 3 | 40.46% | 2 | 0.692854 | 1 | 0.9936 | 2 | 0.1568 | 3 |
| Forecast horizon is 1 to 4 weeks ahead, for all forecast period | | | | | | | | | | |
| Model/measure | MAE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank | Scaled MSE | Rank |
| Base-lift | 22.669 | 7 | 46.24% | 7 | 0.761699 | 7 | 1.1365 | 7 | 0.2186 | 7 |
| ADL-own | 15.630 | 5 | 40.45% | 6 | 0.690272 | 6 | 1.0000 | 6 | 0.1548 | 5 |
| ADL-intra | 15.157 | 2 | 40.12% | 3 | 0.686329 | 4 | 0.9913 | 3 | 0.1514 | 2 |
| ADL-own-EWC | 15.546 | 4 | 40.31% | 5 | 0.688358 | 5 | 0.9950 | 5 | 0.1540 | 4 |
| ADL-own-IC | 15.942 | 6 | 40.25% | 4 | 0.683757 | 2 | 0.9948 | 4 | 0.1553 | 6 |
| ADL-intra-EWC | 15.089 | 1 | 40.01% | 2 | 0.684993 | 3 | 0.9876 | 2 | 0.1509 | 1 |
| ADL-intra-IC | 15.211 | 3 | 39.93% | 1 | 0.681286 | 1 | 0.9871 | 1 | 0.1517 | 3 |
| Forecast horizon is 1 week ahead, for all forecast period | | | | | | | | | | |
| Model/measure | MAE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank | Scaled MSE | Rank |
| Base-lift | 24.990 | 7 | 45.415% | 7 | 0.762 | 7 | 1.1279 | 7 | 0.2261 | 7 |
| ADL-own | 16.662 | 5 | 39.873% | 6 | 0.689 | 6 | 1.0000 | 6 | 0.1561 | 6 |
| ADL-intra | 15.661 | 3 | 39.434% | 3 | 0.686 | 4 | 0.9883 | 3 | 0.1529 | 3 |
| ADL-own-EWC | 16.588 | 4 | 39.720% | 5 | 0.686 | 5 | 0.9955 | 5 | 0.1549 | 4 |
| ADL-own-IC | 17.015 | 6 | 39.519% | 4 | 0.680 | 2 | 0.9902 | 4 | 0.1552 | 5 |
| ADL-intra-EWC | 15.595 | 1 | 39.329% | 2 | 0.684 | 3 | 0.9850 | 2 | 0.1523 | 2 |
| ADL-intra-IC | 15.653 | 2 | 39.148% | 1 | 0.679 | 1 | 0.9804 | 1 | 0.1520 | 1 |

We also investigate the models’ forecasting performance for the time periods depending on whether the focal product is being promoted since the corresponding sales tend to exhibit very different levels of variations. We refer these two periods as the promoted period and non-promoted period respectively. Table 4 shows the forecasting performance of the models for the promoted period and the non-promoted forecast period respectively for one to eight-week forecast horizon[[10]](#footnote-11). The results are similar compared to those in Table 2. From these comparisons, 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 used for the correction gets 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 model between these two models, named as the ADL-EWC-IC model. The ADL-EWC-IC model will be identical to the ADL-intra-EWC model for the promoted period and 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 from the same 28 product categories but from a set of different 28 stores). Table 5 shows the forecasting performance of the ADL-EWC-IC model compared to other three models[[11]](#footnote-12). The 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.

Table 6 shows the percentage reduction of the MASE by the ADL-intra-EWC model and by the ADL-intra-IC model compared to the ADL-intra model for each individual product category for one to eight-week forecast horizon. We focus on the ADL-intra-EWC model and the ADL-inter-IC model because they have the best forecasting performance overall and the ADL-intra model has a similar model specification expect that it overlooks the problem of structural change. The comparison results for other error measures and horizons are similar and we do show them for simplicity. The ADL-intra-EWC model and the ADL-intra-IC model outperforms the ADL-intra model for most of the product categories (e.g., 20 and 17 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 (e.g., Ma et al., 2016). Figure 3 depicts the previous findings using boxplots for their best performing product categories. In these charts, positive values indicate the percentage reduction of the MASE by the ADL-intra-EWC model and by the ADL-intra-IC model compared to the ADL-intra model.

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 |

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

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Forecast horizon is 1 to 8 weeks ahead, for the promoted period | | | | | | | | | | |
| Model/measure | MAE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank | scaled MSE | Rank |
| Base-lift | 119.330 | 7 | 87.26% | 7 | 1.915 | 7 | 1.3705 | 7 | 2.4742 | 7 |
| ADL-own | 65.272 | 5 | 47.56% | 5 | 1.329 | 5 | 1.0000 | 4 | 1.0719 | 5 |
| ADL-intra | 63.100 | 2 | 46.04% | 2 | 1.307 | 2 | 0.9795 | 2 | 1.0265 | 2 |
| ADL-own-EWC | 65.010 | 3 | 47.43% | 4 | 1.325 | 3 | 0.9955 | 3 | 1.0662 | 4 |
| ADL-own-IC | 69.677 | 6 | 47.95% | 6 | 1.354 | 6 | 1.0208 | 6 | 1.1299 | 6 |
| ADL-intra-EWC | 62.737 | 1 | 45.91% | 1 | 1.303 | 1 | 0.9756 | 1 | 1.0196 | 1 |
| ADL-intra-IC | 65.013 | 4 | 46.30% | 3 | 1.327 | 4 | 1.0035 | 5 | 1.0651 | 3 |
| Forecast horizon is 1 to 8 weeks ahead, for the non-promoted period | | | | | | | | | | |
| Model/measure | MAE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank | scaled MSE | Rank |
| Base-lift | 8.837 | 7 | 41.10% | 7 | 0.609 | 7 | 1.0083 | 7 | 0.0973 | 7 |
| ADL-own | 8.523 | 6 | 39.83% | 6 | 0.605 | 5 | 1.0000 | 6 | 0.0921 | 5 |
| ADL-intra | 8.475 | 5 | 39.70% | 4 | 0.606 | 6 | 0.9986 | 4 | 0.0922 | 6 |
| ADL-own-EWC | 8.467 | 4 | 39.70% | 3 | 0.604 | 3 | 0.9963 | 1 | 0.0920 | 3 |
| ADL-own-IC | 8.427 | 2 | 39.71% | 5 | 0.598 | 1 | 0.9995 | 5 | 0.0916 | 1 |
| ADL-intra-EWC | 8.433 | 3 | 39.61% | 2 | 0.605 | 4 | 0.9964 | 2 | 0.0921 | 4 |
| ADL-intra-IC | 8.377 | 1 | 39.61% | 1 | 0.600 | 2 | 0.9976 | 3 | 0.0918 | 2 |

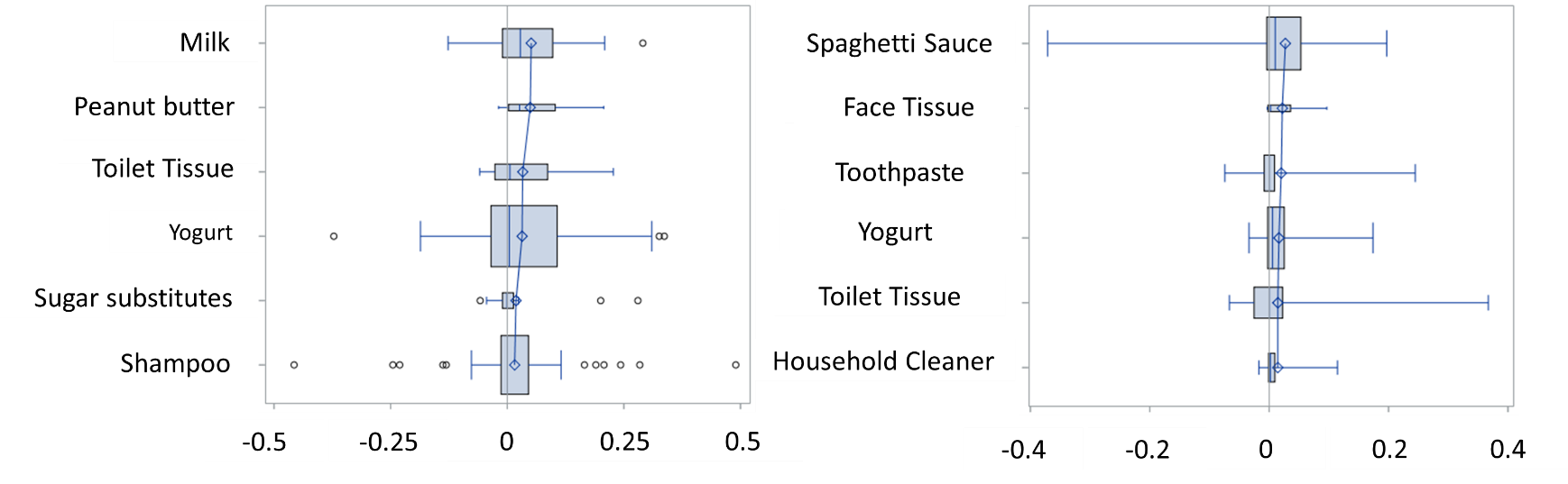
Table 5. The forecast results based on previously unseen data from a different set of 28 stores.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| All forecast period, for 1 to 8 weeks ahead | | | | | | | | | | |
| Model/measure | MAE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank | scaled MSE | Rank |
| ADL-intra | 13.441 | 3 | 40.01% | 4 | 0.770 | 4 | 1.0000 | 4 | 0.1689 | 3 |
| ADL-intra-EWC | 13.473 | 4 | 39.89% | 3 | 0.769 | 3 | 0.9964 | 3 | 0.1690 | 4 |
| ADL-intra-IC | 13.339 | 1 | 39.60% | 2 | 0.762 | 2 | 0.9885 | 2 | 0.1674 | 1 |
| ADL-EWC-IC | 13.387 | 2 | 39.59% | 1 | 0.762 | 1 | 0.9876 | 1 | 0.1677 | 2 |
| promoted period, for 1 to 8 weeks ahead | | | | | | | | | | |
| Model/measure | MAE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank | scaled MSE | Rank |
| ADL-intra | 55.110 | 1 | 45.96% | 3 | 1.569417 | 4 | 1.0000 | 3 | 1.2509 | 2 |
| ADL-intra-EWC | 55.549 | 3 | 45.90% | 1 | 1.568883 | 1 | 0.9960 | 1 | 1.2549 | 3 |
| ADL-intra-IC | 55.112 | 2 | 45.99% | 4 | 1.569142 | 3 | 1.0090 | 4 | 1.2477 | 1 |
| ADL-EWC-IC | 55.549 | 3 | 45.90% | 1 | 1.568883 | 1 | 0.9960 | 1 | 1.2549 | 3 |
| non-promoted period, for 1 to 8 weeks ahead | | | | | | | | | | |
| Model/measure | MAE | Rank | SMAPE | Rank | MASE | Rank | AvgRelMAE | Rank | scaled MSE | Rank |
| ADL-intra | 8.296 | 4 | 39.27% | 4 | 0.67148 | 4 | 1.0000 | 4 | 0.1047 | 4 |
| ADL-intra-EWC | 8.279 | 3 | 39.15% | 3 | 0.670104 | 3 | 0.9963 | 3 | 0.1047 | 3 |
| ADL-intra-IC | 8.182 | 1 | 38.81% | 1 | 0.66279 | 1 | 0.9871 | 1 | 0.1036 | 1 |
| ADL-EWC-IC | 8.182 | 1 | 38.81% | 1 | 0.66279 | 1 | 0.9871 | 1 | 0.1036 | 1 |

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 each product category for one to eight-week forecast horizon

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Category/MASE | ADL-intra-EWC | ADL-intra-IC | Category/MASE | ADL-intra-EWC | ADL-intra-IC |
| Beer | 0.12% | -0.58% | Mayonnaise | 0.07% | 0.58% |
| Blades | 0.20% | 2.19% | Milk | 1.04% | 6.25% |
| Carbonated Beverages | 0.40% | 0.10% | Mustard & Ketchup | 0.64% | -1.04% |
| Cigarette | 0.17% | 1.29% | Peanut butter | -0.15% | 5.11% |
| Coffee | -0.01% | 0.38% | Photo | 1.16% | 0.20% |
| Cold Cereal | 0.11% | -2.29% | Salty snacks | 0.02% | 0.25% |
| Deodorant | -0.01% | 1.74% | Shampoo | 0.38% | 1.56% |
| Face Tissue | 1.80% | -0.47% | Soup | 1.03% | -3.29% |
| Frozen Dinner | -0.67% | -0.70% | Spaghetti sauce | 1.61% | 1.67% |
| Frozen pizza | -1.71% | -1.73% | Sugar substitutes | 0.39% | 3.41% |
| Household Cleaner | 1.25% | 0.72% | Toilet Tissue | 0.04% | 2.45% |
| Hotdog | -0.44% | -4.05% | Toothbrush | -0.02% | -2.12% |
| Laundry Detergent | 0.43% | 0.62% | Toothpaste | 1.66% | -1.80% |
| Margarine/Butter | -0.57% | -0.76% | Yogurt | 1.78% | 4.47% |

Figure 3. forecasting performance comparison: for the MASE, and for one to eight-week forecast horizon.



The box widths are proportionate to the number of SKU’s for each product category. The square symbols, which are joined by lines for illustration, indicates the group means. (positive numbers indicate higher performance of the proposed models)

1. the ADL-intra-EWC model (b) the ADL-intra-IC model,

## Exploring the determinants of the improvement in the forecasts

The results in Table 6 show that our proposed models generate more accurate forecasts for some product categories (e.g., Yogurt, Milk, Toilet Tissue etc.). We further explore the determinants of the improvement of the forecasting performance of our proposed models at SKU level. This provides insights into for what types of SKUs we may get most benefit by using the proposed models. We consider the following data characteristics as potential determinants: 1) basic statistical measures for both the prices and sales variables including the average, standard deviation, skewness, range, kurtosis, and coefficient of variation; 2) the frequency of the feature and display promotions for each SKU; 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 . 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) approximates 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 fourteen explanatory variables described above, which mitigates the issue of multicollinearity[[12]](#footnote-13). Table 6 shows the correlation between the original fourteen explanatory variables and the constructed factors[[13]](#footnote-14). We may interpret factor 1 as “Outliers and general variations”, factor 2 as “Sales level and variation”, factor 3 as “Central tendency of sales”, factor 4 as “Price level and variation”, and factor 5 as “Randomness and growth”.

Thus, we can develop regression models to explore potential determinants of the forecasting improvement by the proposed models. Specifically, we construct dependent variables including the percentage reductions of the error measures by the ADL-intra-EWC model and the ADL-intra-IC model compared to the ADL-intra model, and the percentage reductions of the MASE by the ADL-own-EWC model and the ADL-own-IC model compared to the ADL-intra model. For robustness, we develop regression models with and without dummy variables for each product category.

Table 6. The pattern of the factors

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Factor1 | Factor2 | Factor3 | Factor4 | Factor5 |
| Proportion of outliers | 0.855 |  |  |  |  |
| Coefficient of variation (price) | 0.759 |  |  |  |  |
| Coefficient of variation (sales) | 0.714 |  |  |  |  |
| Frequency of Feature | 0.703 |  |  |  |  |
| Standard deviation of sales |  | 0.964 |  |  |  |
| Range of sales |  | 0.929 |  |  |  |
| Average sales |  | 0.807 |  |  |  |
| Frequency of Display |  | 0.281 |  |  |  |
| Kurtosis of sales |  |  | 0.973 |  |  |
| Skewness of sales |  |  | 0.881 |  |  |
| Standard deviation of price |  |  |  | 0.987 |  |
| Average price |  |  |  | 0.831 |  |
| Randomness |  |  |  |  | 0.992 |
| Trend |  |  |  |  | 0.703 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Table 7 reports the estimated parameters of the regression models for the MASE for the one to eight weeks ahead horizon[[14]](#footnote-15). For example, for the percentage reduction of the MASE by the ADL-intra-EWC model and by the ADL-intra-IC model compared to the ADL-intra model, the estimates for “Randomness and growth” are positive (e.g., 0.38 and 0.63) and statistically significant (e.g., a p-values smaller than 0.001, displayed as “0.000”, and 0.004). This indicates that, using the ADL-intra-EWC model and the ADL-intra-IC model lead to higher percentage reductions of the MASE for the SKU’s with higher randomness and trend (e.g., being difficult to forecast and exhibit a trend in sales), possibly because the SKUs 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 and the ADL-own-IC model tend to have less advantages compared to the ADL-intra model and the ADL-own model respectively for the SKUs with a higher proportion of outliers and higher variations, possibly because that the ‘intercept correction’ for the bias can be submerged by high sales spikes which are usually ‘outliers’ and caused by promotions and it is more challenging to estimate the forecast bias under higher sales variations. Overall, the results here may indicate a possibility of determining the optimal sales forecasting method specifically for an SKU. However, the findings are only exploratory, and we leave it to future research.

Table 7 The determinants of reductions of the MASE for one to eight weeks ahead horizon\*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Horizon = 1 to 8 weeks ahead | ADL-intra-EWC | | ADL-own-EWC | | ADL-intra-IC | | ADL-own-IC | |
| Parameter/estimate and p-values | Estimate | P-value | Estimate | P-value | Estimate | P-value | Estimate | P-value |
| Outliers and promotional variations | 0.07 | 0.434 | 0.11 | 0.303 | -1.09 | 0.000 | -1.45 | 0.000 |
| Sales level and variation | 0.12 | 0.173 | 0.16 | 0.105 | -0.21 | 0.340 | -0.93 | 0.000 |
| Central tendency of sales | -0.06 | 0.460 | -0.07 | 0.511 | -0.68 | 0.002 | -0.84 | 0.001 |
| Price level and variation | -0.12 | 0.149 | -0.17 | 0.092 | 0.07 | 0.742 | -0.09 | 0.721 |
| Randomness and growth | 0.38 | 0.000 | 0.45 | 0.000 | 0.63 | 0.004 | 0.80 | 0.001 |
| Intercept | 0.30 | 0.001 | 0.37 | 0.000 | -0.38 | 0.082 | -0.46 | 0.060 |
| Horizon = 1 to 8 weeks ahead | ADL-intra-EWC | | ADL-own-EWC | | ADL-intra-IC | | ADL-own-IC | |
| Parameter/estimate and p-values | Estimate | P-value | Estimate | P-value | Estimate | P-value | Estimate | P-value |
| Outliers and promotional variations | 0.21 | 0.119 | 0.41 | 0.009 | -0.45 | 0.000 | -0.60 | 0.000 |
| Sales level and variation | 0.12 | 0.172 | 0.20 | 0.055 | -0.12 | 0.595 | -0.85 | 0.001 |
| Central tendency of sales | -0.04 | 0.662 | 0.03 | 0.804 | -0.45 | 0.061 | -0.55 | 0.041 |
| Price level and variation | -0.12 | 0.338 | -0.30 | 0.046 | -0.10 | 0.761 | -0.39 | 0.284 |
| Randomness and growth | 0.32 | 0.000 | 0.38 | 0.000 | 0.48 | 0.039 | 0.56 | 0.033 |
| Intercept | 1.48 | 0.001 | 1.64 | 0.001 | 2.40 | 0.031 | 4.06 | 0.001 |

\*The estimates are all multiplied by 100.

The top half of the Table shows the parameter estimates for the model without category dummy variables. The bottom half of the Table shows the parameter estimates for the model with category dummy variables (the estimate for the dummy variables are omitted for simplicity).

## Conclusions, limitations and future research

Grocery retailers need to effectively manage their inventory 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 all assume that the effect of marketing activities such as price reductions and feature and display promotions are constant over time. This assumption may not hold because of the impact of external factors such as a change in economic conditions, a change in consumer taste and the entry of new retailers. The data on these factors are typically not always available. Or, we do not actually know which of these external factors is causing the structural change. Conventional models that do not account for structural change may be subject to the problem of structural change and generate biased and less accurate forecasts.

Table 8. The percentage reductions for different error measures

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Candidate 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% |

Our research focuses on how to mitigate this problem using data on variables that retailers have control over. That is, data on variables relating to marketing activities which retailers use to influence sales in their stores. We propose models which take into account the potential forecast bias caused by structural changes. The ADL-intra-EWC model generates forecasts which are the combination of various sets of forecasts by the ADL-intra model with different estimation windows under a condition when structural changes are detected. It tries to achieve an effective trade-off between the forecast bias and the forecast error variance. The ADL-intra-IC model attempts to offset the potential forecast bias by adding the estimated forecast bias back to the error term at a cost of inflated forecast error variance when structural changes are detected. In a retailer context, the data at SKU level exhibit very different characteristics across different product categories and usually exhibit high levels of variations. Based on our empirical results, we find that our models outperform the ADL-intra model across all the 28 product categories. Table 8 shows the percentage reductions of various error measures by the ADL-intra-EWC model and the ADL-intra-IC model for one to eight-week forecast horizon[[15]](#footnote-16). Specifically, by using the ADL-intra-EWC model, we can reduce the MASE by 10.6% compared to the current practice of using the Base-lift method.

In this study, we have also evaluated the forecasting performance of the ADL-own-EWC model and the ADL-own-IC model. These methods are particularly valuable to manufacturers under certain circumstances where competitive promotional information may not be available (M. Ali & Boylan, 2011; M. M. Ali, Babai, Boylan, & Syntetos, 2017). In our study, the ADL-own -EWC model and the ADL-own -IC model both outperform the ADL-own model across all the product categories. Table 8 also shows the percentage reductions of various error measures by the ADL-own-EWC model and the ADL-own-IC model for one to eight-week forecast horizon.

In this study, the ADL-intra-EWC model has the best performance for the promoted forecast period while the ADL-intra-IC model dominates the non-promoted forecast period. We, therefore, develop an exploratory model which is effectively a combination of the ADL-intra-EWC model and the ADL-intra-IC model based on if the focal product is being promoted. The resulting ADL-EWC-IC model thus generates the most accurate forecasts across all the candidate models for the original data and even for previously unseen data from another set of 28 stores.

We also explore the relationship between the relative advantage of the proposed models and the data characteristics of the product SKU. We find that the ADL-intra-EWC model and the ADL-own-EWC model forecasting performances are generally better compared to the ADL-intra model and the ADL-own model respectively for the SKU’s with high level of variation and trend. The ADL-intra-IC model and the ADL-own-IC model tend to have better forecasting performances compared to the ADL-intra model and the ADL-own model respectively for the SKU’s with the following characteristics: high variation and trend, with a low proportion of outliers and low level of general variations, and with a low level of sales central tendency.

The approach that we propose in this study is new to the area of product sales forecasting at a retailer level but we have also identified areas where we feel further improvements in forecasting performance could be achieved. For example, there are studies which use splines smoothing method to model seasonality, which were found useful for electricity data (Nagbe, Cugliari, & Jacques, 2018). For the EWC method, we combine five sets of forecasts based on ten different estimation windows using equal weights. The forecasting performance could potentially be improved by changing the number of the estimation windows, by changing the minimum length of the estimation windows and by exploring alternative forecasting combination schemes (e.g., based on k-fold evaluation). In the case of IC method, Clements and Hendry (1999) summarize various correction schemes each of which may have different effects on the trade-off between the bias and the error variance[[16]](#footnote-17). Also, Ma et al. (2016) have proposed models which integrate both the intra- and the inter-category promotional information. So, 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. A method alternative to the ADL-intra-EWC method and the ADL-intra-IC method is to directly model the changing process of the effect of the marketing activities. For example, the time-varying parameter model. However, a disadvantage of this method is that we need to make very strong assumptions of how the effect of the marketing activities change overtime. e.g., Foekens, Leeflang, and Wittink (1999) modelled the effect of the marketing activities as a linear function of previous promotional activities. The model has a sophisticated structure and was not developed for forecasting. Therefore, we leave the exploration of the potential of this type of model for future research. assumed that the effect of the marketing activities is a linear function of previous promotional activities. In summary, the models we proposed in this study produce consistently accurate forecasts. They also suffice 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.

**Acknowledgements**

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.

**Reference:**

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

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>

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.

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

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).

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.

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

Clements, M. B., & 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. (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).

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

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.

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., Nikolopoulos, K., Crone, S., & Syntetos, A. (2008). Forecasting and operational research: A review. *Journal of the Operational Research Society, 59*.

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.

Goodwin, P. (2002). Integrating management judgment and statistical methods to improve short-term forecasts. *Omega, 30*(2), 127-135. doi: Doi: 10.1016/s0305-0483(01)00062-7

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).

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.

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>

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

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

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

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

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, L. V. (2009). Forecasting Economic and Financial Variables with Global VARs. *International Journal of Forecasting, 25*, 642-675.

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>

Petropoulos, F., Makridakis, S., Assimakopoulos, V., & Nikolopoulos, K. (2014). ‘Horses for Courses’ in demand forecasting. *European Journal of Operational Research, 237*(1), 152-163. doi: <http://dx.doi.org/10.1016/j.ejor.2014.02.036>

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*, 374-382.

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

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

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

Trusov, M., 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. 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 paer, University of Waikato, Department of Marketing*.

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).

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 that 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 in a sudden and abrupt way. We thank one of the anonymous reviewers for this suggestion. [↑](#footnote-ref-2)
3. We include in the supplementary material a simulation example with the intercept term to demonstrate the impact of the structural change on the forecasting performance. [↑](#footnote-ref-3)
4. We select the SKUs with positive movements for at least 90% of the time. [↑](#footnote-ref-4)
5. Huang et al. (2014) used alternative schemes such as Akaike’s Information Criterion. In this study, we find little difference in the results between different these two schemes. [↑](#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. However, 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 reconduct the entire evaluation using a sequential Chow test for up to 70% of weeks and we find little difference in the results. [↑](#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 not shown here for simplicity. [↑](#footnote-ref-11)
11. Other models including the Base-lift method, the ADL-own model, the ADL-own-EWC model, and the ADL-own-IC model are all outperformed by the four models in Table 5 and we do not show them for simplicity. [↑](#footnote-ref-12)
12. We choose to retain five factors based on the Scree plot and 77% of the original information have been retained. [↑](#footnote-ref-13)
13. In Table 6, we omit all small values for simplicity. [↑](#footnote-ref-14)
14. The results are consistent for other error measures and forecast horizons. [↑](#footnote-ref-15)
15. The results are similar for other forecast horizons. [↑](#footnote-ref-16)
16. For example, one of the alternative 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. [↑](#footnote-ref-17)