Elsevier Editorial System(tm) for European Journal of Operational Research
Manuscript Draft

Manuscript Number: EJOR-D-18-00185R2

Title: Forecasting retailer product sales in the presence of structural change

Article Type: Innovative Application of OR

Section/Category: (T) Forecasting

Keywords: Forecasting; OR in marketing; Analytics; Retailing

Corresponding Author: Dr. tao huang, Ph.D

Corresponding Author's Institution: University of Surrey

First Author: tao huang, Ph.D

Order of Authors: tao huang, Ph.D; Robert Fildes, PhD; Didier Soopramanien, PhD

Abstract: Grocery retailers need accurate sales forecasts at the 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 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. Based on data from a well-known US retailer, we show that our methods outperform conventional forecasting methods that ignore the possibility of such changes.

Dr. Tao Huang
Surrey Business School, University of Surrey
Guildford, Surrey, GU2 7XH, UK

April 2019

Dear Professor Teunter,

We would like to thank you for the opportunity to resubmit a revised copy of this manuscript.

In this study, we propose more effective forecasting methods for retailer product sales at SKU level by taking into account the structural change problem.

The manuscript has been revised to address all the comments made by the reviewers. For example, we follow the reviewer's suggestion and modify our models to capture the seasonal effect using trigonometric variables. We re-run the entire analysis and we find the results consistent. The modified models generate slightly more accurate forecasts compared to previous models which capture the seasonal effect using deterministic four-week dummy variables (we thank the reviewer for this helpful advice). We simplify the previous section 8 and we also revise the manuscript by reorganizing the paragraphs, clarifying ambiguities, and revising grammar/typing mistakes, etc. We carefully proofread the manuscript and in that respect have a professional proofreading editor help us with the task.

We would like to express our thanks to the reviewers for the helpful comments and suggestions for the revision and also for the positive feedback.

We believe have resulted in an improved revised manuscript, and we very much hope the revised manuscript is accepted for publication in Journal.

Thank you for your consideration of this manuscript.

Sincerely,

Dr. Tao Huang on behalf of the authors (Dr. Tao Huang, Professor Robert Fildes, and Dr. Didier Soopramanien).

A list of responses to the reviewers' comments

We would like to thank the two anonymous reviewers for their helpful comments and valuable suggestions. We have carefully read through the report and revised the manuscript by taking into account all the comments. We believe that the paper has improved substantially with their contributions.

In addition to the modifications based on the reviewers' suggestions, we highlight the following major changes in the revised manuscript:

1. We now modify the models to capture the seasonal effect using trigonometric variables, and we re-run the entire analysis. We find that all the results are consistent. Also, the models now generate more accurate forecasts compared to the previous models which capture the seasonal effect using deterministic four-week dummy variables. We thank the reviewer' helpful suggestions.

2. We now reorganize the literature review section and critically review those previous forecasting studies by highlighting their limitations.

3. We clarify and improve the Tables by adding captions and revising their formats.

4. We make clear that the Base-lift method is not the industrial standard and some retailers are taking advantages of econometric models.

5. We briefly summarize the intuition we found in section 8 in the previous draft and we put it to the last section. We make it clear that the results are exploratory and we leave further analysis to future research.

6. We have carefully proofread the manuscript and have in that respect asked a professional proofreading editor to help us with the task.

Please see our detailed responses to the reviewers' comments as follows:

Reviewer comments:

Reviewer #1: I am not very happy with the authors' response to my criticism of their use of 12 four-week dummies to capture seasonality. First, they note that this parameterization assumes the seasonal effect to be constant *within* each four-week bucket, which is good - but they do not seem to note the other effect I noted: that in this parameterization, *different* four-week buckets are treated as *completely unrelated*. This parameterization assumes that there is no reason why June and July should be any more similar than June and December, which is prima facie ecologically very doubtful. Also, they claim in their response that "the models are estimated with a comparably large sample (e.g., 160 weeks), where the loss of 12 degrees of freedom is not an issue". First, 12 df are indeed a problem for models with just 160 observations. A good rule of thumb is having 20 observations per df, and here we have about 15. But we actually have far fewer than 15 observations per df!

The truly monumental four-line equation (7) alone seems to imply 52 dfs if we consider only a single interacting product (M=N=P=1). The entire system is hopelessly overparameterized, and no, using the Lasso does not mean that an ecologically invalid parameterization (see above) suddenly becomes ecologically valid.

I would definitely suggest that the entire analysis be re-run with a far more parsimonious and ecologically valid seasonal model. However, this would be a lot of work, and I believe that the paper is useful even with the current seasonal model. Nevertheless, I do not want readers to come away with the impression that this approach is defensible. (It isn't. There is no argument in favor, except for "we already did it this way", and that it is easy to interpret - but so is the humour theory in medicine.) Consequently, the least I urge is a *much* stronger statement about the limitations of the seasonal model.

We thank the reviewer for pointing out the limitation of the four-week dummy variables and the potential benefit of the trigonometric variables in capturing the seasonal effect. We have taken this on board and we realize that this is an important suggestion and therefore we have conducted the entire analysis using trigonometric variables to account for the seasonal effect. The results are all consistent with the results which we have obtained in the previous submissions of the paper and do not change the main contributions of the paper. Also, we find that the models with trigonometric variables do have higher forecasting accuracy compared to the models that provided earlier where we capture the seasonal effects using deterministic four-week dummy variables. We thank the reviewer for the suggestion which is indeed helpful.

Section 2, the literature review, should be revisited. Section 2.1 is one long paragraph. Please break it up and structure paragraphs logically. Section 2.2 is entitled "The effect of marketing activities including price and promotions", but all of section 2.1 is already about promotional modelling. Please reorganize. In addition, please critically review the literature cited and consider cutting literature that does not consider *forecasting*.

Following your suggestion, we now reorganize section 2.1 into two paragraphs: the first paragraph introduces the studies which deal with the problem of forecasting retailer product sales for the promoted period and the non-promoted period separately; the second paragraph introduces the studies which do not split the forecast period.

We also revise section 2.2 to highlight evidence from previous studies which suggests that the effects of marketing activities are likely to change over time thus justifying the need for the main contribution of our research. We now shorten the review of the studies which focus on previously well-known effects of the marketing activities.

We also provide a more critical review of previous studies by highlighting their limitations in relation to our research contributions.

Eqs. 12 & 13: since all the MASEs have the same denominator, the formulas can equivalently be written in terms of MAEs, which may be a bit more intuitive.

We thank the reviewer for pointing this out. We now add a note “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”.

Table 6: please indicate in the table caption that positive numbers refer to forecast improvements with respect to the benchmark (same for Figure 3). By which logic are entries bolded?

Revised: we have added notes to indicate that positive numbers refer to forecast improvement.

We now remove the bolded format in Table 6. Instead, we explain the logic of why we select the product categories and show their results in Figure 3(a) and 3(b). e.g., "Figures 3(a) and 3(b) show the boxplots for the percentage reduction in the MASE for selective product categories where the two methods respectively produce the greatest improvement in forecasting performance compared to the ADL-intra model."

Here is another possible explanation for structural changes in response to markting activities: the IRI dataset spans four years. It is quite possible that promotions have changed during this time, even if they are all labeled "feature" and "display". For instance, display racks may have been redesigned, or features moved from weekly flyers to mobile apps. (Condensing the enormous amount of different promotions at a typical retailer into just two categories "feature" and "display" is a bit of a laugh, too, but that is a feature of the IRI dataset.)

We thank the reviewer for highlighting this point. We add this in section 2.2 as one of the reasons marketing activities may change. In particular, we add the following sentence:

"The effect of the marketing activities can also change depending on how retailers communicate their marketing events. For example, retailers may promote products through mobile applications and adopt new prominent promotional shelf tags, which can make promotions more effective (Dinner, Heerde, & Neslin, 2015). The effect of the marketing activities can also change due to an update of their content and format. For example, retailers tend to launch promotional events of a wide range of types such as multi-buy promotions, store flyers, mobile apps, billboard advertising, and temporary price reduction (TPR), or TPR for shopper-card holders only. Retailers may initially promote a product with 'Buy One Get One Free' but then update the content to 'Buy One Get the Second for Half Price" months later. They may change the format of the feature advertising from weekly store flyers to mobile apps and also redesign the racks of their display. These changes in the content and format of marketing activities can be expected to lead to changes in consumer response."

The authors keep referring to the base-lift method as the industry standard. I beg to differ. There are lots of econometric methods (see section 5.3 in Fildes et al., 2018), which should definitely outperform base-lift. In particular since the authors' description of base-lift implies that the lift factor is taken from the last promotion, regardless of whether it had a feature or a display, or whether the price reduction was in any way comparable to the price reduction in the promotion we are forecasting for. The base-lift method, in this particular setup, is little more than a straw man that may improve on the simplest naive forecast, but certainly not by much, and it is certainly not state of the art, and Appendix B of Fildes et al. (2018) indicates that it is not all that common: none of the five cases tabulated use it.

We now revise it by introducing the Base-lift model as "a model which has been used as the benchmark model in previous studies", and we also change the sentence in the last section "Our models significantly outperform the industrial practice." to "Our models significantly outperform the Base-lift model.". We also cite Fildes, et al. (2018) in the last section of the study and highlight that nowadays practitioners tend to take advantages of econometric models (and this is why we compared the forecasting performance of our proposed methods with those conventional econometric models which have similar specifications but overlooks the problem of structural change). We add the following sentence in the last section of the paper:

“In this study, we also compare the forecasting performance of our proposed methods with conventional econometric models which have similar specifications but overlook the structural change problem. The ADL-intra-EWC method and the ADL-intra-IC method outperform the ADL-intra model, and the ADL-own-EWC method and the ADL-own-IC method outperform the ADL-own model. We conduct the comparison to highlight the benefit of taking into account the problem of structural change as some retailers have tried to take advantage of conventional econometric models (Fildes, et al., 2018).”

I am still not convinced by section 8. By the time I had arrived here, I was well and truly confused, and this would have been a good point to stop the paper. Instead, we get a PCA (is this a PCA?) that adds yet another layer of abstraction, and all of it post hoc. I find this section weak and unconvincing. Please consider cutting it.

We now remove this section but we briefly summarize some of the findings in the last section. We make it clear that the exploratory results themselves are worth considering in relation to what we are proposing and worthy of further analysis.

Please proofread the manuscript carefully. There are many mentions of "Mariana" which should be "Mariano". Often, citations indicate that an author has different first names and/or initials in the literature database, which leads EndNote to believe that these refer to different people, and to erroneously disambuguate them. In the list of references, Wildt et al. seems to have been edited by someone named "E. proceedings".

Revised. We have proofread the manuscript and have in that respect asked a professional proofreading editor to help us with the task. We have also corrected the list of references.

Reviewer #2: I am happy to see that the authors managed to overcome many deficiencies of the original draft. Although, originally I thought that the manuscript was to be rejected, to my surprise this version of the paper is much improved in all aspects hopefully through my comments and those of the other reviewer.
The literature review is now on spot, contributions are clearly explained, methodology is vastly easier to read through and it is easy to understand the process that the authors follow. As a result, the whole positioning of the manuscript in the sales forecasting respective literature along with the OR one overall is successful. Therefore, I would now recommend that this manuscript is accepted with minor revisions, as long as the editor agrees.
Those revisions are merely the polishing of the paper in terms of the writing. I would suggest one more check for typos and reduction of 'we...' across the manuscript.

We thank the reviewer for his/her comments and appreciation of our work and have proofread the document for typos whilst also revising the changes suggested by reviewer 1.

***Highlights (for review)**

- ✓ We propose novel forecasting methods for retailer product sales at the Stock Keeping Unit level.
- ✓ Our proposed methods generate accurate forecasts by taking into account the problem of structural change.
- ✓ We evaluate the forecasting performance of the Autoregressive Distributed Lag models with the Estimation Window Combining method and the Intercept Correction method.

Forecasting retailer product sales in the presence of structural change

Tao Huang[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 the 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 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. Based on data from a well-known US retailer, we show that our methods outperform conventional forecasting methods that ignore the possibility of such changes.

Keywords:
Forecasting; OR in marketing; Analytics; Retailing

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

1. Introduction

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

Some recent studies have proposed effective methods to forecast retailer product sales at the SKU level. For example, Gür Ali, SayIn, van Woensel, and Fransoo (2009) proposed the regression tree method with a range of variables constructed from the sales, price, and promotion of the focal product. Huang, Fildes, and Soopramanien (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, and Huang (2016) further developed three-stage forecasting methods which integrate the promotional information of the products across related product categories. The various methods in the literature have been explicitly surveyed by Fildes, Ma, and Kolassa (2018).

These studies assume that the impact of marketing activities such as the price and promotions on product sales remains constant over time. However, in practice, the effect of prices and promotions may change due to many uncontrollable external factors. For example, customers may become more sensitive to prices and promotions during an economic crunch period (Wildt, 1976; Wildt & Winer, 1983). Also, customers may change their tastes due to a change in their familiarity with the product, or with a change in their 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 of the marketing activities launched by the new competitor but also because customers seek variety. In 2014, the German discount retail chain Aldi opened more than 400 stores in the United States, leading to changes in customer grocery purchasing habits which then exerted severe competitive pressure on other retail chains (Loeb, 2014).

Under any of these circumstances described above, these forecasting models assume constant effects of the price and promotions but may potentially be subject to the problem of structural change (Allen & Fildes, 2001). As a result, the forecasts generated by these models might be biased and less accurate. The structural change problem has been addressed by previous studies (see a summary in Clements & Hendry, 1999) but overlooked in the marketing domain of forecasting retailer product sales. In this study, we design novel methods to forecast retailer product sales by taking into account the problem of structural changes. Specifically, we examine the forecasting performance of the Autoregressive Distributed Lag (ADL) models with the Intercept Correction (IC) method and the ADL model with the Estimation Window Combining (EWC) method for retailer product sales. The EWC method is to combine different sets of forecasts generated by the same model but with different estimation windows (Pesaran & Timmermann, 2007). The IC method is to make corrections to the final forecasts of the model based on an estimate of the forecast bias (Clements & Hendry, 1998, 1999).

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 offer operational guidance to not only retailers but also to manufacturers when competitive promotional information becomes unavailable. Finally, our methods are fully automatic (e.g., the specification of the model does not rely on human intervention but algorithms) and are easy to implement, which meets the requirement by retailers who nowadays sell tens of thousands of products.

The remainder of the paper is organized as follows. Section 2 initially summarizes previous studies which forecast retailer product sales at the SKU level: we then discuss those findings which justifiy why the effect of marketing activities, including price and promotions, may change over time. Section 3 describes the structural change problem and the methods which can be applied to mitigate the problem. Section 4 explores the data that we use for empirical analysis. In section 5, we introduce our proposed three-stage forecasting methods. Section 6 describes the experimental design for evaluating the alternative models. Section 7 summarizes and discusses the results to compare the methods' performances. In the last section, we provide recommendations for retailers, address various research limitations, and highlight directions for future research.

2. Literature review

2.1 Forecasting retailer product sales at the SKU level

In practice, some retailers forecast their product sales at the SKU level using a two-stage 'Base-lift' method (Cooper 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 may use simple univariate methods to generate the 'baseline' forecasts for the non-promoted period and then make adjustments for the 'lift' effect of any upcoming promotional events. The adjustment could be estimated by relying on the experience of brand/category managers or based on the lift effect by the previous promotional event (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009; Fildes, Nikolopoulos, Crone, & Syntetos, 2008). A stream of research studies has been devoted to helping retail managers effectively tackle their own cognitive biases when they make the adjustment typically reflecting their understanding of the market conditions (Fildes, Goodwin, & Önkal, 2018; Petropoulos, Fildes, & Goodwin, 2016). 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 (Cooper et al., 1999; Cooper & Giuffrida, 2000; Trusov, Bodapati, & Cooper, 2006). Aburto and Weber (2007) used Neural Network models to estimate the 'lift' effect from sales promotions on the product though their evaluation is only based on a very limited number 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 holistic 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 were evaluated based on a very limited number of products. Gür Ali et al. (2009) proposed the regression tree method and the support vector regression (SVR) method to forecast retailer product sales at the SKU level for the non-perishable food categories. Their methods incorporated variables that were constructed based on statistical measures of past information (e.g., the sales, prices, and promotions) of the focal product and showed overall superior forecasting performance. Their methods did not perform better than the Base-lift method for the time period when the focal product was not being promoted. One of the limitations of their methods was that they overlooked the effect of competitive promotions on the sales of the focal product. Divakar, Ratchford, and Shankar (2005) proposed the CHAN4CAST method to forecast product volume sales for beverage manufacturers. Their method incorporated the promotional information for a small number of known competitors of the focal product (e.g., the main competitors, Coca *versus* Pepsi). Their method, however, is not applicable to retailers where there are hundreds of competitive products. Huang et al. (2014) proposed two-stage Autoregressive Distributed Lag (ADL) methods to forecast retailer product sales at the SKU level, which was the first to account for the competitive promotional information from 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 modeling strategy based on these selected variables. Their methods had superior forecasting performance for five grocery categories such as Bottled Juice, Soft Drinks, and Bath Soap. However, their methods relied on intervention by human experts and thus do not directly meet the requirements for automatic modeling which is considered essential by today's retailers. Ma et al. (2016) proposed three-stage ADL methods which further integrate the promotional information not only from the same product category but also from other related product categories. Their methods were extensions of those in Huang et al. (2014) and also benefited from an automatic model specification procedure. Their methods outperformed the Base-lift benchmark model for 15 food product categories. These studies suggest that promotional information is valuable in forecasting retailer product sales, and this is reflected in new 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 from the marketing activities.

2.2 The changing effect of marketing activities

Previous studies of retail demand have suggested that the effect of marketing activities can change over time. Wildt (1976) and Wildt and Winer (1983) found that the effect of the marketing activities may change due to a change in economic conditions, consumer tastes, and the competition environment. Customers may find price reductions and promotions more attractive during an economic crunch compared to other time periods. Mahajan, Bretschneider, and Bradford (1980) found that the effect of prices and promotions changes during different stages of the product lifecycle. Meeran et al. (2017) find that customers have different tastes and preferences when they accumulate more knowledge about the product, when they seek variety, and when they reach a different social status and then decide to adopt a different lifestyle. Changes in the behavior of individual customers may eventually lead to substantial change in the aggregate effect of the marketing activities on product sales. Pauwels and Srinivasan (2004) found that the introduction of store-own brands in a product category reduces the price elasticities of premium national brands and increases price elasticities of 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 products through mobile applications and adopt new prominent promotional shelf tags, which can make promotions more effective (Dinner, Heerde, & Neslin, 2015). The effect of the marketing activities can also change due to an update of their content and format. For example, retailers tend to launch promotional events of a wide range of types such as multi-buy promotions, store flyers, mobile apps, billboard advertising, and temporary price reduction (TPR), or TPR for shopper-card holders only. Retailers may initially promote a product with ‘Buy One Get One Free’ but then update the content to ‘Buy One Get the Second for Half Price” months later. They may change the format of the feature advertising from weekly store flyers to mobile apps and also redesign the racks of their display. These changes in the content and format of marketing activities can be expected to lead to changes in consumer response.

3. The problem of structural change

The problem of structural change has been addressed by previous studies in the forecasting literature[2] (e.g., Castle, Doornik, & Hendry, 2008; Hendry, 2018; Pesaran & Timmermann, 2007). 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 $[1{:}T]$, the unobserved data generating process (DGP) is:

$$y_{t+1} = 1_{\{t\le T_1\}}\beta_1' x_t + \left(1 - 1_{\{t\le T_1\}}\right)\beta_2' x_t + u_{t+1} \qquad (1)$$

where, y_{t+1} and x_t are respectively the vectors of the dependent variable at week *t*+1 and independent variable at week *t*. u_{t+1} is the vector of the error term at week *t*+1. β_i (where *i*=1,2) are the vectors of the parameter coefficients. $1_{\{t\le T_1\}}$ is an indicator which equals to 1 when $t \le T_1$(where $1 < T_1 < T$) and 0 otherwise. Therefore, the DGP has a structural change where the true parameter of the independent variable changes from β_1 to β_2 after T_1. We can estimate a model with a functional form congruent with the DGP (e.g., $y_{t+1} = \hat{\beta}' x_t + \hat{u}_{t+1}$) based on the data before and after the structural change, e.g., $[m{:}T]$, where $1 \le m < T_1 < T$. Thus, the OLS estimate of the parameter is:

$$\hat{\beta}_T(m) = (x_{m,T}' x_{m,T})^{-1} x_{m,T}' y_{m,T} \qquad (2)$$

where $y_{m,T}$ is the vectors of the dependent variable for the time period from week *m* to week *T*, and $x_{m,T}$ 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., $y_{t+1} = \beta_2' x_t + u_{t+1}$, when $t > T$. Thus, the one-step ahead forecast error is:

[2] The term ‘structural change’ is used interchangeably with the term ‘structural break’ in the literature. In this study, we use the term ‘structural change’ as in the retailer context we expect the effects of the marketing activities to change gradually rather than in a sudden and abrupt way. We thank one of the anonymous reviewers for pointing this out.

$$\hat{e}_{T+1}(m) = \left(\beta_2 - \hat{\beta}_T(m)\right)' x_T + u_{T+1}$$

$$= \left(\beta_2' - y_{m,T}' x_{m,T} (x_{m,T}' x_{m,T})^{-1}\right) x_T + u_{T+1}$$

$$= (\beta_2 - \beta_1)' x_{m,T_1}' x_{m,T_1} (x_{m,T_1}' x_{m,T_1})^{-1} x_T - u_{m,T}' x_{m,T} (x_{m,T_1}' x_{m,T_1})^{-1} x_T + u_{T+1} \quad (3)$$

where x_{m,T_1} is the vector of the independent variable for the time period from week *m* to T_1. $u_{m,T}$ is the vector of error term for the time period from week *m* to *T*. u_{T+1} is the error term at week $T+1$. The conditional mean of equation (3) is:

$$E[\hat{e}_{T+1}(m)|x_T] = \left(\beta_2 - \beta_1\right)' x_{m,T_1}' x_{m,T_1} \left(x_{m,T_1}' x_{m,T_1}\right)^{-1} x_T \qquad (4)$$

Equation (4) is unequal to zero as β_2 is unequal to β_1, which indicates that the forecast at week $T+1$ is biased. The forecast bias may subsequently lead to lower forecast accuracy (Clements & 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 & Hendry, 1999; Pesaran & Timmermann, 2005, 2007)[3].

In this study, we implement two methods to mitigate the problem of structural change. The first method is the Intercept Correction (IC) which specifies non-zero values for the model's errors in the forecast period given that the model is subject to structural change (Clark & McCracken, 2007; Clements & Hendry, 1994, 1999). 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., $\widehat{\mathrm{Bias}}_{IC} = \frac{1}{\lambda}\sum_{i=1}^{\lambda} \hat{e}_{T-\mathrm{i}}$, where λ is the number of residuals. When $\lambda = 1$, the bias will be estimated to be the residual at the forecast origin, i.e., $\hat{e}_{T-1}$, (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 & Hendry, 1999). Also, in practice, product sales at the 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 predominantly represent random variations rather than the bias caused by the structural change.

The second method is the Estimation Window Combining (EWC) which combines the forecasts generated by the same model but with different estimation windows (e.g., Pesaran & Pick, 2011; Pesaran, Schuermann, & Smith, 2009; Pesaran & Timmermann, 2005). The forecasts can be

[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.

combined based on equal weights, which have been found effective and easy to implement (Elliott, Granger, & Timmermann, 2006). For the example as depicted in equation (1), we may estimate the model using the most recent ω observations to generate the first set of forecasts, e.g., $\hat{y}_{T+1,1} = \hat{\beta}_{T-\omega+1,T} x_T$, where $\hat{\beta}_{T-\omega+1,T}$ represents the parameter estimated based on the observation window $[T-\omega+1, T]$. The value of ω 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., $\hat{y}_{T+1,2} = \hat{\beta}_{T-\omega,T}\, x_T$, and so forth, until we generate the $(T-\omega+1)^{th}$ set of forecasts based on the estimation window $[1, T]$. Thus, we may obtain the final forecast by equally combining all the $T-\omega+1$ sets of forecasts:

$$\hat{y}_{T+1}(T,\omega) = (T-\omega+1)^{-1} \sum_{m=1}^{T-\omega+1} \hat{y}_{T+1,m} = (T-\omega+1)^{-1} \sum_{m=1}^{T-\omega+1} \hat{\beta}_{m,T} x_T \tag{5}$$

Pesaran and Timmermann (2007) show 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 the models with smaller estimation windows tend to ignore some of the data before the structural change (which may potentially be more informative compared to the data after the structural change). The EWC method thus tries to generate more accurate forecasts by making 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 previous studies. For example, the EWC method has shown superior forecasting performance for exchange rate, inflation, and equity index futures (e.g., Pesaran & Pick, 2011; Pesaran et al., 2009; Rapach & Strauss, 2008). Meanwhile, the IC method has been applied to forecast the likes of wages, unemployment, and CPI inflation (e.g., Clark & McCracken, 2007; Clements & Hendry, 1996). However, in the case of retailer product sales, whether we should account for structural change and which of the two methods (i.e., the IC method, and/or the EWC method) would generate more accurate forecasts remain empirical questions.

4. The data

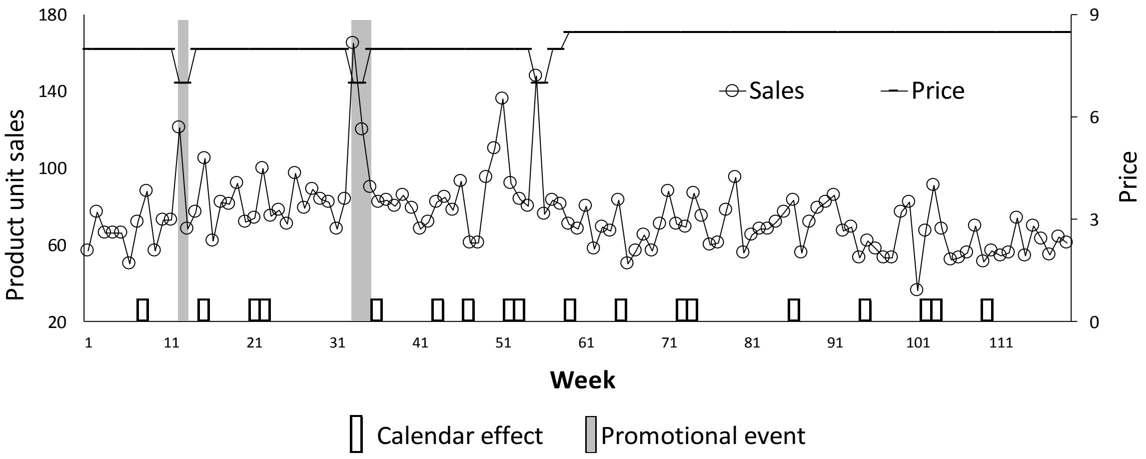
In this study, we use the retail dataset which is publicly available from the Information Resources, Inc. (IRI) company. A more comprehensive description of the dataset can be found in Bronnenberg, Kruger, and Mela (2008). The dataset contains weekly data at the SKU level with variables including product unit sales, price, features, and displays. We initially evaluate the forecasting performance of various models based on 1831 SKUs for 28 product categories from 28 different stores. We select the SKUs for the same category from the same store, and with positive movements for at least 90% of the time. Table 1 shows basic statistical measures for the selected SKUs 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 price reductions, feature, or display of the focal product, as well as calendar events such as Halloween, Thanksgiving, and Christmas.

Table 1. Statistical descriptions for each product category

Category	Price mean	Sales mean*	Display percentage**	Feature percentage***	Number of SKUs
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 across all the SKUs 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.

5. Methodology

We propose two novel methods to forecast retailer product sales at the 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 SKUs in a single 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 & Kolassa, 2009). Therefore, we select the most relevant competitive explanatory variables using the Least Absolute Shrinkage and Selection Operator (LASSO) procedure (Huang et al., 2014; Tibshirani, 1996). That is, we construct the following model for each SKU:

$$\ln\,(y_{0,t}) = X\beta + u, subject\ to\ \sum_{j=1}^{N}\left|\beta_j\right| = \eta\,, \eta \leq \eta_0 \qquad (6)$$

where ln $(y_{0,t})$ represents log sales of the focal product for a store at week *t*. X 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. β represents the vector of the parameter coefficients. *N* is the total number of SKUs for the category. η_0 is the shrinkage factor. The LASSO procedure thus imposes a constraint on 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, 2017; Ma et al., 2016)[4].

During the second stage, we construct the General Autoregressive Distributive Lag (ADL) model following Huang et al. (2014) based on the variables retained by the LASSO procedure during the first stage. The LASSO procedure has a limitation in that it may potentially miss important variables especially under the condition of high multicollinearity (Fan & Lv, 2008; Ma et al., 2016). Previous studies suggest that product sales are usually mostly influenced by the prices and promotions of the products themselves (Bucklin, Gupta, & Siddarth, 1998). Thus, we intentionally incorporate the prices and promotion variables of the focal product into the general ADL model even if these variables were not retained by the LASSO procedure during the first stage. We also incorporate the dynamic effect 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 written as follows:

$$
\begin{aligned}
\ln(y_{0,t}) = {} & intercept + \tau * t + \sum_{j=1}^{L} \alpha_j \ln(y_{0,t-j}) + \sum_{j=0}^{L} \beta_{0,j} \ln(p_{0,t-j}) + \sum_{j=0}^{L} \gamma_{0,j} Feature_{0,t-j} \\
& + \sum_{j=0}^{L} \gamma_{0,j} Display_{0,t-j} + \sum_{m=1}^{M} \sum_{j=0}^{L} \beta_{m,j} \ln(p_{m,t-j}) \\
& + \sum_{n=1}^{N} \sum_{j=0}^{L} \gamma_{n,j} Feature_{n,t-j} \sum_{n=1}^{P} \sum_{j=0}^{L} \gamma_{n,j} Display_{n,t-j} + \theta_1 \sin\left(\frac{2\pi t}{52}\right) + \theta_2 \cos\left(\frac{2\pi t}{52}\right) \\
& + \theta_3 \sin\left(\frac{2\pi t}{4}\right) + \theta_4 \cos\left(\frac{2\pi t}{4}\right) \\
& + \sum_{c=1}^{9} \sum_{v=0}^{1} \delta_{c,v} CalendarEvent_{c,t-v} + \varepsilon_t \qquad (7)
\end{aligned}
$$

where $\ln(y_{0,t})$ is the log sales of the focal product at week t. We include the time t as a variable to capture any potential trend during the estimation period (Song & Witt, 2003). $\ln(p_{0,t-j})$ and $\ln(p_{m,t-j})$ respectively represent the log price of the focal product and the log price of a competitive product, m, at week $t-j$. $Feature_{0,t-j}$ and $Display_{0,t-j}$ represent the feature and the display dummy variables for the focal product at week $t-j$. The trigonometric variables of $\sin\left(\frac{2\pi t}{52}\right)$ and $\cos\left(\frac{2\pi t}{52}\right)$ capture the month of the year effect, and the trigonometric variables of $\sin\left(\frac{2\pi t}{4}\right)$, and $\cos\left(\frac{2\pi t}{4}\right)$ capture the week of the month effect (A. Harvey, 2006)[5]. $CalendarEvent_{c,t-v}$ is the

[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.
[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

dummy variable for the c^{th} calendar event at week $t-v$. The dummy variable represents the week of the calendar event if $v=0$, and the week before the event if $v=1$. c takes the values from 1 to 9 representing all the calendar events[6]. $\alpha_j, \beta_{0,j}, \gamma_{0,j}, \beta_{m,j}, \gamma_{n,j}, \theta_1, \theta_2, \theta_3, \theta_4, \delta_{c,v}, \tau$ are the parameters. ε_t is the error term and we assume that $\varepsilon_t \sim iid(0, \sigma^2)$. L is the order of the lags and is set as 2. M, N, and P are the numbers of selected competitive price, feature, and display variables for the product category.

The general ADL model, as shown in equation (7), contains too many explanatory variables and lacks parsimony. Therefore, we simplify the model using the LASSO procedure following Ma et al. (2016) (we refer to the resulting 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 have shown that models simplified by the LASSO procedure can have good forecasting performance and outperform traditional models based on statistical significance (Epprecht, Guegan, & Veiga, 2013; Ma et al., 2016). Also, the LASSO procedure enables the automation of the statistical forecasting task which becomes essential as typically grocery retailers stock a large number of SKUs (Cooper et al., 1999). To mitigate the limitation of the LASSO procedure in that it may potentially miss important variables, we specify 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:

$$\ln(y_{0,t}) = intercept + \tau * t + \sum_{j=1}^{L} \alpha_j \ln(y_{0,t-j}) + \sum_{j=0}^{L} \beta_{0,j} \ln(p_{0,t-j}) + \sum_{j=0}^{L} \gamma_{0,j} Feature_{0,t-j} + \sum_{j=0}^{L} \gamma_{0,j} Display_{0,t-j} + \theta_1 \sin\left(\frac{2\pi t}{52}\right) + \theta_2 \cos\left(\frac{2\pi t}{52}\right) + \theta_3 \sin\left(\frac{2\pi t}{4}\right) + \theta_4 \cos\left(\frac{2\pi t}{4}\right) + \sum_{c=1}^{9} \sum_{v=0}^{1} \delta_{c,v} CalendarEvent_{c,t-v} + \varepsilon_t \qquad (8)$$

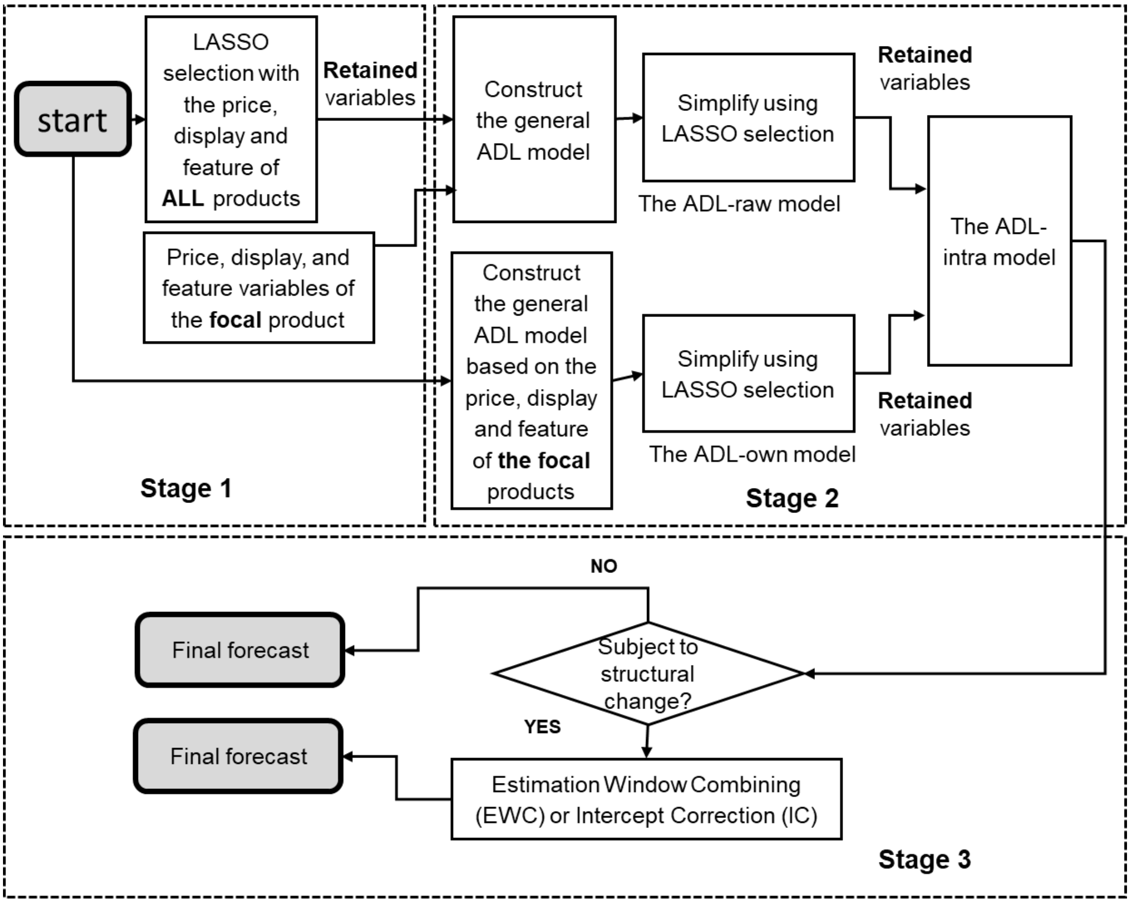
We simplify the supplementary parallel ADL model by using the LASSO procedure (we refer to the resulting model as the ADL-own model thereafter). We then incorporate the explanatory variables retained in the ADL-own model into the ADL-raw model (we refer to the resulting model as the ADL-intra model hereafter). This enables us to selectively retain potentially important variables 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

accuracy compared to models which capture the seasonal effect using four-week dummy variables (e.g., Huang et al., 2014).

[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.

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].

Figure 2. An illustration of 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 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]. For instance, suppose we have an estimation period of 160 weeks. We would then conduct the Chow test for 152 times and each time we assume a structural change has occurred at a specific week from week 5 to week 156 and 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

[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.
[8] We keep at least 5% of the weeks for the estimation of the test.

threshold, i.e., 0.001[9]. Previous studies have proposed alternative tests which focus on estimating multiple structural changes and their locations but they are usually associated with stringent assumptions (e.g., Andrews, 1993; Andrews & Ploberger, 1994; Bai & Perron, 1998, 2003; Brown, Durbin, & Evans, 1975). In our study, we only need to identify the presence of structural change. 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.

6. The experimental design

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

$$Forecast_t = \begin{cases} M_t, & \text{if the focal product is not being promoted} \\ M_t + \text{ adjustment}, & \text{if the focal product is being promoted} \end{cases}$$

$$M_t = (1 - a)M_{t-1} + aS_{t-1} \tag{9}$$

where M_t represents the baseline forecast for week t 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, S_{t-1} represents the sales of the focal product for the previous time period when the focal product was not promoted. a 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.

[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 fewer observations (e.g., 70% of weeks) and we find the final results consistent.

We specify the models with an estimation window of 160 weeks, and 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. The value of the price and any promotional information is considered 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-H weeks ahead forecasts, where H is 1, 4, and 8, to approximate the situation retailers face in practice. For the EWC method, the final forecasts are generated by equally combining the forecasts using the same model with 10 estimation windows (e.g., suppose we have an estimation period of 160 weeks, the estimation windows for the models will be [1, 160], [3, 160], and so forth, until [19, 160]). For the IC methods, we estimate the forecast bias as the average value of the 16 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 perspectives (Kolassa, 2016; Petropoulos & Kourentzes, 2015). 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]. We also include relative measures such as 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:

$$\mathrm{MASE}(H) = \frac{1}{S}\frac{1}{H}\frac{1}{K}\sum_{s=1}^{S}\sum_{h=1}^{H}\sum_{k=1}^{K}\left|\frac{y_{s,h,k}-\hat{y}_{s,h,k}}{\frac{1}{T_0-1}\sum_{t=2}^{T_0}|y_{s,t,k}-y_{s,t-1,k}|}\right| \tag{10}$$

$$\mathrm{AvgRelMAE}(H) = \left(\prod_{s=1}^{S}\mathrm{RelMAE}_{s,H,k}\right)^{\frac{1}{S}}, \text{ where } \mathrm{RelMAE}_{s,H,k} = \frac{\mathrm{MAE}_{s,H,k}^{C}}{\mathrm{MAE}_{s,H,k}^{B}},$$

$$\mathrm{MAE}_{s,H,k}^{C} = \frac{1}{H}\frac{1}{K}\sum_{h=1}^{H}\sum_{k=1}^{K}\left(\left|y_{s,h,k}-\hat{y}_{s,h,k}\right|\right) \tag{11}$$

[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.

where MASE(H) and AvgRelMAE(H) are the MASE and the AvgRelMAE based on one-to-H weeks ahead forecast horizon (H=1, 4 and 8) across S SKUs (e.g., S= 1831) for K rolling events (e.g., K=18). $y_{s,h,k}$ and $\hat{y}_{s,h,k}$ are respectively the h-step ahead actual value and forecast value for data series s based on the k^{th} rolling event. T_0 is the total number of observations in the estimation window (i.e., $T_0 = 160$). The AvgRelMAE measures the forecasting performance of one model relative to another and the corresponding $\mathrm{MAE}^{C}_{s,H,k}$ and $\mathrm{MAE}^{B}_{s,H,k}$ are the MAE by these two models based on one-to-H weeks ahead forecast horizon across S SKUs for K rolling events. In this study, we use the AvgRelMAE to measure the forecasting performance of each model relative to the ADL-own model. Thus the $\mathrm{MAE}^{C}_{s,H,k}$ is the MAE by the candidate model and the $\mathrm{MAE}^{B}_{s,H,k}$ 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 mitigates the bias caused by the logarithm transformation (e.g., Cooper et al., 1999; Ma & Fildes, 2017; Ma et al., 2016).

7. 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 & Mariano, 1995; D. Harvey, Leybourne, & Newbold, 1997)[11]. The following findings emerge from this analysis:

(i) The Base-lift model generates the least accurate forecasts across all the error measures.

(ii) 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).

(iii) The ADL-own-EWC model outperforms the ADL-own model for all the error measures.

(iv) The ADL-own-IC model generally outperforms the ADL-own model except for the MAE.

(v) The ADL-intra-EWC model outperforms the ADL-intra model for all the error measures.

(vi) 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 week ahead and one-to-eight weeks ahead).

(vii) Overall, the ADL-intra-EWC model and the ADL-intra-IC model generate the most accurate forecasts.

[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.

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

Forecast horizon is one-to-eight weeks ahead					
Model/measure	MAE	SMAPE	MASE	AvgRelMAE	scaled MSE
Base-lift	22.92	46.98%	0.7753	1.1508	0.2234
ADL-own	15.70	40.74%	0.6932	1.0000	0.1552
ADL-intra	15.36	40.39%	0.6915	0.9934	0.1530
ADL-own-EWC	15.61	40.61%	0.6907	0.9954	0.1542
ADL-own-IC	16.14	40.67%	0.6899	0.9986	0.1570
ADL-intra-EWC	**15.27**	**40.29%**	0.6900	**0.9893**	**0.1525**
ADL-intra-IC	15.54	40.37%	**0.6896**	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 performances 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 et al., 2009). We refer to these two periods as the promoted period and non-promoted period respectively thereafter. 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 weeks ahead forecast horizon[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 average performances 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 refer to 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

[12] The results for other forecasting horizons are similar and are omitted for simplicity.

unseen data (e.g., the data for 1605 SKUs for the same 28 product categories but from a different set of 28 stores). Table 5 shows the forecasting performance of the models[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 i can be demonstrated as follows[14]:

$$\mathrm{PctRed}(\mathrm{ADL}-\mathrm{intra}-\mathrm{EWC}, i) = \frac{\mathrm{MASE}(\mathrm{ADL}-\mathrm{intra}, i) - \mathrm{MASE}(\mathrm{ADL}-\mathrm{intra}-\mathrm{EWC}, i)}{\mathrm{MASE}(\mathrm{ADL}-\mathrm{intra}, i)} \times 100\% \tag{12}$$

$$\mathrm{PctRed}(\mathrm{ADL}-\mathrm{intra}-\mathrm{IC}, i) = \frac{\mathrm{MASE}(\mathrm{ADL}-\mathrm{intra}, i) - \mathrm{MASE}(\mathrm{ADL}-\mathrm{intra}-\mathrm{IC}, i)}{\mathrm{MASE}(\mathrm{ADL}-\mathrm{intra}, i)} \times 100\% \tag{13}$$

We then take the average value of $\mathrm{PctRed}(\mathrm{ADL}-\mathrm{intra}-\mathrm{EWC}, i)$ and $\mathrm{PctRed}(\mathrm{ADL}-\mathrm{intra}-\mathrm{IC}, i)$ respectively across all the SKUs for each product category. Table 6 shows the results for each product category for one-to-eight weeks ahead forecast horizon[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 et al., 2016). Figures 3(a) and 3(b) show the boxplots for the percentage reduction in the MASE for selective product categories where the two methods respectively produce the greatest improvement in forecasting performance compared to the ADL-intra model.

[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.
[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.
[15] The comparison results for other error measures and horizons are similar and thus omitted for simplicity.

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 for one-to-eight weeks ahead forecast horizon

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 week 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 weeks ahead forecast horizon for 1605 SKUs for the same 28 product categories from a different set of 28 stores

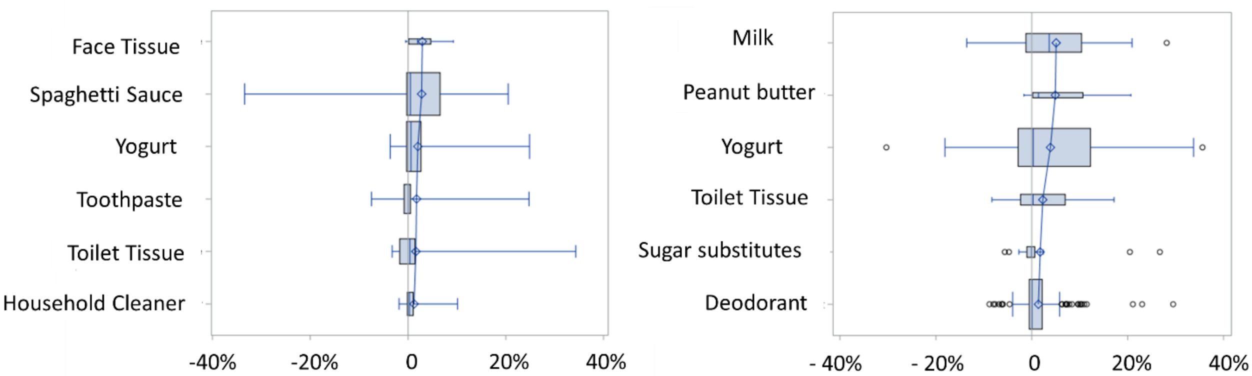
All forecast period, for 1 to 8 weeks ahead					
Model/measure	MAE	SMAPE	MASE	AvgRelMAE	scaled MSE
ADL-intra	13.46	39.91%	0.7669	0.997	0.1674
ADL-intra-EWC	13.47	39.79%	0.7650	0.993	0.1674
ADL-intra-IC	**13.39**	39.50%	0.7592	0.986	**0.1660**
ADL-EWC-IC	13.41	**39.49%**	**0.7588**	**0.985**	0.1661
promoted 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 weeks ahead 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%
Cigarettes	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 ahead forecast horizon for selected product categories.



(a) the ADL-intra-EWC method (b) the ADL-intra-IC method

The box widths are proportionate to the number of SKUs 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.

8. 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 have focused on incorporating additional information to build better forecasting models (e.g., Gür Ali et al., 2009; Huang et al., 2014; Ma et al., 2016), but 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 changes in economic conditions, changes in consumers' tastes, and new entrants into the market. The data on these external factors are typically not available. Thus, conventional models that assume constant effects of 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 7. The percentage reductions of different error measures compared to the Base-lift method for one-to-eight weeks ahead forecast horizon

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 the forecasts generated by ADL-intra models 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 by changing the estimation window. We also propose the ADL-intra-IC method which attempts to offset the potential forecast bias. The method adds the estimate of the recent 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 Base-lift model. Table 7 shows the forecasting improvement by the ADL-intra-EWC method and the ADL-intra-IC model compared to the Base-lift method averaged over a one-to-eight weeks ahead forecast horizon. Specifically, by using these methods 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. Table 7 also shows the forecasting improvement by the ADL-own-EWC method and the ADL-own-IC method compared to the Base-lift method for one-to-eight weeks ahead 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 also compare the forecasting performance of our proposed methods with conventional econometric models which have similar specifications but overlook the structural change problem. The ADL-intra-EWC method and the ADL-intra-IC method outperform the ADL-intra model, and the ADL-own-EWC method and the ADL-own-IC method outperform the ADL-own model. We conduct the comparison to highlight the benefit of taking into account the problem of structural change as some retailers have tried to take advantage of conventional econometric models (Fildes, Ma, et al., 2018).

We also evaluate the models' forecasting performance depending on whether the focal product is being promoted. We find that the ADL-intra-EWC method has the best performance for the promoted forecast period and that 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 SKUs from a different set of 28 stores, and find that this combined model generates the most accurate forecasts overall. We note that the results are post hoc and based on the same dataset. However, this may suggest a potential for more effective forecasting strategies, and we leave further analysis to future research.

In this study, our proposed methods deliver greater accuracy improvements compared to conventional models for some product categories. This may further raise the question whether our methods lead to greater accuracy improvements for SKUs with some specific characteristics. For example, in an exploratory analysis, we regress the improvement of the forecasting performance (e.g., as defined in equation 12 and 13[16]) on a wide range of measures such as the mean and standard deviation of product sales and price, the intensity of promotion, the proportion of outliers, randomness, and trend (see Fildes, 1992). We find that both of our proposed methods have greater accuracy improvements compared to the ADL-intra models for SKUs 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). The ADL-intra-IC method tends to have smaller accuracy improvements for SKUs with higher proportions of outliers and higher levels of promotion intensity, possibly because it becomes more difficult to make adjustments for the forecast bias when there are too many outliers which are likely associated with promotional activities. This finding is consistent with the forecasting performance of the ADL-intra-IC model for the non-promoted period. Thus, the post hoc results suggest a potential for more effective forecasting strategies where we select the forecasting models based on the data characteristics of the SKU, an interesting question which we also leave to future research.

[16] We have also tried dependent variables for other error measures and we have consistent findings.

The methods proposed in this study are new in the domain of forecasting retailer product sales at the SKU level, but there are areas where further improvements in forecasting performance can be achieved. For the ADL-intra-EWC method, we equally combine the forecasts generated by the ADL-intra model with 10 different estimation windows. It is possible to further explore the model's forecasting performance with different numbers of 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 & 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 attention to the problem of structural change. 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 model is that we need to make strong assumptions concerning the effect of the changing marketing activities. For example, Foekens, Leeflang, and Wittink (1999) modeled the effect of marketing activities as a linear function of previous promotional activities. Their models were not developed for forecasting purposes. In summary, the methods we have proposed in this study produce consistently more accurate forecasts than established alternatives. They also satisfy the practical requirements of retail forecasting in that they are intuitive, they can be developed and operated automatically and can also use readily available data on marketing activities.

Acknowledgments

We would like to thank the IRI company for making the data available. All the analyses and findings in this paper based on the IRI dataset are those solely of the authors and not those of the IRI company.

References

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

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.

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.

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. P., & Hendry, D. F. (1994). Towards a theory of economic forecasting. In C. P. Hargreaves (Ed.), *Nonstationary Time Series Analysis and Cointegration*: Oxford University Press.
Clements, M. P., & Hendry, D. F. (1996). Intercept corrections and structural change. *Journal of Applied Econometrics, 11*(5), 475-494.
Clements, M. P., & Hendry, D. F. (1998). *Forecasting Economic Time Series*: Cambridge University Press.
Clements, M. P., & Hendry, D. F. (1999). *Forecasting non-stationary economic time series*. London: The MIT Press.
Cooper, L. G., Baron, P., Levy, W., Swisher, M., & Gogos, P. (1999). Promocast: A new forecasting method for promotion planning. *Marketing Science, 18*(3), 301-316.
Cooper, L. G., & Giuffrida, G. (2000). Turning datamining into a management science tool: new algorithms and empirical results. *Management Science, 46*(2), 249.
Corsten, D., & Gruen, T. (2003). Desperately seeking shelf availability: an examination of the extent, the causes, and the efforts to address retail out-of-stocks. *International Journal of Retail & Distribution Management, 31*(12), 605-617.
Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: the case of judgmental adjustments to SKU-level demand forecasts. *International Journal of Forecasting, 29*(3), 510-522.
Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics, 13*, 253-263.
Dinner, I. M., Heerde, H. J. v., & Neslin, S. (2015). Creating customer engagement via mobile apps: how app usage drives purchase behavior. *Working paper, 10.2139/ssrn.2669817*.
Divakar, S., Ratchford, B. T., & Shankar, V. (2005). CHAN4CAST: A multichannel, multiregion sales forecasting model and decision support system for consumer packaged goods. *Marketing Science, 24*(3), 334-350.
Elliott, G., Granger, C. W. J., & Timmermann, A. G. (2006). *Handbook of Economic Forecasting* (Vol. 1): North-Holland.
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., Goodwin, P., & Önkal, D. (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., Ma, S., & Kolassa, S. (2018). *Retail forecasting: research and practice*. Working paper. Lancaster University Management School. Lancaster University.

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

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.

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

Harvey, A. (2006). *Seasonality and unobserved components models: an overview*. Paper presented at the Eurostat Conference on Seasonality, Seasonal Adjustment and their Implications for Short-Term Analysis and Forecasting, Luxembourg.

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.

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.

Kolassa, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting, 32*(3), 788-803. doi: https://doi.org/10.1016/j.ijforecast.2015.12.004

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.

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.

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.

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., & Srinivasan, S. (2004). Who benefits from store brand entry? *Marketing Science, 23*(3), 364-390.

Pesaran, M. H., & Pick, A. (2011). Forecast combination across estimation windows. *Journal of Business & Economic Statistics, 29*(2), 307-318. doi: 10.1198/jbes.2010.09018

Pesaran, M. H., Schuermann, T., & Smith, V. (2009). Forecasting economic and financial variables with global VARs. *International Journal of Forecasting, 25*, 642-675.

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

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

Petropoulos, F., Fildes, R., & Goodwin, P. (2016). Do ‘big losses’ in judgmental adjustments to statistical forecasts affect experts’ behaviour? *European Journal of Operational Research, 249*(3), 842-852.

Petropoulos, F., & Kourentzes, N. (2015). Forecast combinations for intermittent demand. [journal article]. *Journal of the Operational Research Society, 66*(6), 914-924. doi: 10.1057/jors.2014.62

Rapach, D. E., & Strauss, J. K. (2008). Structural breaks and GARCH models of exchange rate volatility. *Journal of Applied Econometrics, 23*(1), 65-90.

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

Song, H., & Witt, S. F. (2003). Tourism forecasting: the general-to-specific approach. *Journal of Travel Research, 42*, 65-74.

Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: theory, practice, their gap and the future. *European Journal of Operational Research, 252*(1), 1-26.

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

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

Trusov, M., Bodapati, A. V., & Cooper, L. G. (2006). Retailer promotion planning: improving forecasting accuracy and interpretability. *Journal of Interactive Marketing, 20*(3-4), 71-81.

Wildt, A. R. (1976). *The empirical investigation of time dependent parameter variation in marketing models*. Paper presented at the Marketing Educators' Conference.

Wildt, A. R., & Winer, R. S. (1983). Modeling and estimation in changing market environments. *The Journal of Business, 56*(3), 365-388.

Supplementary Material

Click here to download Supplementary Material: Huang Fildes Soopramanien 2019 SUPPLEMENTARY MATERIALS resubmissio