**View Letter**

Top of Form

|  |
| --- |
|  |
| **Date:** | May 08, 2018 |
| **To:** | "tao huang" t.huang@surrey.ac.uk,sphere082175@hotmail.com |
| **cc:** | ;d.soopramanien@lancaster.ac.uk;r.fildes@lancaster.ac.uk |
| **From:** | Ruud Teunter eesserver@eesmail.elsevier.com |
| **Reply To:** | Ruud Teunter r.h.teunter@rug.nl |
| **Subject:** | Your Submission EJOR-D-18-00185 |
| Ref.:  Ms. No. EJOR-D-18-00185 Forecasting Retailer Product Sales in The Presence of Structural Breaks European Journal of Operational Research      Reviewer #1: This is a nice paper that investigates the influence of structural change (see one point I make below) on retail data in the IRI data set. It proposes to augment established Autoregressive Distributive Lag (ADL) models, by either fitting multiple models to shorter and shorter terminal segments of the series, then averaging the forecast (so data points after the conjectured structural break are weighted more heavily), or by attempting to estimate a bias from the terminal in-sample fits and correcting for this estimated bias. The forecast accuracy is improved in either way. I have few important points to make and mostly recommend toning down some overly enthusiastic claims.    Medium points:  - Highlights: "Retailer product sales from a wide range of product categories" is not very useful.  We now remove this point from the highlights.  - "Structural breaks" suggests a sudden and abrupt change in a parameter, like a step change in the overall level of the time series. I find this choice of words somewhat unhappy. On the one hand, in a retail environment, I would typically not expect a sudden abrupt change (except in exceptional cases, like a new store opening next door), but rather a gradual one, which would also be more consistent with drivers of such change that the authors discuss, e.g., shifts in lifestyle. On the other hand, there is a vast literature on detecting structural changes in time series (e.g., look at the documentation of the strucchange R package on CRAN), but the authors nowhere apply such a test (which would not be overly useful in my opinion, per above). Thus, maybe "structural change" instead of "structural break" would be a more useful word for the concept.  We thank the reviewer for this helpful advice. We change the term from structural break to structural change and add the following footnote:  “*The term ‘structural change’ is also used interchangeably with the term of ‘structural break’ in the literature. In this study, we use the term “structural change” as in the retail context we may 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 to point this out*.”  Previously we included the description for the structural change test in the Appendix of the supplementary material- now we describe the details about how we conduct the sequential Chow test in section 6.  - The EWC is very similar to estimating a single model with weighted observations (or residuals), with the weights decreasing as we go further into the past. I wonder whether a direct approach like this would be feasible. It would have the advantage of returning a single parameter vector that could be examined and analyzed, whereas in EWC, one has to look at multiple such vectors, one from each of the models we average.  This is possible- we have evaluated the forecasting performance of ADL models with time-varying parameters (TVP), which has the same rationale described here. However, the preliminary results are mixed. One possible explanation is that the model may have too sophisticated structures which are not robust to the nasty retailer data at SKU level. We add the following to the last section:  “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 et al. (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 to future research.  ”  - None of the accuracy measures used is minimized in expectation by an unbiased forecast (see the introduction of Kolassa, 2016, IJF). I am somewhat concerned that the increase in accuracy may have come at the cost of bias. I would recommend using either a measure that is minimized in expectation by an unbiased forecast, e.g., a scaled RMSE, or assessing bias in some way.  We add the scaled MSE as one of the error measures. The increase in the accuracy come from the reduction of the bias but at a cost of increased forecasting error variance. The IC method adds the estimated bias back to the forecasts, which directly reduces the bias but increase the forecast error variance. The EWC method combines the forecasts from different estimation windows. Compared to the original full estimation window, other estimation windows are all smaller but more recent (e.g., closer to the forecast origin). Forecasts based on these ‘smaller’ estimation windows are less biased (e.g., contain fewer pre-structural change data) but potentially with higher forecast error variance (e.g., fewer estimation information). Thus, the final forecasts will be less biased and potentially with higher forecast error variance. Whether or not the final forecasts by the EWC method can be more accurate depends on the trade-off between the reduced forecast bias and the potentially increased forecast error variance.    - I see no discussion of how the estimation was actually carried out. Was some third-party software used, e.g., an R package? Please cite software and packages used (with version numbers). Software authors are entitled to recognition.  We estimate the model parameters using the OLS estimator, and we implement the estimation using the MODEL procedure in SAS. 9.4. We add this in section 6 and section 7.  - p. 11: what is "four-week seasonality"? Does this refer to cutting the year into 13 four-week periods (13\*4=52), then expending 12 dummies? If so, this corresponds to fitting a periodic \*step\* function, which stays constant for four weeks and then may change sharply when we enter the next four-week bin. This is discretizing a continuous variable (namely, time). Don't do this. The disadvantages of discretization have been well documented over the years - apart from the almost certainly ecologically invalid step fit I discuss, it also expends far too many degrees of freedom. Instead, use a number of periodic spline transforms of time. Three or four parameters expended here would be much better invested than twelve in discretization. Further reading here: https://stats.stackexchange.com/questions/230750/when-should-we-discretize-bin-continuous-independent-variables-features-and-when https://stats.stackexchange.com/questions/41227/justification-for-low-high-or-tertiary-splits-in-anova (see the links in Glen\_b's answer)  In this study, we use the deterministic four-week dummy variables (e.g., 12 four-week dummy variables for the 52 weeks) to capture the seasonality which cannot be captured by the holiday event dummy variables (e.g., Christmas, New Year’s Day etc.). This approach has the limitation that the effect is assumed to stay constant within the four-week bin and has a cost of degree of freedom. However, in the retailer context, the models are usually estimated with a comparably large sample (e.g., 160 weeks in our study), where the loss of degree of freedom is not an issue. Also, the effect is in theory allowed to change sharply across different four-week bins, but it empirically depends on the data.  In this study, we focus on real retail data which have unique characteristics. e.g., nasty and of high variations, and the sales may change abruptly and cannot be explained by other known factors. Also, we use a sophisticated model specification process where we have a vast number of independent variables initially and then recursively simplify the model. During this process, we remove some of the four-week dummy variables using the LASSO procedure. Therefore, the performance of alternative spline methods for seasonality becomes another empirical question.  We thank the reviewer to point out the possibility to further improve the model’s performance using alternative methods for seasonality. e.g., using a periodic spline function, which were proven useful in modelling seasonality in other fields such as electricity demand. We also thank the reviewer for the reference where the natural splines method is compared with the traditional bin method based on a simulated data. We address the limitation of our current approach in the last section and we leave the evaluation of those alternative methods to future research.  - p. 19: To be honest, I do not find the introduction and discussion of the ADL-EWC-IC model convincing, since it was created after analyzing the performance of the separate models on subsamples. It is always easy to build a new model ex post that appears to perform well, but this is little better than data snooping. Can the ADL-EWC-IC model be compared to the other models on previously unseen data? If not, please label this discussion explicitly as exploratory. Similarly, please revisit the discussion in the first paragraph on p. 25.  We thank the reviewer for the advice. We now evaluate the forecasting performance of the ADL-EWC-IC model based on previous unseen data. That is, for the data from another set of 28 stores. The results are consistent. We claim the ADL-EWC-IC model as “exploratory”.  - p. 23: In discussing a statistically significant positive coefficient for "Randomness and growth" in ADL-intra-EWC, the authors write that "This suggests that our proposed models tend to be more advantageous for the SKUs which are difficult to forecast and exhibit a trend in sales". I do not understand this. The dependent value modeled is MASE, so a positive coefficient of 0.4 should mean that a unit increase in "Randomness and growth" should be associated with a 0.4 unit (up to multiplication by 100, per the footnote to Table 7) \*increase\* in MASE. That is, ADL-intra-EWC should perform \*worse\* than the benchmark for high-"Randomness and growth" series, not \*better\*. Please clarify. Same for the rest of the discussions on the same page, and the third paragraph on p. 25.  The dependent variable is not the MASE but the percentage reduction of the MASE by the ADL-intra-EWC model or the ADL-intra-IC model compared to the ADL-intra model. e.g., it is calculated as [MASE(ADL-intra) - MASE(ADL-intra-EWC)]/ MASE(ADL-intra).  Thus, a positive coefficient of 0.4 indicate that one unit increase in "Randomness and growth" would cause a 0.4 units reduction in the MASE. We now make the description more explicit:  Please note we now use the figure of two decimals below (e.g., we report 0.38 rather than 0.4 below).  “For the percentage reduction of the MASE by the ADL-intra-EWC model and by the ADL-intra-IC model, the estimates of the parameter “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.”  - p. 23, "All the results here indicate that we may pre-test these features for each SKU and then determine the optimal sales forecasting method specifically for that SKU." This is a very exploratory finding, not guided by prior hypotheses. Unless this can be verified on previously unseen data, please emphasize the tentative nature of this recommendation.  We rewrite the sentence as “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”.  - p. 24, " The improved forecasting accuracy for product sales substantially contributes to retailers' profit". This is a very strong claim, and not backed by anything the authors did in their paper. Whether forecast accuracy improvements actually translate into better profits is not this clear-cut, since forecasts must still be translated into operational plans, which are constrained by logistics. In addition, supply chain operations rely far more on quantile forecasts than on point forecasts, since the total order includes safety amounts, and whether these quantile forecasts can be improved by the authors' proposals is not obvious. Please tone this claim down.  We thank the reviewer for this comment and we tone down the claim of our contribution- we remove this sentence and add “Therefore, our study provides retailers more effective forecasting methods.”.  Minor points:  - p. 6, 1st equation: there is no intercept in the formula. If u\_t is assumed to have mean 0, this implies that a price of x=0 is associated with sales of y=0. I assume the authors meant to include an intercept parameter.  We no longer consider the variables to be price or price reductions (as pointed out by the reviewer, the sales will not be zero even there is no price reduction). We only use this example to analytically demonstrate the forecast bias caused by the structural change. We highlight that more sophisticated scenarios (e.g., with endogenous variables) can be proved using simulation and we include the simulation example in the supplementary material where the model has an intercept.  - Table 3: why do different models serve as benchmarks?  All the candidate models significantly outperform the basic benchmark model, i.e., Base-lift model, which is widely used in practice. We compare the ADL-intra-EWC model and the ADL-intra-IC model with the ADL-intra model so that we know how much improvement is contributed by taking into account the problem of structural change. We now highlight this in the manuscript.  - Figure 3: please provide more information in the figure caption instead of in the text or even in footnotes - the reader should not need to hunt through the text to understand the figure. Are the diamonds joined by lines group means? Are box widths \*proportional\* to numbers of SKUs in each category? "Are determined by" can be a log transform, a square root or anything else. Please ensure that the horizontal axes have the same extension so the plots are comparable, and that the whiskers are not cut off by the figure bounding box.    We now describe the details explicitly. The diamonds joined by lines represent the group means and the widths of the box are proportionate to the numbers of the SKUs for each product category. We now show the original boxplots without outliers being clipped.  - Table 7: please indicate in the table caption what "their counterparts" are. The entire table is unclear to me; how do the top and the bottom half differ? What does "Model with 5 factors and category dummy variables" as a caption to the bottom half (or is it?) refer to, in contrast to the top half? The text says on p. 23 that "the horizon is one to eight-week ahead", but the table says "Horizon = 8" - please clarify whether the horizon is eight weeks or \*up to\* eight weeks.  We now rewrite the description for the corresponding section, and we also clarify the description for the horizons. e.g., they are one to eight week ahead forecasts.    - References: please provide full details for Loeb (2015) revised  Typos:  - p. 3, 2nd para, l. 3: "The model which is subject to structural break" - either add "a" or change to "breaks"  Revised  - p. 3, 2nd para, l. 7: remove "in" after "including"  Revised  - p. 4, l. 2: "mention" -> "mentioned"  Revised  - p. 4, l. 3: "values" -> "value"  Revised  - p. 4, 2nd para, l. 2: "the change of" -> "changes in" Revised  - p. 5, l. 5: "forecast" -> "forecasts" Revised  - p. 6, 2nd para, l. 3: "structure" -> "structural" Revised  - p. 6, third equation block: italic and upright versions of beta, X and Q are mixed, which is painful to me (also on p. 7)). Please ensure proper mathematical typesetting. Lowercase x\_{T+h} and uppercase X\_{T+h} is used inconsistently here.  We thank the reviewer to point this out. We now revise those inconsistent versions of letters and symbols. We now revise by discarding the symbol “Q” and using all lowercase and uppercases consistently.  - p. 7, equation for the bias correction: lowercase omega is used both as the starting index of the summation (with the ending index indicated by W - why mix Greek and Latin?) and as the summation index itself.  We now use alternative letters to make it more readable.  - p. 7: "The estimated bias are" should be "is" Revised - p. 10, equation: can be slightly simplified by removing "=\eta, \eta"  We thank the reviewer- perhaps it is more readable if with \eta  - p. 19/20: Figure 3(c) is not "in the bottom-right corner", but in the bottom left.  Revised. Also, we no longer show the boxplot for the ADL-EWC-IC model using Figure 3(c) because we now consider the ADL-EWC-IC model as an exploratory model.    Reviewer #2: This is an interesting paper that is trying to investigate the forecasting performance of several ADL models over retailer product sales. The study is particularly focusing on the effects of structural breaks originating from marketing activities over the products.  Main Comments: 1.     I found the structure of this paper very confusing. For example, the introduction and literature review section are very poorly written with many overlaps and repetitions that are not at all informative for the reader. The contribution of the paper, as outlined, is very weak.  We now revise the structure of the paper and highlight the contribution.  2.     In the introduction the authors should clearly indicate what is the model they are introducing and why. On top of that, it should be made clear to the reader why the specific models seem appropriate for the retail forecasting exercise. It is quite astonishing that the reader does not get a gist of what he/she is going to see unless he/she reaches page 10.  We now introduce the proposed models in the introductions and their rationale.  3.     The literature review is very short and several references are packed all together without any meaningful commentary (for example in page 5). I would really be interested to see what are the findings of studies assuming constant marketing activities, as this would highlight/clarify/validate potentially comparisons with the models at hand.    We now revise the structure and make the logic more explicit. We summarize the findings by the studies which overlook the change of the marketing activities. We add:  “The studies described above try to generate accurate forecasts by capturing the 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 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).”    4.     The text is characterized by some generalizations that make the reader confused on what the authors are claiming. For example, in page 3 '… The data in retailer product sales… macroeconomics).' I would suggest a thorough read-through to the authors in order to make the text more to the point.  Revised  5.     Section 3 and 4 seem a bit redundant the way they are presented. I would expect to see a methodology section, where these two sections could motivate/inform the selection of models in section 6. The ideal approach would be that section 3 and 4 are reduced substantially and included in the commentary of section 6 or if needed in an appendix.  We now merge and reduce section 3 and section 4. The new section demonstrates that structural change leads to forecast bias using a simple example, and also indicate that the IC method and the EWC method may potentially improve the forecasting performance. We now move the analytical demonstration of the EWC method to the appendix.  6.     In section 5, it should be explained what the display and feature percentage is along with the motivation of the selection of these inputs. In my view, the data section should have been after the intro/literature review.  We now add the description below Table 4.  7.     In terms of the analysis during structural breaks, it is interesting that the authors do not make explicitly clear how they detect, test or analyse their results for structural breaks. It would be expected that in that type of paper where forecasting performance is evaluated within the presence of structural breaks that vast emphasis would be given on that issue. The general information provided in section 4 are not enough in my opinion. What tests have been done and what was the result? Except from the Chow test, have the authors investigated the Andrews approach (2003 , Econometrica) or Fixed Regressor Boostrap? The vague analysis on that aspect is a main shortcoming of this paper.  We previously described the details of how we conduct the sequential Chow test in the Appendix in the supplementary material - we now include the description in the section 6. We now conduct the sequential Chow test for up to 95% of the weeks in the estimation period. For example, imagine we have an estimation period of 160 weeks, we conduct the Chow test for each of the 160\*0.95= 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 reserve at least 5% of the weeks for the estimation of the test. Thus, we may obtain up to 152 p-values in total. The null hypothesis of no structural change will be rejected only when none of these p-value is below the threshold. To mitigate the multiple comparison problem, we adopt a very small threshold, i.e., 0.001.  We also address the fact that previous studies have proposed alternative tests (e.g., Donald W K Andrews, 1993; Donald W. K. Andrews & Ploberger, 1994; Bai & Perron, 1998, 2003; Brown, Durbin, & Evans, 1975). However, these tests have different focuses (e.g., the size and the location of the structural change) and stringent assumptions (e.g., a known number of multiple changes as a priori knowledge). The estimation of the locations and the sizes using these tests were not satisfactory (Pesaran and Timmerman, 2005). In our study, the purpose is neither to detect the location nor the number of structural changes, but focusing on investigating the existence of any structural change, so that we can estimate and then offset the bias (using the IC method) or to take a trade-off between the forecast bias and the forecasting error variance (using the EWC method). Therefore, we conduct a sequential Chow test which serves for this purpose and has the benefit of simple implementation. The empirical results suggest that our models generate the more accurate forecasts.  The Andrews’ approach published in *Econometrica* in 2003 (e.g., the end‐of‐sample instability test) is more generally applicable because it can be used when there are very limited data before or after structural change within the estimation sample. In comparison, the sequential Chow test can only be applied when there are enough observations before and after the structural change (e.g., it is an F-test which compares the fit of the model before and after the structural change). However, in this study, we find that it does not matter how far we can apply the test towards each end of the estimation sample. We previously conduct the sequential Chow test for up to 70% of the weeks in the estimation period. We now conduct the sequential Chow test for up to 95% of the weeks in the estimation period. The final results (e.g., the superior forecasting performance of our proposed models) are all consistent. There may be two explanations: first, our evaluations are based on rolling events for nearly two thousand SKU’s. Thus, it is unlikely that the results are sensitive to a specific time period when sequential Chow test fails to detect the structural change for all the SKU’s. Second, as pointed out by the other reviewer, the effects of the marketing activities tend to change in a gradual way rather than an abrupt way. Thus, the sequential Chow test would be able detect the existence of the structural change even the Chow test is not applied when there are very limited data before or after the structural change. Also, the Andrews’ instability test assumes that explanatory variables must be strictly stationary, otherwise the test will be associated with a distorted inference and mix the instability of the explanatory variables with the instability of the regression model. In the retailer context, the assumption of stationarity of the explanatory variables may not always be true as the in general prices increase gradually in a two-to-three-year timeframe due to inflation, though intermittently associated with price reductions.    We thank the reviewer for pointing out the possibility of gaining further improvement if a set of alternative tests can be used (as there is always an opportunity to improve by trying alternative candidate methods) but we leave that to future research.  8.     In the modelling part, I would expect to see a small description at least of the benchmark model, the base-lift model. Also, it is not well explained why Lasso is used two times. What motivates the authors to double apply Lasso shrinkage? Is there any similar study suggesting that? What are the expected benefits of this approach? Given that this paper comes down to a horse-racing application between ADL models constrained by Lasso operators, it is important to clarify why the ADL-raw models is combined with the ADL-own model? Wouldn't be a logical question from the reader as to why the Lasso operator is not used in a third stage, for example? A lot of discussion is provided before-hand (in cases unnecessary), but at section 6 that the reader wants to understand the reasoning and process of modelling, he/she struggles to do so because of the lack of explanations. Another important shortcoming for this work.  We thank the review to point this out. We now describe explicitly the Base-lift benchmark model. We also explain explicitly the modelling process and its rationale. The LASSO was initially used as a variable selection method to identify the important marketing variables (e.g., Huang et al, 2014). It was then used as a model simplification strategy following Ma et al. (2016), and it leads to the ADL-raw model. We choose the LASSO procedure because it proves to be effective, and it is also automatic. However, the LASSO procedure runs the risk of missing important variables (e.g., potentially the price and promotions of the focal product). Thus, we try to bring them back to the ADL-intra model but only if they are retained in the ADL-own model by the LASSO procedure, which suggests that they are useful/important. The supplementary parallel ADL (i.e., model 6, in the revised manuscript), by definition, has fewer explanatory variables compared to the ADL-raw model, thus we suffer less from multicollinearity when implementing the LASSO procedure to simplify the supplementary parallel ADL compared to the ADL-intra model. Therefore, it is unlikely for any truly important variable to be missed by the LASSO procedure for both the ADL-own model and for the ADL-raw model.  During the model specification process, we can keep some of those redundant variables in the ADL-intra model and just bear the cost of reduced efficiency. However, if we miss some important variables, the final forecasts will be biased and the forecast error variance will both rise (and this bias is not the bias incurred by the structural change but from the bias of the parameter estimate). Missing important variables is far more serious compared to the cost of efficiency (Davidson and MacKinnon, 2004). Thus, when we simplify the general ADL model, we bring back the potential important variables as much as possible (e.g., the marketing variables of the focal product were initially included in the general model and later brought back again if they can be retained by the ADL-own model). We are willing to bear the cost of efficiency. However, if we choose to implement the LASSO procedure for the ADL-intra model (i.e., in a third stage), we will be choosing the efficiency at the cost of bias.    9.     In the experimental design section, the representation of the competing models is not clear. The authors should have included perhaps a table or a clear text on what is the differences between models in a concise and transparent way. Additionally, I am not sure why the roll-forward period is equal to two weeks, while the forecasts are done for 1, 4 and 8 weeks. The authors should explain why they used sMAPE over MAPE or why MASE is more informative. In terms of statistical accuracy, RMSE and Theil-U statistics are also quite often used. One final point is why are the authors inclined to use the adjustment of Cooper et al. (2009). This was not quite clear to me.  We now revise this section to add a clear text which highlight the difference between the models.  We implement the models with 18 rolling events, and for each time we roll the full estimation window forward for two weeks. Therefore, we will be able to evaluate our models for longer periods of time span (e.g., under this condition, the initial full estimation window is [1:160] and the last estimation window is [35:194], thus the data we used have a time span from week 1 to week 194, compared to shorter time span from week 1 to week 167 if we only roll forward one week each time). This may potential make our results more robust as we evaluate our models for longer time span.  We include traditional error measures in the evaluation. These error measures capture different aspects of the unobserved loss function for the retailer but all of them have limitations. Therefore, they need to be used as a whole to form the broad picture of the model comparison. We now add the scaled MSE as an additional error measure. It conveys the information of the RMSE and the Theil’s U (i.e., the relative RMSE when the benchmark is a random walk with one-step forecasts).  We use the adjustment of Cooper et al. (2009) to mitigate the bias due to the logarithm transform (e.g., the expected value of the log transformed variable does not equal to the log transform of the expected value).  10.     The section 8 is quite extensive in terms of results in tables, but the comments are quite limited and badly written. For example, is there consistency in ranking across statistical measures? What exactly is the importance of Table 3? At such a forecasting exercise, a Diebold-Mariano test is needed to assess the significance of the differences in performances. The Wilcoxon Sign Rank is not enough. It is also not quite clear what Table 4 is offering. In Figure 3, is there a particular reason why these six product categories are selected for analysis?  We now highlight that the results are consistent for all the error measures.  We now replace the Wilcoxon SR test with Diebold-Mariano (DM) test in Table 3. We include Table 4 to demonstrate the performance of the various models depending on whether the focal product is being promoted. This is because the product sales variations are much higher for the promoted period compared to the non-promoted period. We choose the six product categories for which the proposed models have highest advantage over the ADL-intra model.  11.     Table 5 should probably be within the analysis of the determinants of the forecasting performance. I also find the explanation of the process confusing. Why is this factor selection followed? Have the authors explored another factor analysis?  We rewrite this section explicitly. Table 5 consists of the parameter estimates for two regression models: one with category dummy variables and one without. The parameter estimates for the category dummy variables are all omitted for simplicity.  Minor comments: 1.     The writing of the whole paper is not very good. A lot of proof-reading is required. Some examples: \*     'Under such a circumstance…activities' in the abstract.  Revised \*     '…the generated forecasts may potentially…' in the abstract  Revised \*     SKU abbreviation should be defined in the abstact and in text and then explained perhaps on a footnote. \*     '…proposed holistic methods to generate…' page 4, line 57  Revised \*     Footnote 7 should refer to figure 1.  Revised 2.     Sometimes writing lack of academic standard. For example, there is no numbering in equations, there is extensive use of bullet points within text, tables are not explained with footnotes, many equations appear within text, the tables style is not uniform etc.  Revised 3.     The paper is not well positioned in the OR forecasting literature.  Overall, based on the above I am inclined to reject the paper due to its lack of solid contribution, convincing results, presentation and academic rigour. | |
|  | |

Bottom of Form