Analysis of judgmental adjustments in the presence of promotions

Juan R. Trapero[a,*], Diego J. Pedregal[a], R. Fildes[b], N. Kourentzes[b]

[a] *Universidad de Castilla-La Mancha*
Departamento de Administracion de Empresas, Ciudad Real 13071, Spain
[b] *Lancaster University Management School*
Department of Management Science, Lancaster LA1 4YX, UK

Abstract

Sales forecasting is increasingly complex due to many factors, such as product life cycles that have become shorter, more competitive markets and aggressive marketing. Often, forecasts are produced using a Forecasting Support System that integrates univariate statistical forecasts with judgment from experts in the organization. Managers add information to the forecast, like future promotions, potentially improving accuracy. Despite the importance of judgment and promotions, the literature devoted to study their relationship on forecasting performance is scarce. We analyze managerial adjustments accuracy under periods of promotions, based on weekly data from a manufacturing company. Intervention analysis is used to establish whether judgmental adjustments can be replaced by multivariate statistical models when responding to promotional information. We show that judgmental adjustments can enhance baseline forecasts during promotions, but not systematically. Transfer function models based on past promotions information achieved lower overall forecasting errors. Finally, a hybrid model illustrates the fact that human experts still added value to the transfer function models.

Keywords: Demand forecasting, Judgmental adjustments, Promotions, Transfer function, Intervention analysis

*Corresponding author.
Email addresses: juanramon.trapero@uclm.es (Juan R. Trapero), diego.pedregal@uclm.es (Diego J. Pedregal), r.fildes@lancaster.ac.uk (R. Fildes), n.kourentzes@lancaster.ac.uk (N. Kourentzes)

1. Introduction

Manufacturing firms are fundamental in supporting most modern economies. Such companies have to face an increasingly competitive environment due to world globalization. Under such circumstances, improvements in supply chain management can lead to competitive advantage. In this context, Stock-Keeping Unit (SKU) demand forecasting is of paramount importance to reduce inventory investment, enhance customer satisfaction and improve distribution operations. Demand forecasting at SKU level relies on a particular type of a Decision Support System, known as a Forecasting Support System (FSS), (Fildes, Goodwin, & Lawrence, 2006). This FSS integrates a univariate statistical forecasting approach (system forecast) delivering the baseline forecast with managerial judgment from forecasters in the organization.

Univariate forecasting methods are based on time series techniques that analyze past history sales in order to extract a demand pattern that is projected into the future (Makridakis, Wheelwright, & Hyndman, 1998). An example of such techniques is the family of exponential smoothing methods (Gardner, 2006; Hyndman, Koehler, Ord, & Snyder, 2008). This kind of technique is very well-suited to companies that handle numerous SKUs where forecasts must be made semi-automatically. Nonetheless, these methods have some weaknesses, the most important of which for this paper is that they are not able to include potentially relevant additional information, like promotions. In fact, promotional campaigns aim at modifying customer behaviour in order to increase sales. These changes affect customers demand making univariate forecasting algorithms inadequate for predicting promotional sales, since they are based on previous demand patterns, which do not include promotional periods. In order to solve that problem and improve forecast accuracy, managerial adjustments are employed. In fact, Franses & Legerstee (2009) presented a case study where about 90% of al cases were adjusted. Among the reasons to adjust forecasts, Fildes & Goodwin (2007) indicated promotional and advertising activity as the main drivers behind the judgmental adjustment of statistical forecasts.

Given the important relationship between judgmental forecasting and promotions, the first objective of this work is to analyze the accuracy of judgmentally adjusted forecasts applied to promotional campaigns. As far as the authors are concerned, this is the first case study that employs organisational data to verify whether judgmental forecasts during promotional periods achieve lower forecasting errors than its statistical counterpart.

An alternative approach to the problem of promotional forecasts is to use multivariate statistical models that use past promotions information to formulate causal models based on multiple linear regression whose exogenous inputs correspond to the promotion features (price discounts, type of display, type of advertising, etc.), (Özden Gür Ali, Sayin, van Woensel, & Fransoo, 2009; Cooper, Baron, Levy, Swisher, & Gogos, 1999). In this sense, several Promotional Support Systems (PSS) have been developed providing promising results, see SCAN*PRO in Leeflang, van Heerde, & Wittink (2002), PromoCast^{TM} in Cooper et al. (1999), and CHAN4CAST in Divakar, Ratchford, & Shankar (2005). Nonetheless, these methods do not compare their results to judgmentally adjusted forecasts provided by experts. Therefore, our second objective is to develop an enhanced and automatically identified Exponential Smoothing model with intervention analysis and transfer function terms operating on a dummy promotion variable. We can then compare that model with judgmentally adjusted forecast in order to assess the extent of improvement that multivariate models can provide. Weekly real data from a manufacturing company is used to illustrate our findings.

Our results show that judgmentally adjusted forecasts on promotional periods may improve system forecasts, though not systematically, because when adjustments are relatively too large, forecasting accuracy can be reduced. Interestingly, a transfer function model based on past promotions information achieve on average lower forecasting errors than system and judgmentally adjusted forecasts. In addition, a hybrid model is developed based on the transfer function model and expert's judgment to provide the lowest error among the methods considered.

This article is organized as follows: Section 2 summarizes the literature on promotional modeling and the role of judgmental adjustments. Section 3 introduces the case study and carries out an exploratory data analysis. Section 4 explores the main sources of judgmental bias. Section 5 explores different causal models to improve the judgmental forecasts. Section 6 proposes a hybrid statistical-judgmental model to combine the best elements of both approaches and finally, main conclusions are drawn in Section 7.

2. Literature review

Judgmental forecasting is an active research area, where interest has grown increasingly over the last 25 years. It has been recognised that human judgment may lead to important benefits in terms of forecasting accuracy but it can also be subject to many biases (Lawrence, Goodwin, O'Connor, & Önkal, 2006). It should be noted that the topic of managerial adjustments in the particular context of supply chain demand forecasting has remained overlooked until recently (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009). In order to find out what are the variables responsible for bias, Fildes et al. (2009) analyzed adjustments sign and size as bias drivers. They found that negative judgmental adjustments, i.e. those adjustments which judgmentally reduce statistical forecasts, are based on more reliable information and consequently obtain more accurate forecasts than positive adjustments, although this does seem to depend on the choice of error measure (Davydenko, 2012). Also, relatively larger adjustments tended to produce greater improvements than smaller adjustments although part of this effect arises from the particular error measures used (Davydenko, Fildes, & Trapero, 2010).

In order to reduce the bias and make judgmental forecasting more efficient some models with different levels of complexity have been developed to integrate statistical and managers' judgment. For instance, Blattberg & Hoch (1990) equally weighted the importance of each approach. Fildes et al. (2009) suggested an optimal model based on linear regression that distinguished the weight of statistical and judgmental forecasting considering the adjustments sign. Those differences between positive and negative adjustments motivated the use of non-linear models in Trapero, Fildes, & Davydenko (2011), where State Dependent Parameter estimation methods were employed. The results concluded that: i) negative adjustments could be modeled by linear models; ii) positive adjustments followed a non-linear pattern; and iii) the managerial weight should be different depending on the adjustment size.

The effort to find the appropriate balance between statistical and judgmental forecast can help us to understand the biases introduced by forecasters (Goodwin, 2000, 2005). Unfortunately, even after finding an equation to mechanically integrate judgmental adjustments with the baseline forecasts, some implementation issues might arise. For instance, forecasters may find it less motivating to adjust the forecast, putting less effort into performing the task (Belton & Goodwin, 1996); or they may attempt to pre-empt the corrections by modifying their adjustments. Furthermore, the origin of the biases can also be time-varying (Fildes et al., 2009). Therefore, the reduction of judgmental adjustments is usually recommended for managerial adjustments to be beneficial for improving statistical forecasts (Goodwin, 2005). Behind this tendency to over adjust there might be a trend for associating a false pattern to noise randomness. A possible way to reduce the number of adjustments to be made by judgmental forecasters is to use the available information more efficiently. Thus, if we look for a reduction in the number of adjustments, a potential solution would be to model the effects of promotions on sales forecasts, for those promotions where past information is available. Adjustments would then be made to take into account factors excluded from this enhanced model.

Promotions modelling to enhance sales forecasting is not a new topic. Various Decision Support Systems have been designed to accomplished such a task. Cooper et al. (1999) designed a promotion-event forecasting system (PromoCast^{TM}) where historical data was used to build a type of regression model on the basis of 67 variables. Divakar et al. (2005) developed CHAN4CAST, a Web-based decision support system to forecast sales consumer packaged goods. Essentially, most of the literature is centered on comparisons between extrapolation and causal methods; see a literature review by Özden Gür Ali et al. (2009). These methods assume, to some extent, that extrapolation and causal methods improve on the results achieved by judgmental adjustments made by managers, but there is no empirical work that tests this assumption. Therefore, there is too limited a research literature regarding the accuracy of judgmental adjustments when managers respond to promotional information. Whether there are accuracy gains from either causal or extrapolation methods over and above any improvements from managerial adjustments remains a moot point.

3. Case study

Data from a manufacturing company specialized in household products have been collected. The data comprises: i) shipments; ii) one-step-ahead system forecasts; iii) one-step-ahead adjusted or final forecasts; and iv) a dummy variable whose value is 1 if there is a promotion and 0 elsewhere. Our focus is to investigate whether knowing in advance that a SKU is to be promoted can be used to improve managerial forecasts.

The data previously described contains 169 SKUs. In total, 25,012 complete triplets that have been sampled weekly between October 2008 and July 2011. Basically, the final forecast produced by the company is the result of two sources of information (Fildes et al., 2009). On the one hand, there is a computer software which provides the statistical system forecasts. On the other hand, the company forecasters meet with personnel in sales, marketing, and production to share pieces of information that cannot be included in the statistical model. Thus the previous system forecast is adjusted accordingly by the meeting group decisions leading to an agreed forecast, the final forecast.

Observations with shipments, system forecast or final forecast equal to zero have been removed from the original dataset in order to be able to use percentage error measures. After removing those observations, our sample size results in 18,096 cases.

3.1. Research questions

In this section, a descriptive analysis of the data is carried out using common error measures to answer some key questions. The first question we consider when studying judgmental adjustments is:

Q1:*Is judgmental forecasting more accurate than statistical forecasts?*

Table 1 shows some common error measures to assess the accuracy of the system forecast (SF) compared to final forecasts (FF). To achieve that aim the mean absolute percentage error, $MAPE = mean(|PE_t|)$ and the median absolute percentage error, $MdAPE = median(|PE_t|)$ were chosen as accuracy measures, where PE_t is the percentage error given by:

$$PE_t = 100(y_t - F_t)/y_t, \qquad t = 1, \ldots, N \tag{1}$$

Here, y_t stands for the actual value and F_t is the forecast, both of them at time t. The MdAPE is a more robust implementation of MAPE in the presence of outliers, Fildes (1992). These measures are those most commonly used in practice (Hyndman & Koehler, 2006) due to their simplicity of interpretation and applicability to this particular type of dataset. The percentage errors of these forecasts are used to calculate the MAPE and MdAPE of each individual SKU across time, which are afterwards aggregated in dataset average figures, obtaining the Mean(MAPE), Mean(MdAPE) as overall error measures over all SKUs. Based on this overall performance (last row of Table 1), different conclusions may be reached depending on the selected error measure. For instance, the SF Mean(MAPE) is less than the FF. Conversely, Mean(MdAPE) indicates that FF performs better than SF. Therefore we have contradictory results. Nevertheless, this kind of outcome is common in the literature (Trapero et al., 2011). A possible explanation might be the presence of extreme absolute percentage errors that can affect Mean(MAPE), being more sensitive to outlying values than its median counterpart. These extreme values could originate from judgmental adjustments to account for certain promotional information. Then, if promotions may distort the FF performance, the first question should be rephrased as:

Q2:*Is judgmental forecasting more accurate than statistical forecasts when there are promotions?*

The results in the last row of Table 1 have been broken down according to whether there is a promotion or not. The first row (No promo) shows the results when no promotion is applied and the second row (Promo) when there is a promotion. The first column (N) indicates the number of observations in each case. It is interesting to note that the number of observations that are subject to promotions represents just 8% of the whole sample. Table 1 shows that forecasting errors are higher in the presence of promotions. More importantly, FF forecasts, which represent managerial adjustments, lead to worse predictions than system forecasts (SF), and such a conclusion is reached both by Mean(MAPE) and Mean(MdAPE).

Table 1: Mean of MAPE and MdAPE to assess the forecasting error

	N	Mean(MAPE)		Mean(MdAPE)	
		SF	FF	SF	FF
No promo	16581 (92 %)	52.8	58.1	30.7	30.7
Promo	1515 (8 %)	59.1	97.4	47.1	49.9
Overall	18096	52.2	60.6	31.8	31.3

According to these results managerial adjustments increase forecast errors rather than decrease it. Then, how managers reduce forecasting accuracy? The reason behind such a bad performance achieved by FF may be that judgmentally adjusted forecasts appear to be subject to bias (Fildes et al., 2009). In order to explore that possibility, the following question is proposed:

Q3: *Are judgmental forecast adjustments biased?*

The analysis of the forecasting error bias can be accomplished by means of percentage error measures without taking absolute values. Such measures are defined as Mean Percentage Error (MPE) and Median Percentage Error (MdPE). Their advantage is that this way it is possible to aggregate the measures across SKUs. Considering equation (1), a positive value of the percentage error means that actual values are greater than forecasts and vice versa.

The last row in Table 2 shows the overall bias achieved by SF and FF. Regarding judgmental forecasts (FF), both error measures are negative indicating a bias towards optimism. Note that optimism bias has been previously found in literature (Fildes et al., 2009; Trapero et al., 2011). However, the bias associated with SF is not clear because Mean(MPE) is negative but Mean(MdPE) is positive. Therefore, we can only conclude that judgmental forecast are biased towards optimism. In that sense, such an optimism may be consequence of a sales increase expected by managers in promotional periods. Then, the previous question may be rewritten as:

Q4: *Are judgmental forecast adjustments biased in the presence of promotions?*

Overall results in Table 2 are disaggregated depending on whether there are promotions or not. Analyzing SF performance in promotional periods, the actual

Table 2: Mean of MPE and MdPE to assess the forecasting error bias

	Mean(MPE)		Mean(MdPE)	
	SF	FF	SF	FF
No promo	-21.6	-34.5	2.5	-6.2
Promo	8.7	-67.1	30.1	-16.2
Overall	-18.1	-36.3	5.3	-5.9

values are larger than forecasts providing positive values for Mean of both MPE and MdPE. Recall that SF is based on exponential smoothing models that only consider past data and so, it is unable to foresee a higher level of sales due to promotions. Such inability is driving managers to try to correct it by adjusting the SF. Nonetheless, as the Mean of MPE and MdPE of FF shows, their adjustments are too optimistic.

4. Exploring the bias: sign and size of adjustments

The results in the last section showed that judgmentally adjusted forecasts are biased. The literature suggests various factors that may explain the judgmental forecasting bias, such as the size and sign of adjustments (Fildes et al., 2009; Syntetos, Nikolopoulos, Boylan, Fildes, & Goodwin, 2009; Trapero et al., 2011). For example, small adjustments are expected to be less effective than large adjustments (Fildes et al., 2009). Furthermore, positive adjustments, those which increase the system forecast, frequently are less accurate than negative adjustments (Fildes et al., 2009; Trapero et al., 2011). In this section we explore the accuracy of the SF and FF as they depend on the size and sign of adjustments. Additionally, we also investigate those variables in case observations that are subject to promotions.

4.1. Sign of adjustments

Table 3 shows the Mean(MAPE) and the Mean(MdAPE) according to the adjustments sign. The second column shows that positive adjustments are more frequent than negative ones (59.1% against 22.9%). Regarding positive adjustments, it is not totally clear which method is more accurate, since it depends on the error measure. In turn, negative adjustments improve the forecasting accuracy. These conclusions agree with those found in the literature (Fildes et al., 2009; Trapero et al., 2011). Differences found between SF and FF accuracy are more attenuated by using the Mean(MdAPE).

Since the accuracy of the forecasts depends on both promotions and adjustment sign, the next step is to identify the relation between both factors. Table 4 shows the

Table 3: Mean of MAPE and MdAPE to assess the forecasting error according to the adjustment sign

Adjustments	N	Mean(MAPE)		Mean(MdAPE)	
		SF	FF	SF	FF
Positive	10691 (59.1 %)	47.9	65.7	32.5	31.8
Negative	4144 (22.9 %)	69.0	60.6	33.2	32.0
None	3261 (18.0 %)	49.7	49.7	30.8	30.8

Mean(MAPE) and the Mean(MdAPE) according to both the adjustments sign and whether there is a promotion. It can be seen that most of promotion observations are subject to positive adjustments.

In relation to positive adjustments, FF is less accurate than SF for promotional periods. On the other hand, during non promotional observations, the results achieved by Mean(MdAPE) indicate that FF is more accurate than SF. In addition, negative adjustments in both situations improve the forecasting accuracy provided by SF. Finally, it should be noted that just a few promotions have not been judgementally adjusted.

4.2. Size of adjustments

In our dataset there are SKUs with different sales level and variability. Thus, it is convenient to provide a framework where it is possible to compare them. This can be done by means of data normalization. In particular, each product can be normalized with respect to its sales standard deviation, which can be interpreted as measuring the intrinsic difficulty of forecasting the particular SKU. Note that other normalization alternatives are possible. Nonetheless, in this article the SKU sales standard deviation has been chosen as a normalization factor in order to be able to compare our results with previous published works (Fildes et al., 2009; Trapero et al., 2011).

Table 5 shows the forecasting errors measured by normalized MAE. It should be pointed out that usually MAE values are less than one, which means that forecast error variability is less than sales variability. However, this does not hold for promotional periods. This is an indicator of the difficulty associated to forecast promotional sales. During non promotional periods, negative adjustments improve forecasting accuracy whereas positive adjustments do not. On the other hand, errors

Table 4: Mean of MAPE and MdAPE to assess the forecasting error depending on adjustment sign and whether it is subject to promotions

Adjustments		N	Mean(MAPE)		Mean(MdAPE)	
			SF	FF	SF	FF
Positive	Promo	1245 (6.9%)	54.1	104.0	46.9	52.7
	No Promo	9446 (52.2%)	47.6	61.8	31.4	32.0
Negative	Promo	185 (1%)	71.4	68.6	54.0	53.4
	No Promo	3959 (21.9%)	68.6	59.7	32.8	31.6
None	Promo	85 (0.5%)	71.4	71.4	66.9	66.9
	No Promo	3176 (17.6%)	49.9	49.9	31.6	31.6

in promotional periods reveal larger errors for FF than SF for both positive and negative adjustments.

One advantage of normalizing the dataset is that all the observations can be treated as cross-sectional data. This transformation allows us to sort the data as function of size and sign. A useful visual interpretation can be obtained by plotting the MAE versus normalized adjustments, see Fig 1. In this figure, the MAE has been separated according to the adjustment sign. In relation to positive adjustments it is interesting to note that FF is more accurate than SF when normalized adjustments are lower than approximately 3 and SF performs better for bigger adjustments. In the

Table 5: MAE to assess the forecasting error depending on adjustment sign, size and whether it is subject to promotions

	Overall		Promotions		No promotions	
Adjustments	SF	FF	SF	FF	SF	FF
Positive	0.80	0.86	1.19	1.40	0.75	0.78
Negative	0.72	0.69	1.11	1.22	0.70	0.66
None	0.66	0.66	0.74	0.74	0.66	0.66
Total	0.76	0.78	1.15	1.34	0.72	0.73

same figure we provide a histogram of the normalised adjustments by size, indicating that there are mostly small size adjustments, in favour of positive ones.

Regarding negative adjustments some conclusions can be extracted from Table 5 and Fig 1 : i) on average MAE is lower than for positive adjustments; ii) adjustments are smaller than positive ones, for instance the maximum negative adjustment is around -2.5 whereas positive adjustments can reach values close to 10 (i.e, adjustments can be 10 times higher than the variability of shipments for that SKU); iii) FF gives a lower error than SF.

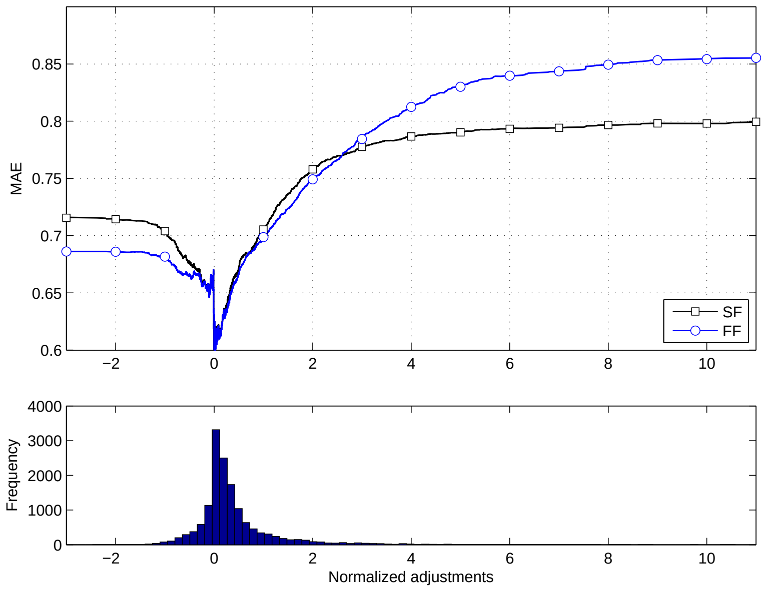

Figure 1: MAE vs. normalized adjustments. Histogram of normalized adjustments

One explanation for the improvement achieved by judgemental forecasting related to positive adjustments is the fact that managers know when there is a promotion and they can increase the values provided by the SF. In order to verify the managers judgment when there are promotions Fig. 2 depicts the MAE obtained for positive and negative adjustments in the presence of promotions. Here, we can see clearly that positive adjustments beat SF when adjustments are lower than approximately 3. Furthermore, larger positive adjustments do not improve SF results. On the other hand, the smaller adjustments lead to improvements in accuracy. In relation to negative adjustments in the presence of promotions, they yield worse results than SF, however their impact is lower than positive adjustments because they are less frequent and smaller as it is shown in its histogram. It is interesting to note that most of the positive adjustments are located between 0 and 3 that is the range of normalized adjustments where FF performs better than SF. That means that in general adjustments improve forecasting accuracy, however, as a consequence of a few large positive adjustments the average accuracy of FF is reduced.

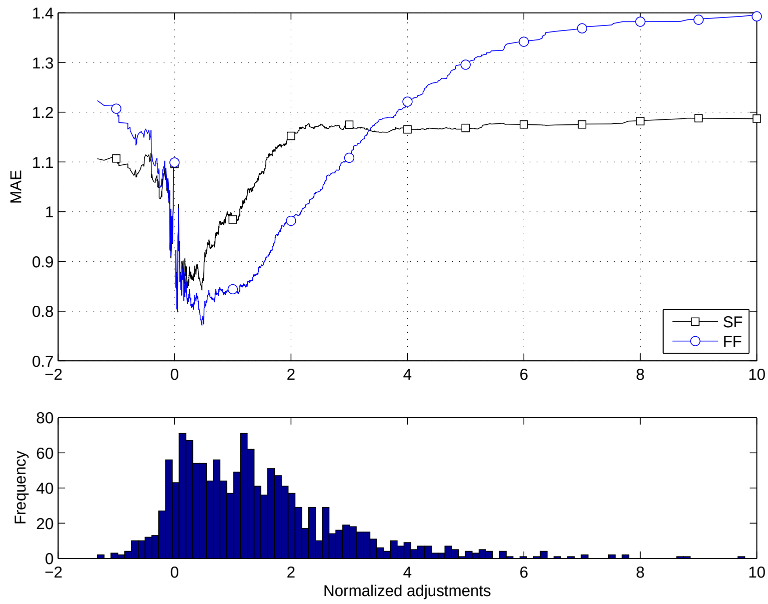

Figure 2: MAE vs. normalized adjustments with promotions. Histogram of normalized adjustments with promotions

Finally, Fig. 3 shows the MAE vs. normalized adjustments and its histogram when there are no promotions. Essentially, positive adjustments do not significantly improve SF. It should be remarked the good performance of negative adjustments included in FF compared to SF.

In summary, positive adjustments are larger and more frequent than negative adjustments. Moreover, positive adjustments may improve forecast accuracy when there are promotions, however their improvement is reduced as adjustments get larger. Finally, negative adjustments when there are no promotions reduce the error considerably. This can be explained by the managers knowing when a promotion is finished and adjusting the system forecast back to its pre-promotion levels.

5. Promotional models

In the previous section we discussed how managers can include information about promotions that their System Forecast is unable to process and consequently they

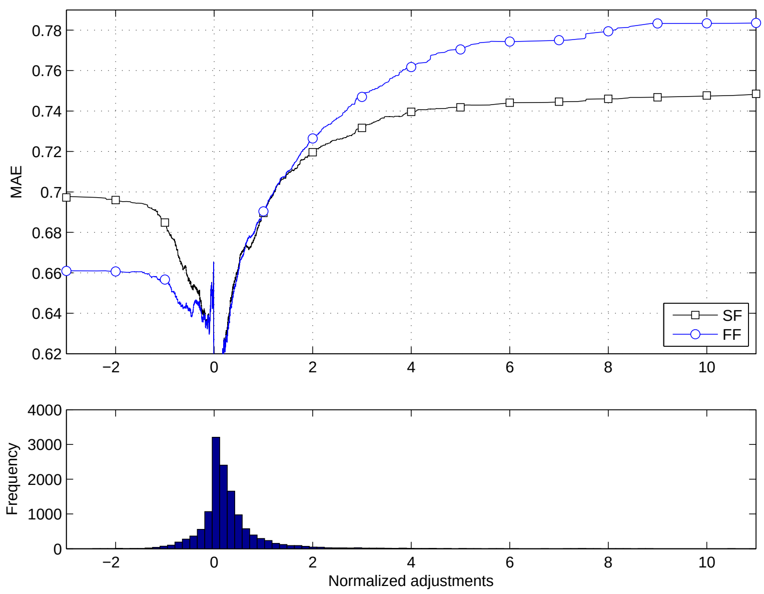

Figure 3: MAE vs. normalized adjustments without promotions. Histogram of normalized adjustments without promotions

might improve forecasting accuracy; see Fig. 2. This result opens the door to the following question: could the managers have obtained valuable information from a systematic analysis of past promotions? In other words, managers could have analyzed past promotions patterns and try to project them for similar future product promotions. In this case, multivariate statistical models could substitute managers adjustments since human minds are not well suited to coping consistently with lots of information (Lawrence et al., 2006) as in this case study.

In order to test whether managerial adjustments can be replaced by multivariate models when dealing with promotions, a simple approach is proposed. This method is based on transfer function models operating on dummy variables that indicate whether there is a promotion or not. Moreover, that transfer function is combined with Exponential Smoothing models. The structure of the transfer function is identified automatically by the Schwarz Information Criterion (Schwartz, 1978).

5.1. Automatically identified Transfer Function

The simplest version of the model is an Exponential Smoothing for the non promotions situations, see (2).

$$\begin{aligned} y_t &= l_{t-1} + e_t \\ l_t &= l_{t-1} + \alpha e_t \end{aligned} \tag{2}$$

where l_t is a time varying level of sales and e_t is a white noise sequence with zero mean and constant variance.

The model is expanded with a transfer function term that operates on a dummy variable consisting of a step for the samples where the promotion is activated, see (3).

$$\begin{array}{l} y_t = l_{t-1} + \frac{B(L)}{A(L)} P_t + e_t \\ l_t = l_{t-1} + \alpha e_t \end{array} \tag{3}$$

where L is the backshift operator such that $L^j y_t = y_{t-j}$, $B(L) = (b_0 + b_1 L + b_2 L^2 + \ldots + b_m L^m)$ is a polynomial in the backward shift operator of order m, $A(L) = (1 + a_1 L + a_2 L^2 + \ldots + a_n L^n)$ is a polynomial of order n and P_t is a binary dummy variable with ones in the weeks where there is a promotion. The time and length of promotions are known in advance.

Previous to the estimation of the transfer function the specific orders of the numerator and denominator polynomials have to be identified. In this particular case it is achieved by minimising the well known Schwartz Information Criterion on a range of possible models. The models include combinations of polynomials for the numerator of orders one to five and zero to one in the denominator.

The amount of information used for the identification of the transfer function models differs depending on the situation. For those SKUs with only one promotion or the first promotion of those SKUs with several promotions, no prior information is available and the model has to be identified dynamically. In those SKUs where several promotions are implemented the model takes advantage of the immediate past promotion, i.e. the model identified for the previous promotion is used as a starting point for the next and all the information concerning the previous promotion is included into the identification stage for the promotion that follows.

This case study is based on one step (week) ahead forecasts. It implies that all models are always estimated on the information available up to the forecast origin in order to produce the best forecasts possible. This means that as a promotion advances in real time different transfer function orders and/or different parameters are used for each of the forecasts produced. All models are estimated by Exact Maximum Likelihood with the ECOTOOL Matlab toolbox, see Pedregal, Contreras, & Sanchez (2010).

5.2. Experimental setup

In sections 3 and 4, the whole dataset was used to analyze the accuracy of the system and final forecasts given by the company. In this section that dataset is split to carry out a predictive empirical evaluation experiment. Since the number of observations with and without promotions are different, we have separated the sample in the following way: i) the estimation sample to determine the coefficient of the exponential smoothing was 50% of the data constituted by non-promoted observations where enough data is available; and ii) given that only 8% of the whole dataset is affected by promotions, all those observations are considered as hold-out sample, i.e., all the promotions will be forecast.

In order to show whether the Transfer Function (TF) model proposed is able to reduce promotional forecasting errors, Figure 4 depicts the MAE associated to the System Forecast (SF), Final Forecast (FF) and Transfer Function model (TF) as a function of the normalized judgmental adjustments on the hold-out sample. It is important to note that the TF model is capable of capturing part of the knowledge that experts have included in their forecast when adjustments are not too large, as well as avoiding large FF errors in those situations when large adjustments were made. In that sense, the TF model achieves the lowest error of the three approaches on average. These results suggest that adjustments applied to promotions can be substituted by statistical models obtaining lower forecasting errors and reducing the workload of managers when judgmentally analysing and adjusting system forecasts.

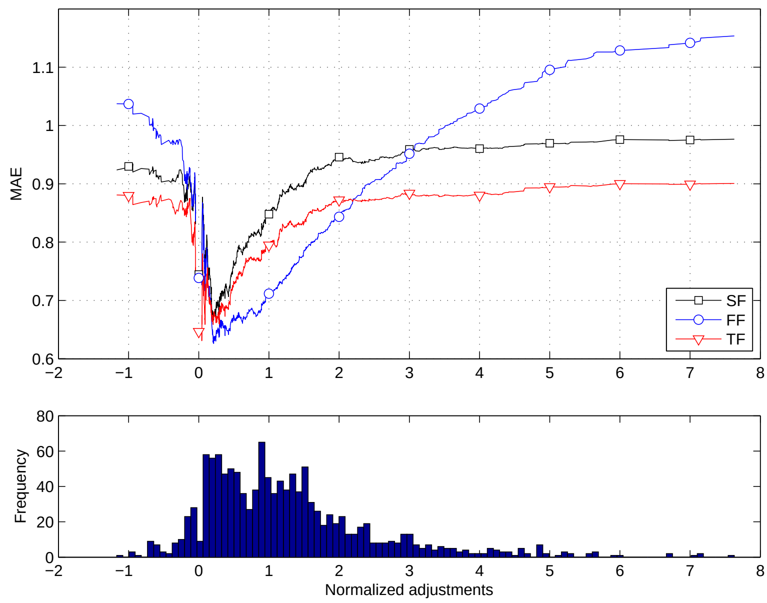

Figure 4: MAE of normalized adjustments with promotions on the hold-out sample

Figure 5 shows the three forecasts considered in the absence of promotions. It can be seen that negative adjustments in FF considerably reduce the forecasting errors. This result agrees with previous references (Fildes et al., 2009; Trapero et al., 2011) where forecasters achieved better forecasts when making negative adjustments because they might handle more realistic information besides promotional one. Additionally, TF forecasts are more accurate than SF ones. A possible explanation is that SF forecasts do not distinguish promotional and non-promotional periods and after several promotional weeks, the univariate algorithm implemented in SF needs some transitory time to reach the average non promotional sales.

In those cases of positive adjustments, generally adjustments also achieve lower forecasting errors. However, because of adjustments whose size is greater than 4 (when normalised by the SKU standard deviation), FF performs worse on average than SF and TF.

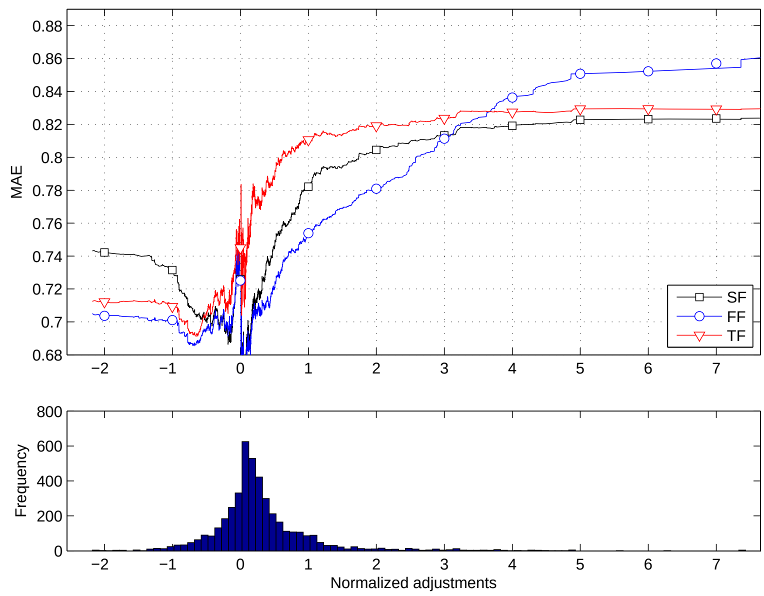

Figure 5: MAE of normalized adjustments without promotions on the hold-out sample

5.3. Experimental results

Table 6 shows the forecasting errors on the hold-out dataset achieved by the SF, FF and TF (Transfer Function) models. The first two rows show the percentage errors Mean(MdAPE) and Mean(MdPE)[1], and the third row shows MAE on the

[1]Note that other percentages error measures as Mean(MAPE) and Mean(MPE) have not been included in Table 6 because they might lead to a misunderstanding given that they are less robust

Table 6: Forecasting errors of SF, FF and TF on the hold-out dataset. The error measures chosen are Mean(MdAPE), Mean(MdPE), MAE and ARMAE

	Overall			Promotions			No promotions		
	SF	FF	TF	SF	FF	TF	SF	FF	TF
Mean(MdAPE)	38.6	40.0	35.8	47.7	50.7	43.1	27.6	26.9	26.9
Mean(MdPE)	19.2	-10.3	9.5	32.0	-16.8	14.1	3.6	-2.3	3.9
MAE	0.819	0.858	0.805	0.956	1.119	0.886	0.786	0.797	0.787
ARMAE	-	1.04	0.97	-	1.14	0.93	-	1.01	1.00

normalized hold-out sample. Essentially, all measures agree that overall SF performance is better than FF as a consequence of the greater forecasting errors made during promotions. Moreover, Mean(MdPE) indicates that judgmental adjustments included in FF are biased towards optimism, particularly when there are promotions.

In relation to the TF accuracy, TF performs better than SF and FF. It is interesting to point out the good results obtained during promotions by TF. Moreover, since TF works as an exponential smoothing during non promotional periods, TF and SF provides similar errors when there are no promotions.

5.4. An alternative error measure: ARMAE

Percentage errors as Mean(MdAPE) and Mean(MdPE) together with MAE on normalized data have been used previously to measure the accuracy of judgmental forecasts and the influence of promotions on such errors. Nonetheless, even when the aforementioned error metrics are commonly accepted, Makridakis (1993) and Hyndman & Koehler (2006) point out the limitations of percentage errors because they overweight the large errors resulting when the actual value y_t is relatively small compared to the forecast error.

In order to overcome those disadvantages of percentage errors the MASE (mean absolute scaled error) was proposed by Hyndman & Koehler (2006). Nevertheless, MASE introduces a bias towards overrating the performance of a benchmark forecast as a result of arithmetic averaging (Davydenko et al., 2010). To avoid that overrating the Average Relative Mean Absolute Error (ARMAE) based on a geometric average is proposed by Davydenko et al. (2010). Here, the system forecast is employed as

to extreme values than their versions based on the median (Mean(MdAPE) and Mean(MdPE)), as we explained earlier.

benchmark, i.e:

$$ARMAE = (\Pi_{i=1}^{m} r_i^{n_i})^{1/\sum_{i=1}^{m} n_i}, \qquad r_i = \frac{MAE_i^a}{MAE_i^b} \tag{4}$$

where MAE_i^b is the MAE for baseline statistical forecast, and MAE_i^a is the MAE for the alternative forecasting methods. The alternatives considered are: i) Judgmentally adjusted forecast (FF); and ii) Transfer function forecast (FF). Both are computed for each SKU i. n_i stands for the number of available errors for the i^{th} SKU and m is the number of SKUs under study. MAEs in (4) are computed as:

$$MAE_i^a = \frac{1}{n_i} \sum_{t \in T_i} |e_{i,t}^a| \tag{5}$$

$$MAE_i^b = \frac{1}{n_i} \sum_{t \in T_i} |e_{i,t}^b| \tag{6}$$

where $e_{i,t}^a$ and $e_{i,t}^b$ represent the errors for the alternative forecasts and the baseline statistical forecasts, respectively. T_i is a set containing time periods for which $e_{i,t}^a$ are available. ARMAE can be interpreted in the following way: values of $ARMAE < 1$ indicate that on average $MAE_i^a < MAE_i^b$ and consequently alternative methods performs better than baseline forecasts.

The last row in Table 6 shows the ARMAE achieved by the various methods we have discussed. It agrees with the rest of error measures that adjustments (FF) do not improve forecasting accuracy when there are promotions. More importantly, the ARMAE agrees with Mean(MdAPE), Mean(MdPE) and MAE that the TF model proposed delivers the lowest forecasting error and shows that when there are promotions, a significant error reduction is also achieved by TF.

6. A Hybrid model to forecast promotional sales

The previous section showed that TF can reduce considerably the forecasting error on promoted weeks. Nonetheless, Figure 4 also shows that FF might achieve the lowest forecasting error for an interval between 0 and 2.5 of normalized adjustments. It is interesting to assess whether the managers' judgmental adjustments still contain useful information for TF. Focusing on the sample where TF forecasts are available and there are promotions, a forecast encompassing test (Fang, 2003) is carried out:

$$y_t = \beta_0 + \beta_1 FF_t + \beta_2 TF_t + e_t, \tag{7}$$

where β_1 and β_2 are constrained to add up to 1 and β_0 permits for the possibility of bias. A significant coefficient of the FF model implies that it captures additional information, which is currently missing in TF. We find that this is the case with a p-value less than 0.01.

Therefore, a potential model that might improve the results could be the combination of TF and FF. In this sense, a hybrid model in terms of a linear regression is defined such as:

$$y_t = \alpha_0 + (\alpha_1 TF_t + \alpha_2 FF_t) + (\alpha_3 TF_t + \alpha_4 FF_t)X_1 + \epsilon_t, \tag{8}$$

where α_0 in stands for the bias and ϵ_t is the error term. X_1 is a dummy variable that allows the weights of TF and FF to vary depending on the adjustments size, i.e:

$$X_1 = \begin{cases} 0 \; if \; 0 \leq adjustments \leq 2.5 \\ 1 \; otherwise \end{cases}$$

In order to evaluate the hybrid model performance, the following experiment is designed. Let the total sample be determined by those observations where TF forecasts are available and under promotions. Then, that sample is divided in two parts. Firstly, the estimation sample which comprises 60% of the total sample, and secondly, the hold-out sample with the rest of observations.

The results of the estimated hybrid model applied to the hold-out sample can be seen in Fig. 6, where the hybrid model provides the lowest error. We compare this model with the rest of methods as well as the 50%-50% model (BH) defined in Blattberg & Hoch (1990). We can see that the hybrid model demonstrates the lowest errors by combining the two sources of information, indicating that there is still useful information in the managers' adjustments.

It is arguable how applicable is such a model in practice. It assumes that the judgmental adjustments will happen the same way, even if experts know that their forecasts will be consequently combined with a statistical model, which is not to be expected. However, it provides evidence that combining statistical models of promotions with judgmental information can lead to substantial gains in accuracy. Furthermore, it illustrates that experts add value to the forecasting process, even if we move away from simple baseline forecasting models to promotional models, as in the TF.

7. Conclusions

Judgmental forecasting has been commonly employed to modify system forecasts when promotions are taking place in order to achieve a lower forecasting error. Interestingly, little empirical research has been done to analyze the efficiency of those

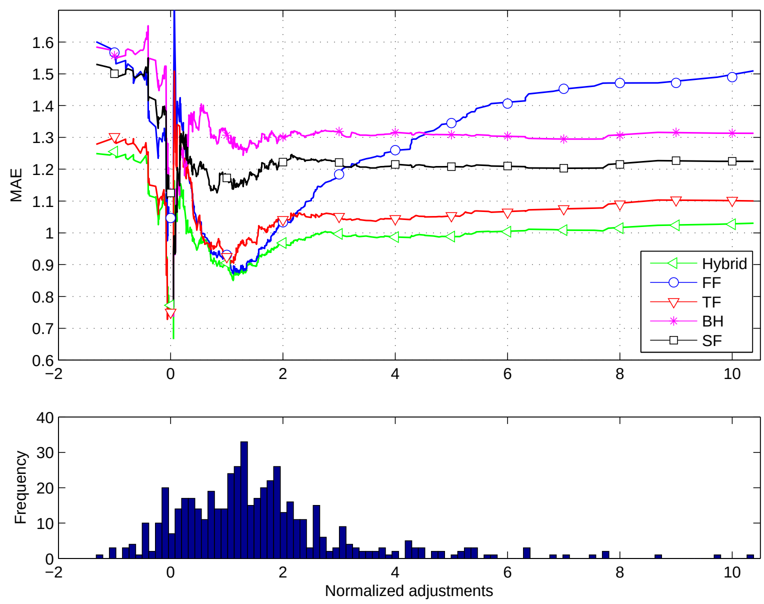

Figure 6: MAE of considered models vs. normalized adjustments. Histogram of normalized adjustments

adjustments. This paper investigates the accuracy of judgmental forecasting in the presence of promotions at SKU level. The first finding shows that judgmentally adjusted forecasts may enhance forecasts under certain circumstances. In order to understand the conditions that lead to better forecasts, the analysis of the sign and size of adjustments has been shown to be crucial. A superficial exploratory data analysis could lead to the conclusion that experts adjustments reduce forecasting accuracy and therefore, this could lead to the suggestion to eliminate judgmental forecasting in promotional periods. A deeper analysis concluded that when adjustments size was not too large experts reduced forecasting error. Additionally, an optimistic bias was also found since positive adjustments tended to provide higher errors than negative ones.

Since experts made their adjustments on the basis of analyzing past promotional demand, an alternative is to substitute judgmental forecasting by a mathematical model to forecast promotional sales. This research has presented a simple model based on a transfer function, automatically identified, combined with a single exponential smoothing. The aim of this model was not to provide optimal forecast but to show that if a simple model could beat judgmental forecasting, investing more resources and effort to develop a sophisticated model would be a worthy objective.

The results showed that this simple model on average could perform better than the expert adjustments.

Modelling promotions have three main advantages. Firstly, experts can benefit from a reduction of effort devoted to judgmentally adjust forecast. Secondly, more accurate forecast can be achieved. Finally, modelling promotional effects can also improve accuracy when forecasting for non-promotional periods, since unusual uplifted sales data as a consequence of a promotion leads to overforecasts carrying over to subsequent periods after the promotion is ended.

Given the potential benefits of mathematical modelling further research should investigate models capable of including efficiently particular information about promotions such as type of advertising, display, price discount, category, etc (Ramanathan & Muyldermans, 2011). Furthermore, this analysis should be extended to longer forecasting horizons and to other companies to make the results more general.

Finally, a simple hybrid model that integrated judgmental adjustments and the transfer function forecasts showed that experts still added value to the forecasts. A limitation of such endeavours is that there is very limited understanding of how experts will change their adjustments in the light of a forecasting support system that adjusts further their forecasts.

Acknowledgment

The authors also would like to acknowledge the support provided by The International Institute of Forecasters and SAS research grant and the Junta de Comunidades de Castilla-La Mancha project number PII1I09-0209-6050

References

Belton, V. & Goodwin, P. (1996). On the application of the analytic hierarchy process to judgmental forecasting. *International Journal of Forecasting*, *12*, 155–161.

Blattberg, R. C. & Hoch, S. J. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science*, *36*, 887–899.

Cooper, L. G., Baron, P., Levy, W., Swisher, M., & Gogos, P. (1999). PromoCast trademark: A new forecasting method for promotion planning. *Marketing Science*, *18*(3), 301.

Davydenko, A. (2012). *Integration of Judgmental and Statistical Approaches for Demand Forecasting: Models and Methods*. Ph.D. thesis, Lancaster University.

Davydenko, A., Fildes, R., & Trapero, J. R. (2010). Measuring the accuracy of judgmental adjustments to sku-level demand forecasts. Tech. rep., Lancaster University Management School Working Paper 2010/026.

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*, 334–350.

Fang, Y. (2003). Forecasting combination and encompassing tests. *International Journal of Forecasting*, *19*(1), 87 – 94.

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

Fildes, R. & Goodwin, P. (2007). Against your better judgment? how organizations can improve their use of management judgment in forecasting. *Interfaces*, *37*, 70–576.

Fildes, R., Goodwin, P., & Lawrence, M. (2006). The design features of forecasting support systems and their effectiveness. *Decision Support Systems*, *42*(1), 351 – 361.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and jugdmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, *25*, 3–23.

Franses, P. H. & Legerstee, R. (2009). Properties of expert adjustments on model-based SKU-level forecasts. *International Journal of Forecasting*, *25*, 35–47.

Gardner, E. S. (2006). Exponential smoothing: The state of the art, Part II. *International Journal of Forecasting*, *22*, 637–666.

Goodwin, P. (2000). Correct or combine? mechanically integrating judgmental forecasts with statistical methods. *International Journal of Forecasting*, *16*, 261–275.

Goodwin, P. (2005). How to integrate management judgment with statistical forecasts. *Foresight: The International Journal of Applied Forecasting*, *1*, 8–12.

Hyndman, R. J. & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*, 679–688.

Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Berlin, Heidelberg: Springer-Verlag.

Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, *22*, 493–518.

Leeflang, P. S., van Heerde, H. J., & Wittink, D. (2002). How promotions work: SCAN*PRO-based evolutionay model building. *Schmalenbach Business Review*, *54*, 198–220.

Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, *9*, 527–529.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: Methods and applications. Third Edition*. John Wiley & Sons.

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

Pedregal, D. J., Contreras, J., & Sanchez, A. (2010). *Handbook of Networks in Power Systems*, chap. ECOTOOL: A general MATLAB forecasting toolbox with applications to electricity markets, pp. 69–104. Springer Verlag.

Ramanathan, U. & Muyldermans, L. (2011). Identifying the underlying structure of demand during promotions: A structural equation modelling approach. *Expert Systems with Applications*, *38*, 5544–5552.

Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, *6*, 461–464.

Syntetos, A. A., Nikolopoulos, K., Boylan, J. E., Fildes, R., & Goodwin, P. (2009). The effects of integrating management judgement into intermittent demand forecasts. *International Journal Production Economics*, *118*, 72–81.

Trapero, J. R., Fildes, R., & Davydenko, A. (2011). Nonlinear identification of judgmental forecasts effects at SKU level. *Journal of Forecasting*, *30*, 490–508.