Lancaster University Management School
Working Paper
2010/003

Judgmental Adjustments to Demand Forecasts: Accuracy Evaluation and Bias Correction

Andrey Davydenko, Robert Fildes and Juan Trapero Arenas

The Department of Management Science
Lancaster University Management School
Lancaster LA1 4YX
UK

Judgmental Adjustments to Demand Forecasts: Accuracy Evaluation and Bias Correction

Davydenko A., Fildes R., Trapero J. R.

a.davydenko@lancaster.ac.uk

Lancaster University
Department of Management Science
LA1 4YX, UK

January 17, 2010

Abstract

Judgmental adjustments to statistically generated forecasts have become a standard practice in demand forecasting, especially at a stock keeping units level. However, due to the subjective nature of judgmental interventions this approach cannot guarantee optimal use of available information and can lead to substantial cognitive biases. It is therefore important to monitor the accuracy of adjustments and estimate persistent systematic errors in order to correct final forecast.

This paper presents an appropriate methodology for such analysis and focuses on specific features of source data including time series heterogeneity, skewed distributions of errors, and generally nonlinear patterns of biases. Enhanced modelling and evaluation techniques are suggested to overcome some imperfections of well-known standard methods in the given context.

Empirical analysis showed that a considerable proportion of final forecast error is formed by a systematic component which can be predicted. Proposed bias correction procedures allowed to substantially improve the accuracy of final forecasts. In particular, one-factor models of the relationship between forecast error and adjustment were found to be a simple, robust and efficient tool for the given purpose.

Keywords: *demand forecasting, judgmental adjustments, judgment under uncertainty, bias correction, accuracy measurement*

Contents

1 Introduction

Judgmental adjustments to statistically generated forecasts have become a standard practice in demand forecasting, especially at a stock keeping units (SKU) level (Sanders and Manrodt, 2003; Fildes et al., 2009). While providing adequate short-term extrapolations of general trends, statistical methods cannot ensure efficient handling of special events due to the limitations of historical data. Manually overriding baseline extrapolations is an easy and fast way to incorporate additional information known to experts into a final forecast. However, the disadvantage of this approach is that the highly subjective nature of judgmental interventions can lead to biases and non-optimal use of available information (Sanders and Manrodt, 2003).

Practically important tasks therefore are to monitor the accuracy of adjusted forecasts, to reveal the degree and the patterns of persistent systematic errors, and to find ways of eliminating them in the given context.

The properties of errors of adjusted demand forecasts have been examined in a number of studies (Fildes et al., 2009; Fildes and Goodwin, 2007; Nikolopoulos et al., 2005; Fildes et al., 2006; Nikolopoulos, 2008; Mathews and Diamantopoulos, 1986, 1989, 1990, 1992). Most of publications showed that adjustments on average lead to the improvements of accuracy, however the accuracy of final forecasts heavily depended on the sign and size of

adjustment. As for the models for optimal correction of adjusted forecasts the existing literature is represented by several works only (Fildes et al., 2009; Fildes and Goodwin, 2007; Fildes et al., 2006) which were based on the same sets of data. It was generally shown that adjusted forecasts suffer from systematic errors and can be efficiently corrected by means of statistical modelling techniques.

However, the results published to date contain mainly descriptive analysis and are confined to the examined datasets only. Moreover, methodologically analysis was carried out in different ways with different assumptions about the statistical properties of the data. These reasons make it difficult to generalise existing results and to form a consistent and coherent set of practical recommendations for companies.

The present paper focuses on common data features and suggests a general methodology for handling judgmental adjustments in demand forecasting systems. In particular, most important data features which were identified and taken into account involve time series heterogeneity, non-negative domain of actual sales data and skewed distributions of forecasting errors, and generally nonlinear patterns of biases.

One of important tasks arising with regard to adjustments data lies in the comparison of forecasts across many SKUs. Though this topic is not new (Hyndman and Koehler, 2005), analysis showed that the methodology for error evaluation when applied to judgmental adjustments is still not sufficiently developed and supported. Some of well-known error measures can give misleading conclusions due to inadequate data transformations which distort the original dependencies between variables. Other measures introduce biases and outliers as a result of arithmetic operations per se. The paper describes the appropriateness of some well-known measures and provides recommendations on constructing more reliable accuracy evaluation schemes.

As for the bias correction task, the paper proposes enhancements to existing approaches in order to develop a more general, quick and robust way of modelling systematic errors. Since theoretical model is not known, empirical data was studied with the use of flexible non-parametric procedures such as local polynomial smoothing (Cleveland and Devlin, 1988) in order to adequately capture the relationship between variables. In addition, appropriate non-linear parametric model was specified and evaluated.

The next section describes the process of using judgmental adjustments along with introducing necessary terminology and notation. Subsequent sec-

tions describe recommended procedures for accuracy evaluation and bias correction. The concluding section summarises the results achieved so far.

2 Source Data and Process of Making Judgmental Adjustment

The research employed data collected from two companies specialising on distribution of fast-moving consumer goods (FMCG). The data contains observed monthly values of actual demand at SKU level, corresponding one-step-ahead statistical forecasts, and judgmentally adjusted forecasts. A dataset for the first company (Company A) relates to 254 SKUs and includes 3012 cases of forecasts and corresponding actual outcomes pertaining to the period from March 2004 to December 2005. For the second company (Company B) the data relates to 413 SKUs, contains 7544 cases pertaining to the period from January 2004 to December 2007. These datasets are used in the paper to illustrate the identified features of data and to evaluate the performance of the suggested procedures for the correction of systematic errors. It is assumed throughout the paper that all forecasts have a fixed constant horizon. Modelling and empirical analysis was based on one-step-ahead monthly forecasts.

The employed data is representative for most FMCG manufacturing or distribution companies. In such settings it is usually needed to deal with a large number of time series of different lengths related to different products with different scales and units of measurement. In order to illustrate the heterogeneity of SKU-level data Fig. 1 shows real series (Company B) containing one-step-ahead statistical forecasts, corresponding adjustments, adjusted forecasts, and actual observations.

Most typically, the process of making judgmental adjustment is performed sequentially and only includes the two following steps (Fildes et al., 2009).

At first, for a given period in future t and a given product or SKU i a statistical forecast $F^{\mathrm{s}}_{i,t}$ is generated by means of a special software package. Usually it is accomplished by applying a simple univariate forecasting method, and the source dataset for that method concerns only past values of sales:

$$F^{\mathrm{s}}_{i,t} = f(\text{past data}).$$

After that the model-based forecast $F^{\mathrm{s}}_{i,t}$ is reviewed by experts (repre-

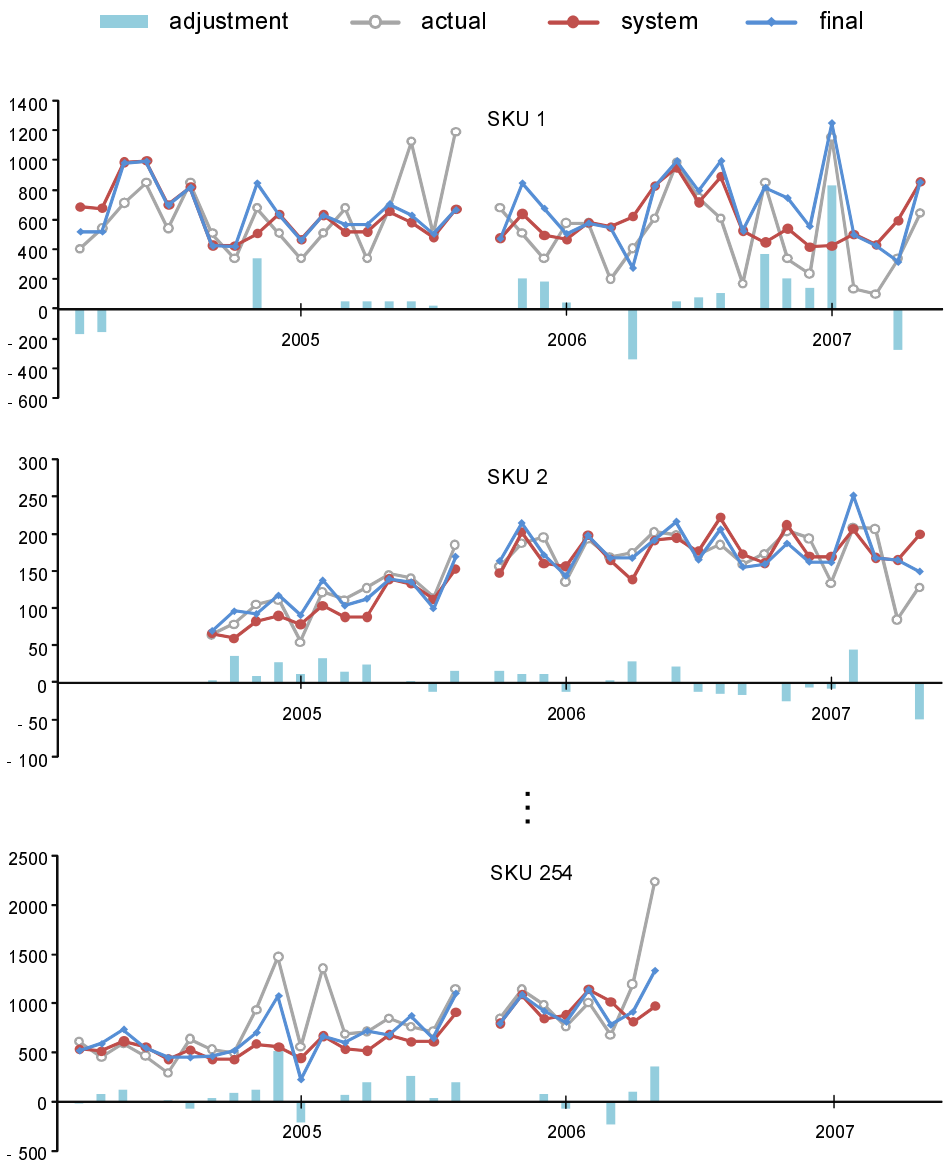

Figure 1: Examples of time series (Company B). The data contains observed monthly values of actual demand at SKU level, corresponding one-step-ahead statistical forecasts, and judgmentally adjusted forecasts.

sentatives from marketing, sales, logistics or production departments). As a result of their revision, the statistical forecast may be adjusted in order to take into account information about exceptional circumstances. The final forecast becomes

$$F^{\mathrm{f}}_{i,t} = F^{\mathrm{s}}_{i,t} + a_{i,t},$$

where i – SKU index, t – a given period in future, $a_{i,t}$ – corresponding adjustment.

If experts are fully satisfied with the statistical forecast and have no additional knowledge about the environment, it is assumed that $a_{i,t} = 0$.

The same procedure is repeated to prepare final forecasts for each SKU i.

This approach is currently very widely adopted because i) it is simple to use, to understand and to implement, and ii) it allows to incorporate the latest information rapidly (Sanders and Manrodt, 2003).

3 Accuracy Evaluation

The major difficulty in measuring the accuracy of judgmental adjustments is caused by the heterogeneity of source data. Usually it is needed to compare forecasting performance across many time series related to different SKUs. Most of the well-known error measures cannot give reliable results in this case because of special features of SKU-level data. In particular, popular measures based on absolute percentage errors can occur inappropriate due to the characteristics of distribution of forecast errors and their correlation with the actual values. This section illustrates the imperfections of several most widely used error measures and provides recommendations on constructing more appropriate criteria and tests.

3.1 Appropriateness of Known Error Measures

A traditional way to compare the accuracy of forecasts across multiple time series is based on using absolute percentage errors (Hyndman and Koehler, 2005). Let the forecasting error for a given time period t and SKU i be

$$e_{i,t} = Y_{i,t} - F_{i,t},$$

where $Y_{i,t}$ – demand value for SKU i observed at time t, $F_{i,t}$ – the forecast of $Y_{i,t}$.

The percentage error (PE) is calculated as

$$p_{i,t} = \frac{100e_{i,t}}{Y_{i,t}}.$$

The most commonly spread PE-based measures are mean absolute percentage error (MAPE) and median absolute percentage error (MdAPE) which are defined as

$$\text{MAPE} = \text{mean}(|p_{i,t}|),$$

$$\text{MdAPE} = \text{median}(|p_{i,t}|),$$

where $\text{mean}(|p_{i,t}|)$ denotes the sample mean of $|p_{i,t}|$ over all available values, and $\text{median}(|p_{i,t}|)$ denotes the sample median.

The disadvantage of these measures is that percentage errors cannot be computed when $Y_{i,t} = 0$ and have skewed distributions when $Y_{i,t}$ is relatively small compared to $e_{i,t}$ (Hyndman and Koehler, 2005). Such situations occur quite commonly with SKU-level data and therefore PE-based measures cannot be used efficiently for the given task.

Alternatively, instead of using actual value in the denominator the forecast error can be divided by standard deviation of all known elements within a time series (Billah et al., 2005). This approach was also used in (Fildes et al., 2009) in order to transform data for modelling the features of adjustments.

Prediction error as a proportion of the in-sample standard deviation of actual values can be written as

$$g_{i,t} = \frac{100e_{i,t}}{s_i}, s_i = \sqrt{\frac{1}{n_i - 1}\sum_{j \in T_i}(Y_{i,j} - \overline{Y_i})^2},$$

where n_i – length of observed time series for SKU i, T_i – a set containing time indexes of observed series elements, $\overline{Y_i}$ – sample mean of observed elements $Y_{i,j}$ for a given SKU i.

Mean absolute prediction error as a percentage of the standard deviation will be further denoted as

$$\text{MAPES} = \text{mean}(|g_{i,t}|).$$

It was noted in (Hyndman and Koehler, 2005) that dividing by standard deviation is not desirable because the denominator grows with the sample

size in time series containing trend. Instead, it is recommended to scale errors by in-sample mean absolute error (MAE) from some benchmark forecasting method.

In case of using system forecast as a benchmark forecast such scaled errors can be found as

$$q_{i,t} = \frac{e_{i,t}}{\frac{1}{n_i}\sum_{j \in T_i} \left|Y_{i,j} - F^{\mathrm{s}}_{i,j}\right|},$$

where $F^{\mathrm{s}}_{i,j}$ – system forecast for SKU i, period j.

Absolute scaled error $|q_{i,t}|$ is interpreted as follows. If $|q_{i,t}| < 1$ then error $e_{i,t}$ arises from a better forecast than the average system forecast computed in-sample, whereas $|q_{i,t}| > 1$ means the opposite.

Analogously to previous measures mean absolute scaled error is defined as

$$\mathrm{MASE} = \mathrm{mean}(|q_{i,t}|).$$

Scaled errors and scaled actual values can be used to illustrate some important properties of adjustments data (Fig. 2, Fig. 3). The shown data relates to one of the companies and includes history of observations of about two years. It can be seen that final forecast errors have truncated and skewed distribution, correlate both with actual values and adjustments, and a substantial proportion of errors is comparable to actual demand values. Errors arising from positive adjustments are on average negative and correspond to low actual values of demand, while negative adjustments on average lead to positive errors and relate to higher actuals. Transition to percentage measures magnifies the errors of positively adjusted forecasts (Fig. 3) due to low values of actuals. Excluding observations with low values (for instance, less than 10 as was done in (Fildes et al., 2009)) still cannot sufficiently improve the properties of percentage errors since a large proportion of data resides in the area where the actual value is less than absolute error.

While having advantages over percentage measures, calculating MASE introduces a bias towards overrating the accuracy of a benchmark forecasting method. It is a well-known fact that as a result of taking arithmetic mean of ratios of loss functions the penalty for bad forecasting becomes larger than the reward for good forecasting (Thompson, 1990). For example, suppose a comparison of accuracy of two forecasting methods is performed across two time series. For the first series the MAE ratio is $r_1 = {}^1/_2$ and for the second series the MAE ratio is the opposite: $r_2 = {}^2/_1$. Averaging the ratios gives

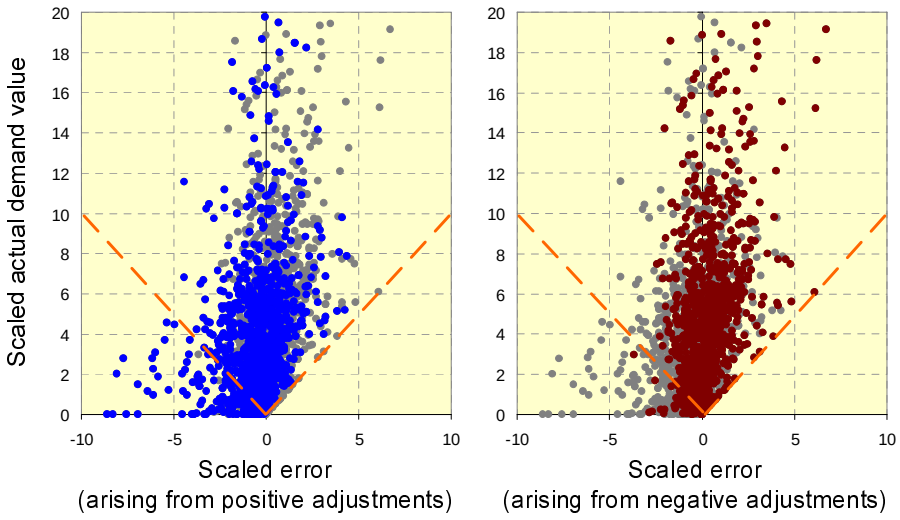

Figure 2: Dependencies between final forecast error, actual value, and sign of adjustment. Absolute errors in the area below the dotted line are higher than actual demand value and therefore lead to substantial distortions of error properties when using percentage measures (Fig. 3).

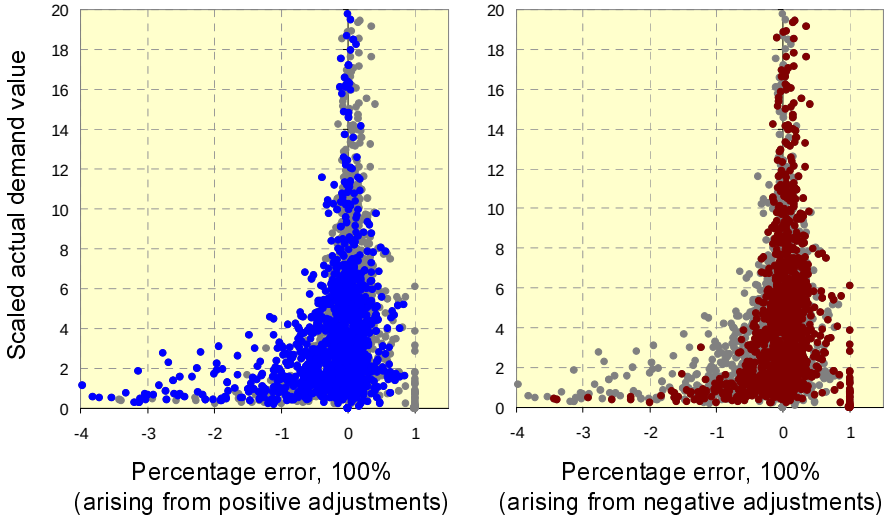

Figure 3: Percentage errors patterns for positive and negative adjustments.

MASE $= \frac{1}{2}(r_1 + r_2) = 1.25$, which indicates that two methods have different accuracy and the benchmark method is superior regardless of the choice of this method (since $r_1 = {}^1/_{r_2}$). This bias was found to be substantial for available datasets, especially in case of short series and large differences in accuracies. Moreover, using MASE (as well as MAPE or MAPES) results in unstable estimations as the arithmetic mean is severely influenced by extreme cases arising from dividing by relatively small values. If such outliers are present, the arithmetic mean can be very different from the mode or median.

To ensure correct additive properties of relative error measures it is possible to apply logarithmical transformations to ratios (Thompson, 1990). A recommended algorithm based on log-transformed ratios of MAEs is described in the next subsection.

Geometric mean of relative absolute errors (GMRAE) (Fildes, 1992) can also be used to overcome the disadvantages of the arithmetic mean. If forecasts $F_{i,j}$ are compared against benchmark forecasts $F^{\mathrm{s}}_{i,j}$ then

$$\mathrm{GMRAE} = \sqrt[k]{\prod_{i=1}^{m}\prod_{j\in T_i}\left|\frac{Y_{i,j}-F_{i,j}}{Y_{i,j}-F^{\mathrm{s}}_{i,j}}\right|}, k = \sum_{i=1}^{m} n_i,$$

where m – total number of series.

However, this measure shows only relative improvement not depending on units of measurement even for the same SKU. Therefore GMRAE is not sufficiently informative with regard to decision-making in the areas of operations management and planning. For example, the error ratio ${}^{1\text{ unit}}/_{10\text{ units}}$ is treated in the same way as the ratio ${}^{100\text{ units}}/_{1000\text{ units}}$, but the implications of these quantities for decision-making process differ dramatically. Error difference in 9 units of a product could probably be acceptable, whereas accuracy reduction of 900 units can lead to serious losses of investments. Thus, the major disadvantage of GMRAE is the inability to take into account the absolute error value at a level of an individual series. In addition, it cannot cover cases of zero forecasting errors (if either $F_{i,t} = Y_{i,t}$ or $F^{\mathrm{s}}_{i,t} = Y_{i,t}$), which reduces the evaluation sample.

3.2 Recommended Accuracy Evaluation Procedure

To ensure the desirable properties of error measures instead of averaging MAE ratios according to MASE scheme it can be recommended to use

weighted mean of log-transformed MAE ratios. Alternatively, it is possible to find weighted geometric mean of MAE ratios. Weighting is necessary for SKU-level data since lengths of time series can differ substantially.

Let the MAE ratio for a given time series i be

$$r_i = \frac{\sum_{j \in T_i} \left| Y_{i,j} - F_{i,j} \right|}{\sum_{j \in T_i} \left| Y_{i,j} - F^{\mathrm{s}}_{i,j} \right|},$$

where $F_{i,j}$ – forecast to be evaluated, $F^{\mathrm{s}}_{i,j}$ – system forecast for SKU i and period j, T_i – a set of time indexes for which actual values and forecasts are known, $Y_{i,j}$ – observed demand value for SKU i and period j.

The recommended measure for forecast comparison is a weighted mean of log-transformed MAE ratios (WLR):

$$\mathrm{WLR} = \frac{\sum_{i=1}^{m} n_i \ln r_i}{\sum_{i=1}^{m} n_i},$$

where m – number of SKUs, n_i – the length of time series for SKU i.

A similar approach was proposed in (Thompson, 1990) where log mean squared error ratios (LMR) were averaged across series. However, the recommended here WLR measure allows the comparison of accuracy across series of different lengths and changes quadratic loss to sums of absolute errors. In general, measures for other loss functions can be defined analogously (unless the loss function for a benchmark forecast over some series is zero).

Obtaining $\mathrm{WLR} < 0$ means that the forecast being evaluated was better than the benchmark forecast for the given dataset, $\mathrm{WLR} > 0$ indicates the opposite.

When applying other measures described above the following precautions should be taken into account: i) results based on percentage errors can be misleading due to correlation between errors, actuals, and adjustments, ii) measures based on arithmetic mean of ratios introduce biases and outliers arising due to the calculation procedure per se.

3.3 Results of Empirical Data Analysis

To evaluate the efficiency of adjustments the described above measures were applied to the dataset for Company A. Overall results are presented in Table 1. Trimmed values of arithmetic means were calculated in order to eliminate the influence of outliers. To ensure robust estimations 1% of largest errors were excluded.

Table 1: Evaluation results for nonzero adjustments data

	System forecast	Final forecast
MAPE, % (1% trim)	30.99	**24.63**
MdAPE, %	18.82	**15.92**
MAPES, % (1% trim)	77.19	**62.67**
MASE (1% trim)	0.97	**0.87**
GMRAE	1.00	**0.84**
WLR	0.00	**-0.16**

According to the results on average adjustments lead to improvements according to all used measures with regard to the given dataset. This agrees with some studies published previously (Fildes et al., 2009; Fildes and Goodwin, 2007; Nikolopoulos et al., 2005; Fildes et al., 2006; Nikolopoulos, 2008; Mathews and Diamantopoulos, 1986, 1989, 1990, 1992) which relied mainly on trimmed MAPE and MdAPE.

To assess the statistical significance of the improvements for various error measures the following tests were applied. Following the approach reported in (Fildes et al., 2009) the difference in MAPE was compared against zero with the use of Wilcoxon's signed paired rank test. The same test was applied to compare MASE measures in a similar fashion to (Hyndman and Koehler, 2005). In all cases the improvements were found to be significant. Importantly, the distribution of the difference between APEs (and the same for absolute scaled errors) was far from Gaussian. This means that applying the t-test as done in (Nikolopoulos, 2008) is generally not advisable for adjustments analysis.

A remarkable fact reported in some recent publications was that the accuracy of adjustments differed depending on their direction. In particular, based on the comparison of pairs of APEs it was found that positive adjustments did not lead to significant improvements (Fildes et al., 2009;

Nikolopoulos, 2008).

Here the same analysis was carried out using additional measures not dependent on actual values. Table 2 and Table 3 present results for subsets of the given dataset.

Table 2: Evaluation results for positive adjustments data

	System forecast	Final forecast
MAPE, % (1% trim)	**25.98**	26.11
MdAPE, %	20.13	**17.00**
MAPES, % (1% trim)	77.80	**66.49**
MASE (1% trim)	0.97	**0.92**
GMRAE	1.00	**0.85**
WLR	0.00	**-0.13**

Table 3: Evaluation results for negative adjustments data

	System forecast	Final forecast
MAPE, % (1% trim)	34.25	**20.74**
MdAPE, %	16.87	**14.96**
MAPES, % (1% trim)	98.93	**68.25**
MASE (1% trim)	0.97	**0.85**
GMRAE	1.00	**0.80**
WLR	0.00	**-0.22**

It can be seen that positive adjustments did not improve accuracy in terms of trimmed MAPE. In the same time, other measures including MdAPE showed improvements yielded by adjustments of both types. Relying on percentage errors in this situation can give misleading conclusions for the following reasons.

Firstly, percentage errors can distort the original features of errors due to dividing by actual values which are correlated with the direction of adjustment (Fig. 2, Fig. 3). Essentially, errors arising from positive adjustments become highly magnified, while errors of negative adjustments are diminished.

Secondly, the distribution of differences between APEs of final and system forecast is skewed since such difference cannot be less than -100%, but has no

upper limit. Applying tests against median becomes therefore problematic due the asymmetry. Moreover, the difference between the MdAPEs is not the same as the median of differences between APEs because errors of final and system forecasts are highly correlated.

It is therefore more advisable to rely on distributions of differences of scaled errors or logarithms of relative measures to assess the significance of improvements in accuracy. For the given dataset MASEs of final and system forecasts differed significantly for both types of adjustments, whereas APEs could indicate significant improvements only for negative adjustments.

4 Correction for Biases

It is well-known that judgments under uncertainty are affected by various types of biases. Particularly, this relates to adjustments employed in numerical predictions (Tversky and Kahneman, 1974). Biases can arise due to inadequacies of human information-processing and motivational factors. In order to improve the quality of judgments it is possible to detect, predict, and compensate systematic errors.

Revealing the influence and the patterns of judgmental biases should be done on the basis of statistical analysis of available data. However, in the current case the data includes many time series relating to different products with different measurement units. The amount of data within a single time series is usually insufficient for finding consistent statistical estimations. Thus, bias correction becomes complicated because of the same problem of data heterogeneity which was addressed in the previous section.

In order to eliminate the differences in the levels of time series for subsequent statistical analysis some works used relative measurements of source variables (Fildes et al., 2009; Fildes and Goodwin, 2007; Nikolopoulos et al., 2005; Fildes et al., 2006). In particular, a model used in (Nikolopoulos et al., 2005) was based on dividing all source variables by system forecast to explore dependencies between percentage errors and percentage adjustments. However, it can be shown that such modelling procedure leads to heteroscedastic errors and distortions of original data features, which results in inefficient and misleading estimations.

Another approach was based on normalising data using standard deviation of actual values within each time series (Fildes et al., 2009). Linear models were then built to describe dependencies between actual values, sys-

tem forecast and adjustments. A major limitation here lies in assuming the type of existing dependencies to be locally linear with arbitrary splitting data into subsets. Choosing the standard deviation of time series elements as a normalisation factor can also lead to inefficiencies and difficulties in interpretation since the standard deviation grows with trend in non-stationary time-series. Moreover, it was found that in this case for some datasets normalised final forecast error becomes correlated with the normalisation factor.

This paper presents a more flexible way to model systematic error of final forecast based on non-linear regression methods and scaling schemes with better properties. In particular, it is suggested to perform scaling with the use of system forecast MAE to avoid dependency between scale and trend.

The major factor that exerts influence on final forecast error was found to be the value of adjustment. This dependency can be described by the following general model:

$$E_{i,t} = f(A_{i,t}) + \varepsilon_{i,t}, \tag{1}$$

where $E_{i,t}$ – scaled final forecast error for SKU i, period t, $A_{i,t}$ – scaled value of corresponding adjustment, $\varepsilon_{i,t}$ – error (noise) which has zero mean.

According to the chosen transformation procedure scaled variables are found using the following formulae:

$$E_{i,t} = \frac{e_{i,t}}{\eta_i}, A_{i,t} = \frac{a_{i,t}}{\eta_i},$$

$$\eta_i = \frac{1}{n_i} \sum_{j \in T_i} |Y_{i,j} - F^{\mathrm{s}}_{i,j}|,$$

where $e_{i,t}$ – the error of final forecast for SKU i and period t, $a_{i,t}$ – the value of adjustment which corresponds to that final forecast, η_i – in-sample MAE of system forecast for SKU i, n_i – number of observations for SKU i, T_i – a set containing time indexes of observed series elements, $F^{\mathrm{s}}_{i,j}$ – system forecast for SKU i and period j, $Y_{i,j}$ – corresponding observed demand value.

Empirical data reveals that this regression function has non-linear form (Fig. 4). In order to find the required conditional expectations with minimal assumptions made about the form of relationship it is possible to choose from a range of nonparamtric smoothing techniques.

Here local polynomial estimators (Cleveland and Devlin, 1988) were used to produce fitted values by locally weighted regression. According to this

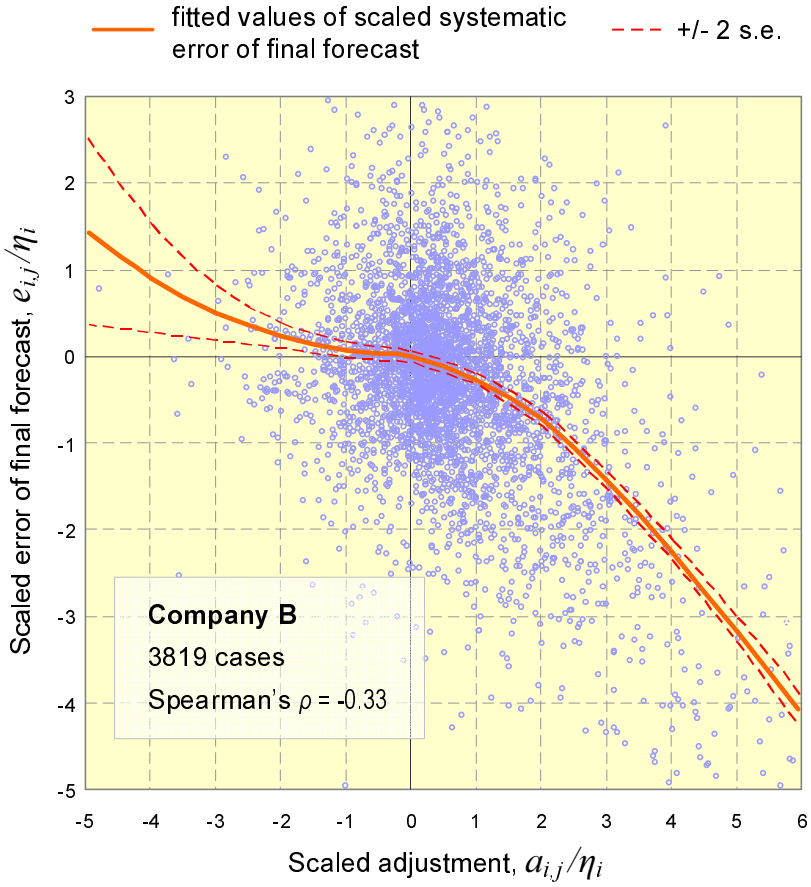

Figure 4: Observed errors of final forecast and estimations of systematic error based on non-parametric smoothing.

method the polynomial is fitted using weighted least squares, giving more weight to points near the point whose response is being estimated and less weight to points further away. The results are shown in Fig. 4. The implementation of this method is available in many statistical packages, the current modelling was done using 'stats' package for R language. Here default recommended span value of 0.75 was used with locally quadratic fit. Confidence intervals for given data should be treated with caution as the regressors are estimated with errors.

Model (1) and the same non-parametric estimation technique can be extended by including more factors such as the system forecast, the previous

forecast error, scaling factor itself, and others.

Alternatively, it is possible to use a parametric approach to specify the regression model. According to empirical data the following analytical expression can be used to approximate the relationship between systematic error and adjustment:

$$f(A_{i,t}) = \begin{cases} \beta_1 (A_{i,t})^{\gamma_1} & A_{i,t} \geq 0, \\ \beta_2 |A_{i,t}|^{\gamma_2} & A_{i,t} < 0, \end{cases} \tag{2}$$

where β_1, β_2, γ_1, γ_2 are model parameters to be estimated (as an option by means of least squares method).

The systematic error is adequately described by model (2) as long as forecast errors are relatively small compared to adjustments. However, it was found that the regression function has an oblique asymptote $E_{i,t} = -A_{i,t}$, which means that the absolute systematic error cannot be higher than the absolute adjustment. Therefore the following more precise formula can be used:

$$f(A_{i,t}) = \begin{cases} \dfrac{\beta_1 (A_{i,t})^{\gamma_1 + 1}}{\beta_1 (A_{i,t})^{\gamma_1} + 1} & A_{i,t} \geq 0, \\ \dfrac{\beta_2 |A_{i,t}|^{\gamma_2 + 1}}{\beta_2 |A_{i,t}|^{\gamma_2} + 1} & A_{i,t} < 0. \end{cases}$$

As it is seen from Fig. 4 for the given dataset the distribution of adjustments is skewed. Positive adjustments result in higher average bias (and consequently higher overall absolute error) since they on average are larger. It is also notable that fitting a linear model for the data shown would produce coefficients very close to the '50% model, 50% manager heuristic' proposed in (Blattberg and Hoch, 1990).

The correction of the judgmental forecast for the presence of systematic error is performed by adding the predicted systematic error to the final forecast:

$$\widehat{Y}_{i,t} = F^{\mathrm{f}}_{i,j} + \eta_i \widehat{E}_{i,t},$$

where $\widehat{Y}_{i,t}$ is a new corrected forecast for SKU i, period t, $F^{\mathrm{f}}_{i,j}$ – corresponding judgmentally adjusted forecast (final forecast), $\eta_i \widehat{E}_{i,t}$ – estimation of systematic error found using one of the aforesaid approaches.

The dataset for Company B was used to evaluate the forecasting performance of the proposed procedure. For this purpose approximately 80% of

sample was used for fitting the model, while the remaining part of the data was used for out-of-sample evaluation. Results are shown in Table 4.

Table 4: Accuracy of forecasts before and after bias correction (out-of-sample)

			Corrected final forecast	
	System forecast	Final forecast	Power regression	Non-par. smoothing
MAPE, % (2% trim)	25.96	25.64	22.57	**22.22**
MdAPE, %	17.88	16.01	14.86	**14.44**
MAPES, % (1% trim)	88.29	83.08	74.44	**73.82**
MASE (1% trim)	0.97	0.95	0.84	**0.83**
GMRAE	1.00	0.89	**0.79**	0.80
WLR	0.00	-0.07	-0.17	**-0.18**

According to the results a considerable proportion (approximately 10%) of final forecast error was predictable. The proposed models allowed to estimate and compensate such persistent errors thereby improving the quality of final forecast.

5 Conclusions and Recommendations

The discovered properties of judgmental adjustments make it possible to draw the following conclusions about the choice of appropriate means for handling them in demand forecasting systems.

It was found that SKU-level data exhibit complex features which can render existing methods for forecast evaluation and correction inappropriate. In particular, due to correlation between errors, actual values, and adjustments it is unadvisable to evaluate adjustments accuracy only using percentage errors such as MAPE or MdAPE. Measures based on relative errors such as MASE and GMRAE occurred to be more suitable with regard to adjustments data. However, it was shown that they can be either not sufficiently informative, or lead to outliers and considerable biases. To ensure more efficient and robust comparison of forecasts an additional error measure was introduced based on weighted average of ratios of forecasts MAEs. With regard to available datasets the recommended accuracy evaluation procedure showed improvements yielded by both positive and negative adjustments.

Empirical evidence suggests that systematic error of final forecast depends mainly on the value of adjustment. The revealed pattern suggests that both for positive and negative adjustments the absolute value of systematic error rises with the increase of the size of adjustment. Adjustments of negative sign lead to positive bias and vice versa. For the given datasets a considerable proportion (approximately 10%) of final forecast error was formed by a systematic and predictable component.

The proposed methods for bias correction involve flexible non-parametrical procedures as well as their parametrical alternatives based on power functions. The analysis carried out showed that such procedures can be recommended as a quick and efficient tool for revealing and eliminating systematic errors in final forecasts.

References

Billah, B., M. King, R. Snyder, and A. Koehler (2005). Exponential smoothing model selection for forecasting. Working Paper 6/05, Department of Econometrics and Business Statistics, Monash University, Australia.

Blattberg, R. and S. Hoch (1990). Database Models and Managerial Intuition: 50% Model+ 50% Manager. *Management Science 36*(8), 887–899.

Cleveland, W. S. and S. J. Devlin (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association 83*(403), 596–610.

Fildes, R. (1992). The evaluation of extrapolative forecasting methods. *International Journal of Forecasting 8*(1), 81–98.

Fildes, R. and P. Goodwin (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces 37*(6), 570.

Fildes, R., P. Goodwin, M. Lawrence, and K. Nikolopoulos (2006). Producing more efficient demand forecasts. Lancaster University Management School Working Paper 2006/054.

Fildes, R., P. Goodwin, M. Lawrence, and K. Nikolopoulos (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.

Hyndman, R. and A. Koehler (2005). Another look at measures of forecast accuracy. Working Paper 13/05, Department of Econometrics and Business Statistics, Monash University, Australia.

Mathews, B. and A. Diamantopoulos (1986). Managerial intervention in forecasting: An empirical investigation of forecast manipulation. *International Journal of Research in Marketing 3*(1), 3–10.

Mathews, B. and A. Diamantopoulos (1989). Judgmental revision of sales forecasts: A longitudinal extension. *Journal of Forecasting 8*, 129–140.

Mathews, B. and A. Diamantopoulos (1990). Judgemental revision of sales forecasts: Effectiveness of forecast selection. *Journal of Forecasting 9*(4), 407–415.

Mathews, B. and A. Diamantopoulos (1992). Judgmental revision of sales forecasts – The relative performance of judgementally revised versus non-revised forecasts. *Journal of Forecasting 11*(6), 569–576.

Nikolopoulos, K. (2008). On the accuracy of judgmental interventions on Statistical Forecasts. Working Paper 0021, University of Peloponnese, Department of Economics.

Nikolopoulos, K., R. Fildes, P. Goodwin, and M. Lawrence (2005). On the accuracy of judgmental interventions on forecasting support systems. Lancaster University Management School Working Paper 2005/022.

Sanders, N. and K. Manrodt (2003). Forecasting software in practice: Use, satisfaction, and performance. *Interfaces 33*(5), 90–93.

Thompson, P. (1990). An MSE statistic for comparing forecast accuracy across series. *International Journal of Forecasting 6*(2), 219–227.

Tversky, A. and D. Kahneman (1974). Judgment under uncertainty: Heuristics and biases. *Science 185*(4157), 1124–1131.