Measuring the behaviour of experts on demand forecasting: a complex task

Nikolaos Kourentzes[a,*], Juan R. Trapero[b], Ivan Svetunkov[a]

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

Abstract

Forecasting plays a crucial role in decision making and accurate forecasts can bring important benefits for organizations. Human judgement is a significant element when preparing these forecasts. Judgemental forecasts made by experts may influence accuracy, since experts can incorporate information difficult to structure and include in statistical models. Typically, such judgemental forecasts may enhance the accuracy under certain circumstances, although they are biased given the nature of human behaviour. Although researchers has been actively looking into possible causes of human bias, there has been limited research devoted to empirically measuring it, to the extent that conclusions can be totally divergent depending on the error metric chosen. Furthermore, most of the error metrics are focused on quantifying the magnitude of the error, where the bias measure has remained relatively overlooked. Therefore, in order to assess human behaviour and performance, an error metric able to measure both the magnitude and bias of the error should be designed. This paper presents a novel metric that overcomes the aforementioned limitations by using an innovative application of the complex numbers theory. The methodology is successfully applied to analyse the judgemental forecasts of a household products manufacturer. This new point of view is also utilized to revisit related problems as the mechanistic integration of judgemental forecasts and the bias-accuracy trade-off.

[*]Correspondance: N Kourentzes, Department of Management Science, Lancaster University Management School, Lancaster, Lancashire, LA1 4YX, UK. Tel.: +44-1524-592911
Email address: `n.kourentzes@lancaster.ac.uk` (Nikolaos Kourentzes)

1. Introduction

Forecasting is of paramount importance to decision making for organisations. For example, in the context of supply chain management accurate forecasts can affect positively the operational management of companies, leading to enhanced customer satisfaction, lower inventory investment and reduced product obsolescence; among other advantages (Moon et al., 2003; Trapero et al., 2012). Such forecasts are often the result of the integration of statistical forecasting with managerial judgement from the forecasters in the organisation. The behaviour of the latter is crucial for the overall performance of the forecasting process.

When those forecasts are computed at the Stock Keeping Unit (SKU) level, a particular type of Decision Support System, known as a Forecasting Support System (FSS) is commonly employed (Fildes et al., 2006). A complete FSS includes a statistical part and a judgemental part, allowing the incorporation of human expertise in the forecasts. Usually, the statistical part is based on univariate time series techniques that analyse past historical data in order to extract a demand pattern that is then projected into the future (Ord and Fildes, 2012). This is often implemented using exponential smoothing methods, due to their relatively good performance, reliability and transparency (Gardner, 2006; Hyndman et al., 2008). Furthermore, such forecasting methods are well-suited to companies that handle numerous forecasts and require some automation.

The statistical part of a FSS is complimented by human judgement to overcome limitations of statistical algorithms. Human judgement is involved in different ways in the forecasting process (Lawrence et al., 2006). Firstly, human judgement is applied in the preliminary time series exploration, where the human expert is trying to identify patterns in a time series in order to select an appropriate statistical method (Harvey, 2007). In some cases model parameters may be selected judgementally, as is often the case in intermittent demand forecasting (Kourentzes, 2014). It is also employed to directly modify the quantity being forecast when the FSS statistical models do not include potentially relevant information (Fildes et al., 2009). In other cases, forecasts are entirely based on human judgement, something that is common for instance with analysts predictions for financial markets (Bozos and

2

Nikolopoulos, 2011). The use of judgemental forecasting in different organizations has been analysed by Sanders and Manrodt (2003), who investigated 240 U.S. corporations and found that 30.3% of them mainly used judgemental methods and 41% employed both quantitative and judgemental methods. Regarding more disaggregated case studies at SKU level Franses and Legerstee (2009) analysed a multinational pharmaceutical company where about 89.5% of all cases were adjusted by experts and Trapero et al. (2011) investigated a chemical company specialized in household products where 65% of SKUs were also judgementally adjusted. The reasons behind such adjustments were explored by Fildes and Goodwin (2007), who identified promotional and advertising activity as the main drivers.

Since judgemental forecasting is based on human inputs, such decisions are subject to heuristics and biases that suggest their behaviour should be analysed (Hämäläinen et al., 2013). Mello (2009) analyses the biases introduced by means of forecast game playing, defined as the intentional manipulation of forecasting processes to gain personal, group, or corporate advantage. Eroglu and Croxton (2010) explore the effects of particular individual differences and suggest that a forecaster's personality and motivational orientation significantly influence the forecasting biases. Kremer et al. (2011) conclude that forecasters overreact to forecast errors in relatively stable environments, but underreact to errors in relatively less stable environments. These findings build on research by psychologists on human behaviour, where several biases have been recognised that are applicable to forecasting such as: anchoring, availability, over-optimism, recency and underestimation of uncertainty (Fildes et al., 2009; Kahneman, 2011; Petropoulos et al., 2014). Other aspects of judgement biases are associated with 'wishful thinking' related to the success of activities that a manager is personally involved or responsible for.

Apart from behavioural aspects, forecasters' bias has also been studied by analysing the direction and magnitude of the judgemental adjustments (Fildes et al., 2009; Trapero et al., 2011, 2013). Recent literature suggests the existence of a bias towards making overly positive adjustments, i.e. to increase the forecast value provided by the statistical baseline forecast (Fildes et al., 2009). Trapero et al. (2013) analysed the forecasting performance of judgemental forecasting in the presence of promotions, being one of the main reasons to judgementally adjust forecasts, and concluded that experts may improve the forecasting accuracy but not systematically. In particular the direction and size of adjustment were found to affect the accuracy of the

adjustments, highlighting the importance of exploring expert behaviour for different situations and bias preferences.

In the literature different mechanistic models have been proposed to counter such forecasting biases. Blattberg and Hoch (1990) proposed to equally weight the statistical and judgemental forecast. Fildes et al. (2009) provided a linear optimal model, where the weight associated to each part were optimized based on past information. Trapero et al. (2011) investigated potential nonlinearities associated with those weights, while Trapero et al. (2013) proposed a hybrid model to combine expert and statistical forecasts depending on the size of the judgemental adjustment.

Although most of the literature agrees with the fact that experts to a certain extent improve forecasting accuracy, the results are inconclusive (Kremer et al., 2011). One of the main causes is due to the utilized forecast error metrics (Davydenko and Fildes, 2013). In fact, some studies arrived at different conclusions depending on the error metric chosen. For instance Fildes et al. (2009) and Trapero et al. (2011) presented case studies where the adjustments improved the forecasting accuracy when Median Absolute Percentage Error (MdAPE) was used, while the opposite conclusion was reached, i.e. the adjustments reduced the forecasting accuracy, if the Mean Absolute Percentage Error (MAPE) was used. In order to overcome these problems with traditional error metrics Hyndman and Koehler (2006) proposed the Mean Absolute Scaled Error (MASE), which is a relative error measure based on the ratio given by the proposed technique's Mean Absolute Error (MAE) divided by the benchmark forecasting method MAE. Recently, Davydenko and Fildes (2013) refined the MASE and proposed the Average Relative MAE (AvgRel-MAE), where the arithmetic means were replaced by geometric means. The latter measure was applied to a supply chain dataset, where the benchmark was the statistical baseline forecast with respect to the judgementally adjusted forecast.

Nevertheless, both the MASE and the AvgRelMAE assess the error magnitude, whereas how to best measure the forecast bias is overlooked in the literature. Since the bias error is a key variable to understand the behavioural aspects of adjustments (Trapero et al., 2013), other error measures are required. In this work, we propose a novel error metric that is capable of providing measures of both error magnitude and bias at the same time. In particular, the new metric relies on the properties of complex numbers to yield a set of metrics and graphs that facilitate the understanding and representation of the experts' forecasting behaviour and performance.

4

The rest of the paper is organised as follows: section 2 discusses the limitations of existing bias metrics and introduces the proposed one; section 3 introduces a company case study and demonstrates the use of the new metric, followed by section 4 that demonstrates some of its modelling benefits. The paper concludes with remarks on the new metric and further research.

2. Measuring forecasting bias

2.1. Existing metrics and limitations

To better illustrate the construction and advantages of the new metric let us assume that we have collected forecast errors from three experts: $E_A = (-5, +6, -2)$, $E_B = (-6, +50, -50)$ and $E_C = (+13, +2, -3)$. We can measure the type and size of their bias by calculating the Mean Error (ME) as $ME = n^{-1} \sum_{j=1}^{n} e_j$, where n is the number of errors e_j, for each expert. This results in $ME_{E_A} = -0.33$, $ME_{E_B} = -2.00$ and $ME_{E_C} = 4.00$ for each expert respectively. The mean errors describe the positive or negative bias of the experts, as well as the magnitude, and assuming that the errors are of the same scale we can rank their performance in terms of bias.

At closer inspection we can observe that there are three issues with this calculation. First, the size of errors is lost. In our example, expert E_B makes the largest forecast errors, yet the resulting ME_{E_B} is smaller than expert E_C who has made much smaller forecast errors. Note that we operate under the assumption that all errors are of the same scale and units, therefore the predictions of expert E_2 are indeed the least accurate. To illustrate the point further, any expert could achieve a zero bias by simply making an adjustment that would incur high error yet cancel out any previous bias, resulting in a ME of zero. A second issue, is that if we calculate the mean bias of all experts it is impossible to decompose back to the performance of individual experts, or appreciate how large errors are cancelled out, resulting in small apparent biases. The next issue is associated with the interpretability of the bias measurement: ME does not provide an insight whether the bias of the experts is high or low. Scale independent variations of ME have attempted to address this issue. It is common to either normalise the errors before ME is calculated, thus adjusting the scale of errors, or to use either the Mean Percentage Error (MPE) or the scaled Mean Error (sME). The latter two are calculated as follows:

$$MPE = \frac{1}{n} \sum_{j=1}^{n} \frac{e_j}{y_j}, \qquad (1)$$

$$sME = \frac{m}{n} \sum_{j=1}^{n} e_j / \sum_{k=1}^{m} y_k, \qquad (2)$$

where y_j are the actual observations at period j and m is the fitting sample size and n the number of errors considered in the metric.

The MPE expresses the bias as a percentage, therefore is scale independent and easy to interpret. Although this helps convey better the size of the bias it has a number of limitations. MPE does not have an upper bound, therefore it still offers limited insight in terms of how large or small a bias is. Furthermore, its measurement is biased as negative and positive errors do not contribute equally. The following example illustrates this: let us assume that the forecast for a period is 90, while the observed demand is 100. The error, measured as the difference between actual and forecasted values will be 10 and the MPE = 10%. If on the other hand the forecast was 100 and the demand was 90 the error would be -10, but the MPE = -11.1%, even though the bias is actually of the same size. The overall MPE over the two cases would imply an average negative bias, where in fact there the negative and positive errors are equal and therefore there should be no bias. Furthermore, should the observed value be zero for a period, MPE cannot be meaningfully calculated.

The sME is the ME scaled by the mean of the observed actuals. That avoids many of the problems of MPE, but assumes that the time series for which the bias is measured is stationary and still does not have an upper bound, thus complicating its interpretation. In addition, for time series that the mean of the observed actuals is zero, or very close to zero, the calculation of sME becomes problematic. Other variations of ME exist, with similar limitations.

2.2. A new metric based on complex numbers

To overcome these limitations we will introduce a new class of bias metrics. Given errors $e_j = y_j - f_j$, where y_j is the observed value and f_j the forecasted value for period j, instead of using the raw errors we can calculate their square root:

$$z_j = \sqrt{e_j} = a + bi. \tag{3}$$

Since errors can be negative z_j can be a real or imaginary number and i is the imaginary unit that satisfies the equation $i^2 = -1$. In Eq. (3) a is the real part and b is the imaginary part of the complex number. For positive errors $a = \sqrt{e}$ and $b = 0$, while for negative $a = 0$ and $b = \sqrt{|e|}$. We name this Root Error (RE). Using this we can define the *Sum Root Error* (SRE) and *Mean Root Error* (MRE) to summarise across several errors:

$$SRE = \sum_{j=1}^{n} \sqrt{e_j} = \sum_{j=1}^{n} a_j + i \sum_{j=1}^{n} b_j, \tag{4}$$

$$MRE = \frac{1}{n} SRE = \frac{1}{n} \sum_{j=1}^{n} a_j + \frac{i}{n} \sum_{j=1}^{n} b_j. \tag{5}$$

Note that Eqs. (4) and (5) are relatively robust to outliers as their impact is reduced due to the square root. The resulting mean root errors for the three experts are: $MRE_{E_A} = 0.82 + 1.22i$, $MRE_{E_B} = 2.36 + 3.17i$ and $MRE_{E_C} = 1.67 + 0.58i$. For the first two experts the real part of the complex number is smaller than the imaginary part, implying that the negative errors are larger and therefore the existence of a negative bias. The opposite is true for the third expert. The results agree with ME with respect to the direction of the bias. If both real and imaginary parts are equal then the forecasts are unbiased. Note however that for MRE_{E_B} both a and b parts are larger than those of other experts, implying that this MRE is the result of larger errors and this expert is less accurate than either E_A or E_C.

Fig. 1 illustrates the differences between the representations of conventional bias and bias using complex errors. In Fig. 1a any errors on the diagonal line have equal real (positive) and imaginary (negative) errors and are unbiased. Any complex errors on the left side of the diagonal exhibit negative bias, while the opposite is true for errors on the right side of the diagonal. Expert B is closer to the diagonal in comparison to Expert A, therefore is less biased. However, in contrast to ME results, illustrated in Fig. 1b, the magnitude of the complex error for Expert B is much larger in comparison to the other experts, revealing the high errors that are cancelled out in the case of ME. This example illustrates that the root error does not
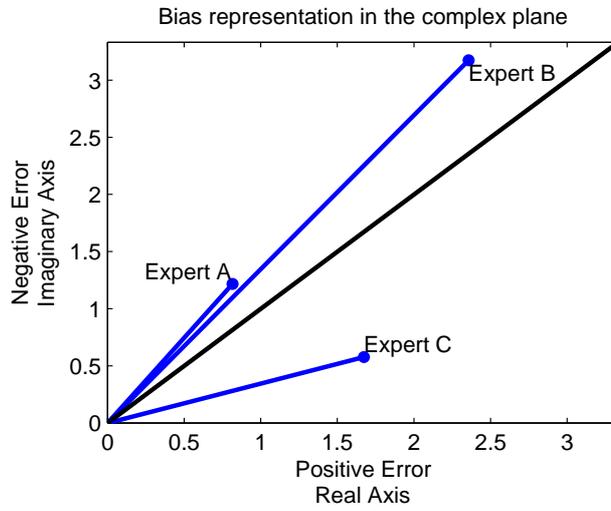
(a) MRE plotted on the complex plane.
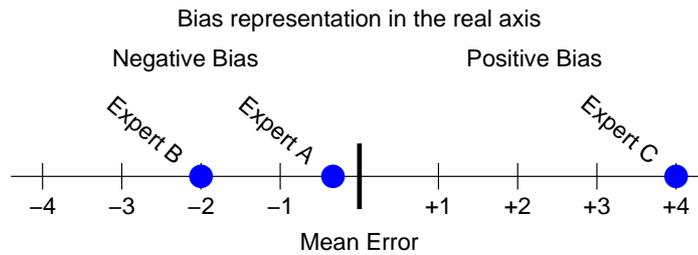
(b) ME plotted on the real number axis.

Figure 1: Complex and conventional representation of bias.

uncouple bias and accuracy, overcoming one of the limitations of conventional bias metrics.

We can take advantage of the complex nature of the errors to simplify their interpretation by expressing them in their polar form. In this case we can separate a complex number to its magnitude r and angle γ, the latter also known as argument or phase:

$$r = \sqrt{a^2 + b^2}, \tag{6}$$

$$\gamma = \begin{cases} arctan(b/a), & \text{if } a > 0 \\ \pi/2, & \text{if } a = 0 \text{ and } b > 0 \\ 0, & \text{if } a = 0 \text{ and } b = 0, \end{cases} \tag{7}$$

where γ is expressed in radians and arctan is the inverse tangent. Note that usually the angle of a complex number is calculated with $+2\pi k$, but in the context of complex error we can set $k = 0$ as other values do not add any valuable information. The connection between the representation of the complex error in the complex plane and its polar form is illustrated in Fig. 2.
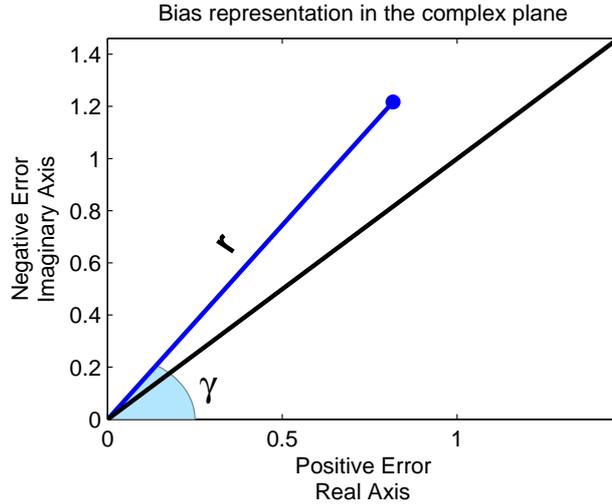

Figure 2: An example of a complex error and its polar form (γ, r).

For the three experts of our example $r_{E_A} = 1.47$, $r_{E_B} = 3.95$ and $r_{E_C} = 1.77$ with $\gamma_{E_A} = 0.31\pi$, $\gamma_{E_B} = 0.30\pi$ and $\gamma_{E_C} = 0.11\pi$. The magnitudes r, which capture the error size, demonstrate again that this information is retained, in contrast to ME. The angle γ reveals the direction and size of bias, irrespective of the magnitude r of the error. By definition it is bounded between $[0, \pi/2]$, as $e_j \in \mathbb{R}$. Unbiasedness will result when $\gamma = \pi/4$, i.e. when $a = b$. Therefore, since $\gamma_{E_A}, \gamma_{E_B} > \pi/4$ both experts are negatively biased, while the third one is positively biased. Up until now we have not been able

to fully describe whether the bias of an expert is large or not as there was no maximum bound on the value of ME and its variants. However, for complex errors γ is bounded, allowing to do precisely that. We can express the bias as the difference from the unbiased state, resulting in a value between -1 (maximum negative bias) up to 1 (maximum positive bias), with 0 being the unbiased result. We define the *Bias Coefficient* κ as:

$$\kappa = 1 - \frac{4\gamma}{\pi}. \tag{8}$$

The bias coefficient is a unit-free metric. For the experts of our example: $\kappa_{E_A} = -0.25$, $\kappa_{E_B} = -0.19$ and $\kappa_{E_C} = 0.58$. A forecast that is always over the observed values will have a $\kappa = -1$, always over-forecasting, while $\kappa = 1$ for the opposite case. Note that since the bias coefficient κ is both unit free and bounded it makes it easy to characterise the bias of experts, either for a single expert, or relatively to others. Crucially the size of the bias coefficient has a clear interpretation that existing bias metrics lack.

Using complex errors permits to easily decompose and aggregate the behaviour of forecasters in terms of bias and error, using the addition and subtraction of complex numbers. Fig. 3 visualises this. Given the forecasting behaviour of two (or more) experts we can aggregate their behaviour to the overall organisational behaviour. Alternatively, given the organisational behaviour and of some experts, we can infer the individual behaviour by subtracting from the overall, both in terms of bias direction and size, as well as error magnitude. This overcomes the third limitation of ME and its variants, which do not provide transparency how each individual behaviour contributes to the overall behaviour.

2.3. Connection with traditional error metrics

So far the discussion has been mostly focused on the bias aspect of the new metric. Here we discuss its magnitude characteristics. Sometimes it is required to provide the error metric in the same units of the actual values. In this sense, we can compute the Squared Mean Root Error (SMRE):

$$SMRE = \left(\frac{\sum_{j=1}^{n} \sqrt{e_j}}{n} \right)^2, \tag{9}$$

so as to retain the units of the metric equal to the observed and forecasted values. The SMRE is a complex number with the main difference to the
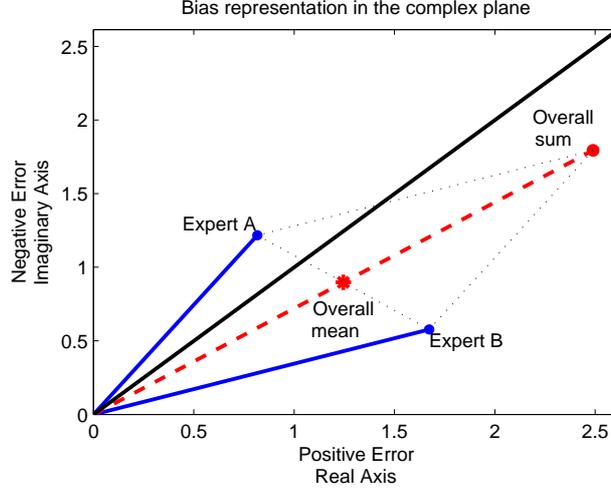
Figure 3: An example of a complex number sum and mean.

well-known Root Mean Squared Error (RMSE) being that the root error is computed first and then the result is squared to return the metric to the original units. Note that, in contrast to RMSE, since we first compute root of the error we retain the information about its sign.

To directly connect the complex metric with conventional ones we can use the Geometric Root Mean Squared Error (GRMSE). This metric has favourable statistical properties being robust to outliers and able to deal with aggregating forecast errors across time series of different scales when comparing the accuracy of alternative forecast sources (Fildes, 1992). The GRMSE is defined as:

$$GRMSE = \left(\prod_{j=1}^{n} e_j^2 \right)^{\frac{1}{2n}}. \tag{10}$$

Calculating the geometric mean of SMRE we construct the Geometric Squared Mean Root Error (GSMRE), which is the complex version of GRMSE:

$$GSMRE = \left(\prod_{j=1}^{n} \sqrt{e_j} \right)^{\frac{2}{n}}. \tag{11}$$

The following properties are true: i) $|GSMRE| = GRMSE$, i.e. the magnitude of the complex number GSMRE is equal to the value of the conventional

11

GRMSE and thus, it provides exactly the same information about the error magnitude and a direct translation of the complex error to conventional ones; ii) Unlike the GRMSE, the GSMRE also offers information about the bias. Particularly, the angle of the GSMRE indicates the proportion of positive or negative errors in the whole sample. It should be noted that the bias bounds found in the previous section for the MRE $[0,\pi/2]$ will be scaled to $[0,\pi]$ due to the square operation.

To illustrate further the benefits of the proposed complex errors in capturing the behaviour of forecasters we use a case study introduced in the next section.

3. Case study

Data collected from a manufacturing company specialising in household products will be used to demonstrate how the root error captures the forecasting behaviour and the insights it offer. The dataset contains historical sales and final forecasts. The final forecasts are produced by human experts adjusting statistical baseline forecasts. The difference between the historical sales and the final forecasts highlight the behaviour of the forecasters in terms of forecasting preferences, which are an amalgam of individual and company level targets, biases and information availability. The forecasters draw upon information from sales, marketing and production personnel to adjust the baseline forecasts to the final values that are eventually used to base decisions made by the company.

The dataset contains weekly forecasts for 156 products. In total 18,096 forecasts are available, resulting in an average of 116 records per product. In this study we will treat each product separately. We do this since the available external information used to adjust the forecasts is often relevant at a product-by-product level and the company does not treat all products equally in terms of importance and expectations.

The sales of different products are on different scale. To avoid using percentage metrics, for the reasons highlighted before, all data are normalised by dividing the sales and the final forecasts by the standard deviation of the sales of each respective product. Each error is now scale and units free and can be used to calculate meaningful statistics across products. This normalisation has been done previously in similar studies (Fildes et al., 2009; Trapero et al., 2011, 2013).

An initial description of the dataset is provided in Table 1. The overall forecast bias as measured by Mean Error (ME) is -0.0623. This clearly shows that there is a negative bias, i.e. on average the company experts are over-forecasting. However, given that it is scaled by the standard deviation of the sales of each respective product, it is difficult to appreciate whether it is a small or a large bias. The Mean Absolute Error (MAE) measures the accuracy of the final forecasts. Again, due to scaling, it is difficult to appreciate whether this is a large or small error. Next, their percentage equivalent metrics are provided. The Mean Percentage Error (MPE) agrees with the ME that there is a negative overall bias, with a size of 36.34%. The Mean Absolute Percentage Error (MAPE) places the magnitude of the errors at 60.61%. Note that the percentage metrics in both cases are misleading in the sense, that an error of 100% is not the maximum and they are biased.

Table 1: Descriptive statistics of the case data

Metric	Value
ME	-0.0623
MAE	0.7934
MPE %	-36.34%
MAPE %	60.61%
MRE	0.3635 +0.4337i
Bias coefficient	-11.19%
GRMSE	0.5839

Finally, in Table 1 the Mean Root Error (MRE) is provided, with a value of 0.36 + 0.43i. This tells us that the imaginary part that corresponds to negative errors is larger than the real part that corresponds to positive errors, thus overall the forecasts are negatively biased. The bias coefficient κ is provided, which is found to be -11.19%. Note that as κ is bounded up to $\pm 100\%$ we can describe the overall observed forecast bias as small. To illustrate the measurement bias introduced in MPE due to the division by sales, if we were to calculate the bias coefficient on percentage errors the resulting κ would be -37.66%, substantially different from its unbiased counterpart. The magnitude r of MRE is 0.57, which when considered only on its own merely provides a measurement of the size of the errors included in the bias calculation. The GRMSE is found to be 0.5839, which as discussed before can be calculated either from the complex errors of the various products of the case company, or conventionally from the errors of the forecasts.

Fig. 4 presents histograms of the forecast bias per product as measured using ME and the bias coefficient κ. In both histograms the vertical thick line represents the unbiased forecasting behaviour, while the dotted line represents the overall company bias. As discussed above the ME histogram is not bounded. Therefore, products that their forecasts are highly biased appear as outliers, distorting the distribution. In the case of the bias coefficient histogram all measurements are within [-1, 1], resulting in an easier to describe distribution and revealing in more detail how the forecast bias of the various products differ, as any outliers are bounded.

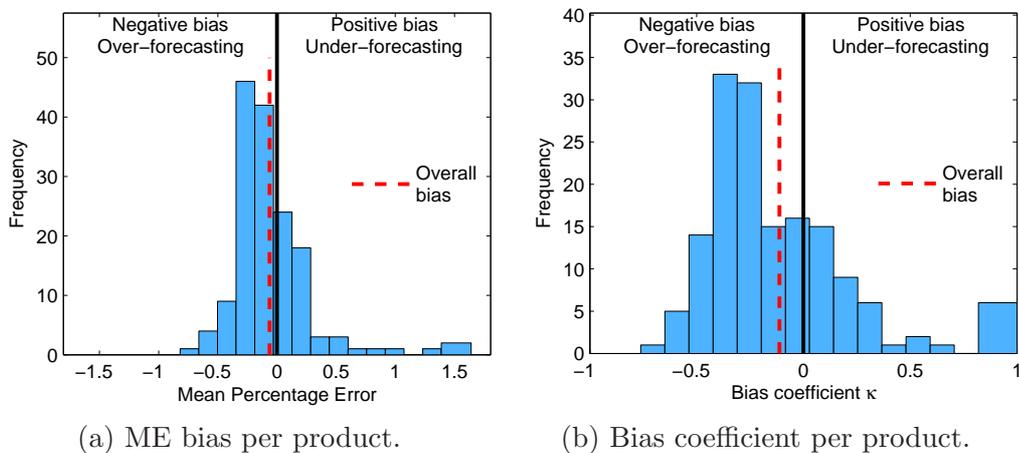(a) ME bias per product.

(b) Bias coefficient per product.

Figure 4: Histograms of forecast bias per product as measured by ME and the bias coefficient.

We can summarise the information contained in Fig. 4b with a box-plot, as in Fig. 5, which captures the distribution of the bias coefficient of the products. The unbiased behaviour is again represented by a vertical thick line. Given the bounded nature of κ we describe cases with $|\kappa| > 0.5$ as strongly biased and the rest as weakly biased, providing a simple and intuitive description of the forecast bias behaviour. Note that this characterisation of bias is not dataset dependent and can be used to easily compare and benchmark behaviours of different experts and companies. The non-symmetric forecast bias behaviour of the case company is apparent. We can clearly see that most products (about 50%) exhibit negative weak bias and about a quarter of the products having strong negative bias. The 3rd quantile of the bias coefficient shows mostly weak negative and unbiased forecasting

14

behaviour, while the remaining 25% of the products exhibits positive bias of increasing strength. A number of products with outlying behaviour have very strong positive biases, which influence the overall company bias, making it appear less biased. This becomes apparent when we consider the median bias coefficient, as illustrated in Fig. 5.
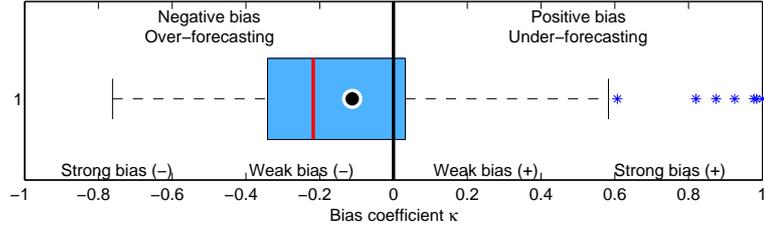

Figure 5: Bias coefficient boxplot. The mean bias coefficient is represented by a (●).

Both Figs. 4b and 5 are constructed to provide a clear view of the forecast bias behaviour at an overall company level and at a product level, disregarding the magnitude of the errors that is captured by the root error, so as to provide a comparison with the existing conventional bias metrics. However, the additional information contained in the complex errors can provide interesting insights. We propose a new plot to visualise the richness of information contained in this metric. Fig 6 illustrates the proposed 'Bias plot'. This plot is based on the polar coordinates of complex errors, using the error magnitude r and its angle γ that is expressed in radians. Forecasts errors can only result in complex numbers with $0 \leq \gamma \leq \pi/2$, hence the plot is bounded accordingly. The correspondence between the bias coefficient κ and angle γ is highlighted. This visualisation demonstrates that since the value of κ depends solely on γ, as in Eq. (8), it is correctly interpreted as having a linear scale in Figs 4b and 5. The vertical line represents the unbiased case, while any values on the left are negatively biased and any values on the right positively biased. The further a point is from the origin of the axes the larger is the magnitude of the error r.

The MRE of each product is plotted, as well as the MRE that represents the average overall company behaviour. Observe that since we are dealing with complex numbers, the company's behaviour can be geometrically constructed from the forecasters' behaviour for each individual product and vice-versa, as in Fig. 3. We can observe that the majority of products have a negative bias with relatively small but similar associated errors. A smaller
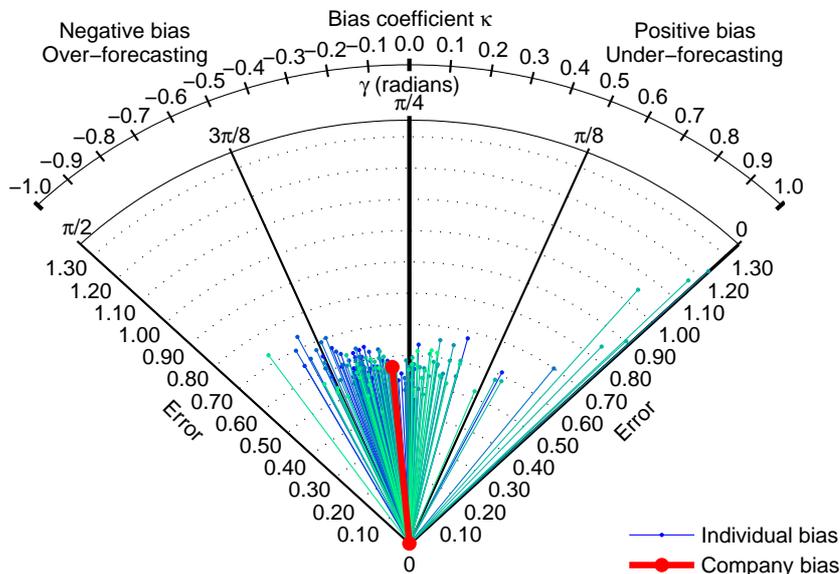
Figure 6: Forecast bias plot.

number of positively biased products can be seen, in agreement with the exploration in Fig. 5. With the bias plot it is evident that these highly biased products are also associated with higher than average errors. Therefore, this plot permits to easily characterise the forecasting behaviour of both individual products and overall, in terms of bias and errors. Note that following Eq. (10) the geometric mean of the squared magnitudes of the individual products will result in the GRMSE of the company, presented in Table 1.

## 4. Applications of MRE

4.1. Describing the impact of experts' bias on accuracy

In the literature a question that has been raised is whether we can describe the connection between bias of experts and forecasting accuracy, with the assumption that high biases in judgemental adjustments will lead to disproportionately high forecast errors (Trapero et al., 2011). It is worthwhile to explore whether complex errors can help us answer this question better. To address this question we need a measure of bias and a measure of accuracy. We have demonstrated so far that complex errors provide both, as

16

captured by their magnitude r and angle γ, while avoiding the pitfalls of conventional metrics. Fig. 7a provides a scatterplot of the bias direction and size and the error magnitude. The company bias is highlighted and a 2nd order polynomial fit is estimated on the data:

$$\hat{r} = 0.6773\gamma^2 - 1.1249\gamma + 1.0096. \tag{12}$$

The coefficient of determination of the polynomial is $R^2 = 0.66$. Observe that while the polynomial captures part of the apparent nonlinear relationship between bias and error, strongly biased observations are not modelled adequately. Higher order polynomials are not used since they do not fit better to the data, overfitting the weak bias region.
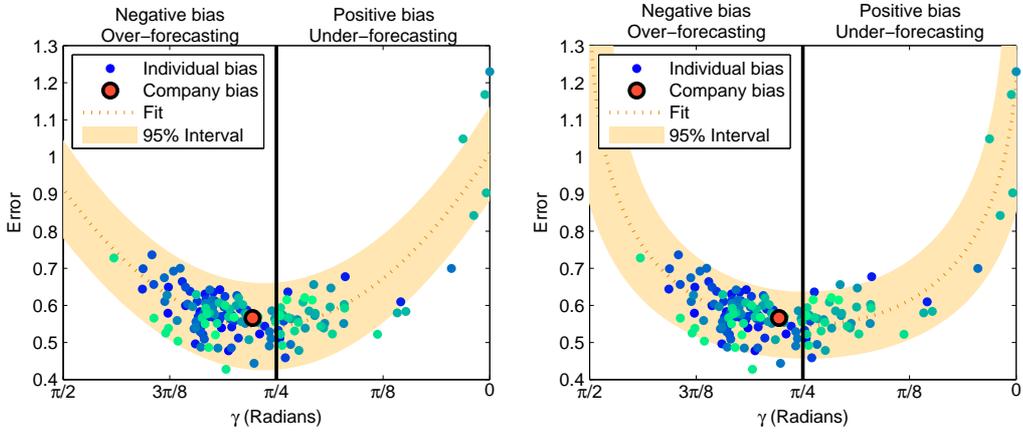
(a) 2nd order polynomial fit in angle-error coordinates.

(b) 2nd order polynomial fit in *bias plot* coordinates.

Figure 7: Forecast bias vs. error magnitude scatterplots with polynomial fits.

A modelling limitation of the polynomial in Eq. (12) is that we do not take advantage of the connection between the magnitude and angle of complex numbers. In Fig. 7b we fit a 2nd order polynomial on the real and imaginary part of the complex errors, which is subsequently projected on the scatterplot:

$$\hat{b} = 0.4379a^2 - 0.0155a + 0.5465. \tag{13}$$

This can be visualised by fitting the polynomial in the bias plot in Fig. 6. Expressed in terms of angle and magnitude Eq. (13) becomes:

17

$$\hat{r} = \sqrt{cos\,(\gamma)^2 + (0.4379cos(\gamma)^2 - 0.0155cos(\gamma) + 0.5465)^2}. \qquad (14)$$

The resulting fit is superior with $R^2 = 0.82$, an increase of about 23% in comparison to the fit of Eq. (12). It describes well both products with weak and strong forecasting bias and correctly models the extreme values. The polynomial fit, as seen in Eq. (13), does not need to resort to highly-nonlinear models that are hard to identify, estimate and extract inference from. Furthermore, the prediction interval of the polynomial can also be projected, providing additional insight in the forecast behaviour of the case company. We attempted to approximate the projected fit using angle-error coordinates, as in Fig 7a, and polynomials up to 15th order, which were all found to be inappropriate. Even if a fit were found to be adequate, the complexity of the model would prohibit any useful inference. Using complex errors the superior fit is feasible because we take advantage of the full information encoded in the root error, which is not possible with conventional metrics.

Note that the constant of both polynomials is positive, capturing a structural negative bias in the behaviour of the organisation of the case study. However, in the case of Eq. (12) part of the nonlinearity that is not captured by the polynomial is interpreted as bias, providing erroneous description of the forecasting behaviour. Finally, it is easy to express the connection between bias and error is terms of bias coefficient instead of γ, using the connection in Eq. (8).

Using the properties of proposed root error we are able to fit a low order polynomial that fits well to the observed data and describes clearly and accurately the forecasting behaviour of the case company. As was demonstrated above, such a result is not possible if we do not consider the additional properties of complex errors.

4.2. Limits of bias-accuracy improvements

The separation of bias and the error magnitude r that is done by the root error is helpful in exploring whether it is possible to correct the bias without affecting the magnitude of errors, i.e. the forecasting accuracy. An unbiased behaviour results when $\gamma = \pi/4$ or simply $a = b$. Let z be the resulting root error of a biased forecast that we want to adjust, which has magnitude r_z. Let u be an unbiased error. Equal magnitudes are necessary if forecasting accuracy is to stay the same, therefore the magnitude of u is set to be equal

to r_z. Let d be a forecast adjustment (positive or negative) that will result in the desired unbiased result u and $v = \sqrt{d}$. In terms of complex errors the adjustment is $u = z + v$. Solving for v we find:

$$v = u - z = r_z \left(cos\left(\frac{\pi}{4}\right) + isin\left(\frac{\pi}{4}\right) \right) - a - bi. \tag{15}$$

Separating the real and imaginary parts of the adjustment we get:

$$Re(v) = \sqrt{a^2 + b^2}cos\left(\frac{\pi}{4}\right) - a, \tag{16}$$

$$Im(v) = \sqrt{a^2 + b^2}sin\left(\frac{\pi}{4}\right) - b. \tag{17}$$

Given that $d \in \mathbb{R}$, since it is a forecast adjustment, only one of the real or imaginary parts of v can be non-zero. Therefore, if such an adjustment exists:

$$\begin{cases} cos\left(\frac{\pi}{4}\right) & \geq \frac{a}{\sqrt{a^2+b^2}} \\ sin\left(\frac{\pi}{4}\right) & \geq \frac{b}{\sqrt{a^2+b^2}} \\ ab & = 0 \end{cases} \tag{18}$$

These conditions cannot be met by any positive or negative adjustment of size a^2 or $-b^2$ respectively. This proves that it is not possible to correct a biased behaviour without accepting some additional forecast error. This is analogous to the bias-variance trade-off and provides an intuitive proof using complex errors.

5. Conclusions

In this paper we proposed a new error metric, which takes advantage of the properties of complex numbers, to help us better describe the forecasting behaviour of experts and organisations. We demonstrated its advantages over conventional metrics using a case study.

The new metric is capable of capturing the forecast bias, overcoming limitations of existing metrics, and the magnitude of the errors. Based on this metric we introduced the bias coefficient κ that is unit free and scale independent, allowing to easily compare the forecasting behaviour between different experts or organisations. Furthermore, because κ is bounded between [-1,

1] by construction, it allows us to characterise the measured bias, providing an non-relative description of the forecasting behaviour as unbiased, weakly or strongly biased. This is a property that existing bias metrics lack. The error magnitude associated with complex errors can be summarised using the geometric mean, resulting in the Geometric Squared Mean Root Error. Interestingly, the magnitude of this metric matches with the Geometric Root Mean Squared Error, which has been proposed in the literature as an accuracy metric with desirable properties (Fildes, 1992).

Error metrics, depending on the objective of their use, should have various qualities including robustness, interpretability and scale and unit independence (Armstrong and Collopy, 1992). The proposed root error, and the various metrics that we can construct from it, score favourably in these terms. The root error due to the squared root used in its calculation is robust to the effect of outliers, which can otherwise dominate error measure summaries for a single or across multiple time series. The bias coefficient, which is based on the root error, provides an intuitive and interpretable measure of bias. This is also unit and scale free, which permits summarising across multiple time series and benchmarking. The error magnitude as summarised by the GSMRE has similar properties. To these we add the connection between bias and accuracy that root error retains, while existing metrics do not, therefore more clearly representing the holistic behaviour of forecasters. This can also help us to understand how individual expert behaviours or choices in the forecasting of individual items result in the overall organisational forecasting behaviour.

Although the root error is based on complex number analysis, its calculation is relatively simple. We anticipate that in practice forecasters will not report the root error directly, but rather the bias coefficient and the GSMRE. We provided novel visualisations of the forecasting behaviour, capturing the bias and error magnitude, in order to enhance the communication of the results to experts, addressing a need indicated by Hämäläinen et al. (2013).

We demonstrated that the proposed metric provides estimation benefits when we try to model the effects of forecasting bias on accuracy. The additional information contained in the complex errors permits capturing complicated relationships without requiring highly-nonlinear models. This is desirable both for the identification and estimation of the models, as well as providing more transparent intelligence on the forecasting behaviour.

Further research should address the utilization of the complex error metrics for comparing forecasting approaches in different fields. Additionally,

judgemental forecast observations coming from other industries should also employed to consolidate the results presented here, in particular given the scale free properties of the bias coefficient that allows drawing cross-industry results and benchmarks.

References

Armstrong, J. S., Collopy, F., 1992. Error measures for generalizing about forecasting methods: Empirical comparisons. International journal of forecasting 8 (1), 69–80.

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

Bozos, K., Nikolopoulos, K., 2011. Forecasting the value effect of seasoned equity offering announcements. European Journal of Operational Research 214 (2), 418–427.

Davydenko, A., Fildes, R., 2013. Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. International Journal of Forecasting 29 (3), 510–522.

Eroglu, C., Croxton, K. L., 2010. Biases in judgmental adjustments of statistical forecasts: The role of individual differences. International Journal of Forecasting 26, 116–133.

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

Fildes, R., Goodwin, P., 2007. Against your better judgment? How organizations can improve their use of management judgment in forecasting. Interfaces 37, 570–576.

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.

Hämäläinen, R. P., Luoma, J., Saarinen, E., 2013. On the importance of behavioral operational research: The case of understanding and communicating about dynamic systems. European Journal of Operational Research 228 (3), 623–634.

Harvey, N., 2007. Use of heuristics: Insights from forecasting research. Thinking & Reasoning 13 (1), 5–24.

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. Springer-Verlag, Berlin.

Kahneman, D., 2011. Thinking, fast and slow. Macmillan.

Kourentzes, N., 2014. On intermittent demand model optimisation and selection. International Journal of Production Economics 156, 180–190.

Kremer, M., Moritz, B., Siemsen, E., 2011. Demand forecasting behavior: System neglect and change detection. Management Science 57 (10), 1827–1843.

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.

Mello, J., 2009. The impact of sales forecast game playing on supply chains. Foresight: The International Journal of Applied Forecasting 13, 13–22.

Moon, M. A., Mentzer, J. T., Smith, C. D., 2003. Conducting a sales forecasting audit. International Journal of Forecasting 19 (1), 5–25.

Ord, J. K., Fildes, R., 2012. Principles of Business Forecasting, 1st Edition. South-Western Cengage Learning, Mason, Ohio.

Petropoulos, F., Makridakis, S., Assimakopoulos, V., Nikolopoulos, K., 2014. 'Horses for Courses' in demand forecasting. European Journal of Operational Research 237 (1), 152–163.

Sanders, N. R., Manrodt, K. B., 2003. The efficacy of using judgmental versus quantitative forecasting methods in practice. Omega 31 (6), 511–522.

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

Trapero, J. R., Kourentzes, N., Fildes, R., 2012. Impact of information exchange on supplier forecasting performance. Omega 40 (6), 738–747.

Trapero, J. R., Pedregal, D. J., Fildes, R., Kourentzes, N., 2013. Analysis of judgmental adjustments in the presence of promotions. International Journal of Forecasting 29 (2), 234–243.