Published in "Journal of Advances in Management Research", 2021, vol. 18, no. 5, pp. 661-683, which should be cited to refer to this work.
DOI: 10.1108/JAMR-05-2020-0080

Evaluating demand forecasting models using multi-criteria decision-making approach

Yvonne Badulescu[a,b], Ari-Pekka Hameri[a] , Naoufel Cheikhrouhou[b]

[a] Faculty of Business and Economics, University of Lausanne, Switzerland

[b] Geneva School of Business Administration, University of Applied Sciences Western Switzerland (HES-SO), 1227 Carouge, Switzerland

emails:

corresponding author: yvonne.badulescu@unil.ch

ari-pekka.hameri@unil.ch

naoufel.cheikhrouhou@hesge.ch

Abstract. Evaluating appropriate error measures to determine demand forecast accuracy is essential in model selection, however there is no approach that simultaneously evaluates different model classes and several inter-dependent error measures. Furthermore, error measures may yield conflicting results making it more difficult to select the 'best' forecasting model when considering several error measures simultaneously. This paper proposes a novel process of evaluation of demand forecasting models using the analytical network process combined with the technique for order of preference by similarity to ideal solution (ANP-TOPSIS) which incorporates interdependence amongst error measures. The methodology is validated through an implementation case of a plastic bag manufacturer demonstrating that the use of the ANP-TOPSIS approach, avoided the selection of an inappropriate forecasting model due to conflicting error measurements. Moreover, a sensitivity analysis finds that the interdependence between the error measures is found to impact the relative closeness to the ideal solution, even though it plays a minimal role in the final ranking of the forecasting models.

Keywords: Decision making, Forecasting, Multicriteria, ANP.

1 Introduction

Many demand forecasting models have been developed to accurately determine the real future demand of products acting as the foundation for efficient supply chain planning. Given the amount of different forecasting quantitative, qualitative and hybrid forecasting models available, researchers and practitioners alike are confronted with the issue of selecting the most appropriate model. Moreover, selecting a demand forecasting model can be a challenging task if forecasters rely mainly on the results of error measures to support selection. Although using several error measures to evaluate forecasting models is recommended, they can oftentimes yield conflicting results; a forecasting model may have a better performance using one error measure and an alternative forecasting model shows better performance using another error measure. Furthermore, the Akaike Information Criteria and Schwartz's Bayesian Criterion approaches to selecting demand forecasting models are limited in their ability to compare forecasting models within different classes, for example, a time-series forecast against a causal model or a hybrid model combining both quantitative and qualitative methods.

Practitioners usually select a forecasting model based the performance of one or several error measures applied to all product families and categories. The use of an inappropriate error measure could lead to the selection of less effective forecasting model, highlighting the importance of selecting a number of forecast error measures to evaluate the models. Although the high similarity in the calculation of many error measures, often times, different error measures have yielded conflicting results when evaluating alternative forecasting models, making it more difficult for practitioners to select the best forecasting model for their purpose. This paper proposes a hybrid multi-criteria decision making (MCDM) approach, the Analytic Network Process coupled with the Technique for Order of Preference by Similarity to Ideal Solution (ANP-TOPSIS) to support the selection of the best forecasting model out of a set of alternatives based several error measures. In contrast to the Analytical Hierarchical Process (AHP) approach, the advantage of the ANP approach is that the interdependence amongst the evaluation criteria, the error measures, is included in the calculation of criteria weights. In addition, ANP-TOPSIS is a straight forward approach that can be adopted in industry.

The methodology is demonstrated on an implementation case of a plastic bag manufacturing firm for the selection of the best forecasting model out of five alternatives of different classes: two autoregressive time-series models, one exponential smoothing model and two hybrid models that include judgmental adjustment of an ARIMA model. The final ranking is tested for robustness via a sensitivity analysis of the evaluation criteria weights (error measures) as well as by omitting the interdependence between the error measures. The results show that despite conflicting error measurements, the ANP-TOPSIS approach presented in the paper enables the decision maker to select the best forecasting model for their particular case.

This paper is divided into 6 sections: the next section provides an overview of forecasting model selection. Section 3 outlines the proposed ANP-TOPSIS approach that supports forecasting model selection using several error measures. The methodology is validated through an implementation case of a plastic bag manufacturer in Section 4. Section 5 provides a critical discussion of the results, followed by the conclusion in Section 6.

2 State of the Art

2.1 Background

Exponential smoothing models and autoregressive models are the most commonly used time-series forecasting in demand planning (Petropoulos et al., 2018). These models have high applicability to time-series data and ease of understanding in business contexts (Alvarado-Valencia et al., 2017).

In addition to purely statistical models, judgmental forecasting is frequently employed as a stand-alone, such as for new product market introductions, or is incorporated into the statistical forecast to create a hybrid forecast which includes both quantitative and qualitative information, through judgmental adjustments, bootstrapping, combination, and decomposition (Arvan et al., 2019). Marmier and Cheikhrouhou (2010) develop a hybrid forecast based on a systematic approach that structures and integrates judgment into demand forecasting using event-based factors and extend the model to adjust

the forecast based on collaborative human judgment improving the forecasting model MAPE and MAE (Cheikhrouhou et al., 2011). Van den Broeke et al. (2018) use several error measures to evaluate hybrid forecasting models and find that accuracy can either improve, remain the same or decrease depending on the case and time horizon. Therefore, hybrid forecasting models should be carefully considered together with statistical forecasting models when evaluating alternative models.

Forecast accuracy is the most important criterion to determine the performance of a forecast (Ha et al., 2018) and the choice of the error measure is highly important in evaluating the forecasting models (Davydenko & Fildes, 2013). Although error measures provide quantitative support in the performance evaluation of a forecasting model, the selection of a forecasting model cannot be solely selected based on any one specific error measure due to the risk of ignoring an important aspect of the forecast that contributes to its accuracy. Scale dependent methods, such as the mean error (ME), mean squared error (MSE), root mean squared error (RMSE), and the MAE can indicate biasness of the forecast as well as how spread out the forecasted values are from the actuals. In the case of ME, it is possible that a negative error on one data point would counterbalance a positive error on another data point. The median could be used instead of the mean to counter this effect (Franses, 2011) or by using the absolute values of the error. On the other hand, the MAE may skew the mean when confronted with large outliers (Davydenko & Fildes, 2016). RMSE is generally preferred to MSE as it is in the same scale as the data, even though it is more sensitive to outliers than MAE (Hyndman & Koehler, 2006). The RMSE gives extra weight to large errors due to the squaring function and is sensitive to scale, and the results frequently differ when applied to various sets of data. Hassani et al. (2015) use the RMSE to evaluate the performance of 17 univariate and multivariate forecasting models to determine the price of gold. The results of the RMSE showed that no model outperformed another in the short run and the long run. However, the exponential smoothing model had the lowest average RMSE over the full forecasting horizon of 24 months. Using more than one error measures can also yield conflicting accuracy results. Fildes et al. (2011) evaluate the accuracy of several airline traffic forecasting models using four error measures

which result in conflicting results, particularly between RMSE and Geometric Mean Relative Absolute Error (GReIAE). They conclude that several error measures must be considered when comparatively comparing different forecasting models. Similarly, in determining the effectiveness of judgmental forecast adjustments, several papers used multiple error measures to determine the accuracy of their models and arrived and opposing conclusions (Fildes et al., 2009; Franses & Legerstee, 2010; Trapero et al., 2012). Conflicting conclusions resulting from using more than one error measure can be confusing and damaging to decision-makers (Davydenko & Fildes, 2013) and none of the abovementioned articles propose an approach in selecting a forecasting model with conflicting error measurements.

Accuracy measures based on percentage errors, such as mean percentage error (MPE), root mean squared percentage error (RMSPE) and MAPE, allow for comparison across different data sets as they are scale independent. The MPE is a good measure of the relative size and direction of the bias and the RMSPE takes only positive values due to the squaring function and therefore provides an average relative size of the error. On the other hand, both are very sensitive to large outliers (Davydenko & Fildes, 2016). The MAPE is one of the most commonly used and highly recommended error measures in demand forecasting. However, the MAPE is often criticised for being asymmetrical, meaning it is pulled upward due to the heavier penalty of positive errors. In addition, the MAPE assumes a meaningful zero which is not considered an issue in demand forecasting when referring to quantity, except potentially in the case of returns. The MAPE is therefore best used when dealing with positive actual observations (Ren & Glasure, 2009).

Furthermore, the R^2 should not be used independently without considering its result along with other error measures as R^2 ignores the forecast bias and the results may show a perfect result of 1 even when the forecasted values are very different from the actuals (Armstrong, 2001). Table 1 provides a summary of the error measures referred to in the literature and the result when considered independently. The error (e_i) is the difference between the forecast (F_i) and the actual observation (O_i) for point i in the time series and N is the number of data so that, $e_i = F_i - O_i$. The error measures in Table 1 are comparatively

simple accuracy measures as they are intuitive and relatively easy to calculate and are most likely to be used in industry. Many more sophisticated error measures have been developed in the literature but are omitted in the context of this article due to their specificity and likelihood of being widely used in industry.

Table 1: Summary of the literature on error measures in their application to forecasting model selection

Error measure	Equation	Advantages	Constraints	References
Mean Error (ME)	$ME = \frac{\sum e_i}{N}$	• Indicates bias, range & variance	• Positive and negative values negate each other Scale dependent	Franses, (2011)
Mean Squared Error (MSE)	$MSE = \frac{\sum e_i^2}{N}$	• Indicates bias, range & variance	• Scale dependent	
Root Mean Squared Error (RMSE)	$RMSE = \sqrt{MSE}$	• Indicates bias, range & variance • Generally preferred to MSE as same scale as data	• Sensitive to outliers Sensitive to Scale Adds weight to large outliers	Hyndman and Koehler, (2006) Davydenko & Fildes, (2016)
Mean Absolute Error (MAE)	$MAE = \frac{\sum \lvert e_i \rvert}{N}$	• Indicates bias, range & variance	• Outliers create skewed mean Scale dependent	Davydenko & Fildes, (2016)
Mean Percentage Error (MPE)	$MPE = \frac{\sum \frac{e_i}{O_i}}{N}$	• Accuracy based on percentage • Allow comparison across data sets • Scale independent • Relative size and direction of bias	• Sensitive to large outliers	Armstrong and Davydenko & Fildes, (2016)
Root Mean Squared Percentage Error (RMSPE)	$RMSPE = \sqrt{\frac{\sum \left(\frac{e_i}{O_i}\right)^2}{N}}$	• Accuracy based on percentage • Allow comparison across data sets • Scale independent	• Does not consider negative values due to squaring function, therefore provides relative error • Sensitive to large outliers	Armstrong and Davydenko & Fildes, (2016)
Mean Absolute Percentage Error (MAPE)	$MAPE = \frac{\sum \left\lvert \frac{e_i}{O_i} \right\rvert}{N}$	• Accuracy based on percentage • Allow comparison across data sets • Scale independent	• Asymmetrical due to heavier penalties on positive error	Ren and Glasure, (2009)
Coefficient of determination (R^2)	$R^2 = 1 - \frac{\sum (O_i - F_i)^2}{\sum (O_i - \bar{O})^2}$	• Coefficient of determination • Measures proportion of variability in the linear function	• Ignores forecast bias	Armstrong, (2001)

2.2 Forecasting model selection

As there are a large number of feasible forecasting models to predict demand, demand planners in industry often find it difficult to know which is the best model for their demand forecast, particularly in the case of conflicting error measurements. The Akaike Information Criterion (AIC) is one of the known methods used in automatic model selection, introduced by Akaike (1974). Although it is very effective in selecting models within the same model class and even comparing non-nested models (e.g., linear vs non-linear models), this method cannot be used for the automatic model selection of models from different forecasting model classes, such as between exponential smoothing and autoregressive models. The same is true with the Schwartz's Bayesian Criterion (SBC), which, similar to AIC, also considers the goodness of fit of the data with a complexity penalty. Although the SBC penalises the complexity more than AIC, the approach is still limited to assessing models within the same class.

Villegas et al. (2018) propose using support vector machines (SVM) to select the best forecasting model, out of a pool of alternatives at each moment in time, as the model variables change such as its relative accuracy performance and the fitted parameters of the model. They find that using SVM results in a higher overall forecast accuracy. Ghobbar and Friend (2003) develop a predictive error forecasting method to evaluate demand forecasting models in the airline manufacturing industry based on their factor levels. They use MAPE as the dependent evaluation criteria. Neither of them consider hybrid forecasts, which integrate judgmental information. Oh and Morzuch (2005) evaluate eight competing demand forecasts for tourism in Singapore based on six performance measures measuring bias and forecast error including MAPE, MAE, RMSE, AIC and SBC. Their findings show that using different performance measures leads to the selection of different forecasting models. Taylor and McSharry (2007) evaluate six forecasting models to determine the electricity demand based for ten European countries using the MAPE and MAE which, when ranked, yielded conflicting results, except for the highest ranked model which was consistently first. Petropoulos et al. (2018) and W. Han et al. (2019) explore forecasting model selection based only on subjective expert judgment and find that the selected models perform better based on the error measures used (MAE, MAPE, MASE and MPE) than the forecasting models selected using AIC, and find that collaborative judgment is better than single judgment and statistical selection. Davydenko & Fildes (2013) investigate the use of the MAPE and

Median average percentage error (MdAPE) on evaluating judgmental adjustment of statistical forecasts and conclude that due to conflicting results of the MAPE and several other error measures, forecasters in companies cannot just rely on the MAPE to determine the model's performance. However, Davydenko & Fildes (2013) and the aforementioned literature do not propose an approach to select the most appropriate model when using multiple error measures for evaluation and particularly in the case of conflicting error measurements.

Xu and Ouenniche (2012) evaluate forecasting models for crude oil prices based on trade-offs between several error measures using elimination and choice expressing reality (ELECTRE) I & II or preference ranking organisation method for enrichment evaluation (PROMETHEE) and find that the best performing models are selected using both methods. Velasquez & Hester (2013) find that although PROMETHEE is easier to use than ELECTRE, it lacks clarity in weight determination for the criteria and ignores potential interdependencies between them. Mehdiyev et al. (2016) propose the use of the PROMETHEE approach to evaluate the relative performance of several multiclass classification algorithms: Bayesian Networks, Artificial Neural Networks, SVMs, Logistic Regression and decision trees, based on several error measures. However, neither Mehdiyev et al. (2016) or Xu and Ouenniche (2012) consider the evaluation of multiclass demand forecasting models based on interdependent error measures.

ANP applies pairwise comparisons to compare alternatives as well as estimate weighting to the criteria and priority scales and are relatively straightforward to use (Saaty, 2001; Velasquez & Hester, 2013). On the other hand, the ANP method is susceptible to rank reversal at the end of the process, which could result in the final ranking being reversed in order. Using TOPSIS addresses the issue of rank reversal when a non-optimal alternative is introduced and allows for easier ranking between alternatives (Sipahi & Timor, 2010). However, the TOPSIS method does not consider criteria interrelationships, nor does it provide an easy method for determining criteria weights, and is often paired with the ANP method (Tao et al., 2012).

The hybrid MCDM approach of ANP-TOPSIS is robust, based on the constant number of calculation steps for TOPSIS regardless of the number of attributes. It is also scalable and takes into consideration the inherent interdependence between the evaluation criteria (Velasquez & Hester, 2013). Lastly, it is also able to analyse both quantitative and qualitative information (Kara & Cheikhrouhou, 2014) and has the ability to consider the interdependence between evaluation criteria (error measures) of which none of the aforementioned MCDM approaches do.

Table 2 includes some of the different approaches in the literature for selecting forecasting models based on various evaluation and information criteria. The methods using information criteria such as those proposed by Akaike (1974), Petropoulos et al. (2018) and Villegas et al. (2018) only allow a comparison of forecasting models within the same model class as previously mentioned signifying that hybrid forecasting models cannot be compared to a purely statistical model using these approaches. MCDM approaches have been applied to evaluate forecasting models relative to each other in Mehdiyev et al. (2016) and Xu and Ouenniche (2012) and show promising results when comparing time-series models with regression models and classification algorithms such as decision trees. However, neither have considered hybrid forecasting models which combine quantitative and qualitative information to develop a forecast, nor have they considered the interdependence between error measures.

The state of the art presents several approaches to evaluating different forecasting models in diverse applications primarily using error measures and information criteria, like AIC and SBC, as the evaluation criteria. The use of AIC and SBC for model evaluation limits the comparison with models within the same model class, such as exponential smoothing time-series models. In addition, selection of the best forecasting model may be improved considering several error measures simultaneously (Xu & Ouenniche, 2012). There has not been any research yet proposing an approach on how to evaluate multiclass demand forecasting models based on several error measures that have interdependencies. This paper proposes an MCDM approach, ANP-TOPSIS, to support the selection of multiclass demand forecasting models with conflicting performances.

Table 2: Overview of the methods for forecasting model evaluation and selection

Reference	Evaluation and selection method	Forecasting models: Time-series	Forecasting models: Regression	Forecasting models: Hybrid	Evaluation criteria
Akaike (1974)	AIC	✓			• Information criteria
Ghobbar and Friend (2003)	PEFM	✓	✓		• MAPE
Oh and Morzuch (2005)	Comparison of: Error Measures, Information Criteria, Biasness	✓	✓		• 3 Error Measures: MAE, MAPE, RMSE • AIC & SBC • Biasness
Taylor and McSharry (2007)	Comparison of: Error Measures	✓	✓		• 2 Error Measures: MAPE & MAE
Xu and Ouenniche (2012)	ELECTRE I ELECTRE II PROMETHEE	✓	✓		• 13 Error Measures
Mehdiyev et al. (2016)	PROMETHEE		✓		• 11 Error measures • Classification performance measures
Villegas et al. (2018)	SVM	✓			• Information criteria (AIC & SBC) • Estimation information (autocorrelation) • Statistical tests (P-values)
Petropoulos et al. (2018)	Judgmental selection	✓		✓	• MAE for evaluation • MAPE, MPE and Mean Absolute Scaled Error (MASE) to measure performance

3 Methodology and Calculation

The literature review highlights the lack of an approach that evaluates demand forecasting models of different classes based on several error measures and their interdependencies as evaluation criteria. An MCDM approach acts as an effective tool to aid in the decision-making process when several evaluation

criteria must be considered. In addition, it allows for a comparison between models of different classes as each alternative model is represented by its performance in the evaluation criteria.

We propose an ANP-TOPSIS approach which is able to 1) structure the problem into a network of evaluation criteria and competing forecasting models, 2) consider multiple evaluation criteria and the influence they have on one another (interdependence), 3) provide a trade-off between evaluation criteria to compensate for poor results (compensatory method) and 4) compare forecasting models of different classes (quantitative and hybrid models) via the normalisation of the evaluation criteria results.

Figure 1 illustrates the seven main steps in the ANP-TOPSIS process. Steps 1-4 represent the ANP part and steps 5-7 represent the TOPSIS part of the hybrid MCDM approach.

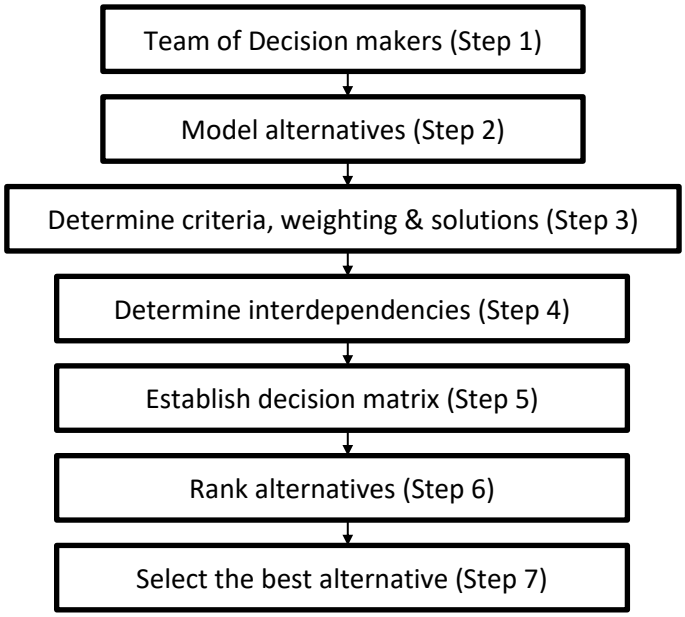

Figure 1: Proposed ANP-TOPSIS framework for forecasting model selection

The first step is to gather a team of experts that will use their knowledge and experience to: select the alternative forecasting models in step 2; decide upon the evaluation criteria; and allocate priorities to the individual criteria using pairwise comparisons in step 3. This team also determines the interdependencies between criteria and reprioritises the criteria based on their dependence on each other in step 4. The

decision matrix is calculated in step 5 which is the basis for determining the relative closeness to the ideal solution that provides the final ranking in steps 6 and 7. Each step is explained in detail in the following sections.

The methodology is demonstrated in an empirical implementation case based on data and additional information collected from a plastic bag manufacturer in the south of Spain taken from (Cheikhrouhou et al. 2011). The implementation case is based on three years of daily sales data of plastic bags from a manufacturer in Spain. The company has four major supermarkets as customers that make up the majority of their sales. The daily sales data are aggregated to monthly buckets for analysis and forecasting purposes. The time-series used are composed of the aggregate monthly demand collected over a period of 36 consecutive months (mth_1 to mth_{36}) from January to December representing 3 years (Y_1 to Y_3) seen in Figure 2. The objective is to plan the demand for the fourth consecutive year (Y_4) from January to December and select the forecast which best fits the actual results based on a number of evaluation criteria.

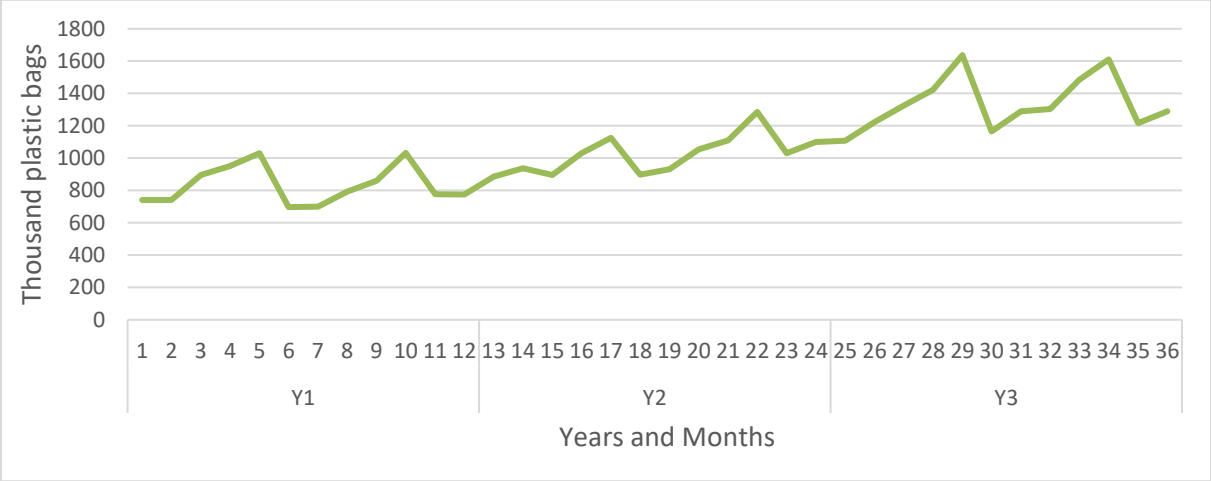

Figure 2: Historical data of polyethylene bag sales over three years

The time-series for the historical sales of the polyethylene bag (Figure 2) shows two seasonal peaks per year just before summer in the 5th month of each year (May), and the 10th month of each year (October)

just before the winter period. The seasonal pattern of 12 months is confirmed via the autocorrelation function applied to the 36 months of data.

Step 1: Preparation and formation of group of decision makers and data treatment

The first step is to compose a team of decision makers (DMs) of experts in the field that can define and assess the evaluation criteria. The DMs should have common interests related to the objectives of the analysis and aim to obtain a consensus in the decision making process (Al-Harbi, 2001). This team of experts will also be involved in the evaluation and approval of the final ranking results and therefore should be composed of those with the knowledge and capability of making decisions related to demand forecasting within the organization. Two decision makers (DM) are responsible for the prioritisation and weighting of the evaluation criteria through pairwise comparisons for the plastic bag manufacturer. The decision makers include two of the authors of this paper who performed pairwise comparisons of the seven error measures to determine attributed weightings. Both DMs have an academic and professional experience in forecast performance evaluation.

Step 2: Selection of forecasting model alternatives

The DMs agree upon alternative forecasting models that will be compared against one another other. It is not recommended to select too many alternatives due to the difficulty in conducting pairwise comparisons with a high number of alternatives, and criteria. A commonly used rule of thumb is to choose between 5-9 alternatives in order to optimize judgmental decision making.

The historical sales data of the polyethylene bag is plotted as a time-series in Figure 2 and shows a strong linear trend and seasonality with peaks in summer and winter due to demand increases in plastic bags during the months before summer and the Christmas holidays. Five forecasting models are calculated based on the historical sales data and judgmental forecast adjustments: ARIMA, SARIMA, Holt-Winter, ARIMA with single judgmental adjustment and ARIMA with collaborative judgmental adjustment. These models are considered most appropriate since the moving average and simple exponential

smoothing implies that neither seasonal variation nor tendency exists. In addition, the data-series cannot be assimilated to a random process, such as random walk, because of the observed seasonality and trend. The performance of each forecast is calculated using seven error measures.

Step 3: Determine the criteria on which to evaluate alternatives & calculate solutions per alternative

The DMs select between 5 and 9 criteria on which to evaluate the forecast alternatives. As the result of using different error measures may result in diverging measures of accuracy, several different error measures are selected as the criteria for the implementation case, which are considered important to the evaluation of the forecasts, but which at times may provide conflicting information and must be considered simultaneously. These are ME, MAE, MPE, RMSPE, RMSE, MAPE and R^2 as can be seen in Table 1.

Several key aspects of these criteria need to be taken into account. First of all, the criteria are for most cases incomparable, in the sense that they are not in the same units (for example, RMSPE and MAPE are in percentages, and R^2 is a ratio and therefore has no unit). An R^2 measure of 1.0 can be understood in that the regression line perfectly fits the data. Negative values of R^2 may occur when fitting non-linear trends to data. In seasonal time-series with non-linear trends, we can expect to see negative values of R^2.

Secondly, a common point amongst the criteria is that their values should be as close to zero as possible for smallest error, except for R^2 which should be closest to 1. In this paper, RMSPE and RMSE were chosen instead of MSE and MSPE as they use the same units as the original data (as opposed to squared units). R^2 is a different type altogether as it does not represent a mean computed using errors, but a measure of how well future outcomes are likely to be predicted by the model. The solutions of each criterion are calculated per alternative forecasting model creating the solution matrix F, where f_{mn} is the solution of criterion n based on model m.

$$F = \begin{pmatrix} f_{11} & \cdots & f_{1n} \\ \vdots & \ddots & \vdots \\ f_{m1} & \cdots & f_{mn} \end{pmatrix}$$

The models are determined based on a product's historical data and other relative qualitative information. The historical data should be cleaned of any outliers and the qualitative information integrated in a structured manner.

The Holt-Winter forecasting model is plotted against the actual demand in Figure 3 as the continuous grey line. The y axis on Figure 3 is magnified to show values between 1000 and 1700 thousand bags to better visualise the variations between the actual results (bars) and the forecasting models (lines). The Holt-Winter forecasted line follows the demand seasonality well although with a lower amplitude than the actuals. The Holt-Winters method provides a forecast with few outliers shown by the close values of the MAE and RMSE, 75.9 and 92.5 respectively, in Table 4. Figure 3 illustrates that this forecast has values that are relatively close (both above and below) to the actual results. Table 4 presents the solutions of the error measures for each of the alternative forecasts. The error measures in Table 4 quantify these discrepancies and show that the ME and MPE (both non-absolute measures) are relatively low at 36.9 (from a demand of >1200) and 3.27%, respectively. In addition, the R^2 measure of 0.61 indicates that the fit to the regression is not too low for a time-series presenting strong seasonality.

The second forecasting method analysed is the ARIMA (5,0,4) method. Figure 3 shows that the ARIMA method (dotted line) follows the actual demand relatively well, including the seasonality peaks (albeit with a slight delay on the second peak). The fit is better at the beginning of the time-series than at the end where there is a more noticeable positive error. The visual analysis is supported by the error measures in Table 4, where the ME and MPE are very low (1.33 and 0.51% respectively) suggesting an equal distribution of positive and negative errors. The RMSE is marginally larger than the MAE, which can explain the larger positive error in the later months. The R^2 measure of 0.24 is very low due to the high variance of error along the time-series.

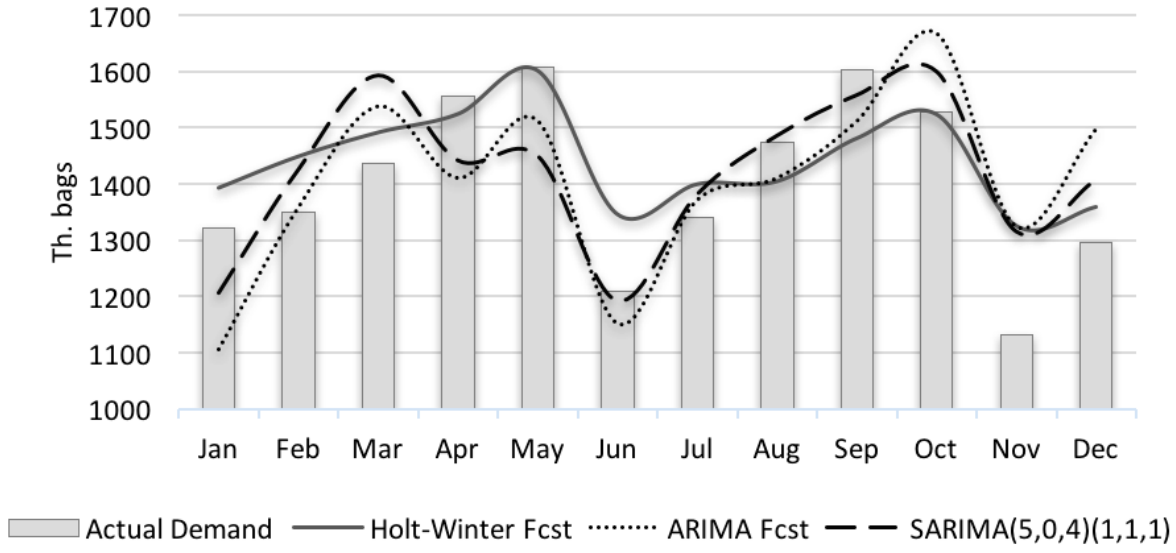

Figure 3. Actual sales values (demand) versus Holt-Winter, ARIMA and SARIMA forecasts

The SARIMA forecast, which introduces seasonality to the ARIMA (5,0,4) model, is used as the basis. The three parameters representing the orders of the seasonal autoregressive and moving average parts of the model are determined by simulating various configurations, using the R programming language and free software, to yield the lowest ME and RMSE, which result in (1,1,1). The results for the SARIMA (5,0,4)(1,1,1) are illustrated in Figure 3 showing strong adherence to the real data and their seasonality.

As the time-series in this implementation case explicitly show a seasonal character, it was expected that SARIMA would provide a better forecast than ARIMA. The forecast initially has some difficulty following the curve at the beginning of the 12-month period, followed by a very good fit in the later months (June to October) and then another deviation in the final two months of the forecast. The analysis of the SARIMA forecast error measures in Table 3 shows a relatively low ME and MPE of 17.50 and 1.70%, respectively, showing a near-equal positive and negative error distribution. In addition, the closeness in the values of MAE and RMSE indicates the absence of outliers.

Table 3. Evaluation criteria solutions per forecasting model

		Models				
		Holt-Winter	ARIMA	SARIMA	Single Adjusted	Collaborative Adjusted
Criteria	ME	36.9	1.3	17.5	-52.4	5.6
	MAE	76	112	92	97	33
	RMSE	93	130	107	172	43
	MPE	3.3 %	0.5 %	1.7 %	-3.3 %	0.5 %
	RMSPE	7.3 %	9.7 %	8.0 %	11.1 %	3.0 %
	MAPE	5.8 %	8.2 %	6.7 %	6.7 %	2.4 %
	R^2	0.61	0.24	0.48	-0.34	0.92

The two judgmentally adjusted forecasts utilize both time-series and qualitative information to determine the forecast. The first uses ARIMA(5,0,4) single judgment adjustment in which the expert judgment is integrated in a structured manner as complementary information to the ARIMA(5,0,4) forecast. The forecasting expert identifies and classifies the events into four adjustment factors: transient, transferring, jump or change trend factors, and then attributed a weight to each.

The first event in Y_4 is a special discount offer that will run in January, implying that there will be an advancement to January of the expected sales in February and March, thereby increasing the volume in January and decreasing volume in February and March. This is identified as a transferring factor. The second event relates to technical design changes to the plastic bag so that it meets environmental ISO requirements which are estimated to increase the expected demand over several months via new customers. This increase is identified as a jump adjustment factor and is applied from February through August. The third event in $Y_{4,}$ which is expected to affect the forecast, is the closure of the facilities of one of the four major customers in September. The expectation is that the sales will decrease only in that month. This is identified as a transient factor and is represented by the dashed line in Figure 4 as the large dip in the single judgmentally adjusted forecasting model in September. The final event in Y_4 impacts the full 12 months as it relates directly to the cost of the raw material - plastic - which increases

the sales price of the plastic bag and is expected to negatively impact the quantity sold. This event is identified as a trend change and results in a monthly decrease in the forecast.

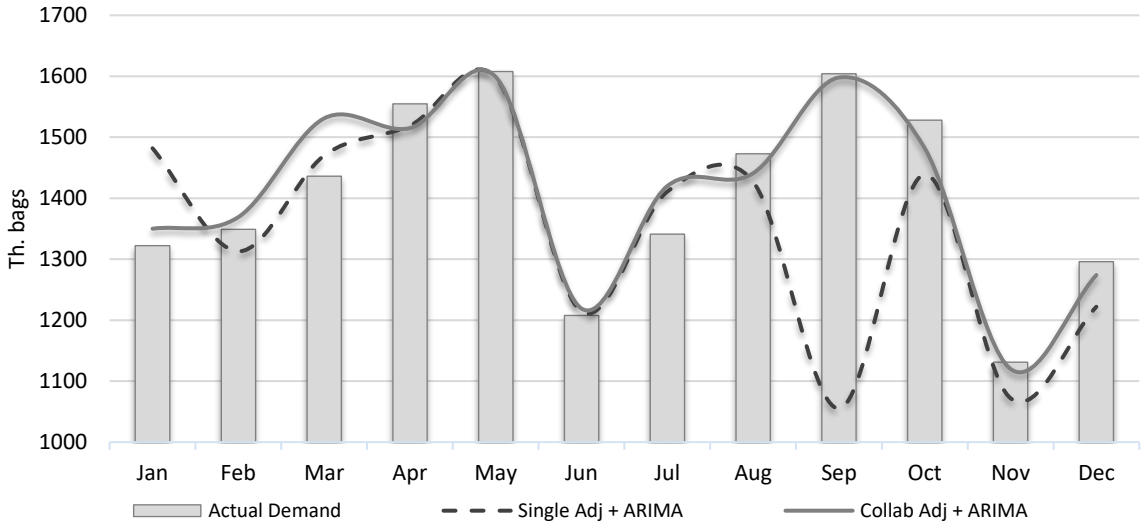

Figure 4: Actual sales versus hybrid forecasting models

Figure 4 shows that this forecast (dashed line) has a very good fit with the actual data, including month-over-month trend, seasonality, and peaks except for the under-forecasted value in September (-549). In addition, there are very large differences observed between the MAE and RMSE of 97.42 and 171.94, respectively (Table 4), which shows the strong impact of the September value. However, as the remaining fit is very good, other error measures mitigate the outlier's impact. The R^2 measure is negative which occurs when the fit of the forecasted line is worse than just fitting a horizontal line. This is due to the extreme September value and thus the non-linear error trend.

The second judgmentally adjusted forecast utilizes the collaborative expert judgment of several people and is based on ARIMA(5,0,4). The method for the collaborative judgmentally adjusted forecast builds upon the single judgmentally adjusted forecast presented previously by using the expert opinion of three experts instead of only one. The process is comprised of four parts: first, mathematical forecasts are created based on cleaned data (done in section 4.2), then factors are identified and classified using the forecasters' knowledge related to future events. Third, the different information collected is integrated

into a fuzzy inference engine to obtain a global fuzzy judgment. One of the major differences with the single judgmentally adjusted forecasting model is the adjustment for September due to the closure of one of the company's major customers. The single expert believed that the volume would decrease by the customer's monthly purchase quantity. However according to the team of three experts, that customer will order almost double their monthly volume to replenish their stock. Another observed difference between the two judgmentally adjusted forecasting models is in January based on the special discount offer that is expected to advance expected sales from February and March to January. This was also the case in the collaborative judgmentally adjusted forecast. However, the volume advanced from February and March to January is 34% less than in the case of the single judgmentally adjusted forecast. Both the ME and MPE are small showing a quasi-equal distribution of positive and negative errors in Table 4, and the MAE and RMSE are also very low due to the impact of the two small peak errors in March and July. The R^2 measure of 0.92 shows that there is an extremely good fit between forecast and data; there is very little difference in the error trend along the time-series.

Step 4: Determine interdependencies between criteria, prioritisation through pairwise comparisons, weighting and normalisation

The criteria are evaluated by calculating their weights in terms of priority, initially disregarding the interdependencies between the criteria. Accordingly, a pairwise comparison matrix is formed (A) using an "expert" judgment, on the basis of Saaty's Fundamental 1-9 scale which classifies the relative importance of one criterion over another (Saaty, 1990). The normalised weight vector (w) is then obtained by determining the maximum eigenvalue λ_{max} of the comparison matrix (A) and finding the solution to the equation below.

$$Aw = \lambda_{max} w$$

In the implementation case, the DMs compare the interdependent criteria pairwise to firstly prioritise the criteria based on the amount of useful information conveyed (for example, for ME where positive

errors can counteract negative errors, priorities in comparison to other error measures are very low). RMSE has a little higher overall priority to MAE due to its use in identifying the presence of outliers (square power). The team then compares them to determine by how much more one criterion is more important than another using Saaty's 1-9 scale (Saaty, 1990) shown in Table 4. The weights and importance corresponding to each criterion depend on the chosen MCDM method. At the end of pairwise comparisons, we obtain a normalised weight vector *w* by calculating the eigenvectors of the priority matrix.

Table 4. Criteria priorities and weights

Criteria	ME	MAE	RMSE	MPE	RMSPE	MAPE	R^2	**Weight (*w*)**
ME	1	1/5	1/4	1/3	1/5	1/5	1	**0.046**
MAE	5	1	1/3	1	1/3	1/3	1	**0.098**
RMSE	4	3	1	3	1	1/3	1	**0.167**
MPE	3	1	0	1	1/3	1/5	1	**0.080**
RMSPE	5	3	1	3	1	1/3	1	**0.173**
MAPE	5	3	3	5	3	1	1	**0.308**
R^2	1	1	1	1	1	1	1	**0.129**

The consistency ratio is calculated to determine the consistency in the expert judgments by comparing the consistency index (CI) to the Random Consistency Index (RI) and should be less than 10% (Saaty, 2001):

$$CI = \frac{\lambda_{max} - n}{n-1} \quad CR = \frac{CI}{RI}$$

The CR for this case is 0.0935, and is therefore considered acceptable as it is < 0.1 (Saaty, 2001).

CI generally increases as the number of alternatives increases (S. Han, 2016) and a method was developed in Ergu et al. (2011) identifying the inconsistent elements in a pairwise comparison matrix with the objective of improving the CR.

In order to determine the impact of the criteria on each other, the network of influences among the criteria is established through the pairwise comparisons of each criterion from the perspective of the control criterion. To illustrate how to integrate the interdependence into the criteria weightings, let us

consider *n* criteria named B_1 to B_n. Firstly, the dependence among each criterion with respect to itself must be determined. Does B_1 depend on B_2? Vice-versa, does B_2 depend on B_1? Continuing all the way to B_n. If, for example, there are *x* numbers of dependencies, the pairwise comparison matrices are determined for the impacted criteria with respect to the control criteria, *A*, and the normalized eigenvectors are calculated for these and placed in the columns of a supermatrix for network (*B*). The criteria which are not dependent on any others use the values of the pairwise comparison matrix *A*.

$$B = \begin{pmatrix} B_{11} & \cdots & B_{1n} \\ \vdots & \ddots & \vdots \\ B_{n1} & \cdots & B_{nn} \end{pmatrix}$$

B is subsequently multiplied by the weight vector *w* to determine the normalised weight vector, ω_{ANP}, of the criteria including interdependence.

$$\omega_{ANP} = \begin{pmatrix} B_{11} & \cdots & B_{1n} \\ \vdots & \ddots & \vdots \\ B_{n1} & \cdots & B_{nn} \end{pmatrix} \times \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} \omega_1 \\ \vdots \\ \omega_n \end{pmatrix}$$

ANP allows for a structured methodology of criteria identification, prioritisation and weighting based on expert judgment. The ranking of the alternative solutions uses the second part of the hybrid MCDM model, TOPSIS.

In the implementation case, all the error measures except R^2 are dependent on each other simply because of their calculation method. Their interdependencies are illustrated in Figure 5.

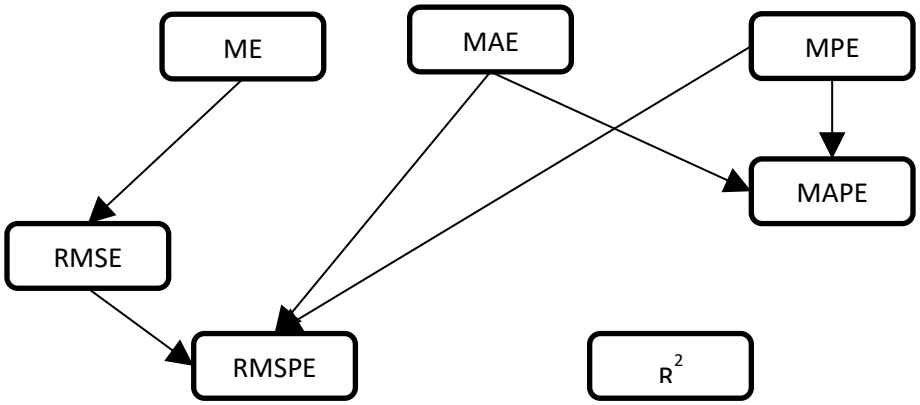

Figure 5: Interdependencies between MCDM criteria

Both ME and MAE have an impact on RMSE, as squaring the error removes any negative, and the RMSE is in the same units as ME and data. Due to the way the error measures are calculated, the difference between MAE and RMSE will highlight the presence of outliers. MAPE is dependent both on MAE and MPE equally as it is calculated using both the absolute value of errors and the percentage of total demand. Additionally, MPE and RMSE both influence RMSPE as it is comprised of the squared error aspect as well as showing a unit-less percentage of total demand. One thing to note is that MAE for example influences the RMSPE *indirectly* through RMSE.

Local priorities result from the eigenvectors of the priority matrix in Table 4. The supermatrix of the network is calculated in Table 5 with the normalised eigenvectors calculated for each criterion in the columns. The final column is the normalised weight vector ω_{ANP} determined by multiplying the supermatrix with the priority weight vector *w*.

Table 5. Normalised interdependences and weights (ANP-TOPSIS)

Criteria	ME	MAE	RMSE	MPE	RMSPE	MAPE	R^2	Weight (ω_{ANP})
ME	1	0	0.106	0	0	0	0	**0.064**
MAE	0	1	0.100	0	0	0	0	**0.184**
RMSE	0	0	0.745	0	0.138	0	0	**0.148**
MPE	0	0	0	1	0.172	0.200	0	**0.171**
RMSPE	0	0	0	0	0.690	0	0	**0.119**
MAPE	0	0	0	0	0	0.600	0	**0.185**
R^2	0	0	0	0	0	0	1	**0.129**

Step 5: determine TOPSIS decision matrix & calculate distances to ideal solutions and rank

TOPSIS allows for the evaluation of the forecasting models compared to each other based on the criteria weightings calculated by the ANP process in steps 3 and 4. Firstly, the decision matrix r_{mn} is determined normalising the solutions matrix, *F*, and multiplying it with the ANP weight vector ω_{ANP}.

$$r_{mn} = \frac{f_{mn}}{\sqrt{{f_{mn}}^2}} \times \omega_{ANP}$$

A decision matrix for the ranking of five alternatives in the implementation case is created on the basis of the criteria per alternative. The decision table is then normalised to allow for comparisons between values, shown in Table 6. It is worth noting that the absolute values of the normalised decision values for the cost-based criteria are taken to determine the positive and negative ideal solution vectors as these error measures also represent bias in the forecast results versus the actual. The "best" value is that which is closest to zero rather than the smallest. For example, in the case of ME, the worst value is the single judgmentally adjusted forecast even though it is the lowest on a numerical scale. However, it is the value farthest from zero which represents the highest error. This step is not necessary for R^2 as it does not consider a meaningful zero value.

Table 6. Normalised decision matrix ANP-TOPSIS

	Holt-Winter	ARIMA	SARIMA	Single Adjustment	Collaborative Adjustment
ME	0.036	0.001	0.017	-0.050	0.005
MAE	0.072	0.107	0.087	0.093	0.032
RMSE	0.052	0.073	0.061	0.097	0.024
MPE	0.112	0.018	0.058	-0.113	0.016
RMSPE	0.047	0.063	0.051	0.071	0.020
MAPE	0.076	0.108	0.088	0.088	0.031
R^2	0.062	0.024	0.049	-0.035	0.093

The decision matrix r_{mn} is used to determine the positive-ideal and negative-ideal solutions. It then calculates each alternative's distance to them and establishes a ranking based on these distances. To establish the positive-ideal solution vector (noted V+) and the negative-ideal solution vector (noted V-), the minimum value is determined from the alternatives for each criterion. Therefore, the maximum value for V- is determined for each criterion.

Subsequently the separation measures are calculated using the Euclidian distance. The separation of each alternative from the positive-ideal V+ is noted D+. Similarly, the separation of each alternative from the negative-ideal solution V- is noted D-.

$$D_m^+ = \sqrt{\sum\nolimits_n (r_{mn} - V_n^+)^2}$$

$$D_m^- = \sqrt{\sum\nolimits_n (r_{mn} - V_n^-)^2}$$

The relative closeness to the ideal solution is calculated as C_m and the performance order is ranked.

$$C_m = \frac{D_m^-}{D_m^- + D_m^+}$$

A larger index value means that the performance of the alternative is better.

In the implementation case, to establish the positive-ideal solution vector (noted *V*+) we determine which value is the minimum for the cost-based criteria in which a value closer to zero in error results is considered as better (ME, MAE, RMSE, MPE, RMSPE, MAPE) and the maximum for benefit criteria, which favours a higher solution as in the case of R^2. Inversely, to establish the negative-ideal solution vector (*V*-), the maximum values for the cost-based criteria are selected and the minimum for the benefit criteria. Table 7 shows the positive and negative ideal solutions. Note that the negative ideal solution for ME and MPE are the results of the single judgmentally adjusted forecast from the normalised decision matrix (Table 6) but with opposite sign. This is because they are the values which are farthest from the ideal solution of 0. The same is not true for R^2 as the ideal solution is 1, as R^2 does not have a meaningful zero.

Table 7. Positive-ideal (V+) and negative-ideal (V-) solutions for ANP-TOPSIS

Criteria	V+	V-
ME	0.001	0.050
MAE	0.032	0.107
RMSE	0.024	0.097
MPE	0.016	0.113
RMSPE	0.020	0.071
MAPE	0.031	0.108
R^2	0.093	-0.035

The relative closeness, *C*, is calculated based on the relative distances, *D+ and D-*, from the positive and negative ideal solutions, *V+ and V-*, in Table 8 and equations from section 3.5. Results can be seen in Table 8 in order of the closest model for the ideal solution to farthest and the respective ranking. The forecast selected using the ANP-TOPSIS approach is the collaborative judgmentally adjusted model, with a much greater relative closeness to the ideal solution, *C*, than all the other alternatives based on error measures as the evaluation criteria. The least desirable forecasting model is the single judgmentally adjusted forecast with an extremely low relative closeness to the ideal solution. The SARIMA and Holt-Winter models both integrate seasonality into their parameterisation which is one of the initial observations made on the historical data. Both of these models are ranked higher than ARIMA, which ignores the seasonal component, although by very little: a difference of 0.042 and 0.015. This leads us to believe that the judgmental adjustment of a forecast can either make or break the forecast and that a collaborative approach is the best method.

Table 8. Distances and final ranking (ANP-TOPSIS)

	Relative Closeness *C*	**Rank**
Collaborative Adjustment	0.982	**1**
SARIMA	0.508	**2**
Holt-Winter	0.481	**3**
ARIMA	0.466	**4**
Single Adjustment	0.099	**5**

3.1 Effect of interdependence between evaluation criteria

In order to validate the final ranking of the implementation case results and to determine whether the initial assumption of interdependence between variables is sound, the process is repeated omitting the interdependencies. Instead of using the weight vector including interdependence, ω_{ANP}, the priority weight vector *w* is used to determine the normalised decision matrix r_{mn}.

By ignoring the interdependence, the model alters the weight vector by 39%, based on the absolute values of the differences (Table 9). The most notable difference is the increase in the weight attributed to the MAPE and the decrease in MAE and MPE, which partly explains the different ranking of ARIMA and Holt-Winter between ANP-TOPSIS with and without considering interdependence between criteria in Table 10. No difference is observed in R^2 as it does not depend on the results of the other criteria and measures how close the forecasted values are to the fitted regression line.

Table 9. Comparison of weight vectors with (ω_{ANP}) and without (w) interdependence between criteria

Criteria	ω_{ANP}	w	% increase by omitting interdependence	Absolute difference
ME	0.064	0.046	-28%	0.018
MAE	0.184	0.098	-47%	0.086
RMSE	0.148	0.167	13%	0.019
MPE	0.171	0.080	-53%	0.091
RMSPE	0.119	0.173	45%	0.054
MAPE	0.185	0.308	67%	0.123
R^2	0.129	0.129	0%	0.000

The positive and negative ideal solutions are calculated in the same way. They are used to calculate the relative closeness and determine the ranking of alternative forecasts. The relative closeness and ranking of the results omitting the interdependence between criteria are shown in Table 10 against that of ANP-TOPSIS. The only change in the ranking is between SARIMA and the Holt-Winter models due to minor changes in their relative closeness's to the ideal solution. The relative closeness of SARIMA decreases by 0.046 and increases by 0.054 for the Holt-Winter model. These small changes are sufficient to swap the ranking of the two models as they were very close to begin with.

The ARIMA model remains ranked second last when omitting interdependence but is displaced farther from the ideal solution, losing 0.13 in relative closeness. On the other hand, the single judgmentally

adjusted forecast comes closer to the ideal solution when ignoring the interdependencies between criteria.

Table 10. Relative closeness to ideal solution and final ranking with and without inter-dependence between criteria

	Including interdependence between evaluation criteria		**Omitting interdependence between evaluation criteria**	
RANK	**Relative Closeness *C***	**Ranking of alternatives**	**Relative Closeness *C***	**Ranking of alternatives**
1	0.982	Collaborative Adjustment	0.987	Collaborative Adjustment
2	0.508	SARIMA	0.534	Holt-Winter
3	0.481	Holt-Winter	0.462	SARIMA
4	0.466	ARIMA	0.336	ARIMA
5	0.099	Single Adjustment	0.142	Single Adjustment

3.2 Sensitivity Analysis

To analyse the quality of the methodology in demand forecast selection, a sensitivity analysis is conducted to determine the effect of the criteria weighting on the final ranking of results. Therefore, experiments are run in which the weighting of one criterion is increased by 50% and the weighting of the remaining criteria is reduced proportionally. Table 11 shows the ranking based on the relative closeness of the alternatives for seven scenarios. Each scenario represents the ranking results when the weight of one specified criterion is increased by 50% and the weight of the other criteria is decreased proportionally so that the sum of the normalised weights equals 1. This is done for seven scenarios based on the seven criteria. In every case, the collaborative judgmentally adjusted forecasting model is always selected first with the largest relative closeness to the ideal solution, and the single judgmentally adjusted forecast is always the least favourable. This is in line with the original results and indicates that the forecast selection process is robust when selecting the best forecast out of a set of alternatives.

On the other hand, the ranking of the purely statistical models, ARIMA, SARIMA and Holt-Winter switch between each other. The most obvious is Scenario 4, where the ranking of ARIMA is better than the other two, and Scenario 1, where ARIMA outranks the Holt-Winter model. This is expected since the solutions for the ME and MPE of ARIMA are the best out of the five alternatives for ME and second best after the collaborative judgmentally adjusted forecast for MPE (Table 3). Scenario 3 and 5 have the same ranking results as those of the original results.

3.3 Statistical Significance Test of Predictive Accuracy

The Kolmogorov-Smirnov Predictive Accuracy (KSPA) test from (Hassani and Silva, 2015) is performed on the errors the two most highly ranked forecasting models: the Collaborative adjusted and SARIMA forecasting models, in order to determine whether there is a statistical significance between their distributions. Firstly, the absolute forecast error is calculated and squared for the 12-month forecasting period from January to December in Y_4 which is used in the two-sample two-sided and one-sided KSPA test. The two-sided KSPA test of the errors of the two models yield a P-value of 0.0337 which is less than 0.05, thus supporting the rejection of the null hypothesis and confirming that there is indeed a statistical significance between the errors of the two forecasting models. The one-sided KSPA tests whether the forecast with the lowest error also has a stochastically smaller error than the other forecasting model thereby testing whether there is a statistical significance between the forecasts (Hassani and Silva, 2015). The one-sided test yields a P-value of 0.01685, also less than 0.05, and confirms that the Collaborative adjusted forecasting model does provide a forecast with a lower stochastic error than SARIMA.

Table 11. Ranking based on relative closeness of the alternatives to the ideal solution based on an increase in each individual criterion

	Scenario 1 (ME +50%)	Scenario 2 (MAE +50%)	Scenario 3 (RMSE +50%)	Scenario 4 (MPE +50%)	Scenario 5 (RMSPE +50%)	Scenario 6 (MAPE +50%)	Scenario 7 (R^2+50%)
Collaborative Adjustment	1	1	1	1	1	1	1
SARIMA	2	3	2	3	2	3	3
Holt-Winter	4	2	3	4	3	2	2
ARIMA	3	4	4	2	4	4	4
Single Adjustment	5	5	5	5	5	5	5

The results of the sensitivity analysis show that in all cases, the model selected to forecast the polyethylene bag demand is the collaborative judgmentally adjusted model.

4 Discussion and Managerial Insights

This section presents additional discussion points regarding the implementation case results. The results from the implementation case consistently rank the collaborative judgmentally adjusted forecasting model as the best model by a large margin (relative closeness to the ideal solution) which supports the findings in (Cheikhrouhou et al., 2011) in which collaborative judgmentally adjusted forecasting models outperform single adjusted. Conversely, the approach also consistently ranks the single judgmentally adjusted forecasting model last, worse than ARIMA, which differs from the results in (Marmier & Cheikhrouhou, 2010) that select the single judgmentally adjusted forecasting model over ARIMA due to the better performance in MAE and MAPE. However, the ANP-TOPSIS approach uses several error measures for evaluation and integrates the experts' opinions into the weighting of importance for each criterion, which results in ARIMA ranking higher than the single judgmentally adjusted forecasting model.

The collaborative judgmentally adjusted forecast is followed by SARIMA and the Holt-Winter models which both take seasonality into consideration in the model parameters. This may seem obvious when reviewing the seasonal nature of the historical data.

The effect of excluding the interdependence of the criteria is determined by comparing the results of the relative closeness and ranking the alternatives with and without the inclusion of interdependence between error measures. The final ranking of alternatives does not largely differ when ignoring the interdependence between criteria indicating that the difference in the criteria weight vectors, as a result of the interdependence, does not have a large impact on the selected model.

The results in Table 11 show that the pure time-series autoregressive models, ARIMA and SARIMA, both decrease in relative closeness to the ideal solution when interdependence between criteria is ignored. Conversely, the exponential smoothing model, Holt-Winter, increases in relative closeness to the ideal solution. In addition, both judgmentally adjusted forecasts increase in their relative closeness; the single judgmentally adjusted forecast more so than the collaborative approach (0.043 and 0.005 respectively). This is most likely due to the weights attributed to the criteria.

The results from the sensitivity analysis validate the robustness of the selected model, consistently selecting the collaborative judgmentally adjusted forecast as the highest ranked model. Including the consolidated judgment from more than one expert is an integral part of obtaining a good forecast. The sensitivity analysis also consistently results in the single judgmentally adjusted forecasting model being the worst model. The only changes in the rankings were between the purely statistical models ARIMA, SARIMA and Holt-Winter models which show greater sensitivity to the criteria weights. However, this is probably due to the fact that there is a minor difference in the relative closeness to the ideal solution to begin with (Table 8) and not necessarily because they were sensitive.

Even though the ANP-TOPSIS process is highly subjective and founded in human judgment, there were some additional manual manipulations required in the implementation case to calculate the positive and negative ideal solutions ($V+$ and $V-$) for the ME and MPE. This is because the MCDM method does not

consider a meaningful zero and these values were below zero. In this case, the distance from zero is used as a measure to determine the absolute values of the errors. In this implementation case, it is only valid for the single judgmentally adjusted forecasting model for ME and MPE where their negative values show a conservative bias of under-forecasting the demand. Therefore, it is recommended that experts carefully analyse the results when there is a criterion that includes a non-arbitrary zero point.

The ANP-TOPSIS approach makes it possible to structure the issue into a network of criteria and alternatives, takes into account the interdependence of the evaluation criteria, allows a trade-off of poor results with good results due to its compensatory nature, and can compare the performance of forecasting models of different classes.

4.1 Managerial Insights

The plastic bag manufacturer had difficulty in selecting a forecasting model based solely on the results of the error measures they consistently used: ME and MAPE, which in the implementation case yield conflicting results. The results of the ME suggested the ARIMA model should be selected and the MAPE implied the collaborative judgmentally adjusted forecasting model should be selected. The error measurements also highlight the importance of considering a larger set of error measures as the evaluation criteria. The proposed ANP-TOPSIS approach provides the possibility to reduce the risk in selecting a forecasting model by considering several error measures simultaneously and can support the management team of the plastic bag manufacturer to select a new forecasting model, even in the case of conflicting error measurements. They historically used ARIMA to forecast future sales of all of their products, evaluating the forecast's performance using the ME and MAPE however the final results rank the collaborative judgmentally adjusted forecasting model as the best model. Consequently, these results support the justification for the change management necessary to implement a more structured forecasting process that includes a team of experts who adjust the forecast based on their knowledge of future events. Furthermore, using the ANP-TOPSIS approach to evaluate the alternative models enables

the company to avoid selecting the incorrect forecast, particularly in the case of the comparison between the single judgementally adjusted forecasting model and ARIMA. If only the MAE and MAPE were considered in evaluating these forecasts, the experts would have selected the single judgementally adjusted forecasting model over ARIMA, however by using additional error measures and weighting the evaluation criteria based on subjective sense of comparative importance, the ARIMA model was higher ranked in all cases.

Another beneficial impact of using the collaborative judgmentally adjusted forecast instead of ARIMA, is avoiding the potential loss of sales due to the frequent under-forecasting of the ARIMA model. This is not quantifiable as it depends on periodic decision making by the production and inventory planners that determine the level of inventory using a min-max system, based on inventory targets and actual sales of the previous month.

The selection of the best forecasting model amongst alternatives may also improve the budgeting cycle conducted at the end of every year for the following year by harmonising the bottom-up sales forecasting model with the top-down sales targets.

Conversely, the ANP-TOPSIS process should be facilitated by a knowledgeable person, which could be trained with regards to the approach to use in order to repeat it when deemed necessary, for example, on a yearly basis.

5 Conclusion

This paper presents an approach that can be a useful tool in industries to choose from several demand forecasting models of different classes. Using the ANP-TOPSIS approach presented in this paper, these companies are able to determine which forecasting model they should select using a set of error measures as evaluation criteria. The scientific contribution of the approach is that it allows the comparative evaluation of forecasting models of different classes using several interdependent error measures and

shows that the use of this approach can help avoid the selection of an inappropriate or 'worse' forecasting model.

The methodology was demonstrated using an implementation case of a plastic bag manufacturer. Five forecasting models of different classes were evaluated with seven error measures. The results showed conflicting results particularly between the ME, MAE and the MAPE. The ANP-TOPSIS approach enabled ranking of the alternative forecasting models taking into consideration all seven error measures as well as the interdependence between them. The interdependence between error measures showed to have an impact on the relative closeness to the ideal solution and therefore should be considered when evaluating alternative forecasting models using error measures. Nevertheless, the collaborative judgmentally adjusted forecasting model was consistently ranked first for the implementation case as demonstrated by means of a sensitivity analysis. Additionally, a KS predictive accuracy test confirmed the statistical significance of the errors between the collaborative judgmentally adjusted forecasting model and the second highest ranked model, SARIMA. The ANP-TOPSIS approach allowed the avoidance of selecting the inappropriate model; between the single judgmentally adjusted forecasting model and ARIMA, which show opposing results in MAE and MAPE versus the other measures. By considering all seven error measures and their interdependencies, the single judgmentally adjusted forecasting model is ranked as the worst model out of the alternatives.

A limitation was observed during the implementation case: the MAPE has a meaningful zero where the best solution is not simply the smallest value, but rather that which is closest to zero. This required a manual manipulation for calculating the minimum and maximum distances to the ideal solution. Therefore, this approach should be used in conjunction with a critical perspective when calculating the TOPSIS part of the approach. Another limitation is the subjective weighting by experts of the evaluation criteria. It is not so intuitive to weight error measures using Saaty's scale in a business concept as it may be influenced by several factors, such as the knowledge of the expert themselves and what they are accustomed to using or the level of clarity when it comes to the strategy of the company with regards to

forecasting accuracy and inventory. Therefore, it is recommended that future research should look into how to manage uncertainty in the weighting of the evaluation criteria.

It would also be interesting to see how qualitative evaluation criteria, such as the expertise required for a particular forecasting model, could be combined with the quantitative criteria of error measures. Another future direction is to investigate the impact of the implicit biases of the experts involved in weighting the evaluation criteria as they might favour a forecasting model that requires less expertise or effort, and weight it higher than the accuracy criteria favoured by the company's strategy.

6 Acknowledgements

This work was supported by the Swiss National Science Foundation under project n° [176349].

7 Declaration of interest

No potential conflict of interest was reported by the authors.

8 References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Al-Harbi, K. M. A.-S. (2001). Application of the AHP in project management. *International Journal of Project Management*, *19*(1), 19–27. https://doi.org/10.1016/S0263-7863(99)00038-1

Alvarado-Valencia, J., Barrero, L. H., Önkal, D., & Dennerlein, J. T. (2017). Expertise, credibility of system forecasts and integration methods in judgmental demand

forecasting. *International Journal of Forecasting*, *33*(1), 298–313. https://doi.org/10.1016/j.ijforecast.2015.12.010

Armstrong, J. S. (2001). Evaluating Forecasting Methods. In J. S. Armstrong (Ed.), *Principles of Forecasting* (Vol. 30, pp. 443–472). Springer US. https://doi.org/10.1007/978-0-306-47630-3_20

Arvan, M., Fahimnia, B., Reisi, M., & Siemsen, E. (2019). Integrating human judgement into quantitative forecasting methods: A review. *Omega*, *86*, 237–252. https://doi.org/10.1016/j.omega.2018.07.012

Cheikhrouhou, N., Marmier, F., Ayadi, O., & Wieser, P. (2011). A collaborative demand forecasting process with event-based fuzzy judgements. *Computers & Industrial Engineering*, *61*(2), 409–421. https://doi.org/10.1016/j.cie.2011.07.002

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. https://doi.org/10.1016/j.ijforecast.2012.09.002

Davydenko, A., & Fildes, R. (2016). Forecast Error Measures: Critical Review and Practical Recommendations. In *Business Forecasting: Practical Problems and Solutions*. John Wiley & Sons. http://rgdoi.net/10.13140/RG.2.1.4539.5281

Ergu, D., Kou, G., Peng, Y., & Shi, Y. (2011). A simple method to improve the consistency ratio of the pair-wise comparison matrix in ANP. *European Journal of Operational Research*, *213*(1), 246–259. https://doi.org/10.1016/j.ejor.2011.03.014

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in

supply-chain planning. *International Journal of Forecasting*, *25*(1), 3–23. https://doi.org/10.1016/j.ijforecast.2008.11.010

Fildes, R., Wei, Y., & Ismail, S. (2011). Evaluating the forecasting performance of econometric models of air passenger traffic flows using multiple error measures. *International Journal of Forecasting*, *27*(3), 902–922. https://doi.org/10.1016/j.ijforecast.2009.06.002

Franses, P. H. (2011). Averaging Model Forecasts and Expert Forecasts: Why Does It Work? *Interfaces*, *41*(2), 177–181. https://doi.org/10.1287/inte.1100.0554

Franses, P. H., & Legerstee, R. (2010). Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? *Journal of Forecasting*, *29*(3), 331–340. https://doi.org/10.1002/for.1129

Ghobbar, A. A., & Friend, C. H. (2003). Evaluation of forecasting methods for intermittent parts demand in the field of aviation: A predictive model. *Operations Research*, 18.

Ha, C., Seok, H., & Ok, C. (2018). Evaluation of forecasting methods in aggregate production planning: A Cumulative Absolute Forecast Error (CAFE). *Computers & Industrial Engineering*, *118*, 329–339. https://doi.org/10.1016/j.cie.2018.03.003

Han, S. (2016). How can we handle too many criteria/alternatives? : A study on AHP structural design. *NUCB Journal of Economics and Information Science*, *60*(2), 10.

Han, W., Wang, X., Petropoulos, F., & Wang, J. (2019). Brain imaging and forecasting: Insights from judgmental model selection. *Omega*, *87*, 1–9. https://doi.org/10.1016/j.omega.2018.11.015

Hassani, H., Silva, E. S., Gupta, R., & Segnon, M. K. (2015). Forecasting the price of gold. *Applied Economics*, *47*(39), 4141–4152. https://doi.org/10.1080/00036846.2015.1026580

Hassani, H., and Silva, E. S., (2015) "A Kolmogorov-Smirnov Based Test for Comparing the Predictive Accuracy of Two Sets of Forecasts," *Econometrics* 3(3), 590–609. https://doi.org/10.3390/econometrics3030590

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Kara, S. S., & Cheikhrouhou, N. (2014). A multi criteria group decision making approach for collaborative software selection problem. *Journal of Intelligent & Fuzzy Systems*, *26*(1), 37–47. https://doi.org/10.3233/IFS-120713

Marmier, F., & Cheikhrouhou, N. (2010). Structuring and integrating human knowledge in demand forecasting: A judgemental adjustment approach. *Production Planning & Control*, *21*(4), 399–412. https://doi.org/10.1080/09537280903454149

Mehdiyev, N., Enke, D., Fettke, P., & Loos, P. (2016). Evaluating Forecasting Methods by Considering Different Accuracy Measures. *Procedia Computer Science*, *95*, 264–271. https://doi.org/10.1016/j.procs.2016.09.332

Oh, C.-O., & Morzuch, B. J. (2005). Evaluating Time-Series Models to Forecast the Demand for Tourism in Singapore: Comparing Within-Sample And Postsample Results. *Journal of Travel Research*, *43*(4), 404–413. https://doi.org/10.1177/0047287505274653

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., & Siemsen, E. (2018). Judgmental selection of forecasting models. *Journal of Operations Management*, *60*, 34–46. https://doi.org/10.1016/j.jom.2018.05.005

Ren, L., & Glasure, Y. (2009). Applicability of the Revised Mean Absolute Percentage Errors (MAPE) Approach to Some Popular Normal and Non-normal Independent Time Series. *International Advances in Economic Research*, *15*(4), 409–420. https://doi.org/10.1007/s11294-009-9233-8

Saaty, T. L. (1990). How to make a decision: The analytic hierarchy process. *European Journal of Operational Research*, *48*(1), 9–26. https://doi.org/10.1016/0377-2217(90)90057-I

Saaty, T. L. (2001). *Decision Making with Dependence and Feedback: The Analytic Network Process : the Organization and Prioritization of Complexity*. Rws Publications.

Sipahi, S., & Timor, M. (2010). The analytic hierarchy process and analytic network process: An overview of applications. *Management Decision*, *48*(5), 775–808. https://doi.org/10.1108/00251741011043920

Tao, L., Chen, Y., Liu, X., & Wang, X. (2012). An integrated multiple criteria decision making model applying axiomatic fuzzy set theory. *Applied Mathematical Modelling*, *36*(10), 5046–5058. https://doi.org/10.1016/j.apm.2011.12.042

Taylor, J. W., & McSharry, P. E. (2007). Short-Term Load Forecasting Methods: An Evaluation Based on European Data. *IEEE Transactions on Power Systems*, *22*(4), 2213–2219. https://doi.org/10.1109/TPWRS.2007.907583

Trapero, J. R., Kourentzes, N., & Fildes, R. (2012). Impact of information exchange on supplier forecasting performance. *Omega*, *40*(6), 738–747. https://doi.org/10.1016/j.omega.2011.08.009

Van den Broeke, M., De Baets, S., Vereecke, A., Baecke, P., & Vanderheyden, K. (2018). Judgmental forecast adjustments over different time horizons. *Omega*. https://doi.org/10.1016/j.omega.2018.09.008

Velasquez, M., & Hester, P. T. (2013). An Analysis of Multi-Criteria Decision Making Methods. *International Journal of Operations Research*, *10*(2), 11.

Villegas, M. A., Pedregal, D. J., & Trapero, J. R. (2018). A support vector machine for model selection in demand forecasting applications. *Computers & Industrial Engineering*, *121*, 1–7. https://doi.org/10.1016/j.cie.2018.04.042

Webby, R., & O'Connor, M. (1996). Judgemental and statistical time series forecasting: A review of the literature. *International Journal of Forecasting*, *12*(1), 91–118. https://doi.org/10.1016/0169-2070(95)00644-3

Xu, B., & Ouenniche, J. (2012). Performance evaluation of competing forecasting models: A multidimensional framework based on MCDA. *Expert Systems with Applications*, *39*(9), 8312–8324. https://doi.org/10.1016/j.eswa.2012.01.167