On the Robustness of ARIMA-based Benchmarks for Corporate Financial Planning Quality

Jochen Martin and Thomas Setzer
Karlsruhe Institute of Technology
{jochen.h.martin | thsetzer}@gmail.com

Abstract

Corporate financial planning relies on thousands of financial forecasts generated by human forecasters with varying performance (forecast errors). Previous work proposes ARIMA prediction as a competitive benchmark for manual forecasts. However, ARIMA can also produce large errors, and a company needs to understand sensitivity of ARIMA-outcome to time series characteristics before ARIMA-benchmarks can be established. Using forecast data provided by a global corporation, we present sensitivity analysis of ARIMA on shifts in fitting periods including the financial crises. Results show that ARIMA leads to rather robust performance, on average dominating human forecasters, with some huge errors not made by forecasters. We conclude that ARIMA can be applied to generate benchmarks in financial planning, which can then be refined to reflect novel expectations.

1. Introduction

Corporate financial planning (CFP) in multinational enterprises is a time-consuming and challenging task, in particular in heterogeneously structured companies with decentralized data generation processes. Typically, CFP is based on financial forecasts generated and delivered individually by subsidiaries over a period of 12 months or longer. Forecasts are mostly based on forecasters’ opinions and consequently biased by his/her cultural background, the organizational structures of a company, or the company’s line of business, amongst other dimensions. In order to make a consolidated corporate-wide financial liquidity or hedging plan, the heterogeneous financial forecast data with hundreds of thousands of predicted items (planning items) delivered by many subsidiaries must be integrated in standardized processes to allow for efficiently validating the data, with the final goal of increasing the quality of the forecasts and resulting corporate financial planning.

For large, global companies, it is particularly challenging to compose high-quality planning [1]. For instance, in central currency-specific liquidity planning, non-centralized planning and revisioning processes have to be coordinated within local partitions and internal transactions between these partitions have to be tracked and consolidated to ensure a proper and consistent overall data standard and corporate financial planning model.

To cope with such challenges global companies are faced with, corporate financial portals have turned out to be a very efficient means of enhancing the process of centralized liquidity risk management [2]. In this spirit, IT-based business analytics services can be offered to support planning processes, such as market-based prediction, the identification of complex events as a sequence of defined activities. Such decision support and validation services are usually based on knowledge extraction or pattern detection techniques, especially if vast volumes of data are supposed to be processed for this task.

However, detecting patterns in large data sets can be tedious since it requires a variety of upstream and downstream efforts and may be quite different for different data sets delivered [3]. Moreover, quality of financial forecasting activities is typically assessed by its outcome using broadly accepted ex-post measures such as accuracy or alternative derivatives of forecast versus actual deviances – i.e., forecasting errors. However, cash flow positions are often forecasted months or even years before they actually take place, with frequent revisions of their monetary quantification based on additional knowledge and adapted expectations.

Unfortunately, today corporate financial controllers have little guidance on how to assess forecasting processes’ quality. This is especially true because of the complex data structure in financial forecasting processes accompanied by often unknown dynamics.

However, to assess, rate, and finally improve forecasting quality, ‘fair’ quality metrics are required that do not solely consider the prediction errors but also relate the error to the difficulty of the prediction. Consider the case of our industry partner, a global company with subsidiaries in divisions such as
material science (MS), crop science (CS), health care (HC), and diverse (DV). The last cluster contains entities that can not exactly be assigned to one of the previous divisions. Each division exhibits individual characteristics, which heavily impact the difficulty of deriving accurate predictions.

For instance, MS supplies a huge amount of materials or upstream products to the automotive industry. Sales volumes, and, hence, orders and respective invoices issued strongly depend on macro economic developments, which typically exhibits a rather volatile behavior leading to reduced predictability and predictive accuracy. Another example of stochastic external data are weather conditions and seasonal patterns that heavily impact actual items in CS. On the other hand, according to experts within the company, HC has no explicit underlying influences but simply depends on the maturity of licenses; hence the actual time series are much smoother and rather steady, hence, somewhat easier to predict. Using the same quality metric to assess and rate forecasting quality for all companies despite the individual characteristics of their lines of business (such as absolute ex-post planning errors) clearly disadvantages companies with more volatile and stochastic developments of actual items over time and does not necessarily reflect the effort and quality of the forecasts at company level.

The absence of fair metrics considering the difficulty of the prediction is an important issue in the forecasting community in general and has lead to more sophisticated evaluation statistics. Arguable the most well known candidate of such a metric goes back to the seminal work of Henry Theil in 1966, who introduced the $U$-statistics [4]. $U$-statistics have a number of attractive properties. Those do not provide information on forecasting bias measured by mean errors (the most basic one of all forecast evaluation statistics). In contrast, $U$-statistics are derived by dividing the Root Mean Square Error (RMSE) of a forecasting method by the RMSE of a dummy (or no-change) prediction model, where simply the last known actual value is also the prediction of the subsequent one. A $U$ of 1 means that the model is as good as the dummy. If $U$ exceeds 1, the dummy would produce lower prediction errors. Hence, a model should only be used when $U$ is significantly smaller than 1, indicating that more accurate forecasts than a no-change model can be obtained. In general: the smaller $U$, the better the model. The intuition is to normalize forecasting errors by the volatility of a time-series (as the RMSE of the dummy prediction increases with volatility). Hence, smaller prediction errors in rather smooth time-series are penalized harder, assuming that the prediction is supposed to be easier with such time series.

However, the application of $U$-statistics or derivatives comes along with other biases questioning its appropriateness for the evaluation of financial forecasting process quality. That is because the metric benefits companies with actual time series exhibiting systematic patterns such as cyclic or trend components, as those can be found for instance in CS with seasonal structures, and current actual values are typically auto-correlated with a lag of twelve months. Obviously, using this knowledge can improve forecast accuracy measured with $U$-statistics significantly, in particular if the actual values differ significantly between subsequent months of a year. Recent work on group differences in corporate financial forecasting quality indeed revealed a systematic impact of these (time series) characteristics on forecast/actual deviations, which emphasizes the need for a 'fair' measure to assess the quality of financial forecasting processes [5].

Consequently, the same authors argue that 'fairer' metrics can be derived by replacing the dummy forecast by a statistical forecast that considers the major structures present in historic time series of actual items. The authors argue 'fairness' of the resulting statistics by the fact that statistical prediction purely operate on known historical observations, information that is also available for forecasters. The authors have shown that the use of statistical prediction produced by ARIMA models instead of dummy forecasts as benchmark not only fulfills the major criteria of fairness in terms of objectivity, but also derives surprisingly accurate predictions. This makes it also a valuable tool for forecasters, who can use the forecasts for preliminary orientation which can then be manually adjusted to reflect novel or simply different expectations not captured by autoregressive, moving-average time-series prediction models such as ARIMA.

Unfortunately, although their analytical results are very promising, the predictive power of ARIMA was rather low for certain subsets of actual items as the success of auto-regressive prediction models depends on the right fitting of the model on the historical data set (the training data); in other words, the re-occurrence of learned patterns in the prediction period.

As a consequence, before ARIMA-based quality metrics can be established in CFP, corporates need to better understand the covariates of the predictive accuracy of such models in the CFP domain, in particular the sensitivities regarding changing (and moving) fitting periods and the impact of structural breaks in empirical financial time series data. In particular, more insights are required regarding the dependency between the predictive power of a statistical prediction versus the forecast accuracy of human forecasters.
Based on historical data of a global company, in this paper we present first results of sensitivity analysis of the ARIMA model performance with reduced fitting period durations, including fitting periods entailing the structural breaks during the financial crises 2008 and 2009. In particular, we address the research question RQ 1: How robust is the performance of ARIMA forecasts in financial planning regarding different fitting periods and structural breaks in fitting periods? In addition, we aim at empirically evaluating interdependencies between time series characteristics such as seasonality and the influence of fitting periods to address a second research question RQ 2: Is the relative performance of human forecasting processes compared to ARIMA predictions depending on division-specific characteristics?

The outcomes show that although we can clearly identify such division-specific dependencies, the quality of statistical predictions within a division as well as the ratios of manual and statistical forecast MAPEs are surprisingly robust over time and division.

The remainder of this article is structured as follows: in Section 2, we will describe the design of our empirical analysis. First, we characterize the empirical data sets from our industrial partner. Second, we will introduce the metrics used for quality assessment of financial forecasts. Third, we will describe the design of our analysis and detail on the particular research questions. In Section 3 analytical results are presented and discussed. Finally, in Section 4 we summarize our major findings, their managerial impact and limitations.

2. EMPIRICAL ANALYSIS

In this section we will describe the design of the empirical analysis. Outcomes are then presented and discussed in the next section. First, we will briefly characterize the available empirical data and introduce the metrics used to measure the quality of financial predictions. Second, we will describe the evaluation framework and the different forecasting scenarios considered in our analysis.

2.1. Available Data

The data set to be evaluated is the cash flow-oriented financial forecast data and the corresponding actual values from our industry partner. The forecast data contains data sets delivered quarterly from November 2007 (fourth quarter 2007) to November 2011 (fourth quarter 2011). Each data set contains around 200,000 planned values delivered by more than 100 subsidiaries that have handed in forecast data during this delivery. The forecast data per entity reports expected volume of issued and received invoices, cash flows, tax payments and so on for one to 15 months into the future. Due to the data delivery on a quarterly basis with a maximum forecasting horizon of 15 months, each item is planned five times with only one corresponding actual value. The actual value and all forecast values have the same value date, however, with different plan dates. The structure of value and forecast data is illustrated schematically in Figure 1.

![Figure 1. Actual items and planning items](image)

As an example, the figure shows the volume of invoices received from a certain partner, in a certain currency per month. The actual values of invoices received in the three months of a particular value quarter (Q-Values; the real invoices received in that quarter) are shown on the bottom of the figure. The first prediction (usually also called planning) for the Q-Values is delivered five quarters earlier (the values at the top of the figure (the upper line). Subsequently, the predictions for the Q-Values are potentially revised every subsequent quarter. The last prediction is then delivered one quarter before the Q-value quarter. Hence, for the same individual actual items we find five predictions with different prediction horizons within the data, and in each quarter values are predicted with horizons of up to 15 months. The actual data contains 60 months from January 2008 to December 2012.

2.2. Metrics for Planning Quality

Accuracy measures can be calculated for pairs of actual values with corresponding forecasted values. Each of the resulting five accuracy values (reconsidering that each actual value is forecasted five times) is characterized by the forecasting horizon of the underlying forecast data lasting from 1–3 to 13–15 months.
2.2.1. Assessment based on Forecast-Actual Deviation. One popular measure for forecast error is the Absolute Percentage planning Error (APE), that we compute for each month \( t \) using (1).

\[
APE_t = \frac{|AE_t|}{|A_t|} = \frac{|E_t^p - A_t|}{|A_t|} \tag{1}
\]

Please notice that the APE can be computed for sub-samples such as company, division, currency, business partner, the actual item type (invoice issued, invoice received, etc.), which would result in many indices to identify a certain subsample. As in this work we conduct analysis either on the total samples, grouped by \( t \) or on one of the five business divisions, for reasons of brevity we will avoid the introduction of indices beyond \( t \). The absolute error \( AE_t \) per month in (1) is calculated as the absolute difference between an actual value \( A_t \) and its prediction \( E_t^p \). Based upon that, the absolute percentage error per month \( t \), \( APE_t \) is calculated as the ratio of \( AE_t \) and corresponding actual exposure \( A_t \).

A major problem with APE is that two different effects can boost the indicator: first, an increased absolute error, and second, a very small actual value approaching to zero. Unfortunately, preliminary analysis derived that the second effect can dramatically drive the calculated APE although planning errors where low even when only few actual items approached zero.

As we are only interested in the first driver and want to mitigate the second effect, we calculate the Mean Absolute Percentage Error (MAPE) for time series not as average APE, but base our evaluation on the average absolute actual exposure in the denominator. Consequently, in this paper MAPE is always calculated for time series with the length \( s \) as shown in (2).

\[
MAPE_s = \frac{1}{|E_{mean}|} \sum_{t \in T} |AE_t| = \frac{1}{\sum_{t \in T} |A_t|} \sum_{t \in T} |E_t^p - A_t| \tag{2}
\]

In (2), \( T = \{t_1, \ldots, t_s\} \) is the set of time series points and \( |E_{mean}| \) denotes the mean absolute value of actual values in the time series. For practical implications it is important to note that an accuracy measure in the presented way is a rather pessimistic indicator in the context of our industrial partner. Since accuracy is calculated separately for each currency and subsidiary, inter-company errors in planning data are taken into account twice and netting effects resulting from positive and negative errors are completely ignored. In addition, the planning currency differs from the booking currency in some special cases, for instance, due to legal requirements. We controlled for many of such peculiarities but might not know all cases that lead to increased forecasting errors according to our metric. However, despite all these effects the achieved results reveal high overall accuracy in the financial forecasting data of our industrial partner.

2.2.2. ARIMA in Corporate Financial Planning. Today, Auto-Regressive Integrated Moving Average (ARIMA) time series predictions models are considered state-of-the-art and are widely used as those typically lead to lower prediction errors than other prediction models [6, 7].

ARIMA models have a long track of successful application in a broad area ranging from economic modeling [8] to electricity price modeling [9] often outperforming other forecast methods [10].

The model is referred to as ARIMA\( (p,d,q) \) with non-negative integer parameters \( p, d, \) and \( q \) referring to the order of the autoregressive, integrated, and moving average parts of the model. They can be applied even in cases where time series show evidence of non-stationarity, where initial differencing steps (the "integrated" part) is used to eliminate non-stationarity.

However, the optimal fitting of the parameters of an ARIMA model is challenging and requires deep knowledge of time series analysis. The intuition of using ARIMA outcomes as benchmarks is that it can be derived automatically with reasonably good results independent of a managers skills in statistical forecasting and time series prediction. We use a variant of the Hyndman and Khandakar algorithm that combines unit root tests and minimization of the AICc to obtain a proper parameterization of the ARIMA model to be used for a particular time series [11].

As it has been shown in many other domains, in extensive evaluations with corporate financial planning data [5] have shown that predictions based on such auto-fitted ARIMA models lead to comparable, often smaller MAPEs than planning solely based on experts opinions. In their evaluation, the authors split actual data of various subsamples into 48 months fitting period (January, 2008 to December, 2011) and twelve months evaluation period from January 2012 to December 2012 (in case of twelve months prediction). The good predictive performance of ARIMA makes it a very valuable tool for planners as it might give first orientation points.

However, their work also revealed that for some subsets of data ARIMA performed rather poor and there is clearly a need for better understanding when ARIMA
based benchmark forecasts are expected to derive good prediction and when it’s not.

The observation that ARIMA models typically (on average) perform well in a certain domain but also produce exceptionally high prediction errors for specific time series has been observed and described in many papers and is typical for any auto-regressive prediction model [12, 13].

The strength of ARIMA is that components in time series such as trend and seasonality are learned from given data and projected into the future, often leading to better predictive performance compared to models that ignore the historic patterns in time series.

However, the strength of ARIMA can turn into weakness, as structural breaks may be present in a time series, strongly affecting the performance of ARIMA. In case of structural breaks, the model learned by ARIMA is not valid as, after the break, it can’t anticipate future development of a time series. In addition, a model must be re-learned using more recent data point, while many observations are required to have a sufficient number of recent values to learn a novel model to forecast future values. Hence, depending on the duration of the time interval used to learn the ARIMA model, a single but strong structural break in a time series can strongly affect the ARIMA performance for longer periods of time.

In this paper we will study two such effects: first, the influence of a reduced or extended fitting period; second, the development of forecasting accuracy with moving fitting periods.

2.3 Research Design

We will address the first research question RQ 1: How robust is the performance of ARIMA forecasts in financial planning regarding different fitting periods and structural breaks in fitting periods? To evaluate the influence of the fitting period, we will compare a longer, four-year fitting period (used to predict time series of the subsequent year) with shorter, two-year fitting periods, also predicting the subsequent year.

In addition, we want to compare the characteristics influencing the performance of ARIMA with those of manual forecast as expressed in RQ 2: Is the relative performance of human forecasting processes compared to the accuracy of ARIMA predictions depending on division-specific characteristics?

From the cooperation with our industrial partner we know that parts of the business can be characterized in different ways due to different lines of business: first we have seasonal developments caused by natural seasons and second we have entities depending on macro economy, underlying stochastic volatility supposed to be more difficult to predict.

In our analytical analysis we chose a fitting period of 24 months followed by a prediction and evaluation period of 12 months. The prediction is compared to forecasted values with an analogous forecasting horizon lasting from 1–12 months. Moreover, we benchmark the short fitting period results against the results for a 2012 prediction based on the 48-month fitting period, which was the longest time series available in our data (please notice that in our analysis fitting periods of only one year or less resulted in much higher MAPEs and are not considered in this article).

As additional benchmark, we calculate dummy forecasts for all evaluations. This dummy forecast is calculated as \( f_t = a_t \), hence, the forecast \( f_t \) for \( t \) equals the actual in \( t-1 \) (the no-change forecast used for example by U-statistics). However, we did not let the dummy forecast predict the whole twelve months in advance (as planners and ARIMA do), but we set the forecast horizon of the dummy forecast always to one. For instance, while ARIMA makes prediction for example in January for the next December, the dummy forecast was allowed to take the value of November to predict December values. We are aware of the fact that especially for a long forecasting horizon of 12 months in human and ARIMA forecasting, such a comparison is not really fair as prediction difficulty and resulting foreasting errors increase with planning horizon. However, we decided to take the predictions of the dummy forecast as described as a more competitive benchmark.

In total, we calculate nine basic evaluation scenarios. These scenarios result from our basic setting: 24 months fitting plus twelve months evaluation require 36 months in total per scenario. With 60 months actual data overall and a step width of three months between the scenarios we get nine different scenarios with different fitting and forecasting periods. For each scenario, we compared the performance of ARIMA forecasts against human planning and dummy forecasts.

To increase the robustness of our result and to identify division-specifics, we conduct all evaluations for five data samples: these are the complete data sample always including the entire data set and four disjoint sub-samples, determined by the division a specific plan item delivering entity belongs to, namely material science (MS), crop science (CS), health care (HC), and diverse (DV).

This subgroup differentiation allows for clustering the division specific time series characteristics. A few examples for such characteristics have already been described earlier in this paper, in particular the dependence of MS on macro-economic development, or the seasonal patterns in CS. A major goal of our
work is to relate group-differences with differences in fitting sensitivity.

3 EMPIRICAL RESULTS

Using the empirical dataset provided by our industry partner, in this section we present and discuss evaluation results clustered into the evaluation of ARIMA forecasting in corporate financial planning through statistical measures (Table 1) and the comparison to dummy forecasts and forecasts generated by human forecasters (Table 2). Please notice that for compliance reasons, only relative values and linear transformations of actual and forecast values are shown in the tables.

The results underlying the discussion of ARIMA forecasting accuracy are presented in Table 1. The table is structured into five result groups depending on the subgroups (divisions) and nine evaluation settings named with the corresponding fitting period (24 months) in the column headers. In addition, column ‘4 years 1 2008 12 2012’ shows the MAPE derived by ARIMA for the final year with a four-year fitting period from 1 2008 to 12 2011. The outer-right column in the table shows the average MAPE of the nine samples with two-year fitting periods to provide an aggregated view. For each evaluation setting we calculate MAPE per entity and display mean, median, and variance for the respective sample.

For sample Complete, we observe a rather constant median over time with a MAPE of 64% on average, ranging between 63% and 73%. We observe a comparable, slightly lower median of 60% with a four-year fitting-period. However, with two-year fitting periods, average mean (74%) and variance (96%) values are significantly lower compared to the long, four-year fitting, with a mean MAPE of 105% and variance of 282%. When we look at the individual fitting periods, especially variance exhibits strong peaks for fitting periods 10 2008 to 9 2010 and 10 2009 to 9 2011, both periods lead also to significantly higher mean MAPE values.

We will now drill-down on division-level to further analyze the general findings and in particular the two peaks. For all divisions but CS, two-year fitting lead (on average) to significantly lower sample median and variance compared to a longer fitting period. Interestingly, for each of the four divisions we observe comparable, division-typical median values throughout the years and fitting periods. For instance, with HC we obtain median values ranging between 32% and 45%, while with CS we observe median values between 67% and 85%. Surprisingly, with our data set, fitting the model during the crises with its unique and highly volatile curves in 2008 and 2009 does not necessarily decrease the accuracy of the learned prediction model. Furthermore, taking the average accuracy level into account, it seems that predicting CS and DIV data is hardest. Moreover, the variance-peaks in the complete data set results from huge variance peaks in the DIV sample of 919% in 10 2008 to 9 2010 and 779% in 10 2009 to 9 2011.

As aforementioned, while for HC, MS, and DIV two-year fitting dominated a longer four-year fitting, the opposite is true for CS, where the longer four-year fitting clearly outperformed the shorter fitting periods and resulted in very low variance of only 7% (compared to 282% with the Complete sample and

<table>
<thead>
<tr>
<th>Table 1. ABSOLUTE ARIMA PERFORMANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>CS</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>HC</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>MS</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>DIV</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
10% for CS with two-year fitting periods), low median of 49% (compared to 60% with the Complete sample and 75% for CS with two-year fitting) and low mean value of 56% (compared to 105% with the Complete sample and 76% for CS with two-year fitting). In addition, for CS with two-year fitting periods we observe a strong increase in mean (20% points) and median (26% points) compared to the Complete sample. The higher values with CS and two-year fitting periods are rather constant over all fitting periods and not the result of individual outliers.

In summary, we can not confirm that longer fitting periods lead to better results in general even with auto-fitted ARIMA models. It seems the appropriate length of the fitting period depends on the division, i.e., on the individual division characteristics. We found strong indication for the positive influence of longer fitting on accuracy in seasonal time series, while volatile time series like in MS do not necessarily require long fitting periods, as prediction accuracy even improves with short fitting. Finally, the results indicate the stronger influence of outliers in short fitting periods.

Besides the absolute performance of financial forecasts produced by ARIMA, we are primarily interested in the relative performance of dummy and human forecasts to the statistical ARIMA forecast, as we want to study the relation or ratio of the corresponding predictive outcomes. Comparative results are summarized in Table 2.

The layout of Table 2 is similar to the one in Table 1 with five samples, nine evaluations with different two-year fitting periods, one benchmark scenario, one aggregated column showing the average of the two-year fittings and the column with results with the four-year fitting period. However, each entry in the table now shows the ratio of companies where ARIMA outperforms dummy forecast (ARIMA vs. Dummy) and human planning (ARIMA vs. Plan). For instance, with the fitting period "1 2008 to 12 2009" ARIMA forecasting is more accurate in 50 out of 86 company/currency combinations (58%) in the Complete sample compared to human forecasting, and leads to lower MAPE values in 67% of company/currency combinations on the Complete sample compare to a dummy or no-change forecast.

Although not directly relevant for this work, we included the ratio of ARIMA versus dummy in the table to see whether predicting with ARIMA leads to lower MAPE compared to the dummy forecast (although we know that the comparison with the dummy forecast is of limited value since the dummy forecasts benefits from short forecast horizons, the dummy results provide another benchmark to assess the predictive power of human and statistical forecasts). However, despite the 'unfair' comparison of ARIMA and forecaster performance with the dummy forecast,

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete ARIMA vs. Dummy</td>
<td>67%</td>
<td>67%</td>
<td>65%</td>
<td>65%</td>
<td>70%</td>
<td>72%</td>
<td>69%</td>
<td>66%</td>
<td>63%</td>
<td>57%</td>
<td>67%</td>
</tr>
<tr>
<td>ARIMA vs. Plan</td>
<td>58%</td>
<td>59%</td>
<td>58%</td>
<td>63%</td>
<td>59%</td>
<td>57%</td>
<td>66%</td>
<td>55%</td>
<td>57%</td>
<td>51%</td>
<td>59%</td>
</tr>
<tr>
<td>CS ARIMA vs. Dummy</td>
<td>67%</td>
<td>60%</td>
<td>53%</td>
<td>60%</td>
<td>60%</td>
<td>60%</td>
<td>60%</td>
<td>53%</td>
<td>47%</td>
<td>40%</td>
<td>53%</td>
</tr>
<tr>
<td>ARIMA vs. Plan</td>
<td>47%</td>
<td>40%</td>
<td>40%</td>
<td>60%</td>
<td>60%</td>
<td>53%</td>
<td>47%</td>
<td>40%</td>
<td>33%</td>
<td>33%</td>
<td>47%</td>
</tr>
<tr>
<td>HC ARIMA vs. Dummy</td>
<td>81%</td>
<td>73%</td>
<td>69%</td>
<td>69%</td>
<td>85%</td>
<td>88%</td>
<td>85%</td>
<td>77%</td>
<td>77%</td>
<td>65%</td>
<td>78%</td>
</tr>
<tr>
<td>ARIMA vs. Plan</td>
<td>65%</td>
<td>58%</td>
<td>54%</td>
<td>46%</td>
<td>58%</td>
<td>38%</td>
<td>62%</td>
<td>46%</td>
<td>54%</td>
<td>42%</td>
<td>53%</td>
</tr>
<tr>
<td>MS ARIMA vs. Dummy</td>
<td>55%</td>
<td>55%</td>
<td>100%</td>
<td>64%</td>
<td>73%</td>
<td>64%</td>
<td>64%</td>
<td>45%</td>
<td>64%</td>
<td>55%</td>
<td>65%</td>
</tr>
<tr>
<td>ARIMA vs. Plan</td>
<td>73%</td>
<td>64%</td>
<td>82%</td>
<td>73%</td>
<td>55%</td>
<td>73%</td>
<td>91%</td>
<td>64%</td>
<td>64%</td>
<td>55%</td>
<td>71%</td>
</tr>
<tr>
<td>DIV ARIMA vs. Dummy</td>
<td>62%</td>
<td>71%</td>
<td>56%</td>
<td>62%</td>
<td>62%</td>
<td>68%</td>
<td>62%</td>
<td>71%</td>
<td>59%</td>
<td>41%</td>
<td>64%</td>
</tr>
<tr>
<td>ARIMA vs. Plan</td>
<td>53%</td>
<td>68%</td>
<td>62%</td>
<td>74%</td>
<td>62%</td>
<td>68%</td>
<td>71%</td>
<td>65%</td>
<td>68%</td>
<td>56%</td>
<td>66%</td>
</tr>
</tbody>
</table>
ARIMA prediction with a longer forecast horizon still dominated dummy predictions in 67% of cases (57% on average within the longer four-year fitting period). Again, the table shows rather constant ratios over time, ranging between 63 and 72%. Moreover, benefits of using two-year fitting periods are found in all subgroups except CS, with a ratio of 58% on average for two-year fitting. On the other hand, CS with a four-year fitting period lead to a ratio of 80%, the highest ratio in the table at all.

One potential statistical explanation for the benefits obtained with longer fitting periods in CS is the strong seasonal pattern in CS financial data, where over 60% of the variation can be explained by an annual time series component. Here, fitting periods of two years might be too short to learn a seasonal pattern, which strongly degrades the robustness of seasonal prediction and consequently limit the absolute ARIMA performance as well as the performance against dummy forecasting. Nevertheless, ARIMA dominated the dummy forecast even in CS in 58% of cases.

With these first results in mind, we can now focus our evaluation on the comparison of ARIMA predictions with human forecasters. In contrast to dummy benchmarking, ARIMA and human forecast have the same forecast horizon of 1–12 months and consequently have access to the same information. With the Complete sample, ARIMA leads to better results on average with two-year fitting in 59% of cases (51% with a four-year fitting period), while with two-year fitting the ratios ranged from 47% to 71%. Except for CS (again), these results support the application of the shorter two-year fitting periods with a maximum increase of 16% point for MS, where the highest degree of volatility in the time series is found. Hence, it seems that handling insecure economic times and high volatility in time series is a challenge not only for statistical forecasting, but –as expected– also for human forecasters.

This finding supports the idea of normalization of forecasters’ prediction errors by those of an ARIMA-based forecast to better reflect the quality of financial forecasting processes.

The fact that the “ARIMA vs. Plan” ratios shown in Table 2 are typically above 50% is surprising as an implication might be that the performance of human forecasting processes is relatively poor and should be replaced by ARIMA-prediction. This conclusion is oversimplified and misleading as the results are based on Mean Absolute Percentage Error (MAPE) as criterion for forecasting performance. For instance, when computing MAPE two forecast-actual deviation of 10% are weighted equally independent of their actual volumes, although the monetary error increases linearly with the actual volume. From corporate financial controlling we know that forecasters spend significantly more effort and time to predict very large values while smaller values are often considered less important. The same arguments hold for the relatively good performance of dummy forecasting.

However, in this paper we are interested in the robustness of predictive performance of ARIMA models in the context of corporate financial planning. As one of the objectives of such a benchmark would be the identification of poor planning processes that should be improved, MAPE seems to be an adequate criterion precisely because it is not volume-weighted and allows for identifying forecasting processes with low performance independent of the actual volume. Remarkably, the huge peaks in variance and mean MAPE values with ARIMA within particular evaluation samples (see Table 1) seem not to influence the ratios. However, we found no evidence in human forecasting for extraordinary forecasting errors in the ’outlier’ samples. Hence, although on average ARIMA forecasts dominate human forecasting, human forecasting might have had used novel information for their prediction not extractable from past time-series behavior and could better compensate these anomalies in contrast to ARIMA.

4. Conclusions and Future Work

Typically, corporate financial planning is based on financial forecasts generated individually by subsidiaries based on forecasters’ opinions. Previous work has shown that quality-assessment of manual financial forecasts by using statistical forecasts produced by state-of-the-art ARIMA as benchmark is meaningful as ARIMA often derives surprisingly accurate predictions, which make it a promising candidate for benchmarking human forecasting quality.

However, before such quality metrics can be established in CFP, a company needs to better understand the sensitivity of the prediction quality to the length of the fitting periods as well as influences of structural breaks and anomalies typically existing in corporate financial time series.

In this paper, we presented sensitivity analysis of forecast quality of statistical forecasts derived by automatically parameterized ARIMA-models based on historical data of a global company. The outcomes show that in general the quality of ARIMA and the derived metric for forecasting quality are surprisingly robust over time within divisions, even when fitting the model with time series during the time span of the financial crisis. It turns out that for all divisions but CS, a fitting period for the statistical forecast of two years leads to low planning errors. However, for CS with strong seasonal patterns, longer periods are advised.
The results also show that although on average \textit{ARIMA} derives a forecasting quality comparable or even dominating human forecasting, human forecasters can compensate very huge anomalies in contrast to the statistical forecast. A potential explanation would be that forecasters might have had additional knowledge for their prediction not extractable from past observations. These findings support the idea of using \textit{ARIMA} as basic benchmark, which then has to be corrected or adjusted by forecasters using their additional knowledge.

As of today, there is almost no literature available describing the planning accuracy of corporate financial planning, nor are there benchmark datasets that would allow to generalize our findings. Without doubt, in order to establish benchmarks for planning quality in corporate financial planning or alternative domains, more empirical evidence would be of great value. We hope, with the encouraging findings, more research groups and companies will follow this line of research in order to establish and enhance the usage of the proposed metric (or derivatives) in CFP.

5. References


