Tales from tails

Spiliotis, Evangelos; Nikolopoulos, Konstantinos; Assimakopoulos, V.

International Journal of Forecasting

DOI:
10.1016/j.ijforecast.2018.10.004

Published: 01/04/2019

Tales from tails: On the empirical distributions of forecasting errors and their implication to risk

Evangelos Spiliotis[a,*], Konstantinos Nikolopoulos[b], Vassilios Assimakopoulos[a]

[a] *Forecasting and Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, 9 Iroon Polytechniou Str, 15773 Zografou Athens, Greece*

[b] *ForLAB, Bangor University, Bangor Business School, Hen Coleg 2.06, Bangor, Gwynedd, LL57 2DG, United Kingdom*

Abstract

When evaluating the performance of time series extrapolation methods, both researchers and practitioners typically focus on the average or median performance according to specific error metrics, such as the Absolute Error or the Absolute Percentage Error. However, from a risk assessment point of view, it is far more important to evaluate the distributions of such errors and especially their tails. For instance, lack of normality and symmetry in error distributions can have significant implications in decision making, such as in stock control. Moreover, these distributions can frequently only be empirically constructed as they may be the result of a computationally intensive non-parametric approach, such as an artificial neural network. In this study, we propose an approach for evaluating the empirical distributions of forecasting methods and use it to assess eleven popular time series extrapolation approaches across two different datasets (M3 and ForeDeCk). The results highlight some very interesting tales from the tails.

Keywords: Forecasting, Performance, Error Distribution, Tails, Risk

1. Introduction

Forecasters have traditionally been evaluating forecasting performance with various error metrics, such as Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE),

[*] Corresponding author
Email address: spiliotis@fsu.gr (Evangelos Spiliotis)

Mean Error (ME) and Mean Absolute Scaled Error (MASE) (Makridakis et al., 1998; Hyndman & Koehler, 2006), all of them having in common that the average performance of the metric is being assessed. On the other hand, practitioners would only use metrics that make sense to them, and this is why MAPE historically prevailed over the rest. Yet, all of these metrics measure the average forecasting performance and, therefore, are seriously influenced by outliers. Recently, many academic articles have been reporting median version of such metrics as well (MdAPE, MdASE, etc.) so as to account for the latter and give a sense of the influence of extreme errors (Nikolopoulos et al., 2016; Davydenko & Fildes, 2013). Quantile regression approaches, as well as trying to reconstruct the full functional form of the forecast rather than a point forecast at any given point of time in the future, constitute other possible approaches, but are much less used in practice (Taylor & Jeon, 2015; Taylor & Yu, 2016).

From a risk assessment point of view, it is far more important to assess the full distributions of forecasting errors and especially their tails. This is because lack of normality and symmetry in the distributions of forecasting errors can have significant implications in decision making, including among others operation management and inventory control. Furthermore, it is frequently impossible to theoretically construct the distributions of the errors as they may be the result of a computationally intensive non-parametric approach, such as an artificial neural network or a complex machine learning method, making their empirical estimation the only way forward.

Swanson et al. (2000) and Coleman & Swanson (2007) discussed some of these issues and proposed two interesting alternative measures, called MAPE-R and MAPE-T, to deal with the exaggeration of the MAPE and make it more robust to outliers. Instead of eliminating large errors or being strongly influenced by them, these metrics exploit a normalization based on the Box-Cox transformation to mitigate their effect while preserving the original information. In this regard, MAPE-R and MAPE-T provide a representative estimator of the median of the APE distribution, but still fail to assess the distribution as a whole. To deal with this problem, in this study we introduce the Risk Measure (RM), a metric which expands the basic properties of MAPE-R/MAPE-T by considering both the median and the deviation of the errors observed. We claim that RM offers significant advantages over other alternatives as it captures both the risk of large errors and the average accuracy of the examined forecasting method, describing the whole error distribution though a single interpretable value.

In order to throw some light into this dimension of forecasting performance, we evaluate the empirical error distributions of eleven popular time series extrapolation methods. First, we consider the Naive (forecasts are equal to the last known observation) and the Naive 2 (seasonally adjusted Naive) methods which are commonly used in the literature as benchmarks. Then, we examine Simple (Gardner, 1985), Holt and Damped Exponential Smoothing (Gardner, 2006), as well as an equal weighted combination of them, which, despite being relatively simple, are robust and frequently difficult benchmarks to beat (Makridakis et al., 2018). The Theta method (Assimakopoulos & Nikolopoulos, 2000), which was the best performing method of the M3 Competition (Makridakis & Hibon, 2000), is also involved together with auto.arima (Hyndman & Khandakar, 2008) and ETS (Hyndman et al., 2002) (from *ExponenTial Smoothing*), two general forecasting algorithms for automatically selecting the best ARIMA and exponential smoothing model, respectively, that have become popular during the last years for displaying accurate results in many applications. Finally, we examine the Bagged ETS (Bergmeir et al., 2016) and the Multi Aggregation Prediction Algorithm (MAPA) (Kourentzes et al., 2014) which exploit bagging and temporal aggregation approaches, respectively, to deal with uncertainty and improve the forecasting accuracy of typical exponential smoothing.

The evaluation is performed in two different datasets, as follows:

- the M3 Competition data
- the ForeDeCk (http://fsudataset.com/) from which the 100,000 time series of the M4 Competition (Makridakis et al., 2018) has been sourced

The remainder of this paper is structured as follows. In Section 2 the review of the respective literature is presented while in Section 3 we describe the methodological approach followed. Section 4 shows the setup of the empirical experiments and then displays and discusses the results. Finally, Section 5 concludes the paper.

2. Literature review

In the operations management and management science literature, the importance of large forecasting errors is often revisited. Syntetos et al. (2010) find out that small gains in forecasting accuracy may lead to considerable inventory reductions, Jakubovskis (2017) stresses the importance of robustness for strategic facility location, capacity acquisition and technology choice decisions, while Ma & Fildes (2017) suggest that accuracy highly affects retail promotional optimization. Yet, the research done in the field of risk and forecasting is still quite limited and few studies propose robust methods capable of mitigating uncertainty.

Phillips (1979) was the first to discuss how the statistical dependence between the parameter estimates and the final period observations used to generate forecasts affect the distributions of the forecasting errors, stressing the importance of utilizing robust forecasting methods. Chen (1997) provides some useful insights regarding the robustness properties of the Holt-Winters, ARIMA, structural component and regression models for the case of seasonal time series. Boudt et al. (2013) discuss robust approaches for applying GARCH models, Cheng & Yang (2015) propose a technique for combining forecasts with outlier protection, while Castle et al. (2015) claim improvements in accuracy using a class of robust devices exploiting equilibrium-correction models. More recently, Athanasopoulos et al. (2017) provided an understanding as to why multiple temporal aggregation leads to improved forecasts and Petropoulos et al. (2018) explained why does bagging for time series forecasting works.

The above mentioned studies examine interesting ways to deal with data, parameter and model uncertainty, boosting the average forecasting accuracy achieved. However, none of them indicates how the risk of large errors is affected nor proposes criteria to assess forecasting performance and facilitate model selection accordingly. Are the gains reported in accuracy due to the reduction of the large forecasting errors or the small ones? Since in many business activities omitting large errors is much more important than being extremely accurate, assessing forecasting performance beyond average accuracy is essential (Ahmadi-Javid et al., 2017; Luo et al., 2018).

The most common way to evaluate the performance of forecasting methods, is to define an error metric or a loss function and use it to compute their average accuracy. Numerous metrics have been proposed in the literature while searching for the "holy grail", i.e. an interpretable metric of fine statistic properties (Makridakis et al., 1998; Fildes, 1992; Armstrong & Collopy, 1992; Davydenko & Fildes, 2013). Scale-dependent measures, percentage errors, relative errors, relative measures and scaled errors are just some examples of the metrics suggested by the forecasting community (Hyndman & Koehler, 2006). Some of them display desirable statistical properties, e.g. are scale independent, less sensitive to outliers, cannot become infinite/undefined or lead to highly skewed results, while others are easily interpreted and widely used in the business world. These issues have long been being discussed, with the researchers coming into a general agreement that the use of percentage errors is highly inadvisable, relative errors and relative measures are more robust but still inappropriate for many problems, and that scaled errors, as well as median operators of the latter, are probably the best existing alternatives (Hyndman & Koehler, 2006; Davydenko & Fildes, 2013; Fildes, 1992). Nevertheless, such metrics provide no evidence regarding deviations around the mean of the forecasting errors, leading to mixed conclusions in terms of risk. This is a key issue, closely related with the way the forecasting errors are usually distributed.

In practice, forecasting errors are non-normally distributed and asymmetric (Coleman & Swanson, 2007). Depending on the characteristics of the data, some series might be easier to predict than others, with forecasting accuracy being among others subjected to their domain (Kang et al., 2017) or even the error metric used (Swanson et al., 2000). Moreover, each method is good at capturing different time series features (Petropoulos et al., 2014), meaning that the errors generated, as well as their distribution, may vary even when the same dataset is used depending on the forecasting method utilized. For instance, Barrow & Kourentzes (2016) examined the forecast error distributions of base and combination forecasts and their implications for inventory control and found out that the latter led to more normal ones, enhancing safety stock calculations. Yet, in both cases the error distributions were far from normal, indicating that the assumption of normality on which inventory control systems are based is frequently violated and that the risk associated with such forecasts might be greater than expected. The necessity of considering the asymmetry of forecasting error distributions was also stressed by Lee & Scholtes (2014) who concluded that prediction intervals that are based on empirical out-of-sample evaluations are more robust and manage to capture uncertainty more effectively.

In this respect, average or even median accuracy metrics may be misleading and appropriate criteria investigating the distribution of the forecasting errors are necessary to enable meaningful comparisons and effectively capture risk. Some types of metrics, such as the squared ones, are often used to deal with this issue by penalizing larger errors. Alternatively, percentiles (e.g. 95% or 99% of the errors observed) are used to approximate the maximum expected error. In the first case, the errors might once again be non-normally distributed among the series leading to questionable results, while also lacking in interpretability. The latter one has first the disadvantage of arbitrary selecting a percentile to describe uncertainty and second the deficiency of considering only the "worst-case" scenario without taking into consideration the level of accuracy typically achieved by the method.

Undoubtedly, the risk of large errors cannot be reliably assessed (Makridakis & Taleb, 2009). Yet, in this study we attempt to provide an approach for doing so. First, we empirically examine the distribution of the errors generated by eleven popular time series forecasting methods across two different datasets: M3 and ForeDeCk. Then, we apply a power transformation to normalize them and exploit our findings to theoretically evaluate their performance considering both the average accuracy and the standard deviation. In this respect, the method that displays the best accuracy, both for typical and rare cases, is effectively identified.

3. Methodological approach

Assessing the risk of large errors when utilizing forecasting methods is a complex task. This of course does depend on the time series extrapolation methods being employed and to that end detailed descriptions of the eleven aforementioned approaches to be used in this study are presented in detail in Section 3.1.

Furthermore, depending on the type of the time series examined, e.g. based on their frequency and domain, different levels of accuracy might be feasible (Kang et al., 2017). Time series characteristics (trend, randomness, seasonality, linearity, etc.) and strategic decisions, such as the forecasting horizon considered, further affect forecasting performance (Petropoulos et al., 2014). Moreover, some methods might be more appropriate for extrapolating specific types of series depending on the information they are capable of capturing (Fildes & Petropoulos, 2015). Last but not least, structural changes, such as variations in the non-stationary nature or the seasonal regularities of the series, may further aggravate the divergence between in-sample and post-sample performance and lead to large errors (Tashman, 2000). In this regard, significant variations might be observed in forecasting accuracy, both between the methods and the time series examined.

The main difficulty in measuring the risk of large errors derives exactly from the variations described above. Generated errors are usually non-normally distributed and asymmetric, meaning that the performance of a method might significantly vary among a dataset depending on the time series being extrapolated or even the metric used (Coleman & Swanson, 2007). As the diversity and the size of the dataset grows, this phenomenon expands further. Thus, using average accuracy measures as an indicator of risk, that do not take into account abnormalities and asymmetries, is an ineffective choice that may lead to mixed conclusions (Swanson et al., 2000). This applies more or less to any error metric considered, even to those of fine statistical properties: The distribution of the errors is not symmetric as it is bounded on the left by zero (perfect forecast) and is usually unbounded on the right ("bad" forecast); thus, the distribution becomes right-skewed, with skewness being determined by the number and values of outliers present.

Figure 1 provides an illustrative example of this issue. Two methods, A and B, are used to predict 100 randomly selected time series of the M3 Competition (Makridakis & Hibon, 2000) and the histograms of their forecasting errors are accordingly constructed. As seen, in both cases the errors of the methods are non-normally distributed and asymmetric, indicating that some time series are indeed more difficult to predict. For instance, we observe that over half of the errors are lower than 10%, but in may cases they exceed 40%. Some notable differences are also identified between the two methods. For example, in most cases method A is more accurate than method B. Moreover, method A tends to generate larger errors than method B (80.64% versus 46.28%), meaning that the latter could be a better alternative for minimizing risk. However, the fat tail observed for the case of method A does not have a significant impact on its average accuracy (12.27%) which is lower than the one computed for method B (12.42%). In this regard, according to the average accuracy metric used (sMAPE), method A would be preferred over B, leading to larger forecasting errors and increasing risk.

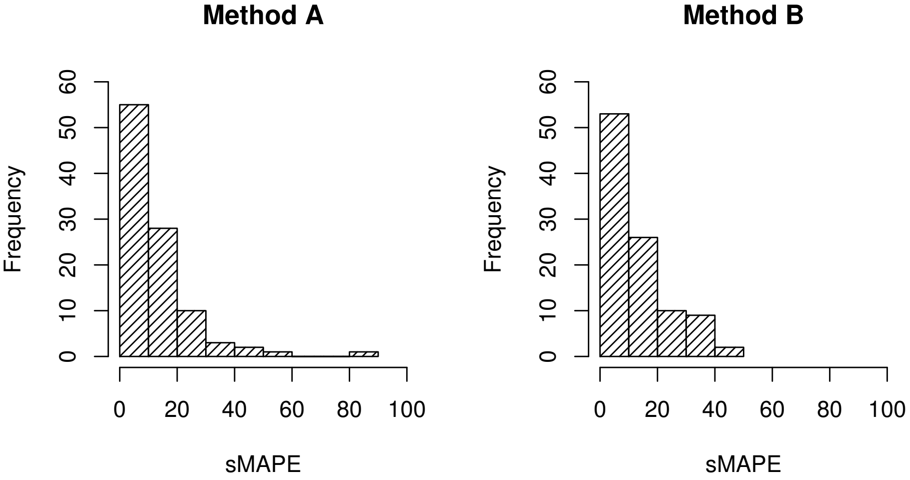

Figure 1: Histograms of the forecasting errors (sMAPE) generated by two different methods across 100 randomly selected time series of the M3 Competition. Lack of normality and symmetry is evident.

Taking everything into consideration, in our point of view (i) exploiting interpretable error metrics to measure the average accuracy achieved per series, (ii) investigating the distribution of the errors computed to properly model their variation and, finally, (iii) considering the accuracy achieved, both on typical and rare cases, are steps in the right direction to effectively evaluate the risk of large errors and improve decision making. The following sections describe how the proposed methodology approaches such an objective.

3.1. Forecasting methods

This section provides details for the time series forecasting methods and respective implementations which has been used for the empirical experiments of this study. The criteria for selecting the methods were mainly (i) their popularity, (ii) the probability of being applied in businesses and organizations, (iii) their ease of use and replicability (e.g. being already implemented in forecasting packages or software), as well as (iv) references in the literature claiming gains in forecasting accuracy by dealing with various sources of uncertainty and potentially reducing the risk of large errors.

In this respect, the forecasting methods considered are chosen broadly to represent both standard and more advanced approaches which demonstrated significant performance in past studies. Moreover, most of them have been considered in well-known forecasting competitions, like the M3 Competition (Makridakis & Hibon, 2000) in which many extrapolation methods were evaluated using a large set of series. All are practical alternatives in commercial applications and statistic software. We note that computer intensive and complex methods, such as neural networks and other machine learning models, were excluded for reasons of simplicity.

The methods used are the following:

- **Naive**: A random walk model and the easiest one to compute. The forecasts are equal to the last known observation of the time series. Naive is mainly used as a benchmark.
- **Naive 2**: Like Naive but the data is seasonally adjusted, if needed, by applying classical multiplicative decomposition by moving averages (Makridakis et al., 1998). A 90% autocorrelation test is performed to decide whether the data is seasonal.
- **Simple Exponential Smoothing (*SES*)**: An exponential smoothing model, aimed at predicting series without a trend (Gardner, 1985). Seasonal adjustments are considered like in Naive 2.
- **Holt Exponential Smoothing (*Holt*)**: An exponential smoothing model, aimed at

predicting series with linear trend (Gardner, 2006). Seasonal adjustments are considered like in Naive 2.

- **Damped Exponential Smoothing (*Damped*)**: An exponential smoothing model, aimed at predicting series with damped trend (Gardner, 2006). Seasonal adjustments are considered like in Naive 2.

- **Combination (*COM*)**: A combination (average) of the three exponential smoothing methods previously described: SES, Holt and Damped, aimed at achieving the possible benefits of averaging the errors of multiple forecasts (Andrawis et al., 2011). Given that many studies claim benefits in risk reduction even when simple combinations are considered (Chan & Pauwels, 2018), this method displays great interest.

- **Theta**: Like Naive 2, time series are first seasonally adjusted if needed. Then, classic Theta is applied for extrapolation, as originally proposed by Assimakopoulos & Nikolopoulos (2000). After predicting, forecasts are seasonally re-adjusted. This method became popular for wining the M3 Competition, outperforming many advanced methods and expert systems, such as ForecastPro and ForecastX. According to the literature, the decomposition approach of Theta can become helpful for identifying long term trends and dealing with data uncertainty.

- **ETS**: Exponential smoothing state space modeling (Hyndman et al., 2002). It is widely used as a general forecasting algorithm as it provides the best exponential smoothing model, indicated through information criteria.

- **Bagged ETS (*bETS*)**: Bootstrapping with aggregation (bagging) using ETS for extrapolating the individual series, as originally proposed by Bergmeir et al. (2016). Bagging has been proven to handle three sources of uncertainty: data, model and parameter uncertainty (Petropoulos et al., 2018). In this respect, evaluating the performance of bETS is of great interest.

- **ARIMA**: Autoregressive integrated moving average models, as implemented in the automated algorithm of Hyndman & Khandakar (2008). A stepwise selection over possible models is performed and the best ARIMA model is returned using appropriate criteria. ARIMA models remain a standard statistical benchmark and, therefore, are considered in this study.

- **Multi Aggregation Prediction Algorithm (*MAPA*)**: A special algorithm for applying multiple temporal aggregation, as originally proposed by Kourentzes et al. (2014). Multiple time series are constructed from the original one using temporal aggregation. In each series, an ETS model is fitted and its respective time series components are forecasted. Finally, the derived components are combined to generate the final forecast. Temporal aggregation helps in strengthening or attenuating the signals of different components, mitigating the importance of model selection. Athanasopoulos et al. (2017) discuss these concepts and evaluate the effect of multiple temporal aggregation when model and/or parameter uncertainty are present.

MAPA is computed using the *mapasimple()* function of the *MAPA* package (Kourentzes & Petropoulos, 2018) for R statistical software (R Development Core Team, 2017). The rest of the forecasting methods are estimated using the *forecast* package (Hyndman, 2017) and the functions *naive()*, *ses()*, *holt()*, *ets()*, *baggedETS()* and *auto.arima()*, respectively. For all of the methods considered the defaults settings are used for reasons of simplicity and reproducibility (Boylan, 2016).

3.2. Selecting an interpretable error metric

As previously discussed, interpretability is mandatory for realizing the level of accuracy achieved in businesses and organizations. Using a benchmark to scale the errors generated by a forecasting method, like done for the case of relative errors, relative measures and scaled errors, may lead to mixed conclusions. For instance, if a method is 20% more accurate than the Naive method, does this mean that the method is accurate enough? Would the conclusion differ if another benchmarks was used as an alternative? How can we confirm in practice that the errors occurred are acceptable (or not) and conclude regarding the risk of generating large errors?

On the other hand, percentage errors are fully interpretable and have the additional advantage of being scale-independent, making them appropriate for comparing forecasting performance across different series and datasets. Undoubtedly, they also come with some drawbacks. To start, they become infinite or undefined if the observed value is zero, or lead to an extremely skewed distribution when the observed value is close to zero. Moreover, they put a heavier penalty on positive errors than on negative ones.

To mitigate these issues, Makridakis (1993) introduced the symmetric Mean Absolute Percentage Error (sMAPE) which is less affected by values close to zero and becomes undefined only if real and forecasted values are both equal to zero. Moreover, sMAPE is bounded to 200%, limiting the skewness of the distribution emerged. sMAPE is defined as follows:

$$sMAPE = \frac{2}{k}\sum_{t=1}^{k}\frac{\left|Y_t - \hat{Y}_t\right|}{|Y_t| + |\hat{Y}_t|} * 100\%, \tag{1}$$

where k is the forecasting horizon, Y_t are the actual observations and $\hat{Y}_t$ the forecasts produced by the model at point t.

Although sMAPE is subject to some cautions, it displays many useful properties compared to other accuracy measures, which are also characterized by similar disadvantages. Moreover, it is widely used for many years and is popular within the forecasting community (Makridakis & Hibon, 2000). In this regard, sMAPE is adopted in this study and used for evaluating the average accuracy achieved per series.

3.3. Investigating the distribution of forecasting errors to evaluate risk

After estimating the accuracy achieved per series according to sMAPE, one must examine how the errors are distributed among the whole dataset and conclude which methods should be utilized to minimize risk. As previously discussed, due to its nature sMAPE is bounded absolutely on the left by zero and partially on the right by 200, leading to right-skewed distributions. Thus, the error distribution emerged is not symmetric, with the degree of skewness being determined by the number and values of outliers, i.e., large errors. The latter are objected to the type of the time series examined (e.g., time series characteristics, frequency and domain), their diversity and the patterns that the forecasting method considered is good at capturing.

In the ideal case that the forecasting errors (u) follow a normal distribution of mean value $\hat{u}$ and deviation σ_u, the risk of generating large errors, noted as Risk Measure (RM), could be defined as follows:

$$RM = \hat{u} + \sigma_u \tag{2}$$

This formula, although quite simplistic, contains some valuable information: First, it determines how accurate the method is on typical cases ($\hat{u}$), indicating that way the average level of accuracy that the method is likely to achieve; Second, it demonstrates how much the error may vary (σ_u) under various sources of uncertainty, illustrating the expected level of accuracy in rare cases; Third, it combines this information in a unique measure, suggesting which method is more appropriate, both for minimizing the risk of large errors and maximizing accuracy.

Note that, in contrast to approaches which exploit percentiles to assess risk, in this case there is no need to determine which percentile will be used in particular as a criteria (e.g. 95%). Thus, comparisons are straightforward and objective.

In practice, however, the assumption of normality is rarely met. In such cases, in order to effectively estimate RM, the shape of the original distribution must be appropriately changed so that the transformed errors come from a symmetric normal distribution. Otherwise, the mean and standard deviation computed will not be representative and proper comparisons will be impossible. This can be achieved by applying a power transformation like the Box-Cox one, as follows:

$$u_{BC} = \begin{cases} (u^{\lambda} - 1)/\lambda, & \text{if } \lambda \neq 0 \\ ln(u), & \text{if } \lambda = 0 \end{cases} \tag{3}$$

where u is the original value of sMAPE, u_{BC} its transformed value and λ the power of transformation. λ is determined to maximize the profile log likelihood (Box & Cox, 1964).

It is noted that $\lambda > 1$ values eliminate negative skewness and $0 < \lambda < 1$ values positive one. Given that this study deals with right-skewed distributions, determined λ values are expected to be lower than 1 and close to zero. The lower the value of λ, the more skewed the original distribution is, and vice-versa. If $\lambda = 1$, then the original distribution is already symmetric and no transformation is required.

Assuming that u_{BC} is normally distributed with a mean value of $\hat{u_{BC}}$ and a deviation of $\sigma_{u_{BC}}$, it is now meaningful to apply equation 2. Moreover, in order RM to be interpretable, the transformed data can be reversed using the same value of λ as follows:

$$RM = \begin{cases} [\lambda(\hat{u_{BC}} + \sigma_{u_{BC}}) + 1]^{1/\lambda}, & \text{if } \lambda \neq 0 \\ e^{(\hat{u_{BC}} + \sigma_{u_{BC}})}, & \text{if } \lambda = 0 \end{cases} \tag{4}$$

In order to better demonstrate the potential benefits of utilizing the suggested approach, the same errors produced by Methods A and B of Figure 1 are normalized using the appropriate Box-Cox transformation ($\lambda = 0.15$). Then, RM is computed and exploited to evaluate their performance in terms of risk. As seen in Figure 2, normality and symmetry have been adequately achieved and meaningful comparisons are now possible as the mean and standard deviation values estimated are quite representative. Moreover, RM of method A (21.76) is higher than that of method B (21.63), indicating that the latter is more appropriate for minimizing the risk of large errors. This is exactly the desirable result.

It should be noted that determining the optimal λ does not guarantee symmetry nor normality. It just helps us generate the distribution which is most likely to be symmetric and normal. In this respect, after determining λ, a nonparametric test must be performed to evaluate the validity of the results. In this study, a two-sample Kolmogorov-Smirnov test (Marsaglia et al., 2003) is used to compare the distance between the Empirical Cumulative Distribution Function (ECDF) of the transformed data and a normal one $N(\hat{u_{BC}}, \sigma^2_{u_{BC}})$. If the p-value computed is high, e.g., $p > 0.05$, one cannot claim statistical support for a

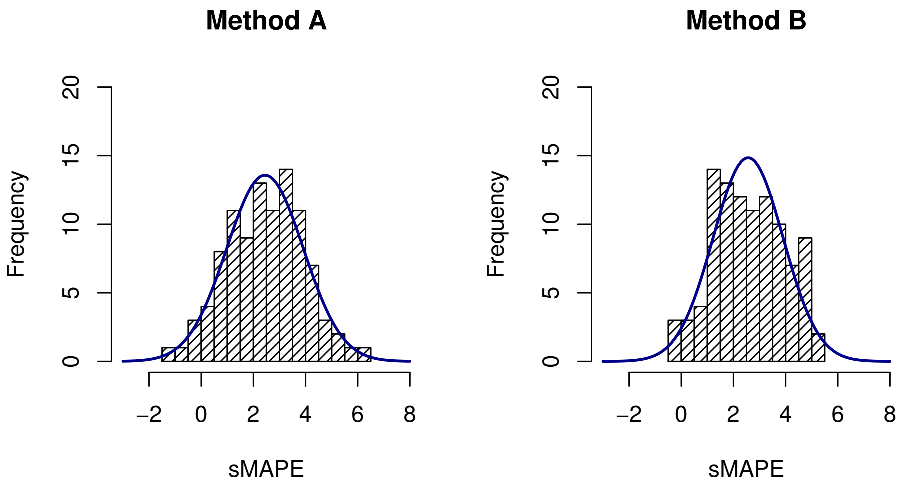

Figure 2: Histograms of the transformed forecasting errors (sMAPE) generated by methods A and B of Figure 1. As seen, normality and symmetry has been adequately achieved. The theoretical normal distribution of each dataset is also presented for reasons of comparison.

difference and the results are assumed valid. In the opposite case, no clear conclusions can be drawn and results should be carefully interpreted.

The methodological approach proposed for evaluating the robustness of forecasting methods is summarized in Figure 3.

Observe that, in contrast to the MAPE-R and MAPE-T metrics proposed by Swanson et al. (2000) and Coleman & Swanson (2007), *RM* is based - for reasons already explained - on the sMAPE and not the MAPE measure and that, apart from capturing the median of the errors, it also considers their standard deviation in order to assess the whole error distribution through a single interpretable value. Note also that the proposed methodology is not necessarily related to the approach of normalizing the original observations using the Box-Cox transformation before extrapolation. As discussed, data and parameter uncertainly, as well as other influencing factors, affect forecasting performance and lead to non-normally and asymmetrically distributed errors. This means that, regardless the forecasting approach used, the generated errors will never be perfectly normal, even if a special pre-processing like the one noted is exploited or a proper forecasting method is used (Barrow & Kourentzes,

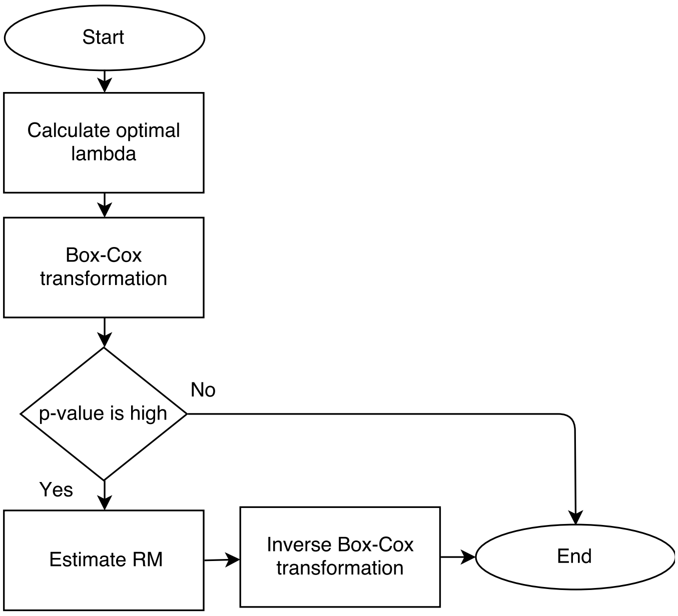

Figure 3: The methodological approach proposed for evaluating forecasting methods beyond average accuracy and identifying those that minimize the risk of large errors.

2016). Yet, since improved parameter estimates lead to better forecasting performance and more normally distributed errors (Phillips, 1979), such approaches may provide less skewed error distributions and, therefore, limit the necessity of utilizing RM to evaluate risk.

4. Experimental design

The first part of this section presents the two datasets which were chosen to demonstrate the use of RM and empirically evaluate the performance of the forecasting methods described in Section 3.1. The second part provides some insights regarding the way the forecasting errors are distributed per case and displays the results.

4.1. The dataset

Two different datasets were chosen to assess the eleven time series forecasting methods considered in this study: M3 Competition and ForeDeCk. The M3 Competition data (Makridakis & Hibon, 2000) is the mostly-used dataset when it comes to business forecasting. This dataset, which includes 3003 time series of five different domains (micro, macro, industry, finance and demographic) and three frequencies (yearly, quarterly and monthly), as well as a small number of series with unspecified domain and frequency (flagged as "other"), was used for the largest forecasting competition that has ever taken place. Its results had a huge impact on forecasting and still inspire a lot of researchers. After almost eighteen years, more than 1000 studies have referenced the results of the competition, and numerous have used its data as a benchmark to evaluate forecasting methods.

However, as recently discussed by Kang et al. (2017), M3 data might not be as indicative of the "real word" as we may think since they display many limitations. For instance, it is known that its time series are unequally distributed according to their domain and characteristics. In this regard, large data collections of diverse time series can help us validate our results and draw safer conclusions.

We decide to exploit ForeDeCk (Forecasting Data Collection), a large dataset of more than 500,000 time series focused on business forecasting applications. Industries, Services, Tourism, Imports & Exports, Demographics, Education, Labor & Wage, Government, Households, Bonds, Stocks, Insurances, Loans, Real Estate, Transportation, as well as Natural Resources and Environment, are some of the domains considered by the dataset. The time series come from trustworthy publicly accessible databases and cover various frequencies, namely yearly, bi-yearly, quarterly, monthly, weekly, daily and hourly. The dataset can be found on-line in *http://fsudataset.com/*.

Yet, forecasting and analyzing half million of time series is extremely time intensive. Thus, a random sample of 10,000 time series is created and used to validate the results of M3. Moreover, to further facilitate comparisons between the two datasets, only yearly, quarterly and monthly series are considered. A summary of the datasets used in this study is presented in Table 1.

As seen, M3 and ForeDeCk display comparable properties as the proportions of their series are quite similar, both per frequency and domain. About half of the data are monthly, while a quarter of them yearly and quarterly. Moreover, most of the time series are of micro, macro and industry domain. A difference can only be identified for the finance data which for the case of ForeDeCk are almost double. Given that the ForeDeCk dataset was constructed

Table 1: Data used for the empirical evaluation. The number of series is displayed per domain and frequency.

Frequency	Micro	Industry	Macro	Finance	Demographic	Other	Total	Total (%)
M3 Competition data								
Yearly	146	102	83	58	245	11	645	*21.5*
Quarterly	204	83	336	76	57	0	756	*25.2*
Monthly	474	334	312	145	111	52	1428	*47.6*
Other	4	0	0	29	0	141	174	*5.8*
Total	828	519	731	308	413	204	3003	-
Total (%)	*27.6*	*17.3*	*24.3*	*10.3*	*13.8*	*6.8*	-	100
ForeDeCk data								
Yearly	647	393	419	712	121	129	2421	*24.7*
Quarterly	654	493	543	579	193	85	2547	*25.5*
Monthly	1173	1048	1014	1154	617	26	5032	*50.3*
Other	0	0	0	0	0	0	0	*0*
Total	2474	1934	1976	2445	931	240	10000	-
Total (%)	*24.7*	*19.3*	*19.8*	*24.5*	*9.3*	*2.4*	-	100

using a random sample of half million time series, its similarity with M3 is very interesting, indicating that some types of data may indeed be more frequently observed in the real word. In this regard, examining the performance of different forecasting across these series, as well as evaluating the risk of large errors, becomes critical.

Note that for the yearly, quarterly, monthly and "other" data, the same rules of M3 Competition are applied, i.e., a forecasting horizon of 6, 8, 18 and 8 points is considered. Thus, direct comparisons between the two datasets examined are possible. The forecasting methods are applied as discussed in Section 3.1 and their performance is evaluated according to RM, as presented in Section 3.3.

4.2. Empirical results

As previously discussed, eleven time series forecasting methods, both standard and more advanced ones, are used to extrapolate the time series of the two datasets presented in this study. Then, sMAPE is applied to measure the accuracy achieved per series. Figure 4 displays the empirical Probability Density Function (PDF) and the Cumulative Distribution Function (CDF) for the case of ETS, both for the M3 and ForeDeCk time series. The rest of the forecasting methods examined display similar distribution functions and therefore are not presented for reasons of simplicity.

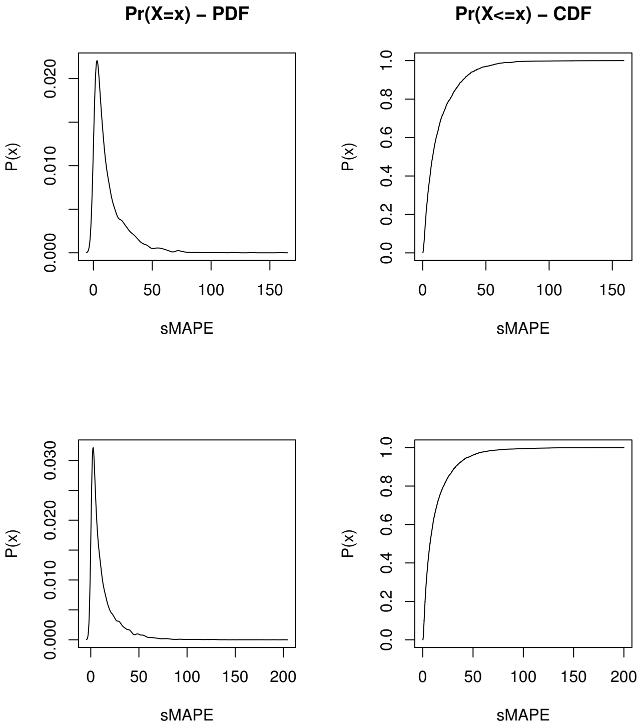

Figure 4: The empirical Probability Density Function (PDF) and Cumulative Distribution Function (CDF) of the ETS method for the case of the M3 (top) and ForeDeCk (bottom) datasets.

As seen, in both datasets the errors are non-normally and asymmetrically distributed, with the majority of the time series being relatively "predictable" and a limited sample of them displaying significantly worse results, probably due to the effect of various sources of uncertainty, such as the data and parameter one. Note for instance that, for the case of

M3, although half of the errors (50% percentile) generated by ETS are lower than 7.8%, in may cases (90% percentile) they exceed 31%, indicating that the nuances of past history may not always persist into the future or effectively captured by the forecasting model considered. This is exactly expected, mainly due to the nature of the error metric used and the particularities of the time series considered that lead to the creation of right-skewed distributions with fat tails. Moreover, we observe that proportionally ForeDeCk consists of more predictable time series compared to M3, although the maximum errors occurred are larger for the former (the 50% and 90% percentiles of sMAPE are 7.2% and 33%, respectively). Thus, despite the forecasting methods perform better on average for the case of ForeDeCk, greater deviations are also observable on their performance affecting the risk of large errors accordingly.

To further investigate the properties of the error distributions emerged, four theoretical CDFs are fitted to the empirical one estimated for the case of ETS and compared with it. This comparison provides some useful insights regarding the way the errors are typically distributed and the law they may follow. As seen in Figure 5, the distribution of sMAPE is far from normal, both for the case of M3 and ForeDeCk time series. Moreover, it displays significant differences when compared to an Exponential or Gamma distribution, mainly due to the effect of the smaller errors observed within the samples. However, the theoretical Lognormal distribution is adequately fitted to the empirical one. This comes into a general agreement with the concepts discussed in Section 3, according to which the utilization of the Box-Cox transformation can effectively change the shape of the original distribution, making it more normal and symmetric. Additionally, according to this finding the λ parameter of the transformation is expected to be close to zero, which is similar to examining the performance of the forecasting methods on a log scale.

Table 2 presents some summary stats on the shape of the distribution considered in Figure 5, confirming the claims made earlier regarding the similarities and differences of the two datasets. The "optimal" shape parameters are selected through the maximum likelihood estimation as implemented in the *fitdist()* function of the *fitdistrplus* R package (Delignette-Muller & Dutang, 2015). The Kolmogorov-Smirnov statistic (KS) and the

Akaike's Information Criterion (AIC) are also displayed to evaluate goodness of fit and, as reported, both criteria indicate that the Lognormal distribution is the preferred one among the candidates.

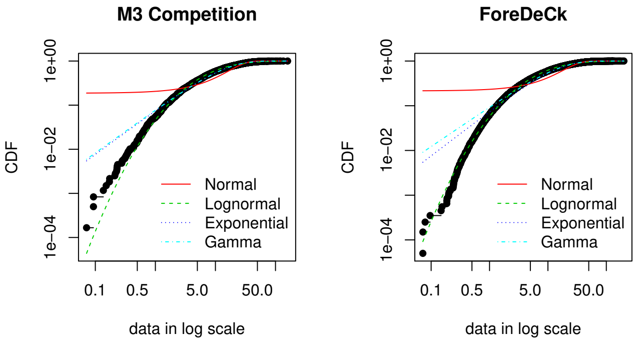

Figure 5: Comparison between the empirical Cumulative Distribution Function (CDF) of ETS's accuracy and four different theoretical ones: Normal (red), Lognormal (green), Exponential (blue) and Gamma (cyan). Both probability and sMAPE values are displayed in log scale.

Table 2: Summary stats on the shape of the empirical Cumulative Distribution Function (CDF) of the ETS's errors considering four different theoretical ones: Normal, Lognormal, Exponential and Gamma. The results are presented both for the M3 and ForeDeCk datasets for reasons of comparison. The Kolmogorov-Smirnov statistic (KS) and the Akaike's Information Criterion (AIC) are also displayed to evaluate goodness of fit.

Dataset	*Normal*		*Lognormal*		*Exponential*		*Gamma*	
	mean	sd	mean	sd	*rate*		shape	rate
M3	13.10	14.70	1.98	1.18	0.08		0.98	0.08
ForeDeCk	13.20	16.70	1.92	1.22	0.08		0.89	0.07
	KS	AIC	KS	AIC	KS	AIC	KS	AIC
M3	0.188	24661	0.035	21406	0.053	21445	0.050	21446
ForeDeCk	0.218	84651	0.026	70671	0.082	71534	0.060	71443

Table 3 presents the forecasting accuracy of the eleven forecasting methods across the M3 and ForeDeCk time series. In both datasets the Theta method is the most accurate one, followed by COM, bETS and MAPA. In contrast, the performance of SES displays the lower deviation, followed by Theta and MAPA. Given that bETS, MAPA and COM incorporate mechanisms for dealing with data, model and parameter uncertainty, it is not surprising that they display top performances both in terms of average accuracy and robustness. Yet, it is quite surprising that SES and Theta, which are way more simplistic, manage to perform better or equally well.

Table 3: Forecasting accuracy achieved for the M3 Competition and ForeDeCk data. The average performance $\hat{u}$ of each method, as well as its deviation σ_u is presented. The 75, 90, 95 and 99 percentiles of sMAPE are also provided.

Method	*M3 Competition*						*ForeDeCk*					
	$\hat{u}$	σ_u	Q(75)	Q(90)	Q(95)	Q(99)	$\hat{u}$	σ_u	Q(75)	Q(90)	Q(95)	Q(99)
Naive	15.70	17.20	20.47	38.18	50.32	75.45	14.62	16.23	19.19	35.25	46.87	78.96
Naive2	14.70	16.60	18.75	35.70	47.70	73.78	14.07	15.93	18.29	34.04	45.48	77.84
SES	13.44	**13.84**	17.82	31.34	41.28	**63.68**	13.64	**15.16**	18.08	32.47	43.45	**71.01**
Holt	14.82	19.11	19.14	35.93	49.18	90.05	14.38	19.60	17.73	35.58	50.85	96.8
Damped	13.06	15.27	17.66	31.23	41.04	69.62	13.23	16.85	16.85	32.63	45.46	79.10
COM	12.93	14.91	17.26	31.18	40.35	67.89	13.04	16.46	16.68	31.86	44.50	76.36
Theta	**12.82**	14.28	**17.11**	31.38	**39.93**	66.01	**12.81**	15.23	16.72	**30.94**	**42.05**	71.79
ETS	13.07	14.68	17.76	31.91	41.37	64.61	13.17	16.70	16.80	32.58	45.18	79.06
bETS	12.95	14.69	17.32	32.02	41.15	68.19	13.04	16.76	**16.59**	31.64	43.35	80.98
ARIMA	13.60	16.04	18.09	33.27	44.15	76.26	13.30	16.67	17.20	33.07	45.03	78.58
MAPA	13.05	14.12	17.83	**31.03**	40.58	67.61	13.01	15.68	17.16	31.59	42.93	72.83

Another finding of the analysis is that, depending on the percentile considered to evaluate the risk of large errors, significantly different conclusions can be drawn. This observation demonstrates the potential benefits of exploiting more objective risk measures, such as the one proposed in this study. For instance, when the 75% of the sample is examined, Theta and bETS display the best performance for the case of M3 and ForeDeCk, respectively. The rank of the methods change for the 90% percentile, where MAPA and Theta perform best. The same stands for the rest of the percentiles examined, which among others indicate that one of the strongest features of Theta and SES is their ability to perform well for the series which are the most difficult to predict, i.e. reduce forecasting errors under conditions of high

Table 4: The risk of large errors for extrapolating the M3 Competition and ForeDeCk time series. The optimal λ parameter of the Box-Cox transformation, the p value of the Kolmogorov-Smirnov test and the *RM* are presented per forecasting method.

Method	*M3 Competition*						*ForeDeCk*					
	λ	p	$\hat{u_{BC}}$	$\sigma_{u_{BC}}$	RM	Rank	λ	p	$\hat{u_{BC}}$	$\sigma_{u_{BC}}$	RM	Rank
Naive	0.10	0.33	2.57	1.36	27.40	11	0.05	0.11	2.30	1.21	25.37	11
Naive2	0.10	0.11	2.48	1.34	25.40	9	0.05	0.13	2.24	1.22	24.34	9
SES	0.10	0.38	2.43	**1.27**	23.33	7	0.05	0.07	2.22	**1.21**	23.66	8
Holt	0.10	0.37	2.35	1.50	25.84	10	0.05	0.10	2.11	1.36	24.59	10
Damped	0.10	0.37	2.27	1.41	22.99	5	0.05	0.08	2.07	1.31	22.83	5
COM	0.10	0.30	2.28	1.39	22.71	2	0.05	0.09	2.07	1.29	22.42	2
Theta	0.10	0.42	2.29	1.36	**22.50**	1	0.05	0.11	2.10	1.25	**22.18**	1
ETS	0.10	0.32	2.27	1.42	23.17	6	0.05	0.10	2.06	1.34	22.92	6
bETS	0.10	0.64	**2.25**	1.42	22.94	3	0.05	0.08	**2.06**	1.32	22.65	3
ARIMA	0.10	0.30	2.28	1.47	24.06	8	0.05	0.05	2.07	1.34	23.29	7
MAPA	0.15	0.47	2.49	1.51	22.99	4	0.05	0.06	2.09	1.29	22.69	4

uncertainty. The robustness properties of these methods significantly affect their average performance and lead to critical gains in forecasting accuracy and risk minimization.

Table 4 presents the results of the proposed methodological approach. This includes the λ values chosen for normalizing the data through the Box-Cox transformation, the p-value calculated by the Kolmogorov-Smirnov test to investigate the normality of the distributions emerged, and finally the *RM* as computed through the mean and deviation of the transformed distributions. As expected, the λ values are relative small across all models and datasets, indicating that the original errors follow a log-normal distribution. Moreover, λ is smaller for the case of ForeDeCk, meaning that, as visualized in Figure 4, greater variations are expected in forecasting accuracy. It is also notable that λ does not significantly change among the methods when the same sample of series is considered. Thus, we conclude that the particularities of each dataset strongly affect the performance of the forecasting methods used, which on the other hand are expected to display similar levels of accuracy per series.

It is also observed that the p-values computed are all higher than the critical threshold set ($p > 0.05$), both for the M3 and ForeDeCk datasets. This demonstrates that the power transformation has been effectively applied for all the forecasting methods considered, making the distributions of the errors initially generated adequately normal and symmetric. Thus, valid conclusions can be drawn if RM is to be used for evaluating the risk of large errors. In general, the p-values are much higher for the case of M3, indicating that the errors originally produced by the methods were more symmetrically and normally distributed compared to those of ForeDeCk and, consequently, that the Box-Cox transformation could be more effectively applied for that case. In other words, as previously noted based on the values of the λ parameter, as well as the shape statistics provided in Table 2, ForeDeCk involves more unpredictable time series, displaying more outliers and complicating the transformation. This reasoning may also partially explain the deviations observed between the p-values of the individual forecasting methods, illustrating that some of them, such as bETS, MAPA and Theta, generate more normally distributed errors than the rest, i.e. are characterized by more robustness properties.

According to RM, Theta is the method that minimizes the risk of large errors, followed by COM, bETS and MAPA. Damped, ETS and ARIMA display close performance, while the rest of the methods significantly worse one. Note that, although on average some methods are more accurate than others ($u\hat{_{BC}}$), due to stronger deviations ($\sigma_{u_{BC}}$) in their accuracy, their final score is significantly affected. It is also encouraging the fact that the ranks of the methods do not change among the datasets, indicating that despite the differences observed between the two samples, results come into a general agreement and objective conclusions are drawn. This conclusion supports among others the use of the M3 Competition data as a standard forecasting benchmark in similar applications and studies.

The RM also helps us validate the conclusions of previous works done in the field, proposing smart ways to deal with data, model and parameter uncertainty and consequently reduce the risk of large errors. For instance, we observe that bETS performs better than simple ETS, demonstrating the benefits of bagging when dealing with various sources of uncertainty. MAPA, which applies ETS on temporally aggregated data, also displays notable gains in robustness, demonstrating the advantages of utilizing temporal aggregation in time series extrapolation. Finally, COM is significantly more robust than the three methods used for averaging their forecasts, verifying the well-known benefits of forecast combination reported in the literature.

An interesting conclusion of this study is that being robust against large errors can play a pivotal role in achieving higher forecasting accuracy, maybe even greater than being more accurate in most of the cases. For instance, according to $\hat{u_{BC}}$ of Table 4, ETS outperforms Theta and COM when common types of time series are examined and limited uncertainty is present. Yet, according to $\sigma_{u_{BC}}$, its performance is less stable than the latter, influencing that way the average level of accuracy achieved. In this regard, identifying methods of good forecasting performance, both for typical and rare cases, can have a major impact on decision making and exploiting the most successful elements and paradigms from the forecasting literature, such as data preprocessing, smoothing, bagging, boosting, temporal aggregation and forecast combination to further enhance forecasting accuracy and minimize risk (Spiliotis et al., 2018) is a step to the right direction.

5. Conclusions

This study assesses the risk associated with using time series extrapolative methods by empirically identifying the ones that minimize the risk of large errors. This is done through a proposed framework which involves constructing the distribution of the generated forecasting errors and applying a Box-Cox transformation to make it more symmetric and resemble a normal one. Then, the mean value and standard deviation of the normalized distribution is computed and used to assess the robustness of the methods considered under various levels of uncertainty.

A major contribution of this work is making available an approach which can be used to effectively measure risk, considering the forecasting accuracy reported both in common and rare cases. Typical approaches are limited in measuring the average or median accuracy achieved through simple error metrics, or the maximum error likely to be observed according to a predefined percentile. These practices display many limitations and may lead to mixed conclusions influenced by error outliers. The suggested approach deals with these issues, enabling straightforward comparisons and facilitating model selection.

Another important conclusion of this study is that methods with robustness properties that display good performance under conditions of high uncertainty, may be more accurate than those performing well for the mainstream majority of the time series. This is because the majority of the series can be adequately extrapolated by any traditional forecasting method, generating relatively low errors. On the contrary, unpredictable series are difficult to deal with and special methods, algorithms and tools are required to improve forecasting performance. Thus, developing forecasting methods that effectively deal with uncertainty and the really difficult to forecast series, could notably reduce the risk of large errors and, therefore, improve decision making under high uncertainty in real life contexts.

For the future we leave the investigation of this approach using more datasets and forecasting approaches that would provide further argumentation for the generalization of the presented empirical results. The sensitivity of the proposed methodology to the existence of possible autocorrelations and heteroscedasticity in the forecasting errors is another fertile area for future research which could expand this analysis from dataset level to time series one.

References

Ahmadi-Javid, A., Jalali, Z., & Klassen, K. J. (2017). Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, *258*, 3–34.

Andrawis, R. R., Atiya, A. F., & El-Shishiny, H. (2011). Forecast combinations of computational intelligence and linear models for the NN5 time series forecasting competition. *International Journal of Forecasting*, *27*, 672–688.

Armstrong, J., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, *8*, 69–80.

Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, *16*, 521–530.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Petropoulos, F. (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research*, *262*, 60–74.

Barrow, D. K., & Kourentzes, N. (2016). Distributions of forecasting errors of forecast combinations: Implications for inventory management. *International Journal of Production Economics*, *177*, 24–33.

Bergmeir, C., Hyndman, R. J., & Bentez, J. M. (2016). Bagging exponential smoothing methods using stl decomposition and boxcox transformation. *International Journal of Forecasting*, *32*, 303–312.

Boudt, K., Danelsson, J., & Laurent, S. (2013). Robust forecasting of dynamic conditional correlation garch models. *International Journal of Forecasting*, *29*, 244–257.

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*, 211–252.

Boylan, J. (2016). Reproducibility. *IMA Journal of Management Mathematics*, *27*, 107–108.

Castle, J. L., Clements, M. P., & Hendry, D. F. (2015). Robust approaches to forecasting. *International Journal of Forecasting*, *31*, 99–112.

Chan, F., & Pauwels, L. L. (2018). Some theoretical results on forecast combinations. *International Journal of Forecasting*, *34*, 64–74.

Chen, C. (1997). Robustness properties of some forecasting methods for seasonal time series: A monte carlo study. *International Journal of Forecasting*, *13*, 269–280.

Cheng, G., & Yang, Y. (2015). Forecast combination with outlier protection. *International Journal of Forecasting*, *31*, 223–237.

Coleman, C. D., & Swanson, D. A. (2007). On MAPE-R as a measure of cross-sectional estimation and forecast accuracy. *Journal of Economic and Social Measurement*, *32*, 219–233.

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

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, *64*, 1–34.

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

Fildes, R., & Petropoulos, F. (2015). Simple versus complex selection rules for forecasting many time series. *Journal of Business Research*, *68*, 1692–1701.

Gardner, E. S. (1985). Exponential smoothing: the state of the art. *Journal of Forecasting*, *4*, 1–28.

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

Hyndman, R., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *26*, 1–22.

Hyndman, R. J. (2017). *forecast: Forecasting functions for time series and linear models*. R package version 8.2.

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

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*, 439–454.

Jakubovskis, A. (2017). Strategic facility location, capacity acquisition, and technology choice decisions under demand uncertainty: Robust vs. non-robust optimization approaches. *European Journal of Operational Research*, *260*, 1095–1104.

Kang, Y., Hyndman, R. J., & Smith-Miles, K. (2017). Visualising forecasting algorithm performance using time series instance spaces. *International Journal of Forecasting*, *33*, 345–358.

Kourentzes, N., & Petropoulos, F. (2018). *MAPA: Multiple Aggregation Prediction Algorithm*. R package version 2.0.4.

Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting*, *30*, 291–302.

Lee, Y. S., & Scholtes, S. (2014). Empirical prediction intervals revisited. *International Journal of Forecasting*, *30*, 217–234.

Luo, J., Hong, T., & Fang, S.-C. (2018). Benchmarking robustness of load forecasting models under data integrity attacks. *International Journal of Forecasting*, *34*, 89–104.

Ma, S., & Fildes, R. (2017). A retail store sku promotions optimization model for category multi-period profit maximization. *European Journal of Operational Research*, *260*, 680–692.

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

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, *16*, 451–476.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, *in press*.

Makridakis, S., & Taleb, N. (2009). Decision making and planning under low levels of predictability. *International Journal of Forecasting*, *25*, 716–733.

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

Marsaglia, G., Tsang, W. W., & Wang, J. (2003). Evaluating Kolmogorov's distribution. *Journal of Statistical Software*, *8*, 1–4.

Nikolopoulos, K. I., Babai, M. Z., & Bozos, K. (2016). Forecasting supply chain sporadic demand with nearest neighbor approaches. *International Journal of Production Economics*, *177*, 139–148.

Petropoulos, F., Hyndman, R. J., & Bergmeir, C. (2018). Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European Journal of Operational Research*, *268*, 545–554.

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

Phillips, P. C. (1979). The sampling distribution of forecasts from a first-order autoregression. *Journal of Econometrics*, *9*, 241–261.

R Development Core Team (2017). *R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria.*

Spiliotis, E., Assimakopoulos, V., & Nikolopoulos, K. (2018). Forecasting with a hybrid method utilizing data smoothing, a variation of the theta method and shrinkage of seasonal factors. *International Journal of Production Economics*, *in press*.

Swanson, D., Tayman, J., & Barr, C. (2000). A note on the measurement of accuracy for subnational demographic estimates. *Demography*, *37*, 193–201.

Syntetos, A. A., Nikolopoulos, K., & Boylan, J. E. (2010). Judging the judges through accuracy-implication metrics: The case of inventory forecasting. *International Journal of Forecasting*, *26*, 134–143.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, *16*, 437–450.

Taylor, J. W., & Jeon, J. (2015). Forecasting wind power quantiles using conditional kernel estimation. *Renewable Energy*, *80*, 370–379.

Taylor, J. W., & Yu, K. (2016). Using autoregressive logit models to forecast the exceedance probability for financial risk management. *Journal of the Royal Statistical Society*, *179*, 1069–1092.