International Journal of Forecasting 36 (2020) 37–53

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Are forecasting competitions data representative of the reality?

Evangelos Spiliotis [a,*], Andreas Kouloumos [a], Vassilios Assimakopoulos [a], Spyros Makridakis [b]

[a] Forecasting and Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece
[b] Institute For the Future (IFF), University of Nicosia, Nicosia, Cyprus, Cyprus

ARTICLE INFO

Keywords:
Forecasting competitions
Time series visualization
Time series features
M4
Forecasting methods

ABSTRACT

Forecasters typically evaluate the performances of new forecasting methods by exploiting data from past forecasting competitions. Over the years, numerous studies have based their conclusions on such datasets, with mis-performing methods being unlikely to receive any further attention. However, it has been reported that these datasets might not be indicative, as they display many limitations. Since forecasting research is driven somewhat by data from forecasting competitions, it becomes vital to determine whether they are indeed representative of the reality or whether forecasters tend to over-fit their methods on a random sample of series. This paper uses the data from M4 as proportionate to the real world and compares its properties with those of past datasets commonly used in the literature as benchmarks in order to provide evidence on that question. The results show that many popular benchmarks of the past may indeed deviate from reality, and ways forward are discussed in response.

1. Introduction

Researchers typically evaluate the performances of new forecasting methods empirically by exploiting well-known datasets that have been examined previously in the literature. This allows the methods proposed in past studies to be used as benchmarks directly, and makes the replication of results much more convenient (Makridakis, Assimakopoulos & Spiliotis, 2018). Another reason for this practice is that the continuous use of such datasets has created the general belief that they represent reality and lead to reliable conclusions that will hold under real conditions. The advantages of using popular datasets for assessing the forecasting accuracy, and especially the confidence that these inspire, have made their use almost mandatory over the years, and led to the assumption that any method that does not perform well across their series should not receive any further attention or be adopted.

However, this widespread approach may also have some critical drawbacks, associated mainly with the limited diversity of the datasets and their relatively small size. If the datasets used in the literature do not represent reality, this calls into question the conclusions of studies that use them. Moreover, given that new methods are forced to improve forecasting accuracy across them, it may be that the present forecasting research focuses on over-fitting methods to specific samples of data rather than solving real world problems.

When it comes to business forecasting, the most commonly used dataset is the M3 competition data (Makridakis & Hibon, 2000). This dataset includes 3,003 series from five different domains (micro, macro, industry, finance and demographic) and of three frequencies (yearly, quarterly and monthly), as well as a small number of series with unspecified domains and frequencies (flagged as

* Corresponding author.
E-mail address: spiliotis@fsu.gr (E. Spiliotis).

https://doi.org/10.1016/j.ijforecast.2018.12.007

"other"). The competition had a huge impact on forecasting and still inspires a lot of researchers. Almost twenty years down the track, more than 1,000 studies have referenced the results of the competition, and numerous have used its data to evaluate other forecasting methods. Given the influence of data from past forecasting competitions, and taking into account the concerns expressed earlier, examining their representativeness becomes critical.

The idea of questioning the data used in forecasting competitions is not new. For instance, it is known that the M3 series are distributed unequally according to their type, with the majority of them being of a monthly frequency and in the micro/macro domains. As a result, any methods that perform well on the larger subsets that arise will be favored over others. This conclusion was reported by Clements and Hendry (2001) and Ord (2001), who commented that the data may not be representative even for the domains considered, and therefore any results should be criticized before generalizing. Thus, Lawrence (2001) suggested that different methods should be preferred for each type of data and Petropoulos, Makridakis, Assimakopoulos, and Nikolopoulos (2014) worked on that direction by introducing an approach for model selection that was based on seven time series features and one strategic decision. More recently, Kang, Hyndman, and Smith-Miles (2017) proposed a method for visualizing data collections of time series in a two-dimensional instance space, in the process identifying some gaps for the case of M3 and representing the effectiveness of different forecasting methods across its space. No doubt similar conclusions could be true for other competitions in the past.

So what is the best way of improving forecasting performances? Should researchers continue to use datasets from past forecasting competitions or seek for other, more representative, ones? If there are some types of time series that we identify more frequently in the real world, such as the M3 dataset, then we might need to keep looking for a dominant method across such benchmarks. If not, we might need to develop special methods for the individual sub-sets that arise or to assess their performances using more appropriate datasets.

Although the suggestion that forecasting competition data might be inadequate for the effective evaluation of forecasting methods appeared quite early, no clear evidence has been provided that such is actually the case. This paper investigates the issue by applying the instance space analysis of Kang et al. (2017) to the M4 data, a large dataset of 100,000 time series sampled randomly from a set of 900,000 series that were collected from multiple and diverse sources. The properties of this dataset, which has been exploited recently for the latest and largest M competition (Makridakis, Spiliotis & Assimakopoulos, 2018), are compared with those of M3 and other popular benchmarks that are used commonly in the forecasting literature. Given the design of M4, this process can be considered equivalent to examining the similarity of data from past forecasting competitions to the real world. The results show that the majority of the well-known benchmarks considered may not represent reality, since, with the exception of M3, their features and their multivariate distributions display some notable variations. Thus, after comparing reality with data from various past forecasting competitions, the factors that influence forecasting performances are demonstrated and ways forward are discussed in order to advance the theory and practice of forecasting.

2. Time series instance spaces

Analyzing the types of numerous time series is a complex task, due mainly to their temporal nature and to differences in length. A smart way of dealing with this issue is by defining indicative and independent statistics for measuring their main features, thus enabling each series to be represented as a static point in a high-dimensional feature space (Hyndman, Wang, & Laptev, 2015). This practice has been used widely for the effective classification and clustering of diverse time series (Wang, Smith, & Hyndman, 2006), with the number and character of the features depending on the application considered (Nanopoulos, Alcock, & Manolopoulos, 2001).

Recently, Kang et al. (2017) exploited the work that has been done in the field and proposed an algorithm for visualizing time series in a two-dimensional (2D) instance space through dimension reduction techniques. This can be shown to be extremely helpful when investigating the properties of a dataset, since, although the series continue to be represented by a vector of n features — originally visualized in a nD space — it becomes possible for humans to inspect them directly in the observable space (i.e., up to 3D, to which we are confined). For example, instances which are close to each other in the 2D space will have similar values across the n features considered, making the detection and analysis of homogeneous data structures much more convenient.

More specifically, Kang et al. (2017) developed a set of six features $F_1, F_2, \ldots, F_6$ that were selected specially because of their ability to provide useful information about the M3 data. The authors used *spectral entropy* (F_1) for measuring the "forecastability" (randomness) of the series, *strength of trend* (F_2) for measuring long-term changes in the mean level, *strength of seasonality* (F_3) for measuring the influence of the seasonal factors, *seasonal period* (F_4) for indicating the length of the periodic patterns in the series, *first order autocorrelation* (F_5) as a measure of the linear relationship between observations, and finally, the *optimal Box–Cox transformation parameter* (F_6) for determining the stability of the series. For more details about these features and how they are computed numerically, please see the original paper.

After estimating the features of the M3 series and computing their feature vectors $\mathbf{F} = (F1, F2, F3, F4, F5, F6)'$, a principal component analysis is used to reduce the dimension of the feature space. Moreover, since the first two principal components with the largest eigenvalues retain almost 70% of the variation in the data, it is reasonable to use their values to visualize the regions of the 2D space that are covered by the M3 data.

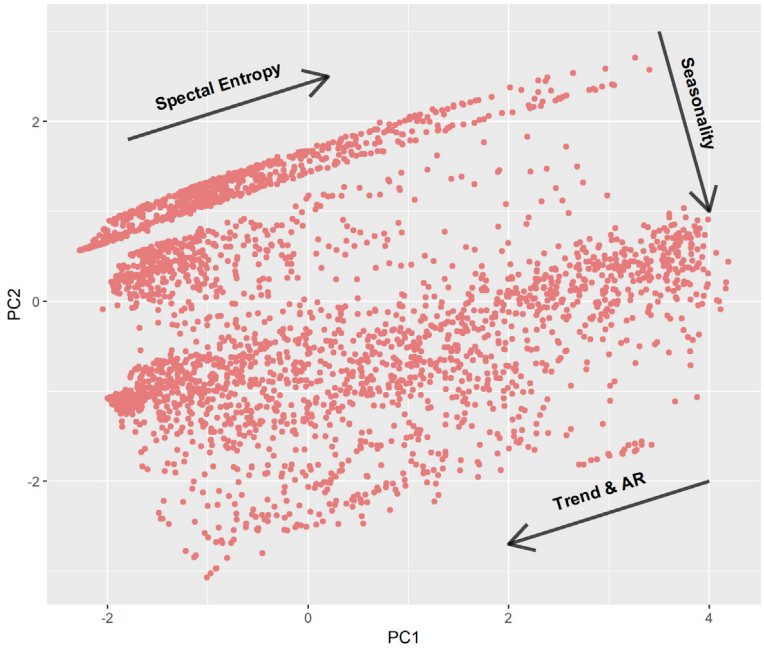

Fig. 1. Instance space of the M3 time series, as proposed by Kang et al. (2017).

The scatterplot of Fig. 1 presents the instance space of the M3 time series using the first two principal components of the methodological approach of Kang et al. (2017), which can be expressed algebraically as:

$$\begin{bmatrix} PC1 \\ PC2 \end{bmatrix} = \begin{bmatrix} 0.54 & -0.54 & 0.24 & 0.25 & -0.54 & -0.11 \\ 0.26 & -0.14 & -0.64 & -0.67 & -0.17 & -0.14 \end{bmatrix} \mathbf{F}$$

Three clusters of data are observable based on the seasonal period of the series: non-seasonal (at the top), quarterly (in the middle) and monthly (at the bottom). Moreover, within each cluster of time series the strength of the seasonality increases as we move from top to bottom. Relatively, the spectral entropy of the data increases from left to right, while the trend and autocorrelation increase from right to left. Thus, instances which are close to each other in this space have similar values across the six features examined. The variability of F_6 cannot be observed in the 2D space, as it requires an additional component to be presented (3D space). It is notable that gaps can be identified in the instance space, indicating for example that it is unlikely that quarterly data will be of low forecastability, or that monthly ones will display low forecastability and high seasonality simultaneously. In addition, some other types of data are observed more frequently, meaning that, as has been reported in the literature, the M3 series are distributed unequally across their features.

3. Using M4 data as proportionate to the "real world"

The instance space analysis of the M3 illustrates that its time series are distributed unequally across their features, with some parts of the space being blank. The same might be true for the space of data from other past forecasting competitions. However, before questioning the use of such datasets in future studies, one should prove that reality is characterized by significantly different properties. The main problem with providing such evidence is that obtaining a complete picture of the "real world" is impossible in practice, due to its unlimited applications and types of data involved. Nevertheless, if a large collection of real and indicative data was available, it would be reasonable to exploit it in order to provide evidence regarding this argument.

In this regard, the present study exploits the M4 data, a large dataset of 100,000 time series that was used recently to support the M4 competition (Makridakis, Spiliotis et al., 2018), the successor to the three previous M competitions and the largest that has ever taken place. This dataset was created on December 28, 2017, when Professor Makridakis chose a seed number to randomly sample ForeDeCk (http://fsudataset.com/), a collection of data consisting of almost 900,000 series from the business world. These series were retrieved from a variety of reliable open-access databases. Some of the domains considered by ForeDeCk were Industries, Services, Tourism, Imports & Exports, Demographics, Education, Labor & Wage, Government, Households, Bonds, Stocks, Insurances, Loans, Real Estate, Transportation, and Natural Resources and Environment. The result of this process was a diverse dataset involving series from six different frequencies and domains, as presented in Table 1.

Several pre-defined filters were also applied beforehand in order to achieve some desired characteristics, relating mainly to the length of the M4 series and their proportions per frequency and domain. For instance, series with fewer than 10 observations, or alternatively fewer than three periods of data, were excluded in order to make sure that an adequate number of observations

Table 1
The M4 competition data used as indicative of real world proportions. The numbers of series are displayed per frequency and domain.

Frequency-Domain	Demographic	Finance	Industry	Macro	Micro	Other	Total
Yearly	1,088	6,519	3,716	3,903	6,538	1,236	23,000
Quarterly	1,858	5,305	4,637	5,315	6,020	865	24,000
Monthly	5,728	10,987	10,017	10,016	10,975	277	48,000
Weekly	24	164	6	41	112	12	359
Daily	10	1,559	422	127	1,476	633	4,227
Hourly	0	0	0	0	0	414	414
Total	8,708	24,534	18,798	19,402	25,121	3,437	100,000

was available for training and testing the participating methods (a minimum of three periods is typically required for calculating the seasonality component effectively). Similarly, series with missing values were excluded in order to avoid the undesirable effect of unavailable observations. Note that methods dealing with missing data and intermittent demand are always tested on datasets other than the M ones, as the competition series are continuous and greater than zero. Thus, the conclusions of this study are expected to hold mainly for business and fast-moving data; i.e., the forecasting applications for which the M competitions have traditionally served as benchmarks.

We would also like to seize this opportunity, by describing in detail how M4 was constructed, to provide some guidelines to facilitate the designing of forecasting competitions in the future, an abstract and difficult-to-automate process which should be performed carefully in order to lead to representative results. These guidelines can be summarized as follows:

1. Construct a large collection of data, including time series that are relevant to the desired applications (frequency and domain).
2. Define which of the applications considered are most important and set their priorities accordingly.
3. Determine the total size of the dataset.
4. Given the priorities and the dataset size, approximate the numbers of series that should be considered for each application.
5. Randomly sample the available series so that the desired amount of data is obtained in each case.
6. Combine the subsets thus created in order to construct the final dataset.

As can be seen, defining the applications (domain and frequency of data) for which the forecasting methods should excel is the most important decision that needs to be made, followed by determining their proportions in the final dataset. For instance, if the objective of the competition is to identify a robust method for general use, then using a data collection like M4 could be a great option. However, if the objective is to identify the best-performing method for specific types of data, such as energy data, another special sample could be a more appropriate choice. Likewise, larger samples of higher frequency data (weekly, daily and hourly) than those considered in M4 might be required if one is dealing with operation problems rather than strategy ones.

In particular, since the aim of M4 was to learn how to improve forecasting performances in businesses and organizations, a large data collection (ForeDeCk) of series of relevant types was first identified so as to cover the applications reported in Table 1. Next, the total size of the dataset was determined (100,000), with the objective of making it large enough to provide statistically significant results. Given the size of the competition's dataset and the priority of each subset, provided in Table 2, the number of series to be considered in each case was estimated and random sampling was performed in order to reach it. Finally, the constructed subsets were combined to create the final dataset for the competition.

Note that the priorities in Table 2 were determined based on how likely it is that a company or an organization will be generating forecasts for each of these individual applications. For instance, when dealing with business forecasting, monthly forecasts are required more frequently than quarterly and yearly ones. Similarly, micro and financial data are more likely to be used to support decision-making than demographic data. Thus, if the priorities "Very Low", "Low", "Medium", "High" and "Very High" account for 0, 1, 2, 3 and 4 points, respectively, Table 2 shows that the 100,000 time series of M4 correspond to a total of 40 points, each representing a sum of 2500 series. As a result, the M4 dataset would require, e.g., approximately 5000 quarterly-finance series, 47,500 monthly ones, and so on.

Once again, we would like to note that considering M4 data as proportionate to the "real world" does not mean that we assume that it covers every single type of business series. Variations on the time series characteristics are possible depending on the forecasting application considered. However, given the number and diversity of the series included in M4, we feel that the major properties of the reality can be captured adequately, leading to representative conclusions and providing useful insights into the research question examined. After all, to the best of our knowledge, no other datasets of comparable size, diversity and quality exist at present. Thus, we believe that M4 could be used in the future as a diverse benchmark when evaluating time series forecasting methods, serving as the standard testing ground for evaluating theoretical developments, assessing new methods and determining their added value.

4. Comparison of past forecasting competitions data with the "real world"

Having assumed that the M4 dataset is proportionate to the "real world", the algorithm of Kang et al. (2017) is applied to its series in order to quantify their features,

Table 2
The priority considered for each forecasting application in terms of the data frequency and domain when constructing the M4 dataset.

Frequency-Domain	Demographic	Finance	Industry	Macro	Micro	Other
Yearly	Low	Medium	Low	Medium	Medium	Low
Quarterly	Low	Medium	Medium	Medium	Medium	Low
Monthly	Medium	Very High	Very High	Very High	Very High	Low
Other	Very Low	Very Low	Very Low	Very Low	Low	Low

Note: The probability of generating relevant forecasts in a business environment is used as a criterion.

Table 3
Distribution of the Kang et al. (2017) features for the M3 and M4 time series.

Frequency	Forecastability	Trend	Seasonality	Linearity	Stability	Observations
M3 Competition						
Yearly	0.71 ± 0.12	0.90 ± 0.17	0.00 ± 0.00	0.69 ± 0.21	0.50 ± 0.44	819
Quarterly	0.69 ± 0.13	0.95 ± 0.10	0.34 ± 0.35	0.70 ± 0.24	0.47 ± 0.44	756
Monthly	0.73 ± 0.18	0.74 ± 0.28	0.33 ± 0.27	0.56 ± 0.34	0.47 ± 0.42	1,428
Hourly	–	–	–	–	–	0
M4 Competition						
Yearly	0.66 ± 0.15	0.93 ± 0.12	0.00 ± 0.00	0.70 ± 0.23	0.41 ± 0.40	27,586
Quarterly	0.63 ± 0.14	0.95 ± 0.10	0.21 ± 0.26	0.76 ± 0.23	0.40 ± 0.41	24,000
Monthly	0.60 ± 0.17	0.88 ± 0.19	0.33 ± 0.29	0.80 ± 0.25	0.43 ± 0.43	48,000
Hourly	0.57 ± 0.09	0.69 ± 0.30	0.89 ± 0.21	0.87 ± 0.12	0.42 ± 0.41	414

Note: The mean of each feature is followed by the standard deviation observed.

visualize them in space and compare M3 with "reality". However, given that the original algorithm mostly examines basic time series characteristics, additional features are considered later so as to support the conclusions made better (Hyndman et al., 2015). Finally, a similar comparison is performed with other popular past forecasting competitions, namely M1 (Makridakis, Andersen, Carbone, Fildes, Hibon, Lewandowski, Newton, Parzen, & Winkler, 1982), M2 (Makridakis, Chatfield, Hibon, Lawrence, Mills, Ord, & Simmons, 1993) and T (Athanasopoulos, Hyndman, Song, & Wu, 2011).

4.1. Exploiting basic time series features

We now compare the properties of the two largest time series forecasting competition datasets, namely M4 and M3, using some basic time series features, as was proposed originally by Kang et al. (2017). A summary of the results is presented in Table 3, which provides the mean value of each feature and its standard deviation across the sample for both datasets. For ease of interpretation and to mitigate any effects due to frequency variations, the results are displayed separately for each seasonal period. Note that, since the "other" data of the M3 and the "weekly" and "daily" data of M4 are non-seasonal, their subsets are merged with that of the yearly series.

The results of Table 3 indicate that the two datasets have quite similar properties, as the values of the features are distributed around similar centers with comparable variances in most of the cases. The main differences observed are that the M4 series are less stable across all frequencies, the M3 yearly series are less forecastable, the M3 quarterly series are more seasonal and less linear, and the M3 monthly series are less forecastable, trended and linear.

Table 3 also provides a clear picture of the time series characteristics that hourly and other high-frequency data may typically display. For instance, by their very nature, the strength of seasonality is dominant, being almost three times as large as those of monthly and quarterly data. Moreover, the hourly data are significantly less trended than the rest of the series, being also rather forecastable and linear. In this respect, one would expect forecasting models that are effective at handling complex seasonal patterns and capturing the running level of the series to perform well for such types of data.

The distribution of the features is studied further in Fig. 2, where kernel density estimates are implemented in order to assess the diversity between the two datasets for each frequency. In addition, the Kullback–Leibler divergence (KL) (Cover & Thomas, 1991) is used to report potential differences in each case and to determine whether these are significant. The KL between two distributions $p(x)$ and $q(x)$ is defined as:

$$KL = \sum p(x) log_2 \frac{p(x)}{q(x)}, \tag{1}$$

where p is the target (true) distribution (in our case, that of M4) and q is the one used to approximate it (here M3), being always nonnegative and zero if and only if $p = q$.

Note that KL is not symmetric and therefore cannot be used as a clear distance measure. However, it can be effective in reporting the entropy increase due to the use of an approximation to the true distribution rather than the true distribution itself. In practice, KL measures the average number of additional bits required to describe the information in p, given q. In this regard, if $H(p)$ is the entropy of the target distribution, the approximate distribution will differ by about $\frac{KL}{H(p)} * 100(\%)$. Both KL and the percentage difference are reported for each feature in Fig. 2.

After examining the density estimates visually, we find out that the features of the two datasets are not only characterized by comparable means and variances, but

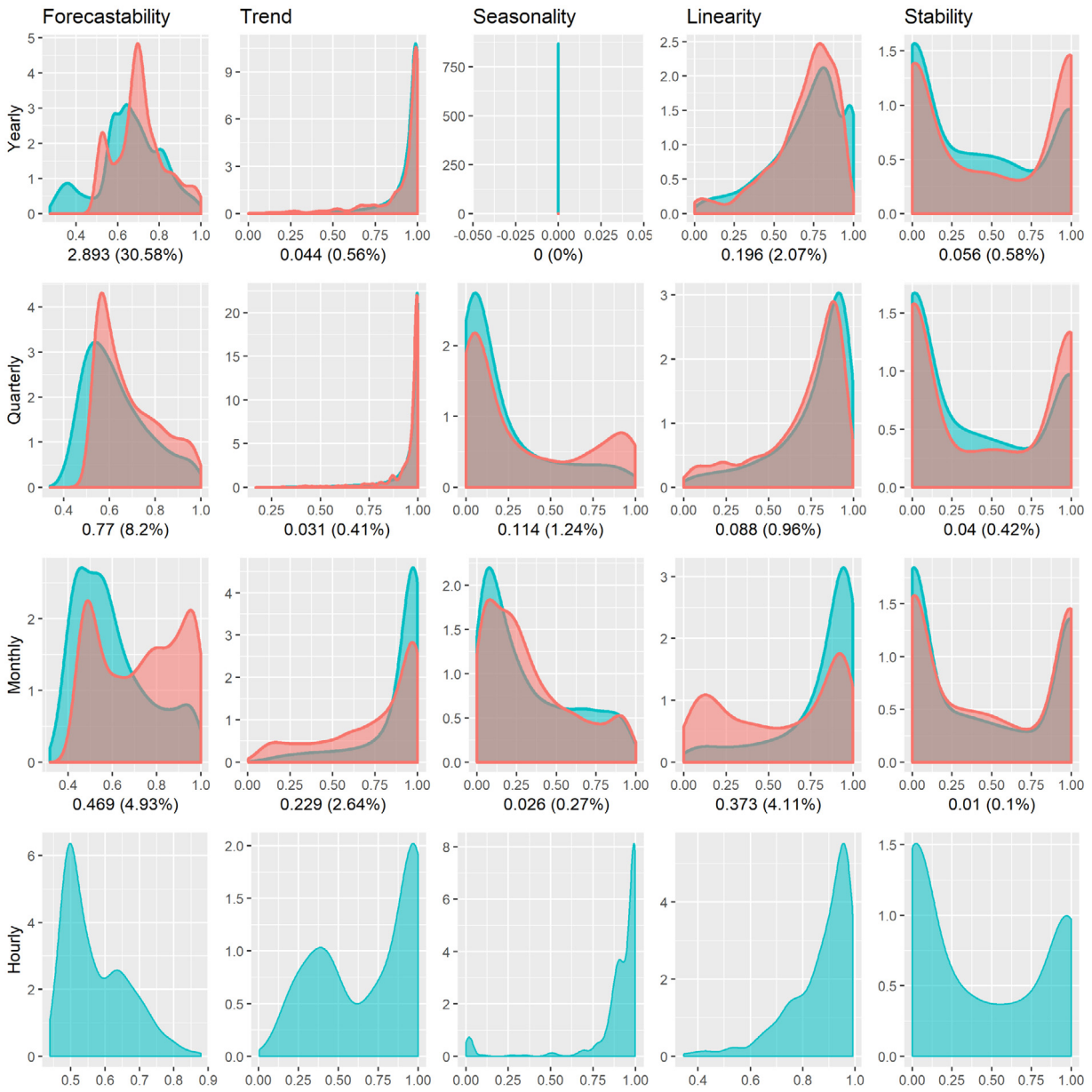

Fig. 2. Kernel density estimates for the Kang et al. (2017) features for the case of M3 (red) and M4 (blue) time series. The distributions are displayed for each frequency, and the Kullback–Leibler divergence is used to report their absolute and percentage differences. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

also distributed in a similar manner. This conclusion is also supported by the estimated KL values, which demonstrate that the differences between the two datasets are insignificant for the majority of the features examined. In fact, the feature of spectral entropy is the only one that reports a KL percentage difference of greater than 5%, which is observed only for the cases of the yearly and quarterly data.

However, the time series features are usually correlated. For instance, trended time series are typically linear, while seasonal ones tend to be less noisy and easier to forecast. Therefore, analyzing one single feature independently of the others may produce biased results. In this regard, the multivariate distribution of the features is also taken into consideration so as to ensure that the two datasets examined behave similarly both for each feature and for each high-dimensional component. To do so, the analysis performed earlier is applied to the first three principal components of the feature space, which are estimated using the combination of the M3 and M4 datasets (103,003 series) as an input, to avoid the potential for a limited or bounded PC space. The first three components of the combined dataset, which retain almost 83% of the variation in the data, are expressed algebraically below, while their distributions are visualized for each frequency

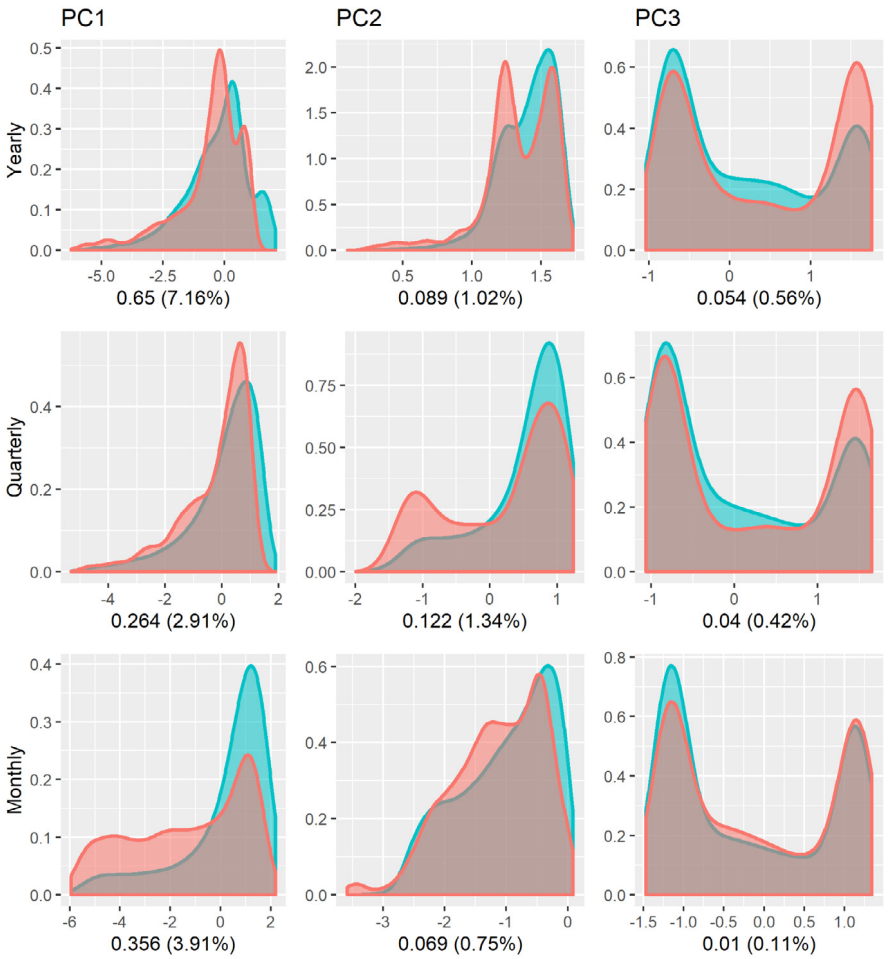

Fig. 3. Kernel density estimates for the first three principal components of the Kang et al. (2017) feature space for the case of the M3 (red) and M4 (blue) time series. The components are estimated using a combination of the M3 and M4 datasets (103,003 series) as input. The distributions are displayed by frequency, and the Kullback–Leibler divergence is used to report their absolute and percentage differences. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in Fig. 3:

$$\begin{bmatrix} PC1 \\ PC2 \\ PC3 \end{bmatrix} = \begin{bmatrix} -0.59 & 0.55 & 0.05 & 0.07 & 0.58 & 0.00 \\ 0.02 & 0.21 & -0.67 & -0.70 & -0.03 & -0.16 \\ -0.01 & 0.06 & -0.02 & -0.18 & -0.05 & 0.98 \end{bmatrix} \mathbf{F}.$$

According to the components, the yearly data from the two datasets display similar multivariate distributions, especially for PC2 and PC3, where the components of seasonality and stability are dominant, respectively. The differences observed for PC1 are caused mainly by the fact that the M3 yearly series are less forecastable and trended. Thus, PC1 obtains slighter lower values in M3 according to the algebraic expression of the components. The differences are even less noticeable for the quarterly and monthly data, where, once again, some minor differences are observable only for the case of PC1. In this regard, when examining the KL values of the density estimates, one may conclude that the multivariate distributions of the two datasets do not differ significantly, and that, in general, the features of the M3 and M4 series are interrelated in a common manner. Thus, based on the six features examined, and despite the differences detected between features in some cases, we conclude that the M3 data are representative of the reality.

It should be noted at this point that, when applying a linear dimension reduction method like PCA, emphasis is placed on keeping dissimilar data points far apart in the low-dimensional space, rather than on keeping the similar ones close together. This approach might result in misleading conclusions when numerous features are being considered or complex nonlinear correlations are present, meaning that the use of nonlinear techniques could be a better alternative in such cases (Kang, Hyndman, & Li, 2018). On the other hand, nonlinear methods do not allow us to understand clearly how the high-dimensional space is constructed, nor which interactions are present. Since this study considers only a limited set of basic time series features for examining the factors that cause any possible variations from reality, PCA is regarded as an effective

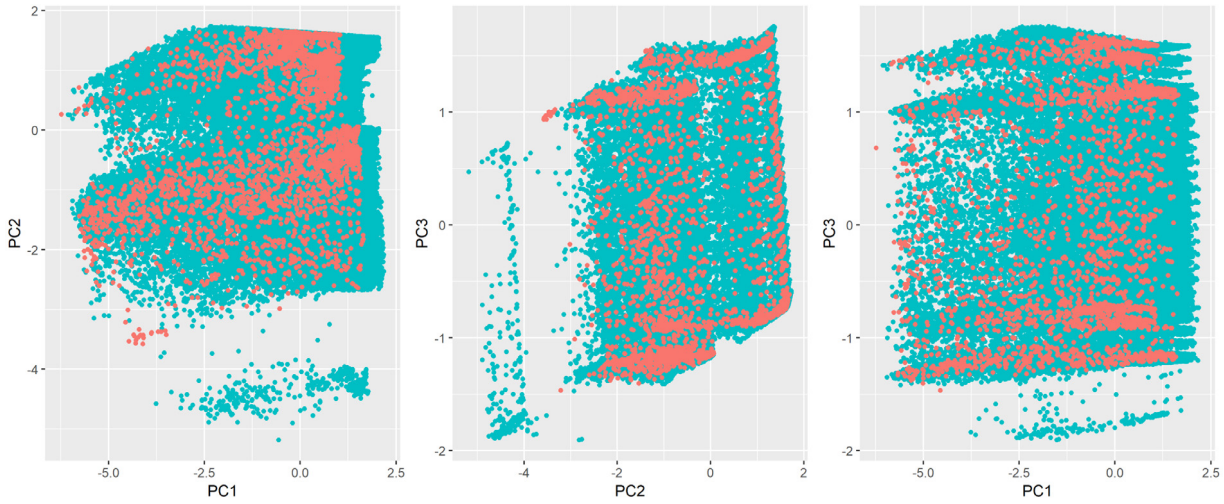

Fig. 4. The instance space of the M3 competition (red) compared to that of the M4 (blue). The six time series features of Kang et al. (2017) are used to construct the space, while the combined dataset of the M3 and M4 series is used for estimating the principal components. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

solution to the problem, and no further analysis is performed. However, the utilization of nonlinear methods for exploring instance spaces is definitely a fertile area for future research.

We enable a better visualization of the properties of the two datasets in total and investigate possible different time series structures by generating the instance space of M3 and M4 and presenting it in Fig. 4 using pairwise scatterplots for the three most significant components of the analysis. The plots are obtained by projecting M4 and M3 data onto the feature space of their combined dataset.

As can be seen in scatterplot PC1-PC2, although M4 includes more highly-trended time series of low spectral entropies (non-overlapping points on the right), the general picture obtained is approximately the same for both sets, meaning that gaps that have been identified previously for the M3 data remain empty. Thus, although M4 provided more time series for spaces that previously were filled only sporadically (e.g., quarterly data of low forecastability), there are still some "uninhabited" territories that time series tend to avoid. Actually, a close look at the instance space shows that this conclusion might be quite reasonable after all. For example, we see for both datasets that few time series "live" on the mid-left of the space (low PC1 - mid PC2), which represents noisy data of low trend and relatively high seasonality. If noise is high, then any seasonal patterns that may exist are difficult to observe, and vice versa, forcing a relative gap in the instance space. Note that the data structures at the bottom of the PC1-PC2 scatterplot and the left of the PC2-PC3 one that are observed only for the case of M4 refer to hourly series.

4.2. Enriching the analysis with additional features

The analysis of Section 4.1 indicates that, although some differences are observed between M4 and M3, these are insignificant and do not affect their properties essentially in the high-dimension situation. As a result, four additional characteristics are introduced and exploited in order to reestimate the primary principal components of the space and evaluate possible variances. These features include *skewness*, *kurtosis*, *non-linear autoregressive structure* and *self-similarity*, as described by Wang et al. (2006).

The first feature is used for measuring asymmetries in the time series. The second is used for measuring whether the data are peaked or flat relative to a normal distribution, and for detecting fat tails. The third is used for expressing non-linearity among the observations, while the last assesses long-range dependences in the data.

It is important to mention that no one can tell for sure whether the additional features can really make a difference in improving the identification of the factors that effectively classify a set of time series and define their properties in detail. This is because the feature vector is used to describe the original time series indirectly and no measure is available for computing objectively the information loss of doing so in order to define its elements optimally. However, the ten features included in this study can definitely provide some trustworthy insights in that direction and help us to conclude whether the principles of M3 are similar to the reality, at least when most of the basic statistical measures are considered.

Following the suggestions in Section 2, we execute a principal components analysis using the combination of the M3 and M4 datasets as the input and examine the first three components. These can be expressed algebraically as follows:

$$\begin{bmatrix} PC1 \\ PC2 \\ PC3 \end{bmatrix} = \begin{bmatrix} -0.44 & 0.46 & 0.01 & -0.01 & 0.43 & 0.06 & -0.27 & -0.35 & -0.03 & 0.45 \\ 0.21 & -0.14 & -0.05 & -0.19 & -0.26 & 0.43 & -0.57 & -0.47 & 0.26 & -0.15 \\ -0.08 & -0.15 & 0.66 & 0.68 & 0.09 & 0.21 & -0.10 & 0.01 & 0.11 & -0.08 \end{bmatrix} \mathbf{F}.$$

The three components retain about 70% of the variation in the data, and each of them emphasizes different

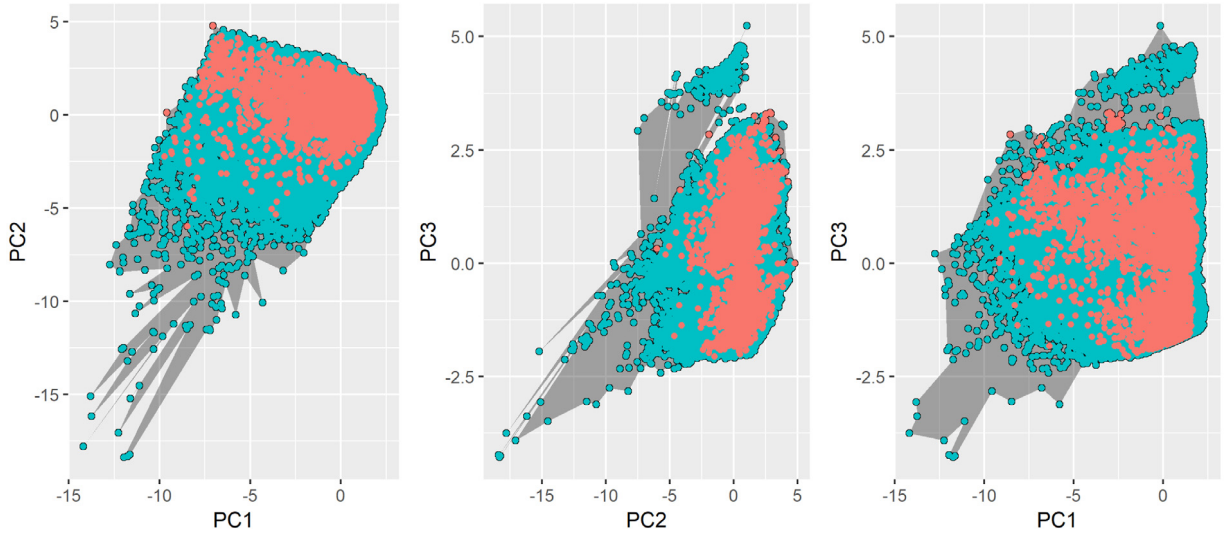

Fig. 5. The instance space of the M3 competition (red) compared to that of the M4 (blue). Ten time series features are used to construct the space and the α-convex hull is used to detect its boundaries empirically. The combined dataset of the M3 and M4 series is used for estimating the principal components. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

time series characteristics. The first one mainly distinguishes trended from non-trended time series, which also tend to be more noisy. As expected, long-range dependences and linearity are related closely to the feature of trend, and influence PC1 accordingly. On the other hand, PC2 focuses on the distribution of the data, emphasizing more skewed and peaked time series. Finally, the third component illustrates the seasonal effect on the data, helping us distinguish time series based on their frequency and seasonality. Thus, we conclude that trend, normality and seasonality are the most indicative time series characteristics, visualized in relative pairwise scatterplots in Fig. 5. Once again, the plots are obtained by projecting M4 and M3 on the feature space of their combined dataset. Moreover, we empirically detect the boundaries of the feature space generated by using the α-convex hull (Pateiro-Lopez & Rodriguez-Casal, 2016) to indicate the smallest convex set that contains "reality".

Fig. 6 suggests that most of the differences that are observed between the two datasets are due to different skewness and kurtosis values, as the time series of M4 tend to be more skewed and pointed. Observe that the scatterplot PC1-PC3 looks very similar to the PC1-PC2 one of Kang et al. (2017) if flipped vertically, suggesting that the two datasets have quite similar properties when abnormality is not examined.

We evaluate the similarity between the two datasets by estimating the distributions of the individual components through kernel density estimates and then computing KL in order to determine whether the multivariate distributions of the two datasets differ significantly. The results are presented in Fig. 6, and, as expected, differences are observed only for the second principal component of the series. However, once again, these differences are minor, displaying a KL percentage difference of less than 5%. Thus, we conclude that, although reality may include more abnormal series, these are observed rarely and do not lead to significant variations overall. This can be confirmed by inspecting the PC2 distributions of Fig. 6, which suggests that the differences that are reported between the two datasets originate mainly from the tails of M4's PC2 distributions, not their center. Thus, even after considering additional time series features, we conclude that the M3 data are representative of the real world.

Undoubtedly, the conclusions drawn in this section are related closely to the features selected for describing the time series, and more variations might be reported if additional or different features were considered. The feature vector is just a mean for describing the original time series indirectly, and no one can tell whether the information loss from using a specific set of features is significant, or objectively quantify the information gain from feature set *A* to an enriched set *B*. In this regard, future related work should include multiple representative features in order to make sure that most of the information is captured effectively. The present study has attempted to do so using some of the most indicative time series statistics.

We provide some guidance for future research by sorting the ten features examined in this study in terms of significance based on their PCA loadings. This is because large coefficients indicate that the corresponding features explain large percentages of the variance, helping us to identify different data structures better. In this respect, we first perform a principal component analysis using all of the features and identify the most important one using their absolute coefficients at PC1 as a criterion. We exclude the feature with the largest coefficient from the analysis, then repeat the procedure in order to determine the second most important feature, and so on. Following this process, we end up with the following list of features: *strength of trend*, *spectral entropy*, *kurtosis*, *first order autocorrelation*, *seasonal period*, *skewness*, *self-similarity*, *optimal Box–Cox transformation parameter*, *strength of seasonality* and *non-linear autoregressive*

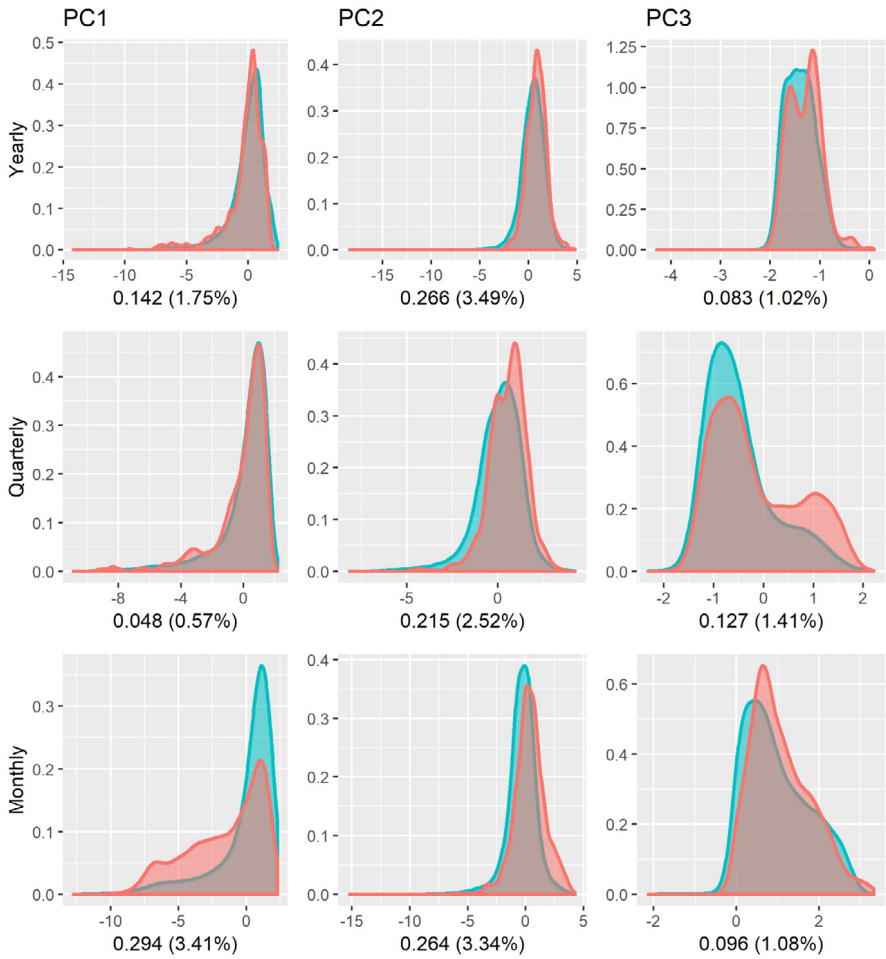

Fig. 6. Kernel density estimates for the first three principal components of the 10-feature space generated for the cases of the M3 (red) and M4 (blue) time series. The components are estimated using the combination of the M3 and M4 datasets (103,003 series) as input. The distributions are displayed for each frequency, and the Kullback–Leibler divergence is used to report their absolute and percentage differences. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

structure. Overall, measures of trend, randomness, abnormality and seasonality seem to be the most important ones for describing a time series effectively. Thus, from our point of view, these features should definitely be considered when clustering time series, and further enrichments could be examined in order to minimize the information loss, with their significance also being determined by the nature of the task under study and its objectives.

4.3. Comparison of other past forecasting competitions datasets with "reality"

It has been shown that M3 may indeed represent "reality", as the properties of its series do not deviate significantly from those of M4's series. However, there are still many other forecasting competition datasets that are used commonly as benchmarks in the literature, and their representativeness should be tested accordingly.

We investigate this problem by applying the principal component analysis of Section 4.2 to the data from the M1, M3, M4 and T competitions and extracting the first three components of each. Then, a pairwise comparison of the estimated components is performed for each data frequency and the KL values are calculated in order to draw relevant conclusions. The analytical results of the analysis are not presented for the sake of simplicity, but the main findings are reported below. Note that no safe conclusions could be drawn for the case of M2, as its sample size is very limited; thus, its series were excluded from this analysis.

In brief, the series from the T competition are the ones that display the most significant differences when compared to M4, especially in terms of PC1, where the KL divergence is greater than 15%, 14% and 22% for the yearly, quarterly and monthly data, respectively. Smaller deviations are reported for the cases of PC2 and PC3, which become notable only for the monthly series. M1 is clearly more representative of the reality, as the properties of its series deviate significantly from those of M4 only for the case of PC1 and for the yearly and monthly series (19% and 7%, respectively). Finally, the results show that M3 is quite

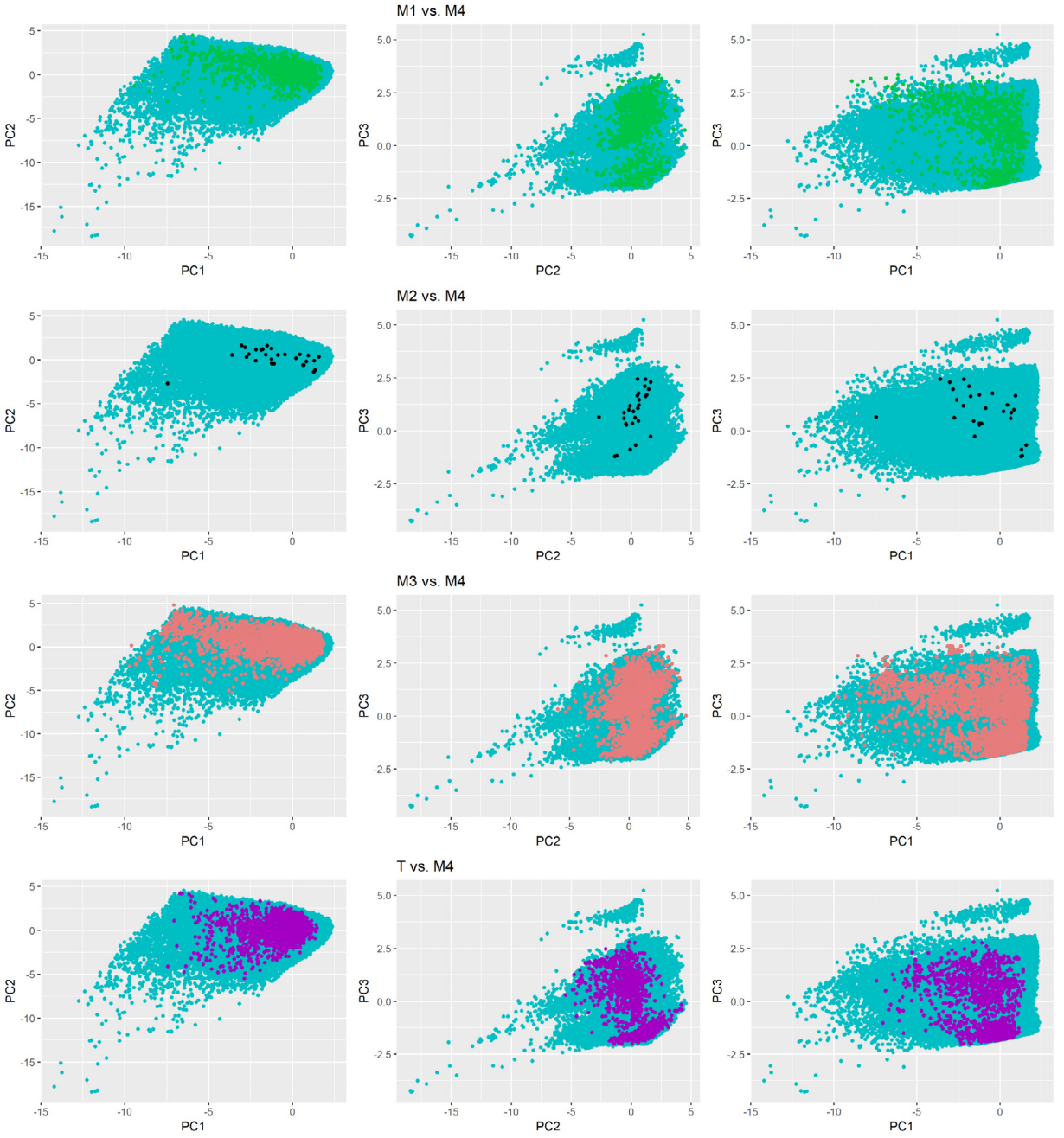

Fig. 7. The instance space of the M4 competition (blue) compared to those of the M1 (green), M2 (black), M3 (red) and T (purple). Ten time series features are used to construct the space. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

similar to M1, although some notable differences can be observed for the first and second principal components of their yearly data. This is not true when comparing M3 to T, where variations similar to those reported when comparing M4 to T are present.

Fig. 7 presents the instance spaces of the datasets examined relative to M4 in order to facilitate comparisons. Observe that the space of M3 displays the fewest gaps relative to reality, in contrast to M1 and T, which consist of less noisy series. Moreover, the T series seem to be much less trended. A common issue that has been identified for the data from all of the past forecasting competitions, and especially T, is that abnormal series are extremely rare or even entirely absent.

In light of the findings above, we conclude that, with the exception of M3, the data used in past forecasting competitions may deviate from "reality", and that the similarities among their findings and results might caused

by resemblances among the time series selected. The differences mainly refer to the trend/randomness and the normality of the data, but no significant differences are observable for the case of seasonality, with the rest of the components displaying rather minor variations. The M3 dataset is also much closer to the reality than the rest of the datasets examined, followed by that of M1.

5. Explaining the performances of time series forecasting methods through the instance space

The analysis performed in Section 4 shows the majority of the past forecasting competitions' datasets seem to deviate from reality, with M3 being an interesting exception. If such is the case, it would be interesting to see how the performance of each forecasting method varies between, say, M3 and M4, and to determine whether the differences observed in their features can explain such deviations in forecasting accuracy. To do so, four well-known forecasting methods are chosen for extrapolating the time series in each dataset as follows:

- Naive 2: For non-seasonal time series, this is equivalent to the Naive method (a forecast is equal to the last known observation). For the rest, time series are seasonally adjusted using classical multiplicative decomposition (Makridakis, Wheelwright, & Hyndman, 1998) and predicted using the Naive method. Then, the forecasts are seasonally re-adjusted.
- ETS: Exponential smoothing state space modeling (Hyndman, Koehler, Snyder, & Grose, 2002). It is used widely as a general forecasting algorithm, as it provides the best exponential smoothing model, indicated through information criteria.
- ARIMA: Autoregressive integrated moving average models, as implemented in the automated algorithm of Hyndman and Khandakar (2008).
- Theta method: Like Naive 2, time series are first seasonally adjusted if necessary. Then, the classic Theta, as originally proposed by Assimakopoulos and Nikolopoulos (2000), is applied for extrapolation. After prediction, the forecasts are seasonally re-adjusted.

We evaluate the forecasting performances using the mean absolute scaled error (MASE) (Hyndman & Koehler, 2006) as our accuracy measure, given that the sMAPE, which was used originally in the M3 competition, has been proved to penalize large positive errors more than negative ones and to increase with the difference in their absolute sizes (Goodwin & Lawton, 1999). However, we also report sMAPE values for the sake of comparison with the original study. The results are summarized in Table 4 for each dataset, method and frequency. The forecasting horizons used were the ones originally defined by the organizers of the competitions.

Regarding the yearly data, Theta outperforms the rest of the participating methods on both the M3 and M4 datasets. In contrast, for the quarterly and monthly series, although the Theta method is the dominant one for M3, it is outperformed by ETS and ARIMA for the case of M4. Note also that the MASE values of M4 are higher on average for all of the methods examined, indicating that the feature differences observed between the two datasets might indeed have an impact on the forecasting accuracy: abnormal and noisy time series are more difficult to forecast than stable, seasonal ones.

We investigate further how time series features influence the forecasting accuracy by using a multiple linear regression (MLR) model to correlate the MASE value achieved for each time series in M4 with the ten features that are chosen to describe its properties, as follows:

$$MASE_i = aF1_i + bF2_i + cF3_i + \cdots . + jF10_i\,,$$

where $MASE_i$ is the error generated for the ith time series of the sample by Naive 2, ETS, ARIMA or Theta. Before estimating the models, both dependent (MASE) and independent (features) variables are 1% trimmed in order to mitigate the effect of extreme values and then scaled within the range of [0,1] so that the results are scale-independent and easier to interpret.

This allows us to detect features that explain the MASE variances statistically significantly, while quantifying their negative or positive effects on the forecasting accuracy approximately. Finally, since each forecasting method generates different MASE values, constructing an individual MLR model for each case and comparing their coefficients will help us to understand the strengths and weaknesses of each method better, and maybe provide some criteria for selecting the most appropriate one effectively. The idea of performing model selection through time series characteristics instead of goodness of fit measures was introduced by Petropoulos et al. (2014), who concluded that there are "horses for courses". This study performs a similar analysis, but exploits real data and more time series characteristics for estimating the coefficients and drawing relative conclusions.

Table 5 presents the coefficients estimated by the MLR models in order to explain the accuracies achieved by the Naive 2, ETS, ARIMA and Theta methods through the selected time series features. The coefficient of determination R^2 is also reported as an indication of their interpretability. As can be seen, the variables considered explain about half of the MASE variance. Some of the factors that influence the value of R^2 are the existence of possible non-linear dependencies, the effects of other variables not included in the models, and the fact that nuances of the past are likely not to continue into the future.

In brief, of the features examined, the frequency, kurtosis, linearity and seasonality are the characteristics that lead in general to an improved forecasting accuracy, while the effects of stability and non-linearity are neutral. In other words, patterns of seasonal, stable and pointed time series can be captured more easily in general by the methods examined and extrapolated in the future. On the other hand, the skewness, self-similarity, spectral entropy and trend have negative impacts on MASE; i.e., abnormal, noisy and trended time series are more difficult to forecast efficiently. Moreover, all variables apart from trend in the ARIMA, ETS and Theta methods were found to be statistically significant, indicating that every single one of

Table 4
MASE and sMAPE values of the four forecasting methods on M3 and M4, presented by frequency.

Frequency	Naive 2	ETS	ARIMA	Theta	Naive 2	ETS	ARIMA	Theta
	M3 Competition							
	MASE				*sMAPE*			
Yearly	3.17	2.89	2.96	**2.77**	17.88	17.11	17.12	**16.73**
Quarterly	1.24	1.19	1.19	**1.13**	9.97	9.60	10.00	**9.30**
Monthly	1.03	0.87	0.88	**0.86**	16.71	14.14	14.88	**13.87**
Daily	–	–	–	–	–	–	–	–
Weekly	–	–	–	–	–	–	–	–
Hourly	–	–	–	–	–	–	–	–
Other	3.09	**1.82**	1.83	2.27	6.30	**4.39**	4.46	4.92
All	1.67	1.44	1.46	**1.42**	14.66	13.07	13.53	**12.82**
	M4 Competition							
	MASE				*sMAPE*			
Yearly	3.97	3.44	3.40	**3.38**	16.34	15.36	15.17	**14.59**
Quarterly	1.37	**1.16**	1.17	1.23	11.01	**10.29**	10.43	10.31
Monthly	1.06	0.95	**0.93**	0.97	14.43	13.53	13.44	**13.00**
Weekly	2.78	**2.53**	2.56	2.64	9.16	8.73	**8.65**	9.09
Daily	3.28	**3.25**	3.41	3.26	**3.05**	3.05	3.19	3.05
Hourly	2.40	1.82	**0.94**	2.46	18.38	17.31	**13.98**	18.14
Other	–	–	–	–	–	–	–	–
All	1.91	1.68	**1.67**	1.70	13.56	12.73	12.67	**12.31**

Table 5
Coefficients of the MLR models relating the MASE values generated by the Naive 2, ETS, ARIMA and Theta methods in the M4 time series with their features.

Method	Forecastability	Trend	Seasonality	Frequency	Linearity	Stability	Skewness	Kurtosis	Non-linearity	Self-similarity	R^2
Naive 2	0.05	0.07	−0.02	−0.26	−0.06	−0.01	0.13	−0.08	−0.01	0.17	0.57
ETS	0.05	0.00	−0.04	−0.22	−0.06	0.00	0.13	−0.06	−0.02	0.21	0.56
ARIMA	0.05	0.00	−0.04	−0.23	−0.05	0.00	0.14	−0.05	−0.01	0.21	0.56
Theta	0.04	0.01	−0.02	−0.24	−0.04	−0.01	0.13	−0.06	−0.01	0.21	0.56

Note: The R^2 values of the models used are reported in order to quantify the interpretability.

them helps to approximate the predictability of the time series.

It is interesting to note that the estimated coefficients are not identical across the models. In this respect, some forecasting methods may be influenced more strongly by specific time series characteristics, but display robustness against others. For instance, the Theta method is affected less by randomness, while the rest of the methods are slightly more vulnerable to it. On the other hand, the performances of ETS and ARIMA are affected less by possible trends, followed by Theta. Given that Naive 2 is a non-trended model, its accuracy decreases significantly when dealing with trended data. Regarding seasonality, ETS and ARIMA display lower coefficients, indicating that using forecasting models with seasonal components might be more effective for extrapolating seasonal series than using a combination of non-seasonal model and external seasonal adjustments, such as the Naive 2 or Theta methods. Moreover, the negative coefficients that are reported for the features of seasonality and frequency confirm that seasonal series and high frequency data are easier to predict than non-seasonal or low-frequency ones. For cases of linearly autocorrelated time series, Naive 2, ETS and ARIMA are the most robust choices, while ETS is best for cases of non-linear series. The impact of stability is insignificant and does not vary across the methods, in contrast to self-similarity, which affects all of the methods strongly. The skewness also increases MASE significantly, with ARIMA being slightly less robust to abnormalities. Finally, pointed time series are forecast better by Naive 2, followed by Theta and ETS.

To conclude, the minor differences observed between the properties of the M4 and M3 datasets seem also to affect the performances of the forecasting methods. This is mainly because each feature has different effect on the forecasting accuracy, and this effect also varies according to the method used. Thus, we find out that there are "horses for courses" in time series extrapolation and that some methods are expected to have an advantage over the others, depending on the structure of the dataset used. In this respect, when attempting to deal with a real world forecasting problem optimally, one can either choose the dominant method across the dataset (aggregate selection) or individually select the most appropriate one per series (individual selection) (Fildes & Petropoulos, 2015). Either way, time series features can provide valuable information for decision-making.

6. Discussion

The present study reveals that the properties of past forecasting competition datasets may deviate from those of reality, with M3 being an interesting exception. The main differences observed refer to the abnormality of the data, with reality consisting of relatively more skewed series that display outliers, as well to their limited

randomness/trend. However, these features may just be examples of such variations, meaning that it might be possible to consider more time series characteristics for describing the properties of the time series effectively, with the potential to reveal new structures that were previously unobservable. For instance, Hyndman et al. (2015) proposed a set of features for the effective identification of unusual series, which may be found more frequently in the real world than in datasets like M3, M1 or T.

Given that there is no objective way of determining which features should be used ideally to classify a set of series, further research and empirical investigations could help us to identify the most significant ones, with their number and nature being subject to the application considered and the objectives of the task under study. For instance, a sensitivity analysis of the results when more features are examined could be proposed, so as to quantify the information gain from using an enriched set. The present study, which has examined the time series characteristics that help us to identify different data structures better, has determined the significance of the features according to their PCA loadings. It is notable that time series characteristics which are observed qualitatively, such as trend, randomness, outliers and seasonality, were found to be the most important ones for describing the properties of the time series. Thus, determining the minimum number and type of features needed to classify a set of series adequately becomes a promising task.

Another interesting outcome of this study was that time series features were found to be related closely to the performances of the forecasting methods used for extrapolation. Although this conclusion has been reported previously in the literature, it is of vital importance because it was drawn using a large, diverse data collection of real time series, serving as a proportion of the real world. Since each method is affected differently by time series characteristics, data classification through features could be exploited in order to select the best-performing method for each series efficiently. This could have a lot of advantages in big-data applications, where the individual validation of forecasting methods requires a lot of time and resources. Moreover, more indicative and robust choices are expected to be made, since model selection is performed using information extracted from a large dataset rather than a unique series. Alternatively, since each dataset has its own particularities, the dominant method could be identified and exploited so as to achieve an adequate forecasting performance, on average. Accuracy, complexity and computational cost would be the main criteria for determining whether individual or aggregate model selection should be performed.

The results also demonstrate significant similarities between the datasets of M3 and M1, which probably explains why the conclusions drawn in the two competitions were identical. Moreover, it is shown that the instance space of the T competition is even more limited, consisting of relatively less noisy and trended series. This conclusion could be evidence that the domain of the data may also determine which time series characteristics will be dominant, thus favoring some forecasting methods over others. In this regard, instead of choosing the best performing method based on the features of the data, it could be meaningful to determine which method performs best in each application field. For instance, the Theta method, which was the winner of the M3 competition, performed the best only for the microeconomic series. In contrast, ARIMA models were found to perform best for financial data, Rule Based Forecasting for macroeconomic data, and forecasting software like ForecastPro and ForecastX for industry and demographic data.

The results of the recently-completed M4 competition, which featured 61 different forecasting methods, shed light on these topics, demonstrating innovative ways and best practices for improving the forecasting accuracy. For instance, the second-best performing method of M4 (Montero-Manso, 2018) involved a combination of statistical methods where the weights are calculated by a learning model over a set of features. Table 6 presents the dominant features of the M4 series by domain of data, to further support such an analysis and provide relevant criteria for model selection. For instance, since Table 5 shows that finance data tend to be noisy, untrended, skewed and not seasonal, the Theta method could potentially be utilized to boost the forecasting performance. Accordingly, ETS and ARIMA could be powerful solutions for extrapolating demographic data that are characterized by strong seasonality, trend and linearity.

Moreover, this paper has presented details of the design of the M4 competition, thus providing some guidelines for the construction of future forecasting competition. The applications involved, in terms of frequency and domain, as well as their proportions in the final dataset, were the main factors considered. These guidelines also help us to answer another major question, which is whether researchers should continue to use well-known forecasting competition data for evaluating new forecasting approaches. The answer seems to be quite relative: since time series features determine forecasting performances but different types of data display different characteristics, the similarities between real-life applications and forecasting competition datasets should be examined carefully to determine whether using the latter is appropriate. M4 includes data from multiple and diverse databases, and thus provides insights about which practices should be used to improve business forecasting in general. In this respect, M4 could become a standard testing ground for evaluating the performances of generic time series methods. However, if the objective was to select the best method for predicting tourism data, using M4 as a benchmark would probably be a suboptimal choice. We consider that forecasting competitions are extremely useful tools for advancing the theory and practice of forecasting as long as their data are used reasonably for evaluating forecasting performances in relevant real life applications. The studies of Petropoulos, Wang, and Disney (2019), who utilized only that part of the M3 dataset that was relevant to their application (industry series), and Kourentzes, Petropoulos, and Trapero (2014), who expanded M3 data according to their needs (including

Table 6
Connecting the features of M4 series with their domain.

Domain	Forecastability	Trend	Seasonality	Frequency	Linearity	Stability	Skewness	Kurtosis	Non-linearity	Self-similarity
Demographic	✗	✓	✓	✓	✓	✓	✗	✗	✓	✓
Finance	✓	✗	✗	✗	✗	✗	✓	✓	✗	✗
Industry	✓	✗	-	-	✗	-	-	✓	-	-
Macro	-	✓	-	-	✓	✗	-	✗	✗	✓
Micro	-	✓	-	✗	✓	✗	-	✗	✓	✓

Notes: ✗, - and ✓indicate low, medium and high significance levels, respectively. The significance levels reported for each feature are relative to the absolute values computed across all domains examined.

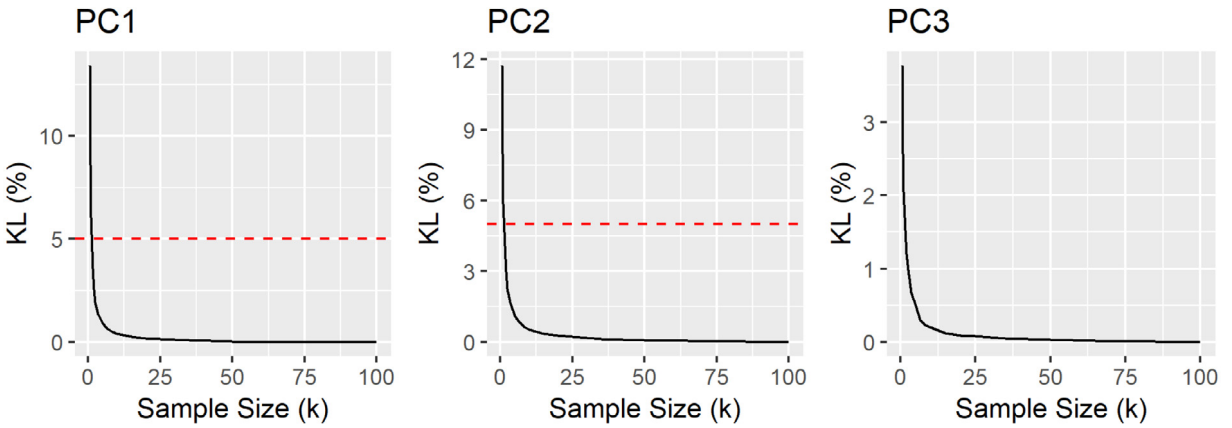

Fig. 8. Comparison of the M4 dataset with random samples of it of various sizes. The Kullback–Leibler divergence is used to report the percentage differences observed for the first three principal components of the enriched feature space. The red dotted line indicates the minimum number of series in the sample that retains the properties of the original dataset (less than 5% difference).

semi-annual series), are some excellent examples of such conservative uses.

As a final point, we would like to note that the methodological framework proposed in this study for assessing the similarity of data from popular forecasting competitions can also be exploited for the construction of minimal datasets that retain the properties of reference ones. For instance, the dataset of M4 could reasonably be shrunk and then used to facilitate the evaluation of complex methods which, otherwise, would be a time-consuming task to undertake. This can be achieved by sampling the reference dataset randomly and comparing its properties in the high-dimensional situation with those of the resulting samples through the Kullback–Leibler divergence. The smallest sample that displays insignificant differences for all the components examined can be considered to be minimal. In order to demonstrate this process in practice, we generated 5000 random samples of various sizes for the case of M4 and estimated the average difference for each sample size for the first three principal components of the enriched feature space. The results are visualized in Fig. 8, and suggest that a random sample of 1,000 time series could be enough to reach similar conclusions. This is because PC1 and PC2 show that smaller samples would contain significantly fewer trended and abnormal series than the reality. Note that this process can succeed the guidelines provided in Section 3 to enable the construction of minimal, yet representative, competition datasets.

7. Conclusions

This study has compared the properties of past forecasting competitions data with those of M4, the latest and largest one available. The purpose of this comparison was to investigate whether these popular datasets, which are often used in the literature as benchmarks, are representative of "reality". To do so, we analyzed the features of the M4, M3, M2, M1 and T series and examined the ways in which these are distributed and interrelated on the instance space proposed by Kang et al. (2017).

The results indicate that, although individually the features of the datasets display quite similar principles, their properties may differ when exploring their multivariate distributions. The variations are more noticeable when abnormality and trend/randomness are taken into account, since no significant differences are observed for the rest of the features. However, the variations reported are larger for some of the datasets, due mainly to their limited size and the special types of data involved. For instance, it was found that M3 did not differ significantly from M4, but that M1, and especially T, displayed major differences when compared to the real word. Thus, different features could potentially be considered in the future for detecting additional variations.

Another outcome of the study is the finding that the differences identified between data from different forecasting competitions also have an impact on the forecasting accuracy achieved. This is because each method

demonstrates its own strengths and weaknesses, which are related closely to the special features of the series being extrapolated. Thus, although a dominant method can be found for every single dataset, and even for reality, it may be possible to exploit the time series features to identify the best alternative for each series, thus further boosting the forecasting performance. It was also found that the domain of the data may determine the characteristics of the series, thus influencing the performances of the methods being utilized.

In light of the findings above, we believe that time series features and domain should definitely be taken into consideration when investigating best practices in forecasting. Thus, instead of questioning the representativeness of forecasting competitions data in general, the right question might be whether these are appropriate for simulating the particular application to which the suggested approach should excel. Organizing competitions like M4, that consist of numerous and diverse series, can help forecasters better capture "reality" and promote forecasting in various fields. This study serves in this direction by presenting how M4 was designed, as well as describing some ideas for the construction of minimal, yet representative, benchmarks.

References

Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, *16*(4), 521–530.

Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting*, *27*(3), 822–844.

Clements, M., & Hendry, D. (2001). Explaining the results of the M3 forecasting competition. *International Journal of Forecasting*, *17*(4), 550–554.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. John Wiley & Sons, Inc..

Fildes, R., & Petropoulos, F. (2015). Simple versus complex selection rules for forecasting many time series. *Journal of Business Research*, *68*(8), 1692–1701.

Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, *15*(4), 405–408.

Hyndman, R., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *26*(3), 1–22.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 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*(3), 439–454.

Hyndman, R., Wang, E., & Laptev, N. (2015). Large-scale unusual time series detection. In *IEEE international conference on data mining, 14–17 November 2015*.

Kang, Y., Hyndman, R. J., & Li, F. (2018). Efficient generation of time series with diverse and controllable characteristics. Working paper available online at https://robjhyndman.com/publications/tsgeneration/.

Kang, Y., Hyndman, R. J., & Smith-Miles, K. (2017). Visualising forecasting algorithm performance using time series instance spaces. *International Journal of Forecasting*, *33*(2), 345–358.

Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting*, *30*(2), 291–302.

Lawrence, M. (2001). Commentaries on the M3-Competition. Why another study? *International Journal of Forecasting*, *17*(4), 574–575.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting*, *1*(2), 111–153.

Makridakis, S., Assimakopoulos, V., & Spiliotis, E. (2018). Objectivity, reproducibility and replicability in forecasting research. *International Journal of Forecasting*, *34*(4), 835–838.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The M-2 Competition: a real-time judgmentally based forecasting study. *International Journal of Forecasting*, *9*(1), 5–23.

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*(4), 802–808.

Makridakis, S. G., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: Methods and applications* (3rd ed.). New York: Wiley.

Montero-Manso, P. (2018). M4metalearning: Metalearning tools for time series forecasting. *R package version 0.0.0.9000*. https://github.com/robjhyndman/M4metalearning.

Nanopoulos, A., Alcock, R., & Manolopoulos, Y. (2001). Feature-based classification of time-series data. *International Journal of Computer Research*, *10*(3), 49–61.

Ord, K. (2001). Commentaries on the M3-competition. *International Journal of Forecasting*, *17*(4), 537–584.

Pateiro-Lopez, B., & Rodriguez-Casal, A. (2016). alphahull: Generalization of the convex hull of a sample of points in the plane. *R package version 2.1*. https://CRAN.R-project.org/package=alphahull.

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

Petropoulos, F., Wang, X., & Disney, S. M. (2019). The inventory performance of forecasting methods: evidence from the M3 competition data. *International Journal of Forecasting*, *35*(1), 251–265.

Wang, X., Smith, K., & Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery*, *13*(3), 335–364.

Evangelos Spiliotis He is a research associate of the Forecasting & Strategy Unit at the National Technical University of Athens, where he also completed his Doctor in Engineering. He has conducted research on time series forecasting, model optimization, forecasting support systems and business resource management in many National and European projects.

Andreas Kouloumos He is an undergraduate student in the Department of Electrical and Computer Enginnering of National Technical University of Athens. Since October 2016 he is a member of the Forecasting and Strategy Unit, where he works among others on time series forecasting, programming and application development.

Vassilios Assimakopoulos He is the professor of Forecasting Systems at the School of Electrical and Computer Engineering of the National Technical University of Athens. He has worked extensively on applications of the Decision Systems in modern problems for business design. Moreover, he has conducted research on innovative tools for management support, in an important number of projects, funded by National and European institutes. He is the author of over than 60 original publications and papers in various national scientific journals and conferences.

Spyros Makridakis is a Professor at the University of Nicosia and its director of the Institute For the Future (IFF) as well as an Emeritus Professor at INSEAD. Spyros is the organizer of the M Competitions that started back in the early 1980s. He has authored, or co-authored, twenty-two books and more than 150 articles. He was the founding editor-in-chief of the Journal of Forecasting and the International Journal of Forecasting.