International Journal of Forecasting 33 (2017) 345–358

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Visualising forecasting algorithm performance using time series instance spaces

Yanfei Kang[a,*], Rob J. Hyndman[b], Kate Smith-Miles[c]

[a] School of Economics and Management, Beihang University, Beijing, 100191, China
[b] Department of Econometrics and Business Statistics, Monash University, Clayton VIC 3800, Australia
[c] School of Mathematical Sciences, Monash University, Clayton VIC 3800, Australia

A R T I C L E I N F O

Keywords:
M3-Competition
Time series visualisation
Time series generation
Forecasting algorithm comparison

A B S T R A C T

It is common practice to evaluate the strength of forecasting methods using collections of well-studied time series datasets, such as the M3 data. The question is, though, how diverse and challenging are these time series, and do they enable us to study the unique strengths and weaknesses of different forecasting methods? This paper proposes a visualisation method for collections of time series that enables a time series to be represented as a point in a two-dimensional instance space. The effectiveness of different forecasting methods across this space is easy to visualise, and the diversity of the time series in an existing collection can be assessed. Noting that the diversity of the M3 dataset has been questioned, this paper also proposes a method for generating new time series with controllable characteristics in order to fill in and spread out the instance space, making our generalisations of forecasting method performances as robust as possible.

1. Introduction

The M3 data (Makridakis & Hibon, 2000) are used widely for testing the performances of new forecasting algorithms. These 3003 series have become the de facto standard test base in forecasting research. When a new univariate forecasting method is proposed, it is unlikely to receive any further attention or be adopted unless it performs better on the M3 data than other published algorithms.

We see several problems with this approach. The M3 dataset was a convenience sample that was collected from several disciplines, namely demography, finance, business and economics. All of the data were positive, with series lengths ranging from 14 to 126, and were observed annually, quarterly or monthly (apart from 174 "other" series, whose frequencies of observation were not provided). Methods that work well on this data set may overfit data with similar data structures. Thus, testing algorithms on this data set will tend to favour forecasting methods that work well with data from these domains, and of these lengths and frequencies. Furthermore, there is no guarantee that the series will be in any way "representative" of the types of data that are found within those domains, as is noted in the subsequent discussion of the M3 data (Ord, 2001). Finally, given that 15 years have elapsed since the M3 results were published, it is highly likely that the patterns seen within typical time series will have changed over time, even within the collection constraints of the competition.

There has been no attempt in the published M3 results to study *why* some methods perform better on certain series than other methods. Is it just chance, or do some time

* Corresponding author.
E-mail addresses: yanfei.kang@outlook.com (Y. Kang), Rob.Hyndman@monash.edu (R.J. Hyndman), Kate.Smith-Miles@monash.edu (K. Smith-Miles).

http://dx.doi.org/10.1016/j.ijforecast.2016.09.004
0169-2070/

series have particular features that make them particularly amenable to being forecast by one method rather than another? In discussing the M3 results, Lawrence (2001) wrote,

> What is needed now is analysis to determine what are the specific time series characteristics for which each technique is generally best and also what are the time series characteristics for which it does not really matter which technique (or set of techniques) is chosen.

Given that the M3 time series might share some specific characteristics, conclusions based on these data might only hold for other series with these particular characteristics (Clements & Hendry, 2001).

Similar comments apply to other collections of time series. How do we know that any time series collection covers the range of possible time series patterns, or is somehow representative of the types of data that we are designing forecasting methods to handle?

This paper proposes a new approach that aims to answer some of these questions. The methodology is an adaptation of previous work by the authors on the objective assessment of combinatorial optimisation algorithms (Smith-Miles, Baatar, Wreford, & Lewis, 2014) and the generation of new test instances (Smith-Miles & Bowly, 2015), which is extended to the time series and forecasting domains for the first time here.

Our approach involves computing the "features" of each time series. For example, we measure the autocorrelation at lag 1, the seasonal period, and the spectral entropy. These, and several other features, are all numerical quantities that are computed on each time series, and we then study the "feature space" of the collection of time series. By studying the feature space rather than the raw time series, we convert the data from temporal to static. We also convert a large collection of time series of different lengths to a data set that comprises a small number of features for each series. Thus, each time series is represented as a point in a high-dimensional feature space, which can be reduced to a two-dimensional instance space using dimension-reduction techniques.

The idea of characterising a time series as a feature vector is not new, and has been used for classifying time series (e.g., Fulcher & Jones, 2014; Fulcher, Little, & Jones, 2013; Nanopoulos, Alcock, & Manolopoulos, 2001), clustering time series (e.g., Fulcher et al., 2013; Wang, Smith, & Hyndman, 2006), and identifying outlying or anomalous time series (e.g., Hyndman, Wang, & Laptev, 2015). This paper generates a two-dimensional instance space of time series and uses it to explore the properties of a given collection of time series, in this case, the M3 dataset. We study the distribution of features across the space in order to obtain an understanding of the similarities and differences between the time series, and to assess the diversity of the collection. We also investigate whether the location in the instance space, given by the features, can predict forecasting method performances. Finally, we identify gaps in the instance space and develop new methods for generating time series with controllable features by evolving time series to lie at given target locations.

The nature and number of the features to be used depends on the problem context and their discriminatory quality (Nanopoulos et al., 2001). We have suggested a small number of features that we think are useful for studying the M3 data. However, there may be many other features that would also be useful and would provide different information from those we have chosen. For other collections of time series, other sets of features will need to be used. For example, Hyndman et al. (2015) use a set of 18 features that are designed for the identification of anomalous time series of web traffic. Fulcher et al. (2013) and Fulcher and Jones (2014) use thousands of features for reduced representations of time series data and their analysis methods, thus extending our ability to capture nuanced characteristics of time series.

Section 2 defines the features that we have chosen for the M3 data and shows how principal components analysis can be used to reduce the dimensions of the feature space in order to enable a visualization of the space of the time series via a two-dimensional instance space.

The scatterplot of the first two principal components suggests that there may be regions of the feature space that are not covered well by the M3 data. Thus, Section 3 uses a genetic algorithm to generate new time series that are designed to fill the "gaps" in the feature space of the M3 data. In this sense, we are contributing a broader and more diverse collection of M3-like time series for testing the performances of forecasting methods.

2. Time series features

Depending on the research goals and domains, previous studies have developed a variety of time series features for the characterisation of time series (e.g., Deng, Runger, Tuv, & Vladimir, 2013; Fulcher & Jones, 2014; Kang, Belušić, & Smith-Miles, 2014, 2015; Mörchen, 2003; Nanopoulos et al., 2001; Wang et al., 2006). This paper considers six features, which are selected because we believe that they provide useful information about the M3 data.

The forecasting methods that performed best on the M3 data were those that modelled the trend and seasonal components of the data explicitly (Makridakis & Hibon, 2000), so we have selected methods that measure those characteristics. In addition, we have also included a measure of "forecastability", as suggested by Goerg (2013), and a measure of variance-stability, based on a Box–Cox transformation.

In the following descriptions, our time series is denoted by $\{x_1, \ldots, x_n\}$, observed at times $1, \ldots, n$.

Spectral entropy F_1. Entropy-based measures have been used widely in non-linear analysis for assessing the complexity of signals (e.g., Bandt & Pompe, 2002; Fadlallah, Chen, Keil, & Príncipe, 2013; Zaccarelli, Li, Petrosillo, & Zurlini, 2013) and measuring the "forecastability" of a time series (Garland, James, & Bradley, 2014; Goerg, 2013; Maasoumi & Racine, 2002). We use the spectral entropy measure that is included in the R package **ForeCA** (Goerg, 2013, 2014), which is an estimate of the Shannon entropy of the spectral density $f_x(\lambda)$ of a stationary process x_t:

$$F_1 = -\int_{-\pi}^{\pi} \hat{f}_x(\lambda) \log \hat{f}_x(\lambda) d\lambda,$$