International Journal of Forecasting 31 (2015) 79–90

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Reproducibility in forecasting research

John E. Boylan [a,*], Paul Goodwin [b,1], Maryam Mohammadipour [a,2], Aris A. Syntetos [c,3]

[a] School of Applied Management & Law, Buckinghamshire New University, Bucks HP11 2JZ, United Kingdom
[b] School of Management, University of Bath, Bath BA2 7AY, United Kingdom
[c] Cardiff Business School, Cardiff University, Cardiff CF10 3EU, United Kingdom

ARTICLE INFO

Keywords:
Forecasting practice
Replication
Seasonal forecasting
Empirical research

ABSTRACT

The importance of replication has been recognised across many scientific disciplines. Reproducibility is a necessary condition for replicability, because an inability to reproduce results implies that the methods have not been specified sufficiently, thus precluding replication. This paper describes how two independent teams of researchers attempted to reproduce the empirical findings of an important paper, "Shrinkage estimators of time series seasonal factors and their effect on forecasting accuracy" (Miller & Williams, 2003). The two teams proceeded systematically, reporting results both before and after receiving clarifications from the authors of the original study. The teams were able to approximately reproduce each other's results, but not those of Miller and Williams. These discrepancies led to differences in the conclusions as to the conditions under which seasonal damping outperforms classical decomposition. The paper specifies the forecasting methods employed using a flowchart. It is argued that this approach to method documentation is complementary to the provision of computer code, as it is accessible to a broader audience of forecasting practitioners and researchers. The significance of this research lies not only in its lessons for seasonal forecasting but also, more generally, in its approach to the reproduction of forecasting research.

Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do (D.E. Knuth, Stanford University).

1. Introduction and research background

Replication is one of the cornerstones of science. With replication, scientific claims may be challenged. In the medical field, Ioannidis (2005) examined 45 highly-cited articles from clinical journals and found that seven were contradicted by subsequent research and another seven were initially found to have stronger effects. Prasad et al. (2013) analysed 363 articles testing standards of care, and found that 146 medical practices were reversed in 10 years of publications.

In the absence of replication, scientific claims rest on the results of single, 'one shot', studies, and hence have various risks and limitations. Researchers may have inadvertently made errors in their applications of methods. They

* Corresponding author. Tel.: +44 0 1494 605130.
E-mail addresses: John.Boylan@bucks.ac.uk (J.E. Boylan), P.Goodwin@bath.ac.uk (P. Goodwin), Mona.Mohammadipour@bucks.ac.uk (M. Mohammadipour), SyntetosA@cardiff.ac.uk (A.A. Syntetos).
1 Tel.: +44 0 1225 383594.
2 Tel.: +44 0 1494 605130.
3 Tel.: +44 0 29 208 76572.

http://dx.doi.org/10.1016/j.ijforecast.2014.05.008
0169-2070/

may have made mistakes in data entry, committed arithmetic or data transcription errors, or written computer code that contains bugs. They may also have made assumptions that are not stated explicitly, and their findings may be sensitive to changes in these assumptions. Other assumptions, and even further errors, may be embedded in commercial software, so that researchers are unaware of them (McCullough, 2000). In addition, results may apply only to the specific data that have been analysed, and hence be subject to sampling error. When statistically insignificant results are obtained, researchers may be tempted to "hunt for *p*-values less than 0.05" (Hubbard & Armstrong, 1994), hence inflating the true probability of committing type I errors. This problem is avoided by replication studies, as statistical significance is not a measure of replicability. Finally, the extent to which the findings generalize to situations or populations beyond those investigated in the original study will be unknown.

These potential risks and limitations suggest a range of approaches to replication. The definition of replicability varies across disciplines, but a special case is reproducibility. If findings are reproducible, then independent researchers are able to obtain the same results as the original study using the same data and the same methods. Reproducibility is a first step towards replication, and therefore, if it cannot be achieved, the generalizability of the findings is likely to be in doubt. Of course, a perfect reproduction of the results may not be possible. For example, improvements in the algorithms embedded in software may lead to differences between the numbers originally reported and those obtained using later versions of the software. However, approximate reproducibility, discussed later in this paper, may still be attainable. Findings that have been reproduced successfully have a much lower risk of being subject to human error. Further, the process of trying to reproduce findings is likely to reveal the extent to which the original results were based on unstated assumptions, and hence the extent to which the findings will change if alternative assumptions are made.

Despite these potential benefits, the frequency of papers reporting the reproduction or replication of results is low in some disciplines. Evanschitzky, Baumgarth, Hubbard, and Armstrong (2007) found that, in marketing, the percentage of papers based on replication studies had halved to 1.2% over the period 1990–2004, relative to the period 1974–1989. A similar study of empirical research papers in forecasting, published between 1996 and 2008, found a rate of 8.4% (Evanschitzky & Armstrong, 2010). Although this was relatively high compared to other areas of management science, the authors argued that the rate needed to increase, given that the findings of about 20% of the original papers were not supported in the replications.

In recent years, there have been several developments which have supported replication in forecasting research. For instance, data sets such as those used in the M1 forecasting competition are easily accessible (Makridakis et al., 1982). The M1 data set has since been used in several other studies. In addition, authors publishing papers in the *International Journal of Forecasting* are requested to make their data publicly available via the journal's website. Indeed, in its inside cover, the journal states that it "encourages replication studies" and that "For empirical studies, the description of the method and the data should be sufficient to allow for replication". However, whether research is truly replicable or not may not be apparent until a full replication is attempted formally. Only then is the absence of important details or the imprecision of definitions or measurements likely to become apparent. For example, Simmons (1986) attempted to reproduce some of the M-competition results for the Naïve2 method. His initial attempt, based on information in the article alone, was unsuccessful. It was only after written communication with Professor Makridakis that sufficient details were clarified for the results to be reproduced. While, in general, it is relatively easy to disclose data, making methods transparent is more problematical. Even the original authors are likely to be unaware of how much documentation of methods is actually required in order to allow an independent researcher to reproduce their results.

This paper is about the process of reproducing results in forecasting research. We describe the process whereby two independent teams of researchers attempted to reproduce the findings of an award winning study, "Shrinkage estimators of time series seasonal factors and their effect on forecasting accuracy" (Miller & Williams, 2003). We then identify issues that arose during the process and discuss how these issues may be resolved.

The remainder of the paper is organized as follows. In the next section, the relationship between reproducibility and replicability is discussed in more detail. In Section 3, the original research is described, the process of reproducing the results and the sources of discrepancies are explained, and the impacts of these differences on Miller & Williams' findings are discussed. A more detailed explanation of this process is given in Appendices A and B. Section 4 compares different approaches to the specification of forecasting methods, and Section 5 concludes the paper. A comprehensive flowchart of the forecasting process is given in Appendix C, and references to supplementary material are provided in Appendix D.

2. Reproducibility vs. replicability

Following on from the discussion in the previous section, we propose the following definitions of reproducibility and replicability in forecasting research. If results are reproducible, then independent researchers are able to obtain the same numerical results by repeating the original study using the same methods on the same data. If findings are replicable, then independent researchers are able to reach the same qualitative conclusions by repeating the original study using the same methods on different data. It should be possible for independent researchers to reproduce or replicate without any additional information from the author(s) of the original study (King, 1995).

Evanschitzky and Armstrong (2010) use the term "reanalysis" to refer to an application of different methods to the same data or a sub-sample of the data. This constitutes a third category, in addition to "reproduction" and "replication", as shown in Fig. 1.