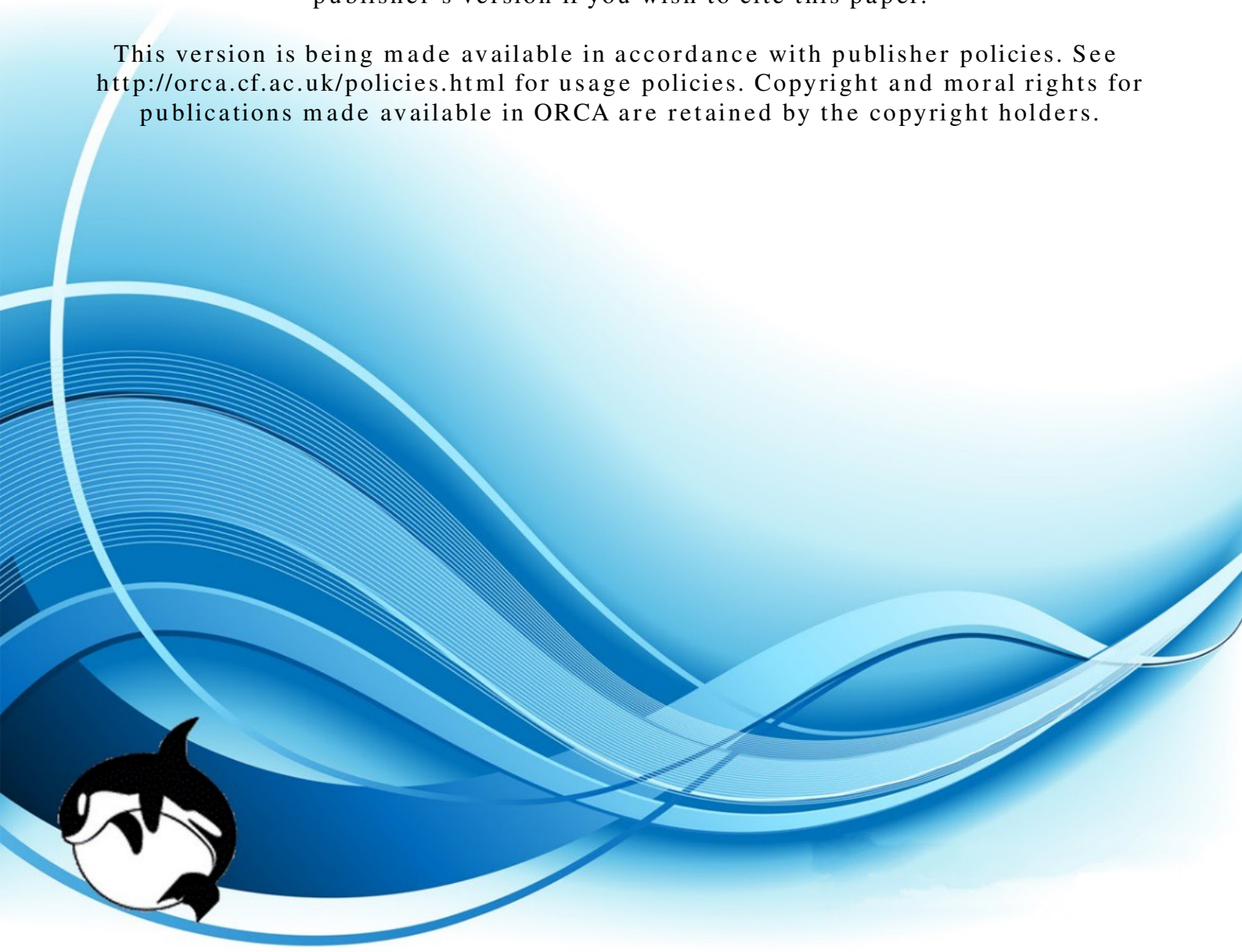

Reproducibility in Forecasting Research

(Accepted for publication in the *International Journal of Forecasting*)
J.E. Boylan, P. Goodwin, M. Mohammadipour, A.A. Syntetos

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 been insufficiently specified, 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, IJF). The teams of researchers proceeded systematically, reporting results 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 & Williams. These discrepancies led to differences in the conclusions on 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, more generally, in its approach to the reproduction of forecasting research.

Keywords

Forecasting practice; Replication; Seasonal Forecasting; Empirical 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, StanfordUniversity).

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 found to have initially stronger effects. Prasad et al. (2013) analysed 363 articles testing standard 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 carry risks and limitations. Researchers may have inadvertently made errors in their application of methods. They 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 will 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) and hence inflate 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. Definitions of replicability vary 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 so, if it cannot be achieved, the generalizability of findings is likely to be in doubt.Of course, perfect reproduction of results may not be possible. For example, improvements in the algorithms embedded in software may lead to differences between the original numbers 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 successfully reproduced 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 reproduction or replicationof results is low in some disciplines. Evanschitzky, Baumgarth, Hubbard, & Armstrong (2007) found that, in marketing, the percentage of papers based on replication studies had halved to 1.2% in the period 1990 to 2004 when compared with 1974 to 1989. A similar study of empirical research papers in forecasting, published between 1996 and 2008, found an 8.4% rate (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 to support replication in forecasting research. Data sets, such as those used in the M1 forecasting competition, are easily accessible (Makridakis, Andersen, Carbone, Fildes, Hibon, Lewandowski, Newton, Parzen, & Winkler, 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 required 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 requires that "For empirical studies, the description of the method and the data should be sufficient to allow for replication." However, whether or not research is truly replicable may not be apparent until a full replication is formally attempted. 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 M1 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 required 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 impact 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 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 "re-analysis" to refer to an application of different methods on the same data or a sub-sample of the data. This constitutes a third category, in addition to "reproduction" and "replication", as shown in Figure 1 below.

	Same Methods	**Different Methods**
Same Data	Reproduction	Re-Analysis
Different Data	Replication	

Fig. 1. Reproduction, Replication and Re-Analysis

Similar distinctions between reproducibility and replicability have been drawn in other scientific disciplines (e.g. in psychology by Asendorpf et al., 2013). However, it should be noted that these terms are sometimes used differently by other authors. For example, Drummond (2009) used the terms in the opposite way to the above definitions.Evanschitzky et al. (2007), used the term "replication with extension" to indicate replication (in our terminology) but with a greater emphasis on generalisation.

Reproducibility is a necessary condition for replicability. An inability to reproduce the numerical results of a study implies that the methods used in that study have been insufficiently specified, thereby precluding replication. However, it is not a sufficient condition because the availability of further data meeting the necessary conditions is also required for a replication study to be conducted and for the qualitative findings to be replicated (e.g., in a forecasting context, *method A* is more accurate than *method B* under certain conditions.)

Another important issue that has not been addressed in forecasting research is 'exact reproducibility'. Does precision to, say, the second decimal place only but not to the third, constitute a reproduction of a previous result or not? Such differences may arise from the use of different optimisation algorithms in different software packages. In this paper, a further distinction is drawn between 'exact reproducibility' and 'approximate reproducibility'. Exact reproducibility corresponds to our previous definition of reproducibility. On the other hand, if it is claimed that findings are approximately reproducible to a certain percentage, then independent researchers should be able to obtain results that differ by no more than that percentage by repeating the original study.

3.The study by Miller & Williams

As previously discussed, the *International Journal of Forecasting* (IJF) is among those journals that support replication studies. Given that reproducibility is a necessary condition for replicability, we have focused on reproducing an important study published in the IJF, namely "Shrinkage estimators of time series seasonal factors and their effect on forecasting accuracy" (Miller & Williams, 2003). This paper won an outstanding paper award, 2002-2003,by the International Journal of Forecasting[1], and has been cited more than 25 times according to Google Scholar. It is also referred to in the well cited reviewbyDe Gooijer & Hyndman (2006) of the most important advancements in the recent history of forecasting.

The paper by Miller & Williams (2003) is not untypical in its documentation of forecasting procedures. The authors give details of their dataset, methods for estimating seasonal factors and accuracy measures. They also provide some information on parameter specification and prediction methods (although further details are needed on these topics, as discussed in Appendix A).

[1] The International Journal of Forecasting Best Paper Award for 2002–2003, *International Journal of Forecasting*, 22(4), p.825 (http://www.sciencedirect.com/science/article/pii/S0169207006000781).

The authors suggested two shrinkage methods to adjust the Classical Decomposition (CD) seasonal factors towards 1.0: the James-Stein (J-S) estimator and the Lemon-Krutchkoff (L-K) estimator (see Miller & Williams, 2003, pp. 671-672).

Using simulation on theoretically generated data, the conditions under which each of these methods are more accurate than Classical Decomposition were identified and guidelines for choice of method (CD, J-S, or L-K) were developed. In the empirical investigation on data series from the M1-Competition, each of the data series were categorized according to the recommended method, based on the proposed guidelines (Miller & Williams (2003), Table 5, p. 678). Forecasting accuracy results were presented for the set of all 55 series and subsets for which each of the methods had been recommended (Table 6, p. 680).

Some of the co-authors of the present paper tried to replicate the forecasting methods suggested by Miller & Williams on different data sets for another project funded by the Engineering and Physical Sciences Research Council (EPSRC, UK). As a precursor to this replication, they attempted to reproduce the results first. However, the results achieved were considerably different from the original ones. Consequently, another independent team was invited to attempt to reproduce the results. Hereafter, these two teams of forecasting experts will be called team A (team that commenced the study) and team B (team invited at a later stage).

This background motivated the following two research questions: i) how feasible is it to reproduce the results of this forecasting research paper?, and ii) how accurate (exact) is the reproduction?The remainder of this section is devoted to answering these questions.

3.1. The process of reproducing Miller & Williams' results

Team A used MATLAB (7.12) while team B used Visual Basic embedded in Microsoft Excel 2003. (Miller & Williams also used Microsoft Excel.) This choice was based on the expertise

of the teams, but later proved to be beneficial to the research, as it allowed for the quantification of the effects of different optimisation routines.

The assumptions and methodological stages of the original research paper are explained in page 679 of Miller & Williams (2003).Both teams fully documented all the working methods and assumptions made in the process of generating the results. The reproduction process is depicted graphically in Figure 2 and is explained below.

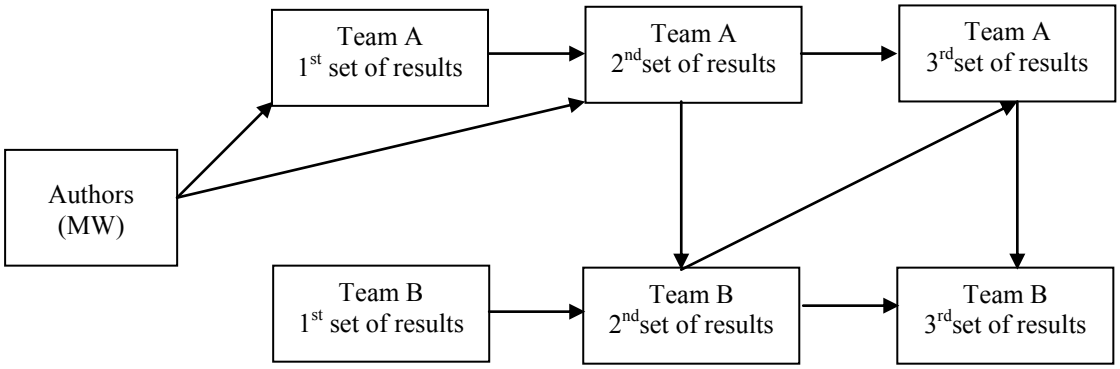

Fig. 2. Reproduction process

First, team A contacted Professors Miller and Williams (MW)seeking clarificationswith regard to the data series. The authors provided team A with the exact 55 series out of the 66 monthly series used in the M1-competition which they used in their study. Subsequently, team A produced the first set of results by making various assumptions regarding those issues about which they were unclear (see *Appendix A.1*). Then, they contacted MW again to resolve the issues raised in the first run and, based on this new information, they produced the second set of results (see *Appendix A.2*).

On the other hand, team B generated their first set of results using only the information given in the original paper and the 111 series from the M1 dataset without any contact with MW (see *Appendix A.4*). They selected the same series used by Team A, showing that MW had

provided sufficient information to allow specification of the exact 55 series. Then, team A provided team B with the additional information gained through their communication with MW and, based on that, team B produced their second set of results (see *Appendix A.5*).

After the first set of results were presented by team B and upon a review of the documented stages of their replication, team A found other discrepancies (see *Appendix A.3*). In their third run, team A attempted to repeat what team B did, by amending their experimental structure to match the assumptions and methods of team B. These stages along with the results produced in each of them are explained in detail in *Appendix A*. The notations used are the same as in Miller & Williams (2003).

As mentioned in *Appendix A.5*, even after further communications among the two teams,there were still discrepancies between their results (team A third set of results and team B second set of results in Table A.1). In order to investigate this issue, each of the 55 series has been checked individually (manually) to identify the series for which the Mean Absolute Percentage Error (MAPE) results produced by the two teams were different. The results are provided in *Appendix B*. It is found that the difference in the results is because different optimisation tools produce different smoothing parameters. The biggest difference related to one series (series 37) for which the Excel Solver optimisation stops at a local minimum. (This issue of Solver stopping at a point that is not a solution but reporting it as a solution was noted by McCullough & Wilson (2005)). As shown in *Appendix B*, the problem was addressed by changing the starting values and the results for team B were once again updated.

Comparing team A's and team B's third sets of results in Table A.1, it can be seen that now the results of team A and B are close. Out of 60 MAPE results, 23 are slightly different (with the absolute difference of two MAPEs ranging from 0.5% to5.0%), 12 of which have an absolute differenceof only 0.5% to 1.5%. It should be mentioned that this remaining difference isalso due to utilisation of different optimisation tools which result in

differentsmoothing parameters (see *Appendix B*). When the same parameters are used, the results of the two teams are exactly the same. Thus, approximate reproduction of results between team A and team B was obtained, and exact reproduction with identical parameters.

However, these final results of the two teams (team A's and team B's third sets of results in Table A.1) are very different from the results of MW (Table 6, p. 680). Reproduction of MW's results was not attained and the next sub-section reviews the reasons for these discrepancies.

3.2. Sources of discrepancies

As discussed in the previous sub-section, a number of issues were raised when trying to reproduce the results by Miller & Williams (2003). These can be classified as follows:

1. Data clarification: Team A had some difficulties identifying the exact 55 out of the 66 monthly series of the 111 series used in M-competition. Therefore, they asked the authors of the original paper for clarification and they were kindly provided with the exact 55 series. Team B, on the other hand, had no problem identifying the 55 series under concern using only the information provided in the original paper (bearing in mind that the 111 M-competition series are publicly available).
2. Methods clarification: As discussed in *Appendix A* (A.1 and A.3), the calculation of the coefficient of skewness, initialisation of the smoothing method and the use of rolling or non-rolling forecasts, which were not clarified in the original study, also resulted in discrepancies among the results.
3. Different software: The use of different software by teams A and B accounted for some of the differences. Results are reported in *Appendix B*showing that optimised parameters may differ between Excel and MATLAB, with Excel sometimes identifying local minima.
4. Accuracy measures: The fact that team A used the out-of-sample MAPE (for the first two sets of results) while team B used the in-sample MAPE for the purpose of selecting which

exponential smoothing method to use, also played an important role in obtaining different results.

All but one of the above issues could have been resolved by the provision of more information in the original study. Using different software is a separate issue, but this may also be accounted for by discussing the details of the package used for producing the results.

3.3. Implications for the findings by Miller & Williams

The emphasis of this paper has been on reproducing the numerical results of Miller & Williams (2003) and not their qualitative conclusions. However, in this section, the impact of the discrepancies (between the results produced by the two teams and the original results) on the conclusions reached by Miller & Williams are discussed.

Miller & Williams' primary hypothesis is that damping of seasonal factors improves on Classical Decomposition (CD), and this hypothesis is supported by our research. In *Appendix A*, Table A.1 (team A's and team B's third sets of results)shows that the Lemon-Krutchkoff (L-K) method produces lower MAPEs than CD when all 55 series are considered. However, Table A.1 (team A's and team B's third sets of results) also shows that the James-Stein (J-S) method produces higher MAPEs than CD for longer horizons, in contradiction to Miller & Williams.

Regarding the magnitude of the improvementsresulting from the use of the seasonal damping methods, MW mentioned that, compared to CD, J-S provided reductions in average MAPE ranging from 0 to 2.2%(which we believe should be 0 to 4.4% for their results). However, the two teams'resultsagree on only one case of improvement, for a 3-month horizon, which is no more than 1.6% . (Team A also identified a 0.12% improvement for a 6-month horizon).

MW also mentioned that L-K provided reductions in average MAPE ranging from 1.6% to 6.7%, when compared to CD, which is different to our results showingreductions in average

MAPE ranging from 0.4% to 5.2%. We do agree with MW that, when applied to all 55 series, L-K is the most accurate method on average.

As shown in *Appendix A* (A.1, A.2 and A.4), the numbers of series in each of the categories (L-K recommended, J-S recommended, and CD or J-S recommended) are different to those suggested by Miller & Williams. Teams A and B both report 30, 9 and 16 series, respectively, compared to 31, 10 and 14 series reported by Miller & Williams. MW concluded that, for the 31 series for which the recommended method of seasonal adjustment is Lemon-Krutchkoff, the use of L-K indeed leads to the smallest average MAPE. However, based on our results, there are two exceptions to this: for 12 and 18-months horizons, CD has the lowest MAPE rather than L-K.

For the 10 series for which the James-Stein method is recommended, MW mentioned that the use of J-S generally produces more accurate forecasts than the other two methods (there is only one exception to the rule which is for the 1-month horizon). On the other hand, our results show that this finding is valid only for the 3-month horizon.

For the 14 series for which the recommended method is J-S, but CD is also considered suitable, MW claimed that the choice of method for seasonal adjustment did not make a substantial difference in forecasting accuracy. Their results also show that the use of J-S leads to the smallest average MAPE. However, our results show that, compared to L-K,J-S provides more accurate forecasts for none of the series (L-K is the best method for all the horizons). Also, the results of the methods for seasonal adjustment are not insubstantial (for MW results, the difference between results are at most 3.7%, but this gets as high as 11.48% for our results).

To conclude this section, the results from teams A and B support the hypothesis that the Lemon-Krutchkoff method is more accurate than Classical Decomposition but not that the James-Stein method is more accurate than Classical Decomposition. Moreover, our results do

not support the guidelines suggested by Miller & Williams (Table 5, p. 678). This is because, based on our results, there are many exceptions for each category: for L-K recommended series there are 2 out of 5 horizons for which L-K is not the best method; for J-S recommended series J-S is best only for 1 out of 5 horizons; and, for CD or J-S recommended series, J-S does not perform better than the other two methods for any of the horizons. More work is needed on understanding the conditions under which seasonal damping methods outperform Classical Decomposition.

4. Specification of forecasting methods

4.1. Comparison of approaches

It is common for authors of forecasting papers to include statements of methods, including assumptions, in words (textual descriptions). However, it may be very difficult for others to translate these words into an unambiguous form for reproduction of results, replication of findings oradaptation of methods. To address this issue, some alternative approaches are discussed in this section.

One way of presenting methods is through the use of flow charts. A flow chart is a type of diagram that presents a method in algorithmic form, showing the steps as boxes of various kinds, and their order by connecting them with arrows. They are used extensively in simulation modelling (e.g.Hayes, Leal, Gray, Holman, & Clarke, 2013) but not so widely in forecasting. Another alternative is that the code itself may be offered alongside an academic paper. The internet is a significant aid to those who wish to make their data and algorithms available, for example by the use of journals' electronic companions[2].

[2]For example, recently, "Information and Inference: A Journal of the IMA", which publishes mathematically-oriented papers, has asked authors of papers with computational simulations / plots / tables to include their code in a format with a reference manual or a brief user guide. To encourage this, the journal has followed an existing standard that papers with accompanying code are marked as 'reproducible', which is indicated by a small diamond containing the letter R (see, for instance, http://imaiai.oxfordjournals.org/content/2/1/69.full.pdf+html).

Both flowcharts and code have advantages and disadvantages in facilitating reproduction and replication. Making code available guarantees exact reproduction of results while a flowchart may allow for only approximate reproduction. Nevertheless exact reproduction using provided code may conceal errors, which might otherwise be revealed if new independent code is developed. In replication some small changes to code may be needed to cater for new data sets (e.g. different sample sizes), but the effort involved in carrying out the replication will be relatively small. Providing a flowchart will necessitate the development of new code with the attendant dangers of introducing programming errors, which may be less easy to identify than in reproduction, given the absence of a set of earlier results based on the same data.

If the methods need to be adapted, using flowcharts and developing new code may be easier than adjusting code developed by other researchers. A flowchart is more accessible than code and requires only a basic understanding of the flowcharting rules and conventions. It is easy and quick to read and apprehend. On the other hand, using code requires an understanding of the language of the code, which may need a significant time to acquire. Another concern about provision of code is that people's knowledge will affect its accessibility. For example, there are fewer people today who are able to read code in APL[3] than 30 years ago.

Flowcharting and provision of code are not mutually exclusive. On the contrary, they are complementary. Some researchers may wish to reproduce or replicate without adaptation of methods. Other researchers may wish to experiment with adaptations of forecasting methods. Provision of flowcharts and code caters for both research audiences.

To summarise, textual description of methods and assumptions has been a common approach in forecasting studies. This approach was also adopted by MW in the research analysed in this

---

[3] A Programming Language (K. E. Iverson, A Programming Language, John Wiley and Sons, Inc., 1962)

paper. However, our results in Section 3 show thatreproduction of the results of MW's research was not possible based on the information provided in the paper. As discussed in this sub-section, alternative approaches,such as flowcharts and provision of code, may facilitate reproduction, replication and adaptation. The application of flowcharting to the research presented in this paper will be discussed in the next sub-section.

4.2. Flowchart for reproducibility

As explained, flowcharts are very accessible and easy to understand and, although they have not been widely used in forecasting studies, they can be easily implemented.

We have presented the detailed flowchart for the methods analysed in this paper in *Appendix C* using the information gathered from the authors of the original paper and the communications between the two teams. The flowchart consists of four blocks as shown below (Figure 3).

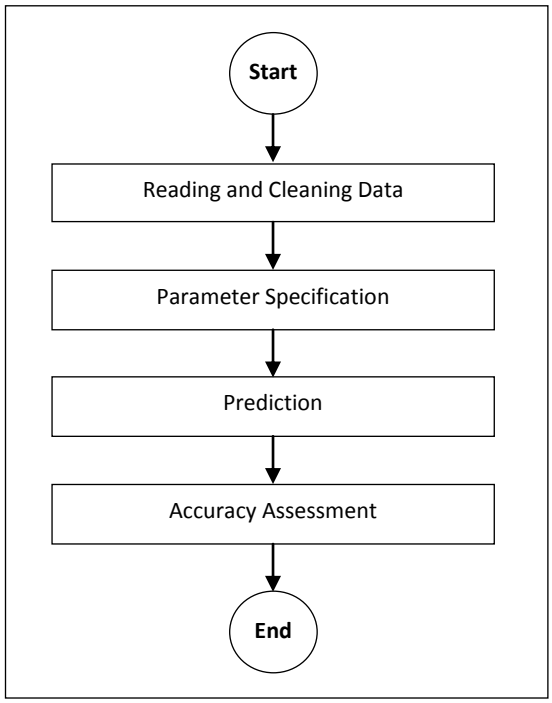

Fig. 3. Flowchart of the forecasting process

Each of the four blocks contains sequences of stages shown in detail in *Appendix C*. The blocks could be used for a variety of forecasting approaches. For example, parameter specification for smoothing methods (which has been used here) includes initialisation and optimisation whereas, forthe Box-Jenkins approach, it contains identification and estimation. Distinguishing these blocks in the code would also increase the clarity of the code and facilitate understanding and adaptation for reproducing the results and/or replicating the findings.

We believe that any independent researcher who wishes to approximately reproduce our results should be able to do so based on the flowcharts (see Figure 3 and Figures C.1, C.2, C.3 and C.4 in *Appendix C*). In addition, in order to facilitate exact reproduction of the methods used in our research, both the MATLAB code used to generate the results by team A and the Visual Basic (embedded in Microsoft Excel 2003) code and the Excel analysis used by team B are available in the electronic companion to this paper (please refer to *Appendix D*).

5. Conclusions and implications

In this study we have attempted to reproduce the results provided by Miller & Williams (MW, 2003). Our aim was to assess the feasibility and accuracy of doing so. It is important to emphasize that the methodsin the MW paper were not untypicalin their fullness of documentation, compared to other papers in the forecasting literature. Hence, the MW paper may be regarded as representative of method documentation in forecasting research.

We have worked in two teams (each of which attempted independently to reproduce the MW results) and in a structured way that allowed for the progressive accumulation of information relevant to the data and methods used in the MW study. Although the two teams reached almost the same results, those were different from the results provided by MW and we have not arrived at the same conclusions as the original paper. This provides an example of where lack of reproduction of results matters in terms of replication of the findings and conclusions.

It is also important to note that the two teams did not achieve exact reproduction of each other's results, because of differences in software optimisation methods.

Based on the outcomes of this work, we believe that there is considerable scope for improving the reproducibility of forecasting research papers in general and papers published by the *International Journal of Forecasting* (IJF) in particular. The IJF requires that "for empirical studies, the description of the method and the data should be sufficient to allow for replication". However, in practice, it is uncommon for the reviewers or the editorial office to request details that are sufficient to reproduce the results. Consequently, there is an overreliance of the academic community on the goodwill of the authors of the original studies to answer simulation related queries, provide empirical data and clarify methodological issues.

In an attempt to enable other researchers to reproduce, replicate or adapt the methods used by MW, we have provided a fully documented flowchart of the methods in the paper. We argued that flow-charts are accessible to a broader audience of forecasting practitioners and researchers than provision of code. However, we suggested that flowcharts and codes are complementary in providing high level understanding and granular appreciation of forecasting methods. To that end, we have supplemented our paper with electronic companions that include both the flowcharts and the code written by the two teams of researchers.

We would like to close our paper by inviting other researchers to attempt to reproduce our results. This would enable the approach to reproducibility proposed in this paper to be tested and commented upon by others. We also acknowledge that the issues discussed in this paper arise from a single research study and we would encourage researchers to attempt to reproduce other important forecasting studies and expand on the recommendations made in this paper. Finally, and most importantly, we would encourage authors (including ourselves)

to consider the issues of reproducibility and replication when documenting forecasting procedures and experimental structures employed for their research.

Acknowledgements

We are grateful to Don Miller and Dan Williams for their support during our attempts to reproduce their work and all the information they have provided. We would also like to thank the participants of the 33rd International Symposium on Forecasting (June 23-26, 2013, Seoul, Korea) for their constructive comments and interesting remarks on this work, as well as the following researchers for comments on earlier drafts of this paper: Zied Babai and Olivier Dupouet (Kedge Business School, France), Nikos Kourentzes (Lancaster University, UK) and Mohamed Naim (Cardiff University, UK). The research described in this paper has been supported by the Engineering and Physical Sciences Research Council (EPSRC, UK) grant no. EP/G003858/1, a project entitled '*Using the Grouping and Shrinkage Approaches to Forecasting Sub-Aggregate Level Seasonal Demand*'.

APPENDIX A. Reproduction of results

This appendix presents the results produced by team A and B in each round as shown in Figure 2.

A.1. Team A first set of results

Team A first attempted to reproduce the empirical results of the original study using the exact 55 series obtained from the authors and by making some assumptions regarding the issues about which they were unclear. In particular, assumptions were made with regards to:

- The formula used to calculate the coefficient of skewness.
- The initialisation of the three exponential smoothing methods, namely: simple exponential smoothing (SES), Holt's method, and damped-trend.
- Whether the J-S and L-K seasonal factors should also be adjusted to average 1.0 (similar to the adjustment by MW for CD).

Team A used the MATLAB function to calculate the coefficient of skewness:

$$\text{Skewness}_{\text{MATLAB}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\right)^3} \qquad \text{(A.1)}$$

where X_i is the observation at time $i = 1,...,n$, and $\bar{X}$ is the sample mean of the n observations. It was also assumed that, by skewness, MW meant absolute skewness.

The SES method was initialized by assuming that the first forecast for the deseasonalised series was the first deseasonalised value, $F_1 = X_1$ (Makridakis, Wheelwright, & Hyndman, 1998). For Holt's method and damped-trend, it was assumed that the initial level (L_1) is the first deseasonalised observation, $L_1 = X_1$, and the initial trend (b_1) is the difference between the first two deseasonalised observations, $b_1 = X_2 - X_1$ (Makridakis et al., 1998).The *fmincon* function in MATLAB was used to obtain the smoothing parameter values that minimize the

in-sample Mean Squared Error (MSE) (with the constraints being identical to the authors' bounds on these parameter values).

Team A assumed that the J-S and L-K seasonal factors should be adjusted to average 1.0. (This was not specified by MW.)

The first set of team A's results, based on the above assumptions and formulae, is presented in Table A.1. All results relate to Mean Absolute Percentage Errors (MAPEs).Comparing this to Table 6 of Miller & Williams (2003), it can be seen that not only are the MAPE results very different, but also the number of series in each category (L-K, J-S, and CD or J-S recommended) are not the same. Given the inconsistencies, team A further contacted the authors to resolve a number of issues which may have resulted in these discrepancies.

A.2. Team A second set of results

In the second communication with MW, some important points were clarified. The authors advised team A that they used the following formula to calculate the coefficient of skewness:

$$\text{Skewness}_{MW} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^3}{S^3} \qquad \text{(A.2)}$$

where S is the sample standard deviation (A.3). The difference between equations (A.1) and (A.2) is that in equation (A.2) the standard deviation is calculated by:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \qquad \text{(A.3)}$$

Also, MW used the following equations to initialise trend and level respectively:

$$b_1 = \frac{\frac{1}{12}\sum_{i=13}^{24} X_i - \frac{1}{12}\sum_{i=1}^{12} X_i}{12} \qquad \text{(A.4)}$$

$$L_1 = \frac{1}{12}\sum_{i=1}^{12} X_i - 5.5b_1 \qquad \text{(A.5)}$$

The second set of team A's results were generated based on the above information (see Table

A.1). Despite the additional information, both the actual MAPE results and the number of series identified in each category were different to those reported by MW. At this point, team A invited an independent team (Team B) to reproduce the results (see *Appendix A.4*).

A.3. Team A third set of results

After team B produced their first set of results and following communications between teams A and B, some other sources of discrepancy were identified:

- Use of in-sample or out-of-sample MAPE to select the exponential smoothing method: Team A had used out-of-sample MAPE in their first two attempts, while team B used in-sample MAPE.
- Use of rolling or non-rolling forecasts for the hold-out data: Team A had used rolling forecasts, while team B used non-rolling forecasts.
- Team A also realised that using different starting values when optimizing the parameters for the smoothing methods would result in different optimum parameters. Therefore they used the same starting value as team B for all parameters (which was 0.1) except for the damping parameter (which was 0.9).

In an attempt to reach agreement with the results produced by team B, team A ran a third experiment using the in-sample MAPE for methods' selection, non-rolling forecasts and the above discussed starting value for optimisation. The third set of team A results are presented in Table A.1.

A.4. Team B first set of results

Initially, team B was provided with only a copy of the paper and the entire M1 dataset and asked to reproduce the empirical results without any further information. They were asked to disclose their working methods and assumptions in doing so. The results provided by team B did not match either team A's results or the ones provided by MW.

Based on the fact that MW used Excel 2003 for their study, team B assumed that the Excel

coefficient of skewness should be used:

$$\text{Skewness}_{\text{Excel}} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S} \right)^3 \qquad \text{(A.6)}$$

where the notation is unchanged. It can be seen that equation (A.6) is different from both equations (A.1) and (A.2). Further, team B assumed that the coefficient of skewness was only measured on the CD seasonal factors and that MW meant absolute skewness when describing their shrinkage selection criteria (otherwise strong negative skewness would be regarded as symmetry according to the authors' stated criterion).

With regards to initialization, team B used the same initialization that team A employed in their first attempt (see *Appendix A.1*). The Excel Solver was used to identify the MSE-minimizing parameter values.(The Solver Options were as follows: Max Time = 100 seconds; Iterations = 100; Precision = 0.000001; Tolerance = 5%; Convergence = 0.0001; do not assume linear model, automatic scaling, or non-negative; Estimates = Tangent; Derivatives = Forward; Search = Newton). Team B used the same approach as team A in adjusting the J-S and L-K seasonal factors to average 1.0.

The first set of team B's results were generated based on the above initialization and skewness related assumptions and are shown in the Table A.1. Both the MAPE results and the number of series identified in each category were different to those of MW and team A's first and second round results. After further communication with team A, team B received the additional information team A had received from MW (discussed in *Appendix A.2*) and made a second attempt to reproduce the results.

A.5. Team B second set of results

In the second round, team B was provided with the coefficient of skewness formula and starting estimates of the levels and trends that MW used. The second set of team B's results is presented in Table A.1.Comparing these results with the third set of team A results, the two

teams managed to achieve the same number of series assigned to different shrinkage methods which is still different to that reported by MW. However, there were discrepancies between their final results and this issue is further investigated in *Appendix B*.

Table A.1

Mean Absolute Percentage Errors for Team A and Team B

The results from team A were produced with their own initialisation procedure and the use of MATLAB for the calculation of the coefficient of skewness. The results for 'All 55 series' are shown to three decimal places, for consistency of presentation with Table 6 of Miller & Williams (2003).

	Horizon	L-K recommended			J-S recommended			CD or J-S recommended			All 55 series		
		CD	J-S	L-K	CD	J-S	L-K	CD	J-S	L-K	CD	J-S	L-K
Team A 1st set of results		**34 series**			**8 series**			**13 series**			**55 series**		
	1	8.99	9.05	8.91	8.22	7.56	7.82	5.37	5.32	5.39	8.024	7.954	7.919
	3	9.53	9.63	9.45	9.38	8.54	8.82	5.55	5.42	5.50	8.566	8.476	8.425
	6	10.20	10.40	10.13	10.82	9.86	10.16	5.94	5.82	5.92	9.284	9.236	9.142
	12	11.08	11.40	10.99	11.46	10.64	10.89	6.77	6.63	6.75	10.113	10.160	9.975
	18	11.04	11.20	10.74	14.00	14.23	13.95	7.86	7.66	7.80	10.715	10.804	10.511
Team A 2nd set of results		**30 series**			**9 series**			**16 series**			**55 series**		
	1	8.95	8.94	8.86	7.48	6.88	7.11	6.76	6.69	6.67	8.071	7.949	7.936
	3	9.49	9.52	9.45	8.58	7.82	8.06	6.99	6.92	6.86	8.615	8.486	8.470
	6	10.26	10.32	10.22	9.95	9.08	9.34	7.60	7.54	7.49	9.434	9.308	9.284
	12	11.05	11.00	11.03	10.85	10.09	10.33	8.61	8.49	8.51	10.306	10.122	10.181
	18	10.97	10.62	10.77	12.92	12.98	12.80	9.62	9.64	9.44	10.896	10.725	10.717
Team A 3rd set of results		**30 series**			**9 series**			**16 series**			**55 series**		
	1	7.15	7.24	6.73	7.53	8.98	8.93	6.47	6.97	6.17	7.016	7.446	6.929
	3	8.00	7.95	7.53	10.96	10.06	10.40	7.34	7.49	7.04	8.291	8.160	7.860
	6	9.49	9.51	9.18	11.48	11.41	11.34	7.45	7.42	7.25	9.222	9.211	8.969
	12	12.35	12.61	12.45	13.42	13.37	13.12	8.66	8.48	8.44	11.451	11.533	11.396
	18	13.96	14.10	14.05	14.92	15.08	14.65	10.14	10.01	9.93	13.004	13.071	12.950
Team B 1st set of results		**36 series**			**8 series**			**11 series**			**55 series**		
	1	7.21	7.32	6.90	8.22	9.38	9.55	5.86	5.74	5.86	7.085	7.301	7.078
	3	7.82	7.88	7.43	12.11	10.84	11.43	6.85	6.64	6.95	8.249	8.065	7.915
	6	9.21	9.25	8.86	12.66	12.53	12.59	7.01	6.88	7.08	9.270	9.251	9.045
	12	11.80	11.66	11.47	14.31	13.78	13.92	8.34	8.20	8.43	11.475	11.276	11.222
	18	13.34	13.06	12.90	15.39	15.19	15.17	9.83	9.71	9.92	12.939	12.698	12.633
Team B 2nd set of results		**30 series**			**9 series**			**16 series**			**55 series**		
	1	7.14	7.24	6.72	7.53	8.98	8.93	6.31	6.97	6.01	6.964	7.446	6.874
	3	7.99	7.95	7.52	10.96	10.06	10.40	7.18	7.49	6.88	8.239	8.160	7.804
	6	9.46	9.51	9.15	11.49	11.41	11.34	7.30	7.42	7.09	9.164	9.211	8.908
	12	12.31	12.61	12.42	13.43	13.37	13.12	8.50	8.48	8.29	11.388	11.533	11.330
	18	13.91	14.10	14.00	14.93	15.08	14.65	9.99	10.01	9.79	12.937	13.071	12.880
Team B 3rd set of results		**30 series**			**9 series**			**16 series**			**55 series**		
	1	7.14	7.24	6.72	7.53	8.98	8.93	6.47	6.97	6.17	7.010	7.446	6.922
	3	7.99	7.95	7.52	10.96	10.06	10.40	7.34	7.49	7.04	8.284	8.160	7.851
	6	9.46	9.51	9.15	11.49	11.41	11.34	7.45	7.42	7.25	9.209	9.211	8.955
	12	12.31	12.61	12.42	13.43	13.37	13.12	8.65	8.48	8.44	11.431	11.533	11.375
	18	13.91	14.10	14.00	14.93	15.08	14.65	10.13	10.01	9.93	12.978	13.071	12.923

Appendix B. Comparison of team A's and team B's results

In order to examine the differences between team A's (third set) and B's (second set) results, each of the 55 series has been checked individually (manually) to identify the series for which the MAPE results produced by the two teams were different. This was the case for 10 series. The accuracy measure used was the absolute difference (AD) of the MAPE produced by the two teams: $|MAPE_{teamA} - MAPE_{teamB}|$.

All the intermediate results were compared for those 10 series and this enabled the classification of the sources of discrepancy into three categories:

- The smoothing parameters, which give the smallest in-sample MSE, are not very different but this affects the selection of the smoothing method which in turn affects the final results (0.01 to 0.09 in terms of absolute difference) (series 46);
- The smoothing parameters are very different but although the same smoothing method is selected, the final results are very different too (up to 2.63 in terms of absolute difference) (series 8 and 37);
- The smoothing parameters are slightly different and this does not affect the selection of smoothing method, but the final results are also different (0.01 to 0.15 in terms of absolute difference) (series 5, 9, 17, 18, 29, 32 and 35).

The two teams realised that using different optimisation tools (*fmincon* in MATLAB by team A and Excel Solver by team B) is the reason for obtaining different smoothing parameters. This is despite the fact that both tools are minimizing the same function and using the same boundaries for the parameters. It has been found that, in all cases, the MATLAB-produced parameters give the smallest in-sample MSE. This could result in a different smoothing method as for series 46. In this case, the 0.0019 absolute difference of the in-sample MAPE for the Holt's method between the two teams, results in team A selecting Holt's method while team B selects the damped trend.

On the other hand, series 32 from the third category which has the highest absolute difference (0.15) may be considered. Both teams have selected the Holt's method for Classical Decomposition but the 0.0044 AD in α (the smoothing constant for smoothing the level in Holt's method) and 0.0144 AD in β (the smoothing constant for smoothing the trend in Holt's method) leads to 0.0056 AD for the in-sample MAPE and 0.15 to 0.07 AD for the out-of-sample MAPEs for horizons 1 to 18 respectively.

For the second category, which has the highest amount of discrepancies, series 37 has the highest difference in MAPEs between the two teams. Again, both teams have selected Holt's method for L-K seasonal adjustment, but there is a 0.3418 AD in α. This leads to 0.0676 AD in the in-sample MAPEs but 2.63 to 2.35 AD in the out-of-sample MAPEs for horizons 1 to 18 respectively.

Further examination reveals that the difference in smoothing parameters for series 37 (CD and LK results) was because the Excel solver optimisation stops at a local minimum when starting from α equal to 0.1 (for Holt's method). It can be seen in Table B.1 that changing the starting value to 0.01 would result in selecting 0.01 which is the global optimum for the specific range of α $(0.01 \le \alpha \le 0.9)$.

Table B.1
The effect of using different starting values in EXCEL optimisation.

	Holt's CD		**Holt's LK**	
Starting Value	0.1	0.01	0.1	0.01
α	0.249703	0.01	0.351805	0.01
β	0	0	0	0
In-sample MSE	963415.8	937052.1	978913.5	964858.5

Correcting for series 37, the third and final set of results by team B is summarized in Table A.1.

Appendix C. Flowchart

The flowchart of the forecasting process(Figure 3) is expanded in this Appendix.

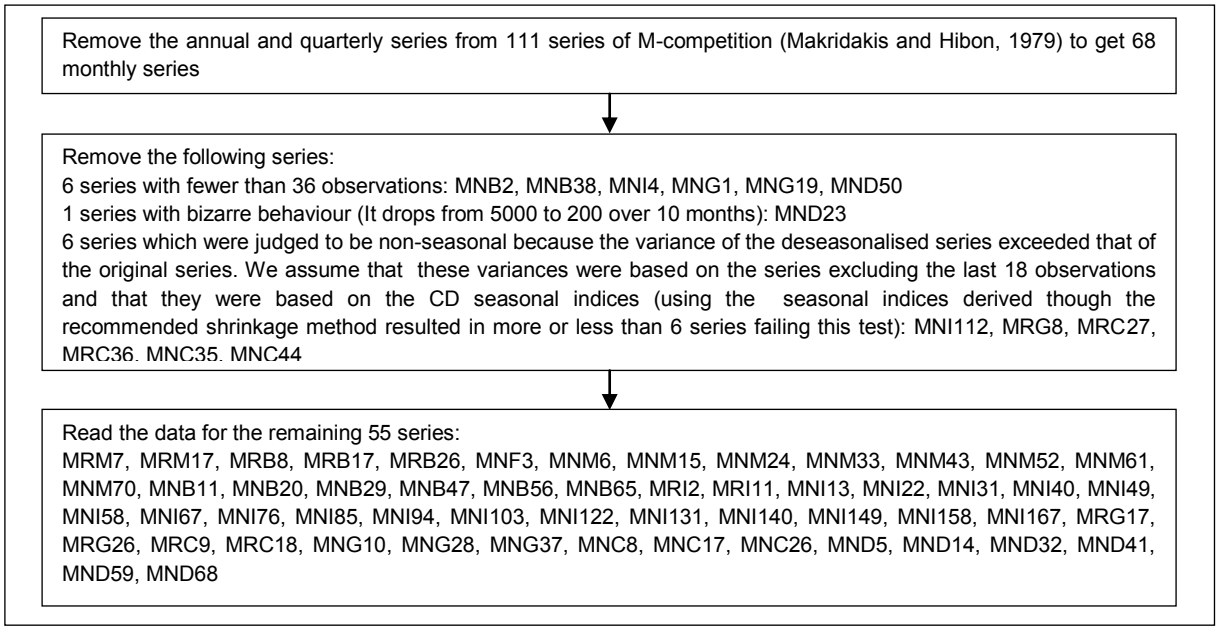

Fig. C.1. Reading and Cleaning Data

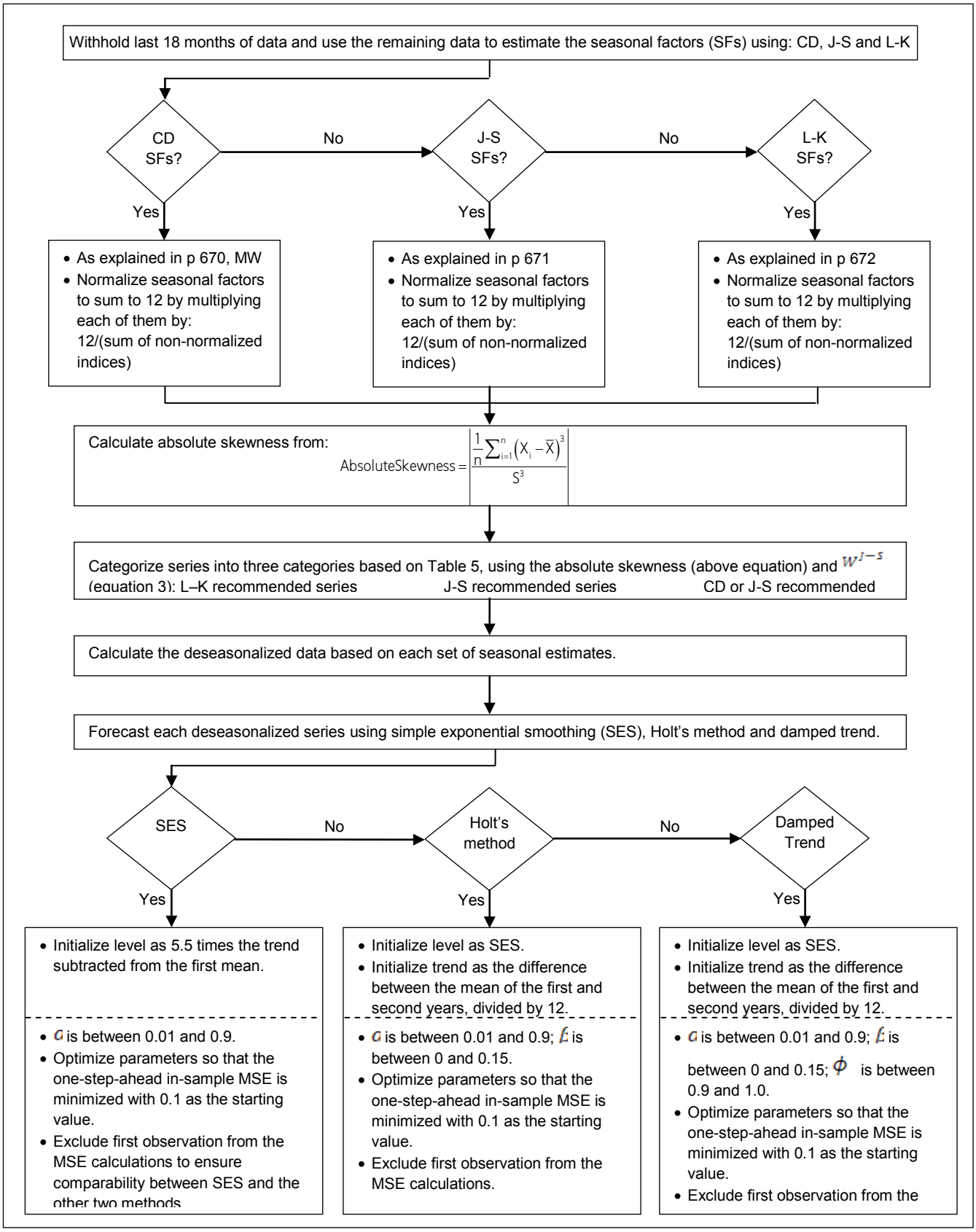

Fig. C.2. Parameter Specification

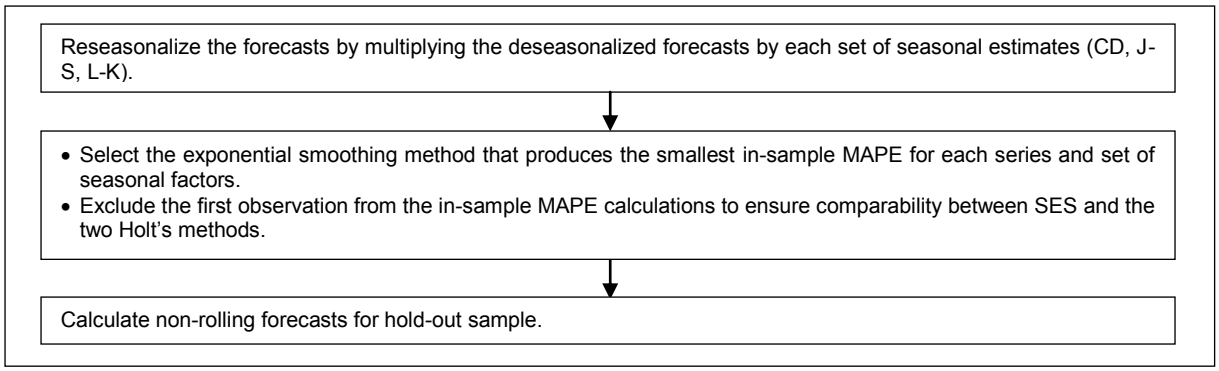

Fig. C.3. Prediction

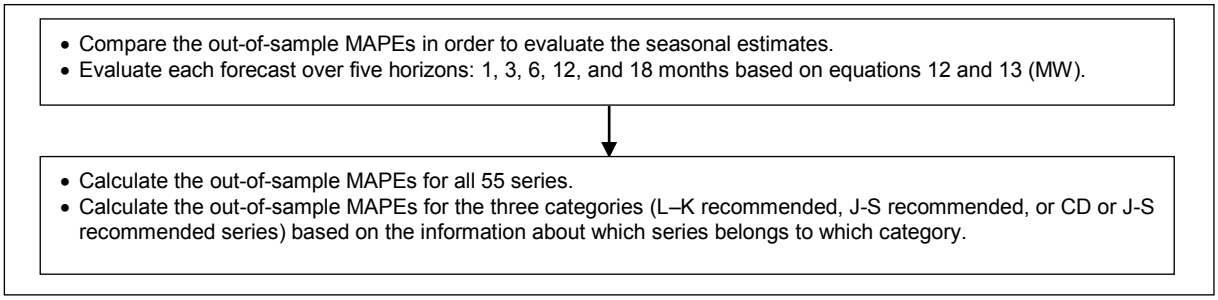

Fig. C.4. Accuracy Assessment

Appendix D. Supplementary material

The following supplementary material related to this article is included in the website of the *International Journal of Forecasting*:

1. A colour version of the flow chart presented in Figure 2 and *Appendix C*;
2. The MATLAB (7.12) code used to generate the results produced by team A;
3. The Visual Basic (embedded in the Microsoft Excel 2003) code and the Excel analysis used to generate the results produced by team B.

References

Asendorpf, J. B., Conner, M., Fruyt, F. D. E., Houwer, J. A. N. D. E., Denissen, J. J. A., Fiedler, K., … Vanaken, M. A. G. (2013). Recommendations for Increasing Replicability in Psychology. *European Journal of Personality*, 27(2), 108–119.

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 Years of Time Series Forecasting. *International Journal of Forecasting*, 22(3), 443–473.

Drummond, C. (2009). Replicability is not Reproducibility: Nor is it Good Science. Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML, Montreal, Canada.

Evanschitzky, H., & Armstrong, J. S. (2010). Replications of forecasting research. *International Journal of Forecasting*, 26(1), 4–8.

Evanschitzky, H., Baumgarth, C., Hubbard, R., & Armstrong, J. S. (2007). Replication Research's Disturbing Trend. *Journal of Business Research*, 60(4), 411–415.

Hayes, A. J., Leal, J., Gray, A. M., Holman, R. R., & Clarke, P. M. (2013). UKPDS Outcomes Model 2: A New Version of a Model to Simulate Lifetime Health Outcomes of Patients with Type 2 Diabetes Mellitus using Data from the 30 year United Kingdom Prospective Diabetes Study: UKPDS 82. *Diabetologia*, 56(9), 1925–33.

Hubbard, R., & Armstrong, J. S. (1994). Replications and Extensions in Marketing – Rarely Published But Quite Contrary. *International Journal of Research in Marketing*, 11(3), 233–248.

Ioannidis J.P.A. (2005) Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. *Journal of the American Medical Association*, 294(2), 218-228.

King G. (1995) Replication, Replication. *PS: Political Science & Politics*, 28(3), 443-452.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1984). The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting*, 1(2), 111- 153.

Makridakis, S., & Hibon, M. (1979). Accuracy of forecasting: an empirical investigation. *Journal of the Royal Statistical Society, Series A*, 142, Part 2, 97-145.

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

McCullough, B. (2000). Is it Safe to Assume that Software is Accurate? *International Journal of Forecasting*, 16(3), 349–357.

McCullough, B.D. & Wilson, B. (2005). On the accuracy of statistical procedures in Microsoft Excel 2003. *Computational Statistics & Data Analysis*, 49(4), 1244-1252.

Miller, D. M., & Williams, D. (2003). Shrinkage Estimators of Time Series Seasonal Factors and their Effect on Forecasting Accuracy. *International Journal of Forecasting*, 19(4), 669–684.

Prasad, V., Vandross, A., Toomey, C., Cheung, M., Rho, J., Quinn, S., ... & Cifu, A. (2013). A decade of reversal: an analysis of 146 contradicted medical practices. *Mayo Clinic Proceedings*, 88(8), 790–798. Elsevier.

Simmons, L. F. (1986). M-Competition - A Closer Look at Naïve2 and Median APE: A Note. *International Journal of Forecasting*, 2(4), 457–460.