Is there a Golden Rule?

Robert Fildes and Fotios Petropoulos

Lancaster Centre for Forecasting, Department of Management Science, Lancaster University, UK

r.fildes@lancaster.ac.uk; f.petropoulos@lancaster.ac.uk

June 2014

Abstract

Armstrong, Green and Graefe (this issue) propose the Golden Rule in forecasting: "be conservative". According to the authors, the successful application of the Golden Rule comes through a checklist of 28 guidelines. Even if the authors of this commentary embrace the main ideas around the Golden Rule, which targets to address the "average" situation, they believe that this rule should not be applied automatically. There is no universal extrapolation method that can tackle every forecasting problem; nor are there simple rules that automatically apply without reference to the data. Similarly, it is demonstrated that for a specific causal regression model the recommended conservative rule leads to unnecessary inaccuracy. In this commentary the authors demonstrate, using the power of counter examples, two cases where the Golden Rule fails. Forecasting performance is context-dependent and, as such, forecasters (researchers and practitioners) should take into account the specific features of the situation faced.

Keywords: forecasting, time series, ARIMA, regression modelling, forecasting performance, model specification

Acknowledgments

The authors would like to thank Scott Armstrong, Kesten Green, Andreas Graefe, Geoffrey Allen, Everette Gardner Jr., Paul Goodwin, Konstantinos Nikolopoulos, and Keith Ord for their constructive comments on earlier versions of this commentary.

Introduction

Students of Scott Armstrong's works, which include the authors of this commentary, will be familiar with the themes presented here. These have, in some ways, been with him throughout his research career, and can be summarized as: simple methods work best. This is most recently articulated in Armstrong (2006) but goes back at least as far as Armstrong and Shapiro (1974). The scepticism shown has many benefits, not least the requirement to always compare new forecasting methods with simple or established alternatives. The results have often been provocative and resisted as some of the tales in the Armstrong, Green and Graefe (this issue, henceforward AGG) paper make clear (e.g. the response of the statistical community to the Makridakis forecasting competitions (Fildes and Makridakis, 1995)). The Armstrong campaign has not been uniformly welcomed: people's careers are often built on proposing new methods, which with editorial bias towards publishing only positive findings and the neglect of replication (Evanschitzky and Armstrong, 2010), can lead to whole communities chasing after red herrings. So, unequivocally the argument proposed by AGG is interesting. However, the devil lies in the detail. The authors of this commentary provide evidence in two specific cases that the Golden Rule does not necessarily deliver forecasts as accurate as are achievable using other (standard) approaches. The strong version of the rule as posed is, they argue, only true for the "average" situation: there are likely to be many exceptions. In the conclusions the purpose of a "Golden Rule" will be revisited asking the question as who might find it helpful.

Extrapolative methods: the Golden Rule and the accuracy of extrapolation methods

Perhaps the most frequent application of statistical forecasting methods is the use of extrapolative methods in demand planning. Much commercial software has been developed that integrates a wide range of methods, including methods that potentially incorporate trend

and seasonal components in the data. Because the forecaster often faces the problem of forecasting many series (a median estimate of 400 time series, according to Fildes and Goodwin 2007), there is little time to analyse each series individually. Standard methods that have been applied include moving averages, exponential smoothing, the ARIMA methods of Box and Jenkins and benchmark (naïve) methods. Various competitions have provided valuable evidence on the accuracy of these methods and these are used by AGG to provide support for various guidelines on following the Golden Rule.

In an attempt to evaluate the Golden Rule as it applies to extrapolative forecasting methods, a subset of the monthly data used in the M3-Competition (Makridakis and Hibon, 2000) is reanalyzed. The lengthier of these series have been considered (those with 126 or more observations), 998 series in total (available on the Journal's web site). The same subset was used in Fildes and Petropoulos (this issue, hereafter FP).

First, the methods that are intrinsically conservative should be defined. In this comparison the Naïve applied on the original and the deseasonalized data (Naive and DNaive respectively), Simple Exponential Smoothing (SES), Holt, Holt-Winters and Damped (applied on the deseasonalized data, hereafter DDamped) methods are considered. The automatic ARIMA algorithm is also employed, as implemented by Hyndman and Khandakar (2008). These methods are typical of the range of methods available in commercial software, but exclude the proprietorial methods, such as Forecast Pro's expert selection, included in the M3 comparisons.

Translating the Golden Rule's recommendations into a model selection strategy is not straightforward when the problem faced is one of automatic selection and nothing is known other than the data are monthly. A key issue is whether the Golden Rule is also an implicit recommendation to choose simple methods or is a recommendation to favor well-performing

empirical methods. Simple exponential smoothing could well be regarded as conservative under either of these interpretations since a weighted average of all past observations (with trend and seasonal components removed) has also proved a strong competitor in the various M-Competitions. The two Naïve methods, performing well in these earlier competitions, also partially fulfil the Golden Rule's criteria (though whether they are conservative rather than merely simple when compared with the unweighted average of the data seems an open question). Clearly, ARIMA, Holt-Winters and Damped (on deseasonalized data) all fail the Rule, as they necessarily include the possibility of both trend and seasonal. Damped however accords with the Rule with respect to trend and past performance in empirical evaluations.

The data have been split into a training/estimation set (observations 1 to 48), a validation set (49 to 90) and a test set (91 to 126), discarding the last observations in the longer series. A rolling-origin design is employed, where the training sample is increasing at every step by one observation. The errors are calculated on the validation and test data. See FP for further discussion. A (Geometric) Average Relative Mean Absolute Error has been used: AvgRelMAE (Davydenko and Fildes, 2013), following AGG's recommendations on using relative error measures. This offers an easy interpretation of whether one method outperforms another and has proved to be both robust and sensitive. The performance of each method is related to that of DNaive, so a value greater than 1 shows worse average performance compared to the Naive on the deseasonalized data.

The series have been segmented into various sub-populations as research has established that comparative performance of different methods varies according to the population of time series being studied (see, for example, Fildes et al. (1998) and their discussion of robust trend). The segmentations are defined on the training and validation data as follows: (i) predictable (versus unpredictable) for those series where the non-seasonal Random Walk forecasting method (Naive1) delivers worse than the median performance of all other

methods under investigation as defined by Mean Absolute Error in the validation data (ii) trending (versus non-trending), where there is a significant trend based on the Cox-Stuart test over the test and validation data (data 1 to 90), (iii) seasonal (versus non-seasonal), defined by significance using Friedman's non-parametric test, and (iv) stable (versus unstable), where there is a high correlation between the ranks of methods over the validation data set. These definitions are given more fully in FP.

Table 1 here

The results (in Table 1) show that Damped on deseasonalized data (DDamped) performs best overall. However, the performance of ARIMA is very good, second best across almost all segments. DNaïve (where the forecast is the same as the last value of the seasonally adjusted data, multiplied by the respective seasonal index) also does well overall. Crucially for the argument, for some segments such as stable (defined as the set of series where performance remains much the same over the validation data set), ARIMA performs much better (about 35%) than simpler methods (e.g. SES or Holt). Damped (on deseasonalized data) and ARIMA are the only methods to outperform the naïve benchmark on seasonal and trending series, suggesting that these characteristics are sufficiently persistent as to overcome the recommendation towards conservatism. This reemphasizes results of some earlier comparisons of forecasting methods such as Newbold and Granger (1974) and the tourism competition of Athanasopoulos et al. (2011) where the performance of ARIMA was particularly strong. There is, however, evidence in the poor performance of Holt-Winters that the recommendation to damp trend is valuable. These conclusions generally hold when other error measures, such as MAPE and MdAPE, are used. But such a conclusion must be tempered by the performance of Theta in FP, a method with a constant trend.

The sub-population of size 224 time series where ARIMA has shown itself to be most accurate on the validation data is now analysed. Table 2 shows the performance of the different methods on the test data for various lead times. Overall ARIMA continues to perform very well, especially for the shorter horizons; however, the DDamped method performs about as well and these two methods are best or second best at all horizons.

Table 2 here

Three important points come from this analysis. First, methods designed for a particular set of data may, as here, continue to perform well (see also Fildes et al., 1998). Second, the conservative strategy recommended by AGG can lead to substantial losses in accuracy. Finally, these losses can be avoided through careful *ex ante* analysis of the time series. Effectively, our argument is that a priori rules such as "be conservative" or "simple is best", whilst providing useful pointers, are not as effective as selection rules based on the characteristics of both the time series and the forecasting methods in play.

Causal methods: on the use of equal coefficients in regression modelling

As a part of their "be conservative" rule, AGG suggest the use of damped estimated weights for causal models. In an extreme case of damping, all (*k*) explanatory variables included in a model will, after being standardized, be assigned the same coefficient:

$$y = a + b\sum_{i=1}^{k} x_i + e$$

The same argument is expanded by Graefe (this issue). Dampening reduces the "sensitivity" of the model, and is an approach to overcoming the over-fitting that may arise as a result of using a large number of predictors. It is recommended in cases with small sample

sizes compared to the number of explanatory variables, poor model fits, or when predictor variables are highly correlated (Graefe, this issue).

In this section a model using equal weighted predictors is assessed and its performance compared to that of a standard multiple regression modelling. Using the data from the recent textbook by Ord and Fildes (2013), the monthly price of unleaded gasoline will be estimated. Lagged values of the price of *Unleaded Gas*, price of *Crude Oil*, *New Housing Starts* and *Unemployment* are used as explanatory variables. The data are available from the Journal's web site. According to Ord and Fildes (2013, p. 288) lagged prices of the explanatory variables *Unleaded Gas* and *Crude Oil* demonstrate very high correlation as a result of causal collinearity. The length of the in-sample (in the previous section's terminology, the estimation and validation samples) upon which models are fitted has been varied (24, 60 or 108 months), so that any differences in performance with regards to the sample sizes can be revealed. The accuracy of the two alternatives over multiple horizons and origins, covering a hold-out period of 24 months (from January 2005 to December 2006), is then calculated. All the variables are lagged so as to correspond to the forecasting horizon. Effectively, a different regression model is estimated for each horizon. In the case of the equal weights models, the explanatory variables are first standardized and then summed. From this, a single variable is derived. In order to reflect the positive or negative impact of each variable, the summation is performed using signs as suggested by the multiple regression's coefficients although a priori analysis could equally well be used. For example, *Unemployment* is given a negative sign, in order to reflect its impact on reduced demand, while the lagged value of *Unleaded Gas* has a positive sign. Finally, the b coefficient, which will be the same for all variables, is estimated (note that if y is standardized as well, the b becomes $1/k$).

Table 3 here

Table 3 presents the empirical results on the performance of equal weights model as measured by the Relative MAE for the different sample sizes and across the various forecasting horizons, where the performance of the multiple regression model is used as benchmark. It is clear that a model implying equal coefficients for all variables results overall in worse performance compared to standard multiple regression (values for Relative MAE are greater than 1). This is more apparent for larger sample sizes and longer horizons. The only case where the accuracy of the simple model is comparable to the multiple regression one is when the ratio of sample size per predictor is less than 6, certainly lower than 100 as suggested by Dana and Dawes (2004). In this specific case, the simple model is marginally better than multiple regression for the shorter horizons (1 to 6), while worse for the longer horizons (7 to 12, -5.9%). Moreover, the simplified model employed in the current example relies on the ad-hoc identification of the variables as positive or negative, an a priori requirement that may complicate the modelling process in other cases where the directional effects are uncertain.

In the light of the above analysis, the "damp estimated weights" guidelines should not be automatically employed. Even if such a simplification is useful in cross-sectional social sciences research, where the coefficient of multiple correlation is small and data are noisy, it is suggested that the modeller experiments with the different options before making a final choice. The second issue is the question of the model's dynamics where for near non-stationary data, the situation where the current price is strongly affected by its previous period's price, the operation of the rule is unclear. Does one first establish the autoregressive nature of the series and then carry out the standardisation that is needed? Overall, it is easier and more informative to develop a standard time-series regression model using the principles laid down in Ord and Fildes (2013). And of course one such key principle is that the coefficients in the resulting model should make intuitive sense.

Other aspects of the recommendations on causal modelling are also conflicting: "use all important variables", a recommendation that is in conflict with the thrust towards simplicity. But what are important variables beyond those tentatively identified from prior research. The developments behind the use of score cards in credit appraisal are such that many plausible predictor variables are valuable, some are not and establishing which to include in a regression type model without the use of advanced statistical models would present a challenge (see the example in Chapter 10 of Ord and Fildes, 2013, pp. 311-326). The reference to the dated work of Einhorn (1972) is uninformative on the current state of knowledge as there has been so much development, both in theory but also in practice in dealing with complex cross-sectional data and the use of methods such as Automatic Interaction Detection and logistic regression. Baesens et al. (2003), for example, demonstrates differences in methods but also these methods always outperform a naïve benchmark.

The development of index models as AGG recommend is likely to help in modelling time series with many weakly relevant interdependent variables and, through the work of Stock and Watson (2002), has become a "hot topic" in econometrics where indices are developed using principal component analysis. Perhaps the simple indices recommended in AGG will do the job – but the research remains to be done to establish the circumstances where this approach will prove effective. In this example from Ord and Fildes (2013), where lags are important, the results are not reassuring. In summary, a number of the AGG recommendations on causal modelling are sensible and accord with the principles put forward by Allen and Fildes (2001). But as with the discussion on extrapolation, the generalisations behind the Golden Rule do not always hold – and the forecaster and researcher will have to establish their value in individual circumstances.

Concluding remarks

The proposed Golden Rule offers generalisations across a wide range of forecasting circumstances. The authors of this commentary broadly accept the recommendation "be conservative" and the guidelines in the checklist. AGG point out that for many recommendations there is limited evidence and generally, as it is shown here, the results will be context dependent. Even if the Golden Rule works for the "average" situation, the forecaster should take into account the specifics of the situation being modelled. This is especially true when dealing with count (intermittent demand) data (see for example, Syntetos and Boylan, 2005). Certainly the checklist offers some research opportunities; however, the evidence on many of the guidelines is slight and the suggested error reductions highly uncertain.

For researchers, the effectiveness of new methods is clearly contingent on the area of application. The priority is the development and evaluation of new methods, a research area that has proved productive – for example, the use of the Theta model in extrapolation (Assimakopoulos and Nikolopoulos, 2000). But as has been demonstrated in the two empirical examples, approaches that are recommended based on these guidelines are, in certain circumstances, quite inadequate; so a research strategy that attempts to develop new methods for specific circumstances should be followed.

Practitioners fall into quite different types, ranging from the expert economist to the novice company sales forecaster. Their jobs have different requirements. The adoption of effective techniques is organisation specific and will aim to utilize the specific problem features and resources faced by the organisation, rather than accepting a "conservative" approach that may well not provide a best fit. Here some of the guidelines are unhelpful. For example, "conceal the purpose of the forecast" does not translate into an operational practice in most organisations. But these are peripheral objections to the authors' core arguments and the proposed framework should bring tangible benefits to those who follow it: be

conservative, combine with simple methods, use evidence, evaluate against simple benchmarks and develop processes which are replicable; advice that, if followed, would lead to improved forecasting.

Unquestionably, forecasting methods have been advanced significantly over the years, offering substantial improvements in forecasting accuracy, while guidelines for their efficient application are widely available, not least in Armstrong (2001) and now AGG. However, the question that still remains is how improvements in forecasting methodology can be translated into measurable gains in practice and why the corresponding methods have not been more widely adopted. The barriers to be overcome were identified by Fildes and Hastings (1994) in a specific multinational conglomerate, and included limited software, data availability and lack of trained forecasters. Innovation too requires the support of key managers, in this case the users of the forecasts. Only software design is readily influenceable: but software companies have proved reluctant to respond to user needs or research (Asimakopoulos and Dix, 2013). Despite the advances AGG identify, these same barriers still seem to hold. Further research is needed towards understanding what users require from forecasters and the forecasting process practiced in organisations would extend the scope of Moon et al.'s (2003) empirical study and would prove beneficial in narrowing the gap between theory and practice.

References

Allen, P. G., & Fildes, R. (2001). Econometric forecasting. In J. S. Armstrong (Eds.), Principles of Forecasting (pp. 303-362). Norwell, MA: Kluwer.

Armstrong, J. S. (2006). Findings from evidence-based forecasting: Methods for reducing forecast error. International Journal of Forecasting, *22*(3), 583-598.

Armstrong, J. S., Green, K., & Graefe, A. (this issue). Golden Rule of Forecasting: Be Conservative. Journal of Business Research.

Asimakopoulos, S., & Dix, A. (2013). Forecasting support systems technologies-in-practice: A model of adoption and use for product forecasting. International Journal of Forecasting, *29*, 322-336.

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.

Baesens, B., Van Gestel T., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen., J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, *54*(6), 627-635.

Dana, J., & Dawes, R. M. (2004). The superiority of simple alternatives to regression for social science predictions. Journal of Educational and Behavioral Statistics, *29*(3), 317-331.

Davydenko, A., & Fildes R. 2013. Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. International Journal of Forecasting, *29*(3), 510-522.

Einhorn, H. J. (1972). Alchemy in the behavioral sciences. Public Opinion Quarterly, *36*(3), 367–378.

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

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. Interfaces, *37*, 570-576.

Fildes, R., Hibon, M., Makridakis, S., & Meade, N. (1998). Generalising about univariate forecasting methods: further empirical evidence. International Journal of Forecasting, *14*(3), 339-358.

Fildes, R., & Makridakis, S. (1995). The Impact of Empirical Accuracy Studies on Time-Series Analysis and Forecasting. International Statistical Review, *63*(3), 289-308.

Fildes R. & Petropoulos F. (this issue). An evaluation of simple versus complex selection rules for forecasting many time series. Journal of Business Research.

Graefe, A. (this issue). Improving forecasts using equally weighted predictors. Journal of Business Research.

Hyndman, R. J. & Khandakar, Y. (2008), Automatic Time Series Forecasting: The forecast Package for R, Journal of Statistical Software, *27*(3), 1-22.

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

Moon, M. A., Mentzer, J. T., & Smith, C. D. (2003). Conducting a sales forecasting audit. International Journal of Forecasting, *19*(1), 5-25.

Newbold P., Granger C. W. J. (1974). Experience with forecasting univariate time series and the combintion of forecasts. Journal of the Royal Statistical Society A, *137*, 131-164.

Ord, J. K., & Fildes, R. (2013). Principles of Business Forecasting. Mason, OH: South-Western, Cengage Learning.

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association, *97*(460), 1167-1179.

Syntetos, A. A. & Boylan, J. (2005). The accuracy of intermittent demand estimates. International Journal of Forecasting, *21*(2), 303-314.

Table 1: Accuracy of different methods measured by AvgRelMAE on the test data averaged across lead times 1 to 18, analysed for the difference segments considered.

Methods	Entire data set	Predictable	Unpredictable	Trended	Non-trended	Seasonal	Non-seasonal	Stable	Unstable
Naive	1.14	1.21	*1.00*	1.13	1.25	1.27	0.97	1.47	1.05
DNaive	1.00	1.00	*1.00*	1.00	1.00	1.00	*1.00*	1.00	1.00
SES	1.06	1.11	**0.96**	1.05	1.15	1.15	**0.93**	1.37	*0.97*
Holt	1.19	1.26	1.07	1.19	1.22	1.27	1.09	1.52	1.10
Holt-Winters	1.08	1.06	1.12	1.08	1.04	1.03	1.15	1.05	1.09
DDamped	**0.94**	**0.92**	*1.00*	**0.95**	**0.89**	**0.90**	1.01	**0.91**	**0.95**
ARIMA	*0.98*	*0.98*	*1.00*	*0.99*	*0.95*	*0.96*	1.02	*0.93*	1.00
No. Series	998	694	304	894	104	608	390	249	749

N.B. The most accurate method is shown in **bold**, second most accurate in *italics*.

Table 2. Accuracy of different methods measured by AvgRelMAE on the test data for the segment of time series where ARIMA is the best performer on the validation data (n=224).

Methods	Horizons						
	1	6	12	18	1 to 6	1 to 12	1 to 18
Naive	1.255	1.282	1.000	1.205	1.282	1.235	1.218
DNaive	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SES	1.146	1.170	1.053	1.119	1.168	1.130	1.123
Holt	1.170	1.281	1.223	1.306	1.243	1.241	1.262
Holt-Winters	0.952	1.029	1.109	1.166	0.987	1.019	1.062
DDamped	*0.906*	*0.941*	**0.980**	**0.943**	*0.921*	*0.929*	**0.933**

ARIMA	**0.897**	**0.937**	*0.998*	*0.951*	**0.918**	**0.928**	*0.939*

N.B. The most accurate method is shown in **bold**, second most accurate in *italics*.

Table 3. Average performance of equal weights model for the retail price of petrol data set, as measured by the Relative MAE over different lead times and for different sample sizes. The standard multiple regression model is used as a benchmark.

In-Sample size	Horizons						
	1	2	3	6	12	1 to 6	7 to 12
24	1.11	0.78	1.04	1.01	1.07	0.99	1.06
60	1.22	0.92	1.03	1.30	1.14	1.11	1.23
108	1.32	0.97	1.06	1.32	1.10	1.14	1.11