Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 3, pp. 569–578
DOI : 10.47974/JSMS-974

Statistical time-series forecast error and bias assessment using hold-out samples and a single benchmark (data-reduction) scoring metric

Dewi Rahardja
U. S. Department of Defense
Fort Meade, MD 20755
U.S.A.

Abstract

In Time-Series (TS) forecast modeling, we utilize a Hold-Out (HO) sample to assess a candidate model forecast error and bias, with the Mean Absolute Percentage Error (MAPE) as a single benchmark (data-reduction) for a scoring metric. The purpose of such HO sampling is to assess error rates and accuracy (unbiasedness) level whether the forecasted values of a candidate TS model is within a pre-specified (targeted) margin-of-error (MOE) rate of α-percent. For instance, for an accuracy expectation of 80% level, the MOE rate of α is tolerated to be, 20%. If the MAPE of a candidate TS model is less than the MOE rate of α, then the model is considered reasonably accurate enough. Otherwise, we select the next available candidate TS model with the smallest-possible MAPE. For an illustration, we apply such method to a TS sales-data example.

Subject Classification: *(2010) 37M10 – Time series analysis, 62M10 – Time series, auto-correlation, Regression, etc., 91B84 – Economic time series analysis.*

Keywords: *Time series, Hold-out sample, Mean absolute percent error (MAPE), Margin of error (MOE).*

1. Introduction

Nowadays in the business world, many organizations are striving to manage their resources and project their future demands (volume) of various items or workloads. For example, to forecast their constant (daily, weekly, monthly, yearly) demands (volume) of their daily newspapers

Disclaimer Statement: This research represents the authors own work and opinion. It does not reflect any policy nor represent the official position of the U.S. Department of Defense nor any other federal agency.

E-mail: rahardja@gmail.com

© Taru Publications

sales, the weekly liquor sales in a mini market, monthly storage rental, annual housing sales, etc.

Past research regarding statistical time-series (TS) forecasting methods have been discussed intensively [1–3]. Additionally, the modern statistical-computing executions of the above TS methods have also been implemented using many statistical software products, for instance, Rahardja (2021a; 2021b) papers [4–5], among many others.

As anyone can make any random guess of projections or forecasts, however, how accurate the resulting volume projections or forecasts is a different matter. Hence, such error rate and accuracy (unbiased) assessment of the forecast modeling is necessary for such TS model to be scientifically (statistically) defensible, valid, sound, and justifiable.

In statistics, generally it is often necessary not only to model data but also to assess the validity of that model. To do so, one needs to withhold the data into two groups: the main data and the validation samples. Specifically, in the statistical TS forecast scenario, a Hold-Out (HO) sample is a sample from a data set that is withheld and not used in the TS-model fitting process [6]. It is a subset of a statistical data set that is set aside during the Model Development process in order to help evaluate the estimated model [7–8]. After the model is fit to the main data, it is then applied to the HO sample. Then, a single (data-reduction) scoring metric is utilized to benchmark model performance. Such benchmarking gives an unbiased assessment of how well the model might fit if applied to new data.

Therefore, in this paper, for the purpose of teaching and training a larger audience of practitioners or forecasters (in various fields of study, such as, business, management, finance, economics, etc.), we present the HO-sample [6–8] technique in TS-forecasting context along with the Mean Absolute Percentage Errors (MAPE) [9–10], as a single (data-reduction) scoring metrics, to benchmark and assess the error rate and accuracy (unbiasedness) of a candidate TS-forecasting model.

In this paper, the sections are organized as follow. In Section 2, we describe the HO-sample concept [6–8]. In Section 3, we present a (data-reduction) scoring metric, such as MAPE [9–10] to benchmark the model performance (error and accuracy). Then in Section 4, we illustrate the application to a sales-data series obtained from Exceljet example [11] using Microsoft Excel (2019) software [12]. Finally in Section 5, we conclude the key points and recommendation.

2. Hold-Out (HO) Samples

A Holdout Sample is a subset of a statistical data set that is set aside during the Model Development process in order to help evaluate the estimated model [6–8].

Next, as prescribed in Rahardja (2021b) paper [5] on using the Excel (version 2016 or above) to forecast ESM-family TS model, here we display arbitrary 204-month series (from 17 years × 12 month) of sales volume (say, from October 2020 to September 2017), for the purpose of illustrating and explaining a sub-sample of 60-month to be withheld as a HO-sample [6–8], as shown in Figure 1. As forecasting rule-of-thumb (ROT), some practitioners suggested to withhold about 30% samples [13], as a HO-sample [6–8] to assess forecast error and bias (inaccuracy). Therefore, for the sake of illustration, here we take about 30% sample for our HO-sample [6–8] but rounding the percentage to the nearest integer-multiplication of the seasonality number. In this case, using the Excel built-in function (FORECAST.ETS.SEASONALITY) as described in the references [5, 14], the seasonality period was calculated to be 12. Hence, we withheld 60-month series (5 year × 12 month) out of total of 204-month series (29.4%) which satisfied the ROT recommendation of about 30% [13].

In Figure 1, the blue graph represents the actual observed data series to forecast while the red graph represents the forecasted values. The overlapping graph (red and blue) are the HO samples [6–8] withheld (pretending we do not have such 60-month or 30% TS-data from October

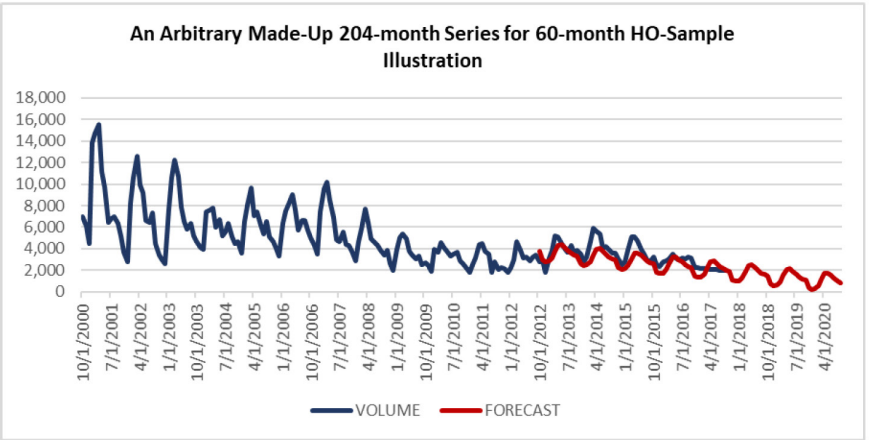


Figure 1

A Made-Up Series for (30%) HO-Sample Illustration.

2012 to September 2017; hence, we have some overlapping red and blue graphs). Then we forecast the future projections from October 2012 onward. Such (overlapping graphs) differences are the forecast errors and bias (inaccuracy) which can be benchmarked by a single (data-reduction) scoring metric, such as MAPE [9–10]. We will discuss such scoring metric MAPE [9–10] in the next section.

3. MAPE – A Single Benchmark

The choice of a single (data-reduction) scoring metric utilized to benchmark model performance is key to any validation exercise [15]. Note that measuring forecast accuracy (or error) is not an easy task as there is no one-size-fits-all indicator. Only experimentation will show which Key Performance Indicator (KPI) is best for their scenario. Note that such KPI are not unique. Note also that there are two categories of KPI: Distance-Based and Likelihood-Based KPIs. The in-depth discussion of these KPIs is beyond the scope of this article. Several examples of Distance-Based KPIs: Mean Absolute Percent Error (MAPE) [9–10], Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Deviation (MAD), etc. Next, examples of Likelihood-Based KPIs are: Akaike Information Criteria (AIC) [16–17], Schwarz Information Criteria (SIC) [18–21], Hannan–Quinn information criterion (HQC) [22–23], etc. Hence, without loss of generalities (WLOG), here we utilize the MAPE [9–10] as one of the most-commonly used benchmarks for a validation exercise (i.e., to measure forecast error and accuracy), due to its simplicity (simple-and-easy to understand) and robustness (i.e., computable quickly-and-easily with the standard Microsoft Excel without even purchasing a more complex-yet-expensive software).

WLOG, MAPE [9–10] as a simple-yet-robust KPI, measures forecast accuracy as a percentage and can be calculated as the average absolute percentage error for each time period forecast minus actual (observed) values divided by actual (observed) values, as follows:

$$\text{MAPE} = 100\% \times (1/n) \sum_t \mid F_t - A_t \mid / A_t$$

where, A_t is the t-th actual (observed) point, F_t is t-th forecasted point and n is the number of fitted points under study. Their difference is divided by the actual (observed) value A_t. The absolute value in this ratio is summed for every forecasted point in time and divided by the number of fitted points n.

Given a pre-specified (targeted) margin-of-error (MOE) rate of α-percent, say typically aiming 80% theoretical accuracy level, then the MOE rate of α can theoretically be tolerated to be 20%. Next, WLOG, we compare the resulting MAPE of the HO samples, from a candidate/selected TS-forecast model. If the MAPE of the HO samples (from a candidate TS-forecast model) is less than (within) the pre-specified MOE rate of α, then such candidate TS-forecast model is considered reasonably accurate (unbiased) enough, as a MAPE "Local Minimum" solution. Otherwise (if exists and if affordable), we can select the next available candidate TS-forecast model with the smallest MAPE possible (as model-search exhaustive as possible), as the MAPE "Global Minimum" solution.

Note that the above discussion regarding forecast error and bias assessment is just purely from a theoretical perspective. However, from a real-world practical perspective, Gilliland (2010) wrote a handbook collection of numerous practitioners discussions regarding forecasting frequently-asked-questions (FAQ) [24]. In such Gilliland (2010) handbook [24], for example, there is a question regarding how accurate should our forecasts be. The answer is as follows: Unfortunately, there is no way to know in advance how accurate our forecasts should be because we do not know the forecastability of the future demand. Therefore, it is not very realistic to set arbitrary forecasting performance goals, such as "Next year MAPE must be less than 20%." If demand is not forecastable to this level of accuracy, it will be impossible to achieve the goal. This can lead to a demoralized forecasting staff that just gives up trying — or makes up numbers. Reasonable goals for your forecasting performance are "to do no worse than a naïve forecast" and to continuously improve the process by making it more efficient. Hence, users need to realize and then balance the existence of both perspectives (theoretical and practical).

4. Example

Among numerous TS examples, Rahardja (2021b) paper [5] utilized the sales-data from Exceljet (2021) example [11]. Note that for a demo purpose, we intentionally chose such a (short) sales-data of an 8-period time-series in order to fit-in all the total package (i.e., the data series, formula, graph, and MAPE calculation) into just one-box display, for ease-of-visualization and illustration.

In the sales-data example on Figure 2, we can see that the gaps between Sales-Data points and the Forecasted Values are extremely close such that we barely see the gaps. Such visual/graphical plot also confirmed

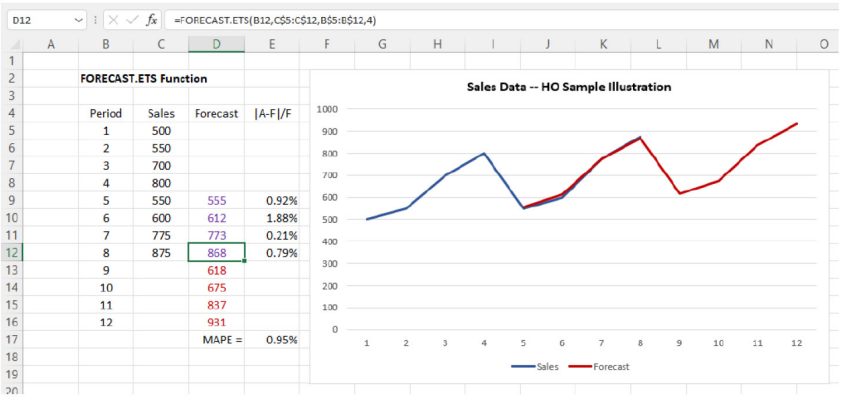


Figure 2

The 4-period HO-Sample of An 8-period (Quarterly-Seasonal) Sales-Data.

with the (overall) numerical calculation of MAPE = 0.96%, (with 4-period HO sample shown on cells C9:C12 of Figure 2) which is an overall (data-reduction) single scoring metric of (error rate) less than 1% differences (average absolute percentage error). Therefore, such TS-model is indeed a very excellently-fitted model. Since there exist both trend (trending-up) and seasonality (quarterly series or 4-period) [5], then according to the prescription from the Rahardja (2020; 2021a; 2021b) papers [3–5], such TS-model is recognized as the "Additive Winters ESM" TS-forecast model.

Next, once the model's name is recognized, for a quality-control (QC) check, one can then replicate such recognized (Additive Winters ESM) TS-forecast model with another reliable/reputable software, say for example, using SAS-ETS 9.4 software [25–26] by executing the "PROC ESM" procedure [27]; or any other reliable/reputable statistical software.

For example, the resulting QC check (i.e., the forecasted-values) from the SAS-ETS 9.4 [25–26] ESM procedure [27] had an overall MAPE = 3.06% (with only 2-period HO sample produced) which is considered as very close to what the above Excel (2019) built-in function produced (MAPE = 0.96% with 4-period HO sample). Therefore, even though for a very-short series of 8-period TS dataset with quarter (4-period) seasonality, both Excel (2019) and SAS-ETS 9.4 were able to produce very-comparably close (overall) MAPE scores (0.96% versus 3.06%). Note that the small percentage difference might came from slightly different estimator optimization computing algorithm used or any other possible micro-step details. For example, to have lower MAPE rate, SAS computing algorithm requires much longer-period series than just 8-period main data series, in order to

produce more than 2-period HO sample. Another example, the default estimator described in SAS-ETS 9.4 under the publicly available SAS online documentation [25–27] is the best/gold-standard estimator: MLE or Maximum Likelihood Estimator due to its 5-criteria of "Best Estimator" definition (i.e., asymptotically consistent, normal, unbiased, efficient, and has invariant property), while in Excel (2019) the online documentation did not mention which estimator optimization algorithm used. For example, whether it is CLS (Conditional Least Square estimator), OLS (Ordinary Least Square), MLE, etc. Also, examining both resulting numerical and graphical forecasted values, they both are within very sensible range of forecasts. Therefore, such resulting QC-check (forecasted-values) produced by both software (MAPE = 0.96% and 3.06%) are very similar (and both MAPE are a much less than the pre-specified MOE, $\alpha = 20\%$) and hence, the 4-period forecasts are considered as sound and justifiable (i.e., valid and reasonably accurate) projections.

5. Conclusion and Recommendation

In this article, we have demonstrated the utilization of HO-Sample [6–8, 13] and a single (data-reduction) score metric such as MAPE [9–10] as benchmarking to assess forecast error and accuracy level. We conclude that such utilization of HO-Sample [6–8, 13] and benchmarking via MAPE [9–10] scoring metric will be very helpful for many forecasters or practitioners (in various fields of study) to assess the forecast error and bias (accuracy level). Using a simple and short (only 8-period) time-series sales-data example with a quarterly (i.e., 4-period) seasonality, we have demonstrated the application of HO Sampling to assess the validity of a candidate TS-forecast model (Additive Winters ESM) via MAPE [9–10], a single (data-reduction) score metric as a benchmark to measure forecast error and bias.

Therefore, in general, given a pre-specified MOE of α-level rate, we recommend to apply HO-Sample [6–8, 13] technique and MAPE [9–10] as a simple, easy to understand, quick to produce, most-commonly used, and reasonably-good benchmark to assess forecast error and accuracy (bias) level. If MAPE [9–10] is less than the pre-specified MOE of α-level rate, then such candidate TS-model is considered to be a good-fit model to the actual (observed) TS data; and the resulting reasonably-accurate forecasts can be helpful/useful for the decision-making.

Disclaimer Statement:

This research represents the authors own work and opinion. It does not reflect any policy nor represent the official position of the U.S. Department of Defense nor any other federal agency.

References

[1] Sypsas, P.T. (1989). Identifying patterns in multiple time series data. *Journal of Information and Optimization Sciences*, 10(3): 471–494.

[2] De Gooijer, J.G., Hyndman, R.J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, 22(3): 443–473.

[3] Rahardja, D. (2020). Statistical methodological review for time-series data. *Journal of Statistics and Management Systems*, 23(8): 1445–1461. DOI: https://doi.org/10.1080/09720510.2020.1727618.

[4] Rahardja, D. (2021a). Statistical Computing for Time-Series Forecasts via SAS Forecast Studio. *International Journal of Applied Science and Mathematics*, 8(4): 38–45.

[5] Rahardja, D. (2021b). Statistical Time-Series Forecast via Microsoft Excel (FORECAST.ETS) Built-In Function. Quest Journals Publisher. *Journal of Research in Applied Mathematics*, 7(11): 69–73.

[6] Holdout Sample, The Institute for Statistics Education, LLC (2021). © Copyright 2021 - Statistics.com, LLC. All Rights Reserved. Retrieve from https://www.statistics.com/glossary/hold-out-sample/.

[7] Holdout Sample, Open Risk Manual (2016). Retrieve from https://www.openriskmanual.org/wiki/Holdout_Sample.

[8] Holdout Sample, Oxford Reference (2021). Oxford University Press. Copyright © 2021. All rights reserved. Retrieve from https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095941196.

[9] Mean Absolute Percentage Error, MAPE, Wikipedia (2021). Last edited on 21 November 2021, at 23:28 (UTC). Retrieve from https://en.wikipedia.org/wiki/Mean_absolute_percentage_error.

[10] Mean Absolute Percentage Error, MAPE (2000). In: Swamidass P.M. (eds) Encyclopedia of Production and Manufacturing Management. Springer, Boston, MA. https://doi.org/10.1007/1-4020-0612-8_580.

[11] Exceljet (2021a). Excel FORECAST.ETS Function. © 2012-2021 Exceljet. Retrieve from https://exceljet.net/excel-functions/excel-forecast.ets-function.

[12] Excel 2019 Support (2019). What's new in Excel 2019 for Windows. © Microsoft 2021. https://support.microsoft.com/en-us/office/what-s-new-in-excel-2019-for-windows-5a201203-1155-4055-82a5-82bf0994631f.

[13] Holdout Sample, KDD Analytics (2018). Practical Time Series Forecasting – Know When to Hold 'em. Posted 8 January 2018. © 2021 KDD Analytics. Retrieve from https://www.kddanalytics.com/practical-time-series-forecasting-holdout-sample/.

[14] Exceljet (2021b). Excel FORECAST.ETS.SEASONALITY Function. © 2012-2021 Exceljet. Retrieve from https://exceljet.net/excel-functions/excel-forecast.ets.seasonality-function.

[15] Polhamus, Aaron (2015). Validation with holdout sampling — why you (or your data provider) better be doing it. Retrieve from https://aaronpolhamus.medium.com/validation-with-holdout-sampling-why-you-or-your-data-provider-better-be-doing-it-bdc3a838c5f5.

[16] Akaike Information Criteria (AIC), Wikipedia (2020). Retrieve from https://en.wikipedia.org/wiki/Akaike_information_criterion.

[17] Akaike, H. (21 December 1981), "This Week's Citation Classic" (PDF), *Current Contents Engineering, Technology, and Applied Sciences*, 12(51): 42 [Hirotogu Akaike comments on how he arrived at AIC].

[18] Schwarz Information Criteria (SIC), Wikipedia (2020). Retrieve from https://en.wikipedia.org/wiki/Bayesian_information_criterion.

[19] Bhat, H. S., and Kumar, N (2010). On the derivation of the Bayesian Information Criterion. Archived from the original on 28 March 2012. https://web.archive.org/web/20120328065032/http://nscs00.ucmerced.edu/~nkumar4/BhatKumarBIC.pdf.

[20] Findley, D. F. (1991). "Counterexamples to parsimony and BIC". *Annals of the Institute of Statistical Mathematics*, 43(3): 505–514. DOI:10.1007/BF00053369.

[21] Kass, R. E.; Wasserman, L. (1995). "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion". *Journal of the American Statistical Associatio*n, 90(431): 928–934. DOI:10.2307/2291327. JSTOR 2291327.

[22] Hannan, E. J., and B. G. Quinn (1979). The Determination of the order of an autoregression. *Journal of the Royal Statistical Society*, Series B, 41: 190–195.

[23] Hannan-Quinn Information Criteria (HQC), Wikipedia (2020). Retrieve from https://en.wikipedia.org/wiki/Hannan%E2%80%93Quinn_information_criterion.

[24] Gilliland, Michael (2010). Forecasting FAQ. The Business Forecasting Deal: Exposing Myths, Eliminating Bad Practices, Providing Practical Solutions. Copyright © 2010 by SAS Institute, Inc. Retrieve from https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119199885.app1. Also available: https://www.coursehero.com/file/115043927/9781119199885app1pdf/.

[25] SAS Institute Inc. (1995). SAS manual SAS/ETS Software: Time series forecasting system, Version 6, First Edition: 225–235. Cary, NC: SAS Institute Inc.

[26] SAS Institute Inc. (2013). Base SAS® 9.4 Procedures Guide: Statistical procedures, Second edition. Cary, NC: SAS Institute Inc.

[27] SAS® Help Center (2017). The ESM Procedure. Copyright © SAS Institute Inc. All rights reserved. SAS® 9.4 and SAS® Viya® 3.3 Programming Documentation. SAS Documentation website: https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.3/etsug/etsug_esm_syntax02.htm.

Received March, 2022