|  |
| --- |
| Two years ago I have written a post “[Naughty APEs and the quest for the holy grail](https://forecasting.svetunkov.ru/en/2017/07/29/naughty-apes-and-the-quest-for-the-holy-grail/)“, where I have discussed why percentage-based error measures (such as MPE, MAPE, sMAPE) are not good for the task of forecasting performance evaluation. However, it seems to me that I did not explain the topic to the full extent – the time has shown that there are some other issues that need to be discussed in detail, so I have decided to write another post on the topic, possibly repeating myself a little bit. This time we won’t have imaginary forecasters, we will be more serious.  **Introduction**  We start from a fact, well-known in statistics. **MSE is minimised by mean** value, while **MAE is minimised by the median**. There is a lot of nice [examples](http://dx.doi.org/10.1080/00031305.1990.10475690), [papers](https://doi.org/10.1016/j.ijforecast.2015.12.004) and [post](https://math.stackexchange.com/questions/2554243/understanding-the-mean-minimizes-the-mean-squared-error/2554276), explaining, why this happens and how, so we don’t need to waste time on that. But there are two elements related to this that need to be discussed.  First, some people think that this property is applicable only to the estimation of models. For some reason, it is implied that forecasts evaluation is a completely different activity, unrelated to the estimation. Something like: as soon as you estimate a model, you are done with “MSE is minimised by mean” thingy, the evaluation is not related to this. However, when we select the best performing model based on some error measure, we inevitably impose properties of that error on the forecasts. So, if a method performs better than the other in terms of MAE, it means that it produces forecasts closer to the median of the data than the others do.  For example, zero forecast will always be one of the most accurate forecasts for intermittent demand in terms of MAE, especially when number of zeroes in the data is greater than 50%. The reason for this is obvious: if you have so many zeroes, then saying that we won’t sell anything in the next foreseeable future is a safe strategy. And this strategy works, because MAE is minimised by median. The usefulness of such forecast is a completely different topic, but a thing to carry out from this, is that in general, **MAE-based error measures should not be used on intermittent demand**.  Are you not convinced? Click here…  Just to make this point clearer, let’s consider the following example in R. We generate data from a mixture of normal and Bernoulli distributions (with probability \(p=0.4\), which means that we should have around 60% of zeroes in the data):  x <- rnorm(150,30,10) \* rbinom(150, 1, 0.4)  The series should look something like this:  plot.ts(x)    An example of artificial intermittent data  We use the first 100 observations for methods estimation and the last 50 for the evaluation of the forecasts. We use two forecasting methods: simple average of the in-sample part of the series and zero forecast (which accidentally corresponds to the median of the data). They look the following way:  plot.ts(x)  abline(h=mean(x[1:100]),col="blue", lwd=2)  abline(h=0,col="purple", lwd=2)  abline(v=100, col="red", lwd=2)    An example of artificial intermittent data and the forecasts. The blue line is the simple average, the purple line is the zero forecast. The red line splits the data into in-sample and the holdout  It is obvious that the average forecast is more reasonable than the zero forecast in this case – at least we can make some decisions based on that. In addition, it seems that it goes “through the data” in the holdout sample, which is what we usually want from the point forecasts. But what about the error measures?  errorMeasures <- matrix(c(mean(abs(x[101:150] - mean(x[1:100]))),  mean(abs(x[101:150] - 0)),  mean((x[101:150] - mean(x[1:100]))^2),  mean((x[101:150] - 0)^2)),  2,2,dimnames=list(c("Average","Zero"),c("MAE","MSE")))  errorMeasures  MAE MSE  Average 15.4360 264.9922  Zero 12.3995 418.4934  We can see that MAE recommends zero forecast as the more appropriate here (it has the lower error of 12.3995 in contrast with 15.4360), while MSE prefers the Average (264.9922 vs 418.4934). This is a simple illustration of the point about the mean and median.  Second, some researchers think that if a model is optimised, for example, using MSE, then it should always be evaluated using the MSE-based error measure. This is not completely correct. Yes, the model will probably perform better if the loss function is aligned with the error measure used for the evaluation. However, this does not mean that we cannot use MAE-based error measures, if the loss is not MAE-based. These are still slightly different tasks, and the selection of the error measures should be motivated by a specific problem (for which the forecast is needed) not by a loss function used. For example, in case of inventory management neither MAE nor MSE might be useful for the evaluation. One would probably need to see how models perform in terms of safety stock allocation, and this is a completely different problem, which does not necessarily align well with either MAE or MSE. As a final note, in some cases we are interested in estimating models via the likelihood maximisation, and selecting an aligned error measure in those cases might be quite challenging.  So, as a minor conclusion, MSE-based measures should be used, when we are interested in identifying the method that outperforms the others in terms of mean values, while the MAE-based should be preferred for the medians, irrespective to how we estimate our models.  As one can already see, there might be some other losses, focusing, for example, on specific quantiles (such as [pinball loss](https://datascience.stackexchange.com/questions/33282/can-pinball-loss-be-used-to-construct-a-prediction-interval)). But the other question is, what statistics minimise MAPE and sMAPE. The short answer is: “we don’t know”. However, [Stephan Kolassa and Martin Roland (2011)](https://foresight.forecasters.org/product/foresight-issue-23/) showed on a simple example that in case of strictly positive distribution the MAPE prefers the biased forecasts, and [Stephan Kolassa (2016)](https://doi.org/10.1016/j.ijforecast.2015.12.004) noted that in case of log normal distribution, the MAPE is minimised by the mode. So at least we have an idea of what to expect from MAPE. However, the sMAPE is a complete mystery in this sense. We don’t know what it does, and this is yet another reason not to use it in forecasts evaluation at all (see the other reasons in [the previous post](https://forecasting.svetunkov.ru/en/2017/07/29/naughty-apes-and-the-quest-for-the-holy-grail/)).  We are already familiar with some error measures from [the previous post](https://forecasting.svetunkov.ru/en/2017/07/29/naughty-apes-and-the-quest-for-the-holy-grail/), so I will not rewrite all the formulae here. And we already know that the error measures can be in the original units (MAE, MSE, ME), percentage (MAPE, MPE), scaled or relative. Skipping the first two, we can discuss the latter two in more detail.  **Scaled measures**  Scaled measures can be quite informative and useful, when comparing different forecasting methods. For example, sMAE and sMSE (from [Petropoulos & Kourentzes, 2015](https://doi.org/10.1057/jors.2014.62)): \begin{equation} \label{eq:sMAE} \text{sMAE} = \frac{\text{MAE}}{\bar{y}}, \end{equation} \begin{equation} \label{eq:sMSE} \text{sMSE} = \frac{\text{MSE}}{\bar{y}^2}, \end{equation} where \(\bar{y}\) is the in-sample mean of the data. These measures have a simple interpretation, close to the one of MAPE: they show the mean percentage errors, relative to the mean of the series (not to each specific observation in the holdout). They don’t have problems that APEs have, but they might not be applicable in cases of non-stationary data, when mean changes over time. To make a point, they might be okay for the series on the graph on the left below, where the level of series does not change substantially, but their value might change dramatically, when the new data is added on the graph on the right, with trend time series.    Two time series examples  MASE by [Rob Hyndman and Anne Koehler (2006)](https://doi.org/10.1016/j.ijforecast.2006.03.001) does not have this issue, because it is scaled using the mean absolute in-sample first differences of the data: \begin{equation} \label{eq:MASE} \text{MASE} = \frac{\text{MAE}}{\frac{1}{T-1}\sum\_{t=2}^{T}|y\_t -y\_{t-1}|} . \end{equation} The motivation here is statistically solid: while the mean can change over time, the first difference of the data are usually much more stable. So the denominator of the formula becomes more or less fixed, which solves the problem, mentioned above.  Unfortunately, MASE has a different issue – it is uninterpretable. If MASE is equal to 1.3, this does not really mean anything. Yes, the denominator can be interpreted as a mean absolute error of in-sample one-step-ahead forecasts of Naive, but this does not help with the overall interpretation. This measure can be used for research purposes, but I would not expect practitioners to understand and use it.  And let’s not forget about the dictum “MAE is minimised by median”, which implies that, in general, neither MASE nor sMAE should be used on intermittent demand.  **Relative measures**  Finally, we have relative measures. For example, we can have relative MAE or RMSE. Note that [Davydenko & Fildes, 2013](https://doi.org/10.1016/j.ijforecast.2012.09.002) called them “RelMAE” and “RelRMSE”, while the aggregated versions were “AvgRelMAE” and “AvgRelRMSE”. I personally find these names tedious, so I prefer to call them “rMAE”, “rRMSE” and “ArMAE” and “ArRMSE” respectively. They are calculated the following way: \begin{equation} \label{eq:rMAE} \text{rMAE} = \frac{\text{MAE}\_a}{\text{MAE}\_b}, \end{equation} \begin{equation} \label{eq:rRMSE} \text{rRMSE} = \frac{\text{RMSE}\_a}{\text{RMSE}\_b}, \end{equation} where the numerator contains the error measure of the method of interest, and the denominator contains the error measure of a benchmark method (for example, Naive method in case of continuous demand or forecast from an average for the intermittent one). Given that both are aligned and are evaluated over the same part of the sample, we don’t need to bother about the changing mean in time series, which makes both of them easily interpretable. If the measure is greater than one, then our method performed worse than the benchmark; if it is less than one, the method is doing better. Furthermore, as I have mentioned in the previous post, both rMAE and rRMSE align very well with the idea of forecast value, [developed](https://www.wiley.com/en-gb/The+Business+Forecasting+Deal:+Exposing+Myths,+Eliminating+Bad+Practices,+Providing+Practical+Solutions-p-9780470574430) by [Mike Gilliland from SAS](https://blogs.sas.com/content/author/mikegilliland/), which can be calculated as: \begin{equation} \label{eq:FV} \text{FV} = 1-\text{relative measure} \cdot 100\%. \end{equation} So, for example, rMAE = 0.96 means that our method is doing 4% better than the benchmark in terms of MAE, so that we are adding the value to the forecast that we produce.  As mentioned in [the previous post](https://forecasting.svetunkov.ru/en/2017/07/29/naughty-apes-and-the-quest-for-the-holy-grail/), if you want to aggregate the relative error measures, it makes sense to use geometric means instead of arithmetic ones. This is because the distribution of relative measures is typically asymmetric, and the arithmetic mean would be too much influenced by the outliers (cases, when the models performed very poorly). Geometric one, on the other hand, is much more robust, in some cases aligning with the median value of a distribution.  The main limitation of relative measures is that they cannot be properly aggregated, when the error (either MAE or MSE) is equal to zero either in the numerator or in the denominator, because the geometric mean becomes either equal to zero or infinity. This does not stop us from analysing the distribution of the errors, but might cause some inconveniences. To be honest, we don’t face these situations very often in real world, because this implies that we have produced perfect forecast for the whole holdout sample, several steps ahead, which can only happen, if there is no forecasting problem at all. An example would be, when we know for sure that we will sell 500 bottles of beer per day for the next week, because someone pre-ordered them from us. The other example would be an intermittent demand series with zeroes in the holdout, where the Naive would produce zero forecast as well. But I would argue that Naive is not a good benchmark in this case, it makes sense to switch to something like simple mean of the series. I struggle to come up with other meaningful examples from the real world, where the mean error (either MAE or MSE) would be equal to zero for the whole holdout. Having said that, if you notice that either rMAE or rRMSE becomes equal to zero or infinite for some time series, it makes sense to investigate, why that happened, and probably remove those series from the analysis.  It might sound as I am advertising relative measures. To some extent I do, because I don’t think that there are better options for the comparison of the point forecasts than these at the moment. I would be glad to recommend something else as soon as we have a better option. They are not ideal, but they do a pretty good job for the forecasts evaluation.  So, summarising, I would recommend using relative error measures, keeping in mind that MSE is minimised by mean and MAE is minimised by median. And in order to decide, what to use between the two, you should probably ask yourself: what do we really need to measure? In some cases it might appear that none of the above is needed. Maybe you should look at the prediction intervals instead of point forecasts… This is something that we will discuss next time.  **Examples in R**  To make this post a bit closer to the application, we will consider a simple example with  smooth  package v2.5.3 and several series from the M3 dataset.  Load the packages:  library(smooth)  library(Mcomp)  Take a subset of monthly demographic time series (it’s just 111 time series, which should suffice for our experiment):  M3Subset <- subset(M3, 12, "demographic")  Prepare the array for the two error measures: rMAE and rRMSE (these are calculated based on the  measures()  function from  greybox  package). We will be using three options for this example: [CES](https://doi.org/10.13140/RG.2.1.3757.2562), [ETS](https://forecasting.svetunkov.ru/en/2017/01/24/smooth-package-for-r-es-function-part-iv-model-selection-and-combination-of-forecasts/) with automatic model selection between the 30 models and ETS with the automatic selection, skipping multiplicative trends:  errorMeasures <- array(NA, c(length(M3Subset),2,3),  dimnames=list(NULL, c("rMAE","rRMSE"),  c("CES","ETS(Z,Z,Z)","ETS(Z,X,Z)")))  Do the loop, applying the models to the data and extracting the error measures from the accuracy variable. By default, the benchmark in rMAE and rRMSE is the Naive method.  for(i in 1:length(M3Subset)){  errorMeasures[i,,1] <- auto.ces(M3Subset[[i]])$accuracy[c("rMAE","rRMSE")]  errorMeasures[i,,2] <- es(M3Subset[[i]])$accuracy[c("rMAE","rRMSE")]  errorMeasures[i,,3] <- es(M3Subset[[i]],"ZXZ")$accuracy[c("rMAE","rRMSE")]  cat(i); cat(", ")  }  Now we can analyse the results. We start with the ArMAE and ArRMSE:  exp(apply(log(errorMeasures),c(2,3),mean))  CES ETS(Z,Z,Z) ETS(Z,X,Z)  rMAE 0.6339194 0.8798265 0.8540869  rRMSE 0.6430326 0.8843838 0.8584140  As we see, all models did better than Naive: ETS is approximately 12 – 16% better, while CES is more than 35% better. Also, CES outperformed both ETS options in terms of rMAE and rRMSE. The difference is quite substantial, but in order to see this clearer, we can reformulate our error measures, dividing rMAE of each option by (for example) rMAE of ETS(Z,Z,Z):  errorMeasuresZZZ <- errorMeasures  for(i in 1:3){  errorMeasuresZZZ[,,i] <- errorMeasuresZZZ[,,i] / errorMeasures[,,"ETS(Z,Z,Z)"]  }  exp(apply(log(errorMeasuresZZZ),c(2,3),mean))  CES ETS(Z,Z,Z) ETS(Z,X,Z)  rMAE 0.7205050 1 0.9707448  rRMSE 0.7270968 1 0.9706352  With these measures, we can say that CES is approximately 28% more accurate than ETS(Z,Z,Z) both in terms of MAE and RMSE. Also, the exclusion of the multiplicative trend in ETS leads to the improvements in the accuracy of around 3% for both MAE and RMSE.  We can also analyse the distributions of the error measures, which sometimes can give an additional information about the performance of the models. The simplest thing to do is to produce boxplots:  boxplot(errorMeasures[,1,])  abline(h=1, col="grey", lwd=2)  points(exp(apply(log(errorMeasures[,1,]),2,mean)),col="red",pch=16)    Boxplot of rMAE for a subset of time series from the M3 |

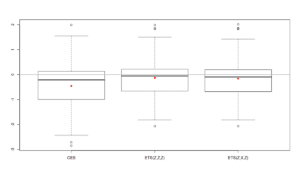
Given that the error measures have asymmetric distribution, it is difficult to analyse the results. But what we can spot is that the boxplot of CES is located lower than the boxplots of the other two models. This indicates that the model is performing consistently better than the others. The grey horizontal line on the plot is the value for the benchmark, which is Naive in our case. Notice that in some cases all the models that we have applied to the data do not outperform Naive (there are values above the line), but on average (in terms of geometric means, red dots) they do better.

Producing boxplot in log scale might sometimes simplify the analysis:

boxplot(log(errorMeasures[,1,]))

abline(h=0, col="grey", lwd=2)

points(apply(log(errorMeasures[,1,]),2,mean),col="red",pch=16)

[](https://i2.wp.com/forecasting.svetunkov.ru/wp-content/uploads/2019/08/M3rMAEBoxplotLog.png?ssl=1)

Boxplot of rMAE in logarithms for a subset of time series from the M3

The grey horizontal line on this plot still correspond to Naive, which in log-scale is equal to zero (log(1)=0). In our case this plot does not give any additional information, but in some cases it might be easier to work in logarithms rather than in the original scale due to the potential magnitude of positive errors. The only thing that we can note is that CES was more accurate than ETS for the first, second and third quartiles, but it seems that there were some cases, where it was less accurate than both ETS and Naive (the upper whisker and the outliers).

There are other things that we could do in order to analyse the distribution of error measures more thoroughly.