Tuning Forecasting Algorithms for Black Swans

S. D. Prestwich *

* *Department of Computer Science, University College Cork, Ireland (e-mail: s.prestwich@cs.ucc.ie).*

Abstract:
Forecasting algorithms based on exponential smoothing have smoothing factors, and it is often recommended that these be tuned to minimise an error measure on observed data. We show that forecasting algorithms such as simple exponential smoothing and Croston's method cannot always be optimally tuned to time series using any of several error measures. We propose a data augmentation approach: adding hypothetical non-stationary time series (which we call "black swans" as they represent unseen pathological cases) to the training data, and minimising a weighted error. The choice of black swans is a form of judgemental forecasting that requires experts to explicitly state their assumptions on unseen data.

Keywords: Forecasting, optimisation, time series, black swan, non-stationarity

1. INTRODUCTION

Single exponential smoothing (SES) and its variants is one of the simplest yet one of the most popular forecasting algorithms, and is provided in many forecasting software packages:

$$s_t = \alpha x_t + (1-\alpha)s_{t-1}$$

where x_t is the demand (other other quantity) at time t, α is the smoothing factor and s_t is the demand forecast. SES has been extended to well-known and more sophisticated forecasting methods that can handle trends, seasonality and intermittency. We mention in particular double exponential smoothing (DES) for data with trends (we use Holt's variant):

$$\begin{aligned} s_t &= \alpha x_t + (1-\alpha)(s_{t-1} + b_{t-1}) \\ b_t &= \beta(s_t - s_{t-1}) + (1-\beta)b_{t-1} \end{aligned}$$

with data smoothing factor α, trend smoothing factor β, and m-step-ahead forecast $s_t + mb_t$. We also consider Croston's method (CR) for intermittent data:

$$\begin{aligned} s_t &= \alpha x_t + (1-\alpha)s_{t-1} \\ y_t &= \beta\tau_t + (1-\beta)y_{t-1} \end{aligned}$$

where τ is an interval between nonzero x_t, y a smoothed version (not updated when 0 occurs), s a smoothed nonzero x, and the forecast is $\frac{s_t}{y_t}$. These methods have parameters (smoothing factors, initial forecasts, initial trends, and initial inter-demand intervals) which must be tuned to give best results.

An eminently practical question for users of these methods is: what values should the parameters have? One could simply choose reasonable values for α, β as recommended by practitioners, for example for α Jacobs and Chase [2013] suggests a range of 0.05–0.5 and Chopra and Meindl [2013] an upper bound of 0.2. However, most textbooks recommend choosing values that minimise one-step-ahead forecast error based on historical data, typically using error measures such as mean absolute error (MAE) or mean squared error (MSE). The tuning task can be treated as an optimisation problem to be solved by hill-climbing, grid search, genetic algorithm or other heuristics. This approach is recommended in many sources:

- [Gardner 1985]: "It is dangerous to guess at values of the smoothing parameters. The parameters should be estimated from the data."
- [Gardner 2006]: "There is no longer any excuse for using arbitrary parameters in exponential smoothing given the availability of good search algorithms, such as the Excel Solver."
- Dielman [2006] used grid search to tune single and double exponential smoothing, but focused on the difference between MAE and MSE.
- Paul [2011] used trial and error to tune the SES smoothing factor to minimise MSE and MAE, and found optimal values greater than 0.8.
- Karmaker [2017] found that Solver gives similar results to trial-and-error but more quickly and easily, and found optimal values of 0.31 for MAE and 0.14 for MSE.
- Snyder [1988] proposed a recursive procedure for estimating parameters, and discussed work by Holt and Winters who proposed expensive numerical procedures.
- Another approach is maximum likelihood estimation [Hyndman et al. 2008].
- The Engineering Statistics Handbook [1] recommends the Marquardt procedure to optimise smoothing factors.
- Stevenson and Hojati [2007] recommends using forecasting errors to choose smoothing factors.

[1] NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

- From the Wikipedia page on exponential smoothing: [2] "There are cases where the smoothing parameters may be chosen in a subjective manner — the forecaster specifies the value of the smoothing parameters based on previous experience. However, a more robust and objective way to obtain values for the unknown parameters included in any exponential smoothing method is to estimate them from the observed data."

Thus practitioners are routinely advised to tune forecasting algorithms on data by some form of optimisation.

At first glance the optimisation approach seems straightforward, with only the problems of possible local minima and computational expense to overcome. Unfortunately things are not so simple: the optimal value for the smoothing factors might be undefined, or unreasonably small or large (even 0 or 1), a result that is probably of no practical use. This point is certainly not unknown but it is rarely mentioned in the literature, and Gardner [2006] wrote: "The research in choosing fixed parameters [...] is not particularly helpful." In forecasting competitions (see Makridakis et al. [1982] Appendix 2) the smoothing factor for single exponential smoothing was chosen to minimise MSE on historical data, but no mention is made of pathological cases. When they occur, practitioners presumably curse the software and use a default value.

A recent article that does tackle the problem is Ravinder [2016]. It used Solver which often gave incorrect results, but this was fixed by using its Multistart option. More seriously, on problems from the literature, several cases were found where smoothing factors of 0 or 1 were recommended for SES and DES using MAE and MSE. Allowing initial forecasts and trends to be tuned increased the number of such cases. A conclusion was that traditional guidelines should be ignored in such cases, and that parameters should be kept within reasonable bounds. (It was also recommended to choose initial forecasts and trends by averaging or regression, as Solver becomes very slow when optimising multiple parameters.)

In this work we provide additional evidence that error measures cannot always be used to tune simple forecasting methods. To fix the problem we propose a *data augmentation* approach, based on a Machine Learning technique but with parallels in the forecasting literature. The paper is organised as follows. Section 2 uses artificial series to show that three well-known forecasting methods do not always give reasonable smoothing factors values under optimisation. Section 3 proposes a fix for this problem. Section 4 concludes the paper with a discussion connecting the proposal with techniques from the literature, and proposes future work.

2. UNTUNABILITY OF FORECASTING METHODS

In this section we consider different forecasting methods and error measures, and devise artificial data to give bad optimal results for smoothing factors.

2.1 The untunability of SES

We shall assume that parameter optimisation is done after a long training period.

First consider the series $1, 0, 1, 0, \ldots$ which we shall call the A-series (Alternating). This is a special case of a series considered in Croston [1972], though no connection with optimal smoothing was drawn in that work. If we apply SES to the A-series with smoothing factor α, what value of α minimises MAE? In Appendix A we show that to minimise the MAE or MSE with $\alpha \in [0, 1]$ we must choose $\alpha = 0$. This is the correct result to the optimisation problem but it is useless as a forecasting method.

Next consider the series $1, 1, \ldots, 1, 0, 0, \ldots, 0$ which we shall call the O-series (Obsolescence): the name is taken from forecasting for inventory control where demand for an item can abruptly cease forever. We shall suppose demand is always 1 in our given data but 0 thereafter. In Appendix B we show that the MAE and MSE are both minimised by choosing $\alpha = 1$. This reduces SES to the naive forecasting algorithm sometimes called *random walk* (RW) or the *naive method*, which simply forecasts the last value encountered. RW is not useless but it is usually a poor choice.

In some cases there is no defined optimal value for smoothing factors. Suppose we try to use the Percent Best (PB) error measure to tune SES, which reports the percentage of times SES has smaller error than a baseline forecasting method, with RW as a popular baseline. On the A-series RW always makes exactly the wrong forecast $(0, 1, 0, 1, \ldots)$ whereas SES makes a forecast between 0 and 1 (unless α is 0 or 1). So SES *always* has a PB of 100% and PB cannot be used to tune it.

The above discussion shows that, on certain series, the SES smoothing factor can not or should not be tuned using error measures including MAE, MSE or PB. Either the errors are independent of the smoothing factors, or the optimal value is 0 or 1 which leads to an over-specialised forecasting algorithm that is unable to handle change. The same is also true of many other error measures that are based on MAE or MSE, for example root mean squared error, relative mean absolute error and Thiel's U2 statistic.

2.2 The untunability of DES

Ravinder [2016] reported several textbook series on which DES has optimal smoothing factors of 0 or 1. The DES analysis is harder, so instead we used grid search on the A- and O-series, with initial trends 0, initial forecasts set to the mean of the given data, long run-ins, and errors measured over 10 periods. For the A-series we obtained optimal values $\alpha = \beta = 0$ under both MAE and MSE. For the O-series the optimal values were $\alpha = 1$ and $\beta = 0$ under both MAE and MSE. The optimal β value under PB is 0, but the optimal α value under PB is undefined. Thus the optimality of α for DES on the A- and O-series is equivalent to that of α values for SES.

2.3 The untunability of CR

Croston's method (CR) [Croston 1972] is designed to handle *intermittent* data which contains many 0s, and is

[2] https://en.wikipedia.org/wiki/Exponential_smoothing#Optimization

generally thought to beat SES and DES on such data. CR uses two smoothing factors: α to smooth nonzero demand size and β to smooth inter-demand interval. We use the original version of CR. This has since been shown to be biased [Syntetos and Boylan 2005] but we believe our results also apply to other CR variants.

Our earlier artificial series can be adapted to CR. The IA-series (Intermittent Alternating)

$$2, 0, 1, 0, 2, 0, 1, 0, \ldots,$$

is based on the A-series. The inter-demand interval is a constant 2, so tuning CR to this series reduces to tuning SES to the series $2, 1, 2, 1, \ldots$ which is simply the A-series with each term incremented by 1. The optimal α value is therefore 0.

The IO-series (Intermittent Obsolescent)

$$2, 0, \ldots, 2, 0, 1, 0, 1, 0, \ldots$$

is based on the O-series. Tuning CR to this series reduces to tuning SES to the series $2, \ldots, 2, 1, 1, \ldots$ which is simply the O-series with each value incremented by 1. The optimal α value is therefore 1.

If we wish to tune CR by PB, consider the A-series which is intermittent. CR always forecasts 1 on this series, while RW always forecasts the wrong value (0 for 1 and vice-versa). Therefore whatever the value of α CR always has a PB of 50% or 100%, depending on how we treat ties. Therefore on the A-series the optimum α value for CR is undefined.

3. DATA AUGMENTATION WITH BLACK SWANS

In this section we propose a solution to the untunability problem.

3.1 A machine learning analogy

We draw an analogy between tuning forecasting methods and machine learning. In many machine learning tasks, such as regression and classification, a method is optimised by learning from *training data*. It is then evaluated on separate *testing data* which it has not previously seen.

Sometimes artificial or transformed data is added to the training set to improve robustness, which is called *data augmentation* (this term has a slightly different meaning in statistics). For images this could be left-right reversal, or random changes in brightness and contrast. This is done despite training datasets having tens of thousands of samples or more, and significantly improves accuracy.

By analogy, when we optimise a forecasting method on one or more time series, they can be considered training data and the process of optimisation as training. If optimisation leads to a "bad" result as above, it is only bad in the sense that it performs poorly on data not yet seen: on the training data it has learned perfectly well. We therefore believe that the problem lies not in the trained forecasting method, but in the training data.

If optimisation tells us to use a 0 or 1 smoothing factor, this is the best solution for the training data: the only drawback with it occurs if we believe that future data might be different. Our recommendation is a simple one: to include in the training data at least one series of a type that we are afraid might occur in future, such as non-stationary series, trending series, or pathological examples such as those in Section 2. Exactly what form this additional data is, and how much weight is placed on it, depends on our beliefs regarding future data, and on how risk-averse we wish to be. We shall refer to these additional series as *black swans* (inspired by Taleb [2007]) as they represent unseen cases that we are afraid might occur in the future.

3.2 Choosing black swans

For specific applications we can use any available expert opinion, or even slightly-related data from other sources, for example in forecasting for financial or mortage data we might use related figures taken from the financial crash of 2008, with a weight derived from our guess of the probability of a future crash. As another example, in forecasting for inventory control we could add time series with obsolescence, even if this does not occur in our historical data.

If no such data is available, and we have no concrete ideas for black swans, we propose using the above artificial series as "generic black swans". This is almost an oxymoron, as the definition of a black swan is something unexpected and non-standard. However, this is a simple way of avoiding useless smoothing factors.

3.3 Using black swans

We typically have a number of given data series available for tuning a forecasting method, possibly of different lengths. If they are also of different magnitudes then we might normalise them to have the same mean. Suppose our only training data is the A-series $1, 0, 1, 0, \ldots$ and we wish to optimise SES using MAE. As shown above, optimisation gives the solution $\alpha = 0$ which is correct but useless.

Now suppose that — based on expertise, history or pure caution – we suspect that future data might be quite different. We might theorise that with probability 0.1 we shall encounter a black swan series with a temporary trend, such as $0, 0, 0, 1, 2, 3, 4, 4, 4, 4$. If we augment the training data with this series, assigning a weight of 0.9 to the observed data and 0.1 to the artificial data, the solution (found by trial and error) is $\alpha = 0.19$ which is well within the bounds of typical recommendations.

If we have no ideas for black swans, we can simply add an A- and O-series to any dataset to prevent optimal smoothing factors from being 0 or 1. However, because the two series have different errors we should weight them differently. For SES the A-series has minimum MAE $\frac{1}{2}$ and MSE $\frac{1}{4}$ while the O-series has minimum MAE and MSE $\frac{1}{N}$ where N is the length of the unseen data (see Appendices A and B). We propose the following procedure:

- Normalising both series to have error 1 by rescaling their values.
- Possibly rescaling the black swans or the other series to normalise means.

- Rescaling the black swans by multiplying all their values by pM ($p \in (0,1)$) where M is the number of given series and $\frac{p}{1-p}$ is the relative weight we want black swans to have with respect to the given data. For example if we want the black swans to each have relative weight $p = 0.1$ then they should be rescaled by a factor of $0.1/(1-0.1) \approx 0.11$.

3.4 Making assumptions explicit

It might be argued that we can obtain any optimal α we like by adjusting the weight or form of a black swan, which is arbitrary. Have we gained anything? We argue that we have: we are now forced to state our assumptions on unseen data explicitly, as black swans with weights. For example, instead of stating:

> *Optimisation with a mean initial forecast I tells us to set the SES smoothing factor to 0.01, but we believe this is unreasonable so we shall use 0.1 as recommended by [Smith].*

we can now state:

> *Augmenting our data with black swan X with weight 0.1, optimisation recommends a smoothing factor of 0.13 (instead of 0.01 on the original data only).*

The advantage is that explicit assumptions can be questioned by experts.

4. DISCUSSION

It is often recommended to choose smoothing parameters for forecasting methods based on exponential smoothing, in order to minimise an error measure on given data. Because of the existence of simple time series on which optimisation fails to give a reasonable result, we believe that this recommendation should be modified. It might be argued that such cases can be treated in special ways, but where should one draw the line between those cases and very similar time series on which optimisation gives a reasonable result? Instead we recommend augmenting the given data by at least some hypothetical series, and we have provided two generic series with optimal smoothing parameters 0 and 1. Adding both to given data has the effect of moving the aggregate optimum smoothing factor away from these extreme values. We suggest that it might also improve forecast robustness if historical data is modified in random ways to generate additional series.

Planning for the unknown is an active field of research. Makridakis and Taleb [2009] point out that early papers by Makridakis, and the M-Competitions, demonstrated our inability to forecast accurately under uncertainty, and recent work by Taleb [2007] emphasises the importance of unpredictable events which are often ignored as outliers. They write:

> The future isn't it what it used to be, or alternatively, history never repeats itself in exactly the same way. This means that statistical models that extrapolate (or interpolate) past patterns/relationships cannot provide accurate predictions, since they presume that such patterns/relationships will not change (assumption of constancy).

They propose several ways of preparing for the unknown, including *protective strategies* such as hedging. We consider adding black swans to known data a form of protective strategy. Inventing useful black swans and their associated probabilities requires skills similar to those used by *superforecasters* [Tetlock and Gardner 2015].

Adding black swans to data can also be seen as a form of *risk management*. A recent IJF special issue call-for-papers on *Forecasting, Uncertainty and Risk Management* notes that we need *understanding that uncertainty exists practically everywhere*, including our given data. It divided risks into four classes, one being "unknown unknowns" or black swans: these are entirely unpredictable as they do not appear in the given data. To prepare for them we need *antifragile strategies* such as adding black swans to data.

Modifying the data by adding black swans can also be seen as a form of *judgemental forecasting*. Syntetos et al. [2009] survey the field and cite an earlier study [Goodwin 2000] that found requiring forecasters to justify their adjustments reduced the number of such adjustments. They also cite Fildes et al. [2009] who found that managerial adjustments improve accuracy. Syntetos et al. [2009] discuss possible ways of integrating judgements into forecasting. There are several forms of judgemental forecasting, and our approach is closest to *scenario analysis* [Hassani 2016] in which optimistic and pessimistic possibilities are considered. Scenario analysis does not rely on historical data and does not expect the future to look like the past. Wright and Goodwin [2009] further consider scenarios of low probability, and state that:

- Predictions must not by restricted by data in the reference class. They should also offer the potential to generate surprises.
- Overconfidence in a single future scenario, or in a narrow range of such scenarios, should be avoided.
- The method should exploit certainties or near certainties about the nature of the future.

In future work we intend to test the above ideas on real data, extend them to other error measures such as MAPE and MASE, and to other forecasting methods such as adjusted CR variants.

ACKNOWLEDGEMENTS

We would like to acknowledge the support of the Science Foundation Ireland CONFIRM Centre for Smart Manufacturing, Research Code 16/RC/3918.

REFERENCES

S. Chopra, P. Meindl. Supply Chain Management: Strategy, Planning and Operation. 5th edition, Prentice Hall, 2013.

J. D. Croston. Forecasting and Stock Control for Intermittent Demands. *Operational Research Quarterly* **23**:289–304, 1972.

T. Dielman. Choosing Smoothing Parameters for Exponential Smoothing: Minimizing Sums of Squared Versus

Sums of Absolute Errors. *Journal of Modern Applied Statistical Methods* **5**(1):118–129, 2006, JMASM, Inc.
R. Fildes, P. Goodwin, M. Lawrence, K. Nikolopoulos. Effective Forecasting and Judgemental Adjustments: an Empirical Evaluation and Strategies for Improvement in Supply Chain Planning. *Int. J. Forecasting* **25**:3–23, 2009.
E. S. Gardner Jr. Exponential Smoothing: the State of the Art. *Journal of Forecasting* **4**:1–28, 1985.
E. S. Gardner Jr. Exponential Smoothing: the State of the Art — Part II. *International Journal of Forecasting* **22**(4):637–666, 2006.
P. Goodwin. Improving the Voluntary Integration of Statistical Forecasts and Judgement. *Int. J. Forecasting* **16**:85–99, 2000.
B. Hassani. Scenario Analysis in Risk Management. Springer, 2016.
R. J. Hyndman, A. B. Koehler, R. D. Snyder. Forecasting with Exponential Smoothing. Springer, 2008.
F. R. Jacobs, R. B. Chase. Operations and Supply Chain Management: the Core. 3rd Edition, McGraw Hill Higher Education, 2013.
C. L. Karmaker. Determination of Optimum Smoothing Constant of Single Exponential Smoothing Method: a Case Study. *Int. J. Res. Ind. Eng.* **6**(3):184–192, 2017.
S. Makridakis, A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen R. Winkler. The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition. *Journal of Forecasting* **1**:111–153, 1982.
S. Makridakis, N. Taleb. Living in a World of Low Levels of Predictability. *International Journal of Forecasting* **25**(4):840–844, 2009.
S. K. Paul. Determination of Exponential Smoothing Constant to Minimize Mean Square Error and Mean Absolute Deviation. *Global Journal of Researches in Engineering* 01, 2011.
H. V. Ravinder. Determining The Optimal Values Of Exponential Smoothing Constants — Does Solver Really Work? *American Journal Of Business Education* **9**(1):1–14, 2016.
R. D. Snyder. Progressive Tuning of Simple Exponential Smoothing Forecasts. *Journal of the Operational Research Society* **39**(4):393–399, 1988.
W. J. Stevenson, M. Hojati. Operations Management. Boston: McGraw-Hill/Irwin, 2007.
A. A. Syntetos, J. E. Boylan. The Accuracy of Intermittent Demand Estimates. *International Journal of Forecasting* **21**:303–314, 2005.
A. A. Syntetos, J. E. Boylan, S. M. Disney. Forecasting for Inventory Planning: a 50-Year Review. *Journal of the Operations Research Society* **60**:149–160, 2009.
N. N. Taleb. The Black Swan: the Impact of the Highly Improbable. Random House Group, 2007.
P. E. Tetlock, D. Gardner. Superforecasting: The Art and Science of Prediction. Crown Publishers, 2015.
G. Wright, P. Goodwin. Decision Making and Planning Under Low Levels of Predictability: Enhancing the Scenario Method. *International Journal of Forecasting* **25**(4):813–825, 2009.

Appendix A. TUNING SES ON THE A-SERIES

Suppose the initial forecast is I. The first 3 forecasts are:

$$I \qquad \alpha + I(1-\alpha) \qquad (1-\alpha)[\alpha + I(1-\alpha)]$$

After running SES for a sufficiently long time we can assume that the first and third forecasts are equal (proof in Appendix C):

$$I = (1-\alpha)[\alpha + I(1-\alpha)]$$

hence

$$I = \frac{1-\alpha}{2-\alpha}$$

Therefore the first two forecasts are

$$\frac{1-\alpha}{2-\alpha} \qquad \alpha + \frac{(1-\alpha)^2}{2-\alpha}$$

and the first two absolute errors are

$$\left|1 - \frac{1-\alpha}{2-\alpha}\right| \qquad \left|\alpha + \frac{(1-\alpha)^2}{2-\alpha}\right|$$

The overall MAE is the mean of these, which reduces to:

$$\frac{1}{2-\alpha}$$

So to minimise the MAE with $\alpha \in [0,1]$ we must choose $\alpha = 0$ giving MAE 0.5 and $I = 0.5$.

Similarly for the MSE. We have the same forecasts but the MSE is the mean of:

$$\left(1 - \frac{1-\alpha}{2-\alpha}\right)^2 \qquad \left(\alpha + \frac{(1-\alpha)^2}{2-\alpha}\right)^2$$

which is

$$\frac{1}{(2-\alpha)^2}$$

Again to minimise MSE we must choose $\alpha = 0$ giving MSE=0.25 and $I = 0.5$.

Appendix B. TUNING SES ON THE O-SERIES

A reasonable initial forecast is the mean of the given data, which is 1. Then the forecasts for the new 0 data are

$$1 \quad (1-\alpha) \quad (1-\alpha)^2 \quad (1-\alpha)^3 \ldots$$

The absolute errors are the same terms, so the MAE is the mean of some finite number N of these terms. Whatever the value of N, this is minimised to $\frac{1}{N}$ by choosing $\alpha = 1$. Similarly, the squared errors are

$$1 \quad (1-\alpha)^2 \quad (1-\alpha)^4 \quad (1-\alpha)^6 \ldots$$

and the MSE is also minimised to $\frac{1}{N}$ by choosing $\alpha = 1$.

Appendix C. A CONVERGENCE PROOF

Suppose we increase I by δ for some real value δ which can be positive or negative. Then the third forecast is

$$\frac{1-\alpha}{2-\alpha}(1 + 2\delta - 3\delta\alpha + \delta\alpha^2)$$

which is $\delta(1-\alpha)^2$ greater than its original value. If $\alpha < 1$ then $|\delta(1-\alpha)^2| < |\delta|$ so increasing the first forecast by δ increases the third forecast by a smaller amount. By

the same argument, the increase to the fifth forecast is even smaller, and so on. Therefore when $\alpha < 1$ the odd-numbered forecasts converge to

$$I = \frac{1-\alpha}{2-\alpha}$$

Convergence only fails when $\alpha = 1$.