Turkish Journal of Forecasting vol. 01 no. 1 (2017) pp. 30-39

Content list available at JournalPark

Turkish Journal of Forecasting

Journal Homepage: tjforecast.xyz

Why and How Does Exponential Smoothing Fail? An In Depth Comparison of ATA-Simple and Simple Exponential Smoothing.

G. Yapar[1,*], I. Yavuz[1], H. T. Selamlar[1]

[1]*Dokuz Eylul University, School of Science, Department of Statistics, Tinaztepe, Izmir, Turkey*

A R T I C L E I N F O

Article history:
Received 02 June 2017
Accepted 20 July 2017
Available online 24 August 2017

Keywords:
Exponential smoothing
Forecasting
Initial value
M3-competition
Time series

A B S T R A C T

Even though exponential smoothing (ES) is publicized as one of the most successful forecasting methods in the time series literature and it is widely used in practice due to its simplicity, its accuracy can be affected by the initialization and optimization procedures followed. It also suffers from some fundamental problems that can be seen clearly when its weighting scheme is studied closely. Exponential smoothing fails to account for the amount of data points that can contribute to the forecast when assigning weights to historical data. ATA smoothing has been proposed as an alternative forecasting method and is shown to perform better than ES when the accuracies are compared on empirical data. In this paper, the properties of ATA that make it stand out from ES models will be discussed by just comparing the simple versions of both models Empirical performance of the two simple models will also be compared based on popular error metrics.

1. Introduction

Whenever there is a scientific, industrial, commercial or economic activity, forecasting usually is a crucial part of the process and doing it accurately makes a big difference. Therefore, the forecasting literature keeps expanding at a fast pace inevitably. A comprehensive review of the progress during the 25 year period until the year 2005 can be found in [1]. Undoubtedly ARIMA ([2]) and exponential smoothing (ES) ([3]) are still the two dominant major forecasting techniques and other methods are frequently derived or inspired from them. Among the two, ES methods are applied more frequently due to their simplicity, robustness ([4]) and accuracy as automatic forecasting procedures especially in the famous M-competitions ([5–7]). Literature reviews on ES are given in [8], [9] and [10].

ES models consist of a family of models which assume that the time series has up to three underlying data components: level, trend and seasonality. In ES the goal is to obtain estimates for the level, trend and seasonal pattern and then to use these final values to forecast the future. Each model contains one of the five types of trend (none, additive, damped additive, multiplicative, and damped multiplicative) and one of the three types of seasonality (none, additive, and multiplicative) in addition to a level. [11] started the initiative to provide a taxonomy of ES methods and [12], [7] and [13] extended and modified these ideas. When different combinations of trend and seasonality are considered, 15 different ES models can be formed. The best known of these are SES (no trend, no seasonality), Holt's linear model (additive trend, no seasonality) and Holt-Winters' additive model (additive trend, additive seasonality). [10] have proposed the ETS state space models, which provide a solid theoretical foundation for ES. In this paper, in addition to what had already been done in the literature, additive and multiplicative error terms were assumed for each of the 15 models in previous taxonomies resulting in 30 potential ES models.

* Corresponding author.
E-mail addresses: guckan.yapar@deu.edu.tr (Guckan Yapar), idil.yavuz@deu.edu.tr (Idil Yavuz), hanife.taylan@deu.edu.tr (Hanife Taylan Selamlar)

Despite the fact that there has been substantial research on ES models, some of its shortcomings and fundamental issues have not been resolved and these keep affecting the quality of forecasts obtained using this approach. First and most overlooked of all is the fact that like in all time series literature it is assumed that future will be a continuation of the past and therefore ES models aim to assign relatively more weight to recent observations compared to older ones. This however does not always happen and as we will present with examples later on ES models fall victim to cases where an optimum smoothing parameter may be chosen such that the already unknown (estimated) initial value receives more weight than the most recent observation. Second, there is still no agreed upon consensus on the initialization and optimization of ES models and this in return yields to accuracy problems. For example, [14] showed that even though for other exponential smoothing models the type of initialization and loss functions that are employed did not result in significant changes in post sample forecasting accuracies, for Holt's linear trend model they were very influential especially for long term forecasting horizons. Even if this was not the case, the fact that trying to find an optimal initial value both complicates and prolongs the optimization process cannot be overlooked. Finally, when smoothing a data set over time, the weights should be distributed to observations taking into account where along the time-line the value being smoothed resides, i.e. the most recent observation can receive more weight when there are fewer data points that are contributing to the smoothed value and a little less weight as we move along the time-line. ES models, on the other hand, always assign the most recent value the same weight no matter where along the time-line smoothing is being carried out. All these issues keep ES from performing well under some circumstances and ATA smoothing helps deal with these problems.

In this paper, we will compare in depth the simple versions of ATA proposed by [15] and [16] and exponential smoothing based on popular metrics that are commonly used for comparing forecasting techniques. Since it is already shown in [16] that generalizing ATA to higher order models is straightforward and the higher order ATA model's accuracy is better than its counter ES model, comparing the models in their simple forms here is sufficient and the results can easily be expanded. Details on the results from the more sophisticated ATA methods on M3-sompetition data ([6]) along with R code and an Excel macro to implement the methods can be found on the website "https://atamethod.wordpress.com/". Here, the resulting accuracy measures obtained by applying simple versions of both methods to the M3-competition data sets will be provided for comparison.

2. The ATA method

For the series $X_t, t = 1, 2, \dots, n$, the model which we will denote by $ATA(p, q)$ throughout the paper can be written as:

$$S_t = \left(\frac{p}{t}\right) X_t + \left(\frac{t-p}{t}\right)\left(S_{t-1} + T_{t-1}\right) \tag{1}$$

$$T_t = \left(\frac{q}{t}\right)\left(S_t - S_{t-1}\right) + \left(\frac{t-q}{t}\right) T_{t-1} \tag{2}$$

$$\hat{X}_t(h) = S_t + hT_t \tag{3}$$

for $p \in (1, \dots, n), q \in (0, \dots, n), t > p \geq q$ and $h = 1, 2, \dots$. For $t \leq p$ let $S_t = X_t$, for $t \leq q$ let $T_t = X_t - X_{t-1}$ and $T_1 = 0$. Here X_t is the value of the original series, T_t is the trend component and $S_t$is the smoothed value at time *t*. p is the smoothing parameter for level, q is the smoothing parameter for trend and $\hat{X}_t(h)$ is the h step ahead forecast value.

Recognize that ATA has similar form to ES but the smoothing parameters are now dependent on the number of observations. The specific ATA model defined in equations (1)-(3) mimics the Holt linear trend model. It is also worth pointing out that when $q = 0$, $ATA(p, q)$ reduces to the ATA-simple model with no trend, i.e. for $t > p$:

$$S_t = \left(\frac{p}{t}\right) X_t + \left(\frac{t-p}{t}\right) S_{t-1} \tag{4}$$

and $S_t = X_t$ for $t \leq p$. This model is very similar to the simple exponential smoothing model (SES) $S_t = \alpha X_t + (1-\alpha) S_{t-1}$ for $\alpha \in (0,1)$.

The forecasts for $ATA(p,0)$ can be computed easily as $\hat{X}_t(h)=\bar{X}$ for $h=1,2,\ldots$. Throughout the rest of the paper the comparisons and applications will be carried out using $ATA(p,0)$ and simple exponential smoothing for brevity and simplicity but the results can easily be generalized to models with more components.

3. Comparison of $ATA(p,0)$ and SES

While the functional forms of ATA models are generally very similar to those of exponential smoothing models, there are distinctive features of ATA that separate it from ES. $ATA(p,0)$ can be thought of as an approach that lies in between moving averages (MA) and simple exponential smoothing (SES). $ATA(p,0)$ attaches weights to only the most recent $(n-p)$ observations and zero weights to the other p observations like a MA model and the weights decrease exponentially like SES for some $p(p\geq 3)$. The weighting scheme of $ATA(p,0)$ however, is more flexible and intuitive than SES. For $p=1$ the h step ahead forecast value for $ATA(p,0)$ $\hat{X}_t(h)=\bar{X}$ i.e. all observations contribute equally to the forecast. This very important estimator, which intuitively should be the starting point for any method, can never be formed with SES. For $p=2$, $ATA(p,0)$ produces weights that decrease linearly with slope $\frac{2}{n(n-1)}$ and intercept $\frac{2}{n}$. This also cannot be achieved with a SES model. For $p\geq 3$, $ATA(p,0)$ produces exponentially decreasing weights similar to but not exactly the same as SES. In this case, $ATA(p,0)$ gives greater emphasis than SES to the most recent history and less emphasis than SES to the more distant past at the same smoothing constant level, i. e. when both models give the same weight to the most recent data point $X_n\left(\alpha=\frac{p}{n}\right)$ to estimate X_{n+1}. See Figure 1 for illustration of weights assigned to observations for various p levels.

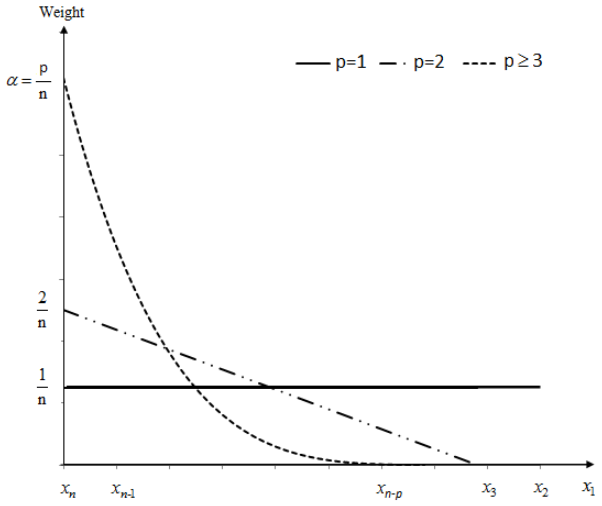

Figure 1. Weights assigned to observations by ATA$(p,0)$ for various p values to obtain $\hat{X}_n(h)$

Not only is $ATA(p,0)$ more flexible but also it is more adaptive to the data at hand. No matter where along the time line smoothing is being carried out, the weights attached to the observations by SES stay fixed. The observations are assigned the weights $\alpha,\alpha(1-\alpha),\alpha(1-\alpha)^2,\ldots$ regardless of the sample size at hand every time. This is not the case for $ATA(p,0)$ as the weights change with respect to the sample size as the weights are:

$$\left(\frac{p}{n}\right),\left(\frac{p}{n}\right)\left(\frac{n-p}{n-1}\right),\left(\frac{p}{n}\right)\left(\frac{n-p}{n-1}\right)\left(\frac{n-p-1}{n-2}\right),\ldots$$

respectively.

Also, the optimization process is easier for $ATA(p,0)$. Even though both SES and $ATA(p,0)$ methods require smoothing constants, $ATA(p,0)$ does not need an initial value unlike SES. When the optimal smoothing parameter is found, the initial value is found simultaneously. With ATA there is no limitation on the number of observations to forecast, only one observation is enough. The smoothing constants for ES are commonly estimated heuristically

(between (0,0.3), [8]) or estimated by minimizing a pre-determined error measure like the mean squared error (MSE), the mean absolute error (MAE) or the mean absolute percentage error (MAPE). After discussing various theoretical and empirical arguments for selecting an appropriate smoothing constant, [8] concludes that it is best to estimate an optimum α from the data. This is generally done by a grid search of the parameter space for α where $\alpha \in (0,1)$. In practice, various α values starting from 0.01 with increments 0.01 are tried and the α value that produces the minimum error is chosen. The number of iterations required to find the optimum smoothing constant for ES is then 100 for any data set. The number of iterations for higher order smoothing models (level, trend and season) to find the optimum smoothing constant combinations will be huge (100^3)On the other hand, for $ATA(p,0)$ the search for the optimal smoothing constant is much easier since the constant depends on the choice of p and $p \in \{1,2,\dots,n\}$. Therefore the total number of iterations needed is only n and it will be n^3 for higher order smoothing models. As a result, when the data size is less than 100, the total number of iterations for $ATA(p,0)$ will always be less than those for ES.

Forecasting models can also be compared on some well-known metrics like the average age of the method and the variance of the forecast. The average age $\bar{k}$ of a model is a measure of the model's ability to utilize fresh data. The smaller $\bar{k}$ the better. [3] defines the average age $\bar{k} = n - \sum_{t=1}^{n} t w_t$ where w_t is the weight given to the t^{th} observation when trying to obtain a forecast. The variance of the forecast can be calculated as $V\left(\hat{X}_t(h)\right) = \sum_{t=1}^{n} w_t^2 \sigma^2 = V\sigma^2$ ([3]). The average age and V of SES are given in [3] as $\bar{k} = \frac{1-\alpha}{\alpha}$ and $V = \frac{\alpha}{2-\alpha}$. In order to compare SES and $ATA(p,0)$ on these metrics, the models will now be given in a compact form as follows. The smoothed value at time t for SES can be represented in the weighted average form as:

$$S_t = \sum_{k=0}^{t-1} \alpha(1-\alpha)^k X_{t-k} + (1-\alpha)^t S_0, \tag{5}$$

and when the model in (4) is applied recursively to all observations in the series, S_t for $ATA(p,0)$ can be written in the alternative form:

$$S_t = \sum_{k=0}^{t-(p+1)} \frac{\binom{t-k-1}{p-1}}{\binom{t}{p}} X_{t-k} + \frac{1}{\binom{t}{p}} S_p, \tag{6}$$

where S_p is the starting value of $ATA(p,0)$ which is simply the p^{th} observation. The weights of $ATA(p,0)$ as given (6) can be thought of as the probabilities from a Negative Hyper-Geometric distribution with parameters $(t,p,1)$ ([17]).

Utilizing the expected value of this distribution, the average age of $ATA(p,0)$ can then be easily found as $\bar{k} = \frac{n-p}{p+1}$. The sum of the squared weights, V, can be written as:

$$V = \sum_{t=1}^{n} w_t^2 = \left(\frac{p}{n}\right)^2 \left[1 + \sum_{t=0}^{n-p-1} \prod_{j=0}^{i} \left(\frac{n-p-j}{n-1-j}\right)^2\right] = \left(\frac{p}{n}\right)^2 {}_3F_2\left((1, p-n, p-n),(1-n,1-n),1\right) \tag{7}$$

From equation (7) it can be seen that V involves the Generalized Hyper-Geometric series ${}_3F_2\left((1, p-n, p-n),(1-n,1-n),1\right)$ ([18]).

At the same smoothing constant level $\left(\alpha = \frac{p}{n}\right)$ the average age of $ATA(p,0)$ is smaller than the average age of SES $\bar{k}_{ATA} = \frac{n-p}{p+1} < \bar{k}_{SES} = \frac{1-\alpha}{\alpha}$, therefore $ATA(p,0)$ should be preferred by researchers since it utilizes fresher data.

In order for the two models to have equal average ages the smoothing constant of SES should be given the value $\alpha = \frac{p+1}{n+1}$ When smoothing constants for the two models are chosen in this fashion to make the average ages equal, $ATA(p,0)$ is still preferable since then $V_{ATA(p,0)} < V_{SES} = \frac{\alpha}{2-\alpha}$.

To present the discussions above in a more organized way, the weights that $ATA(p,0)$ SES and MA assign to the observations when trying to obtain $\hat{X}_{t+1}$ are given in Tables 1 and 2 for sample sizes 12 and 30 respectively. The tables also contain the average ages, variance components and the initial values and weights assigned to them. The initial value for the ES are marked as "?" since there are various ways that the initial value can be assigned.

Table 1. Weights assigned to observations by ATA, ES and MA for $n=12$ to obtain $\hat{X}_{12}(1)$

	$p=1$ and $\alpha=1/12$			$p=2$ and $\alpha=2/12$			$p=3$ and $\alpha=3/12$			$p=6$ and $\alpha=6/12$		
t	ATA	ES	MA	ATA	ES	MA	ATA	ES	MA	ATA	ES	MA
12	0.083	0.083	0.083	0.167	0.167	0.167	0.250	0.250	0.250	0.500	0.500	0.50
11	0.083	0.076	0.083	0.152	0.139	0.167	0.205	0.188	0.250	0.273	0.250	0.50
10	0.083	0.07	0.083	0.136	0.116	0.167	0.164	0.141	0.250	0.136	0.125	-
9	0.083	0.064	0.083	0.121	0.096	0.167	0.127	0.105	0.250	0.061	0.063	-
8	0.083	0.059	0.083	0.106	0.080	0.167	0.095	0.079	-	0.023	0.031	-
7	0.083	0.054	0.083	0.091	0.067	0.167	0.068	0.059	-	0.006	0.016	-
6	0.083	0.049	0.083	0.076	0.056	-	0.045	0.044	-	-	0.008	-
5	0.083	0.045	0.083	0.061	0.047	-	0.027	0.033	-	-	0.004	-
4	0.083	0.042	0.083	0.045	0.039	-	0.014	0.025	-	-	0.002	-
3	0.083	0.038	0.083	0.030	0.032	-	-	0.019	-	-	0.001	-
2	0.083	0.035	0.083	-	0.027	-	-	0.014	-	-	0.000	-
1	-	0.032	0.083	-	0.022	-	-	0.011	-	-	0.000	-
Initial value	X_1	?	-	X_2	?	-	X_3	?	-	X_6	?	-
Weight of initial	0.083	0.352	-	0.015	0.112	-	0.005	0.032	-	0.001	0	-
AA	5.500	6.776	5.500	3.348	4.327	2.500	2.259	2.873	1.500	0.863	1.000	0.50
V	0.083	0.162	0.083	0.116	0.102	0.167	0.164	0.144	0.250	0.347	0.333	0.50

From the tables it can be seen that for p=1 the MA and $ATA(p,0)$ models assign the same weights to observations therefore they have equal average ages, however, SES has larger average age with a slightly smaller variance. For p=2 the weights of the oldest two observations are zero for $ATA(p,0)$ and the average age and variance of the model is now between those of MA and ES with ES having the largest. For p=3, now the weights of the oldest three observations are zero for $ATA(p,0)$ and the average age and variance of the model are again between those of MA and ES with ES having the largest. It is worth drawing attention to the differences between the weights attached to the initial values by the ES and $ATA(p,0)$. For all smoothing levels ES assigns a relatively much larger weight to the initial value compared to $ATA(p,0)$. The fact that ES is not adaptive to the data can be seen by looking at the p=6 column of Table 1 and the p=15 column of 2. Since $\alpha = 0.5$ for both of these cases, even though the weights should be distributed among 12 observations in the first case and 30 observations in the second case, ES assigns 0.5 to the most recent, 0.273 to the second most recent etc., exactly the same weights for both data sets. $ATA(p,0)$ on the other hand takes into account the amount of data that can be utilized and is able to distribute the weights in a fashion that still favours the recent observations but lets the model utilize more recent points at the same time. When the tables are studied closely, it can be seen that $ATA(p,0)$ always assigns more weight to recent observations and less weight to older observations when the models are at the same smoothing level.

Table 2. Weights assigned to observations by ATA, ES and MA for $n=30$ to obtain $\hat{X}_{30}(1)$

	$p=1$ and $\alpha=1/30$			$p=2$ and $\alpha=2/30$			$p=3$ and $\alpha=3/30$			$p=6$ and $\alpha=6/30$		
t	ATA	ES	MA	ATA	ES	MA	ATA	ES	MA	ATA	ES	MA
30	0.033	0.033	0.033	0.067	0.067	0.067	0.100	0.100	0.100	0.500	0.500	0.50
29	0.033	0.032	0.033	0.064	0.062	0.067	0.093	0.09	0.100	0.259	0.250	0.50
28	0.033	0.031	0.033	0.062	0.058	0.067	0.086	0.081	0.100	0.129	0.125	-
27	0.033	0.03	0.033	0.06	0.054	0.067	0.080	0.073	0.100	0.062	0.063	-
26	0.033	0.029	0.033	0.057	0.051	0.067	0.074	0.066	0.100	0.029	0.031	-
25	0.033	0.028	0.033	0.055	0.047	0.067	0.068	0.059	0.100	0.013	0.016	-
24	0.033	0.027	0.033	0.053	0.044	0.067	0.062	0.053	0.100	0.005	0.008	-
23	0.033	0.026	0.033	0.051	0.041	0.067	0.057	0.048	0.100	0.002	0.004	-
22	0.033	0.025	0.033	0.048	0.038	0.067	0.052	0.043	0.100	0.001	0.002	-
21	0.033	0.025	0.033	0.046	0.036	0.067	0.047	0.039	0.100	0.000	0.001	-
20	0.033	0.024	0.033	0.044	0.033	0.067	0.042	0.035	-	0.000	0.000	-
19	0.033	0.023	0.033	0.041	0.031	0.067	0.038	0.031	-	0.000	0.000	-
18	0.033	0.022	0.033	0.039	0.029	0.067	0.033	0.028	-	0.000	0.000	-
17	0.033	0.021	0.033	0.037	0.027	0.067	0.030	0.025	-	0.000	0.000	-
16	0.033	0.021	0.033	0.034	0.025	0.067	0.026	0.023	-	0.000	0.000	-
15	0.033	0.02	0.033	0.032	0.024	-	0.022	0.021	-	-	0.000	-
14	0.033	0.019	0.033	0.030	0.022	-	0.019	0.019	-	-	0.000	-
13	0.033	0.019	0.033	0.028	0.021	-	0.016	0.017	-	-	0.000	-
12	0.033	0.018	0.033	0.025	0.019	-	0.014	0.015	-	-	0.000	-
11	0.033	0.018	0.033	0.023	0.018	-	0.011	0.014	-	-	0.000	-
10	0.033	0.017	0.033	0.021	0.017	-	0.009	0.012	-	-	0.000	-
9	0.033	0.016	0.033	0.018	0.016	-	0.007	0.011	-	-	0.000	-
8	0.033	0.016	0.033	0.016	0.015	-	0.005	0.010	-	-	0.000	-
7	0.033	0.015	0.033	0.014	0.014	-	0.004	0.009	-	-	0.000	-
6	0.033	0.015	0.033	0.011	0.013	-	0.002	0.008	-	-	0.000	-
5	0.033	0.014	0.033	0.009	0.012	-	0.001	0.007	-	-	0.000	-
4	0.033	0.014	0.033	0.007	0.011	-	0.001	0.006	-	-	0.000	-
3	0.033	0.013	0.033	0.005	0.010	-	-	0.006	-	-	0.000	-
2	0.033	0.013	0.033	-	0.010	-	-	0.005	-	-	0.000	-
1	-	0.012	0.033	-	0.009	-	-	0.005	-	-	0.000	-
Initial value	X_1	?	-	X_2	?	-	X_3	?	-	X_{15}	?	-
Weight of initial	0.033	0.362	-	0.002	0.126	-	0.000	0.042	-	0.000	0.000	-
AA	14.500	18.150	14.500	9.333	12.107	7.000	6.757	8.491	4.500	0.938	1.000	0.50
V	0.032	0.015	0.033	0.045	0.034	0.067	0.062	0.053	0.100	0.339	0.333	0.50

The weights these three approaches assign to observations for these two cases (n=12 and n=30) are visualized in Figures 2 and 3. From the figures it can be seen that $ATA(p,0)$ starts with weights equal to those from MA for p=1 and as p increases it starts to produce weights similar to SES with the exception that it keeps assigning more weight to recent observations and less weight to older observations while assigning some of the oldest observations zero weight like MA. As p increases the weights from $ATA(p,0)$ get closer to the weights from SES.

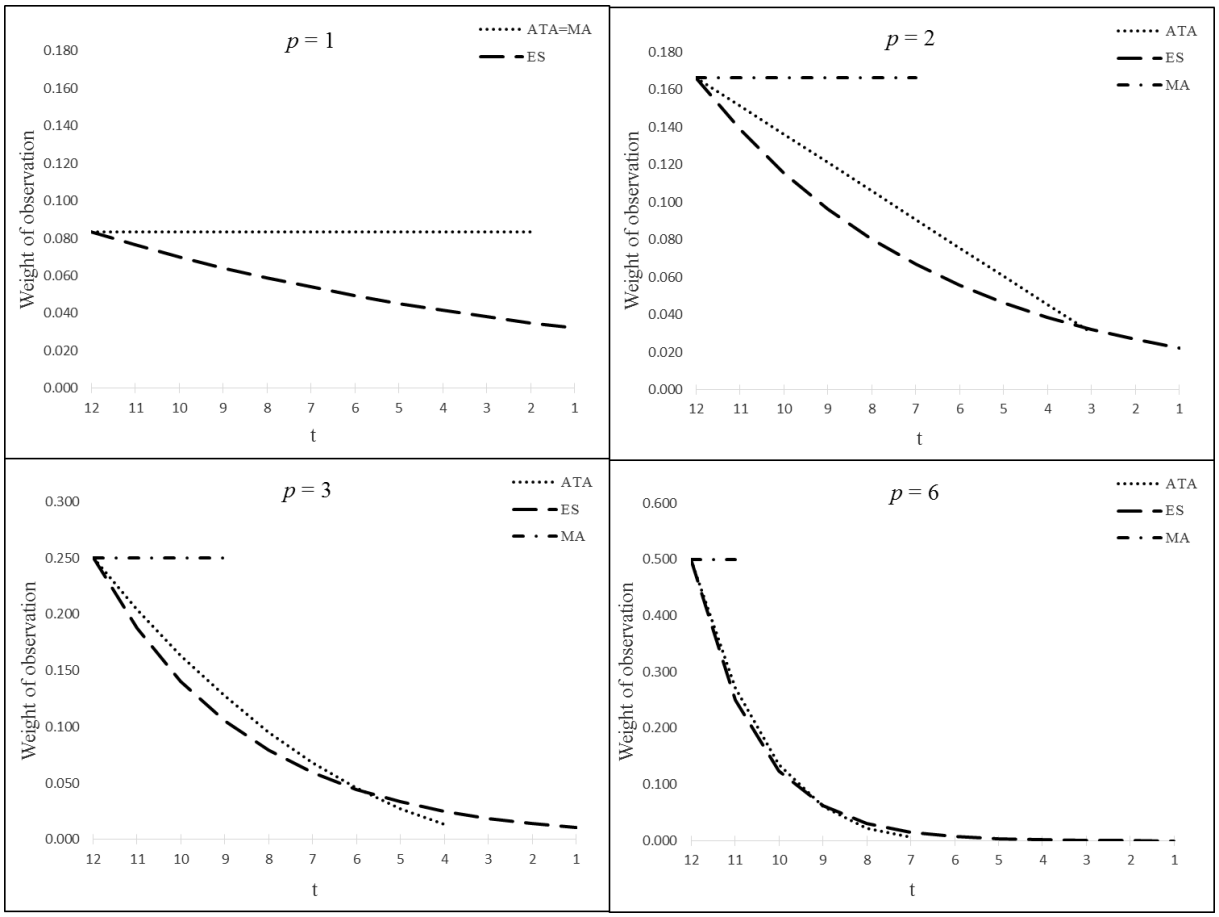

Figure 2. Weights assigned to observations by $ATA(p,0)$, SES and MA for *n*=12 and various *p* values

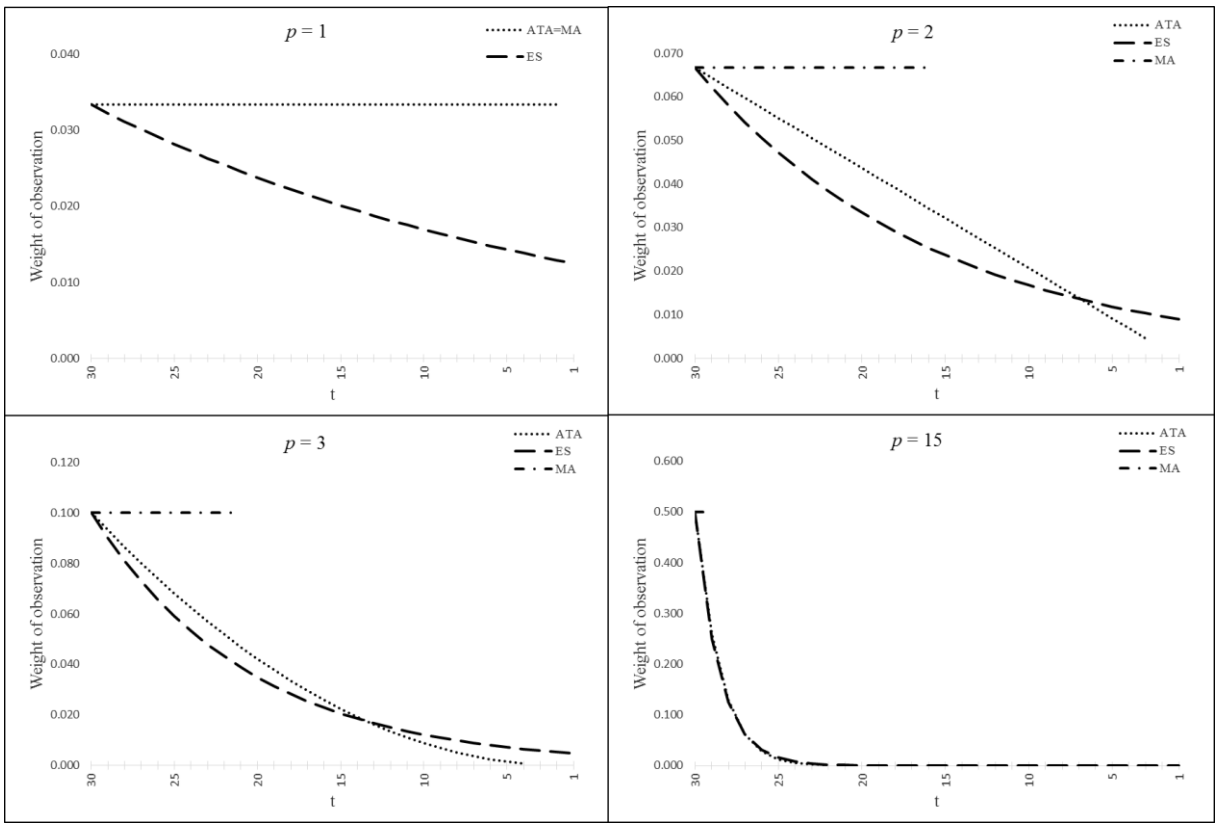

Figure 3. Weights assigned to observations by $ATA(p,0)$, SES and MA for *n*=30 and various *p* values

To summarize, if the goal is to estimate the final value for a component and it is known that it is changing in time, it makes sense to give greater emphasis to the most recent history versus the more distant past, due to fact that the component is changing and therefore the recent history should more accurately reflect current conditions. For this reason it is not surprising that ATA is more accurate than traditional ES methods. Also philosophically, ATA does never violate the concept of exponential smoothing; from the point of view that the recent data is more representative of the future therefore should be assigned more weight. This is guaranteed with ATA since the smallest weight attached to the most recent observation is $1/n$. However, with exponential smoothing models this is not always the case since the weight assigned to the most recent observation can be much smaller than that assigned to the initial value which is highly contradictory of the time series concept that the recent past should receive greater emphasis when predicting the future.

4. M-3 competition results

To compare $ATA(p,0)$ and SES on their forecasting accuracies, we applied both methods to the M3-competition data ([6]) since this collection is the most recent and comprehensive time-series data collection available with verified results. This collection consists of 3003 data sets from various fields. Data sets are of various lengths, with different kinds of trend and seasonality components and each data set consists of in-sample and out-sample data points. When comparing the methods, the optimum smoothing parameters are obtained by minimizing an in-sample error measure and then the forecasts up to 18 steps ahead (the number of steps as specified in the M3-competition) are computed to obtain the average out-sample errors for both models. The data sets are deseasonalized by the classical decomposition method of the ratio-to-moving averages, if necessary and reseasonalized forecasts are produced for as many steps ahead as required.

First, to stay consistent with the rest of the literature, the symmetric mean absolute percentage errors (sMAPE) were used. The in-sample one-step-ahead sMAPE can be defined as:

$$sMAPE = 200 \times mean\left(\frac{\left| X_t - \hat{X}_t \right|}{\left| X_t \right| + \left| \hat{X}_t \right|} \right)$$

where X_t is the actual value and $\hat{X}_t$ is the one-step-ahead forecast. For all data sets, the required numbers of forecasts (for the pre-determined forecasting horizons) were computed and out-sample sMAPEs were averaged across all 3003 series for each forecasting horizon. The results are given in Table 3.

Table 3. Average sMAPE across different forecasting horizons: all 3003 series

	Forecasting horizons										Averages					
Method	1	2	3	4	5	6	8	12	15	18	1--4	1--6	1--8	1--12	1--15	1--18
SES	9.5	10.6	12.7	14.1	14.3	14.9	13.3	14.5	18.3	19.4	11.73	12.68	12.82	13.12	13.66	14.31
ATA	8.9	10.0	12.1	13.7	13.9	14.7	12.8	13.9	17.3	18.9	11.16	12.21	12.34	12.64	13.13	13.77

When the methods are compared based on sMAPE as in Table 3, it can be seen that $ATA(p,0)$ produces smaller average errors for all individual forecasting horizons. The errors are averaged for short and long term horizons on the right side of the table so that the differences between the errors can be more clearly seen. Overall $ATA(p,0)$'s average sMAPE is 13.77 compared to 14.31 for SES which is significantly larger. It is also worth noting that the average sMAPE for all forecasting horizons for the ETS models proposed by [10] is exactly the same (13.77) as the average sMAPE for $ATA(p,0)$. This is very impressive since $ATA(p,0)$, a single model, can produce as accurate forecasts as ETS which performs a model selection on 24 exponential smoothing models for each data set.

Another comparison can be made based on the mean absolute scaled error (MASE) proposed by [19]. MASE can be calculated using

$$q_t = \frac{e_t}{\frac{1}{n-1}\sum_{t=2}^{n} \left| X_t - X_{t-1} \right|},$$

$$MASE = mean\left(\left| q_t \right| \right)$$

where e_t in the numerator of q_t is the error from the method and denominator of q_t can be thought of as the in-sample mean absolute error from the naive forecast method where the forecast for any time point t is simply assumed to be X_{t-1}. MASE is independent of the scale of the data and it will be less than one if the method gives, on average, smaller errors than the one-step ahead errors from the naive method. It is preferred over other scaled measures since it is less sensitive to outliers and scaled errors are starting to be more commonly used for comparing forecast accuracy across series on different scales. We summarize the average MASE for both models in Table 4. When the comparison is based on this metric, $ATA(p,0)$ still performs better than SES on each forecasting horizon and on average for both short and long term forecasting horizons.

Table 4. Average MASE across different forecasting horizons: all 3003 series

	Forecasting horizons										Averages					
Method	1	2	3	4	5	6	8	12	15	18	1--4	1--6	1--8	1--12	1--15	1--18
SES	0.78	1.03	1.36	1.63	1.82	2.05	1.49	0.97	1.25	1.38	1.20	1.45	1.44	1.35	1.32	1.33
ATA	0.75	1.00	1.33	1.61	1.80	2.04	1.47	0.97	1.20	1.37	1.17	1.42	1.42	1.33	1.30	1.30

5. Conclusion

In this paper, the comparisons between ATA and ES are carried out based on the simple versions of the two approaches. These comparisons in addition to the empirical performances of the methods prove that based on accuracy, simplicity, speed and interpretability ATA is better than ES.

ATA is very flexible and the next line of research on ATA will be incorporating dampened and multiplicative trend components to the model in addition to obtaining prediction intervals. The fact that even the simple version of ATA performs so well is very promising and it is inevitable that the more sophisticated versions where combinations and model selection are allowed will perform much better.

References

[1] J.G. de Gooijer, R. Hyndman, 25 Years of IIF Time Series Forecasting: A Selective Review, 2005. http://econpapers.repec.org/RePEc:tin:wpaper:20050068.

[2] G.E.P. Box, G.M. Jenkins, Time series analysis: forecasting and control, Holden-Day, 1970. https://books.google.com.tr/books?id=5BVfnXaq03oC.

[3] R.G. Brown, Statistical forecasting for inventory control, 1959.

[4] P. Goodwin, The Holt-Winters Approach to Exponential Smoothing : 50 Years Old and Going Strong, Foresight Int. J. Appl. Forecast. (2010) 30–33.

[5] S. Makridakis, The Forecasting accuracy of major time series methods, Wiley, 1984. https://books.google.com.tr/books?id=T6qCAAAAIAAJ.

[6] S. Makridakis, M. Hibon, The M3-Competition: results, conclusions and implications, Int. J. Forecast. 16 (2000) 451–476. doi:10.1016/S0169-2070(00)00057-1.

[7] R. Hyndman, A.B. Koehler, R. Snyder, S. Grose, A state space framework for automatic forecasting using exponential smoothing methods, Int. J. Forecast. 18 (2002) 439–454. http://econpapers.repec.org/RePEc:eee:intfor:v:18:y:2002:i:3:p:439-454.

[8] E.S. Gardner, Exponential smoothing: The state of the art, J. Forecast. 4 (1985) 1–28. doi:10.1002/for.3980040103.

[9] E.S. Gardner, Exponential smoothing: The state of the art--Part II, Int. J. Forecast. 22 (2006) 637–666. http://econpapers.repec.org/RePEc:eee:intfor:v:22:y:2006:i:4:p:637-666.

[10] R. Hyndman, A.B. Koehler, J.K. Ord, R. Snyder, Forecasting with Exponential Smoothing: The State Space Approach, Springer Berlin Heidelberg, 2008. https://books.google.com.tr/books?id=GSyzox8Lu9YC.

[11] C.C. Pegels, On Startup or Learning Curves: An Expanded View, A I I E Trans. 1 (1969) 216–222. doi:10.1080/05695556908974435.

[12] E.S. Gardner, E. Mckenzie, Forecasting Trends in Time Series, Manage. Sci. 31 (1985) 1237–1246. doi:10.1287/mnsc.31.10.1237.

[13] J.W. Taylor, Exponential smoothing with a damped multiplicative trend, Int. J. Forecast. 19 (2003) 715–725. doi:10.1016/S0169-2070(03)00003-7.

[14] S. Makridakis, M. Hibon, Exponential smoothing: The effect of initial values and loss functions on post-sample forecasting accuracy, Int. J. Forecast. 7 (1991) 317–330. doi:10.1016/0169-2070(91)90005-G.

[15] G. Yapar, Modified simple exponential smoothing, Hacettepe J. Math. Stat. doi:10.15672/HJMS.201614320580.

[16] G. Yapar, S. Capar, H.T. Selamlar, I. Yavuz, Modified Holt's Linear Trend Method, Hacettepe Univ. J. Math. Stat. (n.d.). doi:10.15672/HJMS.2017.493.

[17] N.L. Johnson, S. Kotz, Urn models and their application: an approach to modern discrete probability theory, Wiley, 1977. https://books.google.com.tr/books?id=ZBfvAAAAMAAJ.

[18] W.N. Bailey, Generalized hypergeometric series, The University Press, 1935. https://books.google.com.tr/books?id=J7VsAAAAMAAJ.

[19] R. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, Int. J. Forecast. 22 (2006) 679–688. http://econpapers.repec.org/RePEc:eee:intfor:v:22:y:2006:i:4:p:679-688.