

Academic Journal of Nawroz University (AJNU), Vol.12, No.4, 2023 This is an open access article distributed under the Creative Commons Attribution License Copyright ©2017. e-ISSN: 2520-789X https://doi.org/10.25007/ajnu.v12n4a1717



# **Box-Cox Transformation for Exponential Smoothing With Application**

Alla Ahmed Ali<sup>1</sup>, Haithem Taha Mohammed Ali<sup>2,3</sup>

<sup>1</sup> Department of Statistics, University of Duhok, Kurdistan Region, Iraq
 <sup>2</sup> Department of Economic Sciences, University of Zakho, Kurdistan Region, Iraq
 <sup>3</sup> Department of Economics, Nawroz University, Kurdistan Region, Iraq

## ABSTRACT

This article introduces a novel algorithm for incorporating power transformation into the estimation process of a Holt-Winters Seasonal model. The algorithm outlines a series of steps aimed at selecting the most appropriate power parameter estimate. This selection is achieved using the conventional Maximum Likelihood Estimation method in combination with various criteria for enhancing statistical modeling efficiency. Supplementary decision rules include assessing Mean Square Error, Mean Absolute Error, and conducting a p-value test for the normality of errors. The algorithm's effectiveness is demonstrated through its application to real-world data. Ultimately, the article affirms the feasibility of obtaining viable solutions for selecting the optimal power parameter.

**KEYWORDS**: Box-Cox Transformation, Exponential Smoothing, Holt-winters.

### **1. INTRODUCTION**

Time series forecasting plays a pivotal role, involving the collection and analysis of historical data related to a singular variable. This data is then leveraged to construct a model that captures the inherent relationships. Projecting this time series into the future is made possible through this model. Forecasting techniques serve as vital tools in decision-making across diverse domains, such as finance, management, the environment, and economics (Lima et al., 2019). Even after more than five decades of consistent application, exponential smoothing remains among the most practically valuable forecasting methods (Goodwin, 2010). The primary goal of transformation is to enhance the statistical analysis of time series data by identifying an appropriate model. In instances where the variance fluctuates alongside the mean measurement level, а straightforward and established class transformation is employed to stabilize the variance (Bartlett, 1947). In 1957, Tukey introduced the power transformer as a means to achieve distribution normality or at the very least, symmetrize error distribution (Tukey, 1957). By 1962, Box and Tidwell had developed the practice of back-transforming data into its original domain. Their focus was on normalizing error distribution and transforming independent variables

without compromising homoscedasticity (Box and Tidwell, 1962). Addressing anomalies such as nonadditivity, non-normality, and heteroscedasticity, Box and Cox presented a parametric power transformation approach in 1964 (Box and Cox, 1964). In 1969, Draper and Cox demonstrated that a power transformation satisfying non-normality would yield an approximately robust estimation of the power parameter, corresponding to a reasonably symmetrical distribution (Draper and Cox, 1969). Poirier's 1978 statement highlighted the challenge of assessing the insignificance of truncation impact, given its dependence on unknown distributional factors (Poirier, 1978). In 1980, Carroll proposed an alternative method for obtaining a robust estimator, distinct from the likelihood method, while also delving into the theoretical exploration of approximate normality (Carroll, 1980). The year 2000 saw Yeo and Johnson introduce a new family of distributions, unrestricted in application, possessing several desirable properties of the Box-Cox transformation (BCT). This new family allowed for use with both positive and negative variable values (Yeo and Johnson, 2000). Adrian, in 2014, combined transforms with exponential smoothing methods in the pursuit of enhanced forecasts. Two transform types were

investigated: the first applied directly to time series data, and the second indirectly to prediction errors. Results indicated that non-transformed time series were notably inferior to some transformed counterparts (Beaumont, 2014). In 2016, Bergmeir et al. utilized bagging exponential smoothing methods coupled with Seasonal, Trend Level (STL) decomposition and BCT. A technique was presented involving components' packing for exponential smoothing methods. This technique employed BCT followed by STL analysis to break down the time series into trend, seasonal components, and residuals. A new sequence was formed using bootstrapped residuals after smoothing the remaining data with an animation block (Bergmeir et al., 2016). In 2019, Voulgaraki applied BCT to sales forecasting. Using historical monthly sales data from Greece's new car retail industry, several time series models were compared for estimation and forecasting. The findings highlighted significantly superior predictions from Box-Cox transformed data compared to untransformed data (Voulgaraki, 2019). In 2021, Aytaç utilized BCT in conjunction with the Prophet algorithm for forecasting Turkey's hazelnut export quantities. The primary goal was to estimate hazelnut export volumes from Turkey over the subsequent 36 months, starting from June 2020. The Prophet algorithm facilitated the forecasts, with adjustments made to enhance prediction accuracy. A booster file was employed to establish data set stability and frequency. The time series data underwent the Shapiro-Wilk test. The Prophet algorithm, aided by BCT, revealed the data set's seasonality, showing monthly oscillations in export volumes. Given that Turkey's hazelnut harvest occurs in August, monthly export figures started climbing, peaking in October (Aytaç, 2021).

The focal point of this paper revolves around leveraging power transformations to enhance the predictive capacity of time series models. The application of BCT within exponential smoothing, particularly the Holt-Winters method, was undertaken. The remaining sections of the paper are structured as follows: The subsequent segment delves into the theoretical underpinnings of exponential smoothing. Following that, the third section explores the theoretical foundations of the BCT. The fourth section outlines the proposed algorithm for the development of a Holt-Winters exponential smoothing model using the BCT. Moving forward, the fifth section provides a practical dimension to the study. Concurrently, the sixth section outlines the conclusions.

## **2.1 EXPONENTIAL SMOOTHING**

The foundational research conducted by Brown (1959, 1962) and Holt (1960), aimed at creating forecasting models for inventory management systems, marked the inception of exponential smoothing techniques in the

1950s (Fomby, 2008). These smoothing methodologies find application in both seasonal and non-seasonal time series analyses, enabling the estimation and provision of reasonably accurate short-term forecasts. One of the most prevalent predictive approaches, exponential smoothing equations for parameter estimation and prediction possess generation an intuitive and easily comprehensible nature, making them widely employed in business contexts. When applying smoothing techniques to data, observed values are utilized to derive smoothed values for the time series. These smoothed values are then employed to predict future time series values, forming the fundamental concept of exponential smoothing (Washington et al., 2003). The term Smoothing" "Exponential itself underscores the exponential decline of weights as observations age (Hyndman et al., 2008). Smoothing is a mechanism to minimize deviations among string values around the curve representing the overall pattern of the string. During data smoothing, previous values are employed to generate a smoothed value for the time series, which is subsequently extended; this process can be seen as the transformation of unruly data into a smoother form, rendering it more manageable. Moving averages and smoothing constitute pivotal tools in the exponential smoothing process (Yaffee and McGee, 2000). This technique is especially suitable for forecasting series that exhibit trends, seasonality, or both (Shastri et al., 2018). Seasonal behavior in a time series denotes its inclination to exhibit recurrent patterns at regular intervals. The term "season" denotes the period preceding the repetition of the behavior. Figure (1) illustrates various non-seasonal and seasonal curves. For added seasonality, the series demonstrates consistent seasonal fluctuations irrespective of its overall level, while for multiplicative seasonality, the extent of seasonal fluctuations varies in relation to the general series level. These characteristics are visually evident in the series graphs themselves (Hyndman et al., 2008). In 1969, Pagels classified exponential smoothing methods as suitable for series featuring constant level or direction (additive or multiplicative), with both non-seasonal and seasonal attributes (Pagels, 1969). In 2002, Hyndman, Rob J, et al. provided a taxonomy for these methods, within which the well-known Holt-Winters Exponential Smoothing (HWES) method emerged. This method is particularly designed for seasonal time series manifesting uniform seasonal patterns. It has even been adapted to accommodate multiple seasonality recently (Hyndman et al., 2002).



Figure 1: Trend types and seasonal indicators Additive and Multiplicative (Gardner, 1985).

In 1960, Winters extended the double exponential smoothing technique, leading to its recognition as the HWES method. This specific smoothing technique is tailored exclusively for seasonal time series, ideally suited for those exhibiting a single seasonal pattern. Widely employed for forecasting within seasonal time series data, this method is structured around three foundational equations: one for level, another for trend, and a third for seasonality. Much like the Holt method, HWES incorporates an additional equation to handle seasonality, contingent upon whether the seasonality is multiplicative or additive (Taylor and Snyder, 2012). The additive approach of HWES is particularly appropriate for time series data where the amplitude of the seasonal effect remains consistent, independent of the mean level of the series. Expressing the h-step-ahead forecasting for a time series at period *t* with seasonality *s*, the additive approach of HWES can be formulated (Setiawan et al., 2017),

$$\hat{y}_{t+h} = \mu_t + b_t h + I_{t+h+s}$$
, h  
= 1,2,... (1)

Where  $\hat{y}_{t+h}$  denotes the forecasted value at period t + h, and,

$$\mu = \alpha (Y_t - I_{t-s}) + (1 - \alpha)(\mu_t + b_{t-1})$$
(2)

$$b_t = \beta(\mu_t - \mu_{t-1}) + (1 - \beta)b_{t-1}$$
(3)

$$I_t = \delta(Y_t - \mu_t) + (1 - \delta)I_{t-s}$$
(4)

Where  $\alpha$ ,  $\beta$ , and  $\delta$  are constants. They need to be estimated in order to minimize the Mean Square Error (MSE),

$$MSE = \frac{\sum (y_t - \hat{y}_t)^2}{n}$$
(5)

and the Mean Absolute Error (MAE),

$$MAE = \frac{\sum |y_t - \hat{y}_t|}{n} \tag{6}$$

In this context,  $y_t$  represents the observed value,  $\hat{y}_t$  signifies the predicted value, and *n* stands for the number

of data points.  $\mu_t$  signifies the smoothed local level at time t after accounting for seasonality and trend.  $b_t$  represents the smoothed local seasonal trend at time t, while  $s_t$  indicates the smoothed local seasonal index at time t. The variable h denotes the forecast horizon, and  $Y_t$  refers to the observation at time t, serving as an index for a given period.

For the initial estimation of seasonal indices, it is imperative to employ data spanning at least one complete season (i.e., a set of periods). Consequently, the initialization of trends and levels takes place at specific periods. The initialization of the level involves computing the average of the first season's data (Nuurhamidah et al., 2020).

$$\mu_s = \frac{1}{s} (Y_1 + Y_2 + \dots + Y_S) \tag{7}$$

Note that this is the average of orders and will eliminate the data's seasonality. To initialize the trend, it is convenient to use two complete seasons (i.e , 2s periods) as follows,

$$b_t = \frac{1}{s} \left( \frac{Y_{s+1} - Y_1}{s} + \frac{Y_{s+2} - Y_2}{s} + \dots + \frac{Y_{s+s} - Y_s}{s} \right)$$
(8)

Each of these terms is an estimate of the trend over one complete season and the initial estimate (Nuurhamidah, et al, 2020),

$$I_t = Y_1 - \mu_s \ I_2 = Y_2 - \mu_s , \dots, I_t = Y_s - \mu_s.$$
(9)

Finding the  $\alpha$ ,  $\beta$  and  $\delta$  parameters value improves the model's performance (Mi et al., 2018).

#### 2.2 BOX-COX TRANSFORMATION

A set of operations referred to as power transformation is utilized to monotonically alter data through power functions. This technique proves effective in reducing variance, rendering data more akin to a normal distribution, and enhancing the validity of association metrics such as the Pearson correlation between variables when compared to other data stabilization approaches. Among the most frequently employed transformations in this domain is the BCT method, which was introduced in 1964 (Sakia, 1992). The foundations of this method are built upon the premise of normalizing the response variable's data through transformation and subsequently determining the distribution of the original data. In this section of the article, we have also reexamined the scenario of the univariate normal random variable, adjusting certain aspects to align with the article's objectives.

Assuming that  $z_1, z_2, ..., z_n$  represents the observations of the transformed data of the random variable Y . let Z be

distributed according to the normal distribution with mean  $\mu$  and variance  $\sigma^2$  and has the following equation,

$$f_Z(z) = (2\Pi\sigma^2)^{-1/2} \cdot exp\left\{\frac{-1}{2\sigma^2}(z-\mu)^2\right\} , Z \in \mathbb{R}$$
(10)

and suppose that the original data  $y_1, y_2, \dots, y_n$  have been transformed using BCT model,

$$Z = Y^{(\lambda)} = \begin{cases} \frac{y^{\lambda} - 1}{\lambda} & , \lambda \neq 0\\ ln y & , \lambda = 0 \end{cases}$$
(11)

Through the derivation of the inverse function of Eq. (11) obtained,

$$y = g(z) = \begin{cases} (z\lambda + 1)^{1/\lambda} & , \lambda \neq 0 \\ exp(z) & , \lambda = 0 \end{cases}$$
(12)

It means that the transformation is from Z to g(z). Using the Change-of-Variables Technique, the following equation represents the probability density function of the random variable *Y*, which represents the original data  $y_1, y_2, \dots, y_n$ ,

$$f_Y(y) = (2\Pi\sigma^2)^{-1/2} \cdot exp\left\{\frac{-[g^{-1}(z) - \mu]^2}{2\sigma^2}\right\} \cdot J(\lambda; y)$$
  
,  $y \in \mathbb{R}^+$  (13)

where  $g^{-1}(z)$  represents the inverse function of Eq. (12).

$$f_Y(y) = (2\Pi\sigma^2)^{-1/2} y^{\lambda-1} . exp\left\{\frac{-1}{2\sigma^2} (\frac{y^{\lambda} - 1}{\lambda} - \mu)^2\right\}$$
(14)

As for the methods of estimating the power parameter  $\lambda$ , one of the criteria is to choose the estimator that maximizes the following log-likelihood function of the PDF of the original variable y,

$$L_{Max}(\lambda, y) = \frac{-1}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} (z - \mu)^2) + \log J(\lambda, y)$$
(15)

Where  $\sigma^{2}(\lambda)$  is the variance estimator of *y*.

#### **3. COMPUTATIONAL ALGORITHM**

In this part of the article, the following proposed computational algorithm was developed in estimation to develop HWES using the BCT on the original data and then to deal with the transformed data. The choice of optimum power parameter  $\lambda$  in this algorithm is based on the three following different criteria; The first is  $L_{Max}(\lambda, y)$  the MLE value of the original random variable *Y* according to Eq. 15. The second is the Select the value of  $\lambda$ , which corresponds to the largest value of the log-

likelihood function random variable *Z* according to Eq.10. The third is to use various test statistics such as the significant or highest p-value of the Shapiro-Wilk test of fit test statistic of original data normality. Therefore, the proposed application algorithm for using the of BCT model to develop a HWES model was as follows:

Step 1: Estimate the HWES of Eq. 1

Step 2: Fix  $\lambda \in \Lambda$ , where  $\Lambda = \{-2, -1.9, ..., 0, ..., 1.9, 2\}$ Step 3: Transform *Y* to  $Z = Y^{(\lambda)}$  using BCT according to Eq 11.

Step 4: Calculate  $L_{Max}(\lambda, y)$ .

Step 5: Estimate the values of HWES model using the inverse transdorm of *z* according to Eq. 12 .

Step 6: Calculate the p-value of the Shapiro-Wilk test of the original data resulting from step 5.

Step 7: Repeat all the steps from 3 to 7 for all values of  $\,\lambda$  in  $\Lambda$  .

#### 4. APPLICATION

BCT was applied to the time series dataset of monthly Electric and Gas Utilities. The time series includes 972 observations in the United States (TSU.S) from 1940 to 2020. We use data on the Industrial production of electric and gas utilities in the United States from 1985–2022. Except for those in U.S. territories, this data measures the actual production of all pertinent facilities in the United States, irrespective of who owns them. Electric and gas utilities industrial production [IPG2211A2N], as downloaded from FRED by the Federal Reserve Bank of St. Louis;

https://fred.stlouisfed.org/series/IPG2211A2N, AICS= 2211,2, Source Code: IP.G2211A2.N This data will be processed using R Software [Figure 2]



Figure 2: The monthly Electric and Gas average series (TSU.S)



Figure 3: The curve of MLE according Eq. 15

The outcomes of the algorithm's application to time series data are presented in Table 1. The table encompasses the forecasting results for the three evaluation criteria utilized in determining the optimal parameter value, namely: MSE and MAE values constitute the first and second criteria. The guiding principle here is that the optimal power parameter value corresponds to the lowest values of these criteria. The third criterion involves the p-value derived from the Shapiro-Wilk statistic for testing data normality. The guiding principle here is that the optimal power parameter value corresponds to the highest value of this indicator. A higher p-value indicates stronger evidence of data closely resembling a normal distribution. Observing Table 1, it becomes evident that the optimal power parameter value is 0.6, aligning with the minimum values of the first and second criteria and the maximum value of the third criterion.

Figure 3 depicts the curve of MLE as per Eq. 15. The peak of the curve corresponds to the optimal power parameter value. This visualization underscores that the three criteria outlined in Table 1 have effectively led to the selection of the optimal parameter, in alignment with the traditional MLE approach.

## Table 1: The values of MSE, MAE, and p-value corresponding to all values of $\lambda \in \Lambda$

λ	-2	-1.9	-1.8	-1.7	-1.6	-1.5	-1.4	-1.3	-1.2	-1.1
MES	21.694	21.847	19.849	19.431	19.422	21.801	21.846	21.847	19.298	21.829
MAE	14.674	14.353	14.717	14.634	14.869	14.765	14.372	14.811	14.748	14.582
P-value	0.8558	0.828	0.795	0.810	0.814	0.801	0.804	0.787	0.757	0.786
λ	-1	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1
MES	18.196	18.712	21.838	21.648	20.459	21.845	16.766	21.845	21.849	21.841
MAE	14.669	14.797	14.720	14.641	14.587	14.776	14.637	14.708	14.687	14.531
P-value	0.930	0.885	0.868	0.836	0.851	0.933	0.878	0.866	0.932	0.933
λ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MES	19.904	29.501	21.000	29.259	21.235	18.487	16.216	21.844	21.842	21.850
MAE	14.729	14.593	14.594	14.647	14.834	14.628	14.504	14.497	14.717	14.729
P-value	0.928	0.918	0.891	0.835	0.953	0.929	0.952	0.952	0.952	0.944
λ	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
MES	18.487	18.487	21.851	20.599	21.849	22.090	17.370	18.161	21.844	21.445
MAE	14.653	14.572	14.692	14.624	14.341	14.641	14.790	14.810	14.650	14.729



Figure 4: MLE plot according to the dataset (TSU.S)



Figure 6: The plot of the p-value of the SWT

Table 2: The estimates of the optimal power parameter according to the three criteria

Dataset	Deserves	$\begin{array}{c} Optimal \\ \lambda \end{array}$	The optimal value			Criteria values		
	Responses		α	β	δ	MSE	MAE	P-value
	Zτ	1.0	0.376	0.000	0.304	18.48793	14.768	0.8084
TSU.S	$Z^\lambda_\tau \mid  BCT$	0.6	0.375	1.82e-11	0.321	16.21658	8.922	0.9520

#### 6. CONCLUSION

Numerous time series exhibit non-stationary variances, necessitating an appropriate variance-stabilizing transformation to address this concern. The utilization of power transformation offers a means to reduce variance. In many instances, this transformation fosters increased normality and variance stabilization. Several techniques exist for selecting the optimal power parameter, categorized into two types: the first involves established estimation methods such as MLE. The second approach employs efficiency criteria within time series modeling as decision guidelines for power parameter estimation. This article endeavors to derive a viable solution by considering various estimation methods and decision rules to identify the optimal parameter, consequently attaining the utmost efficiency improvement for the Holt-Winters' method with additive seasonality in forecasting.

#### REFERENCES

Aytaç, E. (2021). Forecasting Turkey's Hazelnut Export Quantities with Facebook's Prophet Algorithm and Box-Cox Transformation. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 10(1), 33-47.

- Bartlett, M. S. (1947). The use of transformations. Biometrics, 3(1), 39-52.
- Beaumont, A. N. (2014). Data transforms with exponential smoothing methods of forecasting. International Journal of Forecasting, 30(4), 918-927.
- Bergmeir, C., Hyndman, R. J., & Benítez, J. M. (2016). Bagging exponential smoothing methods using STL decomposition and Box-Cox transformation. International journal of forecasting, 32(2), 303-312.
- Box, G. E., & Cox, D. R. (1964). An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological), 26(2), 211-243.
- Box, G. E., & Tidwell, P. W. (1962). Transformation of the independent variables. Technometrics, 4(4), 531-550.
- Carroll, R. J. (1980). A robust method for testing transformations to achieve approximate normality. Journal of the Royal Statistical Society: Series B (Methodological), 42(1), 71-78.
- Draper, N. R., & Cox, D. R. (1969). On distributions and their transformation to normality. Journal of the Royal Statistical Society: Series B (Methodological), 31(3), 472-476.
- Fomby, T. B. (2008). Exponential smoothing models. Mannual sas/ets software: time series forecasting system, 6, 225-235.
- Gardner Jr, E. S. (1985). Exponential smoothing: The state of the art. Journal of forecasting, 4(1), 1-28.
- Goodwin, P. (2010). The holt-winters approach to exponential smoothing: 50 years old and going strong. Foresight, 19(19), 30-33.
- Hyndman, R., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). Forecasting with exponential smoothing: the state space approach. Springer Science & Business Media.
- Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. International Journal of forecasting, 18(3), 439-454.
- Mi, J., Fan, L., Duan, X., & Qiu, Y. (2018). Short-term power load forecasting method based on improved exponential smoothing grey model. Mathematical Problems in Engineering, 2018.
- Nurhamidah, N., Nusyirwan, N., & Faisol, A. (2020). Forecasting seasonal time series data using the holtwinters exponential smoothing method of additive models. Jurnal Matematika Integratif, 16(2), 151-157.
- Pegels, C. C. (1969). Exponential forecasting: Some new variations. Management Science, 311-315..
- Poirier, D. J. (1978). The use of the Box-Cox transformation in limited dependent variable

models. Journal of the American Statistical Association, 73(362), 284-287.

- Sakia, R. M. (1992). The Box-Cox transformation technique: a review. Journal of the Royal Statistical Society: Series D (The Statistician), 41(2), 169-178.
- Setiawan, W., Juniati, E., & Farida, I. (2016, October). The use of Triple Exponential Smoothing Method (Winter) in forecasting passenger of PT Kereta Api Indonesia with optimization alpha, beta, and gamma parameters. In 2016 2nd International Conference on Science in Information Technology (ICSITech) (pp. 198-202). IEEE.
- Shastri, S., Sharma, A., Mansotra, V., Sharma, A., Bhadwal, A., & Kumari, M. (2018). A study on exponential smoothing method for forecasting. International Journal of Computer Sciences and Engineering, 6(4), 482-5.
- Taylor, J. W., & Snyder, R. D. (2012). Forecasting intraday time series with multiple seasonal cycles using parsimonious seasonal exponential smoothing. Omega, 40(6), 748-757.
- Tukey, J. W. (1957). On the comparative anatomy of transformations. The Annals of Mathematical Statistics, 602-632..
- Voulgaraki, M. K. (2019). Box Cox transformation in forecasting sales. Evidence of the Greek Market.
- Washington, S., Karlaftis, M., Mannering, F., & Anastasopoulos, P. (2020). Statistical and econometric methods for transportation data analysis. Chapman and Hall/CRC.
- Yaffee, R. A., & McGee, M. (2000). An introduction to time series analysis and forecasting: with applications of SAS® and SPSS®. Elsevier.

Yeo, I. K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. Biometrika, 87(4), 954-959.