2005

On the Validity of the Geometric Brownian Motion Assumption

Rahul Ratnakar Marathe
Iowa State University

Sarah M. Ryan
Iowa State University, smryan@iastate.edu

On the Validity of the Geometric Brownian Motion Assumption

Abstract
The geometric Brownian motion (GBM) process is frequently invoked as a model for such diverse quantities as stock prices, natural resource prices and the growth in demand for products or services. We discuss a process for checking whether a given time series follows the GBM process. Methods to remove seasonal variation from such a time series are also analyzed. Of four industries studied, the historical time series for usage of established services meet the criteria for a GBM; however, the data for growth of emergent services do not.

Keywords
engineering economics, geometric Brownian motion (GBM), lognormal growth, stock procies, chemical industry, electric utilities, industrial economics, industrial engineering, Brownian movement

Comments
The Version of Record of this manuscript has been published and is available in Engineering Economist 2005, http://www.tandfonline.com/10.1080/00137910590949904. Posted with permission.

On the Validity of the Geometric Brownian Motion Assumption

Rahul R. Marathe
Department of Industrial and Manufacturing Systems Engineering
Iowa State University
Ames, IA 50011-2164

Sarah M. Ryan*
Department of Industrial and Manufacturing Systems Engineering
Iowa State University
Ames, IA 50011-2164

* Corresponding author: smryan@iastate.edu Phone: 515-294-4347

The Version of Record of this manuscript has been published and is available in Engineering Economist 2005, http://www.tandfonline.com/10.1080/00137910590949904. Posted with permission.

On the Validity of the Geometric Brownian Motion Assumption

Abstract

The geometric Brownian motion (GBM) process is frequently invoked as a model for such diverse quantities as stock prices, natural resource prices, and the growth in demand for products or services. We discuss a process for checking whether a given time series follows the GBM process. Methods to remove seasonal variation from such a time series are also analyzed. Of four industries studied, the historical time series for usage of established services meet the criteria for a GBM, however the data for growth of emergent services do not.

The Version of Record of this manuscript has been published and is available in Engineering Economist 2005, http://www.tandfonline.com/10.1080/00137910590949904. Posted with permission.

I. Introduction

Many recent engineering economic analyses have relied on an implicit or explicit assumption that some quantity that changes over time with uncertainty follows a geometric Brownian motion (GBM) process. Below we briefly review a number of applications in different areas. The GBM process, also sometimes called a lognormal growth process, has gained wide acceptance as a valid model for the growth in the price of a stock over time. In fact, [9] refers to it as "the model for stock prices". Under this model, the Black-Scholes formulas for pricing European call and put options, as well as their variations for a few of the more complex derivatives, provide relatively simple analytical evaluation of asymmetric risks. The increasingly numerous and varied applications of the GBM model to processes other than the stock price motivate this paper: to review the assumptions underlying the GBM model, outline established statistical procedures for checking these assumptions, and illustrate their applications to actual data series.

Many recent examples of GBM models have arisen in real options analysis, in which the value of some "underlying asset" is assumed to evolve similarly to a stock price. In some cases, the GBM assumption is stated explicitly, while in others it is implicitly used when options are evaluated by the Black-Scholes formula. In [17], the cost of applying quality control charts was quantified using real option pricing methods, where both the sales volume and the price of a product were assumed to follow GBM processes. The same authors discussed the problem of product outsourcing as a real options problem in [18]. Here, three variables are supposed to follow the GBM process; viz. the unit cost of internally producing the item, the unit outsourcing cost of the item, and the unit delivery cost of outsourced items during the time interval. The GBM process has been also assumed in problems related to natural resources. In [25] the real options theory is applied to decisions of establishing a new forest stand and it is assumed that the future net prices of roundwood follow a GBM process. In [2], the Black-Scholes option pricing formula is applied to the capital allocation for investment. For the machine replacement problem considered in the paper, the present value of the machine cash flows is modeled as a GBM process. The options value of expansion flexibility in evaluating manufacturing investment is studied in [13], wherein the authors use sequential exchange options to value expansion flexibility in justifying the investment. In valuing flexibility an initial investment is considered as being analogous to purchasing an option to exchange one risky asset (subsequent investment, called the delivery asset) for another risky asset (returns accruing from the subsequent asset, called the optioned asset) within a time period from the initial investment. The prices for both of the assets are assumed to follow the GBM.

The GBM model has also been used to represent future demand in capacity studies. In [28], the authors studied capacity utilization over time assuming demand followed a GBM. An indirect validation of the assumption

was provided by [14], which showed in an empirical study of the chemical industry that actual capacity utilization matched the predictions from the model in [28]. In [22] demand for services in rapidly growing industries was assumed to follow a GBM and the expansion policy to minimize cost subject to a service level constraint was developed and analyzed. In this paper, we analyze data to test whether the GBM model is valid for demand; however, the methods we employ are applicable to any data series. Our approach is motivated by Ross [20], who analyzed data for crude oil prices and found that they were inconsistent with a key assumption of the GBM model. In this paper we also consider seasonal effects and show how to remove them before testing for the GBM characteristics. The effect of seasonality may be overcome in financial markets: Samuelson [23] proved that, even when spot prices have systematic seasonal variation, futures prices will not. However, in the demand series we examine in this paper, there are no quantities analogous to futures prices. As pointed out in [25], the GBM process assumption must be subject to test. Where significant financial impacts may result from the decision, it is of utmost importance to verify that a time series follows the GBM process, before relying on the result of such an assumption.

The next section discusses the theory of the GBM process and the parameters involved. The definitions and concepts of the Brownian motion used in the paper are explained in this section. Some data series may contain seasonal variation in addition to exponential growth with uncertainty. Hence before testing the GBM assumption the data series must be deseasonalized. In Section III, two methods of removing the seasonal indices are studied and the unbiased method is selected. Finally, the theory and methods developed in sections II and III are applied to real-life time series in Section IV. The data analyzed in this paper are from varied industries. As the cellular phone industry has been growing multi-fold over short recent intervals, it makes an interesting case to be considered as a GBM process. Also analyzed are airline passenger enplanements, electric power consumption and the growth of the Internet. Finally the results obtained from the data are discussed and summarized. We have our concluding remarks and plan for future work in the last section.

II. Geometric Brownian Motion (GBM) Process

2.1 Preliminaries

A Markov process is a particular type of stochastic process where only the present value of a variable is relevant for predicting the future. The past history of the variable and the way that the present has emerged from the past are irrelevant. A Wiener process is a type of Markov stochastic process in which the mean change in the value of the variable is zero with the variance of change equal to one per unit time. The Wiener process was first applied

in physics to describe the motion of a particle that is subject to a large number of small molecular shocks and was called Brownian motion [9]. The mathematical description of the process was later developed by Wiener [20].

If a stochastic process $\{z(t), t \geq 0\}$ follows a Brownian motion process, it exhibits the following two properties.

- Property I: The change in the value of z, Δz, over a time interval of length Δt is proportional to the square root of Δt where the multiplier is random; specifically, $\Delta z = z(t+\Delta t) - z(t) = \varepsilon\sqrt{\Delta t}$, where ε is a standard normal random variable. Hence values of Δz follow a normal distribution with mean 0 and variance equal to the change in time (Δt) over which Δz is measured.
- Property II: The changes in the value of *z(t)* for any two non-overlapping intervals of time are independent.

Using the principles of ordinary calculus where it is usual to proceed from small changes to the limit as the small changes becomes closer to zero, the Wiener process is the limit as $\Delta t \to 0$ of the process described above for *z(t)*.

A Wiener process is not differentiable with respect to time [15] as seen from the fact that:

$$E\left[\frac{z(s)-z(t)}{s-t}\right]^2 = \frac{s-t}{(s-t)^2} = \frac{1}{s-t} \to \infty, \text{ as } s-t \to 0.$$

However, it is useful to define a term for the expression dz/dt. A term commonly used in engineering to denote this quantity is white noise. The white noise process is the derivative of the Brownian motion process, which does not exist in the normal sense.

The standard Brownian motion process has a drift rate of zero and a variance of one. The drift rate of zero means that the expected value of z at any future time is equal to the current value. The variance of one means that variance of the change in z in a time interval of length T is equal to T. The Brownian motion process is the basis for a collection of more general processes. These generalizations are obtained by inserting white noise in an ordinary differential equation.

A generalized Brownian motion process is of the type: $dx = a\,dt + b\,dz,$ where a and b are constants and z is a Brownian motion process. To understand the equation, each of the components is considered separately. The first term implies that x has an expected drift rate of a per time unit, whereas the second term involving dz can be regarded as adding noise or variability to the path followed by x. The amount of this noise is b times the differential of the Brownian motion process. Hence for a small interval of time, the change in the value of x, Δx, is given by

$$\Delta x = a\Delta t + b\varepsilon\sqrt{\Delta t} \ .$$

Thus Δx has a normal distribution with mean $a\Delta t$ and variance $b^2\Delta t$.

Further generalization of the Wiener process yields the Ito process, where the constants *a* and *b* may depend on the values of *x* and *t*. The Ito process is of the form [9]:

$$dx = a(x,t)dt + b(x,t)dz \ .$$

2.2 Definition of Geometric Brownian Motion Process

The case of stock prices is slightly different from the generalized Brownian motion process. In the case of the Brownian motion process, a constant drift rate was assumed. However, in the case of stock prices, it is not the drift rate that is constant. For stock prices, the return on investment is assumed to be constant, where the rate of return at a given time is the ratio of the drift rate to the value of the stock at that time. Hence the constant expected drift-rate assumption in the case of Brownian motion process is inappropriate and needs to be replaced by an assumption of constant expected rate of return [9].

Let *Y* be the price of the stock at time *t* and assume the expected drift rate is μY for some constant μ. This means that in a short interval of time Δt , the expected increase in *Y* is $\mu Y\Delta t$. The constant parameter μ is the expected rate of return. If the volatility of the stock price is zero, then the model implies that $\Delta Y = \mu Y\Delta t$, and when the limit is taken as $\Delta t \to 0$, the expected stock price at time *T* finally becomes $E[Y_T] = Y_0 e^{\mu T}$, where Y_0 is the original value.

However, the stock prices do have volatility. Hence taking that into consideration, the above model can be written as

$$dY = \mu Ydt + \sigma Ydz, \text{ or } \frac{\Delta Y}{Y} = \mu\Delta t + \sigma\varepsilon\sqrt{\Delta t} \ .$$

The first term of the second equation above is the expected value of the return provided by the stock for a time period of Δt and the second term is the stochastic component of the return. Here σ is the volatility rate.

Taking limits as $\Delta t \to 0$, we have

$$E[Y_T] = Y_0 e^{(\mu + \frac{\sigma^2}{2})T} \ .$$

Geometric Brownian motion is useful in modeling stock prices over time when one believes that the percentage changes over equal length, non-overlapping intervals are independent and identically distributed. For

The Version of Record of this manuscript has been published and is available in Engineering Economist 2005, http://www.tandfonline.com/10.1080/00137910590949904. Posted with permission.

example, if Y_n is the price of the stock at time n = 0, 1, 2… then it is reasonable to suppose that the ratios Y_{n+1}/Y_n, $n \geq 1$, are independent and identically distributed [21]. Let $u_n = \frac{Y_{n+1}}{Y_n}$. After taking the log of both sides and rearranging, we have, $\ln(Y_{n+1}) = \ln(Y_n) + \ln(u_n)$. Now let $w(n) = \ln(u_n)$, that is $w(n) = \ln(Y_{n+1}) - \ln(Y_n)$.

If $w(n)$ for n ≥ 1 are independent and are identically distributed normal random variables with mean μ and variance σ^2, it can be said that the variable u_n will have a lognormal distribution [15]. The successive prices can be found to be [15] $Y_t = u_{t-1}u_{t-2}\cdots u_0 Y_0$. Taking the natural log of this equation, we have

$$\ln[Y_t] = \ln[Y_0] + \sum_{i=0}^{k-1} \ln[u_i] = \ln[Y_0] + \sum_{i=0}^{k-1} w(i) \quad .$$

The term *ln[Y_0]* is constant, and the *w(i)*'s are each normal random variables. Since the sum of normal variables is a normal random variable, it follows that *ln[Y_t]* is a normal random variable. Hence the stock price Y_t has a lognormal distribution, with

$$E\left[\ln\left(\frac{Y_t}{Y_0}\right)\right] = \mu t$$
$$Var\left(\ln\left(\frac{Y_t}{Y_0}\right)\right) = \sigma^2 t \ .$$

Thus, it can be seen that the ratio $\ln\left(\frac{Y_{k+t}}{Y_k}\right)$ has distribution approaching that of a normal random variable with mean μt and variance σ^2t.

The Geometric Brownian Motion process can formally be defined as follows [20]:

We say that the variable Y_k, 0 ≤ k < ∞, follows a GBM (with drift parameter μ and volatility parameter σ) if, for all nonnegative values of *k* and *t*, the random variable $\frac{Y_{k+t}}{Y_k}$ is independent of all values of the variable up to time *k* and if in addition, $\ln\left(\frac{Y_{k+t}}{Y_k}\right)$ has a normal distribution with mean μt and variance $\sigma^2 t$, independent of *k*, where μ and σ are constants.

2.3 Checking for GBM Process Fit

After any seasonal variation is removed from the data, the data can be tested for the GBM process. Referring to the analysis above, there are two assumptions to be satisfied [20]:

The Version of Record of this manuscript has been published and is available in Engineering Economist 2005, http://www.tandfonline.com/10.1080/00137910590949904. Posted with permission.

1. Normality of the log ratios (*w(k)*) with constant mean and variance
2. Independence from previous data (log ratios independent of their past values).

2.3.1 Normality:

The simplest (however not very accurate) way to check for normality is to plot a histogram of the log ratios and compare it to a normal distribution plot. Another graphical method of testing the normality assumption is to examine the **normal probability plot**. A normal probability plot, also known as a normal Q-Q plot or normal quantile-quantile plot, is the plot of the ordered data values against the associated quantiles of the normal distribution. For data from a normal distribution, the points of the plot should lie close to a straight line.

The statistical tests of normality can be conducted in many ways by using any of the goodness-of-fit tests on the *w(k)* values. One way is to run a chi-square test for goodness-of-fit. Another goodness-of-fit test is the **Shapiro-Wilk W** Test [24]. This is the test used in the statistical package JMP for $n \leq 2000$ [12]. In this test, the hypothesis set is:

H_0: The distribution is normal, against

H_a: The distribution is not normal.

The test gives the value of the statistic 'W' and the corresponding p-value. The p-value is compared to the specified level of significance α. If the observed p-value is greater than the level of significance the test statistic is not in the rejection region and the null hypothesis of a normal distribution cannot be rejected. Note that a large p-value does not definitively identify the data as normally distributed; it only means that the data could plausibly have been generated by a normal distribution.

2.3.2 Independence from the past data:

To test the serial independence of the *w(k)*, the **chi-square test on two-way tables** [3] can be used. The chi-square test provides a method for testing the association between the row and column variables in a two-way table. The null hypothesis is

H_0: There is no association between the variables (in other words, one variable does not vary according to the other variable), while the alternative hypothesis is

H_a: Some association does exist. (The alternative hypothesis does not specify the type of association; close attention to the data is required to interpret the information provided by the test.)

The chi-square test is based on a test statistic that measures the divergence of the observed data from the values that would be expected under the null hypothesis of no association.

To test serial independence of the *w(k)* values, the two variables in the chi-square test are *w(k)* and *w(k+1)* for each *k*. To carry out the test the log ratios are segregated into different groups (or intervals) depending on the number of data points and the range of data values. These groups or intervals of the log ratios are formed in such a way that number of observed values in each of the intervals is approximately equal. The two way table is formulated on the concept that the probability of *w(k)* being in state *j* (interval *j*) now after being in state *i* (interval *i*) in the last period is equal for all *j*. Equivalently, if a daily data series follows a GBM process, then tomorrow's state will not depend on today's state. One way to verify that is to see the proportion of time that an observation in state *i* is followed by a state *j* observation [20]. Thus the two way table is constructed with rows of state *i* and columns of state *j*. Under the GBM process model, tomorrow's change would be unaffected by today's change and so the theoretically expected percentages in the two-way table would be same for all rows. The expected value for each cell in a two-way table is equal to $\frac{(row\,total)*(col.total)}{n}$, where *n* is the total number of observations included in the table, row total is the total number of data points in state *i*, and column total is the total number of data points in state *j*. Once the expected values have been computed, the chi-square test statistic is computed as $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected frequency}}$ where the square of the differences between the observed and expected values in each cell, divided by the expected value, are added across all of the cells in the table.

The distribution of the statistic X^2 is chi-square with $(r\text{-}1)(c\text{-}1)$ degrees of freedom, where *r* represents the number of rows in the two-way table and *c* represents the number of columns. The p-value for the chi-square test is $P(\chi^2 \geq X^2)$, the probability of observing a value at least as extreme as the test statistic for a chi-square distribution with *(r-1)(c-1)* degrees of freedom. A small p-value indicates support for the alternative hypothesis; in our case suggesting that successive log ratios are not independent. Note once again that a p-value greater than the chosen level of significance does not positively confirm that the log ratios are serially independent, but it indicates that the data do not contradict that assumption.

III. Seasonality

As mentioned above, the data from various industries were considered for their fit to the GBM process. In some cases the time series exhibited trend and/or seasonal patterns. The usual assumption is that four separate

components – trend, cyclical, seasonal and irregular – combine to provide specific values for the time series data [1]. The GBM process can account for exponential trend via the drift term and irregularity in terms of the white noise process; however, it does not include cyclical or seasonal effects. In this paper, we neglect the cyclic variation and consider the component of the time series that represents the variability in the data due to seasonal influences. It is usual to consider the seasonal movement to be occurring annually, however it should be noted that the season could also be different from a year.

Two common models for decomposing a time series, which aim to isolate each component of the series as accurately as possible, are the additive model and the multiplicative model.

Suppose X_t is the time series value at period t, S_t is the seasonal index at period t, T_t is the trend-cycle component at period t, and E_t is the irregular component at period t,

The additive model has the form $X_t = S_t + T_t + E_t$. That is, the seasonal, trend, and irregular components are added together to give the observed series. In the additive model, the seasonal indices over the periods of a particular season add up to zero [4].

Alternatively, the multiplicative decomposition has the form $X_t = S_t T_t E_t$. Here, the seasonal, trend-cycle and irregular components are multiplied to give the observed series [16]. In multiplicative model, the average seasonal index for a season is unity [1].

An additive model is used if the magnitude of the seasonal fluctuations does not vary with the level of the series. However, if the seasonal fluctuation increases or decreases in the level of the series, then a multiplicative model is more appropriate. As seen from the data analysis, for the data series considered in this paper, the magnitude of seasonal variation increases with time (please refer to Figures 1 and 6 in Section IV). Hence a multiplicative model is used. Often the transformed series can be modeled additively, when the original data are not additive. Logarithms, in particular, turn a multiplicative relationship into an additive relationship [16], since

$$X_t = S_t T_t E_t \Rightarrow \ln[X_t] = \ln[S_t] + \ln[T_t] + \ln[E_t]$$

Suppose we have observations $X_1, X_2 \ldots X_T$ of a process; in particular, for our model, $X_t = Y_t S_t$ or

$$\ln[X_t] = \ln[Y_t] + \ln[S_t] \tag{1}$$

where Y_t are observations at discrete time points of a GBM process and S_t is the seasonal factor.

The observations of the process can also be recorded in terms of the seasons and periods. Let X_{ij} be the observation corresponding to the j th period of the i th season, where $i = 1 \ldots m$ and $j = 1 \ldots p$, that is, we have data

for m seasons, with each season having p periods in it and $T = mp$. Correspondingly, let Y_{ij} be the observation of the j th period of the i th season of a GBM process; and let S_{ij} be the seasonal index for period j of season i, where $S_{ij} = S_j$, for each i. Then, $X_{ij} = S_j Y_{ij}$. Hence Equation (1) can be written as:

$$\ln[X_{ij}] = \ln[Y_{ij}] + \ln[S_j] \tag{2}$$

where in Equation (1), $t = (i-1)p + j$.

In the following, we treat this as a usual additive model, the only difference being that we are using log values, instead of the actual values.

Our goal is to remove the seasonal effects from the time series. This process is referred to as deseasonalization [1]. The first step in deseasonalization is to estimate the values of the seasonal indices for each period. And once the estimates $\hat{S}_j$ of the seasonal variation for each period j are found out, the lognormal variable Y can be estimated using the equation

$$\ln[\hat{Y}_{ij}] = \ln[X_{ij}] - \ln[\hat{S}_j] \tag{3}$$

In the additive model, two estimation methods have been proposed. The analysis of both the methods with respect to the GBM model that is to be tested is included in the following sections. The two methods are compared on the basis of bias and the unbiased one selected. The first method uses moving averages of the consecutive data values [4], [16] and the second one uses averages of all the data values corresponding to each period of the season in turn [8]. In this paper we examine the estimates of the seasonal indices obtained from the series in Equation (1) or (2) using both the methods to see whether the estimates add up to zero and are free of bias.

3.1 Method I [3], [4], [16]

Here, we use the arithmetic centered moving average. The arithmetic moving average of ($2t$+1) data points centered at k is calculated by $\frac{(L_{k-t} + L_{k-t+1} + ... + L_k ... + L_{k+t})}{2t+1}$ where $\{L_i: i = 1, 2, 3\ldots\}$ is the series of data points. The moving averages isolate the seasonal components, which then can be estimated in the case of the additive model by subtracting the moving average from the corresponding data series value. The values thus found are estimates of the seasonal indices for those periods. Hence, first we make sure that these estimated seasonal indices for any season

add up to zero in case of additive model, that is $\sum_{j=1}^{p} E[\hat{S}_j] = 0$. Secondly, we prove that these calculated values are unbiased estimators of the actual seasonal index. That is, $E\left[\hat{S}_j\right] = S_j$.

Now, since Y_t follows a GBM process, continuing from the Section 2.2, from the properties of lognormal distribution [15], given Y_1, we have,

$$E\left[\ln Y_t\right] = E\left[\ln Y_1\right] + \left(\mu + \frac{\sigma^2}{2}\right)(t-1) = E\left[\ln Y_1\right] + \delta\,(t-1), \text{ where } \delta = \mu + \frac{\sigma^2}{2} .$$

So, if we let Y_{11} be the first value of the series,

$$E[\ln Y_{ij}] = E[\ln Y_{11}] + \delta(t-1), \text{ where } t = (i-1)p + j \tag{4}$$

Let P_{ij} represent the arithmetic moving average for the j th period of the i th season (year, for example). Let $\lceil x \rceil$ denote the smallest integer that is greater than or equal to x, and $\lfloor x \rfloor$ denote the largest integer that is less than or equal to x.

From [4], [16], for our model, the centered moving average P_{ij} when p is odd will be given by

$$P_{ij} = \frac{(\ln X_{t-\lfloor p/2 \rfloor} + \ldots \ln X_t + \ldots \ln X_{t+\lfloor p/2 \rfloor})}{p}, \text{where } t = (i-1)p + j .$$

And when the number of periods p is even, the centered moving average is calculated as [16]:

$$P_{ij} = \frac{1}{p}\left(0.5\ln X_{t-\frac{p}{2}} + \ln X_{t-\frac{p}{2}+1} + \ldots + \ln X_{t+\frac{p}{2}-1} + 0.5\ln X_{t+\frac{p}{2}}\right), \text{where } t = (i-1)p + j . \tag{5}$$

After calculating the values of centered moving average, we compute the deviation, $\ln \hat{S}_{ij} = \ln X_{ij} - P_{ij}$, to estimate the log of the seasonal index for the period j based on season i. The log of the estimated seasonal index for a period is calculated as a simple arithmetic average of log of all the seasonal indices for that particular period from all the seasons. That is,

$$E\left[\ln \hat{S}_j\right] = \frac{1}{m}\sum_{i=1}^{m} E\left[\ln \hat{S}_{ij}\right] . \tag{6}$$

In particular, for an odd number of periods p,

$$\ln \hat{S}_j = \frac{1}{m}\sum_{i=1}^{m} \ln \hat{S}_{ij}, \text{ for } j = \lceil p/2 \rceil$$

$$\ln \hat{S}_j = \frac{1}{m-1}\sum_{i=2}^{m} \ln \hat{S}_{ij}, \text{ for } j = 1 \text{ to} \lfloor p/2 \rfloor$$

$$\ln \hat{S}_j = \frac{1}{m-1}\sum_{i=1}^{m-1} \ln \hat{S}_{ij}, \text{ for } j = \lceil p/2 \rceil + 1 \text{ to } p$$

When p is even, the corresponding equations are given as;

$$\ln \hat{S}_j = \frac{1}{m-1}\sum_{i=2}^{m} \ln \hat{S}_{ij} \text{ for } j = 1 \text{ to } \frac{p}{2}$$

$$\ln \hat{S}_j = \frac{1}{m-1}\sum_{i=1}^{m-1} \ln \hat{S}_{ij} \text{ for } j = (\frac{p}{2} + 1) \text{ to } p$$

Lemma 1: For the model in Equation (2), using Method I, the estimates of the logs of the seasonal indices add to 0. That is, $\sum_{j=1}^{p} E[\ln \hat{S}_j] = 0$.

Proof: See Appendix A.

Lemma 2: For Method I, the expected value of the log of the estimate of the seasonal index for a particular period is equal to the log of the seasonal index for the period. That is, $E[\ln \hat{S}_j] = \ln S_j$.

Proof: See Appendix B.

Theorem 1: The expected value of the log of a variable following the GBM process for a particular period, obtained from subtracting the corresponding expected log of the seasonal factor from the log of the observation, is an unbiased estimator of the actual log of that variable. That is $E[\ln \hat{Y}_{ij}] = E[\ln Y_{ij}]$.

Proof: From Equation (3), $\ln[\hat{Y}_{ij}] = \ln[X_{ij}] - \ln[\hat{S}_j]$, and taking expectation of both sides,

$$E\left[\ln \hat{Y}_{ij}\right] = E\left[\ln X_{ij}\right] - E\left[\ln \hat{S}_j\right]$$

However, from Lemma 2, we have $E[\ln \hat{S}_j] = \ln S_j$.

The above equation becomes, $E\left[\ln \hat{Y}_{ij}\right] = E\left[\ln X_{ij}\right] - \ln S_j$. Hence from Equation (2) we can see that, $E[\ln \hat{Y}_{ij}] = E[\ln Y_{ij}]$.

3.2 Method II [8]

In the previous method, the moving average was used. Here the simple arithmetic average of the log values across seasons [8] is examined as an alternative method of deseasonalization. Let P_j denote the average value for the j th period, that is, the average of all the period j values over all m seasons.

$$P_j = \frac{1}{m}\sum_{i=1}^{m} \ln\left[X_{ij} \right] = \frac{1}{m}\sum_{i=1}^{m} \ln\left[Y_{ij} \right] + \frac{1}{m}\sum_{i=1}^{m} \ln\left[S_{ij} \right].$$

The overall average for the season is the average of all the periods of the season, $\overline{P} = \frac{1}{p}\sum_{k=1}^{p} P_k$.

Now the seasonal indices for each period can be calculated by the equation $\ln\left[\hat{S}_j \right] = P_j - \overline{P},\ \forall\, j$.

Using this method, the sum of the estimated logs of the seasonal indices of all the periods of the season is zero. That is, $\sum_{j=1}^{p} E[\ln \hat{S}_j] = 0$. However, the expected value of the ratio of the X variable and the estimated seasonal index for the particular period is not the actual Y variable for the period, as obtained in the previous method. This method can be shown to overestimate the Y_{ij} values by the factor $e^{(\frac{p-(2j-1)}{2})\delta}$ (Note that this factor is greater than 1 for $j < \frac{p+1}{2}$ and less than 1 for $j > \frac{p+1}{2}$; see Figure 2).

We conclude that the method of using moving average (with additive model of log of parameters) is better than the one using simple average. Hence we use Method I to analyze the numerical data.

IV. Data Analysis

The purpose of fitting a model to historical data is to help predict the future, assuming that past and current trends will continue. In trying to fit and forecast demand for services, two difficulties immediately arise. First, the demand will depend on price to varying extents depending on the level of necessity of service and the availability of alternatives for meeting the same need. Secondly, without extensive consumer surveys, the only way to measure past demand is by actual usage, which was limited by the available supply of the service. As a surrogate for the actual demand data, we collected usage data for publicly available sources in the energy, transportation and telecommunication sectors, the analysis for which is given in the succeeding sections.

4.1 Electric Power Consumption

The data were collected from the U.S. Department of Energy's Office of Scientific and Technical Information, which provides access to energy, science, and technology research and development information [7]. The data represent the total monthly sales by electric utilities to all the sectors (namely, residential, commercial, industrial and others). The monthly consumption (in million kilo-watt-hours) for electric power was recorded for each month for 8 years (from 1993 to 2002). Hence the total of 120 data points were used for the analysis of the electric power consumption.

First, the seasonal variation was removed from the data. For this the two methods described in Section III were tested. The results are shown in Table 1, which gives the value for the seasonal index for each month using each of the methods.

Table 1 here

Figure 1 here

The difference between the two methods of evaluating of seasonal variation is seen in Table 1 and Figure 1. For the detailed difference, Figure 2 compares the values before and after deseasonalization by both methods for a representative year, 1998. The deseasonalized values obtained by Method I are seen to have more of an upward trend over the year.

Figure 2 here

The deseasonalized data obtained from Method I were analyzed to check the normality of the log ratios and also their independence.

Even before the normality test, as a visual check for the independence of the log ratios, we observe a scatter plot of log ratios in Figure 3. As there is no apparent pattern to the $w(k)$ values for the data points, we may tentatively say that the $w(k)$ values are independent, which will be examined analytically in the chi-square test of independence. The plot also indicates the plausibility of a constant mean and variance of the $w(k)$ values.

Figure 3 here

Figure 4 shows the histogram and normal probability plot of the log ratios with fitted mean and variance. Since the Shapiro-Wilk test statistic is 0.9844 and the corresponding p value is 0.768, we fail to reject the null hypothesis that the distribution of the log ratios is normal. Hence we can conclude that the data are consistent with the lognormal aspect of GBM.

Figure 4 here

The remaining key characteristic of the GBM process is independent increments. Figure 5 plots the deseasonalized log ratios for years 1994, 1997, 1999, and 2001. The lack of any visible pattern in values for any given year indicates the independence of the successive ratios.

Figure 5 here

Next the independence of the log ratios is checked using a two-way chi-square test. The *w(k)* values were divided into 4 categories as shown in Table 2 and the two way table chi-square test resulted in a p-value for the test of 0.319. The null hypothesis that the variables are independent cannot be rejected.

Table 2 here

Hence we conclude that overall the data are consistent with the periodic observations from a GBM process. The mean log ratio was 0.0025 with a standard deviation of 0.02, indicating the mean growth rate of 3% per annum.

The importance of removing the seasonal variation prior to checking for the normality and independence is stressed from the fact that, for the original time series (before the deseasonalization) the normality test for the log ratios failed (with p-value 0.0004, rejecting the null hypothesis that the distribution for log ratios is normal); also these log ratios were not found to be independent. In fact, the two-way chi-square test on these log ratios gave a p-value of 0.001, indicating that we reject the null hypothesis of independence of the variables. The same fact could also be observed from the scatter plot of log ratios with respect to the prior values (see Figure 6). If the log ratios had been independent, the points of the scatter plot would not have had any trend.

Figure 6 here

4.2 Airline Passenger Enplanement

We collected the historical monthly data on U.S. Revenue Passenger Enplanements for the years 1981 through 2001 from the U.S. Aeronautical Board [26]. Revenue Passenger Enplanement can be defined as the number of paying passengers boarding a flight, including origination, stopovers and connections. It should be noted that each connecting flight between origination point and destination counts as one enplanement.

While analyzing the passenger data, a seasonal trend was observed for which the moving average (Method I) was applied to deseasonalize the log ratios. The final seasonal indices were as given in Table 3.

Table 3 here

The variation in the data values with respect to time is given in Figure 7. It can be seen that as the time increases, the amount of seasonal variation increases (observing the original data); motivating the use of the multiplicative model described in Section 3.1.

Figure 7 here

Figure 8 plots the corresponding log ratios over time. From the plot, it can be seen that there is no visible pattern in the values of log ratios, which indicates their distribution is stationary, and suggests serial independence.

Figure 8 here

The histogram and normal probability plot for the normality test for the passenger data are given in Figure 9. As the p-value of the Shapiro-Wilk test is 0.4416 (greater than 0.05), we cannot reject the hypothesis that the log ratios are normally distributed.

Figure 9 here

Again, as with the electric utility data, the random nature of the deseasonalized *w(k)* values can be visually inspected by the graphs of *w(k)* values given in Figure 10. Here, the changes in *w(k)* values appear to be independent over time, as seen from the randomness of the *w(k)* values for various years. Hence it can be tentatively concluded that the *w(k)* values are independent of each other.

Figure 10 here

To more rigorously test independence by the chi-square test, four intervals of *w(k)* values were selected as shown in Table 4.

Table 4 here

The p-value for the test was found to be 0.058, so we cannot reject the null hypothesis that the *w(k)* values are independent at a 5% significance level.

Once again, as done in the electric demand case, the importance of removing the seasonality factors before checking normality and independence of log ratios is confirmed by performing similar tests with the original log ratios (obtained from the time series without deseasonalization). The normality of the log ratios could not be confirmed (the p-value of the normality test is 0.0001); also the chi-square test gives a p-value, which is very close to zero, forcing the rejection of null hypothesis of the independence test. Hence prior to the deseasonalization, the log ratios are not independent. The same fact could be observed by inspecting the scatter plot of these log ratios with respect to the prior values (Figure 11), which indicates clear trend in the values.

Figure 11 here

Thus, we can conclude that the lognormal ratios after deseasonalization are independent; however, a higher significance level could lead to the opposite conclusion. Hence, the independence test is not as convincing as for the electric power data. The mean log ratio was found to be 0.00271 with a standard deviation of 0.029, and hence the average growth rate was calculated to be 3.3% per annum.

4.3 Cell Phone Revenue

Usage of mobile phone service might be measured by minutes of usage, total connections made, or even the number of handsets sold. Because of the lack of available data on these quantities, the revenue collected from the cellular phone subscribers was analyzed for the period of January 1985 to June 2002, with data collected every 6 months [5].

First of all, the plot given in Figure 12 of log ratios over time was observed. It is seen that there is a decreasing trend in both the mean and the variance of log ratios. Hence visual inspection reveals that the *w(k)* variable may be neither stationary nor independent. Note that, since revenue is the product of sales volume and price, the downward trend could be attributed to price drops rather than flattening growth in demand.

Figure 12 here

The normality test, which includes the histogram of the log ratios and the normal probability plot, is given in Figure 13. From the Shapiro-Wilk test, the p-value is 0.0003, proving that the log ratios are not normally distributed.

Figure 13 here

The result could be influenced by the fact that the number of data points available was only 35. However, the Chi-square test did show independence of the *w(k)* values. The p-value for the independence test is 0.3735. Hence the null hypothesis that the log ratios are independent cannot be rejected. For this independence test the intervals of *w(k)* values used are given in Table 5.

Table 5 here

4.4 Internet Hosts

Internet growth can be measured by changes in either the number of users or number of hosts connected to the network. A host used to be a single machine on the net. However, the definition of a host has changed in recent years due to virtual hosting, where a single machine acts like multiple systems (and has multiple domain names and IP addresses) [11]. Typically, multiple users are connected to a host and the hosts are connected to the network. Since there is no central mechanism for tracking the number of users connected to the network [19], we use number of hosts as a measure of Internet size. In an attempt to gauge the growth of the Internet over the years, The Internet Software Consortium [11] conducted a survey called 'The Domain Survey' and measured the number of hosts. This survey was used in conjunction with the data in [19] to obtain a time series of the number of Internet hosts from

1982 to 2003 with data points recorded every six months. As before, *w(k)* values for the data are calculated and tested for normality and independence.

Figure 14 indicates the values of log ratios over time. It is seen again that the values do not appear to be random. There is visible downward trend in the values of *w(k)*, indicating that the values may not be stationary. One can also observe possible cyclical behavior.

Figure 14 here

The plots for the test of normality are given in Figure 15. The p-value for the Shapiro-Wilk test of normality is less than 0.001; hence the null hypothesis that the log ratios are normal is rejected.

Figure 15 here

To test the independence of *w(k)* values, the Chi-square test cannot be used as before, because the number of data points is too small to create cells such that each holds a positive number of observed values as required by the chi-square test. Hence the *w(k)* scatter plot is analyzed visually to determine the independence of *w(k)*.

From the plot in Figure 16, it can be seen that the *w(k)* values are not random, but rather large (small) log ratios tend to be immediately followed by other large (small) values. Hence we can say that the *w(k)* values are not independent of each other.

Figure 16 here

4.5 Summary of Results

The results of the data analysis for different industries are summarized in Table 6.

Table 6 here

Hence it can be seen that data related to service consumption from different sectors of industry may or may not meet the criteria for the GBM process. Among the services examined, the ones that fail one test or another are in newer industries that perhaps can still be considered emergent. Data on the usage of these services are also less direct and more difficult to obtain. The older and more established services of electric power and airline travel exhibit a better fit to the GBM assumption after deseasonalization. Having ascertained the model's fit to the deseasonalized data, a forecast of future demand can be obtained from the GBM model with the fitted parameters by re-inserting the seasonal factors. How the seasonal patterns would affect decision-making depends on the application, for example, capacity decisions typically consider the peak demand in a season.

We caution that, even when a model appears to closely fit historical data, extrapolation into the future does not carry any guarantee of accuracy. In 1995, logistic growth models showed a very good fit to historical data on

the number of cell phone subscribers [27]. Extrapolation suggested that the number of U.S. subscribers would level out close to 80 million early in the 21st century. As of December 2004, however, the Cellular Telecommunications and Internet Association reported over 173 million current U.S. wireless subscribers [5]. The fit of a model to historical data is a necessary but not sufficient condition for the credibility of its forecasts.

V. Conclusion

From the theory of the Brownian motion discussed in the paper and the subsequent analysis, it can be concluded that the structure for the analysis to check whether a particular time series data follows a Geometric Brownian motion process or not can be applied to varied data types. The result may be different for different data types; for some of the data sets, the GBM process may be appropriate, based on the criteria of normality and independence (for example, electric utility data and passenger data); however for some of the data sets, the assumption of GBM process distribution may not be appropriate (example, cell-phone revenue data and Internet host data). Hence in any given model, caution should be taken before assuming that the particular data set follows the GBM process. It was observed during the analysis of Cellular phone data and the Internet host data that the number of data points may affect the analysis results. Hence attempts need to be made to collect more data points for the given example type.

For cell phone revenue data and Internet hosts' data, it was observed that the log ratios decrease over time. It could be possible that the drift for those time series is dependent on time and the level of the time series. Hence the criteria for the GBM (with assumption of constant drift and volatility) were not being followed in these cases. For these data not following the GBM process, the data can be analyzed for other stochastic diffusion processes [6]. Also to incorporate the dynamic nature of drift (and possibly volatility) parameter, the Ito process for the stock prices can be used. More generalized models can also be studied. In [10], authors discuss some of the extended one-state-variable interest-rate models that involve time dependent parameters. The data for the cell phone revenue and Internet hosts might be analyzed using models similar to the ones given in that paper.

References

1. Anderson, D. R.; Sweeney, D. J.; Williams, T. A. (1994). *An Introduction to Management Science*, 7th ed. West Publishing Company, St. Paul, Minn.
2. Benninga, S.; Tolkowsky, E. (2002). "Real Options- An Introduction and An Application to R&D Evaluation". *The Engineering Economist*, 47(2). 151-168.
3. Blair, M. M. (1952). *Elementary Statistics, with General Applications*, Henry Holt and Company, New York.
4. Brockwell, P. J.; Davis, R. A. (2002). *Introduction to Time Series and Forecasting*, Springer Texts in Statistics, New York.
5. Cellular Telecommunications and Internet Association, *Wireless Industry Survey* (Viewed December 2004) http://www.ctia.org
6. Dixit, A. K.; Pindyck, R. S. (1993). *Investment under Uncertainty*, Princeton University Press. Princeton, N.J.
7. Energy Information Administration. *Electric Power Monthly*. (Viewed December 2003) http://www.osti.gov/servlets/purl/212513-0AAD20/webviewable/212513.pdf (for data from 1993 to 1995) http://www.osti.gov/servlets/purl/584882-h5Z86P/webviewable/584882.pdf (for data from 1995 to 1997) http://www.eia.doe.gov/cneaf/electricity/epm/epmt44p1.html (for data from 1998 to 2000)
8. Hillier, F. S.; Lieberman, G. J. (2001), *Introduction to Operations Research*, 7th ed. McGraw Hill. Boston.
9. Hull, J. C. (2000). *Options, Futures, and Other Derivatives*, 4th ed. Prentice Hall, NJ.
10. Hull, J. C.; White A. (1990). "Pricing Interest-Rate-Derivatives Securities". *The Review of Financial Studies*. 3(4). 573-592.
11. Internet Software Consortium, *Internet Domain Survey* (Viewed December 2003) http://www.isc.org/ds/host-count-history.html
12. JMP (2003). Statistical Discovery Software; http://www.jmp.com/index.html
13. Karsak, E. E.; Ozogul, O. C. (2002). "An Options Approach to Valuing Expansion Flexibility In Flexible Manufacturing Systems Investment". *The Engineering Economist*, 47(2). 169-193.
14. Lieberman, B. M. (1989). "Capacity Utilization: Theoretical Models and Empirical Tests." *European Journal of Operational Research* 40, 155-168.
15. Luenberger, D. (1995). *Investment Science*, Oxford University Press, New York.

16. Makridakis S.; Wheelwright S. C.; Hyndman R. J. (1998). *Forecasting: Methods and Applications*, 3rd ed., John Wiley and Sons Inc. New York.

17. Nembhard, H. B.; Shi, L.; Aktan, M. (2002). "A Real Options design for Quality Control Charts." *The Engineering Economist*, 47(1). 28-50.

18. Nembhard, H. B.; Shi L.; Aktan M. (2003). "A Real Options design for Product Outsourcing." *The Engineering Economist*, 48(3). 199-217.

19. Rai, A.; Ravichandran, T.; Samaddar, S. (1998). "How to anticipate the Internet's global diffusion." *Communications to the ACM*, 41, 97-106

20. Ross, S. (1999). *An Introduction to Mathematical Finance*, Cambridge University Press, Cambridge, U.K., New York.

21. Ross, S. (2000). *Introduction to Probability Models*, 7th ed., Harcourt Academic press, New York.

22. Ryan, S. M. (2004). "Capacity Expansion for Random Exponential Demand Growth with Lead Times." *Management Science*, 50(6). 740-748

23. Samuelson, P. A. (1965). "Proof that Properly Anticipated Prices Fluctuate Randomly," *Industrial Management Review*, 7, 41-49.

24. Shapiro S. S.; Wilk M. B., (1965). "An Analysis of Variance Test for Normality." *Biometrika*, 52, 3/4, 591-611.

25. Thorsen, B. J. (1998). "Afforestation as a Real Option: Some Policy Implications." *Forest Science*, 45(2). 171-178.

26. U.S. Aeronautical Board, *Origin and Destination Survey of Airline Passenger Traffic* (Viewed December 2003) http://www.bts.gov/oai/indicators/airtraffic/annual/1981-2001.html

27. Wang, M.; Kettinger, W. J. (1995). "Projecting the Growth of Cellular Communication." *Communications of the ACM*, 38, No. 10. 119-122

28. Whitt, W. (1981). "The Stationary Distribution of a Stochastic Clearing Process." *Operations Research* 29(2), 294-308.

Appendix A: Proof of Lemma 1

We consider the case where the number of seasons m is even and number of periods p is odd.

$$\sum_{j=1}^{p} E\left[\ln \widehat{S}_j\right] = E\left[\ln \widehat{S}_{\lceil p/2 \rceil}\right] + \sum_{j=1}^{\lfloor p/2 \rfloor} E\left[\ln \widehat{S}_j\right] + \sum_{j=1+\lceil p/2 \rceil}^{p} E\left[\ln \widehat{S}_j\right]$$

$$= E\left[\frac{1}{m}\sum_{i=1}^{m} \ln \widehat{S}_{i\lceil p/2 \rceil}\right] + \sum_{j=1}^{\lfloor p/2 \rfloor} E\left[\frac{1}{m-1}\sum_{i=2}^{m} \ln \widehat{S}_{ij}\right] + \sum_{j=1+\lceil p/2 \rceil}^{p} E\left[\frac{1}{m-1}\sum_{i=1}^{m-1} \ln \widehat{S}_{ij}\right]$$

$$= \sum_{i=1}^{m} E\left[\frac{1}{m} \ln \widehat{S}_{i\lceil p/2 \rceil}\right] + \sum_{j=1}^{\lfloor p/2 \rfloor} \sum_{i=2}^{m} E\left[\frac{1}{m-1} \ln \widehat{S}_{ij}\right] + \sum_{j=1+\lceil p/2 \rceil}^{p} \sum_{i=1}^{m-1} E\left[\frac{1}{m-1} \ln \widehat{S}_{ij}\right]$$

$$\sum_{j=1}^{p} E\left[\ln \widehat{S}_j\right] = \sum_{i=1}^{m} E\left[\frac{1}{m}\left(\ln X_{i\lceil p/2 \rceil} - P_{i\lceil p/2 \rceil}\right)\right] + \sum_{j=1}^{\lfloor p/2 \rfloor} \sum_{i=2}^{m} E\left[\frac{1}{m-1}\left(\ln X_{ij} - P_{ij}\right)\right] + \sum_{j=1+\lceil p/2 \rceil}^{p} \sum_{i=1}^{m-1} E\left[\frac{1}{m-1}\left(\ln X_{ij} - P_{ij}\right)\right],$$

where P_{ij} is the arithmetic moving average, as defined earlier.

Substituting values of P_{ij}, we have that

$$\sum_{j=1}^{p} E\left[\ln \widehat{S}_j\right] = \sum_{j=1}^{\lfloor p/2 \rfloor} \left(-\frac{1}{mp} - \frac{(j-1)}{p(m-1)}\right) E\left[\ln X_{1j}\right] + E\left[\ln X_{1\lceil p/2 \rceil} + \ln X_{m\lceil p/2 \rceil}\right] \left(\frac{p-1}{pm} - \frac{\lfloor p/2 \rfloor}{p(m-1)}\right) + \sum_{j=\lceil p/2 \rceil+1}^{p} \left(\frac{p-1}{p(m-1)} - \frac{1}{mp} - \frac{j-2}{p(m-1)}\right) E\left[\ln X_{1j}\right]$$

$$+ \sum_{\substack{j=1 \\ j\neq\lceil p/2 \rceil}}^{p} \sum_{i=2}^{m-1} E\left[\ln X_{ij}\right] \left(\frac{1}{p(m-1)} - \frac{1}{mp}\right) + \sum_{i=2}^{m-1} E\left[\ln X_{i\lceil p/2 \rceil}\right] \left(\frac{p-1}{mp} - \frac{p-1}{p(m-1)}\right) + \sum_{j=\lceil p/2 \rceil}^{p} \left(-\frac{1}{mp} - \frac{(j-1)}{p(m-1)}\right) E\left[\ln X_{mj}\right]$$

$$+ \sum_{j=1}^{\lfloor p/2 \rfloor} \left(\frac{p-1}{p(m-1)} - \frac{1}{mp} - \frac{j-2}{p(m-1)}\right) E\left[\ln X_{mj}\right]$$

Now we have from Equation (2), $\ln[X_{ij}] = \ln[Y_{ij}] + \ln[S_j]$. Substituting this for each of the X_{ij}, we cancel out the seasonal indices S_j from the above equation. To evaluate the Y_{ij} terms, we use Equation (4) and write all the Y_{ij} in terms of Y_{11} and solve the equation. We get

$$\sum_{j=1}^{p} E[\ln \widehat{S}_j] = 0$$

For the case where the number of season m is odd, and the periods p is also odd, the condition can be found out as:

$$\sum_{j=1}^{p} E\left[\ln \widehat{S}_j\right] = \sum_{j=1}^{\lfloor p/2 \rfloor}\left(-\frac{1}{mp} - \frac{(j-1)}{p(m-1)}\right)E\left[\ln X_{1j}\right] + E\left[\ln X_{1\lceil p/2 \rceil} + \ln X_{m\lceil p/2 \rceil}\right]\left(\frac{p-1}{pm} - \frac{\lfloor p/2 \rfloor}{p(m-1)}\right)$$

$$+ \sum_{j=\lceil p/2 \rceil+1}^{p}\left(\frac{p-1}{p(m-1)} - \frac{1}{mp} - \frac{j-2}{p(m-1)}\right)E\left[\ln X_{1j}\right]$$

$$+ \sum_{j=1}^{p}\sum_{i=2}^{m-1} E\left[\ln X_{ij}\right]\left(\frac{1}{p(m-1)} - \frac{1}{mp}\right) + \sum_{j=\lceil p/2 \rceil}^{p}\left(-\frac{1}{mp} - \frac{(j-1)}{p(m-1)}\right)E\left[\ln X_{mj}\right]$$

$$+ \sum_{j=1}^{\lfloor p/2 \rfloor}\left(\frac{p-1}{p(m-1)} - \frac{1}{mp} - \frac{j-2}{p(m-1)}\right)E\left[\ln X_{mj}\right]$$

The case when the number of periods p is even is also similar to the one formulated above, using Equation (5), which had to be centered because of the even number of periods [5].

Hence, the sum of estimated log of seasonal indices for a season is:

$$E[\ln \widehat{S}_j] = \sum_{j=1}^{p}\left(\frac{u_j}{m-1} - \frac{2j-1}{2(m-1)p}\right)E\left[\ln X_{1j}\right] + \sum_{j=1}^{m}\left(\frac{w_j}{m-1} - \frac{1}{2(m-1)p} - \frac{p-j}{(m-1)p}\right)E\left[\ln X_{mj}\right],$$

$$\textit{where} \quad u_j = \begin{cases} 0 & \textit{for } j \le p/2 \\ 1 & \textit{for } j > p/2 \end{cases}$$

$$\text{And} \quad w_j = \begin{cases} 1 & \textit{for } j \le p/2 \\ 0 & \textit{for } j > p/2 \end{cases}$$

which, when we use Equation (2) for $\ln[X_{ij}]$ and subsequently Equation (4) for $\ln[Y_{ij}]$ as before, comes to zero.

Appendix B: Proof of Lemma 2

To show $E\left[\ln \widehat{S}_j\right] = \ln S_j$, for any j

We have,

$$E\left[\ln \widehat{S}_{ij}\right] = E\left[\ln X_{ij} - P_{ij}\right]$$

$$= E\left[\ln X_{ij} - \frac{1}{p}\sum_{k=t-\lfloor p/2 \rfloor}^{t+\lfloor p/2 \rfloor} \ln X_k\right], \text{ where } t = (i-1)p + j, \text{ therefore}$$

$$E\left[\ln \widehat{S}_{ij}\right] = E\left[\frac{p-1}{p}\ln X_{ij} - \frac{1}{p}\sum_{\substack{k=t-\lfloor p/2 \rfloor \\ k\neq t}}^{t+\lfloor p/2 \rfloor} \ln X_k\right]$$

Since $\ln X_{ij} = \ln S_{ij} + \ln Y_{ij}$, we have

$$E\left[\ln \widehat{S}_{ij}\right] = E\left[\frac{p-1}{p}\ln S_{ij} - \frac{1}{p}\sum_{\substack{k=t-\lfloor p/2 \rfloor \\ k\neq t}}^{t+\lfloor p/2 \rfloor} \ln S_k\right] + E\left[\frac{p-1}{p}\ln Y_{ij} - \frac{1}{p}\sum_{\substack{k=t-\lfloor p/2 \rfloor \\ k\neq t}}^{t+\lfloor p/2 \rfloor} \ln Y_k\right],$$

We know that $\sum_{j=1}^{p} \ln S_{ij} = 0$, hence $\sum_{\substack{k=t-\lfloor p/2 \rfloor \\ k\neq t}}^{t+\lfloor p/2 \rfloor} \ln S_k = -\ln S_{ij}$

$$E\left[\ln \widehat{S}_{ij}\right] = E\left[\frac{p-1}{p}\ln S_{ij} - \frac{1}{p}\left(-\ln S_{ij}\right)\right] + E\left[\frac{p-1}{p}\ln Y_{ij} - \frac{1}{p}\sum_{\substack{k=t-\lfloor p/2 \rfloor \\ k\neq t}}^{t+\lfloor p/2 \rfloor} \ln Y_k\right]$$

$$= E\left[\frac{p-1}{p}\ln S_{ij} - \frac{1}{p}\left(-\ln S_{ij}\right)\right] + \frac{p-1}{p}E\left[\ln Y_{ij}\right] - \frac{1}{p}\sum_{\substack{k=t-\lfloor p/2 \rfloor \\ k\neq t}}^{t+\lfloor p/2 \rfloor} E\left[\ln Y_k\right]$$

$$= \ln S_{ij} + \frac{p-1}{p}\left\{E\left[\ln Y_{11}\right] + \delta\left(t-1\right)\right\} - \frac{1}{p}\sum_{\substack{k=t-\lfloor p/2 \rfloor \\ k\neq t}}^{t+\lfloor p/2 \rfloor} E\left[\ln Y_{11}\right] + \delta\left(k-1\right), \text{where } t = \left(i-1\right)p + j$$

which gives,

$$E\left[\ln \widehat{S}_{ij}\right] = \ln S_{ij} + 0 = \ln S_{ij} \ .$$

From the above equation:

$$\frac{1}{m}\sum_{i=1}^{m} E\left[\ln \widehat{S}_{ij}\right] = \frac{1}{m}\sum_{i=1}^{m} \ln S_{ij}$$

The Version of Record of this manuscript has been published and is available in Engineering Economist 2005, http://www.tandfonline.com/10.1080/00137910590949904. Posted with permission.

From Equation (6),

$$E\left[\ln \widehat{S}_j\right] = \frac{1}{m}\sum_{i=1}^{m} \ln S_{ij} = \frac{1}{m}\left[m \cdot (\ln S_j)\right]$$

because we know that, $\ln S_{ij} = \ln S_j, \ \forall j$

Hence $E\left[\ln \widehat{S}_j\right] = \ln S_j$.

In the case where the number of periods *p* is even, the calculations are similar except for the fact that again Equation (4) for the centered moving average is used instead of the simple moving average. Hence, the values obtained after substituting respective values in the equation $E\left[\ln \widehat{S}_{ij}\right] = E\left[\ln X_{ij} - P_{ij}\right]$ change accordingly, however the concept is similar; and it gives a similar result.

List of Figures

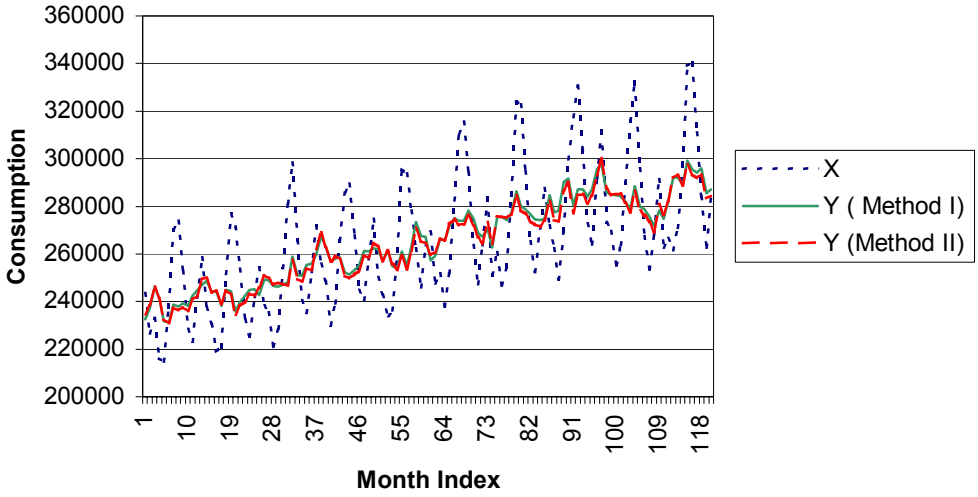

Figure 1. Original and deseasonalized demand data obtained from two different methods (* units: million kWh)

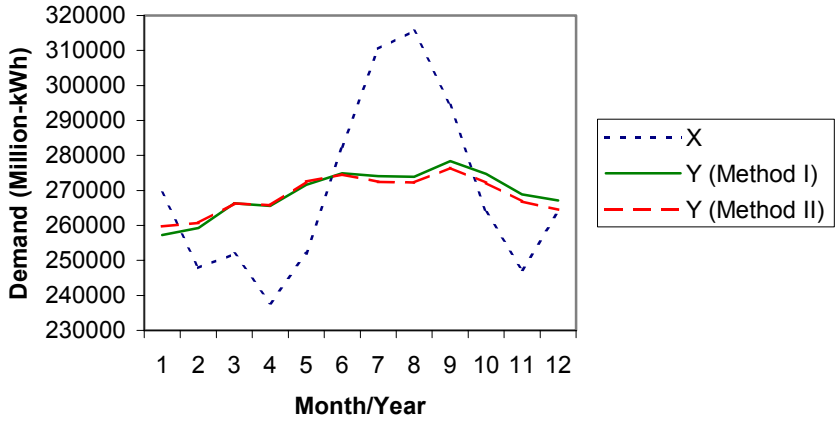

Figure 2. Original and deseasonalized data (obtained from each of the methods) for each month of the year 1998

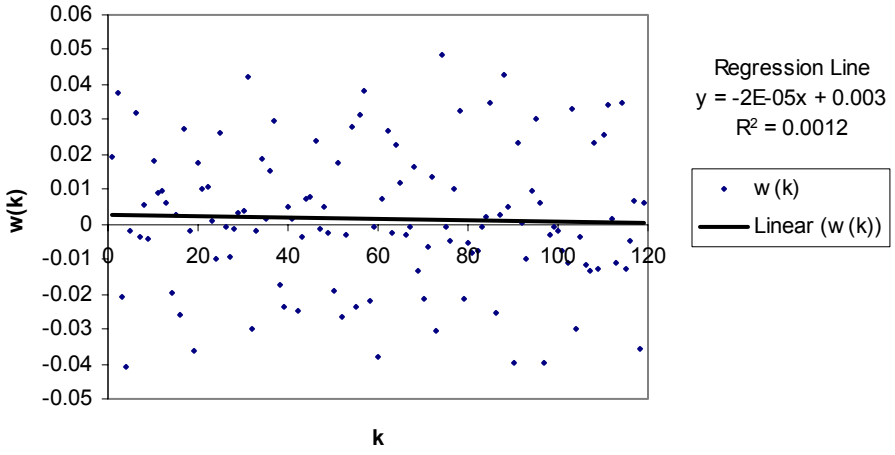

Figure 3. Log ratios (*w(k)*) for all 120 data points indicating the randomness of the *w(k)* values

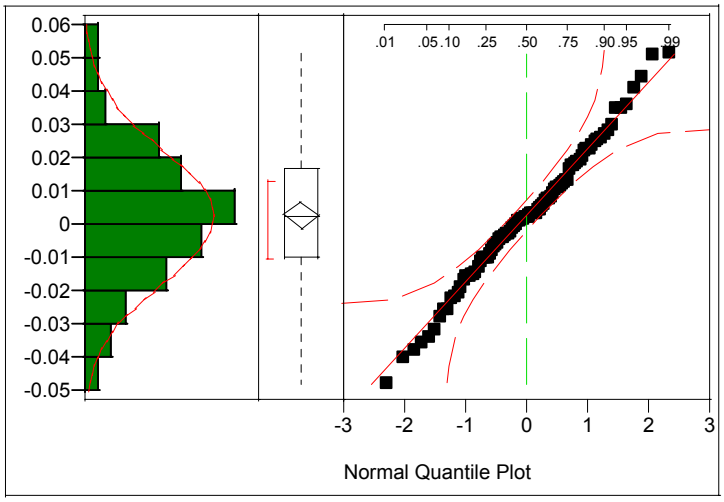

Figure 4. Distributions for the *w(k)* values obtained from the moving average method.

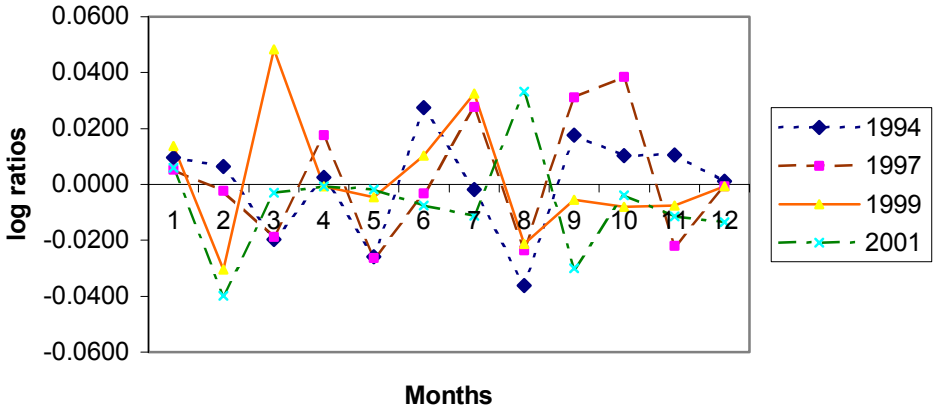

Figure 5. Log ratios for each month of year plotted for 4 representative years indicating randomness

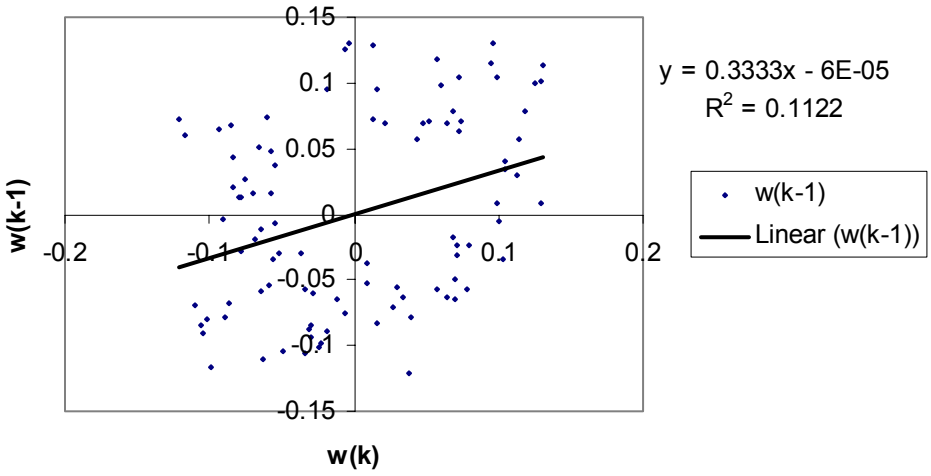

Figure 6. Scatter plot of *w(k)* values for electric demand before deseasonalization

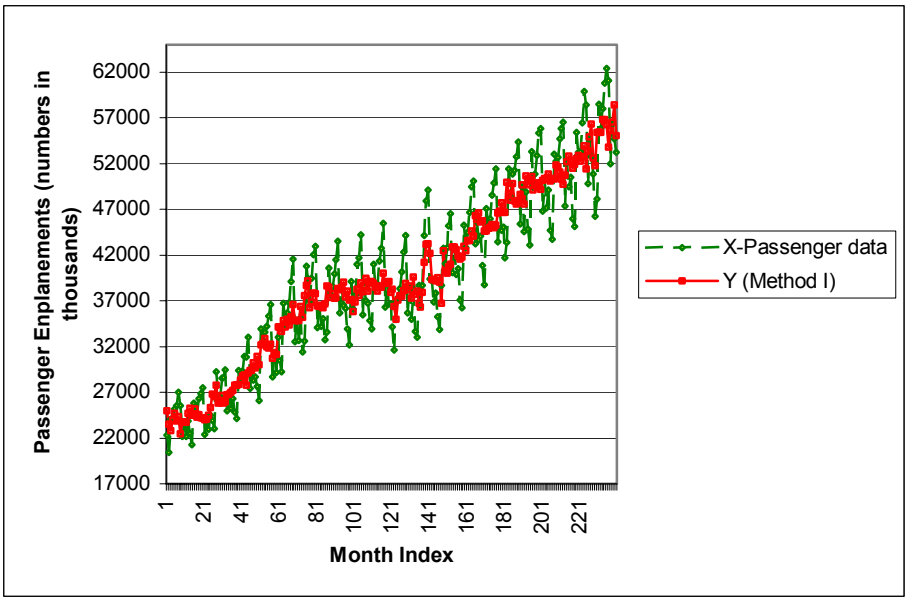

Figure 7. Original and deseasonalized enplanement data obtained from deseasonalization Method I of Section 3.1

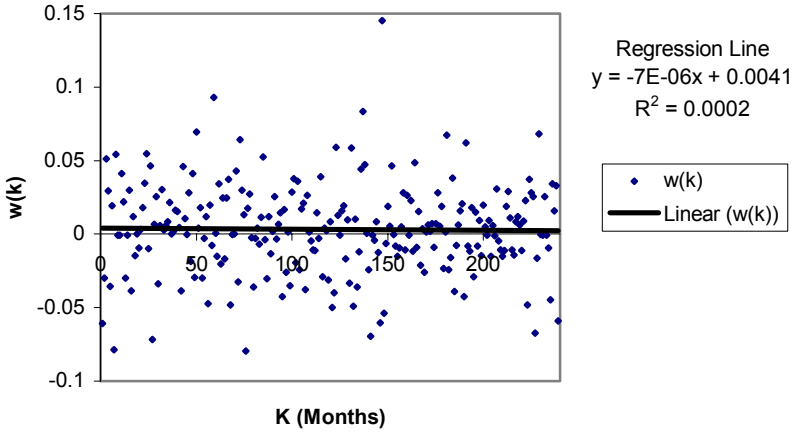

Figure 8. *w(k)* values for all 238 data points indicating the randomness of the *w(k)* values

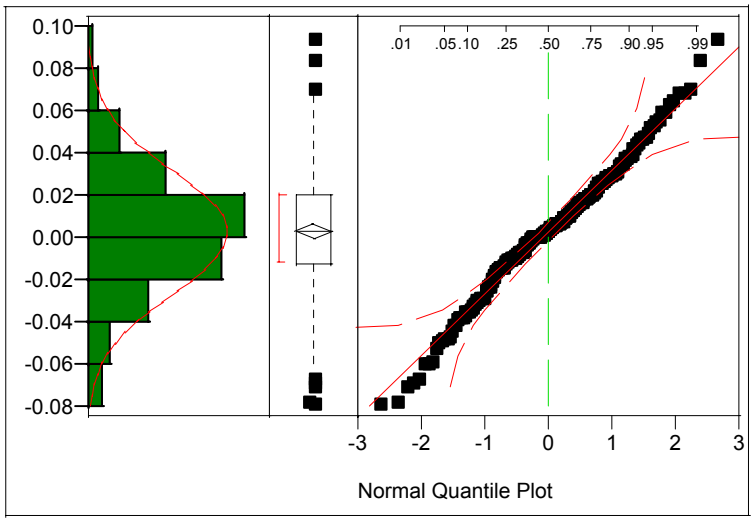

Figure 9. Distributions of *w(k)* for airline passenger enplanement data

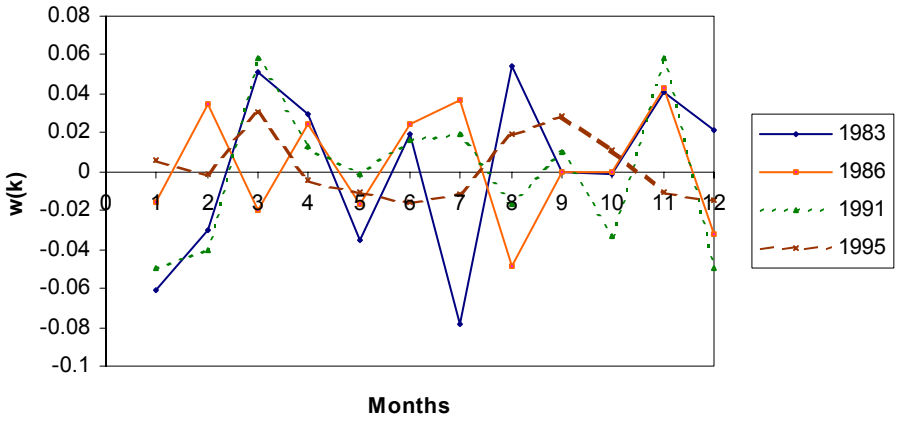

Figure 10. *w(k)* values for each month of year plotted for 4 years indicating randomness

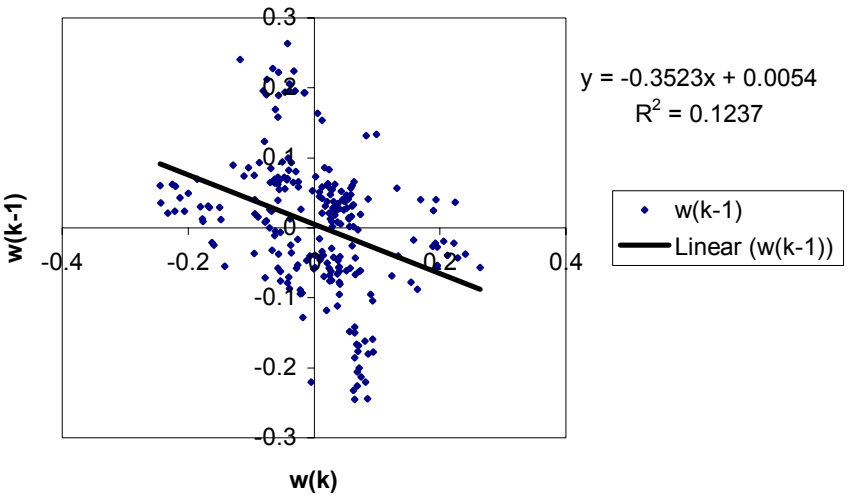

Figure 11. Scatter plot of *w(k)* values for airline passenger data before deseasonalization

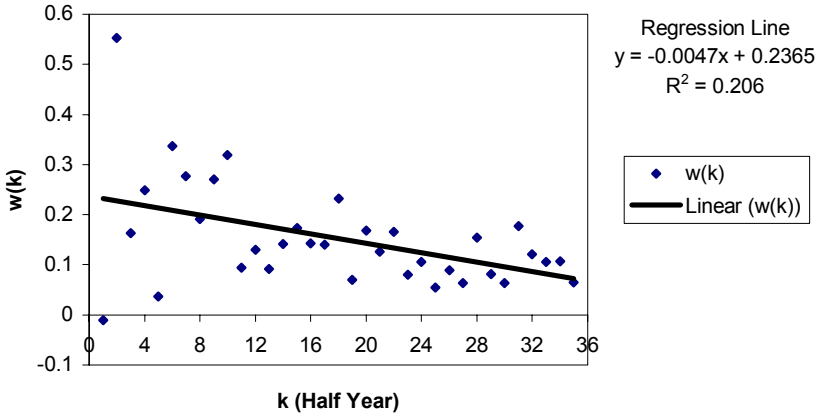

Figure 12. *w(k)* values for data points of cell phone data indicating the randomness of the *w(k)* values

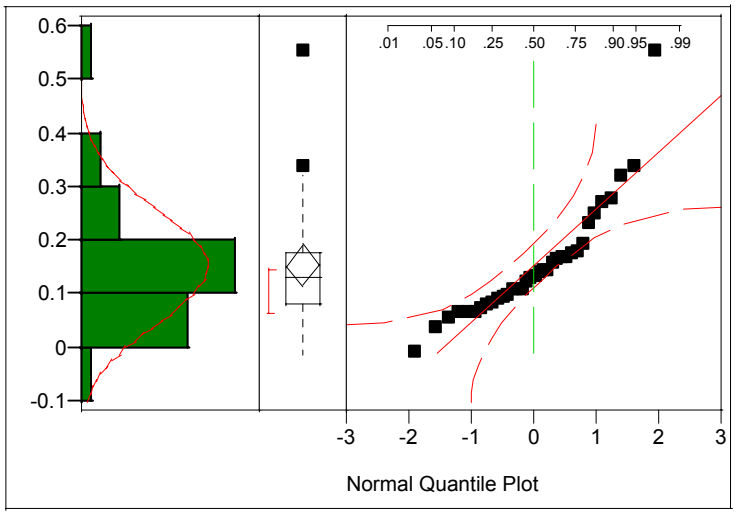

Figure 13. Distributions of *w(k)* for cell phone revenue data

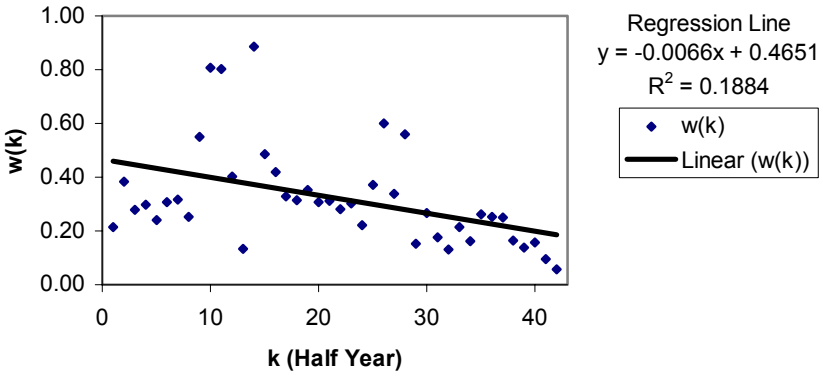

Figure 14. *w(k)* values for data points for internet host data

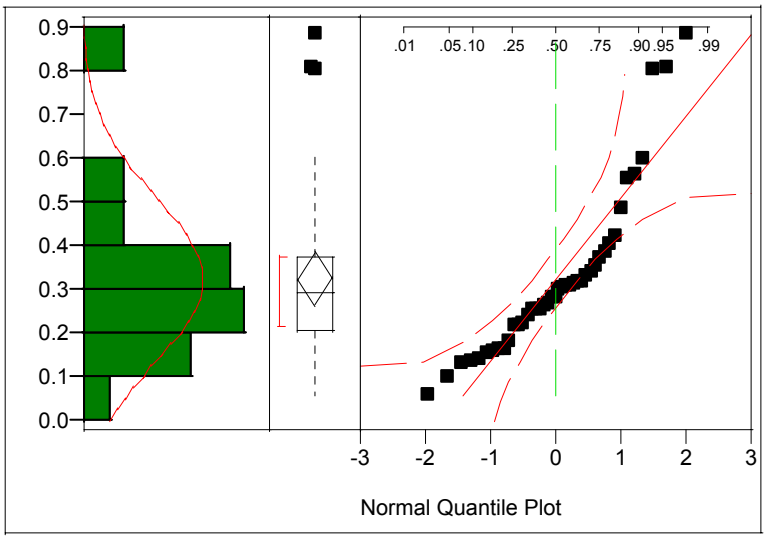

Figure 15. Distributions of *w(k)* for Internet host data

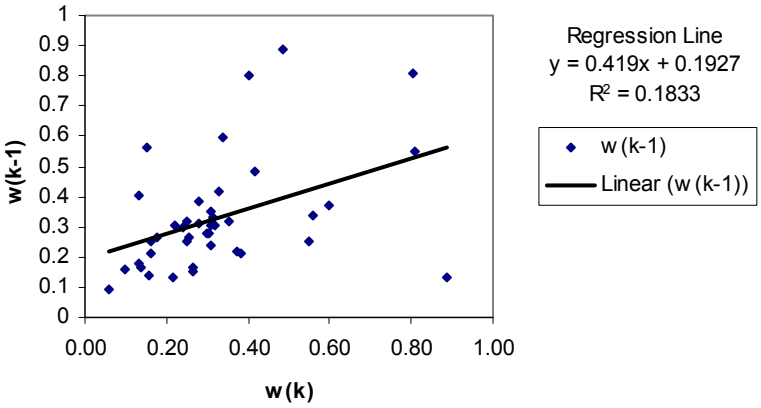

Figure16. *w(k)* scatter plot for Internet host data

List of Tables

Table 1. Seasonal Indices for electric power consumption data from two different methods

Month	Method I	Method II
January	1.0469	1.0372
February	0.9560	0.9507
March	0.9459	0.9458
April	0.8957	0.8952
May	0.9274	0.9244
June	1.0275	1.0288
July	1.1328	1.1396
August	1.1524	1.1594
September	1.0576	1.0649
October	0.9597	0.9686
November	0.9202	0.9269
December	0.9872	0.9971
Sum of Log	-0.00232	0

Table 2. Categories of *w(k)* values for the independence test

Categories	*w(k)* ranges
1	From -0.05 to -0.02
2	From -0.02 to 0
3	From 0 to 0.02
4	From 0 to 0.05

Table 3. Seasonal indices for airline passenger data using moving average method

Month	Sea. Index
January	0.8928
February	0.8686
March	1.0545
April	1.0073
May	1.0203
June	1.0704
July	1.1095
August	1.1361
September	0.9335
October	0.9962
November	0.9368
December	0.9664
Sum of Log	-0.00381

Table 4. Categories of *w(k)* used for airline passenger data

Categories	*w(k)* ranges
1	*w(k)* >0.04
2	*w(k)* from 0.04 to 0.01
3	*w(k)* from 0.01 to -0.03
4	*w(k)* from -0.08 to -0.03

Table 5. Categories of *w(k)* values used for cell-phone revenue data

Categories	*w(k)* range
1	from -0.02 to 0.085
2	from 0.085 to 0.16
3	from 0.16 to 0.25
4	from 0.25 to 0.6

Table 6. Summary of results of the various data sets

Data Set	Time Series	Normality	Independence	Remarks
Electric Utility	Electric Consumption data	Yes $p = 0.768$	Yes $p = 0.319$	Log ratios stationary and independent
Airline	Revenue Passenger Enplanement	Yes $p = 0.4416$	Yes $p = 0.058$	Log ratios stationary and independent
Cell phone	Revenue from Consumer Subscription	No $p = 0.0003$	Yes $p = 0.3735$	Independence test not credible because of fewer data points (Downward trend in log ratios over time)
Internet Industry	Number of Internet Hosts	No $p < 0.001$	No (No chi-square, just scatter plot)	Few Data Points, hence Chi-square independence test cannot be carried out (Downward trend in log ratios over time)