Songklanakarin J. Sci. Technol.
40 (4), 867-877, Jul. – Aug. 2018

http://www.sjst.psu.ac.th

Original Article

A modified Box and Cox power transformation to determine the standardized precipitation index

Tanachot Chaito[1, 2] and Manad Khamkong[2*]

[1] *Multidisciplinary Science Research Centre, Faculty of Science, Chiang Mai University, Mueang, Chiang Mai, 50200 Thailand*

[2] *Department of Statistics, Faculty of Science, Chiang Mai University, Mueang, Chiang Mai, 50200 Thailand*

Received: 12 September 2016; Revised: 27 April 2017; Accepted: 14 May 2017

Abstract

The objective of this research was to create a new transformation method, based on a modification of the Box and Cox power transformation (SMBC) by adding the ratio of skewness to two sample sizes, to characterized rought conditions with a standardized precipitation index (SPI). Along with the various classical data transformations, namely the Box and Cox power transformation (BC), the exponential transformation, the Yeo and Johnson transformation, and a modification to BC by adding range, the results of a simulation study showed that the BC and SMBC methods had similar efficiencies when transforming gamma data, Weibull data, and Pearson type III data to a normal distribution, and notably, SMBC performed particularly well with the latter. Drought conditions were evaluated using the SMBC transformation on real-life data from rain gauging stations at Muang (Lamphun), Mae Prik (Lampang) and Chom Thong (Chiang Mai), Thailand. The SMBC proved to be particularly useful in determining the SPI.

Keywords: drought, non-normal, positively –skewed distribution, rainfall, skewness

1. Introduction

A drought is a natural disaster caused by a water shortage in an area for an extended time, severely affecting the local economy and society. In Thailand, the main cause of drought is insufficient rain in the rainy season (Agro-meteorological Academic Group Meteorological Development, 2011). There are many ways to evaluate drought conditions, and the standardized precipitation index (SPI) is a popular and widely used method to detect and monitor drought.

The SPI was developed by McKee *et al.* (1993) for evaluating drought at different times. Calculation of the SPI uses cumulative rainfall data during a specific time-period of interest, for instance, over 3, 6, 12, 24, or 48-month time period, using the cumulative rainfall amount transformed to standard normal. The SPI can be used to classify conditions as dry or wet in each area (Agro-meteorological Academic Group Meteorological Development, 2011).

While examining the distribution of rainfall, McKee *et al.* (1993) fitted a gamma distribution to the rainfall data in the SPI calculation, whereas other researchers have tried to produce an appropriate model for studying rainfall when applying SPI to a particular area under study. For instance, Zhang *et al.* (2009) applied a log normal distribution to fit the rainfall data from Pearl River, China, whereas Yusof and Hui-Mean (2012) fitted the rainfall data from the state of Johor, Malaysia with a Weibull distribution having a heavy tail. Gabriel (2011) used a Pearson Type III distribution within his calculation of the SPI for Sao Paulo, Brazil, and Khamkong and Bookkamana (2011) fitted a generalized extreme value distribution for the annual maxima of daily and two-day rainfall data for upper Northern Thailand.

*Corresponding author
Email address: manad.k@cmu.ac.th

Environmental variables most frequently display asymmetric distributions with various levels of skewness and kurtosis, and precipitation data are no exception: they tend to be skewed positively (Twardosz & Walanus, 2011). In particular, data transformation methods are used to transform data that is non-normally distributed to become normally distributed. Manly (1976) stated that an exponential transformation is quite effective at turning a skewed unimodal distribution into a nearly symmetric normal distribution, whereas the Yeo and Johnson transformation (2000) is an effective series of power transformations of a skewed distribution. More recently, Watthanacheewakul (2012) proposed a modified Box and Cox transformation as an appropriate method to transform right-skewed data to become normal.

Our focus in this research was to evaluate the available data transformations by applying them to the standardized precipitation index for evaluating drought.

2. Materials and Methods

2.1 Types of distribution

The gamma distribution is a two-parameter family with the probability density function (pdf)

$$f(x;\alpha,\beta)=\frac{x^{(\alpha-1)}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}\text{ , for } x>0 \text{ and } \alpha,\beta>0, \tag{1}$$

where $\Gamma(\alpha)=\int_0^{\infty}x^{\alpha-1}e^{-x}dx$ is the gamma function, α is the shape parameter, and β is the scale parameter.

The Weibull distribution is a two-parameter family with the pdf

$$f(x;\alpha,\beta)=\frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}e^{-\left(\frac{x}{\beta}\right)^{\alpha}}\text{ , for } x>0 \text{ and } \alpha,\beta>0, \tag{2}$$

where α is the shape parameter and β is the scale parameter.

The Pearson Type III distribution (P3) is a three-parameter variation of the gamma distribution with the pdf

$$f(x;\alpha,\beta,\xi)=\frac{(x-\xi)^{\alpha-1}e^{-(x-\xi)/\beta}}{\beta^{\alpha}\Gamma(\alpha)}\text{ , for } \gamma>0,\xi=(\mu-2\sigma)/\gamma, \tag{3}$$

Here ξ is the location parameter, α is the shape parameter, β is the scale parameter, μ is the mean, σ is the standard deviation, and γ is the skewness.

The normal distribution is a two-parameter family with the pdf

$$f(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\text{ , for } x,\mu\in R \text{ and } \sigma^2>0, \tag{4}$$

where μ is the location parameter and σ is the scale parameter.

2.2 Transformation methods

The methods of transforming non-normal data to normal data, tested in our study, are outlined here.

2.2.1 The Box and Cox power transformation (BC)

Box and Cox (1964) proposed a method for transforming non-normal data into normal data with homogeneous variance:

$$Y = \begin{cases} \dfrac{x^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases}, \text{ for } x > 0, \tag{5}$$

where λ is a transformation parameter.

2.2.2 The exponential transformation (EP)

Manly (1976) proposed a data transformation that is quite effective at transforming a skewed unimodal distribution into a nearly symmetric normal distribution:

$$Y = \begin{cases} \dfrac{\exp(\lambda x)-1}{\lambda} & \lambda \neq 0 \\ x & \lambda = 0 \end{cases}, \tag{6}$$

where λ is a transformation parameter.

2.2.3 The Yeo and Johnson transformation (YJ)

Yeo and Johnson (2000) proposed a data transformation that was developed from the Box and Cox transformation and is effective on skewed distributions:

$$Y = \begin{cases} \dfrac{(x+1)^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \ln(x+1) & \lambda = 0 \end{cases}, \tag{7}$$

where λ is a transformation parameter.

2.2.4 Watthanacheewakul's (2012) modified Box and Cox power transformation (MBC)

This is a convenient method for transforming right-skewed data to a normal or nearly normal distribution. Watthanacheewakul (2012) proposed a modified Box and Cox transformation of the form

$$Y = \begin{cases} \dfrac{(x+c)^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \ln(x+c) & \lambda = 0 \end{cases}, \tag{8}$$

where λ is a transformation parameter and c is the range of the data.

2.2.5 The proposed modified Box and Cox power transformation

The proposed modified Box and Cox power transformation (SMBC) is an improvement on the original Box and Cox power transformation, which makes it appropriate for transforming right-skewed data to normal or nearly normal distribution. With the parameter c set to zero, this becomes the Box and Cox power transformation. BC has better efficacy for transforming data but is not appropriate when the data contains zero values. Rainfall data from Thailand include zero values in summer and winter, we proposed an alternative transformation where c is set to the ratio of skewness between two sample sizes.

$$Y = \begin{cases} \dfrac{(x+c)^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \ln(x+c) & \lambda = 0 \end{cases}, \tag{9}$$

where λ is a transformation parameter and $c = \dfrac{(\bar{x} - Q_2)}{2n(sd)}$, in which Q_2 is the 2nd quartile, $\bar{x}$ is the mean, sd is the standard deviation and n is the sample size.The estimation of the parameters in SMBC method is illustrated in the Appendix.

2.3 Estimation of transformation parameter

The value of transformation parameter λ in (5), (6), (7), and (9) can be estimated from the probability density function of a normal distribution as

$$f(y_i \mid \mu,\sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu)^2\right\}, \tag{10}$$

where y_i is the transformed value, μ is the mean and σ^2 is the variance.

The likelihood function of a normal distribution is given by

$$L(\mu,\sigma^2 \mid y_i) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}. \tag{11}$$

If $y_i = \dfrac{x^\lambda - 1}{\lambda}$ is the likelihood function, then

$$L(\mu,\sigma^2,\lambda \mid x_i) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\frac{x_i^\lambda - 1}{\lambda}-\mu\right)^2\right\}\cdot J(y;x), \tag{12}$$

where $J(y;x) = \prod_{i=1}^{n}\left|\dfrac{\partial y_i}{\partial x_i}\right|$ is the Jacobian of the transformation.

For a fixed λ, the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}^2$ are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i^\lambda - 1}{\lambda}\right)$$

and $$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{x_i^\lambda - 1}{\lambda} - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i^\lambda - 1}{\lambda}\right)\right\}^2.$$

Substitute $\hat{\mu}$ and $\hat{\sigma}^2$ into equation (12) to get the log likelihood as

$$\ln L(\lambda \mid x_i) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\left[\frac{1}{n\lambda^2}\left[\sum_{i=1}^{n}\left(x_i^\lambda - 1\right)^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\left(x_i^\lambda - 1\right)\right)^2\right]\right] - \frac{n}{2} + (\lambda-1)\sum_{i=1}^{n}\ln(x_i) \tag{13}$$

The maximum likelihood estimate of transformation parameter λ in (5) is solved from

$$\frac{d\ln L(\lambda \mid x_i)}{d\lambda} = \frac{-n\left(\sum_{i=1}^{n}x_i^{2\lambda}\ln x_i - \left(\frac{1}{n}\right)\left(\sum_{i=1}^{n}x_i^\lambda\right)\left(\sum_{i=1}^{n}x_i^\lambda \ln x_i\right)\right)}{\left[\sum_{i=1}^{n}x_i^{2\lambda} - \frac{1}{n}\left(\sum_{i=1}^{n}x_i^\lambda\right)^2\right]} + \frac{n}{\lambda} + \sum_{i=1}^{n}\ln(x_i) = 0, \tag{14}$$

and similarly the maximum likelihood estimate of λ in(6)is given by

$$\frac{d\ln L(\lambda \mid x_i)}{d\lambda} = \frac{-n\left(\sum_{i=1}^{n}e^{2x_i\lambda}x_i - \left(\frac{1}{n}\right)\left(\sum_{i=1}^{n}e^{x_i\lambda}\right)\left(\sum_{i=1}^{n}e^{x_i\lambda}x_i\right)\right)}{\left[\sum_{i=1}^{n}e^{2x_i\lambda} - \frac{1}{n}\left(\sum_{i=1}^{n}e^{x_i\lambda}\right)^2\right]} + \frac{n}{\lambda} + \sum_{i=1}^{n}x_i = 0, \tag{15}$$

while for (7) the condition is

$$\frac{d\ln L(\lambda|x_i)}{d\lambda}=\frac{-n\left(\begin{array}{l}\sum_{i=1}^{n}(x_i+1)^{2\lambda}\ln(x_i+1)\\-\left(\frac{1}{n}\right)\left(\sum_{i=1}^{n}(x_i+1)^{\lambda}\right)\left(\sum_{i=1}^{n}(x_i+1)^{\lambda}\ln(x_i+1)\right)\end{array}\right)}{\left[\sum_{i=1}^{n}(x_i+1)^{2\lambda}-\frac{1}{n}\left(\sum_{i=1}^{n}(x_i+1)^{\lambda}\right)^2\right]}+\frac{n}{\lambda}+\sum_{i=1}^{n}\ln(x_i+1)=0, \tag{16}$$

and the maximum likelihood estimate of the transformation parameter λ in(9)(details are given in appendix) is solved from

$$\frac{d\ln L(\lambda|x_i)}{d\lambda}=\frac{-n\left(\begin{array}{l}\sum_{i=1}^{n}(x_i+c)^{2\lambda}\ln(x_i+c)\\-\left(\frac{1}{n}\right)\left(\sum_{i=1}^{n}(x_i+c)^{\lambda}\right)\left(\sum_{i=1}^{n}(x_i+c)^{\lambda}\ln(x_i+c)\right)\end{array}\right)}{\left[\sum_{i=1}^{n}(x_i+c)^{2\lambda}-\frac{1}{n}\left(\sum_{i=1}^{n}(x_i+c)^{\lambda}\right)^2\right]}+\frac{n}{\lambda}+\sum_{i=1}^{n}\ln(x_i+c)=0. \tag{17}$$

Watthanacheewakul (2012) recommended a numerical method, such as bisection, to find a suitable value for the transformation parameter that makes the slope of the maximized log likelihood function nearly zero. However, to estimate the parameters, the bisection needs more iterations than either the secant method or some other alternatives. Thus, the secant method may be a preferred alternative over the bisection method.

2.4 Standardized precipitation index (SPI)

The SPI is the transformed value, when the cumulative amount of rainfall is transformed into standard normal distribution. A positive SPI indicates that the observed precipitation is greater than the mean precipitation, whereas a negative SPI indicates the contrary. Interpretations of the SPI values are shown in Table 1.

Table 1. The standardized precipitation index (SPI) categories based on the initial classification of SPI values.

SPI	Category
2.00 and above	Extremely wet
1.50 to1.99	Severely wet
1.00 to 1.49	Moderately wet
-0.99 to0.99	Near normal
-1.00 to-1.49	Moderately dry
-1.50 to-1.99	Severely dry
-2.00 and less	Extremely dry

2.5 Model selection criteria

Two selection criteria for choosing the best-fit distribution, given numeric data, are well-known.

2.5.1 Anderson-Darling (AD) test

Anderson and Darling (1952) proposed a goodness-of-fit test which compares the observed distribution of data with the expected CDF, and is defined as

$$A^2=-n-\frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln F(x_i)+\ln(1-F(x_{n+1-i}))\right], \tag{18}$$

where n is the sample size, $F(\cdot)$ is the expected CDF, and x_i are the ordered data.

2.5.2 Akaike information criterion (AIC)

Akaike (1973) proposed a criterion for selecting a model by comparing the actual model with the proposed model, and it is defined by

$$AIC=2k-2\ln L, \tag{19}$$

where k is the number of parameters in model and L is the maximized value of the likelihood function for the fitted model.

2.6 Numerical studies methodology

2.6.1 Simulation study

1. Generate the data from a gamma distribution, a Weibull distribution and a Pearson type III distribution, using various parameters and sample sizes, with the R program.
2. Investigate whether the data follow a gamma distribution, a Weibull distribution or a Pearson type III distribution by using the AD test at the 0.05 significance level.
3. Check whether the data are normally distributed according to the AD test at the 0.05 significance level.
4. Estimate the transformation parameters by applying the secant method.
5. Apply each data transformation to the simulated data.
6. Re-check to see if the transformed data are normally distributed according to the AD test at the 0.05 significance level.
7. Repeat steps 1-6 1,000 times.
8. Consider the percentages of normal distributions.

Full details are given in Figure 1.

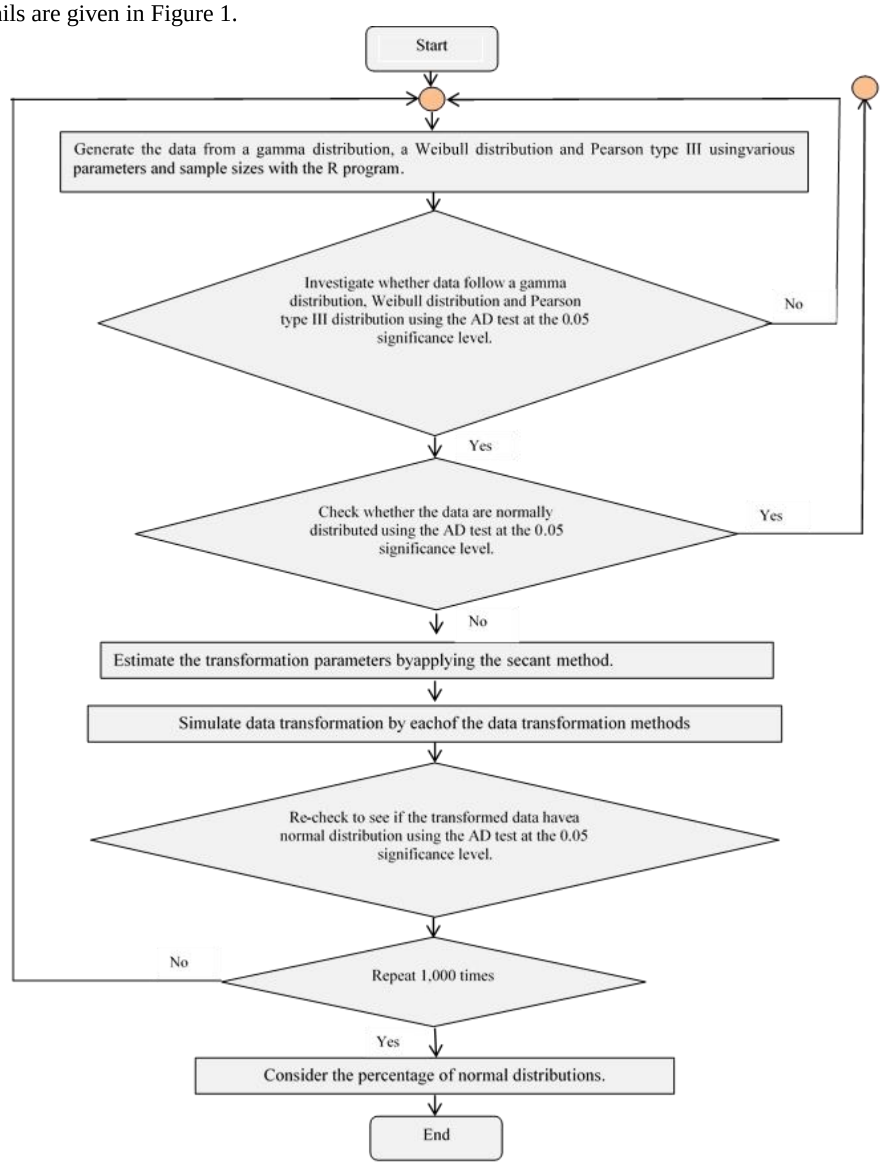

Figure 1. Steps in a simulation study.

2.6.2 Application to the SPI index of rainfall data

1. The rainfall data were separated into the three seasons: summer (February to May), rainy (June to September), and winter (October to January of the next year), each with four months duration. The real-life data were gathered from rain gauging stations at Muang (Lamphun), Mae Prik (Lampang), and Chom Thong (Chiang Mai), Thailand.

2. For each dataset, the unknown parameters of each distribution were estimated by the maximum likelihood method and the best-fit distribution was selected by using the AD test at the 0.05 significance level and by using the AIC.

3. Rainfall data were transformed using SMBC.

4. The transformed rainfall data were re-checked to see if they were normally distributed using the AD test at the 0.05 significance level.

5. If the transformed rainfall data had a normal distribution, they were evaluated as drought indicator values using the SPI.

3. Results and Discussion

3.1 Simulation study

For gamma distribution, the BC transformation obtained the highest percentage followed by the SMBC transformation. When the sample size was small, the EP transformation achieved the highest percentage, and when skewness and sample size increased, the EP the YJ and the MBC transformations showed poorer success rates (Table 2).

Table 2. Percentage normal distributions after transformation of a gamma distribution.

n	α	β	SK	BC	EP	YJ	MBC	SMBC
15	2	1	1.41	99.3	99.3	**99.6**	94.1	99.3
	4	1	1	99.6	**99.9**	99.8	98.8	99.6
	6	1	0.82	99.8	**100.0**	99.8	99.8	99.8
	2	3	1.41	99.3	93.4	98.8	96.7	**99.4**
	4	3	1	99.6	**99.8**	**99.8**	99.6	99.7
	6	3	0.82	99.8	**100.0**	99.9	99.8	99.8
30	2	1	1.41	**99.9**	96.5	99.2	99.8	**99.9**
	4	1	1	99.6	99.6	**99.8**	99.4	99.6
	6	1	0.82	99.1	99.1	99.2	**99.3**	99.1
	2	3	1.41	**99.9**	96.5	98.2	97.3	99.8
	4	3	1	99.6	99.6	**99.8**	99.4	99.5
	6	3	0.82	99.1	99.1	99.1	**99.3**	99.1
50	2	1	1.41	**99.0**	92.1	98.6	93.0	**99.0**
	4	1	1	99.2	97.1	**99.5**	98.3	99.2
	6	1	0.82	**100.0**	98.8	**100.0**	99.2	**100.0**
	2	3	1.41	**99.0**	92.1	96.7	93.0	**99.0**
	4	3	1	**99.2**	97.1	99.0	98.3	**99.2**
	6	3	0.82	**100.0**	97.6	99.9	99.2	**100.0**
100	2	1	1.41	**99.6**	73.7	95.4	77.2	**99.6**
	4	1	1	**99.7**	96.2	99.3	97.1	**99.7**
	6	1	0.82	99.3	99.2	**99.7**	99.3	99.3
	2	3	1.41	**99.6**	73.7	88.0	77.2	**99.6**
	4	3	1	**99.7**	96.2	99.0	97.1	**99.7**
	6	3	0.82	99.3	99.2	**99.7**	99.3	99.3

Note: The bold-face values were considered to indicate the performance transformation.

For a Weibull distribution, the BC and SMBC transformations were similarly efficient. Moreover, when skewness and the sample size increased, the SMBC gave the best performance (Table 3).

For a Pearson Type III distribution, the BC and SMBC transformations were similarly efficient. When the sample size was small, the MBC transformation had the highest success rate (Table 4).

3.2 Rainfall data study

Applications of the SMBC transformation to rainfall data are shown in Table 5.

The rain gauging station at Muang (Lamphun) had maximum/minimum cumulative seasonal rainfalls of 1,054 mm/34.4mmwith a central tendency of 239.4 mm. The Pearson type III distribution gave the best fit to these data, among the distributions tested.

Table 3. Percentage normal distributions after transformation of a Weibull distribution.

n	α	β	SK	BC	EP	YJ	MBC	SMBC
15	0.5	1	5.2	**98.0**	28.6	60.0	24.2	97.3
	1	1	0.65	96.6	81.2	93.2	76.1	**97.4**
	2	1	0.03	**69.4**	15.3	43.9	38.1	68.3
	0.5	3	5.2	98.0	28.6	76.3	24.2	**99.0**
	1	3	0.65	96.6	81.2	96.7	76.1	**96.9**
	2	3	0.03	**69.4**	15.3	54.4	38.0	69.1
30	0.5	1	5.2	**99.2**	1.8	16.7	3.9	85.9
	1	1	0.65	**99.6**	74.4	88.2	78.3	99.5
	2	1	0.03	98.5	96.5	**98.7**	96.9	98.6
	0.5	3	5.2	**99.2**	1.8	31.9	3.9	93.6
	1	3	0.65	**99.6**	74.4	93.8	78.3	**99.6**
	2	3	0.03	98.5	96.5	**99.4**	97.0	98.6
50	0.5	1	5.2	**98.8**	0.0	1.7	0.2	91.5
	1	1	0.65	98.9	46.9	76.5	47.6	**99.3**
	2	1	0.03	**96.4**	87.2	94.1	88.5	**96.4**
	0.5	3	5.2	**98.8**	0.0	9.5	0.2	96.3
	1	3	0.65	98.9	46.9	89.1	47.6	**99.3**
	2	3	0.03	**96.4**	87.2	95.8	88.5	**96.4**
100	0.5	1	5.2	93.8	0.0	0.0	0.0	**97.3**
	1	1	0.65	93.8	6.0	33.3	6.9	**97.4**
	2	1	0.03	**90.9**	73.2	83.4	75.0	68.3
	0.5	3	5.2	93.8	0.0	0.1	0.0	**99.0**
	1	3	0.65	93.8	6.0	64.2	6.9	**96.9**
	2	3	0.03	**90.9**	73.2	87.7	75.0	69.1

Note: The bold-face values were considered to indicate the performance transformation

Table 4. Percentage normal distributions after transformation of a Pearson Type III distribution.

n	α	β	ξ	SK	BC	EP	YJ	MBC	SMBC
15	4	1	0.5	0.47	95.4	94.4	95.6	**96.3**	95.4
	6	1	0.5	0.43	95.1	93.9	95.4	**96.2**	95.1
	4	1	1	0.47	**96.5**	95.3	96.4	96.4	**96.5**
	6	1	1	0.43	**95.4**	93.9	93.5	**95.4**	**95.4**
	4	3	0.5	2.46	**96.3**	94.9	96.1	96.2	**96.3**
	6	3	0.5	2.26	94.9	93.6	94.8	**95.4**	94.8
	4	3	1	2.46	96.2	94.9	96.1	**96.6**	96.2
	6	3	1	2.26	94.5	93.3	94.5	**95.4**	94.5
30	4	1	0.5	0.47	97.7	97.2	97.9	**98.3**	97.7
	6	1	0.5	0.43	97.1	96.6	92.0	**98.1**	97.1
	4	1	1	0.47	97.8	97.2	98.0	**98.2**	97.8
	6	1	1	0.43	97.0	96.6	96.9	**98.1**	97.0
	4	3	0.5	2.46	97.8	97.2	97.7	**98.3**	97.8
	6	3	0.5	2.26	96.8	96.6	97.1	**98.1**	96.8
	4	3	1	2.46	97.8	97.2	97.7	**98.3**	97.8
	6	3	1	2.26	96.9	96.6	96.9	**98.1**	96.9
50	4	1	0.5	0.47	**97.1**	95.5	**97.1**	97.0	**97.1**
	6	1	0.5	0.43	**96.3**	94.6	**96.3**	96.2	**96.3**
	4	1	1	0.47	97.0	95.5	**97.1**	97.0	97.0
	6	1	1	0.43	**96.4**	94.6	96.3	**96.4**	**96.4**
	4	3	0.5	2.46	**97.1**	95.5	**97.1**	97.0	**97.1**
	6	3	0.5	2.26	**96.3**	94.6	**96.3**	96.2	**96.3**
	4	3	1	2.46	**97.1**	95.5	**97.1**	97.0	**97.1**
	6	3	1	2.26	**96.3**	94.6	**96.3**	96.2	**96.3**
100	4	1	0.5	0.47	99.2	**99.6**	99.3	96.9	99.2
	6	1	0.5	0.43	**99.4**	98.3	99.5	98.6	**99.4**
	4	1	1	0.47	**99.3**	96.6	99.1	96.9	**99.3**
	6	1	1	0.43	99.5	98.3	**99.6**	98.7	99.5
	4	3	0.5	2.46	99.0	**99.6**	99.2	96.9	99.0
	6	3	0.5	2.26	**99.4**	98.3	**99.4**	98.6	**99.4**
	4	3	1	2.46	99.1	96.6	**99.2**	96.9	99.1
	6	3	1	2.26	99.4	98.3	**99.5**	98.6	99.4

Note: The bold-face values were considered to indicate the performance transformation

Table 5. Descriptive statistics and model selection criteria of the four–month rainfall period.

Stations	Rainfall		Median	QD	Skewness	Gamma		Weibull		P3	
	Min	Max				AIC	AD	AIC	AD	AIC	AD
Muang Lamphun	34.4	1,054.0	239.4	129.15	0.9818	2,295.5	1.248	2,299.1	1.514	**2,287.0**	**0.754**
Maepik	8.5	1,064.0	302.7	134.80	0.6872	2,261.1	0.647	**2,254.4**	**0.188**	2,259.0	0.324
Chom Thong	27.9	793.0	249.6	115.50	0.8360	**2,227.8**	**0.156**	2,230.8	0.857	2,239.0	0.156

Note: The bold-face values were considered to indicate the best distribution fitted to the real data and QD is quartile deviation.

The rain gauging station at Mae Phrik (Lampang) had maximum/ minimum cumulative seasonal rainfalls of 1,064 mm/8.5mm with a central tendency of 302.7mm. The Weibull distribution provided a very good fit to these data, along with the Pearson Type III and gamma distributions.

The rain gauging station at Chom Thong (Chiang Mai) had maximum/minimum cumulative seasonal rainfalls of 793 mm/27.9mm with a central tendency of 249.6mm. The gamma distribution was the most appropriate in this case.

The evaluation of drought severity from the SPI yielded the following results. In Figure 2, we can see that the Muang (Lamphun) station showed extremely dry conditions in 1974, 1996 and 1999, whereas, as seen in Figure 3, the Mae Phrik (Lampang) station had extremely dry conditions in 1960, 1990, 2005, and 2010. Lastly, data from the Chom Thong (Chiang Mai) station showed extremely dry conditions in 1960, 1992, 2004, and 2010 (Figure 4).

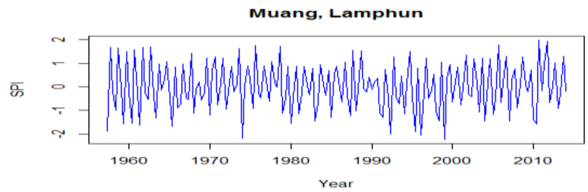

Figure 2. Four – month SPI for Muang, (Lamphun).

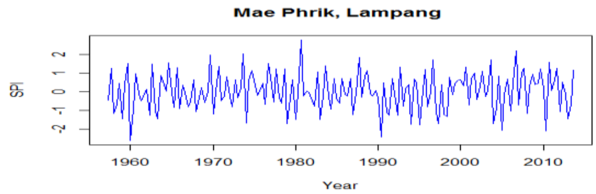

Figure 3. Four – month SPI for Mae Phrik, (Lampang).

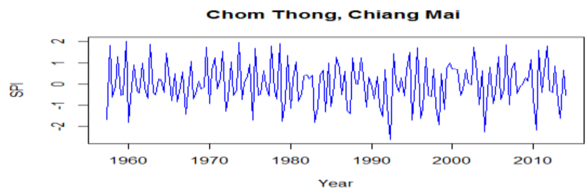

Figure 4. Four – month SPI for Chom Thong, (Chiang Mai).

4. Conclusions

Data transformations that convert a positively–skewed distribution to a normal distribution are essential for the drought analysis using SPI. Overall, the BC transformation is powerful but may not be appropriate when the data contains zero values. We proposed an alternative transformation, the SMBC, by adding the ratio of skewness for two sample sizes. The simulation results showed that the BC and SMBC transformations were similarly efficient for both gamma and Weibull distributed simulated data, and the SMBC performed particularly well with Weibull distributed data. Therefore, our recommendations are that the SMBC transformation is useful when the data follow a Weibull distribution or when the data contain zero values (as on evaluating drought conditions by applying the SPI). Moreover, if one uses the SMBC transformation to convert Weibull distributed data to become normally distributed, then we could construct confidence interval estimates of the population mean, which is useful for further analysis of the data. Future research could apply the SMBC transformation to other kinds of positively–skewed distributions, such as the Gumbel distribution, the generalized exponential distribution, and so on.

Acknowledgements

We are grateful to referees for their thoughtful comments. The authors are grateful for the financial and research facilities support by the Multidisciplinary Science Research Centre, Faculty of Science, Chiang Mai University, and by the CMU Mid-Career Research Fellowship program.

References

Agro-meteorological Academic Group Meteorological Development. (2011). Study on drought index in Thailand. Retrieved from http://www.tmd.go.th/info/info.php?FileID=71

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Proceeding of the *2nd International Symposium on Information Theory*, Tsahkadsor, Armenia, USSR, 267-281.

Anderson, T. W, & Darling, D. A. (1952). Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, *23*, 193–212.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), *26*(2), 211-252.

Gabriel, C. B. (2011). Standardized precipitation index based on Pearson type III distribution. *RevistaBrasileira de Meteorologia*, *26*(2), 167-180.

Khamkong, M., & Bookamana, P. (2011). Development of statistical models for maximum daily rainfall in upper northern region of Thailand. *Chiang Mai Journal of Science*, *42*(4), 1044-1053.

Manly, B. F. J. (1976). Exponential data transformations. *The Statistician*, *25*(1), 37-42.

McKee, T. B., Doesken, N. J., & Kleist, J. (1993). The relationship of drought frequency and duration on time scale. *Proceedings of the Eighth Conference on Applied Climatology*, California, CA, 179-184.

Twardosz, R., & Walanus, A. (2011). A practical method for testing of the significance of difference between average values in precipitation series. *Przeglad Geofizyczny*, *56*(1), 81–97.

Watthanacheewakul, L. (2012). Transformations with right skewed data. *Proceedings of the World congress on Engineering 2012*, London, England, 308-311.

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

Yusof, F., & Hui-Mean, F. (2012). Use of statistical distribution for drought analysis. *Applied Mathematical Sciences*, *6*(21), 1031–1051.

Zhang, Q., Xu, C. Y., & Zhang, Z. (2009). Observed changes of drought wetness episodes in the Pearl River Basin, China, using the standardized precipitation index and aridity index. *Theory of Applied Climatology*, *98*(89), 89 –99.

Appendix

Estimation of transformation parameter for SMBC

The transformation parameter (λ) can be estimated by the probability density function of a normal distribution as

$$f(y_i|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu)^2\right\}, \quad \text{(A1)}$$

where y_i is the observed value, μ is the mean, and σ^2 is the variance.

The likelihood function of a normal distribution is given by

$$L(\mu,\sigma^2|y_i) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\} \quad \text{(A2)}$$

For $y_i = \frac{(x_i+c)^\lambda - 1}{\lambda}$, the likelihood function can be written as

$$L(\mu,\sigma^2,\lambda,c\ |x_i) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\frac{(x_i+c)^\lambda - 1}{\lambda} - \mu\right)^2\right\} \cdot J(y;x) \quad \text{(A3)}$$

where $J(y;x) = \prod_{i=1}^{n}\left|\frac{\partial y_i}{\partial x_i}\right|$ is the Jacobian of the transformation.

For a fixed λ , the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}^2$ are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(x_i+c)^\lambda - 1}{\lambda}\right) \text{ and } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(x_i+c)^\lambda - 1}{\lambda} - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(x_i+c)^\lambda - 1}{\lambda}\right)\right\}^2.$$

Substitute $\hat{\mu}$ and $\hat{\sigma}^2$ into equation (A3). Thus, the log likelihood is

$$\ln L(\mu,,\sigma^2,\lambda|x_i) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\frac{(x_i+c)^\lambda - 1}{\lambda} - \mu\right)^2 + (\lambda-1)\sum_{i=1}^{n}\ln(x_i+c) \quad \text{(A4)}$$

The log likelihood functions of λ becomes

$$\ln L(\lambda|x_i) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(x_i+c)^\lambda - 1}{\lambda} - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(x_i+c)^\lambda - 1}{\lambda}\right)\right\}^2\right] - \frac{n}{2} + (\lambda-1)\sum_{i=1}^{n}\ln(x_i+c) \quad \text{(A5)}$$

The maximum likelihood estimate of the transformation parameter is subject to

$$\frac{d\ln L(\lambda|x_i)}{d\lambda} = \frac{d}{d\lambda}\left[-\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(x_i+c)^\lambda - 1}{\lambda} - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(x_i+c)^\lambda - 1}{\lambda}\right)\right\}^2\right] - \frac{n}{2} + (\lambda-1)\sum_{i=1}^{n}\ln(x_i+c)\right] \quad \text{(A6)}$$

$$\frac{d\ln L(\lambda|x_i)}{d\lambda} = \frac{-n\left(\begin{array}{l}\sum_{i=1}^{n}(x_i+c)^{2\lambda}\ln(x_i+c) \\ -\left(\frac{1}{n}\right)\left(\sum_{i=1}^{n}(x_i+c)^\lambda\right)\left(\sum_{i=1}^{n}(x_i+c)^\lambda\ln(x_i+c)\right)\end{array}\right)}{\left[\sum_{i=1}^{n}(x_i+c)^{2\lambda} - \frac{1}{n}\left(\sum_{i=1}^{n}(x_i+c)^\lambda\right)^2\right]} + \frac{n}{\lambda} + \sum_{i=1}^{n}\ln(x_i+c) = 0\,. \quad \text{(A7)}$$