ANALYSIS OF SAMPLING VARIABILITY OF THE STANDARDIZED PRECIPITATION INDEX

Antonino Cancelliere and Brunella Bonaccorso
Department of Civil and Environmental Engineering, University of Catania,
V.le A. Doria 6, 95125 CATANIA

Abstract

The Standardized Precipitation Index (SPI) is a widely used index for drought monitoring purposes that requires the preliminary fitting of a probability distribution to monthly precipitation aggregated at different time scales. The sampling properties of the SPI are investigated as a function of the sample size adopted for such distribution fitting. In particular, sampling properties of SPI, such as bias and mean square error (MSE), are analytically derived assuming the underlying precipitation series normally distributed, and compared with numerical values obtained by simulation for the cases of gamma and lognormal distributions. Also, the probabilities of correctly or incorrectly classifying drought conditions through the SPI are computed as a function of the available sample size. Results indicate that SPI values are significantly affected by the size of the sample adopted for its estimation. Furthermore, the theoretical MSE computed for the normal case fits well the one obtained numerically in the case of gamma and log-normal distributions, and therefore can find general application to estimate approximate confidence intervals for SPI values.

1 Introduction

Among the several methods proposed in literature for drought identification, the Standardized Precipitation Index (McKee et al.,1993) is one of the most widely applied, especially for drought monitoring. Its major strength stems from the possibility to compare drought events in regions with different climates, taking into account the various time scales that characterize the impacts affecting different water users.

The SPI is based on an equiprobability transformation of precipitation values aggregated at \( k \)-months into standard normal values, with \( k \) properly fixed according to the purpose of the analysis. In practice, the computation of the SPI requires i) fitting a probability distribution to aggregated monthly precipitation series (e.g. \( k = 3, 6, 12, 24, 36 \) months), ii) computing for each value the non-exceedence probability and iii) determining the corresponding standard normal quantile, which is the SPI value.

McKee et al. (1993) assumed aggregated precipitation gamma distributed and estimated parameters using maximum likelihood. Gutman (1998) discussed this assumption by analysing precipitation data from 1035 stations in the U.S. and concluded that, at least for the U.S. it would be preferable to use the Pearson type III distribution.

Regardless of the parametric distribution adopted, as a consequence of the procedure of parameter estimation, the SPI values will exhibit a sampling variability, namely, the SPI for a given year and a given month will depend on the sample size of the observed series of precipitation. This implies a potential limitation when comparing index values based on sample series of different length.
Objective of the present paper is to investigate the variability of the SPI with respect to the size of the sample used for estimating parameters. In particular, sampling properties of the index, such as Bias and Mean Square Error, are derived analytically in the case of precipitation normally distributed and numerically for the case of gamma and lognormal distributions. Finally, the probability of correctly classifying drought conditions as a function of the sample size is also computed for the normal case. The proposed methodology is applied to the SPI series computed on a precipitation series recorded in the Sicilian station of Caltanissetta, characterized by a long period of observation.

2 The Standardized Precipitation Index

In order to derive the sampling properties of the SPI, it is worth recalling its formal definition. With reference to a periodic monthly precipitation series \( X_{\nu,\tau} \), where \( \nu=1,...,n \) is the year and \( \tau=1,...,12 \) is the month, let’s define the backward aggregated series of order \( k \) as:

\[
Y_{\nu,\tau}^{(k)} = \sum_{i=0}^{k-1} X_{\nu,\tau-i}
\]

(1)

Assume \( Y_{\nu,\tau}^{(k)} \), for fixed \( \tau \), identically distributed according to some cumulative density function (cdf) \( F_{Y_{\nu,\tau}^{(k)}}(y) \). Then, the SPI value \( Z_{\nu,\tau}^{(k)} \) corresponding to a given value of \( Y_{\nu,\tau}^{(k)} \) is defined as:

\[
Z_{\nu,\tau}^{(k)} = \Phi^{-1}(F_{Y_{\nu,\tau}^{(k)}}(y))
\]

(2)

where \( \Phi^{-1}(\cdot) \) is the inverse of a standard normal cdf \( \Phi(\cdot) \) with zero mean and unit variance. From such definition it follows that the SPI index is normally distributed with zero mean and unit variance. Practical application of the SPI requires fitting a distribution to an observed sample \( Y_{\nu,\tau}^{(k)} \) for fixed \( \tau \).

From a theoretical standpoint, the distribution of aggregated precipitation \( Y_{\nu,\tau}^{(k)} \) follows directly from the distribution of monthly precipitation \( X_{\nu,\tau} \), although its analytical derivation may not be straightforward for most cases. For example, if \( X_{\nu,\tau} \) is normally distributed, then \( Y_{\nu,\tau}^{(k)} \) will be also normally distributed. Despite the theoretical possibility to derive analytically the distribution of \( Y_{\nu,\tau}^{(k)} \) from that of \( X_{\nu,\tau} \), in practice it is preferable (and easier) to choose a distribution for \( Y_{\nu,\tau}^{(k)} \) from a parametric family and estimate parameters directly from a sample of aggregated values. Following McKee et al. (1993), usually the gamma distribution is employed for such a task.

3 Sampling properties of SPI: normal distribution case

For the sake of simplicity, the series \( Y_{\nu,\tau}^{(k)} \) will be hereafter denoted by \( Y_{\nu}^{(k)} \), since we are interested in analysing the sampling properties of the SPI for fixed \( \tau \) and \( k \).
Let aggregated precipitation $Y_v$, with $v=1, 2, \ldots, n$, be an independent and identically normally distributed series with mean $\mu$ and variance $\sigma^2$, i.e.

$$Y_1, Y_2, \ldots, Y_n \sim \text{i.i.d. } N(\mu, \sigma^2)$$  \hspace{1cm} (3)

Let’s assume the mean $\mu$ and the variance $\sigma^2$ are estimated by the corresponding sample moments, namely:

$$\hat{\mu} = \bar{Y} = \frac{1}{n} \sum_{v=1}^{n} y_v$$  \hspace{1cm} (4)

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum_{v=1}^{n} (y_v - \bar{Y})^2$$  \hspace{1cm} (5)

where $y_1, y_2, \ldots, y_n$ is a sample from (3). With reference to a generic observation $Y$, not included in the estimation sample, the corresponding estimated SPI $\hat{Z}$ is given by the simple equation:

$$\hat{Z} = \Phi^{-1}(Y) = \frac{Y - \hat{\mu}}{\hat{\sigma}} = \frac{Y - \bar{Y}}{S}$$  \hspace{1cm} (6)

as descends directly from eq. (2) when $Y_v$ is normally distributed. The true SPI value $Z$, based on the population mean and standard deviation of the underlying series is:

$$Z = \frac{Y - \mu}{\sigma}$$  \hspace{1cm} (7)

Therefore, the sampling variability of the SPI can be characterized by investigating the distribution of the following random variable as a function of the estimation sample size $n$:

$$D = Z - \hat{Z} = \frac{Y - \mu}{\sigma} - \frac{Y - \bar{Y}}{S}$$  \hspace{1cm} (8)

The random variable $D$ is the sum of two random variables: the first is obviously normally distributed with zero mean and unit variance. To derive the distribution of the second r.v. $\hat{Z}$, it has to be observed that for the normal distribution, the sample mean $\bar{Y}$ and the sample variance $S^2$ are independent (see for example Mood et al. (1974)), and, besides $\bar{Y}$ is independent of both, since it is not included in the estimation sample. Furthermore, for iid normally distributed random variables the following well known results hold (Mood et al. (1974)):

$$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$$  \hspace{1cm} (9)

and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$  \hspace{1cm} (10)
The r.v. $\hat{Z}$ is therefore the ratio of a normal r.v. to the square root of a $\chi^2$ r.v. and thus, after an appropriate rescaling, it is distributed as a Student’s $t$ (Mood et al., 1974). Indeed, it can be shown that:

$$\frac{Y - \bar{Y}}{S} \sqrt{\frac{n}{n+1}} \sim t(n-1)$$

From eq. (11), it follows that $E\left[ \frac{Y - \bar{Y}}{S} \right] = 0$ and $\text{Var}\left[ \frac{Y - \bar{Y}}{S} \right] = \frac{n^2 - 1}{n(n-3)}$.

Exact analytical derivation of the distribution of $D$ is not an easy task, since it is the sum of two dependent r.v., one standard normal and the other one Student’s $t$. Nonetheless, the first two moments of $D$ can provide enough information to characterize the sampling variability of the SPI, since they allow to compute the bias and the Mean Square Error (MSE) of estimation. Indeed:

$$\text{Bias} = E[D]$$

(12)

$$\text{MSE} = E[D^2] = \text{Var}[D] + E[D]^2$$

(13)

In practice it is preferable to use the Root Mean Square Error (RMSE) of estimation which can be computed by taking the square root of the MSE:

$$\text{RMSE} = \sqrt{\text{MSE}}$$

(14)

The bias term $E[D]$ can be computed as:

$$E[D] = E\left[ \frac{Y - \mu}{\sigma} - \frac{Y - \bar{Y}}{S} \right] = E\left[ \frac{Y - \mu}{\sigma} \right] - E\left[ \frac{Y - \bar{Y}}{S} \right] = 0$$

(15)

since both expectations are zero. Thus, in the normal case, the SPI estimator given by eq. (6) is unbiased. As a direct consequence, the MSE of estimation coincides with $\text{Var}[D]$. On the basis of eq. (13), the MSE can be rewritten as:

$$\text{MSE} = \text{Var}\left[ \frac{Y - \mu}{\sigma} - \frac{Y - \bar{Y}}{S} \right] = \text{Var}\left[ \frac{Y - \mu}{\sigma} \right] + \text{Var}\left[ \frac{Y - \bar{Y}}{S} \right] - 2\text{Cov}\left[ \frac{Y - \mu}{\sigma}, \frac{Y - \bar{Y}}{S} \right]$$

(16)

The first term in the above equation is obviously 1. The second term, as it has been shown previously, is equal to $\frac{n^2 - 1}{n(n-3)}$. In order to derive the covariance term in eq. (16), it is worth to capitalize on a well known property of the covariance, thus rearranging it as:

$$\text{Cov}\left[ \frac{Y - \mu}{\sigma}, \frac{Y - \bar{Y}}{S} \right] = \text{Cov}\left[ \frac{Y}{\sigma}, \frac{Y}{\sigma} \right] - \text{Cov}\left[ \frac{Y}{\sigma}, \frac{\bar{Y}}{\bar{S}} \right]$$

(17)

The second term is zero since the observation $Y$ is not included in the estimation sample, and therefore it is uncorrelated with the sample mean $\bar{Y}$. 

A.Cancelliere and B. Bonaccorso
Taking into account conditional expectations concepts, the first covariance term can be rewritten as:

\[
\text{Cov}\left[ \frac{Y}{\sigma} , \frac{Y}{S} \right] = \frac{1}{\sigma} \text{Cov}\left[ Y, Y \right] = \frac{1}{\sigma} \left\{ E\left[ \text{Cov}\left[ Y, Y \frac{1}{S} \right] \right] + \text{Cov}\left[ E\left[ Y \frac{1}{S} \right], E\left[ Y \frac{1}{S} \right] \right] \right\} = \\
= \frac{1}{\sigma} \left[ E\left[ \frac{1}{S} \right] \right] \text{Var}[Y] = \sigma \cdot E\left[ \frac{1}{S} \right]
\]  

(18)

The latter expectation can be computed by reminding that \( S^2 \) is distributed according to a rescaled \( \chi^2 \) distribution (eq. (10)) and therefore, by direct integration of the expectation it follows:

\[
E\left[ \frac{1}{S} \right] = E\left[ \left( S^2 \right)^{\frac{1}{2}} \right] = \frac{1}{\sigma} \frac{\Gamma\left( \frac{n-2}{2} \right)}{\sqrt{n-1}}
\]  

(19)

Finally, combining eqs. (16), (18) and (19), it follows that:

\[
\text{MSE} = 1 + \frac{n^2-1}{n(n-3)} - \frac{n(n-1)}{\Gamma\left( \frac{n-2}{2} \right)} \cdot \sqrt{2(n-1)}
\]  

(20)

thus

\[
\text{RMSE} = \left[ 1 + \frac{n^2-1}{n(n-3)} - \frac{n(n-1)}{\Gamma\left( \frac{n-2}{2} \right)} \cdot \sqrt{2(n-1)} \right]^{\frac{1}{2}}
\]  

(21)

It can be inferred from eqs. (20) and (21) that, in the normal case, the MSE of estimation of SPI does not depend on the parameters \( \mu \) and \( \sigma^2 \) of the underlying variable, but only on the sample size \( n \). In Figure 1, the theoretical RMSE of SPI given by eq. (21) is plotted versus the sample size \( n \). In the same plot, the corresponding RMSE obtained by simulation is also shown. In particular, for a fixed sample length, standard normal series have been generated, and two values of SPI corresponding to a normal variable have been estimated: the first, using the mean and variance of the sample, and the other assuming the population mean and variance, namely 0 and 1 respectively. Then, RMSE has been computed as the average square difference between the two estimates.

It can be inferred from the plot that the RMSE exhibit a power law decrease with the sample size. In particular, the RMSE is about 0.24 for sample sizes of 30, while it is 0.15 for sample sizes of 70. The plot also indicates an excellent agreement between theoretical and numerical RMSE's, thus confirming the validity of the derived analytical expressions.
4 Sampling properties of SPI: other distributions cases

If the $Y_\nu$ are not normally distributed, the SPI must be computed according to eq. (2), which can be rewritten by explicitly taking into account the parameters of the distribution, here indicated generically by a parameter vector $\Theta$:

$$Z = \Phi^{-1}(F_Y(Y; \Theta))$$

(22)

In practice, the parameters $\Theta$ of the underlying distribution are unknown and therefore they will be estimated as $\hat{\Theta}$. As a consequence, the estimated SPI will take the form:

$$\hat{Z} = \Phi^{-1}(F_Y(Y; \hat{\Theta}))$$

(23)

Although in principle, the sampling variability of the SPI can again be characterized by the deviation between the true SPI and the estimated one (see eq. (8)), in practice derivation of the distribution of $\hat{Z}$ is rather cumbersome, which hinders the possibility of an exact analytical approach. Therefore, in order to analyze the statistical properties of $D$, a numerical experiment has been carried out to compute by generation techniques the bias and RMSE of the SPI. Two distributions $F_Y(y)$ have been investigated, namely the two parameters gamma and the two parameters lognormal. For each distribution, the bias and RMSE of the resulting SPI have been estimated numerically, following a similar procedure already outlined for the normal case, considering different sample sizes $n$ and parameters $\Theta$. In particular, different parameter sets have been considered by fixing several coefficients of variation $C_v$ of the distribution, and deriving the corresponding parameters.
The numerical experiment has been set up as follows:

- first, a value of the coefficient of variation $C_v$ is fixed, and the parameter values $\Theta$ are consequently derived according to the analyzed distribution;
- a series of $n$ values are sampled by generation from the selected distribution for fixed $\Theta$;
- a value $Y$ is also sampled from the same distribution, and the corresponding "true" SPI value $Z$ is computed using the parameter set $\Theta$;
- the parameters $\hat{\Theta}$ are estimated from the generated sample;
- the SPI value $\hat{Z}$ corresponding to the generated value $Y$ is then computed using the estimated parameters $\hat{\Theta}$;
- the procedure is repeated by generating 1000 series;
- the mean of the difference and of the square difference between the $\hat{Z}$'s and the "true" values $Z$ are computed, yielding an estimate of the bias and MSE;
- the whole procedure is then repeated for different $C_v$ and $n$.

As already mentioned, two distributions have been investigated, namely the two parameter gamma and lognormal cdf. The two parameter gamma probability density function (pdf) can be written as:

$$f_r(y) = \frac{1}{\Gamma(r)\beta} \left( \frac{y}{\beta} \right)^{r-1} e^{-\frac{y}{\beta}} (24)$$

The two parameters $r$ and $\beta$ are linked to the first two moments of the distribution by:

$$E[Y] = r\beta \quad (25)$$
$$Var[Y] = r\beta^2 \quad (26)$$

The coefficient of variation $C_v$ is therefore:

$$C_v = \frac{\sqrt{Var[Y]}}{E[Y]} = \frac{\sqrt{r}}{r} \quad (27)$$

Since the parameter $\beta$ does not affect the $C_v$, it has been assumed equal to one. Then assuming $C_v = 0.2, 0.4, 0.6$ and 0.8, the corresponding values $r = 25, 6.25, 2.78, 1.56$ have been determined.

The two parameter lognormal pdf can be written as:

$$f_r(y) = \frac{1}{\sqrt{2\pi y\sigma_x}} \exp \left[ -\frac{1}{2} \left( \frac{\ln y - \mu_x}{\sigma_x} \right)^2 \right] (28)$$

where the parameters $\mu_x$ and $\sigma_x$ are linked to the first two moments of the distribution as:

$$E[y] = \exp \left( \mu_x + \frac{\sigma_x^2}{2} \right) \quad (29)$$
Thus, for the lognormal case, the coefficient of variation $C_v$ is given by:

$$C_v = \sqrt{\frac{Var[Y]}{E[Y]}} = [\exp(\sigma_y^2) - 1]^{1/2}$$  \hspace{1cm} (31)$$

In Figure 2, the bias obtained by simulation for the two distributions are plotted versus the sample size for different $C_v$'s. It can be inferred that for practical purposes they are negligible and therefore the estimation can be assumed unbiased. Furthermore, the spread of the bias around the zero value do not seem to depend on the different $C_v$'s.

In Figure 3 the RMSE’s of estimation obtained by simulation are plotted versus the sample size for different $C_v$’s. For the sake of comparison, the theoretical RMSE derived for the normal case (eq. 21) is also plotted by continuous line. It can be inferred that the numerical RMSE's for both investigated distributions are rather similar, although the values related to the gamma distribution are generally slightly higher. Also it is worth pointing out that, as for the bias, even in this case the influence of different $C_v$’s is negligible. Furthermore a comparison between the RMSE’s derived for the normal case and the numerical values obtained for the other two distributions reveals a remarkable fitting of the theoretical line to the numerical RMSE, which suggests the possibility of extending the applicability, at least approximately, of eq. (21) also to the non-normal cases.

**Figure 2** Bias of estimation of SPI for the gamma distribution case (circles) and the lognormal distribution case (dots)
5 Classification probabilities

As already mentioned, the SPI index is generally adopted to classify climatic conditions according to its values. Although McKee et al. (1993) originally proposed a classification restricted only to drought periods, it has become customary to use the index to classify wet periods as well. Table I reports the climatic classification according to the SPI index provided by National Drought Mitigation Center (NDMC, http://www.ndmc.unl.edu/).

<table>
<thead>
<tr>
<th>Index value</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPI ≥ 2.00</td>
<td>Extremely wet</td>
</tr>
<tr>
<td>1.50 ≤ SPI &lt; 2.00</td>
<td>Very wet</td>
</tr>
<tr>
<td>1.00 ≤ SPI &lt; 1.50</td>
<td>Moderately wet</td>
</tr>
<tr>
<td>-1.00 ≤ SPI &lt; 1.00</td>
<td>Near normal</td>
</tr>
<tr>
<td>-1.50 ≤ SPI &lt; -1.00</td>
<td>Moderate drought</td>
</tr>
<tr>
<td>-2.00 ≤ SPI &lt; -1.50</td>
<td>Severe drought</td>
</tr>
<tr>
<td>SPI &lt; -2.00</td>
<td>Extreme drought</td>
</tr>
</tbody>
</table>

Figure 3 RMSE’s of estimation of SPI for the gamma distribution case (circles) and the lognormal distribution (dots)

For practical purposes it may be of some interest to assess how the probability of correctly classifying a drought conditions is affected by the sample size. In other words, assuming for instance that in a given month the true (unknown) drought class is for example severe, the objective here is to evaluate the probability that the estimated SPI (whose parameters are estimated on an observed sample) falls within the same drought class of the true one. Such probability will obviously depend on the distribution of the parameters used for computing the SPI and ultimately on the sample size.
More formally, let the true SPI $Z$ in a given month belong to a given class $C_o$. In general terms we are interested in the probability that the estimated SPI belongs to a class $C_M$, given the true SPI falls in the class $C_o$, i.e.:

$$P[\hat{Z} \in C_M | Z \in C_o] = P[C_{M_l} \leq \hat{Z} \leq C_{M_h} | C_{o_l} \leq Z \leq C_{o_h}]$$

(32)

where $C_{M_l}$ and $C_{M_h}$ represent the limits of class $C_M$ and $C_{o_l}$ and $C_{o_h}$ represent the limits of class $C_O$. Computation of the conditional probability given in eq. (32), requires the joint distribution of $Z$ and $\hat{Z}$. Marginally, it has already been shown that $Z$ is normally distributed, whereas $\hat{Z}$ is distributed according to a Student's t distribution. Therefore, exact derivation of the joint distribution of $Z$ and $\hat{Z}$ is not an easy task. In order to overcome such difficulty, the Student’s $t$ distribution will be approximated here by a normal distribution, assuming therefore a zero mean bivariate normal for the joint of $Z$ and $\hat{Z}$:

$$Z, \hat{Z} \sim \text{BNV}(0, \Sigma)$$

(33)

where $\Sigma$ is the variance covariance matrix of $Z, \hat{Z}$, namely:

$$\Sigma = \begin{bmatrix}
\text{var}[\hat{Z}] & \text{cov}[\hat{Z}, Z] \\
\text{cov}[\hat{Z}, Z] & \text{var}[Z]
\end{bmatrix}$$

(34)

whose elements have been derived previously.

The probability given by eq. (32) can therefore be computed as:

$$P[C_{M_l} \leq \hat{Z} \leq C_{M_h} | C_{o_l} \leq Z \leq C_{o_h}] = \frac{\int_{C_{o_l}}^{C_{o_h}} \int_{C_{M_l}}^{C_{M_h}} f_{\hat{Z},Z}(x,y) \, dx \, dy}{\int_{C_{o_l}}^{C_{o_h}} f_{Z}(x) \, dx}$$

(15)

where $f_{\hat{Z},Z}(x,y)$ is the bivariate normal pdf:

$$f_{\hat{Z},Z}(x,y) = \frac{1}{2\pi|\Sigma|} e^{-\frac{1}{2}(x-x')^\top \Sigma^{-1} (x-x')}$$

(26)

and $x' = [x,y]^\top$. Computation of the double integration in eq. (35) can be carried out by means of the algorithm MULNOR (Schervish, 1984).
In Table II the theoretical probabilities of correctly or incorrectly classify drought conditions according to the SPI are reported for different sample sizes. For simplicity, only drought conditions have been considered, although, due to the simmetry of the normal distribution, by replacing appropriately drought classes with wet classes, the corresponding probabilities for wet conditions can also be inferred. As expected, the probabilities that the sample SPI \( \hat{Z} \) belongs to the same class of the true SPI \( Z \) (see bold values along diagonals) increases as the sample size increases, while the probabilities of misclassification decrease with the sample size. Furthermore, such differences are particularly significant for the extremely, moderately and severely dry conditions, whereas for the near normal conditions the values are very similar.

**Table II.** Probabilities of correctly or incorrectly classify drought conditions using the SPI index for different sample sizes n=20, 50 and 70 years

<table>
<thead>
<tr>
<th></th>
<th>True SPI ( Z )</th>
<th>Sample SPI ( \hat{Z} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Extremely dry</td>
<td>Severely dry</td>
</tr>
<tr>
<td>n=20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extremely dry</td>
<td>0.74</td>
<td>0.25</td>
</tr>
<tr>
<td>Severely dry</td>
<td>0.13</td>
<td><strong>0.55</strong></td>
</tr>
<tr>
<td>Moderately dry</td>
<td>0.00</td>
<td>0.15</td>
</tr>
<tr>
<td>Near normal</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>n=50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extremely dry</td>
<td><strong>0.84</strong></td>
<td>0.16</td>
</tr>
<tr>
<td>Severely dry</td>
<td>0.08</td>
<td><strong>0.71</strong></td>
</tr>
<tr>
<td>Moderately dry</td>
<td>0.00</td>
<td>0.10</td>
</tr>
<tr>
<td>Near normal</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>n=70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extremely dry</td>
<td><strong>0.86</strong></td>
<td>0.14</td>
</tr>
<tr>
<td>Severely dry</td>
<td>0.07</td>
<td><strong>0.76</strong></td>
</tr>
<tr>
<td>Moderately dry</td>
<td>0.00</td>
<td>0.08</td>
</tr>
<tr>
<td>Near normal</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

6 Validation on observed series

Validation of the derived expressions has been carried out with reference to a long historical precipitation series namely the monthly precipitation series observed in the Sicilian station of Caltanissetta during the period 1879-2000 (122 years). Based on such a series, a new precipitation series aggregated on a time scale of \( k=12 \) months has been derived, assumed normally distributed. Then, 100 subseries of \( n \) years have been extracted from the aggregated series with reference to period 1879-1997, by applying a resampling bootstrap technique. Hence, the SPI’s related to the period 1998-2000 have been first computed by using each sample sub-series to estimate parameters and then, compared with the SPI’s computed using the whole 1879-1997 series, which due to the length of the estimation sample plays the role of "true" value.
In Figure 4 and 5 the comparison between SPI’s computed based on the 1879-1997 series and SPI’s based on the sub series of 30 and 70 years, respectively is represented. In the same figures, the approximate 90% confidence intervals are also plotted:

\[ \text{SPI}_{(1879-1997)} \pm \text{RMSE} \cdot \Phi^{-1}(\alpha) \]  \hspace{1cm} (37)

where RMSE ~ 0.23 for a sample size \( n = 30 \), RMSE ~ 0.15 for a sample size \( n = 70 \) and \( \alpha = 0.95 \). As can be inferred from the plots, the box plots are generally centered around the continuous line which represents the "true" value, thus confirming the unbiasedness of the estimation. Furthermore, the confidence intervals seem to reproduce well the sampling variability depicted by the box plots, thus confirming indirectly the validity of the derived expressions.

\[ \text{SPI}_{(1879-1997)} \pm \text{RMSE} \cdot \Phi^{-1}(\alpha) \]

**Figure 4** Comparison between SPI’s computed based on the 1879-1997 series (continuous line), SPI’s based on the sub series of 30 years (box plot) and confidence intervals based on eq. (35) (dashed line), derived from the monthly precipitation series of Caltanissetta

7 Conclusions

The variability of the SPI with respect to the size of the sample used for estimating the parameters of the probability distribution of aggregated precipitation has been investigated in the present paper. In particular, sampling properties of the index, such as bias and Mean Square Error, have been derived analytically for the case of normal distribution and numerically for the case of two parameters gamma and lognormal distributions. For the normal case, it has been shown that the SPI estimation is unbiased and that the RMSE cannot be considered negligible for sample sizes in the order of 20-30.
Figure 5  Comparison between SPI’s computed based on the 1879-1997 series (continuous line), SPI’s based on the sub series of 70 years (box plot) and confidence intervals based on eq. (35) (dashed line), derived from the monthly precipitation series of Caltanissetta

For the other two distributions the numerical investigation revealed that SPI can be assumed unbiased as well, while the resulting RMSE can be approximated well by the analytical expression obtained for the normal case.

Also, the probabilities of classifying correctly or incorrectly drought conditions according to the SPI values, have been computed for the normal case considering different sample sizes. The results have confirmed that the probability of correctly classifying the index increases as the sample size increases, and that it varies significantly depending on the drought class considered namely extreme, moderate or severe.

The theoretical RMSE have been validated on the basis on the long monthly precipitation series observed in the station of Caltanissetta, by means of bootstrap techniques. The derived expressions can find useful application for estimating approximate confidence intervals for SPI, thus allowing the comparison of SPI values computed in locations with precipitation record of different length. Future research will attempt also to derive the sampling properties of the SPI, conditioned on a fixed precipitation value.

Acknowledgements

The financial support from European Commission, through the MEDA project "Mediterranean drought preparedness and mitigation planning (MEDROPLAN", and of MIUR-COFIN2003, "Quantificazione e gestione delle risorse idriche in condizioni di scarsità" is gratefully acknowledged.
References


