Clustering of loglinear models using LRT p-values to assess homogeneous regions relative to droughts class transitions

<table>
<thead>
<tr>
<th>Journal:</th>
<th><em>Journal of Statistical Computation and Simulation</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>GSCS-2011-0413</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Original Paper</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>31-Oct-2011</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Moreira, Elsa; Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, CMA - Research center in mathematics and applications Mexia, João; Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, CMA - Research center in mathematics and applications Pereira, Luís; Instituto Superior de Agronomia, Universidade Técnica de Lisboa, CEER - Research Center of Biosystems Engineering</td>
</tr>
<tr>
<td>Areas of Interest:</td>
<td>CATEGORICAL DATA, CLUSTERING, ENVIRONMENTAL AND ECOLOGICAL, GENERALISED LINEAR MODELS</td>
</tr>
</tbody>
</table>

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

clustering_loglinear_models.tex
Clustering of loglinear models using LRT p-values to assess homogeneous regions relative to drought class transitions
[SPECIAL ISSUE: LINSTAT 2010]

Elsa Moreira\(^{a}\)\(^{\ast}\), João Tiago Mexia\(^{a}\) and Luís Santos Pereira\(^{b}\)

\(^{a}\)CMA - Research Center in Mathematics and Applications, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Campus Caparica, 2829-516 Caparica, Portugal;
\(^{b}\)CEER - Research Center of Biosystems Engineering, Instituto Superior de Agronomia, Universidade Técnica de Lisboa, Tapada da Ajuda, 1349-017 Lisboa, Portugal

(Received 00 Month 200x; in final form 00 Month 200x)

In the present study, a statistical method is used aiming at finding if Alentejo, southern Portugal, could be considered a homogeneous region for drought management purposes. Time series of the Standardized Precipitation Index (SPI) were obtained for 40 locations in the region using precipitation data for the period 1932-1999 (67 years). Contingency tables for the transitions between SPI drought classes were obtained for these time series. Loglinear models were fitted to these contingency tables to estimate the probabilities for drought class transitions. An approach of model clustering where loglinear models were clustered using the asymptotic p-value of a Likelihood Ratio Test (LRT) for the equality of parameters between pairs of models was applied. Two types of LRT were performed: one considering the parameters of the loglinear model all of interest; another considering just some parameters of interest. Two p-value similarity matrices were computed to find similar models that could form clusters, however, the hypothesis of model of clustering were not verified. No clustering was found, thus based on the presented technique, the Alentejo could be considered a homogeneous region relative to drought class transitions.

Keywords: Loglinear models; likelihood ratio test; model clustering; asymptotic p-value.

AMS Subject Classification: 62F03; 62F05; 62H17; 62H30; 62H11; 62P12

1. Introduction

Regions have specific features such as landscape and climate. Drought is a natural but temporary imbalance of water availability, consisting of a persistent lower-than-average precipitation, of uncertain frequency, duration and severity, of unpredictable or difficult to predict occurrence, resulting in diminished water resources availability, and reduced carrying capacity of the ecosystems [20]. Droughts can be characterized by their frequency, intensity and duration [31], as well as by the vulnerability of communities to drought impacts [11, 12]. In terms of droughts, an area is considered a homogeneous region if the drought characteristics of each location inside the region do not differ significantly.

Among indices used to characterize droughts, good results were obtained when adopting the Standardized Precipitation Index (SPI) [13, 14] computed with a 12 months time scale [19–22]. Drought regionalization is advantageous for various

\(^{\ast}\)Corresponding author. Email: efmm@fct.unl.pt
purposes such as planning for droughts and prediction of drought class transitions aimed at management to cope with drought water scarcity and drought risk management [26, 32]. An application to the Iberian Peninsula was developed by Vicente-Serrano (2006). The definition of homogeneous regions in terms of droughts is crucial in order to develop an improved spatial database management system. This kind of system provides for storage, handling, and analyzing spatial and temporal data, mainly long time series of weather variables.

Principal component analysis and cluster analysis are the usual methods used for delineating regions, including in drought studies [2, 3, 5, 24–26]. A recent study by Santos et al. (2010), applying principal component analysis and K-means clustering to re-analysis data for Portugal, assessed the spatial and temporal patterns of droughts in Portugal. Differences in behavior between northern and southern locations were found but no homogeneous regions were identified. Drought frequency, has been analyzed using a variety and probabilistic models. The models that the authors have been using to analyze drought frequency and intensity are loglinear models [16, 17]. In this paper, a different clustering approach was selected for the analysis. Since loglinear models proved well for analyzing and predicting transitions between successive SPI drought classes [16, 17], the adjusted models were used as objects to cluster. The presented statistical approach is clustering method and aims at the clustering loglinear models and assessing when the respective parameters are similar, which would indicate that the region where those models are applied is homogeneous. In [33], a theory for model clustering that applies to any kind of models fitted to data was presented. However before that, the problem of model clustering has been already addressed by [8, 9] with the comparison of multiple logit models, and comparison of among linear regressions [10]. From the Bayesian approach also [23, 29] considered model clustering.

The novelty of present work resides in the fact that is the first time the clustering of models using the asymptotic p-value of a LRT is applied with loglinear models and in particular to assess homogeneous regions relative to drought class transitions. Therefore, the formulation for the specific case of loglinear models was developed and implemented by the authors.

2. Data, SPI and drought classes

For drought monitoring and warning, meteorologists and hydrologists have developed indices, which depend on hydro-meteorological parameters or rely on probabilities of drought occurrence. The SPI, [13, 14], is used for the identification of drought events and to evaluate their severity (Fig. 1). The SPI is widely used because it allows a reliable and relatively easy comparison between different locations and climates. The SPI may be computed on shorter or longer time scales, which reflect different lags in the response of water cycle to precipitation anomalies. The 12-month time scale identifies dry periods of long duration which relate with the global impact of drought on hydrologic regimes and water resources of a region [20].

![Figure 1. SPI 12-months time series for several locations in Alentejo](image-url)

The data used in this study consists of time series of the SPI computed in a 12 month time scale for 40 weather stations located in Alentejo, southern of Portugal. The time period for the study was from September 1932 to September 1999 (67 years). The annual precipitation data sets used in SPI computation were investigated for randomness, homogeneity and absence of trends using the autocorrelation...
test Kendall's, the Mann-Kendall trend test and the homogeneity tests of Mann-Whitney for the mean and the variance [6].

Since the distances from the Atlantic and the north of Africa, could have influence in drought behavior, Alentejo was first divided into four sectors, according to latitude and longitude predefined limits (Fig. 2). In each sector 10 meteorologic stations were included arranged by each sector, ordered by latitude inside each sector and indexed by \( l = 1, \ldots, 40 \) as presented in Table 1. This arrangement was made as a first empirical approach for a regionalization of Alentejo and it will be useful in section 5, in order visualize the existence of clusters......

The SPI severity drought classes are defined in Table 2, which are modified from those proposed by McKee et al. (1993) by grouping the severe and extremely severe drought classes for modeling purposes since transitions referring to the extremely severe droughts class are much less frequent than for other classes; thus, a possible bias in the loglinear fitting is avoided [19, 20].

Using Table 1, the SPI time series for each station were converted into time series of drought classes. Then, the number of one step transitions between any drought classes in consecutive months \( (t \rightarrow t + 1) \) was counted in order to form a 2-dimensional \( 4 \times 4 \) contingency table with \( N = 16 \) cells each one. An example of these contingency tables is presented in Table 3. In these contingency tables, the observed frequencies, denoted by \( n_{ij} \), \( i, j = 1, \ldots, 4 \) on that table, are the number of times that it occurs the drought class \( i \) in a given month, followed by the drought class \( j \) in the next month, i.e. number of transitions between drought classes in successive months, where \( t \) represents a generic month. In each pair of consecutive months, the first one is the entry month, while the other is the exit month.
Table 3. Example of the contingency table for Chouto station

<table>
<thead>
<tr>
<th>Drought class month $t$</th>
<th>Drought class month $t+1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>37</td>
</tr>
<tr>
<td>2</td>
<td>38</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

3. Loglinear models with 2 dimensions

Previous studies, [16, 17], proved the appropriateness of adjusting to these contingency tables quasi-association (QA) loglinear models [equation (1)] [1]. Denoting by $m_{ij}$ the mean value $E(n_{ij})$ of $n_{ij}$, $i, j = 1, \ldots, 4$, also called expected frequency, the QA loglinear model is described by

$$
\log m_{ij} = \lambda + \lambda^r_i + \lambda^c_j + \beta \times i \times j + \delta_i I(i = j)
$$

- $m_{ij}$ is the expected number of transitions between drought class $i$ and $j$;
- $\lambda$ is the constant term;
- $\lambda^r_i$ is the effect of drought class $i$ of the entry month, $i = 1, \ldots, 4$ ($r$ designate the row effect of the contingency table);
- $\lambda^c_j$ is the effect of drought class $j$ of the exit month, $j = 1, \ldots, 4$ ($c$ designate the column effect of the contingency table);
- $\beta$ is the linear association parameter;
- $\delta_i$ is parameter associated to the $i$-th diagonal element of the contingency table, $i = 1, \ldots, 4$;
- $I(i = j)$ takes value 1 when the condition $i = j$ holds and value 0 otherwise.

In adjusting these models, it is assumed that the $n_{ij}$, $i, j = 1, \ldots, 4$ are values taken by independent Poisson distributed variables and the obtained maximum likelihood estimators $\hat{\lambda}, \hat{\lambda}^r_1, \hat{\lambda}^r_2, \hat{\lambda}^r_3, \hat{\lambda}^c_2, \hat{\lambda}^c_3, \hat{\lambda}^c_4, \hat{\beta}, \hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_3, \hat{\delta}_4$ and $\hat{m}_{ij}$, $i, j = 1, \ldots, 4$ [1]. Not all the parameters in the model were linearly independent, since the condition $\sum_{i=1}^4 \lambda^r_i = \sum_{j=1}^4 \lambda^c_j = 0$ has to be respected. Thus, to lighten the model, it was assumed that $\lambda^r_1 = \lambda^c_1 = 0$ as for previous studies in the same region [16, 17].

The QA loglinear model in matrix notation will be

$$
\log m = X\theta = \sum_{k=1}^{12} x_{h,k} \theta_k
$$

where

- $\theta = [\lambda, \lambda^r_2, \lambda^r_3, \lambda^r_4, \lambda^c_2, \lambda^c_3, \lambda^c_4, \beta, \delta_1, \delta_2, \delta_3, \delta_4] = [\theta_1, \ldots, \theta_{12}]$ is the vector of the 12 linearly independent parameters $\theta_k, k = 1, \ldots, 12$ in the model;
- $m = (m_1, \ldots, m_{16})$ are the vectors of expected frequencies ordered by indexes $h = 4i + j - 4, h = 1, \ldots, 16$;
For Peer Review Only

October 31, 2011 10:51 Journal of Statistical Computation & Simulation

cluster- ing˙loglinear˙models

5

\[ X = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 4 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 6 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 8 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 1 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 6 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 9 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 12 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 8 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 6 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 12 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 6 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 1 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 12 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 6 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 1 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix} \]

is the model matrix with elements \( x_{hk} \), \( h = 1, ..., 16 \) and \( k = 1, ..., 12 \) derived from eq. (1).

Since a rather long time span was used, it may also be assumed that:

- \( \hat{\theta} \) of MLE estimates is asymptotically normal with mean value \( \theta \) and variance-covariance matrix

\[
(X^T D(\hat{m})X)^{-1}
\]

where \( D(\hat{m}) \) is the diagonal matrix, whose principal elements are the adjusted expected frequencies (Agresti, 1990);

- \( \hat{\theta} \) is independent from the residual deviance

\[
G^2 = 2 \sum_{h=1}^{16} n_h \log(n_h/\hat{m}_h)
\]

which is asymptotically distributed as a central Chi-Square with 4 degrees of freedom, since there are 16 cells in the contingency tables and 12 linearly independent parameters to be adjusted. So, the Chi-Square test with statistic \( G^2 \) was used to validate the adjustment of the model [1, 18].

In Table A.1 of the Appendix are presented the adjusted parameters and the residual deviations for the different stations grouped according to the considered sectors.

Given that the \( n_h, h = 1, ..., 16 \) are Poisson random variables, the log-likelihood function are

\[
\log L(\theta) = \log\left(\prod_{h=1}^{16} \frac{\exp(-m_h)m_h^{n_h}}{n_h!}\right) = \sum_{h=1}^{16} n_h \log m_h - \sum_{h=1}^{16} m_h
\]

and

\[
\frac{\partial \log L(\theta)}{\partial \theta_k} = \sum_{h=1}^{16} n_h x_{hk} - \sum_{h=1}^{16} x_{hk} \exp(\sum_{k=1}^{12} x_{hk} \theta_k)
\]

since \( m_h = \exp(\sum_{k=1}^{12} x_{hk} \theta_k) \).

4. Clustering of models

In this section, the existence of homogeneous regions in Alentejo regarding drought class transitions is evaluated using a procedure for clustering the loglinear models associate to the 40 stations. The method used for grouping of the models associated
to each object of a data set is described in the following. Before, let’s define the similarity between models through the similarity between their parameters.

### 4.1. Model linking

Let’s now suppose that there are \(N\) objects, that in our case study, are the 40 meteorological stations. For the \(t\)-th object, there is a data set to which a parametric model \(M(\theta_t), t = 1, ..., N\), of the same family was fitted. \(\theta_t\) is the vector of parameters of the \(t\)-th model with \(s\) components. First, all the parameters of the model are considered of interest to our clustering variable; in a second phase only a subset of the total number of parameters are important.

#### 4.1.1. LRT relative to all model parameters

To group the 40 stations (40 objects) in terms of drought behavior, the loglinear models fitted to each station have to be linked two by two. Therefore, the null hypothesis

\[ H_0^{t,l} : \theta_t = \theta_l, t, l = 1, ..., 40 \tag{7} \]

need to be tested for each pairwise combination of models through the use of the well-known Likelihood Ratio Test (LRT) [7, 33].

The joint log-likelihood of the models \(t\) and \(l\) is

\[ \log L(\theta_t, \theta_l) = \log L(\theta_t) + \log L(\theta_l) \tag{8} \]

thus, due to (5),

\[ \log L(\theta_t, \theta_l) = \sum_{h=1}^{16} n_h(t) \sum_{k=1}^{12} x_{hk}\theta_k(t) - \sum_{h=1}^{16} \exp(\sum_{k=1}^{12} x_{hk}\theta_k(t)) \]
\[ + \sum_{h=1}^{16} n_h(l) \sum_{k=1}^{12} x_{hk}\theta_k(l) - \sum_{h=1}^{16} \exp(\sum_{k=1}^{12} x_{hk}\theta_k(l)). \tag{9} \]

Then, under null hypotheses, \(\theta_t = \theta_l = \theta\), the joint log-likelihood reduces to

\[ \log L_0(\theta) = \sum_{h=1}^{16} (n_h(t) + n_h(l)) \sum_{k=1}^{12} x_{hk}\theta_k - 2 \sum_{h=1}^{16} \exp(\sum_{k=1}^{12} x_{hk}\theta_k). \tag{10} \]

The LRT test statistic [7], for our null hypotheses is

\[ 2 \max_{\theta_t, \theta_l} \log L(\theta_t, \theta_l) - 2 \max_{\theta} \log L_0(\theta) \overset{d}{\rightarrow} \chi^2_{12}. \tag{11} \]

In order to maximize the log-likelihood functions, the following 3 systems of
likelihood equations
\[
\frac{\partial}{\partial \theta_k(t)} [\log L(\theta_t)] = 0, \forall k = 1, ..., 12,
\]
\[
\frac{\partial}{\partial \theta_k(t)} [\log L_0(\theta)] = 0, \forall k = 1, ..., 12,
\]
\[
\frac{\partial}{\partial \theta_k} [\log L_0(\theta)] = 0, \forall k = 1, ..., 12,
\]

have to be obtained and solved, in order to each parameter. In matrix notation the
likelihood equations are
\[
X^T(n_t - \hat{m}_t) = 0
\]
\[
X^T(n_t - \hat{m}_t) = 0
\]
\[
X^T(n_t + n_l - 2\hat{m}_l) = 0.
\]

where \(n_t = [n_1(t), ..., n_{16}(t)], t = 1, ..., 40\). Then, after obtaining the estimators
of the parameters by solving the 3 systems of equations, the parameters are re-
placed by their estimators and the expression (11) can be computed. After that,
the asymptotic p-value for our null hypotheses is obtain from the chi-square dis-
tribution with 12 degrees of freedom (expression (11)).

4.1.2. LRT relative to a subset of model parameters

Our main interest for this case study is in the severe/extreme droughts, its du-
ration and frequency. As a result, the drought class transition that more interest
us to analyze is the transition from drought class 4 (severe/extreme drought) to
itself. The number of transitions from drought class 4 to 4 is given by \(m_{44}\) in
equation (1) and \(m_{16}\) in equation (2). The parameters of the loglinear model in-
volved in the computation of that drought transitions are the \(\lambda, \lambda^c, \lambda^r, \beta, \delta\).
Therefore, the vector of parameters of interest was chosen to be the vector
\[\theta = (\lambda, \lambda^c, \lambda^r, \beta, \delta) = (\theta_1, \theta_4, \theta_7, \theta_8, \theta_{12})\]

having 5 components. The remaining
7 parameters, as defined before, will compose the vector of ancillary parameters
\[\alpha = (\lambda^c_2, \lambda^c_3, \lambda^r_2, \lambda^r_3, \delta_1, \delta_2, \delta_3) = (\theta_2, \theta_3, \theta_6, \theta_9, \theta_{10}, \theta_{11})\]

In this case, the fitted model is represented as \(M(\theta_t, \alpha_t), t = 1,...,N\), where
\(\theta_t\) is the vector of fitted parameters of interest, representing a subset of total model
parameters and \(\alpha_t\) is the vector of ancillary parameters, i.e., the parameters that
are supplementary regarding the clustering variable. The null hypotheses that to
test in order to link two models still is
\[H_0^{tl} : \theta_t = \theta_l, t,l = 1, ..., 40.\]

where \(\theta_t\) and \(\theta_l\) are the vectors of parameters of interest for model \(i\) and \(l\) respec-
tively, both with just 5 components.

The joint log-likelihood of the models \(M(\theta_t, \alpha_t)\) and \(M(\theta_l, \alpha_l)\) is
\[
\log L(\theta_t, \theta_1, \alpha_t, \alpha_l) = \log L(\theta_t, \alpha_t) + \log L(\theta_l, \alpha_l)
\]

and the maximized LRT is
\[
2 \max_{\theta_t, \theta_1, \alpha_t, \alpha_l} \log L(\theta_t, \theta_1, \alpha_t, \alpha_l) - 2 \max_{\theta_t, \alpha_t, \alpha_l} \log L_0(\theta, \alpha_t, \alpha_l) \overset{d}{\to} \chi^2_2.
\]
Then, in order to maximize these log-likelihood functions, the following 5 systems of likelihood equations

\[
\frac{\partial}{\partial \theta_k(t)} \left[ \log L(\theta_t, \alpha_t) \right] = 0, \forall k = 1, \ldots, 12,
\]

\[
\frac{\partial}{\partial \theta_k(l)} \left[ \log L(\theta_l, \alpha_l) \right] = 0, \forall k = 1, \ldots, 12,
\]

\[
\frac{\partial}{\partial \theta_k} \left[ \log L_0(\theta, \alpha_t, \alpha_l) \right] = 0, \forall k = 1, 4, 7, 8, 12,
\]

\[
\frac{\partial}{\partial \theta_k(t)} \left[ \log L_0(\theta, \alpha_t, \alpha_l) \right] = 0, \forall k = 2, 3, 5, 6, 9, 10, 11,
\]

\[
\frac{\partial}{\partial \theta_k(l)} \left[ \log L_0(\theta, \alpha_t, \alpha_l) \right] = 0, \forall k = 2, 3, 5, 6, 9, 10, 11,
\]

have to be obtained and solved, in order to each parameter. Written in matrix notation the 5 systems are

\[
X^T_n - X^T \hat{m}_t = 0
\]

\[
X^T n_l - X^T \hat{m}_l = 0
\]

\[
X^T_{int} (n_t + n_l - \hat{m}_{int}(\hat{m}_{anc_t} + \hat{m}_{anc_l})) = 0
\]

\[
X^T_{anc} (n_t - \hat{m}_{int}\hat{m}_{anc_t}) = 0
\]

\[
X^T_{anc} (n_l - \hat{m}_{int}\hat{m}_{anc_l}) = 0
\]

(18)

where

- \( X_{int} \) represents a sub-matrix of \( X \) with only the line vectors 1,4,7,8,12;
- \( X_{int} \) represents a sub-matrix of \( X \) with only the line vectors 2,3,5,6,9,10,11;
- \( \hat{m}_{int} \) are the vector of the expected frequencies obtained from considering the vector \( \theta \) of parameters of interest under \( H_0 \);
- \( \hat{m}_{anc_t} \) are the vector of the expected frequencies obtained from considering the vector \( \alpha_t \) of ancillary parameters of model \( t \);
- \( \hat{m}_{anc_l} \) are the vector of the expected frequencies obtained from considering the vector \( \alpha_l \) of ancillary parameters of model \( l \).

Then, solving the 5 systems of equations and obtaining the parameters estimates, the expression (16) can be computed to obtain the asymptotic p-value for our null hypotheses, \( H_0 \).

### 4.2. Clustering method

There are several strategies of clustering that can be used. However, to start the process, the simplest one, called the manual grouping is used. In this strategy, first the LRT for the null hypothesis \( H_0^{t,l}, t, l = 1, \ldots, 40 \) is performed in a number equal to the number of combinations of 40 objects two by two, i.e, 780 times. If \( H_0^{t,l}, t, l = 1, \ldots, 40 \) is likely to be true, the p-value would be large; otherwise, it would be small [33]. Therefore, this p-value can be used as the similarity measure between pairs the loglinear models. Unlike most similarity measures used in classic cluster analysis, the p-value is not a metric distance.

With the p-values obtained from the LRT for all combinations of pairs of models,
a 40 × 40 p-value matrix is formed. This matrix is symmetric with the principal elements of the diagonal equal to 1. In Tables A.2 and A.3 of the Appendix are presented the obtained p-value matrix for the case of LRT relative to all model parameters and for the case of LRT relative to a subset of the model parameters. Since the matrix is symmetric, just the values up to the diagonal are included. In these matrices, the vertical and horizontal lines limit the 4 sectors predefined by us in Table 1 and Fig. 2. The computation of this two matrices were implemented in Visual Basic with Excel as editor and the Excel Solver Tool to solve the systems of likelihood equations.

To perform the clustering, each model is first considered in a separated cluster [33]. Then, in order to find any visible cluster, a threshold \( \alpha_0 \) is settled. When the p-value is up or equal to this threshold, the examined two clusters are considered that linked, i.e, the correspondent two models are similar. The selection of \( \alpha_0 \) is subjective and different values of \( \alpha_0 \) will lead to different clustering results. In this study, the select threshold was \( \alpha_0 = 0.05 \), since it is not admissible for us less than a probability of 95% for two models to be similar. In Tables A.2 and A.2 the p-values that do not exceed 5% are in bold. The model pairs whose p-values are not smaller than 0.05 are then selected to form clusters from the selected model pairs, since the two models are similar.

5. Discussion of results and conclusions

If the predefined 4 sectors exist (Table 1), then the p-value matrices should have a configuration the one in Table A.4, i.e., the high p-values - larger than 5% - should group around the diagonal. This kind of grouping means that the models inside each sector are linked between themselves, but not are linked with the models of the other sectors. However, when looking to both p-value matrices (Tables A.2 and A.3), most of p-values exceed the threshold \( \alpha_0 = 0.05 \); exceptions consist of few cases dispersed throughout the matrix. This means that the vast majority of models are linked together. The p-values larger than 5% are also spread throughout the matrix, thus not denoting any agglomeration that could indicate different sectors inside the Alentejo. The kind of matrix configuration shown in Tables A.2 and A.3 indicates a nonexistence of clusters.

Moreover, the p-value matrix obtained for the LRT relative to the parameters of interest - parameters that play a role in the calculations of the transition between the drought class severe/extreme to itself - does not differ much of the p-value matrix obtained for the LRT relative to all the parameters of the model. Although the second matrix has less p-values under 5%, the majority of those are for the same pair of models. As a result, when considering just the parameters of interest, unlike it could be expected, the number of pairs of similar models increases, reinforcing the idea of Alentejo as an homogeneous region in terms of drought behavior.

In order to validate the computations of the p-value matrices, the Pearson correlation coefficient \( r \) was computed between the p-values and the total of the differences in absolute value between the homolog parameters of the two models given by

\[
\sum_{k=1}^{12} |\hat{\theta}_l - \hat{\theta}_t|, t, l = 1, ..., 40; t \neq l.
\]

A significant negative correlation at 1% was obtained, indicating that smaller differences between model parameters induce to high p-values. This correlation makes
sense in the context of the methodology used, since higher p-values indicate similarity between the models.

Also, in order to assess if a larger threshold $\alpha_0$ could help to detect a clustering not identified when $\alpha_0 = 0.05$, the histograms for the p-values frequencies were obtained (Fig 3). In both histograms the higher frequency occurs for the range of p-values from 0.9 and 1, and the density probability of the p-values does not allow any conclusion. Finally to end the analysis of the results, regarding the dispersed p-values under 5%, an attempt to understand the reason why those two models are not similar was made. However, based on the factors longitude and latitude it was not possible explain the dissimilarity. Only the fact that the models in these pairs belong to different predefined sectors was noticed.

Therefore, according to the results of the clustering of loglinear models method, the Alentejo could be considered as an homogeneous region relative drought characteristics. This conclusion, however does not mean the no-existence of variability in the region in terms of drought behavior.

Acknowledgments This work was partially supported by the research project PTDC/AGR-AAM/71649/2006 - Droughts Risk Management: Identification, Monitoring, Characterization, Prediction and Mitigation and by CMA/FCT/UNL under the project PEst-OE/MAT/UI0297/2011.

References

REFERENCES


Appendix A. Tables
Table A1. Estimators of quasi-association model parameters fitted to contingency tables and correspondent residual deviances for the 40 sites, grouped by sub-region.

<table>
<thead>
<tr>
<th>Lambda</th>
<th>Beta</th>
<th>Nisa</th>
<th>Cast</th>
<th>Vide</th>
<th>Póvoa</th>
<th>Figueira</th>
<th>Arronches</th>
<th>Aveiro</th>
<th>Palheir</th>
<th>Viana</th>
<th>Vila Nova</th>
<th>V. Vila</th>
<th>Vila Nova</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
Table A2  

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0.02</td>
<td>0.00</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>0.69</td>
<td>0.45</td>
<td>0.34</td>
<td>0.88</td>
<td>0.87</td>
<td>0.34</td>
<td>0.43</td>
<td>0.45</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>1.00</td>
<td>1.00</td>
<td>0.50</td>
<td>0.55</td>
<td>0.23</td>
<td>0.72</td>
<td>1.00</td>
<td>0.94</td>
<td>0.97</td>
<td>0.49</td>
</tr>
<tr>
<td>4</td>
<td>1.00</td>
<td>1.00</td>
<td>0.96</td>
<td>0.80</td>
<td>0.13</td>
<td>0.54</td>
<td>0.97</td>
<td>0.88</td>
<td>0.33</td>
<td>0.68</td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>1.00</td>
<td>0.59</td>
<td>0.73</td>
<td>0.25</td>
<td>0.79</td>
<td>0.84</td>
<td>0.71</td>
<td>0.26</td>
<td>0.73</td>
</tr>
<tr>
<td>6</td>
<td>1.00</td>
<td>1.00</td>
<td>0.93</td>
<td>0.70</td>
<td>0.76</td>
<td>0.95</td>
<td>0.21</td>
<td>0.02</td>
<td>0.46</td>
<td>0.55</td>
</tr>
<tr>
<td>7</td>
<td>1.00</td>
<td>1.00</td>
<td>0.93</td>
<td>0.70</td>
<td>0.76</td>
<td>0.95</td>
<td>0.01</td>
<td>0.02</td>
<td>0.46</td>
<td>0.55</td>
</tr>
<tr>
<td>8</td>
<td>1.00</td>
<td>1.00</td>
<td>0.93</td>
<td>0.70</td>
<td>0.76</td>
<td>0.95</td>
<td>0.02</td>
<td>0.02</td>
<td>0.46</td>
<td>0.55</td>
</tr>
<tr>
<td>9</td>
<td>1.00</td>
<td>1.00</td>
<td>0.93</td>
<td>0.70</td>
<td>0.76</td>
<td>0.95</td>
<td>0.02</td>
<td>0.02</td>
<td>0.46</td>
<td>0.55</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>1.00</td>
<td>0.93</td>
<td>0.70</td>
<td>0.76</td>
<td>0.95</td>
<td>0.02</td>
<td>0.02</td>
<td>0.46</td>
<td>0.55</td>
</tr>
</tbody>
</table>

For completion, please refer to Table A2 in the original document.
|   | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1 | 1.00 | 0.49 | 0.41 | 0.71 | 0.73 | 0.68 | 0.21 | 0.60 | 0.38 | 0.03 | 0.06 | 0.05 | 0.05 | 0.18 | 0.33 | 0.15 | 0.03 | 0.16 | 0.08 | 0.06 | 0.45 | 0.47 | 0.78 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| 2 | 1.00 | 0.85 | 0.85 | 0.05 | 0.50 | 0.15 | 0.98 | 1.00 | 0.01 | 0.41 | 0.54 | 0.80 | 0.98 | 0.88 | 0.99 | 0.98 | 0.31 | 0.99 | 0.95 | 0.66 | 0.98 | 0.52 | 0.76 | 0.21 | 0.05 | 0.11 | 0.96 | 0.79 | 0.92 | 0.89 |
| 3 | 1.00 | 0.91 | 0.91 | 0.73 | 0.89 | 0.21 | 0.95 | 0.93 | 0.47 | 0.12 | 0.74 | 0.09 | 0.65 | 0.58 | 0.67 | 0.66 | 0.46 | 0.74 | 0.95 | 0.91 | 0.97 | 0.62 | 0.98 | 0.75 | 0.06 | 0.78 | 0.72 | 0.93 | 0.94 | 0.72 | 0.43 | 0.63 | 0.99 | 0.08 | 0.65 | 0.95 | 0.81 | 0.47 |
| 4 | 1.00 | 0.92 | 0.92 | 0.60 | 0.50 | 1.00 | 0.92 | 0.33 | 0.06 | 0.86 | 0.87 | 0.75 | 0.80 | 0.76 | 0.75 | 0.76 | 0.43 | 0.86 | 0.88 | 0.96 | 0.98 | 0.84 | 0.91 | 0.84 | 0.58 | 0.89 | 0.67 | 0.86 | 0.68 | 0.88 | 0.34 | 0.90 | 0.92 | 1.00 | 0.31 | 0.71 | 0.72 | 0.31 |
| 5 | 1.00 | 0.58 | 0.83 | 0.89 | 0.70 | 0.11 | 0.01 | 0.30 | 0.92 | 0.45 | 0.80 | 0.33 | 0.46 | 0.58 | 0.42 | 0.68 | 0.97 | 0.93 | 0.98 | 0.97 | 0.99 | 0.60 | 0.95 | 0.89 | 0.48 | 0.74 | 0.45 | 0.95 | 0.38 | 0.43 | 0.66 | 0.94 | 0.28 | 0.20 | 0.73 | 0.18 | 0.01 | 0.05 |
| 6 | 1.00 | 0.09 | 0.60 | 0.57 | 0.13 | 0.02 | 0.14 | 0.96 | 0.19 | 0.18 | 0.37 | 0.70 | 0.60 | 0.76 | 0.84 | 0.77 | 0.34 | 0.41 | 0.62 | 0.37 | 0.53 | 0.36 | 0.09 | 0.21 | 0.54 | 0.86 | 0.49 | 0.41 | 0.53 | 0.05 | 0.01 | 0.07 | 0.02 | 0.13 | 0.09 | 0.43 | 0.91 | 0.50 | 0.31 | 0.44 | 0.77 | 0.80 | 0.44 | 0.01 | 0.02 | 0.41 |
| 7 | 1.00 | 0.97 | 0.58 | 0.15 | 0.81 | 0.89 | 0.84 | 0.84 | 0.86 | 0.84 | 0.86 | 0.55 | 0.93 | 0.52 | 0.97 | 0.90 | 0.75 | 0.93 | 0.79 | 0.68 | 0.91 | 0.86 | 0.83 | 0.83 | 0.90 | 0.42 | 0.93 | 0.97 | 1.00 | 0.44 | 0.77 | 0.80 | 0.44 | 0.01 | 0.02 | 0.41 | 0.91 | 0.50 | 0.31 | 0.44 | 0.77 | 0.80 | 0.44 | 0.01 | 0.02 | 0.41 |
| 8 | 1.00 | 0.86 | 0.41 | 0.62 | 0.90 | 0.98 | 0.91 | 0.96 | 0.99 | 0.98 | 0.43 | 0.99 | 0.94 | 0.69 | 0.99 | 0.48 | 0.86 | 0.53 | 0.95 | 0.54 | 0.59 | 0.93 | 0.99 | 0.93 | 0.17 | 0.96 | 0.95 | 0.96 | 0.83 | 0.82 | 0.97 | 0.41 | 0.01 | 0.02 | 0.41 | 0.91 | 0.50 | 0.31 | 0.44 | 0.77 | 0.80 | 0.44 | 0.01 | 0.02 | 0.41 |

Table A3: p-value matrices for the case of LRT relative to a subset of the model parameters.
Table A4. Configuration of the p-value matrix in case of existing the 4 sectors defined in Table 1 and Fig.2

|   | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  | 13  | 14  | 15  | 16  | 17  | 18  | 19  | 20  | 21  | 22  | 23  | 24  | 25  | 26  | 27  | 28  | 29  | 30  | 31  | 32  | 33  | 34  | 35  | 36  | 37  | 38  | 39  | 40  |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|