

TESTS FOR SPATIAL AND TEMPORAL CORRELATION IN MIXED MODELS FOR CLIMATE DATA

Mercedes ANDRADE-BEJARANO¹
Nicholas Tibor LONGFORD²
Diego Alejandro TOVAR-RIOS³

- **ABSTRACT:** *This paper reports the results of a simulation study of the likelihood ratio test of independence in models with spatial and temporal covariance structure for a given (fixed) regression part of a random coefficient model. Results show that the null distribution of the likelihood ratio test statistic is not distributed as a 50:50 mixture of the constant zero and χ^2 distribution with one degree of freedom.*
- **KEYWORDS:** *χ^2 distribution; likelihood ratio test; mixture of χ^2 ; restricted maximum likelihood.*

1 Introduction

A pair of mixed models with the same regression part but different covariance structures, one a special case of the other, are commonly compared by the likelihood ratio test (LRT) Verbeke & Molenberghs (2000). The two models, the null and the alternative, are fitted by maximum likelihood and the LRT statistic is derived from the values of the likelihood at the two maxima. The restricted maximum likelihood (REML) can be applied instead of the (full) maximum likelihood Patterson & Thompson (1971). Cox & Hinkley (1974) showed that under certain of regularity conditions the LRT statistic has a χ^2 distribution with degrees of freedom given by the difference in the number of parameters in the null and the alternative models.

¹Universidad del Valle, School of Statistics, Associate Professor, P.O. Box: 25360, Santiago de Cali, Valle del Cauca, Colombia. E-mail: mercedes.andrade@correounivalle.edu.co

²Universitat Pompeu Fabra, Department of Economic and Business, Academic Visitor, Barcelona, Spain. E-mail: sntlnick@sntl.co.uk

³Universidad del Valle, School of Statistics, MSc. Student, P.O. Box: 25360, Santiago de Cali, Valle del Cauca, Colombia. E-mail: diego.tovar@correounivalle.edu.co

One of these conditions requires that the null model be contained in an open subset of the alternative model. This is violated by some hypotheses of the covariance structure. For example, a hypothesis of interest may be that a variance σ^2 in the model is redundant, that is $\sigma^2 = 0$; then the null hypothesis is a boundary point of the alternative.

Stram & Lee (1994) used the results from Self & Liang (1987) on nonstandard testing situations and showed that the asymptotic null distribution of the LRT statistic for testing the hypothesis that $\sigma^2 = 0$ is often a mixture of χ^2 distributions rather than a single χ^2 distribution. This result was obtained under the assumption of conditional independence, that the random effects are mutually independent. There are no corresponding results when the effects are correlated.

In this paper, we study models with complex covariance structure for a time series dataset of monthly average temperatures from $I = 28$ sites, collected over the period from 1971 to 2002 ($K = 32$ years), in Valle del Cauca, Colombia, South America. The sites are located at latitudes from $3^\circ 19'N$ to $4^\circ 44'N$, longitudes from $75^\circ 49'W$ to $76^\circ 45'W$ and altitudes of 920 to 1950 meters above sea level (m). Fifteen sites are located in the valley, at altitudes up to 1100 m, and thirteen in the mountains, at altitudes above 1200 m.

We identify two sources of variation in the data, one related to the temporal aspects of the records, and one to the spatial aspects (locations of the sites). In the former, we distinguish three features Andrade-Bejarano (2013):

1. a trend of increasing average temperatures over time at some sites (Figure 1),
2. seasonal variation in monthly average temperatures, with two dry periods (January – February and July – August) and two wet periods (April – May and October – November) (Figure 2), and
3. temporal phenomena, the "El Niño" and "La Niña". The "El Niño" is associated with elevated monthly average temperatures, and the "La Niña" with reduced averages (Figure 3) Andrade-Bejarano (2009).

In the spatial variation, we identify two features:

1. the weather patterns in the valley and in the mountains (different altitudes) differ; the monthly average temperatures tend to be lower at higher altitudes, and
2. the sites at close proximity to one another are more likely to have similar values (and patterns of values) than sites further apart.

The sources of variability motivate our choice of mixed models Henderson (1982). In these models, years and sites are represented by random effects and altitude, indicators of the ENSO (El Niño Southern Oscillation Index) and the geographical position by fixed effects.

In the models we consider, we want to assess the statistical significance of spatial or temporal aspects. To reach this objective, we derive the null-distribution

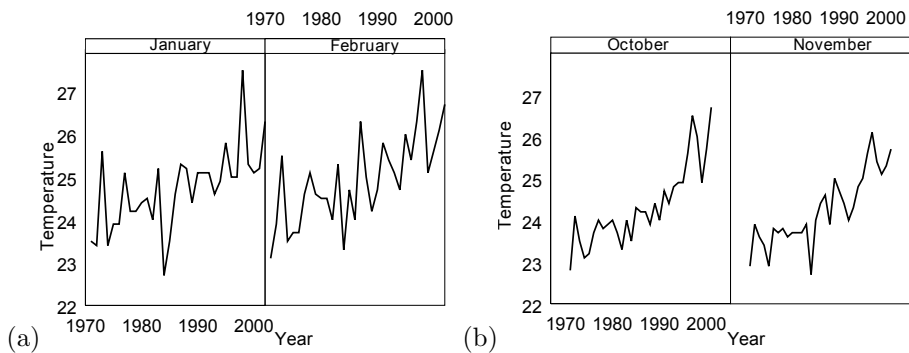


Figure 1 - Time series plots of average monthly temperatures ($^{\circ}\text{C}$) for site Ingenio Central Castilla, located in the valley at the altitude of 1040 m, for (a) January and February (b) October and November.

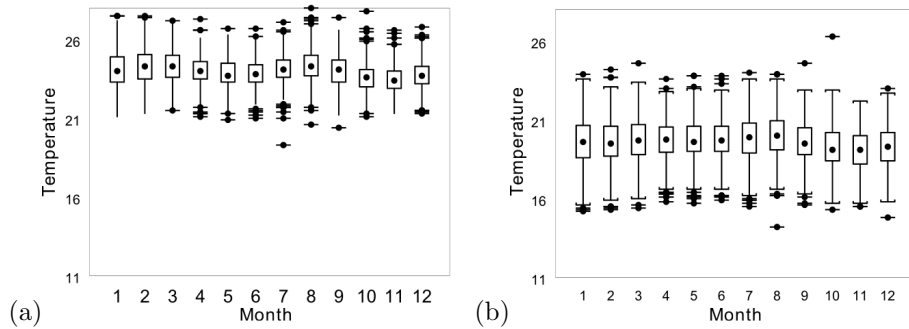


Figure 2 - Box plots of monthly average temperatures ($^{\circ}\text{C}$) for sites in the valley (a) and the mountains (b).

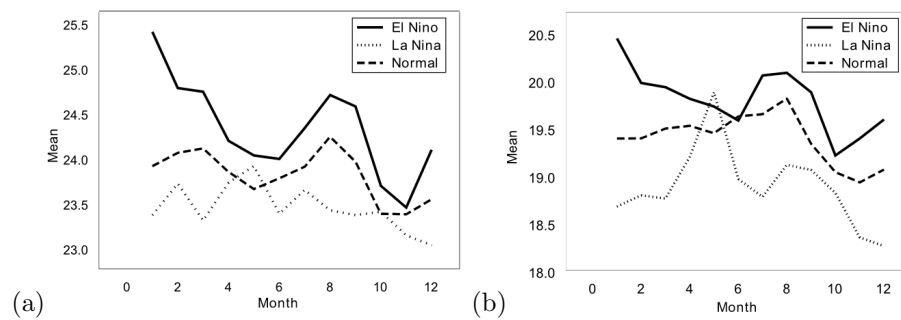


Figure 3 - Graphs of mean values of monthly average temperature ($^{\circ}\text{C}$) for the three phenomena by month, for (a) the valley and (b) mountains. Means obtained over all meteorological stations.

of the LRT statistic for the comparison of a mixed model with spatially and temporally correlated errors with its submodel in which one or several variances are constrained to zero. The two models have the same regression part. They are fitted by REML and the value of the LRT statistic is obtained from the fits of these models.

2 Simulation study

In the simulations, we use four models for the monthly average temperatures. These models, selected by Andrade-Bejarano (2013) using LRT, are

$$y_{ik} = \sum_{m \in V_h} \beta_m x_{mik} + \gamma_{1i} + \gamma_{2k} + \gamma_{3i} x_{3ik} + \varepsilon_{ik}, \quad (1)$$

where $\gamma_{1i}, \gamma_{2k}, \gamma_{3i}$ and ε_{ik} , $i = 1, \dots, I$ and $k = 1, \dots, K$ are random effects; γ_{1i} is associated with the intercept for site i , γ_{3i} with the slope on time (in years) for site i , γ_{2i} with the random effect for year k and ε_{ik} is the residual term for site i and year k . We consider models M1–M4 defined by the sets of variables V_h

- $V_1 = (0, 1, 2, 3, 9)$,
- $V_2 = (0, 1, 3, 4, 9)$,
- $V_3 = (0, 1, 3, 5, 6, 9)$,
- $V_4 = (0, 1, 3, 7, 8, 9)$,

In these models, index 0 corresponds to the intercept ($x_{0ik} \equiv 1$) and the covariates 1–9 are listed in Table 1. We assume that the site-related random vectors γ_{1i}, γ_{3i} are independent of the time-related effect γ_{2k} and they are both independent of the residual terms ε_{ik} . Further,

$$\begin{pmatrix} \gamma_{1i} \\ \gamma_{3i} \end{pmatrix} \sim N \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\gamma_1}^2 & \sigma_{\gamma_1 \gamma_3} \\ \sigma_{\gamma_1 \gamma_3} & \sigma_{\gamma_3}^2 \end{pmatrix} \right\},$$

$\gamma_{2k} \sim N(0, \sigma_{\gamma_2}^2)$ and $\varepsilon_{ik} \sim N(0, \sigma^2)$. We attach to M1–M4 the symbol S or T to indicate the spatial and temporal versions/adaptations of these models.

The alternative models are given by the same sets of variables V_h , $h = 1, \dots, 4$, but for the vector ε of the IK random terms ε_{ik} we assume that $\varepsilon \sim N(0, \sigma^2 F)$, where F is an isotropic covariance function Cressie (1993); Chilès & Delfiner (1999). In this case, F corresponds to the two isotropic covariance functions used for modelling the temporal and the spatial covariance structure:

1. Gaussian model; $\text{cov}(\varepsilon_{ik}, \varepsilon_{i'k}) = \sigma^2 \exp(-d_{ii'}^2 / \rho_s^2)$,
2. exponential model; $\text{cov}(\varepsilon_{ik}, \varepsilon_{i'k}) = \sigma^2 \exp(-t_{kk'} / \rho_t)$,

where $d_{ii'}$ is the distance between sites i and i' , $t_{kk'}$ is defined similarly for years k and k' , and ρ_s and ρ_t are the respective ranges of the spatial and temporal variograms Chilès & Delfiner (1999); Isaaks & Srivastava (1989). These models were selected by LRT and the selection checked by empirical variogram. For a given set of fixed effects in the models with covariates V_h , $h = 1, \dots, 4$, we test the hypotheses of no spatial and no temporal correlation. The null hypotheses are $\rho_s = 0$ and $\rho_t = 0$. We obtain the null distributions for these tests by simulations. We do not consider models in which spatial and temporal correlation are combined, as we regard them as too complex and difficult to fit.

Table 1 - Variables and fixed parameters in the monthly average temperature dataset

m^*	Parameter	Variable	Acronym	V_1	V_2	V_3	V_4
0	β_0	Intercept		x	x	x	x
1	β_1	Altitude (in meters)		x	x	x	x
2	β_2	Southern Oscillation Index	SOI**	x			
3	β_3	Year (centered)		x	x	x	x
4	β_4	Southern Oscillation Index two month ago	SOI.LAG2***		x		
5	β_5	Dummy variable****				x	
6	β_6	Dummy variable****				x	
7	β_7	Dummy variable*****					x
8	β_8	Dummy variable*****					x
9	β_9	Geographical position (valley vs. mountain): Dummy variable		x	x	x	x

Source: Andrade-Bejarano (2013); * index in equation (1); ** The Southern Oscillation Index for the current month was included as a measurement of the ENSO phenomenon; *** The SOI lag 2 (*SOI.LAG2*) from the two previous months was included, as suggested by Madl (2000). He indicates that the normal weather pattern in the months prior to an El Niño event usually breaks down. For some reason not yet well understood, the westward atmospheric pressure gradient decreases; ****Dummy variables for the ENSO in the current month: El Niño : $x_{5k} = 1$, $x_{6k} = 0$; La Niña: $x_{5k} = 0$, $x_{6k} = 0$; Normal: $x_{5k} = 0$, $x_{6k} = 1$; *****Dummy variables for the ENSO two months' lag: El Niño in the two lag months : $x_{7k} = 1$, $x_{8k} = 0$; La Niña in the two months' lag: $x_{7k} = 0$, $x_{8k} = 0$; Normal conditions in the two months' lag: $x_{7k} = 0$, $x_{8k} = 1$.

The simulations are based on the data for two months, February, the driest month in the study zone, and November, the wettest, Andrade-Bejarano (2009). They are conducted in SAS Version 9.1, using the procedure MIXED SAS (2004). For each one of the four models and the two months, we simulate 3000 sets of the outcomes using the estimates of the parameters $\sigma_{\gamma_1}^2$, $\sigma_{\gamma_2}^2$, $\sigma_{\gamma_3}^2$, $\sigma_{\gamma_{13}}$, σ^2 and the vector β , under the null hypothesis that F is the identity matrix.

The sample sizes are $n = 524$ for February and $n = 518$ for November. The SAS procedure used requires as input a set of values for each parameter; the

procedure selects their best configuration for the initial solution. These input values are set by trial and error, taking into account the parameter estimates in the models with independent errors. For example, for model V_1 with spatial dependence of ε , the input values of the variance $\sigma_{\gamma_2}^2$ are set to 0.09, 0.29 and 0.49 and the values of ρ_s set to 1, 2, ..., 7, for February. For November, the same values are used for ρ_s , but values 0.07, 0.27 and 0.47 are used for $\sigma_{\gamma_2}^2$. The procedure implements the Newton-Raphson algorithm to maximise the restricted log-likelihood l_R SAS (2004). Note that the same set of values is used in the 3000 replications; the sets differ slightly across the models, months, and correlation structures for ε .

The null and alternative models are fitted for each simulated dataset and the LRT statistic is obtained as:

$$\Delta l = 2(l_{R1} - l_{R0}),$$

where l_{R0} and l_{R1} are the values of the restricted log-likelihood at the REML fit of the respective null and alternative models. The distribution of the LRT statistic under the hypothesis is compared with the distribution of 50:50 mixture of the constant zero and χ_1^2 distribution. It is practical to compare the frequency of zeros with 0.50, and, separately, the distribution of the positive LRT values with the χ_1^2 distribution.

3 Results

This section describes of the 3000 sets of simulated results related to modelling weather pattern data in Valle del Cauca.

Figures 4 and 5 display the histograms of the positive estimates obtained in the simulations for each of the eight settings (four models by two months). The histograms for the spatial models (M1S–M4S) are in the left-hand column and those for the temporal models (M1T–M4T) in the right-hand column. The subtitle of each panel contains information about the percentage of replicates in which convergence was not reached and the percentage of zero estimates in the remainder. Thus, for M1S in February 11.5% of replicates (345 out of 3000) failed to converge, and 1394 of the remaining 2655 replications (52.5%) yielded variance estimate equal to zero.

The histograms are displayed with scaling for the density on the vertical axis, and all of them have the same range. The density of the χ_1^2 distribution is added by a solid line in each panel for comparison. The diagram does not have a fine resolution for large values (say, greater than 5.0), but there is a discernible excess of simulated values in the range 1.0–4.0 over the expected, with the sole exception for M3T (temporal model M3). Although the largest values of the estimates are close to 15, we curtail the horizontal axes to 8.0, to improve the resolution of the diagrams. The percentage of values that exceed 8.0 is smaller than 1.0%.

The results for November are very similar, although the rate of non-convergence is lower of every model. The probability of zero estimate is greater

February

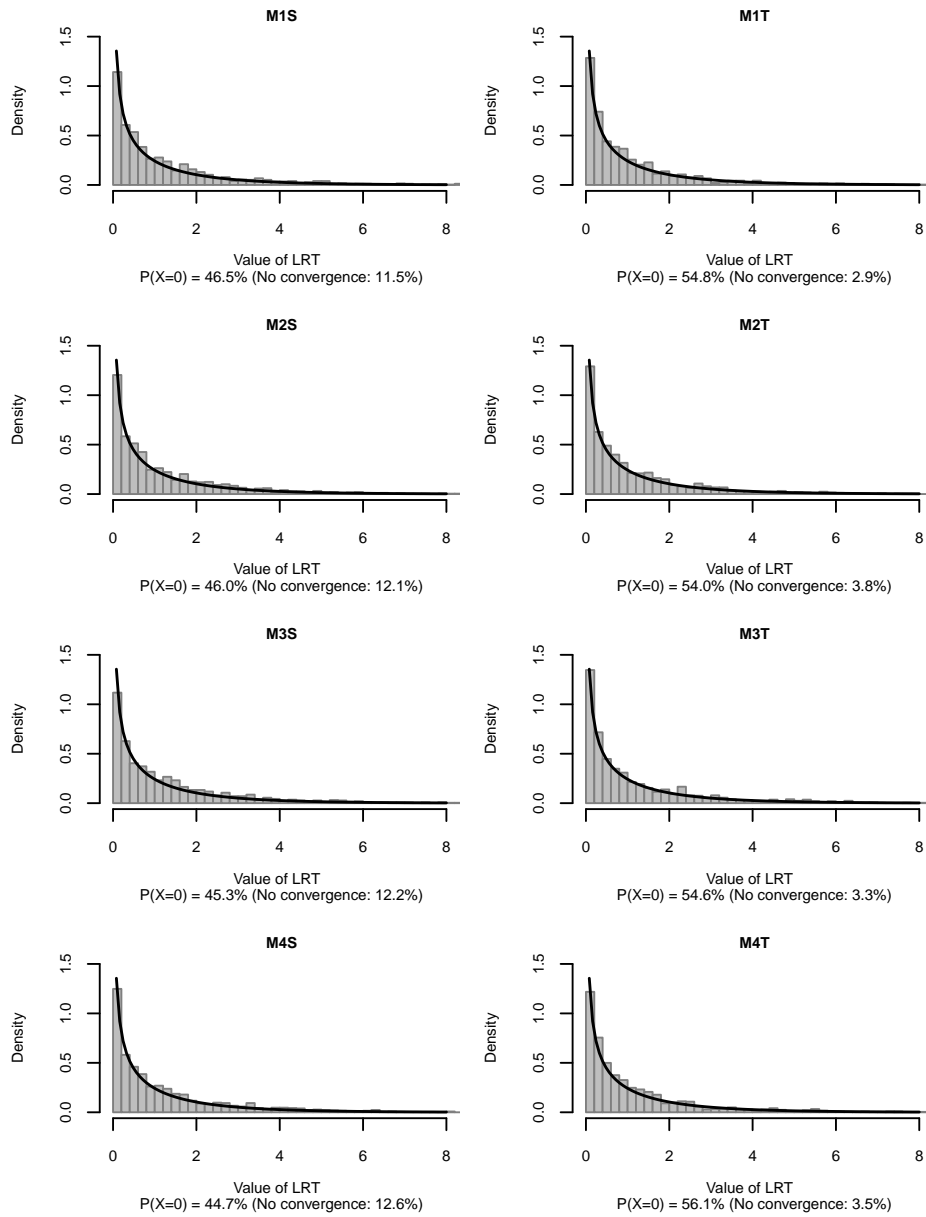


Figure 4 - The empirical and theoretical (χ_1^2 distributed) distributions of the positive estimates of the station-level variance; February.

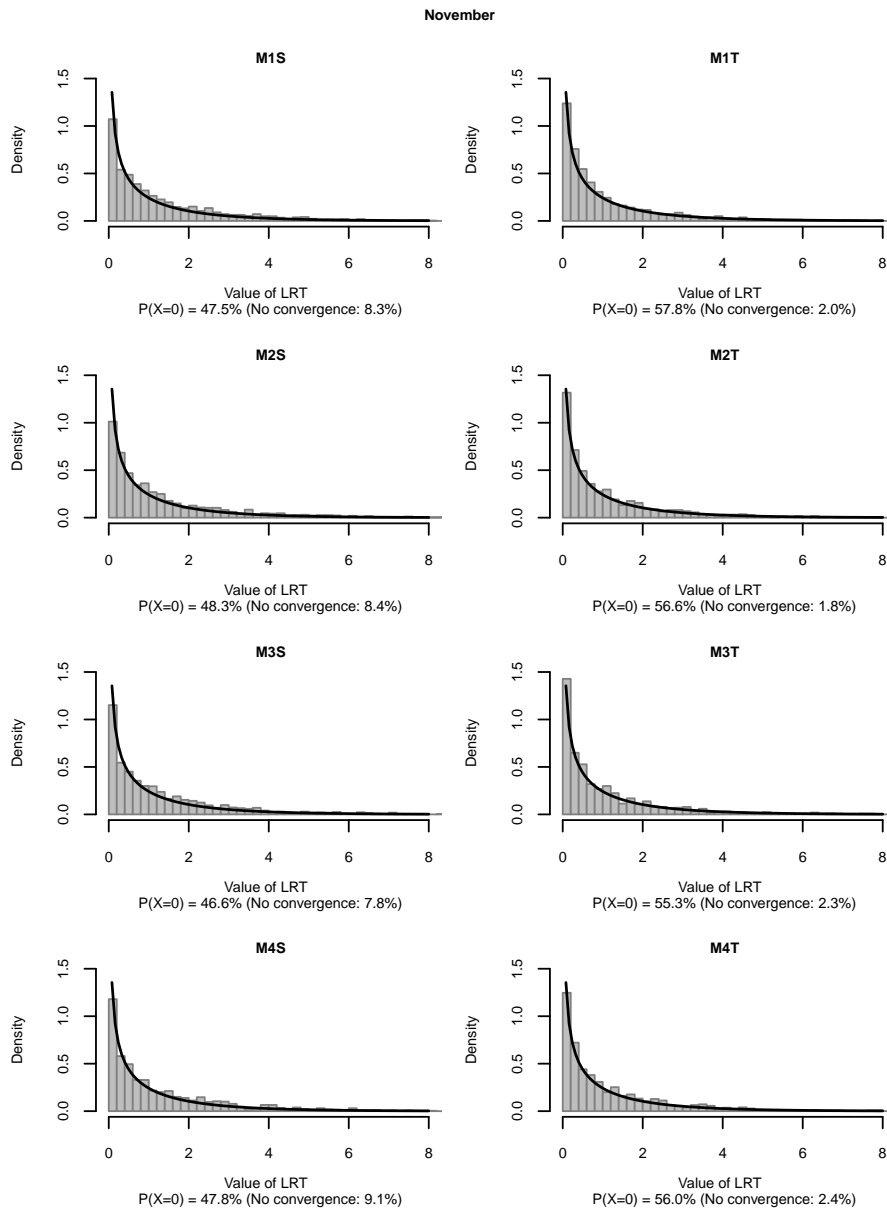


Figure 5 - The empirical and theoretical (χ_1^2 distributed) distributions of the positive estimates of the station-level variance; November.

than 0.5 for all cases (models and months); for the spatial models it is greater by only up to 0.03, whereas for the temporal models it is greater by 0.06–0.09. The simulated distribution of the positive variance estimates can be compared with χ_1^2 also by the Q-Q plot. These are displayed in Figure 6 in sets of four in a panel (for spatial and temporal models by the two months).

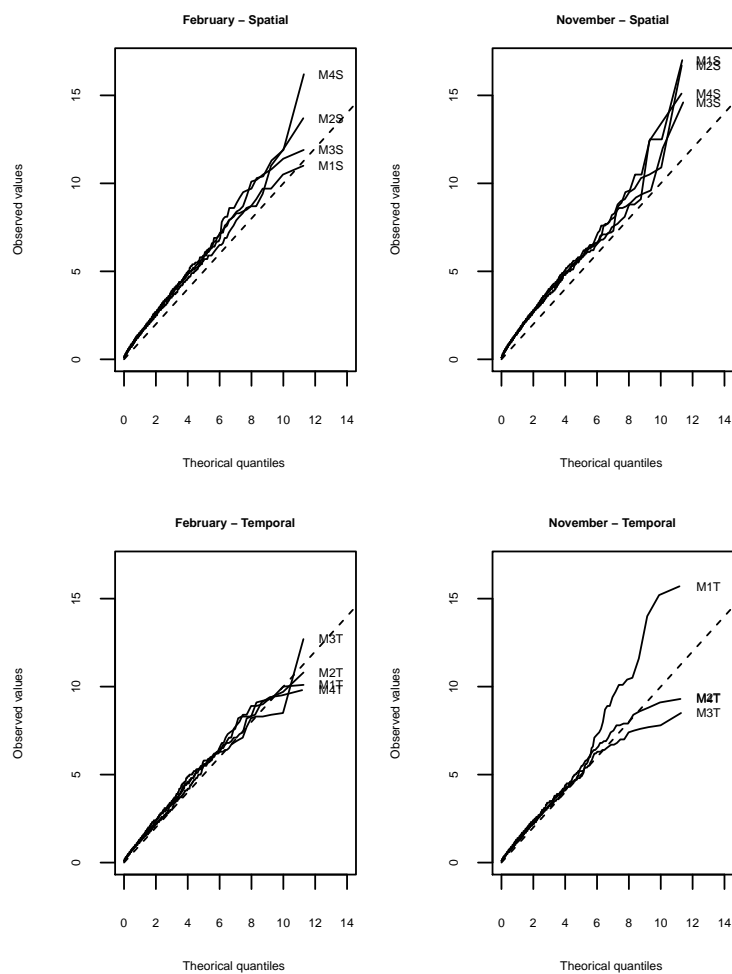


Figure 6 - The Q-Q plots of the the positive estimates of the station-level variance, compared with χ_1^2 distribution.

A Q-Q plot for a set of replicate estimates is generated as follows. The estimates are sorted in increasing order of magnitude. Suppose their number is n .

Then they are plotted against the $i/(n+1)$ -quantile ($i = 1, \dots, n$) of the target (χ_1^2) distribution. The replicates are in agreement with the target distribution if the plot deviates from the identity line, drawn by dashes, insubstantially. To the contrary, we see that the plots for all 16 cases deviate substantially and systematically from the identity line. We note that a Q-Q plot tends to exaggerate the deviations for the few highest values, and so such deviations should be discounted. However, the plotted curves for spatial models are with one exception (for a maximum value) entirely above the identity line. For temporal models, there is a closer agreement for November for much of the distribution, but the right-hand tails of the empirical distribution deviate from the target substantially. Sets of four Q-Q plots are condensed into a single panel to save space and to have all the 16 curves on a single page.

Finally, the deviation of the simulated values from the target is reflected in the empirical means and variances of the positive estimates. These are listed in Table 2. The empirical means are consistently greater than unity, the expectation of χ_1^2 , and most of the empirical variances are much greater than 2.0, the variance of χ_1^2 .

Table 2 - The empirical means and variances of the positive variance estimates

Model	<i>Spatial</i>				<i>Temporal</i>			
	M1S	M2S	M3S	M4S	M1T	M2T	M3T	M4T
<i>February</i>								
Mean	1.35	1.31	1.40	1.37	1.16	1.23	1.24	1.20
Variance	2.51	2.39	2.65	2.84	1.98	2.20	2.24	2.13
<i>November</i>								
Mean	1.44	1.43	1.40	1.38	1.21	1.18	1.14	1.22
Variance	2.77	2.82	2.53	2.61	2.69	1.99	1.82	2.03

Thus, the simulated values of the variance estimates are not distributed as a mixture of a χ_1^2 distribution and the constant zero. The probabilities of zero are quite close to 0.5, but are consistently greater.

4 Discussion

We carried out a simulation study to assess the statistical significance of spatial and temporal correlation in the models M1–M4. Because the null distribution of the LRT statistic is not known and the standard asymptotic theory does not apply for models with spatial and temporal correlated errors, we derived it by simulation. The results obtained show that the null distribution of the LRT statistic is not distributed as a 50:50 mixture of the constant zero and χ_1^2 . This result agrees with Pinheiro & Bates (2000) who found by simulation that the LRT obtained by

comparing two models followed a 65:35 mixture of the constant zero and χ_1^2 in the case of maximum likelihood (ML) method and not a 50:50 mixture. This is not in agreement with Stram & Lee (1994). We concur that the adjustment suggested by Stram & Lee (1994) is not always successful.

The LRT statistic is evaluated by maximising the restricted loglikelihood under the hypothesis and alternative, l_{R0} and l_{R1} , respectively, which yields positive (semi) definite matrices D (random-effects covariance matrix) and Σ (residual covariance matrix) Verbeke & Molenberghs (2000). Non-convergence observed in a small fraction of the simulation is caused by flat or ridged likelihood surfaces SAS (2004) or because the Newton-Raphson procedure oscillates between two solutions. The SAS output does not give the reason for non-convergence. Morrell (1998) indicates some situations in which SAS procedure does not converge:

1. when it cannot find the maximum likelihood value of a positive semidefinite matrix D lying on the constraint surface, and
2. when redundant random effects are included, and the likelihood surface is quite flat.

Greven et al. (2008) found that in many simulations the values of the restricted LRT statistic are smaller than or equal to zero. They explain that this is probably due to internal convergence criteria, multiple local maxima of the likelihood, and they point out that this problem is especially serious in SAS. A recommendation for the future is to use a software other than SAS for similar simulation studies and to record the reasons for non-convergence.

We repeated some of the simulations in R using the `nlme` package and obtained very similar results to their counterpart from SAS. We analysed separately the spatial and the temporal correlation in the errors. A spatio-temporal covariance structure in the errors is not developed. An initial exploration of the spatio-temporal variograms of monthly average temperature was carried out, following the criterion of Cressie & Huang (1999). Spatial variograms for the first time lag show a concave form and spatial continuity behaviour Andrade-Bejarano (2009). Cressie (1993), Gneiting (2002), Ma (2003a), Ma (2003b), Kolovos et al. (2004), Stein (2005) and others have contributed to the derivation of spatio-temporal covariance functions. Banerjee et al. (2004) discuss the nonseparable stationary covariance functions developed by Cressie (1993), Gneiting (2002) and Stein (2005). They indicate the difficulties in the computation of $\text{cov}(s - s', t - t')$. Similarly, Stein (2005) states the need to develop theoretical frameworks for the models, algorithms for fitting them, to compare models and diagnostics to assess the adequacy of the models. At present, no software satisfies this standard, and in the papers reviewed the authors do not specify how they fit spatio-temporal covariance structures.

ANDRADE-BEJARANO, M.; LONGFORD, N. T.; TOVAR-RIOS, D. A. Teste para correlação espacial e temporal para dados climáticos usando modelos mistos. *Rev. Bras. Biom.*, São Paulo, v.32, n.1, p.130-142, 2014.

- RESUMO: Neste artigo apresentamos os resultados de um estudo de simulação para um teste da razão de verossimilhança de independência em modelos com estrutura de covariância espacial e temporal nos erros para a parte dada (fixa) da regressão em um modelo de coeficientes aleatórios. Com os resultados obtidos observou-se que a distribuição sob a hipótese nula do teste não se comporta como uma mistura 50:50 entre a constante zero e uma χ^2 com um grau de liberdade.
- PALAVRAS-CHAVE: Distribuição χ^2 ; teste da razão de verossimilhança; mistura de χ^2 ; máxima verossimilhança restrita.

References

- VERBEKE, G.; MOLENBERGHS, G. *Linear Mixed Models for Longitudinal Data*. Springer Verlag, 2000.
- PATTERSON, H.; THOMPSON R. Recovery of inter-block information when block sizes are unequal. *Biometrika*, v.58, n.545–554, 1971.
- COX, D.; HINKLEY, D. *Theoretical Statistics*. Chapman and Hall, London, 1974.
- STRAM, D.; LEE, J. Variance components testing in the longitudinal mixed effects model. *Biometrics*, v.50, n.4, p.1171–1177, 1994.
- SELF, S.; LIANG, K. Asymptotic properties of maximum likelihood estimators and likelihood ratio test under nonstandard conditions. *Journal of the American Statistical Association*, v.82, p.605–610, 1987.
- ANDRADE-BEJARANO, M. Montly average temperature modelling in an intertropical region. *Weather and Forecasting*, v.28, n.5, p.1099–1115, 2013.
- ANDRADE-BEJARANO, M. Monthly average temperature modelling for Valle del Cauca (Colombia). Unpublished PhD. Thesis in Applied Statistics. The University of Reading, Statistics Section, United Kingdom, 2009.
- HENDERSON, C, R. Analysis of covariance in the mixed model: Higher-level, nonhomogeneous, and random regressions. *Biometrics*, v.38, n.3, p.623–640, 1982.
- CRESSIE, N. *Statistics for Spatial Data*. John Wiley and Sons, 1993.
- CHILÈS, J.; DELFINER, P. *Geostatistics. Modelling Spatital Uncertainty*. John Wiley and Sons, New York, 1999.
- ISAAKS, E.; SRIVASTAVA, R. *An Introduction to Applied Geostatistics*. Oxford University Press, 1989.
- MADL, P. The El Niño (ENSO) phenomenon. *Environmental Physics / Letter VO*, 2000, p.437–503.

- SAS. *SAS/STAT. User's Guide. Version 9.1.*, Cary, NC: SAS: Institute Inc., 2004.
- PINHEIRO, J. C.; BATES, D. M. *Mixed effects models in S and S-PLUS*. Springer Verlag, 2000.
- MORRELL, C. Likelihood ratio testing of variance components in the linear. mixed-effects model using restricted maximum likelihood. *Biometrics*, v.54, n.4, p.1560–1568, 1998.
- GREVEN, S.; CRAINICEANU, C.; KUCHENNOFF, H.; PETERS, A. Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of the Computation and Graphical Statistics*, v.17, n.4, p.870–891, 2008.
- CRESSIE, N.; HUANG, H. Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, v.94, n.448, p.1330–1340, 1999.
- GNEITING, T. Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, v.97, p.590–600, 2002.
- MA, C. Families of spatio-temporal stationary covariance models. *Journal of Statistical Planning and Inference*, v.116, p.489–501, 2003a.
- MA, C. Spatio-temporal stationary covariance models. *Journal of Multivariate Analysis*, v.86, p.97–107, 2003b.
- KOLOVOS, A.; CHRISTAKOS, G.; HRISTOPULOS, D.; SERRE, M. Methods for generating non-separable spatiotemporal covariance models with potential environmental applications. *Advances in Water Resources*, v.27, p.815–830, 2004.
- STEIN, M. Space-time covariance functions. *Journal of the American Statistical Association*, v.100, n.469, p.310–320, 2005.
- BANERJEE, S.; CARLIN, B.; GELFAND, A. *Hierarchical Modelling and Analysis for Spatial Data*. Chapman and Hall, 2004.

Received in 20.07.2013.

Approved after revised in 13.03.2014.