AUSTRIAN JOURNAL OF STATISTICS
Volume 39 (2010), Number 3, 203–221

Outliers in Mixed Models for Monthly Average Temperatures

Mercedes Andrade-Bejarano[1] and Nicholas T. Longford[2]

[1] School of Industrial Engineering and Statistics, Universidad del Valle, Cali, Colombia, and University of Reading, United Kingdom

[2]SNTL and Department of Economics and Business, University Pompeu Fabra, Barcelona, Spain

Abstract: Long-term series of monthly average temperatures taken at 28 sites in Valle del Cauca, Colombia, are studied. Mixed models are applied to cater for the within- and between-site variation. Outliers are inevitable in such studies, due to faulty equipment, slip-ups in the recording process, or unusual weather patterns. We apply a simulation-based approach to the assessment of the outlier status of suspected observations. It is a method based on graphical comparisons of user-defined features, related to large residuals, in the real and simulated data sets. Robustness in the identification of the outliers is achieved by applying the procedure with several alternative models. The impact of the identified outliers is assessed. Two meteorological stations, Zaragoza and Monteloro, are identified as having many outliers, so that all the data from them should be discarded.

Zusammenfassung: Ausreißer in Wetterdaten sind keine Seltenheit, und die häufigsten Ursachen für deren Auftreten sind fehlerhafte Geräte, die fehlerhafte Erfassung der Daten durch das Personal, oder echte, aber ungewöhnliche, Wetterlagen. Wir schlagen die Verwendung von Mixed Models vor, um die Variabilität der Daten zwischen und innerhalb von Wetterstationen zu untersuchen. Die Unterscheidung zwischen realen Extremwerten und fehlerhaft erhobenen Werten erfolgt durch eine simulationsbasierte Methode. Diese Methode basiert auf graphischen Vergleichen von Eigenschaften der Residuen, nachdem die Daten mit mehreren Modellen analysiert wurden wodurch unser Ansatz auch einen gewissen Grad an Robustheit erhält. Die Analyse des Datensatzes Monatliche Durchschnittstemperaturen, die langzeitlich an 28 Stationen im Valle del Cauca (Kolumbien) erhoben wurden brachte zum Vorschein, dass die Daten zweier Wetterstationen (Zaragoza und Monteloro) viele als Ausreißer zu klassifizierende Werte enthalten. Aufgrund dessen empfehlen wir den Ausschluss dieser Stationen von jedweder Analyse.

Keywords: Influence; Long-term time series; Meteorological station; Outlier detection; Residual; Residual variance.

1 Introduction

One of the goals of the Colombian Institute for the Agricultural and Farming Research (CORPOICA) is the comprehensive estimation of climatic and soil variables throughout the country, even at locations where they are not measured directly. Climatic and

soil factors influence the growth of crops and determine the distribution of species (both wildlife and related to agricultural activity) and their adaptation to the changing conditions. (Atherton and Rudich, 1986; Jones, 2000; Villareal, 1980). Our research is connected with this general agenda, together with studying the long-term changes of the climate. Inferences about such changes are difficult to make because of the natural (inexplicable) variability of the weather, not only in the short term (days and weeks), but also across seasons and from year to year. Imperfect measurement and recording are another cause of complications. This article is concerned with a preparatory stage for an analysis of long-term weather patterns, in which we study outliers.

We analyse the monthly average temperatures recorded at a network of meteorological stations (henceforth *sites*) at Valle del Cauca, Colombia (see Figure 1), in the period 1971 – 2002. The 28 sites in the study zone are identified by their names and locations. Their altitudes are also given. The sites are located at latitudes from 3°19'N to 4°44'N, longitudes from 75°49'W to 76°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 1233 m.

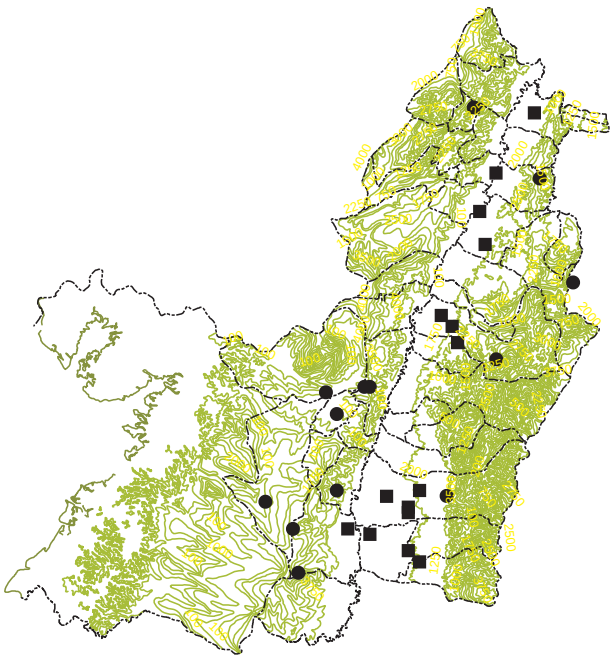

Figure 1: Location of meteorological stations (sites) in Valle del Cauca, Colombia; • — sites located in the mountains; filled box □ — sites located in the valley

In the data, two sources of variation can be identified, one related to the temporal aspects of the records, and one to the spatial aspects (locations of the sites). In the former with distinguish three aspects:

1. the time series for some sites show a trend of increasing average temperatures, see Figure 2;
2. seasonal variation in monthly average temperatures, with two dry periods (Jan-

uary – February and July – August) and two wet periods (April – May and October – November) (Figure 3); 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, (Andrade, 2009).

In the spatial variation, two aspects can be identified:

1. the weather patterns in the valley and in the mountains, or with the altitude, differ; the monthly average temperatures tend to be lower at higher altitude (Figure 4); 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.

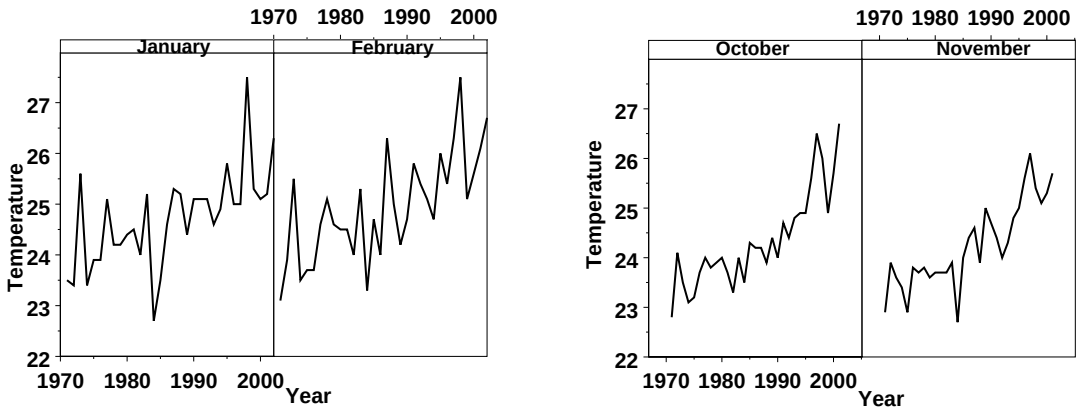

Figure 2: Time series plots of average monthly temperatures (° C) for site Ingenio Central Castilla, located in the valley at the altitude of 1040 m, for January and February (left), and for October and November (right)

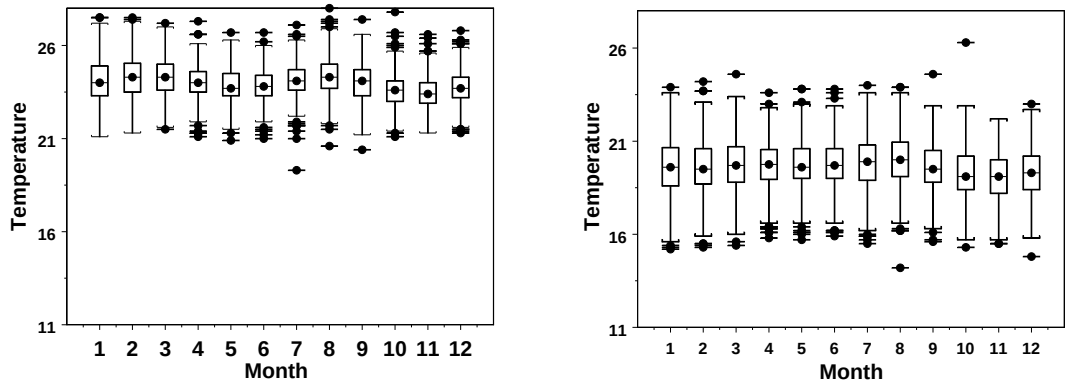

Figure 3: Box plots of monthly average temperatures (° C) for sites in the valley (left) and in the mountains (right)

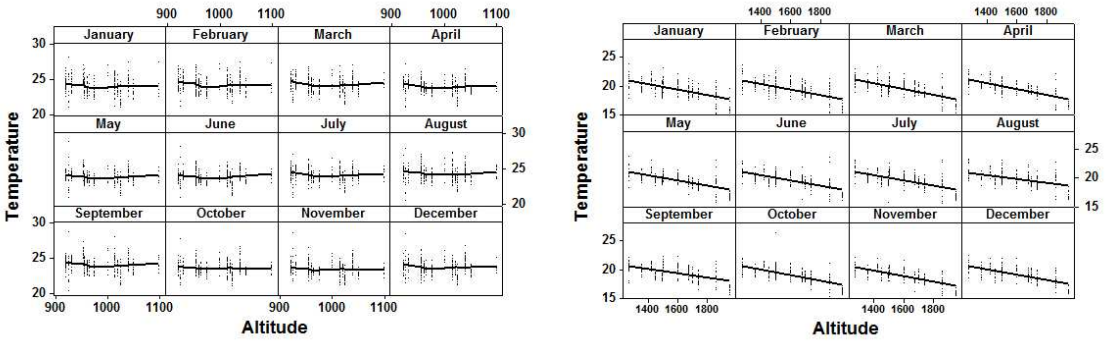

Figure 4: Monthly average temperature (° C) and altitude (m), by month, for sites located in the valley (left) and in the mountains(right)

These sources of variability motivate our choice of mixed models (Henderson, 1982), in which the years and sites are associated with random effects and altitude and indicators of the "El Niño" and "La Niña" phenomena with fixed effects.

The linear mixed model (Laird and Ware, 1982) has the form

$$y_i = X_i\beta + Z_i\gamma_i + \epsilon_i\,, \tag{1}$$

where y_i is the $n_i \times 1$ response vector for subject (individual or cluster) i; X_i is an $n_i \times p$ matrix of the explanatory variables, (covariates); β is the corresponding $p \times 1$ vector of regression parameters; Z_i is an $n_i \times q$ matrix associated with random effects; γ_i is a $q \times 1$ vector of cluster-level random effects; and ϵ_i is an $n_i \times 1$ vector of elementary-level random errors. The matrices X_i and Z_i are observed completely, and at the outset we assume that so are the vectors y_i. The total number of observations is $N = n_1 + \cdots + n_m$, where m is the number of clusters (sites). The $N \times p$ regression design matrix X is formed by vertical stacking of the matrices X_i, $i = 1, \ldots, m$. The variation design matrix Z is composed of the diagonal blocks Z_i, $i = 1, \ldots, m$.

After fitting some of the models we find large residuals, which can reasonably be regarded as outliers, although we prefer to deal with the issue of assigning the outlier status more formally, both to deal with cases that are not obvious, and to avoid being too liberal or too conservative with the assignment of the status. We apply a simulation-based analysis of outliers; for a similar application, see Longford (2001), and for a brief sketch, Longford (1998). The theoretical basis of the method is developed by Rubin (1984).

Hadi and Simonoff (1993) define an outlier as an observed unit which, if excluded from the data set, would yield a much better fit of the adopted model. Hawkins (1980) points out that an intuitive definition of an outlier would be "an observation which deviates so much from other observations as to arouse suspicion that it may have been generated by a different mechanism". Longford (2001) indicates that in ordinary regression $Y = X\beta + \varepsilon$, the quality of the fit is closely linked to the absolute sizes of the residuals $\hat{\epsilon} = Y - X\hat{\beta}$, so outliers are identified among the largest absolute residuals $|\hat{\epsilon}|$. He interprets an outlier as an observation that "spoils" the model assumptions; if this observation were removed from the data set, the estimated residual variance of the ordinary regression model, $\widehat{\mathrm{var}}(\epsilon)$,

would be reduced much more than if an observation that is in accord with the model were deleted.

An observation may be an outlier for the model without a variable included in X, but not be an outlier when the variable is included.

Many diagnostic procedures have been developed for linear regression models, for example (Verbeke and Molenberghs, 2000) and (Atkinson, 1985).

However, the key definitions, of the residual (Cook and Weisberg, 1982), leverage (Belsley, Kuh, and Welsch, 1980; Demidenko, 2004; Schanberger, 2004), and Cook distance (Cook and Weisberg, 1982), do not have any obvious generalisations to linear mixed models. Tan (Personal communication, 19 March, 2007), Verbeke (Personal communication, 19 March, 2007) and Haslett, (Personal communication, 19 March, 2007) indicate that there are no valid tests for outliers and influential observations (Belsley et al., 1980) at present. One reason that hinders this generalisation is that there is no unique definition of residuals for the clusters in linear mixed models given by (1).

Some of the alternative definitions are formulated by Haslett and Haslett (2007) and Verbeke and Molenberghs (2000):

1. the marginal residual, that is, the difference between the (observed) data and the estimated (marginal) mean,
$$r_{\mathrm{mi}} = y_i - X_i\hat{\beta}\,; \tag{2}$$

2. the conditional residual, defined as the difference between the (observed) data and the predicted value of the observation,
$$r_{\mathrm{ci}} = y_i - X_i\hat{\beta} - Z_i\hat{\gamma}_i\,, \tag{3}$$
where $X_i\hat{\beta} + Z_i\hat{\gamma}_i$ is the conditional mean of y_i;
3. the estimated random effect γ_i can also be regarded as a residual since it reflects the deviation of the specific profiles from the population average profile.

The leverage, which in ordinary regression depends solely on the regression design matrix X, does not have a straightforward extension to linear mixed models in which there are two matrices relevant to the concept of leverage: X and the variation design matrix Z. The diagnostic analysis for a linear mixed model cannot be based on the same diagnostic procedures as ordinary least squares regression (Verbeke and Molenberghs, 2000).

Brown and Prescott (1999) use normal plots of residuals and plots of residuals against predicted values to check for outliers. They compare treatment differences and variance and correlation matrices with and without outliers. Verbeke and Molenberghs (2000) use histograms and scatter plots of the empirical Bayes estimates of γ_i for diagnostic purposes. Langford and Lewis (1998) analyse a range of practical procedures for dealing with outliers in multilevel data in the context of educational research. These techniques include the use of deviance reduction, leverage values, hierarchical cluster analysis and the measure called DFITS, defined as

$$\mathrm{DFITS}_{\mathrm{mi}} = D_{\mathrm{mi}} = |p^*_{\mathrm{mi}}|\sqrt{\frac{h^*_{\mathrm{mi}}}{1 - h^*_{\mathrm{mi}}}}\,,$$

where $h^*_{\mathrm{mi}} = h_{\mathrm{mi}} / \sum_k h_{mik}$ is the standardized leverage value,

$$p^*_{\mathrm{mi}} = p'_{\mathrm{mi}} \frac{1}{\sqrt{\frac{n_m - 1 - {p'_{\mathrm{mi}}}^2}{n_m - 2}}}$$

is the studentized residual, and n_m is the number of units in the random part of the model at the level m.

Langford and Lewis (1998) regard as outlying all units for which the absolute studentised residual $|r_s|$ is greater than or equal to 2.0. They exclude the unit under examination from contributing to the random part of the model and introduce separate fixed-effects parameters for the particular unit in the model. At each step, the parameters in the expanded model have to be reestimated (using an iterative procedure), making the data modelling complex and time consuming. Longford (1998) points out that the more procedures are applied, the more outliers are found (including some false negatives).

When observations are clustered, not only elements, but also entire clusters may be outliers. Longford (2001) studies outlying clusters in a two-level random coefficient model. He applies a simulation-based method for outliers, similar to the parametric bootstrap. The method is related to the general proposal of Rubin (1984), which can be paraphrased as follows: "If a particular model fits well, then the realized data set does not stand out among data sets simulated from the fitted model." A linear mixed model given by (1) is considered, assuming that it fits for all clusters except one (h), for which

$$y_h = X_h \delta_h + \varepsilon_h \,, \tag{4}$$

where δ_h is a vector of regression parameters unrelated to β or γ_h in (1) and ε_h is a random sample from $\mathcal{N}(0, \sigma_1^2)$. The mixed model is first fitted to y_{-h}, that is, the data set with the cluster h excluded. Let the resulting parameter estimates be $\hat{\theta}_{[-h]}$. The deviance evaluated for the entire data set y at $\hat{\theta}_{[-h]}$ is compared with the deviances evaluated for the data sets with simulated replacements for cluster h. Although this procedure is computationally intensive, it requires little programming effort, because the same model fitting algorithm is used throughout.

The data set we analyze comprises a matrix with 28 sites as its rows and the 384 months in the period 1971 – 2002 as its columns. Further, the altitude of each site is given, the presence of the "El Niño" and "La Niña" in each month are indicated and the value of the Southern Oscillation Index (SOI) is given.

2 Modelling the Monthly Average Temperatures

We fit six linear mixed models, $M = 1, \ldots, 6$, to the temperature data. To eliminate the seasonal effects, separate analyses are conducted for each month $j = 1, \ldots, 12$ of the year. Therefore, we study $6 \times 12 = 72$ model fits. Each model is fitted to $28 \times 32 = 896$ observations (sites by years), except for a few missing values and, when applicable, excluded outliers.

As an alternative, a single model could be fitted, with the months represented by a categorical variable. The drawback of this approach is that several interactions of this

variable with others would have to be introduced, to take account of the different patterns of the temperatures with regard to the other covariates. Such a proliferation of parameters would hinder the interpretation of the results.

In each model, the altitude of the site is a continuous covariate and the "El Niño" and "La Niña" phenomena are introduced through SOI for the current month as indicator (dummy) variables. These are *prima facie* important predictors. To capture the effect of the "El Niño" and "La Niña" more completely, we also include SOI with lag one (SL1) and lag two (SL2) from the two previous months. Madl (2000) and others identify an influence of the months previous to the "El Niño" phenomenon on the weather. He states that in the months preceding an "El Niño" event, the normal weather pattern breaks down. For some reasons, not yet well understood, the westward atmospheric pressure gradient decreases.

This motivates a sequence of models with increasingly detailed modelling of the "El Niño" and "La Niña" phenomena. The models contain the following covariates:

- x_1 — altitude of the site, in meters;
- x_2 — the Southern Oscillation Index (SOI) for month j and year k;
- x_3 — the year, centred around 2002 (YR – 2002);
- x_4 — the Southern Oscillation Index for the previous month $(j-1)$ of the same year (SL1);
- x_5 — the Southern Oscillation Index for two months earlier $(j-2)$ in the same year (SL2);
- $x_6, \ldots, x_{11}$ — the variables that indicate the following:
  - "El Niño": $x_{6k} = 1$, $x_{7k} = 0$;
  - "La Niña": $x_{6k} = 0$, $x_{7k} = 0$;
  - Normal: $x_{6k} = 0$, $x_{7k} = 1$;
  - "El Niño" in the previous month: $x_{8k} = 1$, $x_{9k} = 0$;
  - "La Niña" in the previous month: $x_{8k} = 0$, $x_{9k} = 0$;
  - Normal conditions in the previous month: $x_{8k} = 0$, $x_{9k} = 1$;
  - "El Niño" two months ago: $x_{10k} = 1$, $x_{11k} = 0$;
  - "La Niña" two months ago: $x_{10k} = 0$, $x_{11k} = 0$;
  - Normal conditions two months ago: $x_{10k} = 0$, $x_{11k} = 1$.

The six models, all of the form (1), include the following covariates:

Model 1: $x_1 - x_3$

Model 2: $x_1 - x_4$

Model 3: $x_1 - x_5$

Model 4: $x_1 - x_7$

Model 5: $x_1 - x_9$

Model 6: $x_1 - x_{11}$

We regard the variables $x_1 - x_3$ as essential for any credible model; with x_4 and x_5, we add more information about "El Niño" and "La Niña", respectively, regarding the former as more important, and the following variables can be regarded as interactions, so they have lower priority for inclusion in the model.

A random effect for the year is included; the random part of the model, Z in (1), comprises the intercept and year (x_3). The two components of γ_i are correlated.

The six models are fitted first under the assumption of independent errors. Since some observations with large residuals are identified, we assess whether these observations are outliers. The temporal and spatial correlation of the errors are analysed and modelled by a posterior analysis(Andrade, 2009).

3 Outliers

As was indicated earlier, if a particular model fits well, then the realised data set does not stand out among data sets simulated from the fitted model. For a given data set, a small number of features associated with the outlier status or any form of model violation is specified. A statistic (such as the largest residual) and a plot (normal plot of the residuals at a given level), or even a combination of plot and statistic, can be defined as features. Such a feature is then evaluated on the analysed data and compared with its versions with data simulated from the fitted model. The steps of the procedure are listed below, with the settings for our analysis given.

1. Features associated with the outlier status are defined. We study the following features:
   (a) the largest (absolute) residual;
   (b) the two largest (absolute) residuals;
   (c) the three largest (absolute) residuals.
2. The models 1 – 6 are fitted assuming no outliers (using all the observed data).
3. In the analysis with feature a, the observation with the largest (absolute) residual is eliminated and the models are fitted again.
4. Two methods are used in the analysis with features b, and c.
   (a) First, the observation with the largest (absolute) residual is removed, as in 3, and then the model is fitted again. The observation with the largest (absolute) residual in this fit is removed, and the models are fitted again. For feature c, this step is repeated, to find the third observation to be removed.
   (b) The observations with the two (or three) largest (absolute) residuals are eliminated simultaneously and the models fitted again without them.
5. 1000 data sets of the outcome variable of the fitted models are simulated, using the original values of X and Z and estimated values of variance components and regression parameters, obtained by fitting the models with the observed data in step 2.

   All the random terms are assumed to be normally distributed.

   The models are fitted to each simulated data set and the steps 3, and 4, applied to the results.
6. The estimated residual variances ($\hat{\sigma}_0^2$) obtained in each simulation in 5. with the complete simulated data set are plotted against their counterparts $\hat{\sigma}_h^2$, $h = 1, 2, 3$, obtained with h observations with the largest (absolute) residuals removed, after applying the appropriate variant of step 4.

7. The estimated residual variances obtained with observed data and with h observations with the largest residuals removed are plotted in the same graphs as in step 6.
8. If the points that correspond to the real data (the observed feature) stand out among the points that correspond to the simulations (the simulated features), the observed data is discordant with the model.

The conditional residuals (3), are used throughout. We prefer to use these residuals because they are contaminated less by the random effects than the alternatives listed earlier. All computing was carried out in SAS version 9.1.

The distances between the observed feature and the simulated features and plots of the distance distribution are also obtained to assess how large are the differences in the residual variances when the $h = 1, 2$ or 3 largest (absolute) residuals are eliminated, to assess the influence that these residuals exercise on the fit, and to determine whether these removed observations are outliers.

A similar analysis of the impact on the other parameter estimates when observations with the largest residuals are eliminated is carried out.

Time series plots for the sites whose data are discordant with the model are also drawn and compared with the corresponding time series plots for nearby sites, following the geostatistics principle that states that two outcomes at locations close to one another are more likely to be similar than outcomes at locations that are further apart (Isaaks and Srivastava, 1989). This is a complement to the decision on the outlier status of the observed data.

4 Results

The sites and years and the corresponding values of each variable for every month are listed in Table 1 for observations that have absolute residuals greater than 2.0. The altitude (*Alt.*) is given in meters above sea level and the average monthly temperature (*Temp.*) in $^\circ$ C. The same observations are identified as outliers for all the six models. The extreme right-hand column (*Res. range*) gives the ranges of the largest absolute residuals for the six models. The narrow ranges attest to the stability of the values of the largest residuals across the models.

The estimated residual variances obtained by fitting the six models with the observed data and with the observation with the largest (absolute) residual removed are very similar for all the months; the one observation, however purposefully selected, has only minor influence on the fit to the remaining 890 or so observations. After removing more than one observation, the results obtained with the two elimination methods do not differ. In the simulations, the two methods of elimination also give very similar results; see Figure 5 for the months of February and November. The simulation analysis identifies the same observations as outliers for all six models; see Figure 6 for the month of November. As discussed later, variables with lagged values, $x_4 - x_{11}$, make a very modest contribution to the model fit.

For the months from January to April, graphs do not provide any evidence that the observations with the largest (absolute) residuals are outliers. For the other months, the observations with the largest (absolute) residuals stand out among the simulated features,

Table 1: The largest (absolute) residuals by month.

	Month/Site	*Year*	*Alt.*	*Temp.*	*SOI*	*SL1*	*SL2*	X_6	X_7	X_8	X_9	X_{10}	X_{11}	*Res. range*
January														
172	El Topacio	1982	1676	21.3	1.3	0.5	0.1	0	0	0	1	0	1	2.76 – 2.78
187	Zaragoza	1983	925	28.3	–4.2	–2.8	–3.2	1	0	1	0	1	0	2.53 – 2.55
169	Acuetulua	1982	1014	21.2	1.3	0.5	0.1	0	0	0	1	0	1	2.49 – 2.51
29	La Teresita	1995	1950	19.6	–0.6	–1.6	–0.7	0	1	1	0	0	1	2.22 – 2.27
418	Zaragoza	1973	925	22.8	–0.5	–1.6	–0.5	0	1	1	0	0	1	2.25 – 2.31
323	Monteloro	1990	1861	20.9	–0.2	–0.7	–0.4	0	1	0	1	0	1	2.14 – 2.16
17	Zaragoza	1972	925	21.1	0.4	0.0	0.5	0	1	0	1	0	1	2.05 – 2.10
February														
117	Zaragoza	1979	925	26.6	0.8	–0.7	–0.3	0	1	0	1	0	1	2.21 – 2.24
174	San Emigdio	1982	1272	22.9	–0.1	1.3	0.5	0	1	0	0	0	1	2.16 – 2.17
March														
324	Monteloro	1990	1861	22.1	–1.2	–2.4	–0.2	0	1	1	0	0	1	2.84 – 2.87
119	Zaragoza	1979	925	26.5	–0.5	0.8	–0.7	0	1	0	1	0	1	2.37 – 2.40
342	Monteloro	1991	1861	21.5	–1.4	–0.1	0.6	1	0	0	1	0	1	2.33 – 2.36
April														
112	Zaragoza	1979	925	27.3	–0.4	–0.5	0.8	0	1	0	1	0	1	3.24 – 3.25
318	Monteloro	1990	1861	22.1	0.0	–1.2	–2.4	0	1	0	1	1	0	3.05 – 3.08
30	Zaragoza	1973	925	21.1	–0.2	0.2	–2.0	0	1	0	1	1	0	2.49 – 2.54
May														
101	Zaragoza	1978	925	28.9	1.3	–0.6	–0.8	0	0	0	1	0	1	4.49 – 4.55
324	Monteloro	1990	1861	23.1	1.1	0.0	–1.2	0	0	0	1	0	1	3.62 – 3.69
105	Monteloro	1978	1861	21.9	1.3	–0.6	–0.8	0	0	0	1	0	1	3.14 – 3.20
116	Zaragoza	1979	925	26.7	0.3	–0.4	–0.5	0	1	0	1	0	1	2.52 – 2.55
June														
104	Zaragoza	1978	925	28.3	0.3	1.3	–0.6	0	1	0	0	0	1	4.21 – 4.30
331	Monteloro	1990	1861	23.6	0.0	1.1	0.0	0	1	0	0	0	1	3.81 – 3.86
349	Monteloro	1991	1861	22.7	–0.5	–1.5	–1.0	0	1	1	0	0	1	2.85 – 2.92
498	El Topacio	1999	1676	16.2	–0.1	0.1	1.4	0	1	0	1	0	0	2.58 – 2.61
July														
327	Monteloro	1990	1861	23.1	0.5	0.0	1.1	0	1	0	1	0	0	3.70 – 3.74
238	I. Manuelita	1985	1020	19.3	–0.3	–0.9	0.2	0	1	0	1	0	1	3.61 – 3.63
103	Zaragoza	1978	925	27.1	0.4	0.3	1.3	0	1	0	1	0	0	3.35 – 3.40
345	Monteloro	1991	1861	22.7	–0.2	–0.5	–1.5	0	1	0	1	0	1	3.07 – 3.15
368	Queremal	1992	1496	15.7	–0.8	–1.2	0.0	0	1	0	1	1	0	3.10 – 3.13
175	Zaragoza	1982	925	21.4	–1.9	–1.6	–0.7	1	0	1	0	0	1	2.51 – 2.54
August														
462	Queremal	1997	1496	14.2	–2.1	–1.0	–2.0	1	0	0	1	1	0	4.89 – 4.94
101	Zaragoza	1978	925	28.0	0.0	0.4	0.3	0	1	0	1	0	1	3.79 – 3.86
175	Zaragoza	1982	925	21.5	–2.5	–1.9	–1.6	1	0	1	0	1	0	3.18 – 3.24
346	Monteloro	1991	1861	22.8	–0.9	–0.2	–0.5	0	1	0	1	0	1	2.86 – 2.89
September														
104	Zaragoza	1978	925	28.8	0.0	0.0	0.4	0	1	0	1	0	1	4.79 – 4.80
440	Acuetulua	1996	1014	28.1	0.6	0.4	0.6	0	1	0	1	0	1	2.77 – 2.80
178	Zaragoza	1982	925	21.8	–2.0	–2.5	–1.9	1	0	1	0	1	0	2.45 – 2.46
351	Miravalle	1991	1233	24.6	–1.8	–0.9	–0.2	1	0	0	1	0	1	2.30 – 2.31
32	Zaragoza	1973	925	20.4	1.4	1.1	0.5	0	0	0	0	0	1	2.28 – 2.30
157	5403502	1981	1600	22.3	0.4	0.4	0.8	0	1	0	1	0	1	2.19 – 2.21
October														
288	La Buitrera	1987	1500	26.3	–0.7	–1.2	–1.5	0	1	0	1	1	0	5.55 – 5.57
105	Zaragoza	1978	925	27.8	–0.7	0.0	0.0	0	1	0	1	0	1	3.80 – 3.81
18	Zaragoza	1972	925	21.3	–1.2	–1.6	–1.0	0	1	1	0	0	1	2.29 – 2.32
352	Monteloro	1991	1861	21.1	–1.5	–1.8	–0.9	1	0	1	0	0	1	2.18 – 2.19
November														
104	Zaragoza	1978	925	28.7	–0.1	–0.7	0.0	0	1	0	1	0	1	4.22 – 4.25
18	Zaragoza	1972	925	21.8	–0.5	–1.2	–1.6	0	1	0	1	1	0	2.14 – 2.17
120	Zaragoza	1979	925	26.4	–0.6	–0.4	0.1	0	1	0	1	0	1	2.09 – 2.11
December														
105	Zaragoza	1978	925	28.5	–0.3	–0.1	–0.7	0	1	0	1	0	1	4.14 – 4.15
317	2608512	1990	954	21.5	–0.5	–0.7	0.1	0	1	0	1	0	1	3.08 – 3.10

providing evidence that these observations are outliers. Subsequently, observations with the two and three largest (absolute) residuals are analysed. Figure 7 shows the graph for July; observation No. 327 is an obvious outlier.

4.1 Observed and Simulated Features

July stands out as the month with the largest distance of the simulated feature ($\hat{\sigma}^2_h$, $h = 1, 2, 3$) from the observed feature; see Figure 8. The diagram also shows the largest variability (standard deviation and interquartile range). The distances from the observed to the simulated feature, when the two largest and the three largest (absolute) residuals are removed from the analysis, are similar to the distance when only the largest (absolute) residual is removed. This indicates that observation 327 is an outlier, but observations with the second and the third largest (absolute) residuals, No.s 238 and 103 in Table 1, are not.

Observations from January and February do not stand out among the simulated features. The distances between estimated residual variances from the observed and the simulated data are very small (Figure 8). March and April also show small distances between observed and simulated features, when the observation with the largest (absolute) residual is removed from the analysis. Only one box plot is drawn for these months because not even the observation with the the largest absolute residual stands out among the simulated features for them.

May, June and August show different distance distributions when the observations with the largest (absolute) residuals are removed from the analysis. The distances between the observed and the simulated features increase when the observations with the two largest (absolute) residuals are removed from the analysis, as compared to when the observations with only the largest (absolute) residuals are removed. The distances also increase when the three largest absolute residuals are removed (Figure 8).

September, October and November show changes in the distances when the observations with the two largest (absolute) residuals are removed compared to when the observation with the largest (absolute) residuals are removed. When the observations with the three largest (absolute) residuals are removed, the distances do not change substantially. There is a substantial change in the distances for December when the observations with the two largest (absolute) residuals are removed (two box plots drawn in 8).

4.2 Profiles of Problematic Sites

The site Zaragoza features in Table 1 for every month, and Monteloro for eight of the months. Since they have many outliers, these two problematic sites deserve a closer examination. By the profile of a site we mean the time series plot of the observations for the site over the period of the study. Figures 9 and 10 display the profiles of Zaragoza and Monteloro, accompanied by profiles of sites that are located nearest to them, and are at similar altitudes. The profiles are broken (discontinued) when observations are not available.

Figure 9 indicates that the monthly average temperatures recorded at Zaragoza differ from their counterparts at the two sites closest to it (24.08 and 38.05 km away, re-

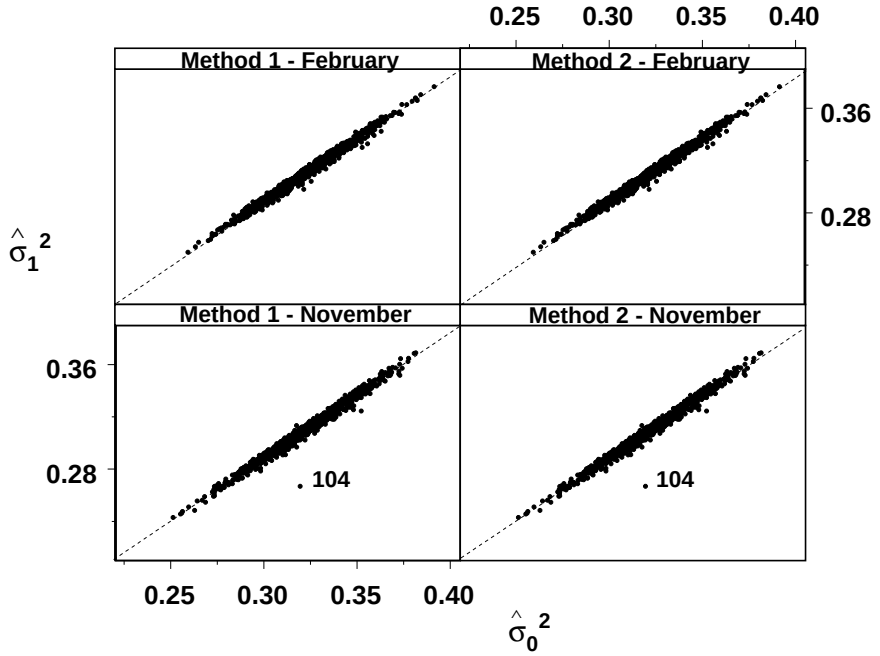

Figure 5: Estimated residual variances for all observations ($\hat{\sigma}_0^2$) and for the observations with the largest (absolute) residual removed ($\hat{\sigma}_1^2$), for the two methods and the months of February and November. The number attached, 104, is the observation identified in Table 1, with the largest (absolute) residual for the month November, and it is an obvious outlier

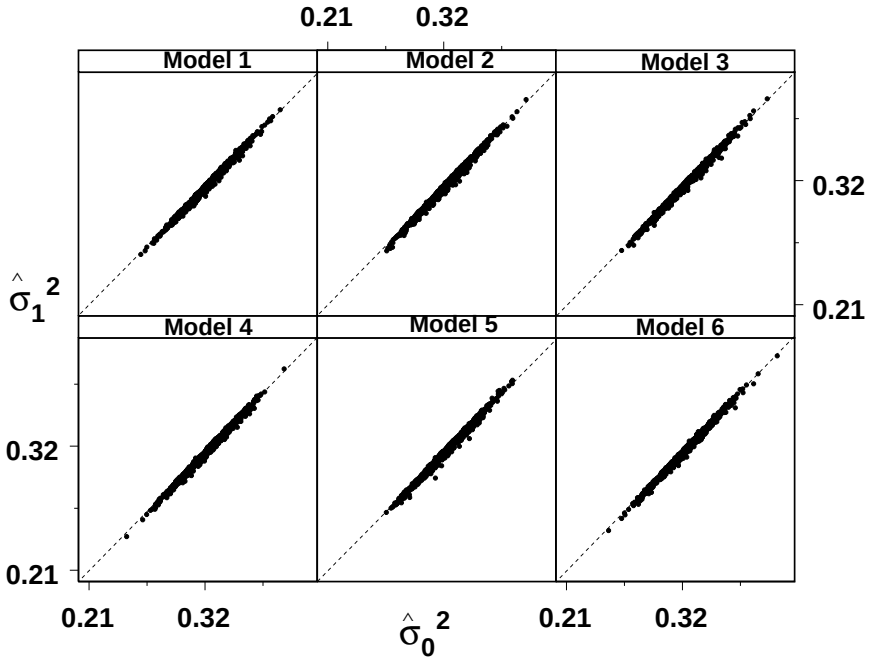

Figure 6: Estimated residual variances for all observations ($\hat{\sigma}_0^2$) and for the observations with the largest (absolute) residual removed ($\hat{\sigma}_1^2$), for the six models, November.

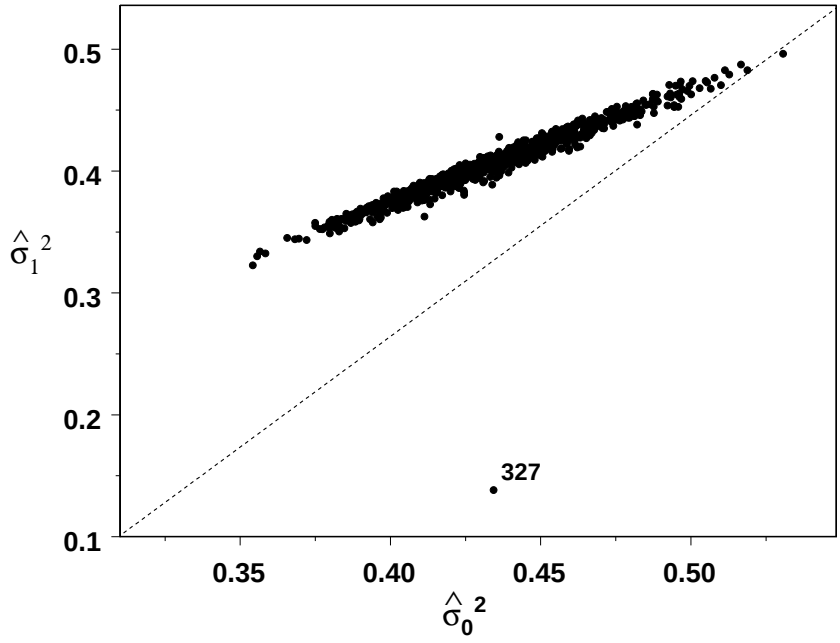

Figure 7: Estimated residual variances for all observations ($\hat{\sigma}_0^2$) and for the observations with the largest (absolute) residual removed ($\hat{\sigma}_1^2$), for the month July, model 1. The number attached, 327, is the observation identified in the Table 1 with the largest (absolute) residual for the month July, and it is an obvious outlier

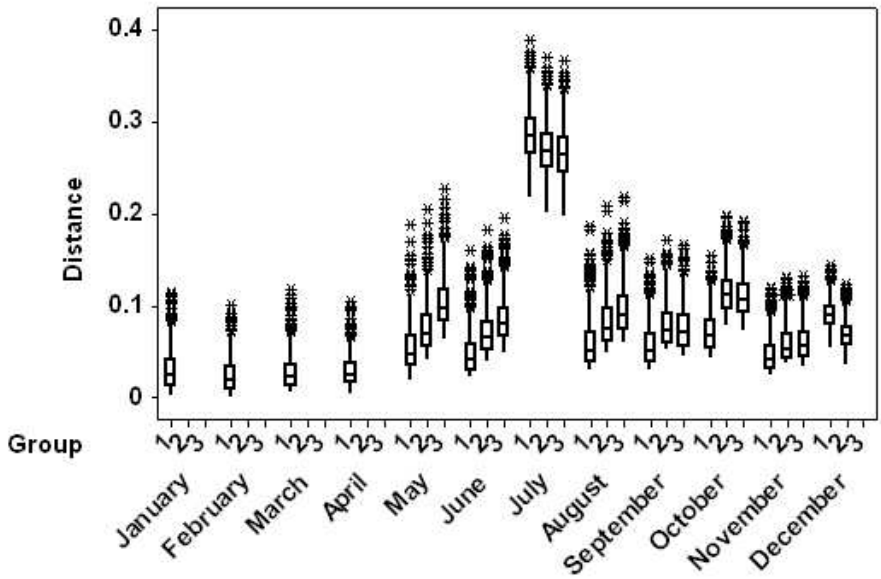

Figure 8: Distances of estimated residual variances by month and group, $h = 1, 2, 3$, marked at the bottom of the diagram. The group indicates the number of the largest (absolute) residuals

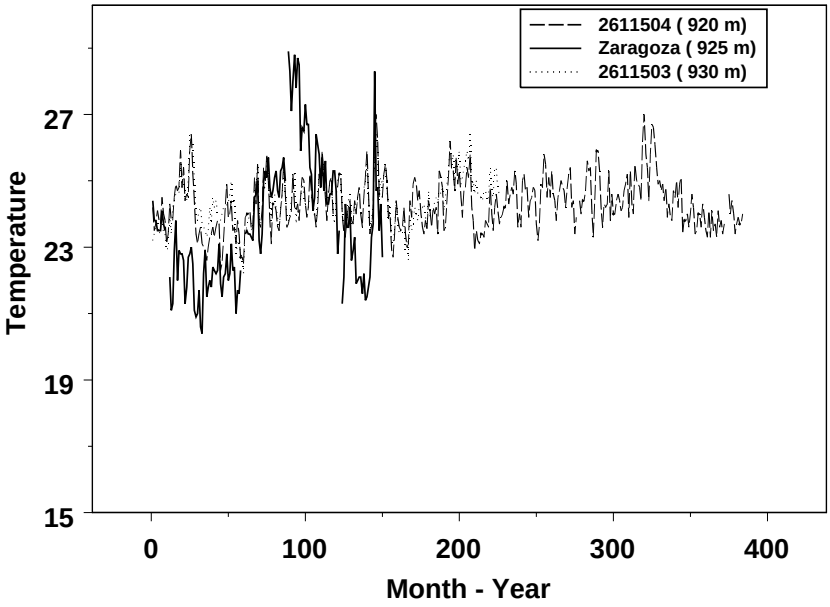

Figure 9: Time series of the sites close to Zaragoza. Month on the x-axis is counted from January 1971, and month 384 corresponds to December 2002

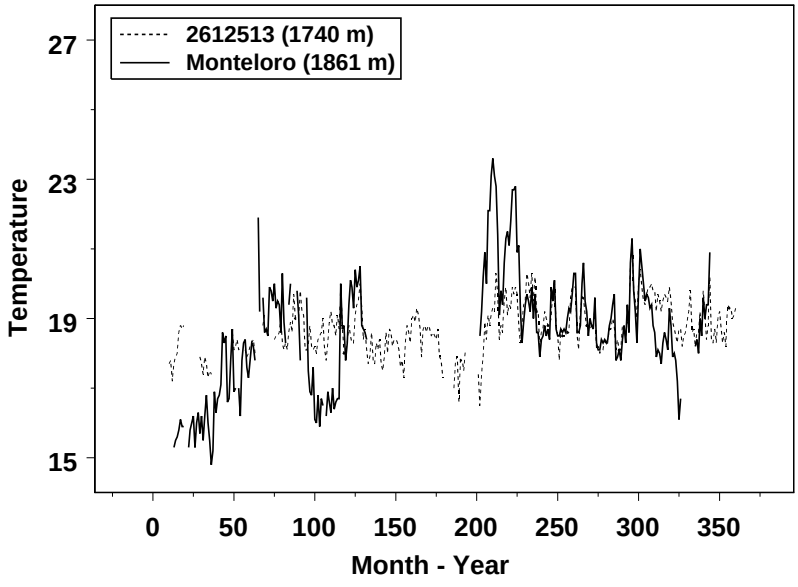

Figure 10: Time series for the sites close to Monteloro. Month on the x-axis is counted from January 1971, and month 384 corresponds to December 2002

spectively). The differences are so large that failure of the measurement instruments at Zaragoza is the only plausible explanation.

Monteloro is compared with the site closest to it, 2612513 (the distance between them is 36.46 km), in Figure 10. The time series of the original data of monthly average temperature show that Monteloro had much larger fluctuations than site 2612513 in the period 1971 – 1991 (months 0 to 250).

We note that these diagrams have an *ad hoc* nature, and are effective only after suspect observations or sites are identified. Otherwise, a lot of diagrams would have to be inspected and subjective judgement exercised in some instances.

Other sites evaluated by similar graphs with their closest sites were Queremal, August; Acuetulua, September; Buitrera, October; and 2608512, December, which show large values, confirming that the observations in these months are outliers.

4.3 Impact of Data Reduction on Parameter Estimates

In this section, we compare the model fits for the original data and for the data reduced by discarding observations adjudged to be outliers. We delete both some individual observations and all the observations from the sites Zaragoza and Monteloro in the period 1971 to 1991 (Andrade, 2009).

For most months, estimates of covariance parameters decrease, and $\hat{\sigma}^2_{\gamma_3}$ increases slightly for May, June, October and November. The residual variance estimate is reduced substantially when the outliers and observations from Zaragoza and Monteloro are removed (Table 2). For models 2 – 6 we observed similar behaviour.

Table 2: Variance component estimates by month, for Model 1, fitted to all the observations (rows marked **A**), and with the outliers removed (**W**)

Month	n	$\sigma^2_{\gamma_2}$		$\sigma^2_{\gamma_1}$		$\mathrm{cov}(\gamma_1,\gamma_3)$		$\sigma^2_{\gamma_3}$		σ^2_{ε}	
		Est.	(*s.e.*)	Est.	(*s.e.*)	Est.	(*s.e.*)	Est.	(*s.e.*)	Est.	(*s.e.*)
January											
A	544	0.376	(0.105)	0.983	(0.308)	0.020	(0.008)	0.0005	(0.0002)	0.399	(0.026)
W	519	0.370	(0.102)	0.923	(0.288)	0.017	(0.007)	0.0004	(0.0002)	0.298	(0.020)
February											
A	548	0.350	(0.097)	1.140	(0.349)	0.026	(0.010)	0.0008	(0.0003)	0.321	(0.021)
W	524	0.337	(0.093)	1.024	(0.315)	0.021	(0.008)	0.0007	(0.0003)	0.251	(0.017)
October											
A	545	0.105	(0.033)	1.124	(0.345)	0.027	(0.010)	0.0008	(0.0003)	0.347	(0.023)
W	522	0.097	(0.029)	1.093	(0.329)	0.027	(0.010)	0.0009	(0.0003)	0.187	(0.012)
November											
A	540	0.111	(0.034)	0.981	(0.310)	0.024	(0.010)	0.0009	(0.0004)	0.319	(0.021)
W	518	0.103	(0.031)	0.827	(0.261)	0.021	(0.009)	0.0012	(0.0004)	0.217	(0.015)

For model 1, in the majority of months the estimates of β_0 are reduced slightly (Table 3). The estimates of β_1 are also reduced, although there are increments between August and October, and also in December. The estimates of β_2 are reduced for some, and for β_3

Table 3: Estimates of the regression parameters for January, February, October and November, Model 1, fitted to all the observations (rows marked **A**), and with the outliers removed (**W**). The p value quoted in the right-hand column is for the t-test of the hypothesis that $\beta_3 = 0$. The corresponding tests for β_0, β_1 and β_2 are all significant at 5% level for every month.

		β_0		β_1		β_2		β_3		
Month	n	Est.	(*s.e.*)	Est.	(*s.e.*)	Est.	(*s.e.*)	Est.	(*s.e.*)	*p value*
January										
A	544	31.952	(0.529)	–0.00775	(0.00035)	–0.364	(0.077)	0.019	(0.013)	0.1372
W	519	31.807	(0.599)	–0.00766	(0.00041)	–0.341	(0.076)	0.015	(0.013)	0.2338
February										
A	548	32.588	(0.558)	–0.00806	(0.00038)	–0.335	(0.066)	0.030	(0.013)	0.0241
W	524	32.361	(0.616)	–0.00789	(0.00043)	–0.335	(0.065)	0.026	(0.013)	0.0446
October										
A	545	30.950	(0.471)	–0.00710	(0.00032)	–0.213	(0.060)	0.035	(0.009)	0.0001
W	522	31.084	(0.563)	–0.00724	(0.00040)	–0.192	(0.055)	0.030	(0.009)	0.0008
November										
A	540	31.264	(0.530)	–0.00740	(0.00038)	–0.147	(0.056)	0.033	(0.009)	0.0007
W	518	30.945	(0.591)	–0.00722	(0.00043)	–0.130	(0.053)	0.025	(0.010)	0.0102

for all the months. As with model 1, estimates of regression parameters decrease in the majority of the months in models 2 – 6.

Table 4 displays the predicted values and their standard errors based on model 1, for the sites Base Aérea and La Esperanza, located at 954 and 1070 m, respectively. When the outliers are removed, the predicted values are reduced (by between 0.01 and 0.17); however, the standard errors are changed only slightly. Similar behaviour is observed for models 2 – 6.

5 Discussion

Outliers are often regarded as undesirable observations among the data we analyse because they distort the results we obtain. Uncertainty about the outlier status of an observation is a common problem. In our analysis, the same observations are identified as outliers with all six models, so we have high confidence that the appropriate observations are singled out and excluded from the analysis. Admittedly, the most important variables, altitude, year and the Southern Oscillation Index (SOI), are included in all the models, but the other variables also have non-negligible effects. We conjecture that failure of the equipment or mislabelling a collected data record are the most frequent causes of an outlier.

The observations with the single largest (absolute) residuals for the months January to April do not stand out among the simulated features; these months do not have large (absolute) residuals in comparison with the rest of the months (see Table 1). For the other months, the observations with the largest (absolute) residuals stand out among the simulated features. Consistent with the characteristics of these observations, the residuals have the largest (absolute) values, greater than 3.25. The profiles (time series plots) of the sites closest to Zaragoza and Monteloro confirm that the observations from these two sites are

Table 4: Predictions of monthly average temperatures by month, for Model 1, for sites Base Aérea and La Esperanza, located at 954 and 1070 m, respectively. Fitted to all the observations (rows marked **A**), and with the outliers removed (**W**)

	Base Aérea		La Esperanza	
Month	*Prediction*	(*s.e.*)	*Prediction*	(*s.e.*)
January				
A	24.719	(0.321)	23.820	(0.310)
W	24.647	(0.326)	23.758	(0.310)
February				
A	25.331	(0.333)	24.127	(0.315)
W	25.264	(0.332)	24.082	(0.309)
October				
A	24.279	(0.267)	23.478	(0.255)
W	24.268	(0.274)	23.448	(0.255)
November				
A	24.300	(0.267)	23.442	(0.251)
W	24.134	(0.263)	23.297	(0.241)

implausible (Table 1 and Figures 9 and 10). These observations also show considerable distances from the observed to the simulated features (estimates of the residual variance σ^2_ε) for these observations are widely separated, and the observations have a strong impact on the parameter estimates, especially on estimated values of σ^2_ε.

The next question is whether the observation with the second and third largest (absolute) residuals are also outliers. In the simulations, we use two elimination methods, eliminating two (or three) observations with the largest (absolute) residuals simultaneously, and one at a time. The two methods yield the same estimated residual variance values, for observed and simulated data. Graphs show that the observations with the three largest (absolute) residuals stand out among the simulated features for May, June and August. These months show changes in some of the (co)variance parameters, σ^2_ε in particular (Andrade, 2009), and also in the distances between the observed and the simulated features, which show increases. Therefore, the observations with the second and third largest (absolute) residuals for May, June and August are declared as outliers. September, October and November also show changes in distances when the two largest (absolute) residuals are removed, and estimates of $\sigma^2_{\gamma_1}$ also decrease, but September and November (Table 2) do not show a substantial decrease in the estimates of σ^2_ε. However, in the case of the second largest residual for September, time series of the site close to Acuetulua (Andrade, 2009) show that observation 440 is out of the range of monthly average temperature values. The observations with the third largest (absolute) residual for these months are not declared as outliers, because they do not show changes in their distances and in parameter estimates when they are removed from the model fitting.

The majority of observations declared as outliers belong to the sites Zaragoza and Monteloro (see Table 1). Time series plots (profiles) for Zaragoza and sites closest to it show a consistent pattern of implausibly large differences. The erratic behaviour we observe is most likely due to a failure of the measurement equipment or the recording

process. Large fluctuations are also identified for Monteloro and its neighbours in the period 1971 – 1991. The parameter estimates are changed substantially when the observations from Monteloro and Zaragoza are excluded from the analysis.

Diagnostic plots confirm that observations Queremal, August (month 462); La Buitrera, October (288); and 2608512, December (317), are outliers (Andrade, 2009). In the majority of months, when these observations, together with the observations from Zaragoza and Monteloro, are eliminated, all covariance parameter estimates decrease, as do the estimates of β's in the majority of the months, for all six models; The predicted values are also reduced.

The measurements from Zaragoza could not be repeated; in any case, this site was closed in 1983. Therefore Zaragoza will be eliminated from the model fitting altogether. The data from Monteloro for the period 1971 – 1991 will also be excluded from the model fitting. In addition to these, four observations were eliminated from the model fitting: Queremal, August (462), Acuetulua, September (440), La Buitrera, October (288), and 2608512, December (317).

Since our analysis involves many models to be fitted, an analyst's propensity to declare observations with large residuals as outliers has a substantial impact on the number of outliers identified. We have presented an approach in which this propensity is greatly ameliorated, and the process of model diagnostics gains greater objectivity and integrity.

References

Andrade, M. (2009). *Monthly Average Temperature Modelling for Valle del Cauca (Colombia).* (Unpublished PhD. Thesis. The University of Reading, United Kingdom)

Atherton, J. G., and Rudich, J. (1986). *The Tomato Crop; A Scientific Basis for Improvement*. Chapman and Hall.

Atkinson, A. C. (1985). *Plots, Transformations and Regression. An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford University Press.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics.* I*dentifying Influential Data and Sources of Collinearity*. John Wiley and Sons.

Brown, H., and Prescott, R. (1999). *Applied Mixed Models in Medicine*. John Wiley and Sons.

Cook, R. D., and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall.

Demidenko, E. (2004). *Mixed Models. Theory and Applications*. John Wiley and Sons.

Hadi, A. S., and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, *88*, 1264-1272.

Haslett, J., and Haslett, S. J. (2007). The three basic types of residuals for a linear model. *International Statistical Review*, *75*, 1-24.

Hawkins, D. M. (1980). *Identification of Outliers*. Chapman and Hall.

Henderson, C. R. (1982). Analysis of covariance in the mixed model: higher-level, nonhomegeneous, and random regressions. *Biometrics*, *38*, 623-640.

Isaaks, E. H., and Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*. Oxford University Press.

Jones, J. B. (2000). *Tomato Plant Culture. In the Field, Greenhouse and Home Garden*. CRC Press LLC.

Laird, N. M., and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963-974.

Langford, I. H., and Lewis, T. (1998). Outliers in multilevel data. *Journal of the Royal Statistical Society, Series A*, *161*, 121-160.

Longford, N. T. (1998). Discussion of the paper Outliers in multilevel data by I. Langford and T. Lewis. *Journal of the Royal Statistical Society, Series A*, *161*, 154-155.

Longford, N. T. (2001). Simulation-based diagnostics in random-coefficient models. *Journal of the Royal Statistical Society, Series A*, *164*, 259-273.

Madl, P. (2000). The El Niño (ENSO) Phenomenon. *Environmental Physics Letter*, 437-503.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*, 1151-1172.

Schanberger, O. (2004). Mixed model influence diagnostics. In *Proceedings of the Twenty-Ninth Annual SAS Users Group International Conference.*

Verbeke, G., and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer-Verlag.

Villareal, R. L. (1980). *Tomatoes in the Tropics*. Westview Press Inc.

Authors' addresses:

Mercedes Andrade-Bejarano
Escuela de Ingenieria Industrial y Estadistica, Edificio 357
Ciudad Universitaria Melendez
Apartado Aereo 25360
Cali, Colombia

South America
E-Mail: mercedes.andrade@correounivalle.edu.co

Nicholas T. Longford
Department of Economics and Business
University Pompeu Fabra
Barcelona, Spain
E-Mail: NTL@SNTL.co.uk