Tree Physiology 22, 973–982
© 2002 Heron Publishing—Victoria, Canada

Evaluation of methods for the combination of phenological time series and outlier detection

JÖRG SCHABER[1,2] and FRANZ-W. BADECK[1]

[1] *Potsdam Institute for Climate Impact Research, Telegrafenberg A51, P.O. Box 60 12 03, D-14412 Potsdam, Germany*

[2] *Author to whom correspondence should be addressed (schaber@pik-potsdam.de)*

Received October 17, 2001; accepted January 22, 2002; published online September 2, 2002

Summary There are several applications of combined phenological time series; e.g., trend analysis with long continuous time series, obtaining a compound and representative time series around weather stations for model fitting, data gap filling and outlier detection. Various methods to combine phenological time series have been proposed. We show that all of these methods can be analyzed within the theory of linear models. This has the advantage that the underlying assumptions for each model become transparent providing a theoretical basis for selecting a model for a particular situation. Moreover, the common theoretical background provides a means of comparing methods by Monte-Carlo simulation and with real data. Additionally, we explored the influences of two outlier detection methods. We show that the error called the month-mistake, whose origin is known and which is one of the few mistakes that can be detected in phenological data because of its large deviation, is best detected by the distribution-free 30-day residual rule in combination with a robust estimation procedure based on the minimization of the sum of absolute residuals (L_1-norm).

Keywords: linear model, month-mistake, robust estimation.

Introduction

Climate change studies have resulted in increased interest in phenological research in recent years. Evidence that the length of the vegetation period has increased in northern latitudes in recent decades (Keeling et al. 1996, Myneni et al. 1997) has prompted research on the effects of climate change and change in growing season length on growth and functioning of ecosystems. The timing of phenological phases is an important factor in analyses of changes in net primary production of trees in response to interannual variation and long-term changes in climate (Goulden et al. 1996, Kramer et al. 1996, 2000, Chen et al. 1999, White et al. 1999).

Fragmentary phenological information is often available at several observational stations and can be combined to provide a continuous time series when single observation series overlap. Besides having the effect of gap filling, a merged continuous time series also reduces the weight of exceptionally early or late observations by the averaging process, thereby increasing its reliability compared with single time series. A combined time series can also be used to find outliers in individual time series (Linkosalo et al. 1996, 2000). Methods for combining phenological time series have been applied in several phenological studies (Häkkinen et al. 1995, Linkosalo et al. 1996, Linkosalo 2000).

Häkkinen et al. (1995) proposed four methods to combine phenological time series, but did not rank the methods because they provided no comparable measure of performance. The objective of our study was to answer the following three questions. (1) Are there theoretical reasons for selecting a certain method of combining fragmentary phenological data into one continuous time series in order to increase the reliability of the phenological information for subsequent analysis? (2) Is there a particular technique that produces the most reliable estimates of the parameters of the combined phenological time series? (3) What is the influence of different outlier detection methods on the combined time series and what is an adequate way to detect and treat outliers?

The answer to the first question was sought by embedding the four methods proposed by Häkkinen et al. (1995) in a common theoretical framework to elucidate the underlying statistical assumptions of each method. The second and third questions were examined by applying several estimation techniques and two outlier detection methods in a Monte-Carlo simulation and to real data.

Methods

A common theoretical background

Häkkinen et al. (1995) proposed four methods of combining phenological time series, hereafter referred to as Methods 1, 2, 3 and 4. We start by analyzing Method 3, which was the method used by Linkosalo et al. (1996, 2000).

The combined time series, y_i $i = 1,\ldots,M$, of Method 3 is defined as (Häkkinen et al. 1995):

$$y_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij} - b_j), \tag{1}$$

where y_i can be interpreted as yearly means of the observations x_{ij} of year i at station j, adjusted by b_j, $n_i \leq N$ is the number of observations, i.e., stations with observations, in year i and N is the total number of stations considered. Some x_{ij} are undefined because no observations were made in year i at station j. The value of b_j is found by an iterative optimization algorithm that minimizes the sum of squared residuals:

$$S_2 = \sum_i \sum_j (x_{ij} - b_j - y_i)^2, \tag{2}$$

where S_2 is the sum of squared residuals.

Defining n_{ij} as 1 if x_{ij} exists, and 0 otherwise, we can redefine Equations 1 and 2 as:

$$y_i = \frac{1}{n_{i\bullet}} \sum_{j=1}^{N} n_{ij}(x_{ij} - b_j) = \bar{x}_{i\bullet} - \frac{1}{n_{i\bullet}} \sum_{j=1}^{N} n_{ij} b_j, \tag{3}$$

and

$$S_2 = \sum_{i=1}^{M} \sum_{j=1}^{N} (n_{ij}(x_{ij} - b_j - y_i))^2, \tag{4}$$

with $n_{i\bullet} = \sum_{j=1}^{N} n_{ij}$ and $\bar{x}_{i\bullet} = \frac{1}{n_{i\bullet}} \sum_{j=1}^{N} n_{ij}\, x_{ij}$.

The minimum of S_2 is found by differentiating S_2 with respect to b_j and y_i and setting the resulting derivatives equal to zero. This results in Equation 3 and:

$$b_j = \bar{x}_{\bullet j} - \frac{1}{n_{\bullet j}} \sum_{i=1}^{M} n_{ij} y_i. \tag{5}$$

with $n_{\bullet j} = \sum_{i=1}^{M} n_{ij}$ and $\bar{x}_{\bullet j} = \frac{1}{n_{\bullet j}} \sum_{i=1}^{M} n_{ij}\, x_{ij}$.

Inserting Equation 3 into Equation 5 and imposing $\sum_{i=j}^{N} b_j = 0$ to obtain a unique solution, we can rearrange to obtain linear equations for b_j:

$$\left(n_{\bullet j} + 1 - \sum_{i=1}^{M} \frac{n_{ij}^2}{n_{i\bullet}}\right) b_j + \sum_{k \neq j}^{N} \left(1 - \left(\sum_{i}^{M} \frac{n_{ij} n_{ik}}{n_{i\bullet}}\right)\right) b_k$$

$$= x_{\bullet j} - \sum_{i=1}^{M} n_{ij} \bar{x}_{i\bullet} \text{ for } j,\ k = 1,\ 2, ..., N, \tag{6}$$

with $x_{\bullet j} = \sum_{i=1}^{M} n_{ij}\, x_{ij}$.

Thus, the solution of Equation 4 with respect to $b = (b_1,\ldots,b_N)$ is Equation 3 and $b = C^{-1}r$, where C is a regular matrix with the set of elements $\{c_{jk}\}$ and r is a vector with the set of elements $\{r_j\}$ for $j,k = 1,\ldots,N$ with:

$$c_{jj} = n_{\bullet j} + 1 - \sum_{i=1}^{M} \frac{n_{ij}^2}{n_{i\bullet}} \text{ and } c_{jk} = 1 - \sum_{i}^{M} \frac{n_{ij} n_{ik}}{n_{i\bullet}} \text{ for } k \neq j$$

$$\text{and } r_j = x_{\bullet j} - \sum_{i=1}^{M} n_{ij} \bar{x}_{i\bullet}. \tag{7}$$

It can be shown (Searle 1971, Rencher 2000) that the same solutions of Equations 3 and 7 are obtained when we formulate the problem of finding y_i and b_j in Equation 4 as a linear model of a two-way crossed classification with fixed effects:

$$x_{ij} = m + a_i + b_j + e_{ij}, \tag{8}$$

setting $y_i = m + a_i$ and applying the method of least squares to estimate the parameters, where m is a general mean, a_i is the effect of year i, $i = 1,\ldots,M$ (hereafter called year effect) and b_j is the effect of station j, $j = 1,\ldots,N$ (hereafter called station effect). The e_{ij} are independent identically distributed random errors with assumed expectancy $E(e_{ij}) = 0$ and common variance σ_e^2. To find a well-defined solution m is set to zero and it is assumed that $\sum_{i=1}^{N} b_j = 0$.

Thus mathematically equivalent solutions are obtained when the problem of finding a combined phenological time series is formulated as in Equations 1 and 2 or as a two-way crossed linear model with fixed effects (Equation 8).

Häkkinen et al. (1995) defined the combined time series according to Method 1 as $y_i = \bar{x}_{i\bullet}$, i.e., simply taking the mean observation each year. We obtain this solution when we define our linear model as:

$$x_{ij} = m + a_i + e_{ij}, \tag{9}$$

i.e., a one-way linear model with fixed effects, and set $y_i = m + a_i$. The e_{ij} are independent identically distributed random errors with assumed expectancy $E(e_{ij}) = 0$ and common variance σ_e^2. To find a well-defined solution m is set to zero.

The combined times series according to Method 2 of Häkkinen et al. (1995) can also be defined within the framework of linear models. But it is not considered here because it requires a long reference time series that is only rarely available.

Method 4 proposed by Häkkinen et al. (1995) formulates Equation 8 as a two-way crossed linear mixed model of randomized block design:

$$x_{ij} = m + a_i + b_j + e_{ij}, \tag{10}$$

where b_j is the random block (station) effect, m is a constant and a_i is the fixed (year) effect. The random terms b_j and e_{ij} are, by assumption, independent identically distributed with an expected zero mean. Moreover b_j is assumed to have the common variance σ_s^2 and e_{ij} has the common variance σ_e^2. The combined time series is defined as $y_i = m + a_i$.

Parameter estimation

The error variance σ_e^2 is an important measure of both the reliability of the underlying data and the resulting combined time series because its magnitude is closely related to the range of confidence intervals for the estimated parameters that were also used by Häkkinen et al. (1995) and Linksalo et al. (1996) to estimate reliability.

Parameters m, a_i and b_j and variance components σ_s^2 and σ_e^2 for the mixed model (Equation 10) can be estimated by various procedures that yield different results when data are missing, i.e., the data are unbalanced, and they also have different statistical properties (Searle 1987, Milliken and Johnson 1992). We explored the differences resulting from use of the Restricted Maximum Likelihood (REML) (Patterson and Thompson 1971, Corbeil and Searle 1976), Maximum Likelihood (ML) (Hartely and Rao 1967, Hemmerle and Hartley 1973), Minimum Variance Quadratic Unbiased Estimation (MIVQUE0) (Hartley et al. 1978, Searle 1987) and TYPE I or Henderson III (T1) (Searle 1971, Milliken and Johnson 1992) techniques. The REML, ML and MIVQUE0 estimates were computed with modules available in the SAS procedure MIXED (SAS Software Release 6.12, Cary, NC). TYPE I was implemented following Thompson (1969) and Searle (1971) and the multiple linear least square regression (LS) for the fixed effect models (Equations 8 and 9) were calculated with the SAS procedure REG (SAS Software Release 6.12).

Robust estimation

The estimation techniques for linear models all use squared residuals to estimate the parameters and variance components. These techniques are susceptible to outliers (Dodge 1987, Rousseeuw and Leroy 1987, Barnett and Lewis 1996, Hubert 1997, Hubert and Rousseeuw 1997) because they emphasize extreme values. A variety of robust estimators have been proposed to identify and accommodate such outliers (Rousseeuw and Leroy 1987, Barnett and Lewis 1996). Hubert (1997) showed that, in the case of binary (dummy) regression variables, the L_1 regression, based on minimizing the sum of absolute residuals (residuals in the L_1-norm) $S_1 = \sum_{ij}\left|e_{ij}\right|$ rather than squared residuals as S_2 in Equation 2, has an optimal breakdown value, although it is not robust in the presence of leverage points, i.e., outliers in the regressors (Hubert and Rousseeuw 1997). In linear models of designed experiments, however, leverage points do not occur because the regressors are not measured values but dummy regression variables that are prescribed by the design of the model. To detect outliers by robust regression, L_1 estimation of the linear model parameters was conducted using the algorithm of Barrodale and Roberts (1973, 1974). The L_1 estimation was applied to the sum of absolute residuals of the two-way linear model with fixed effects (Equation 8). Calculation of the error variance component σ_e^2 of the L_1 estimation was made with the estimator proposed by McKean and Schrader (1987*a*, 1987*b*), where: $\sigma_e = \sqrt{n}(\tilde{e}_{(n-k+1)} - \tilde{e}_{(k)})/(2z)$ and $k = (n+1)/2 - z\sqrt{n/4}$. Here n is the number of non-zero residuals, $\tilde{e}_i$, $i = 1,..,n$ is the ordered set of non-zero residuals and z is an upper tailed standard normal critical value, here $z = 1.96$ after McKean and Schrader (1987*a*, 1987*b*).

Outlier detection

Biological variability of individual plants, differences in microclimate and observational and protocol errors add to the natural variability of phenological data that typically amounts to about 1 to 2 weeks (Baumgartner 1952, Schnelle 1955). One mistake that can be detected despite the natural variability of phenological data is the month-mistake (MM) because of its strong deviation. An MM is a protocol error that occurs when the observation date is noted in a wrong column or the observer uses the wrong column when transforming the actual date to the Julian date (day of year) using a conversion table or when transcribing the phenological information to a database (Schnelle 1955, Menzel 1997, Vasella 1997). Such events result in an observation date that is one or several months too early or too late. Most other mistakes are difficult to detect without having access to the original observation reports or knowledge about the customs of the observers because they vary within the natural variability of phenological observations. Based on the above considerations and the fact that the most straightforward method of detecting outliers in linear models is by considering residuals (Barnett and Lewis 1996), observations were considered as outliers if the estimated residuals of the linear models were larger than or equal to 30 days, i.e. where $\left|e_{ij}\right| \geq 30$. This method of detecting outliers is hereafter called the 30-day rule.

The 30-day rule was compared with the method proposed by Dixon (1950) and modified by King (1953). An extreme value of the ordered observations ($x_{(i,1)},\ldots,x_{(i,n)}$) of year i adjusted for the station effects b_j is considered an outlier when its test statistic T_i exceeds a critical value, where:

$$T_i = \max\left[\frac{x_{(i,n)} - x_{(i,n-1)}}{x_{(i,n)} - x_{(i,1)}}, \frac{x_{(i,2)} - x_{(i,1)}}{x_{(i,n)} - x_{(i,1)}}\right]. \qquad (11)$$

Only critical values for the 1 and 5% significance levels were considered (Barnett and Lewis 1996).

Monte-Carlo analysis

We used Monte-Carlo simulations to compare the methods used to combine phenological time series and to determine the influence of the two outlier detection methods. Observations x_{ij} were randomly generated according to a two-way linear model (Equations 8 and 10) with prescribed parameters m, a_i, b_j, σ_s^2 and σ_e^2. The general mean m was set to 120, and a_i was prescribed to vary according to a normal distribution with zero mean and a standard deviation of 7 days to reproduce the typical range of natural interannual variation of phenological time series. The other parameters depend on the Monte-Carlo simulation and are described below. Because phenological data are usually unbalanced, 50% of the generated data were omitted. This is a representative value as the completeness of the real data demonstrate (see Table 5). To simulate outliers, 1% of the resulting observations were made MMs by adding or subtracting 30 days (50/50 chance) from the generated observations. This is the approximate proportion of outliers that was found by Linkosalo et al. (1996, 2000).

Table 1. Setup of the Monte-Carlo simulations for comparing methods of estimating effects and variance components of linear models (Equations 8–10) with different estimation and outlier detection techniques. The observations x_{ij}, $i = 1,\ldots,M$, $j = 1,...,N$ are generated according to the model $x_{ij} = 120 + a_i + b_j + e_{ij}$, where the a_i, b_j and e_{ij} are prescribed to vary according to a normal distribution $N(x,y)$ with mean x and variance y. The setup and subsequent estimations were repeated 500 times per study.

Study	Average completeness	Year effects (a_i)	Station effect (b_j)	Residual error (e_{ij})	% Month mistakes	Number of stations (N)	Number of years (M)
1	50%	N(0,49)	N(0,15)	N(0,30)	1	10	30
2	50%	N(0,49)	N(0,30)	N(0,60)	1	20	30

Two simulation studies were conducted. In Study 1, which simulates the case when a combined time series is sought for several stations around a weather station, we imposed variances similar to those found in the real data (see Table 5). In Study 2, which simulates a case where many phenological stations within a region are combined for trend analysis, the variance components and the number of stations were doubled compared with those of Study 1. The main parameters of the two studies are summarized in Table 1.

For each study, five runs were conducted. In the first run (Run a) the different estimation methods were applied to the generated data set without imposing MMs. In the second run (Run b) MMs were introduced but they were not detected and corrected. Finally, in the last three simulation runs (Runs c–e), the generated time series were corrected for the detected MMs by the 30-day rule and the Dixon test using 5 and 1% significance levels. Each of the 10 simulation runs, hereafter referred to as 1a–e and 2a–e according to study number and run, was repeated 500 times.

Phenological data

To analyze the effects of discordancy tests on real combined time series we applied them to data of the German Weather Service (DWD). Phenological times series recorded within a radius of 10 km around a weather station on sites that differed less than 50 m in elevation and had at least five years of observations were selected from the phenological and climatological database of the DWD and combined for each weather station. At these stations we used data for bud burst of four deciduous tree species, *Aesculus hippocastanum* L., *Fagus sylvativca* L., *Quercus robur* L. and *Betula pubescens* Roth. Based on results from the Monte-Carlo simulation study, the robust L_1 estimation was used for the outlier tests. After the removal of detected outliers, we applied the T1 estimation procedure to the remaining data for variance component estimation.

Table 2. Phenological stations from the DWD phenological database with a distance of less than 10 km and a difference in elevation of less than 50 m from the DWD Weather Station 2606 (Giessen, 8.7° N, 50.58° E, altitude 186 m, 1951–1999) that have at least five observations of bud burst of horse chestnut.

Station	Distance (km)	Altitude (m)	Observational time span	Number of observations
1	2.12	160	1951–1983	31
2	6.04	190	1951–1973	18
3	4.94	160	1958–1991	33
4	7.80	200	1951–1999	44
5	7.15	190	1951–1978	28
6	6.62	200	1987–1998	12
7	9.79	200	1988–1995	7
8	9.99	180	1952–1998	44
9	5.98	180	1962–1980	18

Additionally, the effect of the different estimation procedures on the estimated combined time series itself was studied using one example from the combined time series. Nine phenological stations around Weather Station 2609 (Giessen, 8.7° N, 50.58° E, altitude 186 m, 1951–1999) were selected. The distance and altitude of the phenological stations and some characteristics are presented in Table 2 and the phenological time series for bud burst of horse chestnut (*A. hippocastanum*) is shown in Figure 1.

Results

Monte-Carlo analysis

For the Monte-Carlo analysis, because all effects, parameters and variance components were prescribed, it was easy to determine whether the models and estimation methods reproduced the values for a large number of simulations. Table 3 presents the mean absolute error (MAE), the mean squared er-

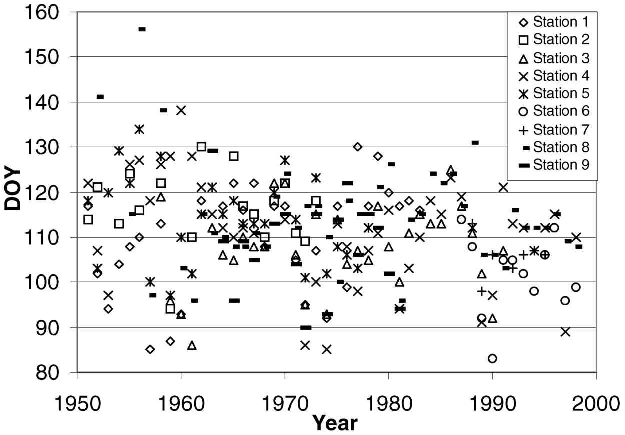

Figure 1. Phenological time series around Weather Station 2609 of the DWD over a 10-km radius as described in Table 2. Observations in Julian day of year (DOY).

Table 3. Summary of the results of the Monte-Carlo simulations described in Table 1. The values are means of 500 runs. Abbreviations: MAE = mean absolute error; MSE = mean squared error; σ_e^2 = estimated error variance according to Equations 8–10; MDTS = mean deviation between estimated and prescribed combined time series with standard deviation; and MM = number of month-mistakes over all 500 simulation runs.

Run	MM	Equation	Estimation method	MAE	MSE	σ_e^2	MDTS ± σ
1a	No MMs	8	L_1	3.5	26.8	88.3	–0.3 ± 3.6
		9	LS one-way	4.8	35.9	44.8	0.1 ± 3.4
		8	LS two-way	3.8	22.4	30.0	0.0 ± 3.1
		10	ML	3.8	22.6	23.9	0.0 ± 3.1
		10	REML	3.8	22.7	30.0	0.0 ± 3.1
		10	MIVQUE0	3.8	22.7	30.0	0.0 ± 3.1
		10	T1	3.8	22.7	30.0	0.0 ± 3.1
1b	736 MMs	8	L_1	3.7	34.4	90.9	–0.2 ± 3.7
	Not corrected for MMs	9	LS one-way	5.1	43.9	54.7	0.1 ± 3.7
		8	LS two-way	4.1	28.8	38.7	0.1 ± 3.4
		10	ML	4.1	29.2	30.8	0.1 ± 3.4
		10	REML	4.1	29.4	38.8	0.1 ± 3.4
		10	MIVQUE0	4.1	29.3	38.8	0.1 ± 3.4
		10	T1	4.1	29.3	38.7	0.1 ± 3.4
2a	No MMs	8	L_1	5.4	56.0	178.9	–0.2 ± 3.4
		9	LS one-way	7.2	80.9	89.8	0.0 ± 3.1
		8	LS two-way	5.7	50.4	60.1	0.0 ± 2.9
		10	ML	5.7	50.9	53.9	0.0 ± 2.8
		10	REML	5.7	51.0	60.1	0.0 ± 2.8
		10	MIVQUE0	5.7	51.0	60.2	0.0 ± 2.8
		10	T1	5.7	51.0	60.1	0.0 ± 2.8
2b	1535 MMs	8	L_1	5.6	64.3	187.2	–0.1 ± 3.5
	Not corrected for MMs	9	LS one-way	7.4	89.7	99.6	0.0 ± 3.3
		8	LS two-way	5.9	58.1	69.2	0.0 ± 3.1
		10	ML	6.0	58.7	62.1	0.0 ± 3.0
		10	REML	6.0	58.8	69.2	0.0 ± 3.0
		10	MIVQUE0	6.0	58.8	69.1	0.0 ± 3.0
		10	T1	6.0	58.8	69.2	0.0 ± 3.0

ror (MSE), the estimated residual variance (σ_e^2) according to Equations 8–10 and the mean deviation between the estimated and imposed combined time series (MDTS) for the simulation runs described in Table 1. The values of MAE, MSE, MDTS and σ_e^2 are the means of over 500 simulation runs. The MDTS is shown with its standard deviation. Differences among the methods with respect to σ_s^2 were qualitatively similar to those for σ_e^2 and are not shown.

The robust L_1 estimation always had the lowest MAE, whereas the non-robust two-way model estimations performed better with respect to MSE. The one-way model estimation (the average times series) performed poorly with respect to MAE and MSE. There were only small differences in simulated values of MAE and MSE between the different non-robust two-way models. The LS estimation for the fixed two-way model had slightly lower MSE in all scenarios. However, the estimation procedures differed in variance component estimation. It is noteworthy that the variance components estimated with the ML method were always lower than those estimated with the other methods. When no MMs were prescribed (Simulations 1a and 2a), the variances determined with the ML method were underestimated, whereas results obtained with the other non-robust methods showed estimated variance components close to the imposed values. The LS, REML, MIVQUE0 and T1 agreed well in their variance component estimates. Among methods, the MIVQUE0 variance estimates always had a higher variability (results not shown), indicating that LS, REML and T1 seem to provide the most reliable variance component estimates. The error variance of the L_1 estimation calculated by the method of McKean and Schrader strongly overestimated the imposed variances. The error variances varied between 15 (16) and 45 (72) in Simulation 1a (1b) and 32 (45) and 74 (93) in Simulation 2a (2b) for the non-robust two-way model estimations. The models did not differ much in terms of the estimated combined time series (MDTS): 99% of the estimates for the combined time series did not differ by more than about 10 days from the imposed values. The robust estimation L_1 was slightly less accurate than the other estimation methods. The fixed (Equation 8) and mixed (Equation 10) two-way models and their respective estimation methods all showed similar accuracy.

The strategy for outlier detection had a substantial impact on the number of outliers detected. The results for the robust L_1 method, LS two-way and T1 estimation are given in Ta-

Table 4. Summary of the results of the Monte-Carlo simulations described in Table 1. The values are means of over 500 runs. Abbreviations and definitions: σ_e^2 = estimated error variance according to Equations 8 and 10; detected outliers = % of imposed month-mistakes (MMs); and MM = detected imposed MMs in % of detected outliers.

Monte-Carlo simulation	Model equation	Estimation method	σ_e^2	Detected outliers	MM
				(% of imposed)	(% of detected)
Run 2c (30 day residuals)	8	L_1	90.9	41	99
	8	LS two-way	37.0	11	100
	10	T1	36.8	12	100
Run 1d (5% Dixon test)	8	L_1	86.4	205	22
	8	LS two-way	33.3	127	34
	10	T1	33.3	127	33
Run 1e (1% Dixon test)	8	L_1	88.7	74	26
	8	LS two-way	36.8	28	51
	10	T1	36.8	30	53
Run 2c (30-day residuals)	8	L_1	179.9	51	89
	8	LS two-way	64.6	29	98
	10	T1	64.4	31	98
Run 2d (5% Dixon test)	8	L_1	182.6	107	36
	8	LS two-way	63.0	77	46
	10	T1	63.0	77	47
Run 2e (1% Dixon test)	8	L_1	185.6	44	50
	8	LS two-way	66.0	31	64
	10	T1	66.0	32	63

ble 4. The results for the LS one-way method are not shown because the method was inferior in terms of estimation accuracy (see above). Because the performance of the mixed model estimation methods was similar in terms of outlier detection, only the results for the T1 methods are shown. In all simulation runs, the highest number of outliers was detected after applying the robust L_1 estimation, but the percentage of detected prescribed MMs was higher for the non-robust methods LS two-way and T1.

Thirty-day rule (Simulations 1c and 2c) The L_1 method fit detected about 41 and 51% of the imposed MMs in run 1c and 2c, respectively. Of the detected MMs, 99 and 89% were imposed MMs. The other estimation procedures detected a substantially lower number of MMs although the proportion of true MMs was higher.

Dixon tests (simulation runs 1d, 1e, 2d and 2e) The 1% and 5% Dixon tests detected more outliers than the 30-day rule in both studies and in combination with any estimation method. Use of the 5% confidence level resulted in detection of two to > 10 times more outliers than the 30-day rule. In simulation run 1d, more outliers were found than the number of prescribed MMs. Around 50% of the outliers detected by the Dixon test had not been imposed. The resulting estimate was similar after removal of 30-day residuals and after removal of outliers detected by the 1% Dixon test in Study 1, although about twice as many outliers were removed by the latter method. In Study 2, almost the same number of outliers was found with the non-robust methods by both the 30-day rule and the 1% Dixon test. However, the Dixon test detected a substantially lower percentage of imposed MMs (63%) than the 30-day rule (98%). This correlates with the lower estimate of σ_e^2 for the latter test.

The highest total number of imposed MMs and a low percentage of non-MMs was found by the combination of the robust L_1-estimation method with the 30-day rule.

Phenological data

Based on the results of the Monte-Carlo simulation, we applied the L_1 estimation to the described data sets to detect outliers followed by the T1 method for variance component estimation. Even with the restriction that a combined time series was calculated only when there were more than five time series per weather station available, there were only about three observations available per year (Table 5). Hence the Dixon rule, which is only defined for $N > 3$, could not be applied in many years. Average completeness was around 50% resulting in unbalanced designs. The Dixon tests found more outliers than the 30-day rule for each species. The impact of the outlier detection procedures on the estimated station variance σ_s^2 was not pronounced and is not considered further. Mean values of σ_s^2 were 17.8, 13.6, 15.1 and 13.3 for horse chestnut, beech, oak and birch, respectively. The 30-day rule found 0.2 to 0.6% outliers on average per species and combined time series decreasing the respective error variance σ_e^2 by 5 to 20%. The 1% (5%) Dixon test found 0.5% (1.1%) to 0.8% (1.5%) outliers decreasing the respective error variance σ_e^2 by 2% (6%) to 6% (10%). Although more values were deleted by the Dixon tests than by the 30-day rule, the resulting estimated σ_e^2 was lower after application of the 30-day rule, except for birch in combination with the 5% Dixon test. The Dixon tests did not identify all observations with 30-day residuals as outliers. Removal of these observations in addition to the outlier detected by the Dixon test led to an additional reduction of the average estimated σ_e^2. The percentage of outliers found by the 5% Dixon test was comparable with the percentages reported by Linkosalo et al. (1996, 2000). Varia-

Table 5. Application of 30-day rule and the Dixon tests after L_1 fit with subsequent T1 variance components estimation with data from the German Weather Service (DWD). The values of the variance components are mean values over all resulting combined time series. Outliers are expressed as percentage of observations (obs.) in the resulting combined (comb.) time series.

Species	*Aesculus hippocastanum*	*Fagus sylvatica*	*Quercus robur*	*Betula pendula*
Number of original observations	9668	5909	7886	7886
Number of resulting time series (No. used weather stations)	69	47	55	57
Number of values in resulting combined time series	3147	2069	2375	2684
Completeness including outliers (%)	50.9	46.7	46.9	50.4
Mean T1 estimate of σ_e^2 including outliers	45.3	36.9	34.7	31.1
Mean estimates of σ_s^2 for all outlier procedures	17.8	13.6	15.1	13.3
30-day residuals (% obs. in comb. time series)	0.6	0.4	0.2	0.2
Mean T1 estimate of σ_e^2 after removal of 30-day residuals	35.7	31.2	31.2	29.6
5% Dixon-type outliers (% obs. in comb. time series)	1.5	1.3	1.1	1.4
Mean T1 estimate of σ_e^2 after removal of 5% Dixon outliers	40.8	34.6	32.4	28.6
1% Dixon-type outliers (% obs. in comb. time series)	0.7	0.5	0.5	0.8
Mean T1 estimate of σ_e^2 after removal of 1% Dixon outliers	42.8	36.1	33.7	30.3
Number of 30-day residuals not detected by the 5% Dixon test	17	2	2	1
σ_e^2 after removal of 5% Dixon-type and 30-day residuals	33.7	29.6	29.5	27.9
Number of 30-day residuals not detected by the 1% Dixon test	22	4	5	5
σ_e^2 after removal of 1% Dixon-type and 30-day residuals	34.7	30.8	30.5	29.0

tion of the error variances was higher than in the simulation study, ranging from 9 to over 120 for horse chestnut before deletion of outliers and from 8 to 80 after deletion of outliers (results not shown).

For an individual combined time series (Table 2), the estimated yearly values differed by up to 10 days among the various procedures (results not shown). The one-way model and the robust fit estimation differed from the two-way models fitted by classical non-robust procedures. This is not surprising because the one-way model does not correct for station effects, and the L_1 estimation is not as susceptible to extreme observations as the other methods. A difference of 1.2 days was observed between the yearly estimates of the combined time series determined by the two-way mixed effect model (Equation 10) and the estimates determined by the mixed models (Equations 8 and 9). This result differs slightly from the Monte-Carlo study, where the differences between the two-way models were negligible; however, means over 500 simulation runs were used in the Monte-Carlo simulation. The estimation procedure in the mixed model did not have a significant influence on the values of the estimated parameters. All 95% confidence intervals for the year effects contained the estimated year effects of all other procedures.

The estimated error variances of the mixed model (estimated by REML, MIVQUE0 and T1) and the fixed two-way model are all 73.6. Only ML and the fixed one-way model estimates differ from this value with 58.5 and 80.5, respectively. This large error variability was probably caused by the large variability of the time series at individual stations between 1955 and 1960 where in some years trends in opposite directions were observed (Figure 1).

The number of values that exceed the 30-day threshold increases from one to four when a robust L_1 fit (Figure 2) is used instead of a T1 fit (Figure 3). Applying the 5% (1%) Dixon test to the L_1 fit, seven (two) outliers are found compared with four (one) outliers for the T1 fit. The outliers detected by the 5% Dixon test are marked with an arrow in Figures 2 and 3. The outliers that were also detected by the 1% Dixon test are additionally labeled with "1%." As seen in Table 5, not all residuals above the 30-day threshold were identified as outliers by the Dixon tests, a typical masking effect (Barnett and Lewis 1996). Because of this masking effect, the T1-estimated error

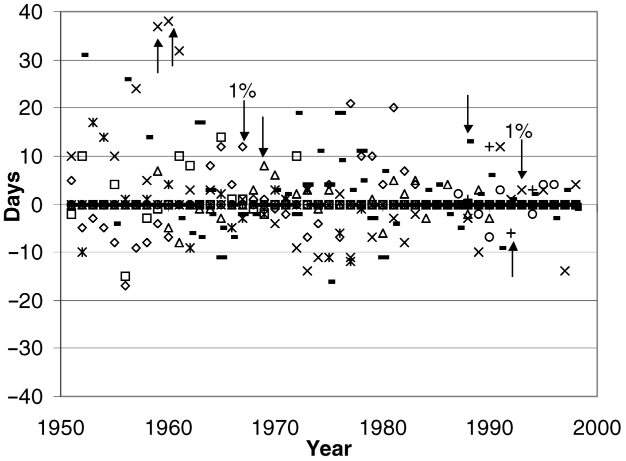

Figure 2. Residuals of the robust L_1 fit for the combined time series at Weather Station 2609. Residuals marked with an arrow have been determined as outliers by the 5% Dixon test. Arrow labeled 1% shows outliers that have also been detected by the 1% Dixon test. The symbols are attributed to the same stations as in Figure 1.

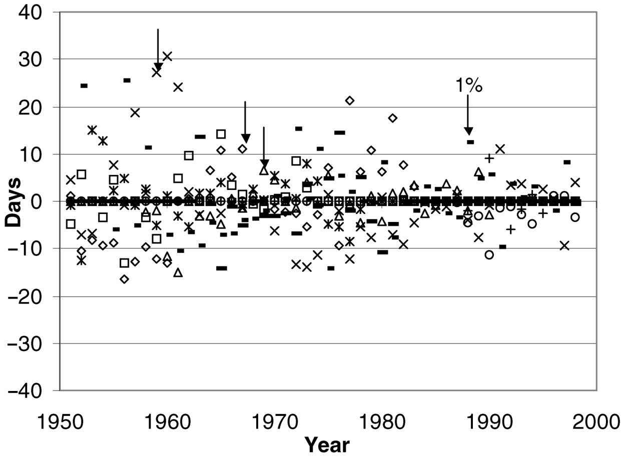

Figure 3. Residuals of the T1 fit for the combined time series at Weather Station 2906. Residuals marked with an arrow have been determined as outliers by the 5% Dixon test. Arrows labeled as 1% indicate outliers that have also been detected by the 1% Dixon test. The symbols are attributed to the same stations as in Figure 1.

variance of the combined time series was higher after removal of 5% Dixon outliers of the L_1 fit than after removal of 30-day residuals, namely 61.1 and 51.1, respectively. This corresponds to the reduction in estimated error variance by 17% through removal of seven Dixon outliers compared with a reduction of 21% through removal of four 30-day residuals. Removing the 30-day residual of the L_1-fit also decreased the maximal absolute difference between the robust combined time series and the T1 time series from 9.3 to 6.2 days. After removal of 5% Dixon outliers, the maximal absolute difference between the T1 time series and the robust time series was 7.5 days. Although more outliers were found by the 5% Dixon test, some extreme observations were missed because of the masking effect, resulting in an unnecessarily high influence of extreme observations on the combined time series. The 1% Dixon test finds outliers with residuals of 3 and 12 days.

Discussion

We showed that the four methods proposed by Häkkinen et al. (1995) for combining phenological time series can be treated within the framework of linear models. Method 3 proposed by Häkkinen et al. (1995) is the same linear model as Method 4, except that the station effects are considered fixed in Method 3 and random in Method 4. Having defined this difference within the framework of linear models gives us theoretical arguments for selecting one method or the other. The difference between the mixed model (Method 4) and the fixed model (Method 3) is of special interest because the other methods perform poorly (Method 1) or are of no relevance because of their restricted applicability (Method 2, see Method section). It can be argued that within the region where a phenology is to be estimated, the distribution of phenological stations is random. If the study objective was to estimate year effects and their variance rather than effects at a single station, which are assumed to be arbitrarily chosen within the region of interest, the mixed model is more appropriate. If the objective of the analysis was to find stations that show extreme behavior, or to check whether station differences are consistent throughout the years, the station effects would be treated as fixed (Searle 1971). However, in terms of parameter estimation, error variances and confidence intervals, the differences between the fixed and mixed models are small when proper estimation methods are used, e.g., LS, REML, MIVQUE0 or T1.

Because phenological data are often not normally distributed (Schnelle 1955, Menzel 1997), the robust ANOVA in the L_1 norm might be a more appropriate method, even though L_1 estimates are not unequivocal (Bloomfield and Steiger 1983) and not as accurate as classical methods (Table 3). The L_1 estimation is considered to be more appropriate for fatter tailed distributions (Dodge 1987). However, a better estimator for the variance component should be found than the one we used. The influence of different strategies for omitting outliers should be explored further. Month-mistakes are independent of other errors. Thus, it can be expected that 50% of the MMs occur in the opposite direction from errors of other origins and are therefore partly masked by these errors. This situation is mimicked in the structure of our Monte-Carlo simulation runs. Therefore, if it is assumed that only about 50% of the MMs can be detected, the performance of the 30-day residual rule combined with a robust L_1 fit is close to optimal. The non-robust procedures combined with the 30-day rule do not perform as well because they decrease large residuals. In the case of normally distributed errors, the Dixon test detects too many outliers, especially at the 5% level. In general, detection of more than 50% of the MMs can only occur at the expense of dismissing values within the natural variability of the data. Even the 1% Dixon test, which found a similar number of outliers as the 30-day rule in Study 2, rejected many false outliers. The Dixon test implies different variances for each year. Although this is reasonable because the rate of phenological development varies from year to year, it can result in the rejection of observations that deviate only slightly from the theoretically estimated value, e.g., the year 1993 in Figure 2, because the other residuals in this year are even smaller. The mean error variance of about 30 found in the real data means that 95% of the residuals do not differ from the theoretical values by more than 10 days. Even the maximal estimated error variance of about 100 means that 95% of the estimated variation is within 20 days. These considerations and the experimental findings of Baumgartner (1952) provide no biological grounds for discarding phenological observations that are less than a week or even 10 days off the estimated values because this is well within the natural unpredictable variability. Moreover, some extreme values can be missed because there might be other rather large residuals in the particular year that mask the outlier, e.g., the year 1952 in Figure 2. Thus, it follows that a distribution-free rule combined with a robust estimation performs better in the Monte-Carlo study and is also more effective in removing large error variances caused by extreme observations when observational data are analyzed. The Dixon rule is not efficient in removing error variance because it misses some extreme observations that contribute significantly to σ_e^2 and hence to the unreliability of the combined time series.

Other types of mistakes have a similar impact as MMs, such as misprints and transposed numbers. However, MM is a distinct feature, because it adds an error component of 30 days or a multiple of 30 days to the deviation of the observed value from the mean. This deviation is large relative to the variance attributed to biological variability. Therefore, a cut-off of 30 days from the expected value is an effective way to remove MMs and all mistakes that induce deviations of more than 30 days. Other observational and protocol errors that cause error components of less than 30 days cannot be distinguished from the variance attributable to biological variability. The Monte-Carlo analysis and inspection of observational time series demonstrate the danger of removing correct observations when distribution-based methods are used to detect outliers in small sample sizes. Improved characterization of the biological variability of phenological series is urgently needed to evaluate the danger of false identification of MMs in cases where a distinct bimodal distribution of phenophases is produced by an intermittant occurrence of environmental conditions unfavorable to phenological development, e.g., a cold spell when the buds in a part of the population have already broken. In addition, studies on the statistics of observational and protocol errors are needed. Such studies should help determine if a cut-off value of less than 30 days can be applied.

We conclude with two recommendations. First, we strongly recommend outlier detection when research is conducted with combined phenological data. In our view, one of the few mistakes that can be detected is MM because its deviation is much larger than the range of natural variability, even though the proportion of MMs to the total number of errors is unknown. The nonparametric 30-day residual rule in combination with a robust L_1 fit is a stable and adequate procedure to detect MMs and other extreme values. Second, after outlier removal, fitting a linear two-way mixed model by the TYPE I or REML estimation method to obtain a reliable continuous time series for further analysis is recommended.

Acknowledgments

The German weather service kindly provided phenological data. We thank Dr. Pincus, University of Potsdam for statistical advice. This work was funded by the HSP3 post-graduate program of the state of Brandenburg, Germany.

References

Barrodale, I. and F.D.K. Roberts. 1973. An improved algorithm for discrete L_1 linear approximation. SIAM J. Numer. Anal. 10: 839–848.

Barrodale, I. and F.D.K. Roberts. 1974. Algorithm 478: Solution of an overdetermined system of equations in the L_1 norm. Comm. ACM 17:319–320.

Barnett, V. and T. Lewis. 1996. Outliers in statistical data. John Wiley, New York, 584 p.

Baumgartner, A. 1952. Zur Phänologie von Laubhölzern und ihre Anwendung bei lokalklimatischen Untersuchungen. Berichte des DWD in der US-Zone 42:69–73.

Bloomfield, P. and W.L. Steiger. 1983. Least absolute deviations: theory, applications and algorithms. Progress in probability and statistics. Vol. 6. Birkhäuser, Boston, 349 p.

Chen, W.J., T.A. Black, P.C. Yang, et al. 1999. Effects of climatic variability on the annual carbon sequestration by a boreal aspen forest. Global Change Biol. 4:41–53.

Corbeil, R.R. and S.R. Searle. 1976. Restricted maximum likelihood estimation of variance components in the mixed model. Technometrics 18:31–38.

Dixon, W.J. 1950. Analysis of extreme values. Ann. Math. Stat. 21: 488–506.

Dodge, Y. 1987. Statistical data analysis based on the L_1-Norm and related methods. Elsevier Science Publishers, Amsterdam, 464 p.

Goulden, M.L., J.W. Munger, S.-M. Fan, B.C. Daube and S.C. Wofsky. 1996. Exchange of carbon dioxide by a deciduous forest: response to interannual climate variability. Science 271: 1576–1578.

Häkkinen, R., T. Linkosalo and P. Hari. 1995. Methods for combining phenological time series: application to bud burst in birch (*Betula pendula*) in Central Finland for the period 1896–1955. Tree Physiol. 15:721–726.

Hartley, H.O. and J.N.K. Rao. 1967. Maximum likelihood estimation for the mixed analysis of variance model. Biometrika 54:93–108.

Hartley, H.O., J.N.K. Rao and L.R. Lamotte. 1978. A simple synthesis-based method of variance component estimation. Biometrics 34:233–242.

Hemmerle, W.J. and H.O. Hartley. 1973. Computing maximum likelihood estimates for the mixed A.O.V. model using the W-transformation. Technometrics 15:819–831.

Hubert, M. 1997. The breakdown value of the L_1 estimator in contingency tables. Stat. Probab. Lett. 33:419–425.

Hubert, M. and P.J. Rousseeuw. 1997. Robust estimation with both continuous and binary regressors. J. Stat. Plan. Inter. 57:153–163.

Keeling, C.D., J.F.S. Chin and T.P. Whorf. 1996. Increased activity of northern vegetation inferred from atmospheric CO_2 measurements. Nature 382:146–149.

King, E.P. 1953. On some procedures for the rejection of suspected data. J. Am. Stat. Assoc. 48:531–533.

Kramer, K., A. Friend and I. Leinonen. 1996. Modelling comparison to evaluate the importance of phenology and spring frost damage for the effects of climate change on growth of mixed temperate-zone deciduous forests. Clim. Res. 7:31–41.

Kramer, K., I. Leinonen and D. Loustau. 2000. The importance of phenology for the evaluation of impact of climate change on growth of boreal, temperate and Mediterranean forests ecosystems: an overview. Int. J. Biometeorol. 44:67–75.

Linkosalo, T. 2000. Analyses of the spring phenology of boreal trees and its response to climate change. Ph.D. Thesis, Univ. Helsinki, Dept. For. Ecol. Publications, No. 22.

Linkosalo, T., T.R. Carter, R. Häkkinen and P. Hari. 2000. Predicting spring phenology and frost damage risk of *Betula* spp. under climatic warming: a comparison of two models. Tree Physiol. 20: 1175–1182.

Linkosalo, T., R. Häkkinen and P. Hari. 1996. Improving the reliability of a combined phenological time series by analysing observation quality. Tree Physiol. 16:661–664.

McKean, J.W. and R.M. Schrader. 1987*a*. Least absolute error analysis of variance. *In* Statistical Data Analysis Based on the L_1-Norm and Related Methods. Ed. Y. Dodge. Elsevier Science Publishers, Amsterdam, pp 297–305

McKean, J.W. and R.M. Schrader. 1987*b*. Coefficients of determination for least absolute deviation analysis. Stat. Probab. Lett. 5: 49–54.

Menzel, A. 1997. Phänologie von Waldbäumen unter sich ändernden Klimabedingungen—Auswertung der Beobachtungen in den Internationalen Phänologischen Gärten und Möglichkeiten der Modellierung von Phänodaten. Forstwissenschaftliche Fakultät der Universität München, München, 150 p.

Milliken, G.A. and D.E. Johnson. 1992. Analysis of messy data. Volume I: Designed experiments. Chapman and Hall, New York, 365 p.

Myneni, R.B., C.D. Keeling, C.J. Tucker, G. Asrar and R.R. Nemani. 1997. Increased plant growth in the northern latitudes from 1981 to 1991. Nature 386:698–702.

Patterson, H.D. and R. Thompson. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58:545–554.

Rencher, A.C. 2000. Linear models in statistics. John Wiley, New York, 578 p.

Rousseeuw, P.J. 1984. Least median of squares regression. J. Am. Stat. Assoc. 79:871–880.

Rousseeuw, P.J. and A.M. Leroy. 1987. Robust regression and outlier detection. John Wiley, New York, 329 p.

Schnelle, F. 1955. Pflanzenphänologie. Geest and Portig, Leipzig, 299 p.

Searle, S.R. 1971. Linear models. John Wiley, New York, 532 p.

Searle, S.R. 1987. Linear models for unbalanced data. John Wiley, New York, 536 p.

Thompson, R. 1969. Iterative estimation of variance components for non-orthogonal data. Biometrics 25:767–773.

Vassella, A. 1997. Phänologische Beobachtungen des Bernischen Forstdienstes von 1869–1882: Witterungseinflüsse und Vergleiche mit heutigen Beobachtungen. Phänologie von Waldbäumen: Historische und aktuelle Beobachtungen. Bern, Bundesamt für Umwelt, Wald und Landschaft, 73:9–75.

White, M.A., S.W. Running and P.E. Thornton. 1999. The impact of growing season length variability on carbon assimilation and evapotranspiration over 88 years in the eastern US deciduous forest. Int. J. Biometeorol. 42:139–145.