Supplementary data for

Identifying phenological phases in strawberry using multiple change-point models

Marc Labadie[1,2], **Béatrice Denoyes**[1] **and Yann Guédon**[2]

[1]UMR BFP, INRA, Université de Bordeaux, 33140 Villenave d'Ornon, France

[2]CIRAD, UMR AGAP and Université de Montpellier, 34098 Montpellier, France

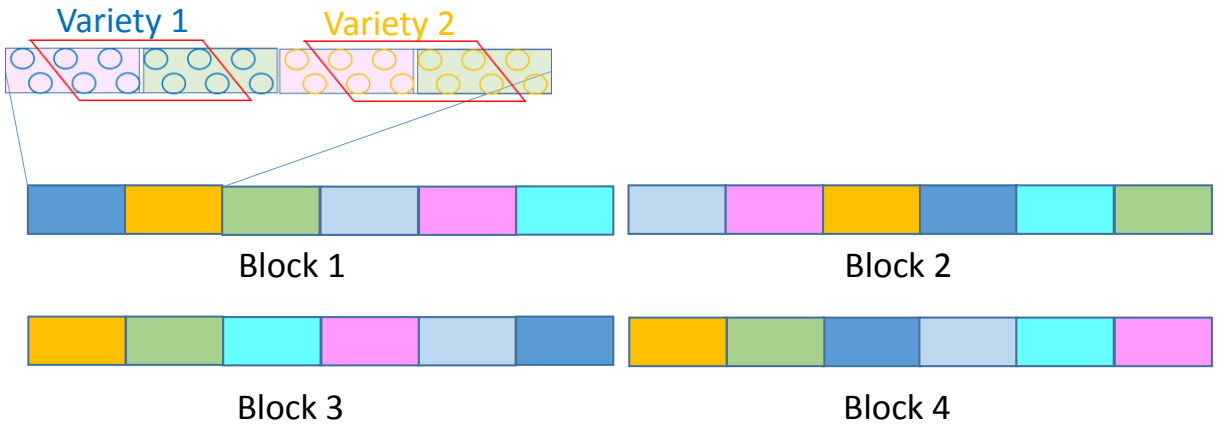

Figure S1. Design of the randomized block experiment. 48 plants were randomly distributed into four blocks for each of the six genotypes (represented by different colors). The 12 plants per block and genotype were distributed in two rows and two breeding ground bags (pink and green). To avoid border effect between neighboring plants belonging to different genotypes, only the height most central plants per block (red parallelepiped) were observed, leading to a total of 32 replicates per genotype.

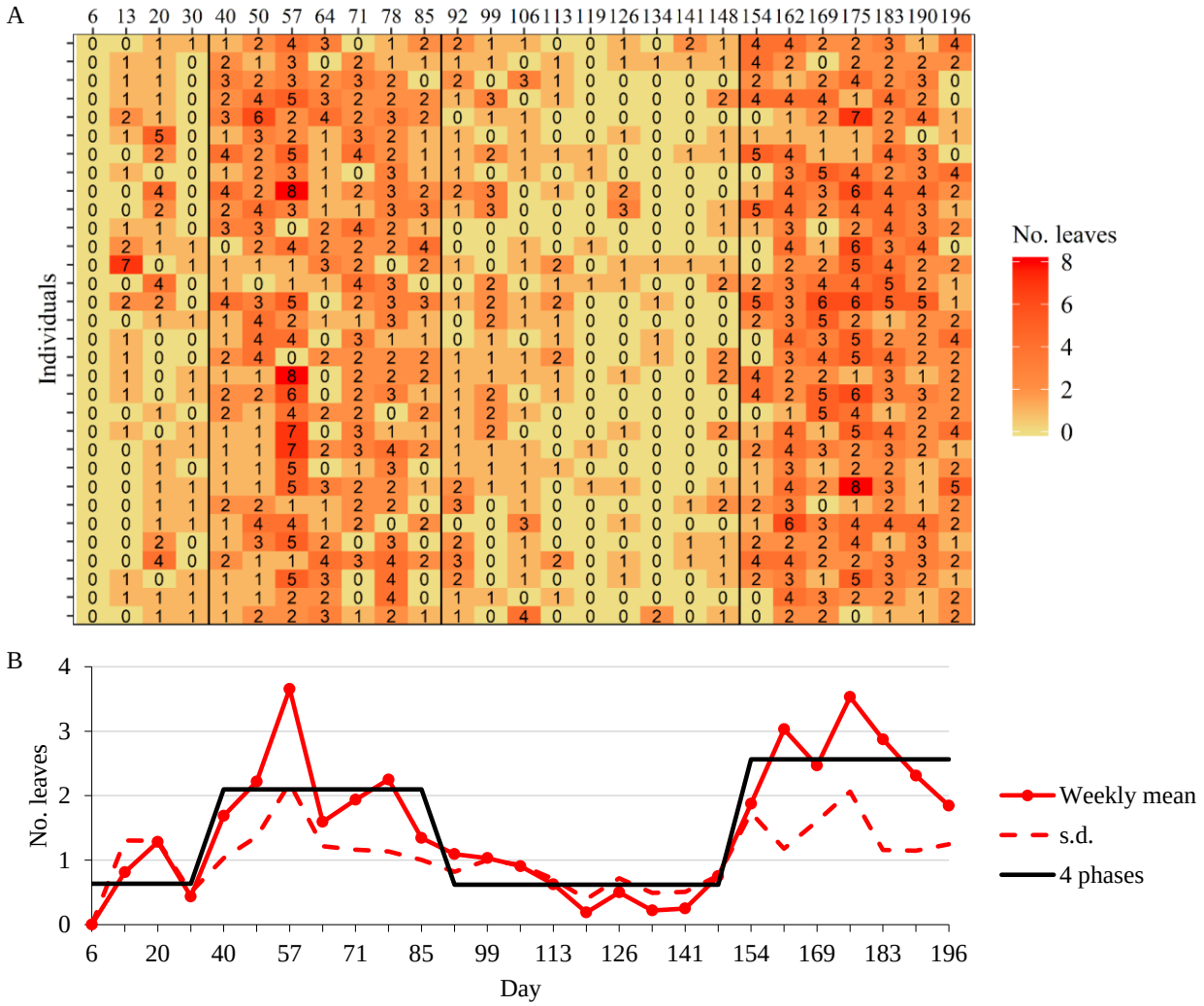

Figure S2. Vegetative development phases identified for Gariguette using univariate categorical multiple change-point models. (A) Heat map of the series of leaf production, the color scale ranging from light orange (low intensity) to red (high intensity). Vegetative development phases are delimited by black lines. (B) The optimal 4-phase segmentation is represented as a piecewise constant functions (black lines), the level of each phase corresponding to the mean number of weekly emerged leaves in the phase. The weekly mean numbers of emerged leaves are represented by red points connected by lines and the associated standard deviations (s.d.) by dashed red lines.

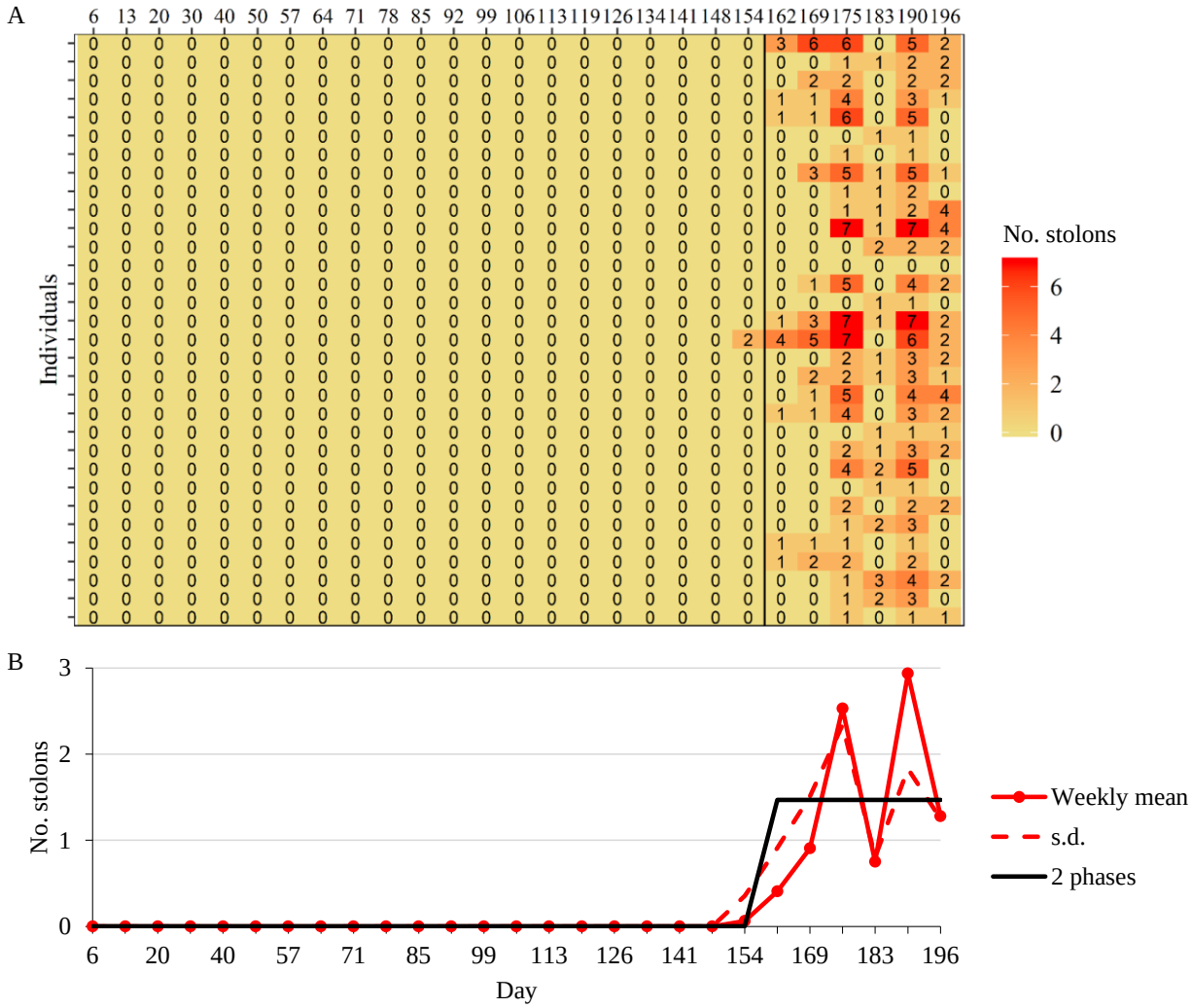

Figure S3. Runnering phases identified for Gariguette using univariate categorical multiple change-point models. (A) Heat map of the series of stolon production, the color scale ranging from light orange (low intensity) to red (high intensity). Runnering phases are delimitated by a black line. (B) The optimal 2-phase segmentation is represented as a piecewise constant functions (black lines), the level of each phase corresponding to the mean number of weekly emerged stolons in the phase. The weekly mean numbers of emerged stolons are represented by red points connected by lines and the associated standard deviations (s.d.) by dashed red lines.

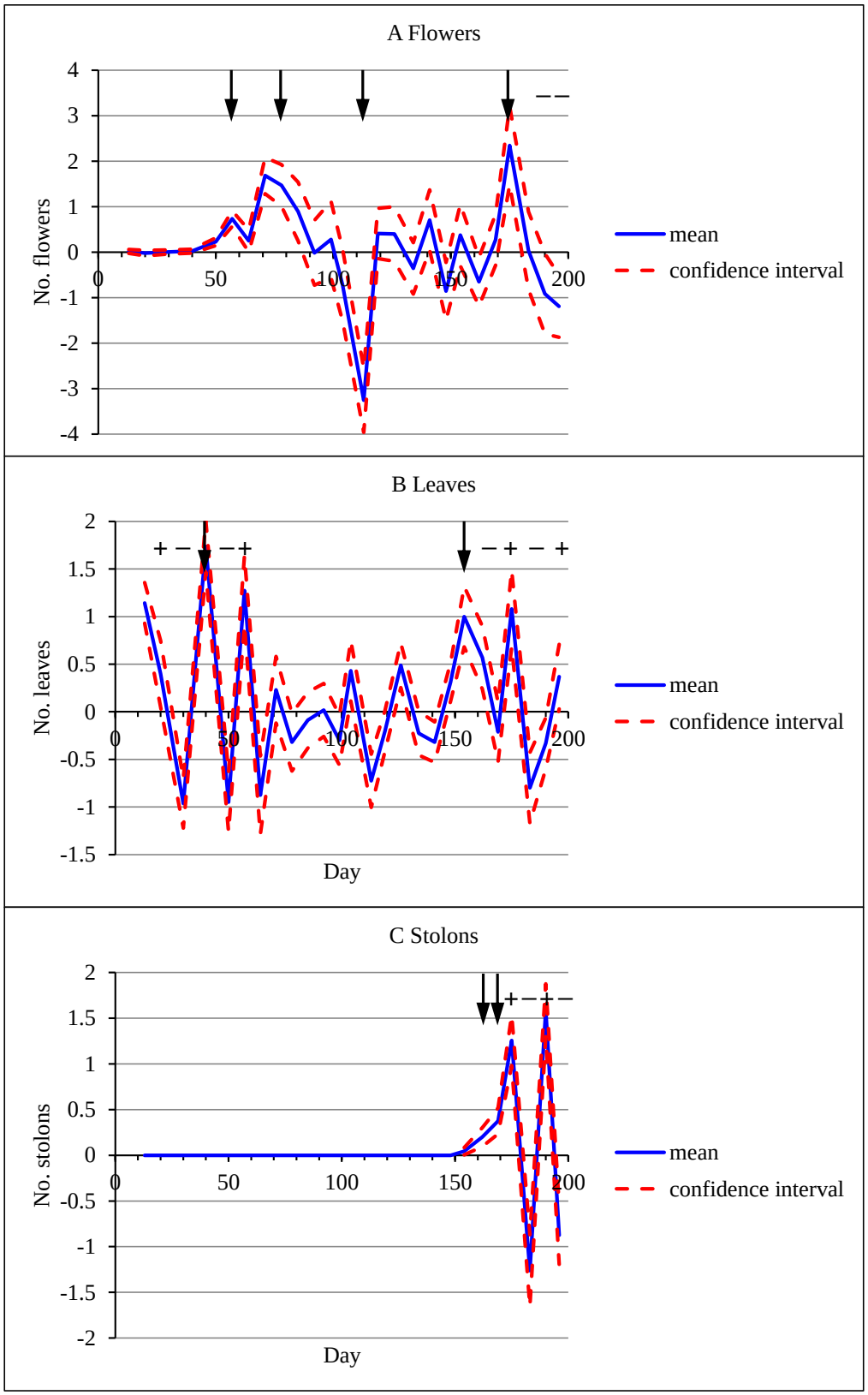

Figure S4. Weekly means with associated confidence intervals computed from the first-order differenced series of the numbers of weekly emerged (A) flowers, (B) leaves and (C) stolons for all the genotypes except Ciflorette. The arrows indicate fluctuations which are for a large part explained by limits between phases for several genotypes. The '+' and '–' indicate respectively other positive and negative fluctuations.

Table S1. Cumulative number of flowers, leaves, crowns and stolons (mean and standard deviation – s.d. – of the frequency distributions) produced per plant during the observation period (i.e. from December 16 2014 to June 24 2015), Spearman rank correlation coefficient between the cumulative number of flowers (respectively number of leaves) and the cumulative number of crowns (n.s. for non-significantly different from 0), chilling requirement (in hours) and flowering earliness (with ordered categories early, median, late) for the six genotypes (http://ciref-agriculture.fr/varietes-fraises-ciref and http://www.ctifl.fr/besoinsenfroid/pages/fraise/Besoins.aspx).

	Flower		Leaf		Crown		Stolon		Correlation		Chilling	Flowering
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	Flower/crown	Leaf/crown	requirement	earliness
Gariguette	72.7	13.7	41.6	8.4	3.1	0.9	8.4	5.9	0.38	0.56	800	early
Cléry	51.2	12.7	40.3	12.2	3.6	0.8	5.5	6.3	0.45	0.6	900	median
Cir107	80.8	23.4	57	11.7	4.5	0.7	11.8	5.9	n.s.	0.46	500	median
Darselect	52.8	11.1	32.6	8.6	3	0.7	3.8	5.2	n.s.	0.44	1000	late
Capriss	56.1	13.7	59.6	13.6	5.7	1.1	3	4.1	0.63	0.73	700	median
Ciflorette	61.3	14.4	46.1	8.6	4	0.8	15	8.3	0.44	0.53	800	early

Table S2. Frequency (Freq.) distributions of the limits between phases computed from the segmentation, asynchronous between individuals, of the series of flower production using hidden semi-Markov chains. The '*' indicate the limits between phases given by the synchronous segmentations of flowering series using multiple change-point models.

	Limit	Freq.	Limit	Freq.	Limit	Freq.	Limit	Freq.	Limit	Freq.
Gariguette	50*	15	64	2	92	6	113	10	169	4
	57	14	71*	27	99*	22	119*	19	175*	17
	64	3	78	3	104	3	126	2	183	11
					113	1	134	1		
Cléry	40	1	64	1			99	1	169	4
	50	7	71	10			104	7	175*	23
	57*	20	78*	18			113*	19	183	5
	64	3	85	2			119	4		
	71	1	92	1			126	1		
Cir107	50	10	64	3			99	1		
	57*	18	71	14			104	12		
	64	4	78*	10			113*	18		
			85	5			119	1		
Darselect	50	7	71	4			104	10		
	57	6	78*	14			113*	18		
	64*	13	85	11			119	4		
	71	5	92	2						
	78	1	99	1						
Capriss	50	2	71	3			104	10		
	57*	19	78*	11			113*	19		
	64	11	85	11			119	3		
			92	7						
Ciflorette	56*	32	70	4			91	4		
			77*	21			98	9		
			84	6			105*	18		
			91	1			110	1		

Table S3. Frequency (Freq.) distributions of the limits between phases computed from the segmentation, asynchronous between individuals, of the series of leaf production using hidden semi-Markov chains. The '*' indicate the limits between phases given by the synchronous segmentations of vegetative development series using multiple change-point models.

	Limit	Freq.	Limit	Freq.	Limit	Freq.	Limit	Freq.
Gariguette	40*	26	85	4			148	4
	50	6	92*	21			154*	13
			99	6			162	15
			104	1				
Cléry	30	7			99	7	154*	7
	40*	23			104	23	162	23
	50	2			113*	2	169	2
Cir107							148	13
							154*	12
							162	7
Darselect							154	2
							162*	17
							169	10
							175	2
							183	1

Table S4. Frequency (Freq.) distributions of the limits between phases computed from the segmentation, asynchronous between individuals, of the series of stolon production using hidden semi-Markov chains. The '*' indicate the limits between phases given by the synchronous segmentations of runnering series using multiple change-point models.

	Limit	Freq.	Limit	Freq.
Gariguette			154	1
			162*	7
			169	5
			175	13
			183	6
Cléry	154*	1	169*	1
	162	1	175	2
	169	6	183	4
	175	11	190	24
	183	13	196	1
Cir107	154	2	175*	4
	162*	7	183	5
	169	8	190	21
	175	13	196	2
	183	2		
Darselect			154	1
			162*	3
			169	5
			175	4
			183	5
			190	14
Capriss			162	1
			169*	2
			175	10
			183	4
			190	15
Ciflorette			154	1
			160*	6
			168	10
			175	15

Protocol S1. Definition of categorical multiple change-point models and associated statistical methods

For a given genotype, multiple change-point models were used to delimit phases within a sample of phenological series of fixed length. Each series corresponds to a plant and may be univariate (emergence rate for a given organ) or multivariate (emergence rates for different organs such as flower, leaf and stolon in our case). These series are indexed by the successive dates of observation (with the convention that the first date of observation is 1 for notational convenience). Let θ denote the parameters of the categorical distributions attached to the successive phases (i.e. the probability masses for the possible number of weekly emerged organs). Let $f_J(\mathbf{s}, \mathbf{x}; \hat{\theta})$ denote the likelihood of the segmentation **s** of the observed series $\mathbf{x} = x_1, \ldots, x_T$. The estimation of the $J-1$ change points $\tau_1, \ldots, \tau_{J-1}$(with the convention $\tau_0 = 1$ and $\tau_J = T + 1$ where T is the last date of measurement), which corresponds to the optimal segmentation $\mathbf{s}^*$ into *J* flowering phases, is obtained using a dynamic programming algorithm (Auger and Lawrence, 1989) that solves the following optimization problem:

$$\hat{\tau}_1, \ldots, \hat{\tau}_{J-1} = \arg\max_{\mathbf{s}} \log f_J\left(\mathbf{s}, \mathbf{x}; \hat{\theta}\right),$$

Regarding the inference of multiple change-point models, one key question is to select the number of phases. In a model selection context, the purpose is to estimate *J* by maximizing a penalized version of the log-likelihood defined as follows

$$\hat{J} = \arg\max_{J} \left\{\log f_J(\mathbf{x}) - \text{Penalty}(J)\right\},$$

where

$$f_J(\mathbf{x}) = \sum_{\mathbf{s}} f_J(\mathbf{s}, \mathbf{x}; \hat{\theta})$$

is the log-likelihood of all the possible segmentations in *J* phases of the phenological series **x** of length *T*. The principle of this kind of penalized likelihood criterion consists in making a trade-off between an adequate fitting of the model to the data (expressed by the log-likelihood) and a reasonable number of parameters to be estimated (controlled by the penalty term). The most popular information criteria such as AIC and BIC are not adapted in this particular context since they tend to underpenalize the log-likelihood and thus select a too large number of phenological phases (Rigaill *et al*., 2016). We thus applied the slope heuristic (SH) given by (Guédon, 2015b)

$$\mathrm{SH}_J = 2\{\log f_J(\mathbf{x}) - 2\,\hat{\kappa}\,\mathrm{pen}_{\mathrm{shape}}(J)\},$$

where

$$\mathrm{pen}_{\mathrm{shape}}(J) = \log\left\{\frac{T^{J-1}}{(J-1)!}\right\},$$

and $\hat{\kappa}$ is the slope of the linear relationship between $\log f_J(\mathbf{x})$ and $\mathrm{pen}_{\mathrm{shape}}(J)$ for overparameterized models estimated by the data-driven slope estimation method (Baudry *et al*., 2012). The posterior probability of the *J*-phase model M_J, given by

$$P(M_J|\mathbf{x}) = \frac{\exp\left(\frac{1}{2}\mathrm{SH}_J\right)}{\sum_{K=1}^{J_{\max}} \exp\left(\frac{1}{2}\mathrm{SH}_K\right)},$$

can be used to assess the relative merits of the models considered.

The posterior probability of the optimal segmentation $\mathbf{s}^*$ given by

$$P(\mathbf{s}^*|\mathbf{x}; J) = f_J(\mathbf{s}^*, \mathbf{x}; \hat{\theta}) / \sum_{\mathbf{s}} f_J(\mathbf{s}, \mathbf{x}; \hat{\theta}),$$

can be efficiently computed by the smoothing algorithm proposed by Guédon (2013). The assessment of multiple change-point models thus relies on two posterior probabilities:

- posterior probability of the *J*-phase model M_J, $P(M_J|\mathbf{x})$ deduced from the slope heuristic computed for a collection of multiple change-point models for $J = 1, \ldots, J_{\max}$, i.e. weight of the *J*-phase model among all the possible models between 1 and $J_{\max}$ phenological phases,
- posterior probability of the optimal segmentation $\mathbf{s}^*$ for a fixed number of phases *J* $P(\mathbf{s}^*|\mathbf{x}; J)$, i.e. weight of the optimal segmentation among all the possible segmentations for a fixed number of phases.

It is often of interest to quantify the uncertainty concerning change-point position. To this end, we computed the posterior change-point probabilities for each change point *j* and each observation date *t* using the smoothing algorithm proposed by Guédon (2013). We define the interval with credibility $1 - \alpha$ for change point *j* as the interval such that,

$$\alpha/2 < \sum_{t=u}^{v} P(S_t = j, S_{t-1} = j-1|\mathbf{x}; J) < 1 - \alpha/2,$$

with $\sum_{t=j+1}^{T-J+j} P(S_t = j, S_{t-1} = j-1|\mathbf{x}; J) = 1$.

Protocol S2. Definition of hidden semi-Markov chains and associated statistical methods

Semi-Markov chains

Let $\{S_t\}$ be a semi-Markov chain with finite-state space $\{0, \dots, J-1\}$. A *J*-state semi-Markov chain $\{S_t\}$ is defined by the following parameters:

- initial probabilities $\pi_j = P(S_1 = j)$ with $\sum_j \pi_j = 1$;
- transition probabilities
  - nonabsorbing state *i*: for each $j \neq i$, $p_{ij} = P(S_t = j | S_t \neq i, S_{t-1} = i)$ with $\sum_{j \neq i} p_{ij} = 1$ and $p_{ii} = 0$ by convention,
  - absorbing state *i*: $p_{ii} = P(S_t = i | S_{t-1} = i) = 1$ and for each $j \neq i$, $p_{ij} = 0$.

An explicit occupancy distribution is attached to each nonabsorbing state:

$$d_j(u) = P(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, \dots, u-2 | S_{t+1} = j, S_t \neq j), \qquad u = 1,2, \dots$$

Since $t = 1$ is assumed to correspond to a state entering, the following relation is verified:

$$P(S_t \neq j, S_{t-v} = j, v = 1, \dots, t) = d_j(t)\pi_j.$$

We define as possible parametric state occupancy distributions binomial distributions, Poisson distributions and negative binomial distributions with an additional shift parameter *d* ($d \geq 1$) which defines the minimum sojourn time in a given state.

The binomial distribution with parameters *d*, *n* and *p* ($q = 1 - p$), B(*d*, *n*, *p*) where $0 \leq p \leq 1$, is defined by

$$d_j(u) = \binom{n-d}{u-d} p^{u-d} q^{n-u}, \qquad u = d, d+1, \dots, n.$$

The Poisson distribution with parameters *d* and λ, P(*d*, λ), where λ is a real number ($\lambda > 0$), is defined by:

$$d_j(u) = \frac{e^{-\lambda} \lambda^{u-d}}{(u-d)!}, \qquad u = d, d+1, \dots$$

The negative binomial distribution with parameters *d*, *r* and *p*, NB(*d*, *r*, *p*), where *r* is a real number ($r > 0$) and $0 < p \leq 1$, is defined by:

$$d_j(u) = \binom{u-d+r-1}{r-1} p^r q^{u-d}, \qquad u = d, d+1, \dots$$

Hidden semi-Markov chain

A hidden semi-Markov chain can be viewed as a pair of stochastic processes $\{S_t, X_t\}$ where the "output" process $\{X_t\}$ is related to the "state" process $\{S_t\}$, which is a finite-state semi-Markov chain, by a probabilistic function or mapping denoted by *f* (hence $X_t = f(S_t)$). Since the mapping *f* is such that a given output may be observed in different states, the state process $\{S_t\}$ is not observable directly but only indirectly through the output process $\{X_t\}$. This output process $\{X_t\}$ is related to the semi-Markov chain $\{S_t\}$ by the observation (or emission) probabilities $b_j(y) = P(X_t = y|S_t = j)$. The definition of the categorical observation distributions expresses the assumption that the output process at time *t* depends only on the underlying semi-Markov chain at time *t*.

The maximum likelihood estimation of the parameters of a hidden semi-Markov chain requires an iterative optimization technique, which is an application of the EM algorithm. Once a hidden semi-Markov chain has been estimated, the most probable state series can be computed for each observed series using the so-called Viterbi algorithm; see Guédon (2003, 2005, 2007) for the statistical methods for hidden semi-Markov chains. In our application context, the most probable state series can be interpreted as the optimal segmentation of the corresponding observed series into successive phenological phases.

References

Auger IE, Lawrence CE. 1989. Algorithms for the optimal identification of segment neighborhoods. Bulletin of Mathematical Biology **51,** 39-54.

Baudry J-P, Maugis C, Michel B. 2012. Slope heuristics: overview and implementation. *Statistics and Computing* **22,** 455-470.

Guédon Y. 2003. Estimating hidden semi-Markov chains from discrete sequences. Journal of Computational and Graphical Statistics **12,** 604−639.

Guédon Y. 2005. Hidden hybrid Markov/semi-Markov chains. Computational Statistics & Data Analysis **49,** 663–688.

Guédon Y. 2007. Exploring the state sequence space for hidden Markov and semi-Markov chains. Computational Statistics & Data Analysis **51,** 2379−2409.

Guédon Y. 2013. Exploring the latent segmentation space for the assessment of multiple change-point models. Computational Statistics **28,** 2641-2678.

Guédon Y. 2015a. Segmentation uncertainty in multiple change-point models. Statistics and Computing **25,** 303-320.

Guédon Y. 2015b. Slope heuristics for multiple change-point models. In: *30th International Workshop on Statistical Modelling* (IWSM 2015). Friedl H, Wagner H. eds., vol. **2**, 103-106.

Rigaill G, Lebarbier E, Robin S. 2012. Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and Computing* **22,** 917-929.