New insights for estimating the genetic value of F1 apple progenies for irregular bearing during first years of tree production

Jean-Baptiste Durand[1,2(*)], **Baptiste Guitton**[3], **Jean Peyhardi**[2,4], **Yan Holtz**[5], **Yann Guédon**[2], **Catherine Trottier**[4], **Evelyne Costes**[5]

Corresponding author: Jean-Baptiste Durand (+33 4 76 63 57 09), jean-baptiste.durand@imag.fr

Supplementary Information

Figures

Fig. S1. Schematic representation of observations performed on 6 years old trees on the trunk, Long Sylleptic Axillary Shoot (LSAS), Long Proleptic Axillary Shoot (LSAS) and Sort Axillary Shoot (SAS). Annual shoots are delimited by "=" and death of the shoot apical meristem is represented by "x". (A) Example of a 5 years long sequence bearing no flower, (B) biennial sequence and (C) regular sequence bearing only flowers.

Fig. S2. Absolute difference between consecutive yields $|Y_t - Y_{t-1}|$as a function of Y_t. Points with $Y_t = 0$ or $Y_{t-1} = 0$ have been removed. The correlation is 0.67, with 95% confidence interval (0.63;0.70), which shows that the implicit hypothesis underlying BBI, i.e. the alternation amplitudes given by the residuals are roughly proportional to the corresponding trend level, is roughly satisfied. Observations associated with regular genotypes are in red, biennial genotypes in green and irregular genotypes in blue. Two main directions are of particular significance: $Y_t \approx 0$ ($|Y_t - Y_{t-1}| \approx Y_{t-1}$ in this case) and $Y_{t-1} \approx 0$ ($|Y_t - Y_{t-1}| \approx Y_t$ in this case), which both are typical cases of alternation (most points aligned on these directions are from biennial or irregular genotypes).

Fig. S3. Empirical and predicted residuals of yields as a function of time for regular bearing genotype g=85.

Fig. S4. Empirical and predicted residuals of yields as a function of time for biennial bearing genotype g=107.

Fig. S5. Empirical and predicted residuals of yields as a function of time for alternate bearing genotype g=108.

Fig. S6. Measurements and predicted yield values in 2010 (year number 5) for regular bearing genotype g=85 (a) and irregular bearing genotype g=108 (b). Circles are the measured values; triangles are the predicted values, located in the middle of prediction intervals (dotted segments).

Fig. S7. Plot of genotypes in the first FDA plane, based on mean entropy and local indices B^{loc} and local genotype AR coefficient γ^{loc}. The three colours indicate to which cluster each genotype belongs, according to the previous clustering performed with genotype AR coefficient and BBI_res_norm.

Fig. S1. Schematic representation of observations performed on 6 years old trees on the trunk, Long Sylleptic Axillary Shoot (LSAS), Long Proleptic Axillary Shoot (LSAS) and Sort Axillary Shoot (SAS). Annual shoots are delimited by "=" and death of the shoot apical meristem is represented by "x". (A) Example of a 5 years long sequence bearing no flower, (B) biennial sequence and (C) regular sequence bearing only flowers.

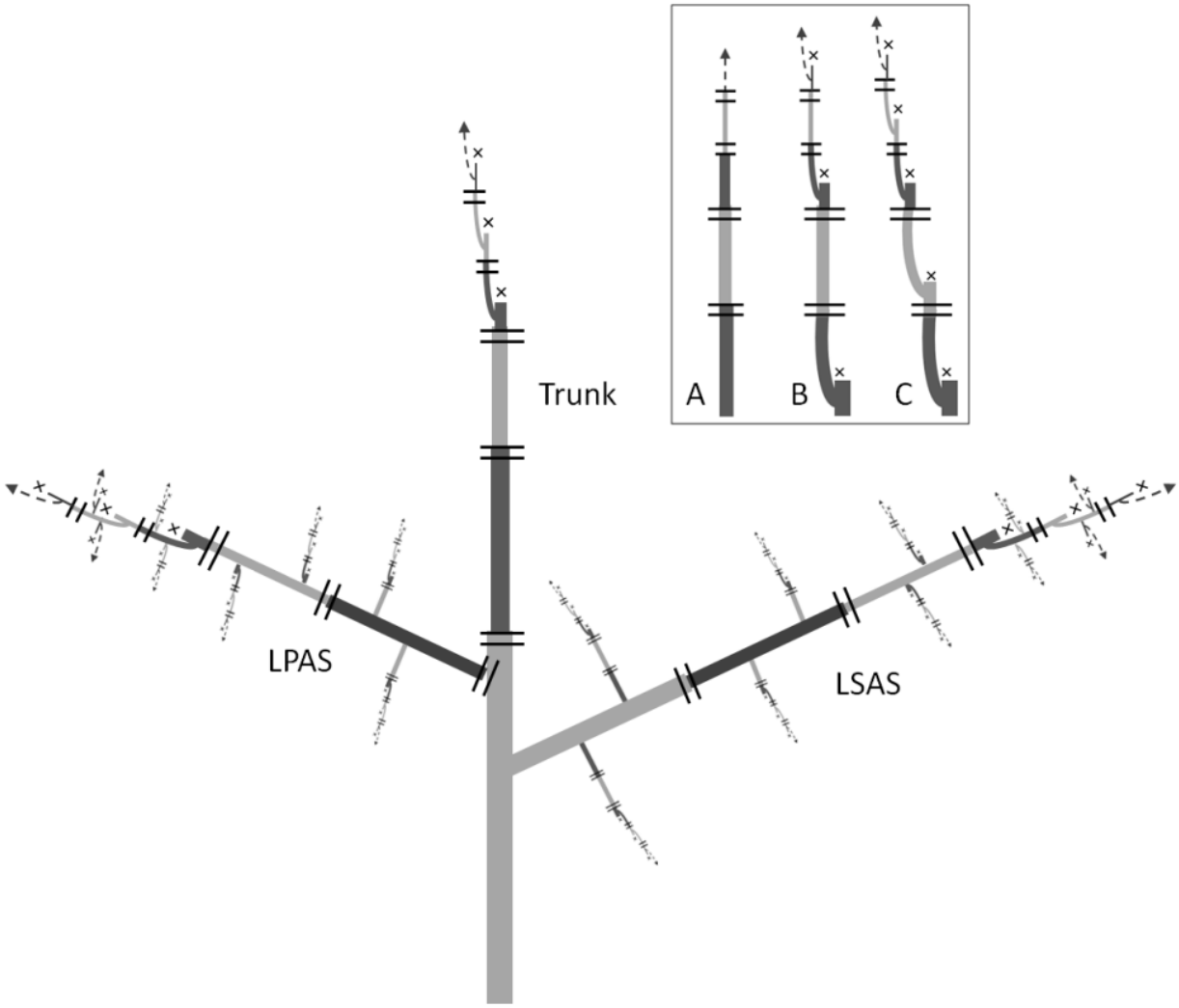


Fig. S2. Absolute difference between consecutive yields $|Y_t - Y_{t-1}|$as a function of Y_t. Points with $Y_t = 0$ or $Y_{t-1} = 0$ have been removed. The correlation is 0.67, with 95% confidence interval (0.63;0.70), which shows that the implicit hypothesis underlying BBI, i.e. the alternation amplitudes given by the residuals are roughly proportional to the corresponding trend level, is roughly satisfied. Observations associated with regular genotypes are in red, biennial genotypes in green and irregular genotypes in blue. Two main directions are of particular significance: $Y_t \approx 0$ ($|Y_t - Y_{t-1}| \approx Y_{t-1}$ in this case) and $Y_{t-1} \approx 0$ ($|Y_t - Y_{t-1}| \approx Y_t$ in this case), which both are typical cases of alternation (most points aligned on these directions are from biennial or irregular genotypes).

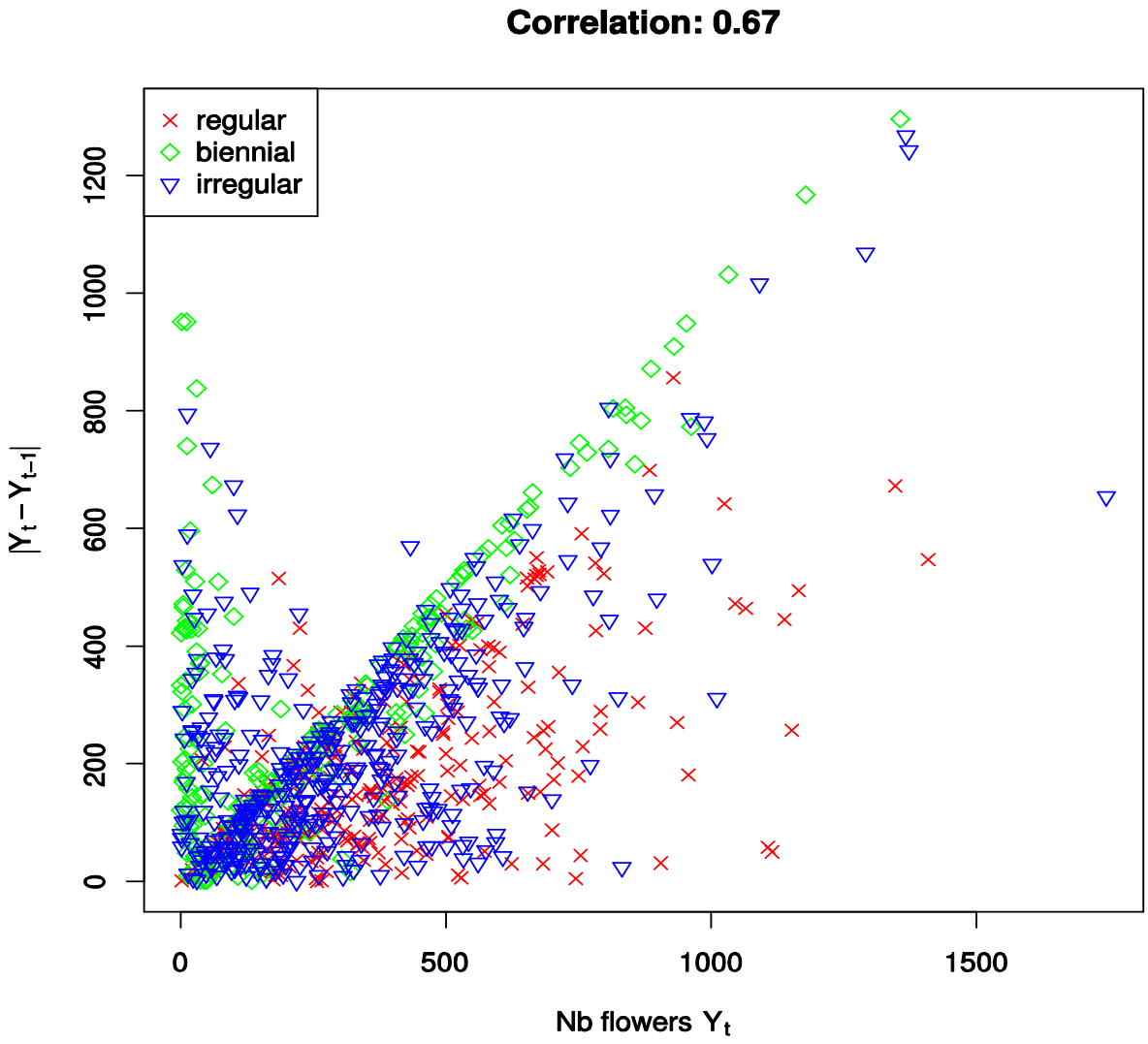


Fig. S3. Empirical and predicted residuals of yields as a function of time for regular bearing genotype g=85.

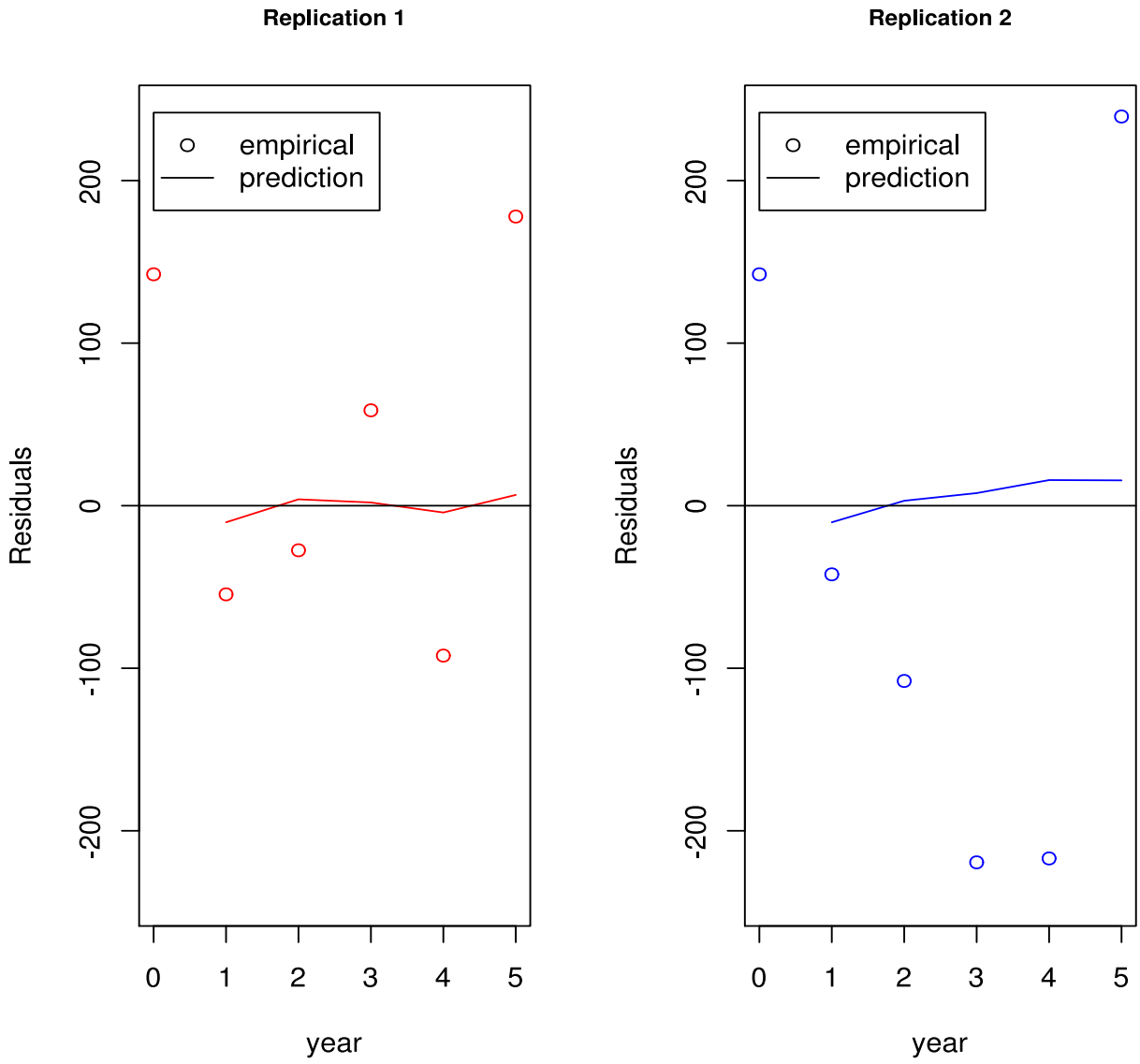


Fig. S4. Empirical and predicted residuals of yields as a function of time for biennial bearing genotype g=107.

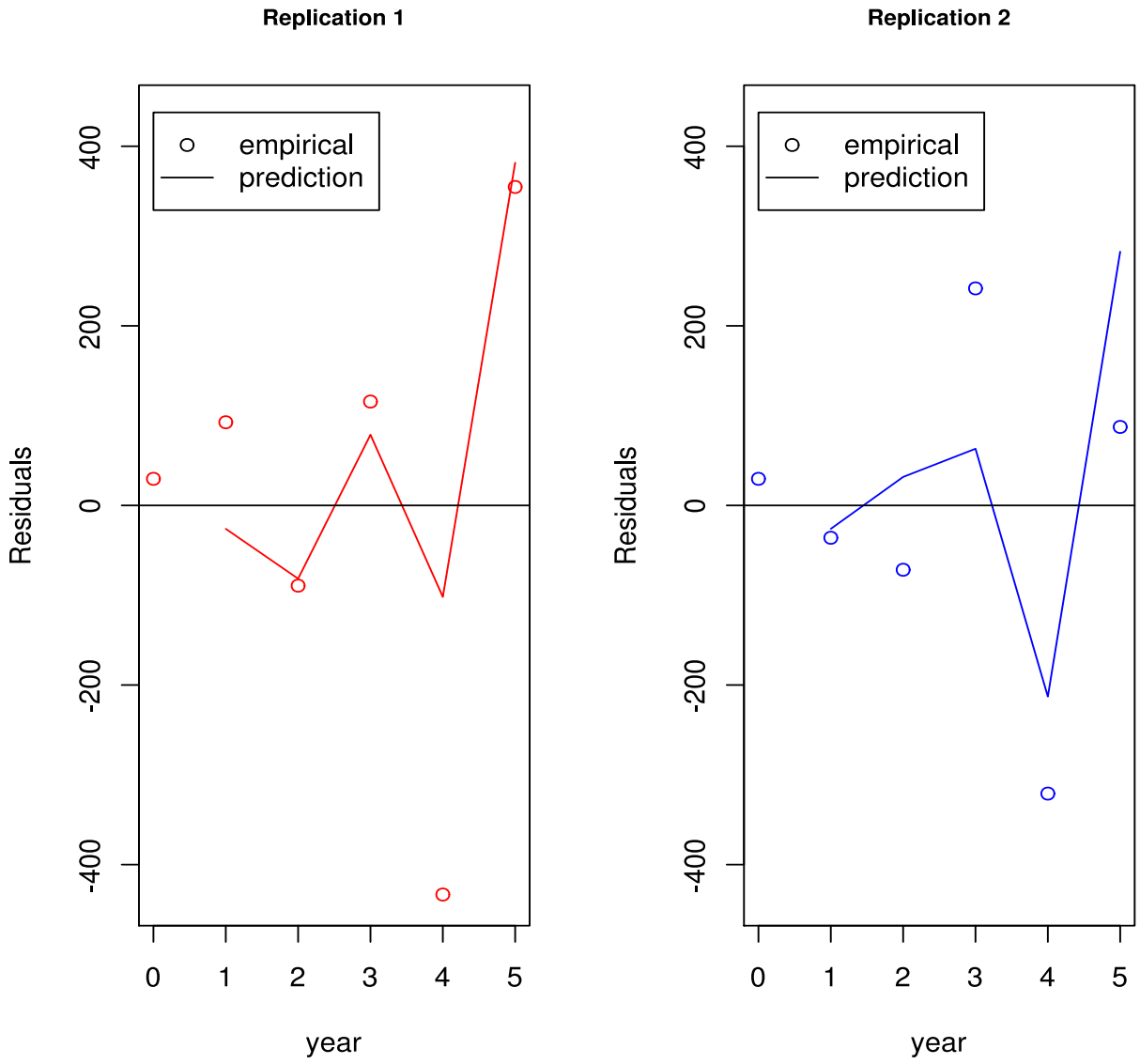


Fig. S5. Empirical and predicted residuals of yields as a function of time for irregular bearing genotype g=108.

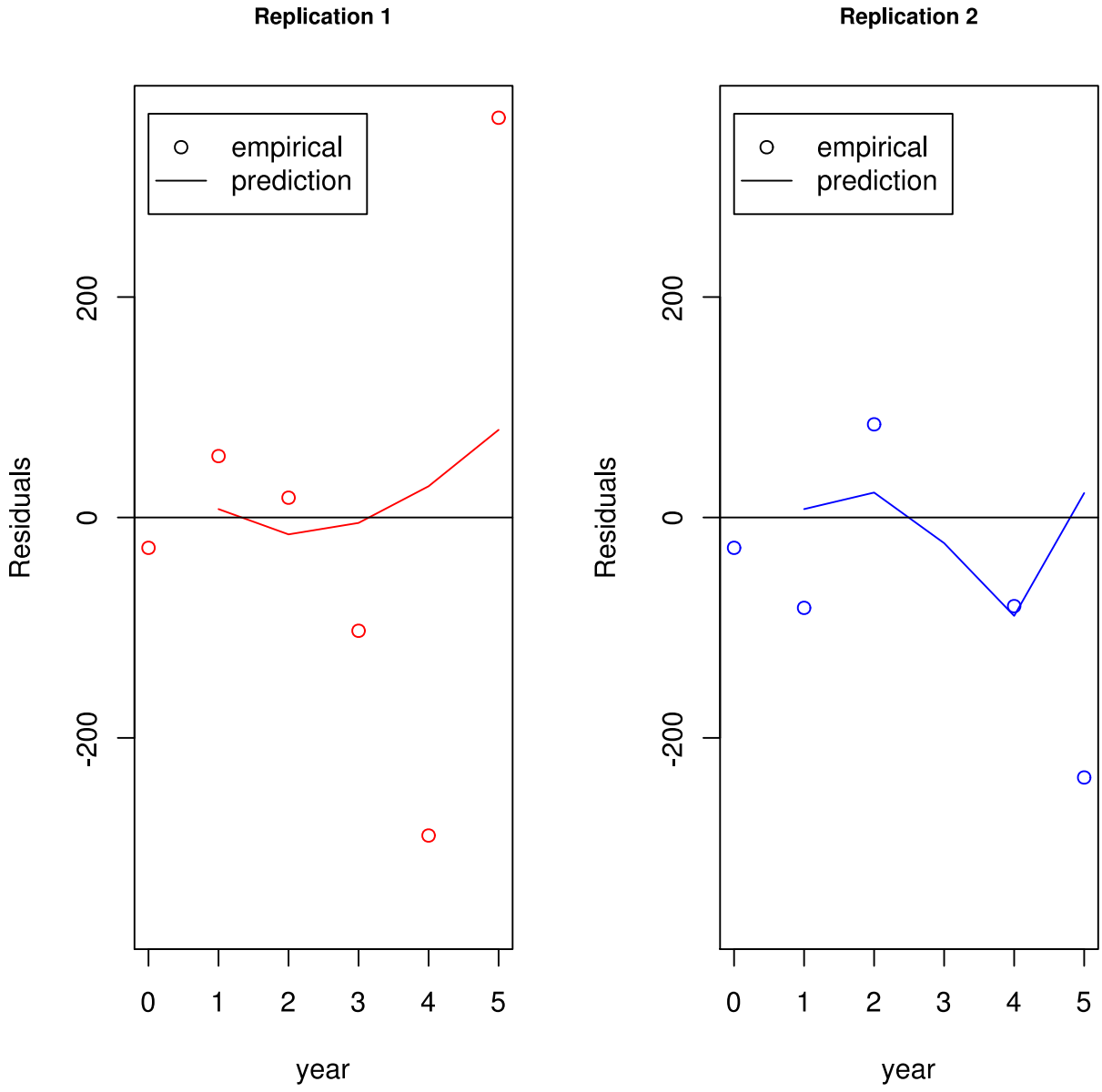


Fig. S6a. Measurements and predicted yield values in 2010 (year number 5) for regular bearing genotype g=85 (a). Circles are the measured values; triangles are the predicted values, located in the middle of prediction intervals (dotted segments).

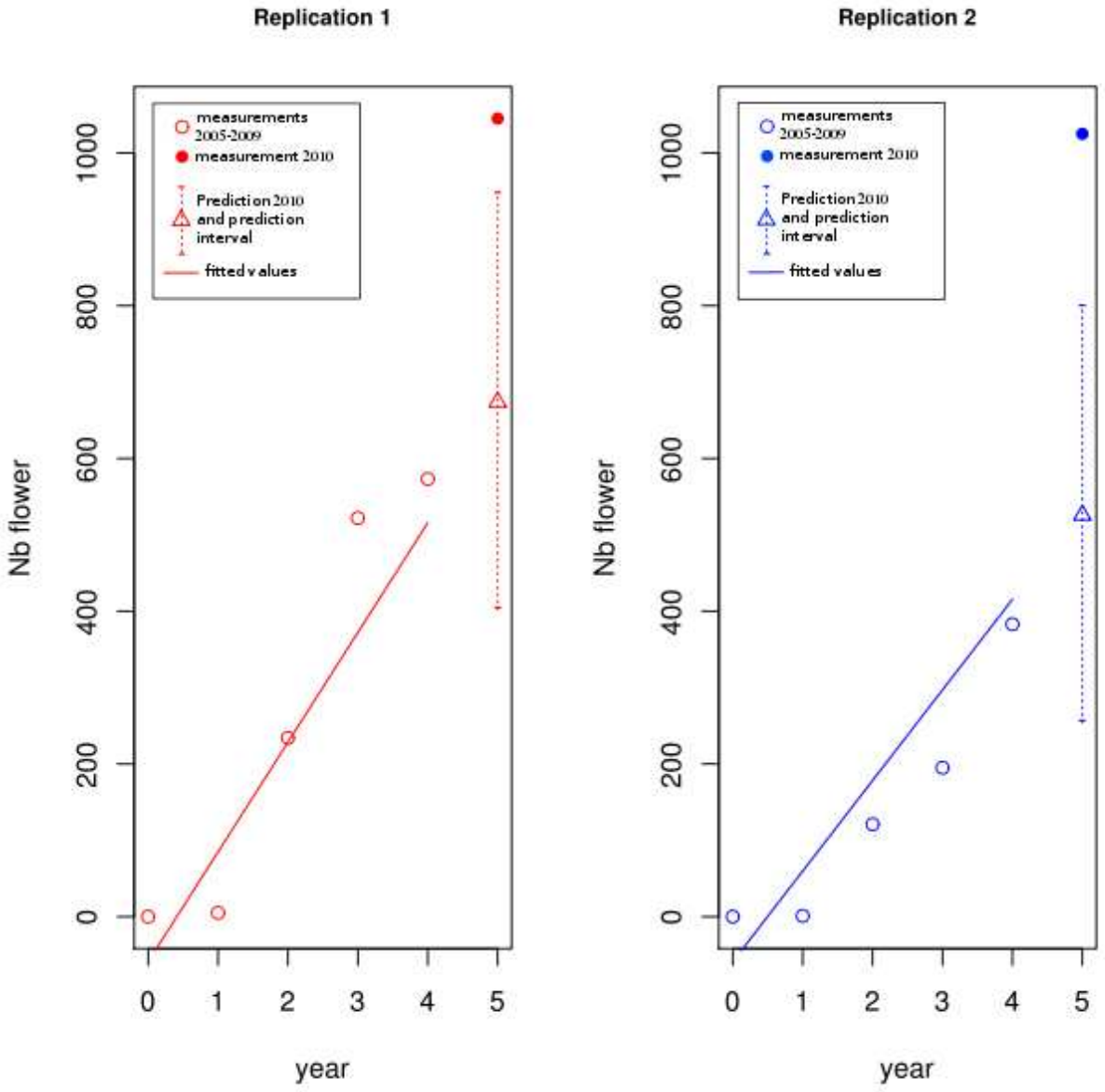


Fig. S6b. Measurements and predicted yield values in 2010 (year number 5) for irregular bearing genotype *g*=108 (b). Circles are the measured values; triangles are the predicted values, located in the middle of prediction intervals (dotted segments).

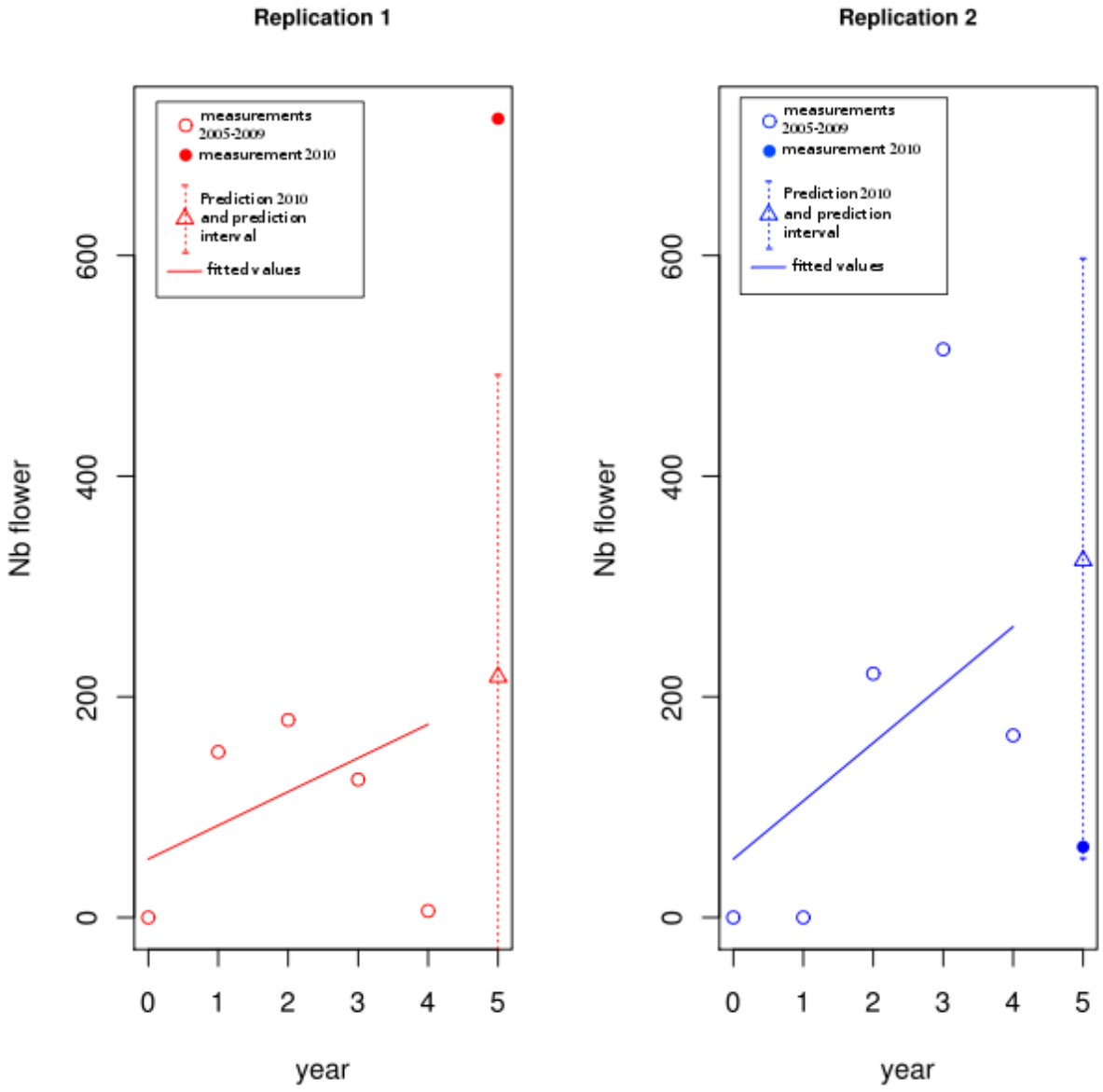
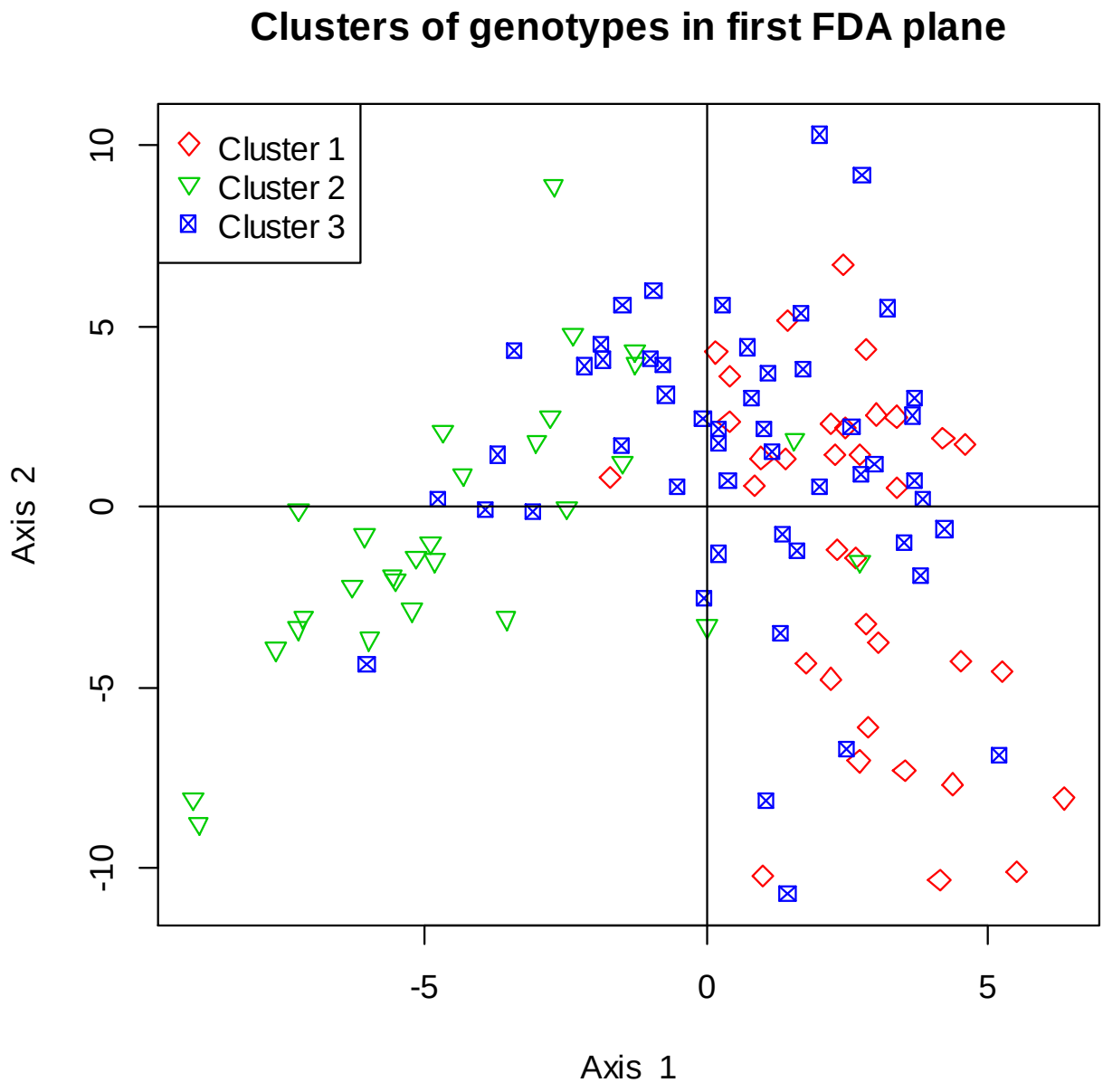


Fig. S7. Plot of genotypes in the first FDA plane, based on mean entropy and local indices B^{loc} and local genotype AR coefficient γ^{loc}. The three colours indicate to which cluster each genotype belongs, according to the previous clustering performed with genotype AR coefficient and BBI_res_norm.



Tables

Table T1. Computation of entropies to quantify synchronism in flowering for three genotypes *g*: regular bearing (*g*=85), biennial bearing (*g*=107) and irregular bearing (*g*=108). For each year, the frequency of flowering $F_{g,r,t}$ and the contribution $Ent_{g,r}$ to the average entropy are given.

Genotypes		**Year**					**Entropy**
		2005	**2006**	**2007**	**2008**	**2009**	
***g*=85**	**Number of GUs**	2	6	10	14	18	Total 50
	$F_{g,r,t}$	0.00	0.00	0.20	0.64	0.72	
	$Ent_{g,r}$	0.00	0.00	0.50	0.65	0.59	0.50
***g*=107**	**Number of GUs**	2	5	9	11	11	Total 38
	$F_{g,r,t}$	0.00	0.60	0.0	1.0	0.0	
	$Ent_{g,r}$	0.00	0.67	0.00	0.00	0.00	0.09
***g*=108**	**Number of GUs**	5	9	13	15	12	Total 54
	$F_{g,r,t}$	0.00	0.22	0.15	0.53	0.00	
	$Ent_{g,r}$	0.00	0.53	0.43	0.69	0.00	0.38

Table T2. Contingency table for the number of genotypes being in cluster c1 in the validation dataset and in cluster c2 in the whole dataset. Clusters are determined by a Gaussian mixture model, based on years 2005-2009 in the case of c1 and 2005-2010 in the case of c2. The numbers in parentheses indicate non-significant switches for genotypes at the boundary between clusters. Cluster 1 corresponds to regular genotypes, cluster 2 to biennial bearing and cluster 3 to irregular genotypes.

		Cluster based on years 2005-2010		
		1	**2**	**3**
Cluster based on years 2005-2009	**1**	**28**	**0**	**3 (2)**
	2	**1**	**25**	**5**
	3	**5 (2)**	**6**	**44**

Table T3. Correlation coefficient between indices at whole tree and AS scales, with 95% confidence intervals. Indices at whole tree scale are computed on the validation set (first 5 years of yield).

	Genotype AR coefficient g_g	**Local BBI_res_norm**	**Local genotype AR coefficient**	**Mean entropy**
BBI_res_norm	-0.62 (-0.72;-0.50)	0.65 (0.54;0.75)	-0.47 (-0.60;-0.31)	-0.52 (-0.65;-0.38)
g_g	1	-0.52 (-0.65, -0.38)	0.55 (0.41;0.67)	0.34 (0.17;0.49)

Table T4. Contingency table for the number of genotypes assigned to class c2 by Gaussian Mixture Clustering on local indices and assigned to class c1 by Gaussian Mixture Clustering on global indices. Cluster 1 corresponds to regular genotypes, cluster 2 to biennial bearing and cluster 3 to irregular genotypes.

		"True" class c1		
		1	**2**	**3**
Predicted class c2	**1**	17	1	9
	2	0	20	5
	3	18	8	37

Table T5. Precision $\left|BBI(t)-\alpha(t)\right|\big/BBI(t)$ of approximation of the BBI by $\alpha(t)=\log t/(t-1)$ in the case of affine growth of Y_t, as a function of the slope *a* and the length *t* of the time series.

t	*a*			
	0.1	1	10	100
5	4.19	0.48	0.07	0.03
25	1.78	0.26	0.04	0.01
400	0.66	0.13	0.02	0.008

Table T6. Precision $\left|BBI\_res\_norm(t)-2\right|/BBI\_res\_norm(t)$ of the approximation of BBI_res_norm by its limit 2 in the case of linear growth of alternate yield Y_t, as a function of the slope *a* and the length *t* of the time series.

t	*a*			
	0.1	1	10	100
5	0.2	0.2	0.2	0.2
25	0.04	0.04	0.04	0.04
400	0.002	0.002	0.002	0.002

A. Analytic properties of BBI

Let us recall that BBI is defined for a sample by:

$$\mathrm{BBI} = \frac{2}{\sum_r (T_{g,r} - 1)} \sum_r \sum_{t=2}^{T_{g,r}} \frac{\left| Y_{g,r,t} - Y_{g,r,t-1} \right|}{Y_{g,r,t-1} + Y_{g,r,t}}$$

where $T_{g,r}$ denotes the number of measurements for replication *r* of genotype *g*, and with the convention that

$$\frac{\left| Y_{g,r,t} - Y_{g,r,t-1} \right|}{Y_{g,r,t-1} + Y_{g,r,t}} = 0$$

if $Y_{g,r,t-1} = Y_{g,r,t} = 0$. Compared to the usual presentation of BBI, a multiplying factor of value 2 is introduced to make BBI comparable in scale to the indices introduced below. The justification is that, in this way, the elementary terms which are averaged take the form of a ratio between an absolute difference $\left| Y_{g,r,t} - Y_{g,r,t-1} \right|$ and a mean $\left(Y_{g,r,t-1} + Y_{g,r,t} \right)/2$.

Using BBI in trended series generates confusion between alternation and trend, as developed in proposition P1. This proposition shows that the BBI of a series on length *T* with affine growth has order of magnitude $\log T / (T-1)$. Moreover, interpreting BBI in the framework of linear filtering (Diggle, 1990; Chatfield, 2003) highlights a restrictive assumption underlying this index, related to the interpretation of BBI as the sum of the absolute values of the residuals obtained by first-order differencing normalized by the sum of the two successive values involved in the differencing (that can be interpreted as a very local trend). The underlying implicit hypothesis is that the alternation amplitudes given by the residuals are roughly proportional to the corresponding trend level (the relevance of this hypothesis for our dataset is highlighted in Fig. S2). In the case where the residuals are independent from the trend level, BBI scales the residuals as a function of the trend level. Each absolute difference will be weighted differently, and BBI will be irrelevant. This should be considered its main shortcoming, as index for alternation.

P. Proofs of propositions

P1/ BBI of a time series with affine growth

If $Y_t = at + b$ then the BBI is asymptotically equivalent to $\log t/(t-1)$. We assume that $a > 0$ and $b > 0$ to ensure that $Y_t > 0$, but the proof can be easily adapted to the other cases, replacing $Y_t + Y_{t-1}$ by $|Y_t| + |Y_{t-1}|$ in the definition of BBI.

Proof

For any *t*, $|Y_t| + |Y_{t-1}| \dfrac{|Y_t - Y_{t-1}|}{Y_t + Y_{t-1}} = \dfrac{|a|}{(at+b)+(at-a+b)}$. We have

$$\sum_{n=1}^{t} \frac{|Y_n - Y_{n-1}|}{Y_n + Y_{n-1}} = \sum_{n=1}^{t} \frac{a}{2an - a + 2b}$$

$$= \sum_{n=1}^{t} \frac{1}{2n(1 - \frac{1}{2n} + \frac{b}{an})} = \frac{1}{2}\sum_{n=1}^{t} \frac{1}{n}(1 + \frac{1}{2n} - \frac{b}{an} + o(\frac{1}{n})),$$

where $\sum_{n=1}^{t} \frac{1}{n}$ is asymptotically equivalent to $\log t$ and where $\sum_{n=1}^{t} \frac{1}{n^2} = \frac{\pi^2}{6}$. Hence, the BBI is asymptotically equivalent to $\alpha(t) = \log t/(t-1)$. The difference between the BBI and this equivalent increases with ratio *a*/*b*. The precision $|BBI(t) - \alpha(t)|/BBI(t)$ of approximation $\alpha(t)$ is given in Table T5 for different values of *a* and *t*, in the case where *b*=1. This table shows that for large values of *a*, the precision of approximation $\alpha(t)$ is quite good even for small values of *t*.

P2/ BBI and BBI_norm for stationary time series with constant amplitude alternation

Recall that BBI and BBI_norm are defined as

$$\text{BBI} = \frac{2}{T-1}\sum_{t=2}^{T}\frac{|Y_t - Y_{t-1}|}{Y_{t-1}+Y_t},$$
$$\text{BBI\_norm} = \frac{\sum_{t=2}^{T}|Y_t - Y_{t-1}|/(T-1)}{\sum_{t=1}^{T}Y_t/T},$$

where it is assumed that these indices apply to series of non-negative values.

For a linear trend with residuals corresponding to an alternation with amplitudes proportional to the trend level ($0 < c \le a$)

$Y_{2t} = (a+c)2t,$
$Y_{2t+1} = (a-c)(2t+1),$

we have

$$\text{BBI} = \frac{2}{T-1}\sum_{t=2}^{T}\frac{a+c(2t-1)}{a(2t-1)\pm c},$$
$$\text{BBI\_norm} = \frac{\sum_{t=2}^{T}\{a+c(2t-1)\}/(T-1)}{a(T+1)/2} = \frac{a+c(T+1)}{a(T+1)/2}.$$

$\text{BBI} \to 2c/a$ and $\text{BBI\_norm} \to 2c/a$ when $t \to +\infty$. While the ranges of possible values of the two indices are similar, we expected BBI_norm to be more robust to outliers. This illustrates the fact that these indices are only relevant when the alternation amplitudes are roughly proportional to the corresponding trend level, a particular case being a stationary series with constant alternation amplitudes.

For a linear trend with residuals corresponding to an alternation with constant amplitude $0 < c \le a$

$Y_{2t} = a2t + c,$
$Y_{2t+1} = a(2t+1) - c,$

we have

$$\begin{aligned}\text{BBI} &= \frac{2}{T-1}\sum_{t=2}^{T}\frac{a+2c}{a(2t-1)}\\ &\approx \frac{(a+2c)\log 2T}{a(T-1)},\\ \text{BBI\_norm} &= \frac{2(a+2c)}{a(T+1)},\\ \text{BBI\_res\_norm} &= \frac{4c}{a(T+1)}.\end{aligned}$$

BBI takes the form of a sub-series of the harmonic series. $\text{BBI} \to 0$, $\text{BBI\_norm} \to 0$ and $\text{BBI\_res\_norm} \to 0$ when $t \to +\infty$.

P3/ Indices for stationary time series with constant amplitude alternation

For a stationary series with average *a* and residuals corresponding to an alternation with constant amplitude $0 < c \le a$

$Y_{2t} = a + c,$
$Y_{2t+1} = a - c,$

we have

$$\begin{aligned}\text{BBI} &= \frac{2}{T-1}\sum_{t=2}^{T}\frac{2c}{2a} = \frac{2c}{a},\\ \text{BBI\_norm} &= \frac{2c}{a},\\ \text{BBI\_res\_norm} &= \frac{2c}{a}.\end{aligned}$$

P4/ BBI_res_norm of an alternate time series with linear growth

If $Y_{2t+1} = a(2t+1)$ and $Y_{2t} = 0$ (with $a > 0$) then the BBI_res_norm is asymptotically independent from *a* and tends towards 2.

Proof

From classical results in linear regression, the least square line has slope

$$\frac{\sum_{n=0,n.odd}^{2t+1} nY_n}{\sum_{n=0}^{2t+1} n^2} = \frac{\sum_{n'=0}^{t} a(2n'+1)^2}{\sum_{n=0}^{2t+1} n^2} = a\frac{4\sum_{n'=0}^{t} n'^2 + 4\sum_{n'=0}^{t} n'^2 + t + 1}{\sum_{n=0}^{2t+1} n^2}$$

$$= a\frac{\frac{2}{3}t(t+1)(2t+1) + 2t(t+1) + t + 1}{\frac{(2t+1)(2t+3)(4t+3)}{6}} = \frac{a}{2} + o(1)$$

Thus, the predictor is $\hat{Y}_t = \frac{a}{2}t + o(t)$ and the empirical residual is $\hat{\varepsilon}_t = \frac{a}{2}t + o(t)$ if *t* is odd and $\hat{\varepsilon}_t = -\frac{a}{2}t + o(t)$ if *t* is even. Consequently, $\left|\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1}\right| = |a|t + o(t)$ and

$$\sum_{n=1}^{t}\left|\hat{\varepsilon}_n - \hat{\varepsilon}_{n-1}\right| = |a|\frac{t(t+1)}{2} + o(t^2) = a\frac{t^2}{2} + o(t^2).$$

Since (for example if *t* is odd and *t = 2k+1*)

$$\sum_{n=1}^{t} Y_n = a\sum_{n'=0}^{k}(2n'+1) = 2a\frac{k(k+1)}{2} + o(t^2) = a\frac{t^2}{4} + o(t^2),$$

we have

$$BBI\_res\_norm(t) = \frac{a\frac{t^2}{2} + o(t^2)}{a\frac{t^2}{4} + o(t^2)} \times \frac{t}{t-1} = 2 + o(1).$$

The precision $\left|BBI\_res\_norm(t) - 2\right| / BBI\_res\_norm(t)$ of the approximation of BBI_res_norm(*t*) by its limit 2 is given in Table T6 for different values of *a* and *t*. This table shows that for any value *a*, the precision of the approximation is good for *t* >25.

M. Supplementary description of statistical models and methods

M1/ Clustering using Gaussian independent mixture models

If $z = z_g$ refers to the two-dimensional vector of indices, the Gaussian mixture model is defined by its probability density function (pdf)

$$f(z) = \sum_{k=1}^{K} \pi_k \, f_k(z; \mu_k, \Sigma_k),$$

where $f_k(z; \mu_k, \Sigma_k)$ denotes the pdf of the bivariate Gaussian distribution with mean μ_k and (diagonal) covariance matrix Σ_k. This pdf corresponds to the assumption that genotypes within cluster *k* follow the distribution $f_k(z; \mu_k, \Sigma_k)$. The clustering is obtained by estimating the model parameters π_k, μ_k and Σ_k, and by associating each genotype z_g with the most likely cluster. The number of clusters *K* was selected using BIC. We used our own implementation of mixture models, developed with the R software.

M2/ Model estimation, selection and validation in neural networks and SVMs

Neural networks and SVMs depend on two kinds of parameters:

- Parameters that can be estimated automatically from the dataset by optimizing a criterion (the likelihood function in the case of NNs, or a geometric criterion in the case of SVMs). Estimation relies on a set of genotypes, referred to as "learning sample", which classes are considered as known. In practice, the classes yielded by Gaussian mixture clustering were considered.
- A so-called *regularisation parameter*, denoted by ν, which controls the ability of the model to predict correctly either the classes of the learning set, or those of future genotypes not in the learning set (and even if possible, classes of both types of genotypes). The regularisation parameter ν has to be specified by the modeller.

In the case of classes comprising reasonably comparable numbers of genotypes, the performance of supervised classification methods can be assessed with the classification error

rate (or *error rate*, in short), which is the frequency of genotypes which class is not correctly predicted. The set of genotypes used to compute the error rate is referred to as "test sample". Usually, a low classification error rate (perfect classification) can be achieved through a particular choice of ν, if the whole dataset is used simultaneously as learning and test sample. However, this is an optimistic prediction of the actual error rate on future genotypes, since the same dataset is used both to estimate the parameters and to compute the error rate. A more reliable way to assess the possibility of classifying future genotypes accurately is the cross-validated error rate (Bishop, 2006, Chapter 1). One half of genotypes, chosen randomly, are used as learning sample and the other half as test sample. Then the roles of both sets are permuted, and this procedure is repeated several times (5 times in our case) to reduce the variability in estimating the average error rate. This variability is related to the random choice of both sets. This algorithm is applied to several values of the regularisation parameter ν, so as to minimise the predicted error rate with respect to ν.

M3/ Factorial Discriminant Analysis (FDA)

FDA is a variant of principal component analysis that aims at providing the plane in which the classes are optimally separated (Tabachnick & Fidell, 2007). This plane is obtained by maximising the separation between the centres of the classes, in regard to the dispersion of the data of each class around their mean. The plane obtained by applying the FDA to each genotype characterised by the three local indices is depicted in Fig. 6, which shows a correct discrimination between the regular genotypes (represented with red diamonds) and the biennial alternating genotypes (green triangles). Note that axis 1 seems sufficient, essentially, to separate both classes, and could be used as a scoring method. The most regular genotypes have maximal coordinates along x-axis, and most alternate genotypes have minimal coordinates along this axis. The irregular genotypes (blue squares) seem to be uniformly

distributed on the plane. Since this is the plane where the classes are optimally separated, genotypes with irregular yields at tree scale cannot be discriminated using the local indices.

Literature Cited in Supplementary Information

Chatfield C. 2003. *The Analysis of Time Series: An Introduction, 6th edition*. Chapman & Hall/CRC Press, Boca Raton.
Diggle PJ. 1990. *Time Series: A Biostatistical Introduction*. Oxford University Press, Oxford.