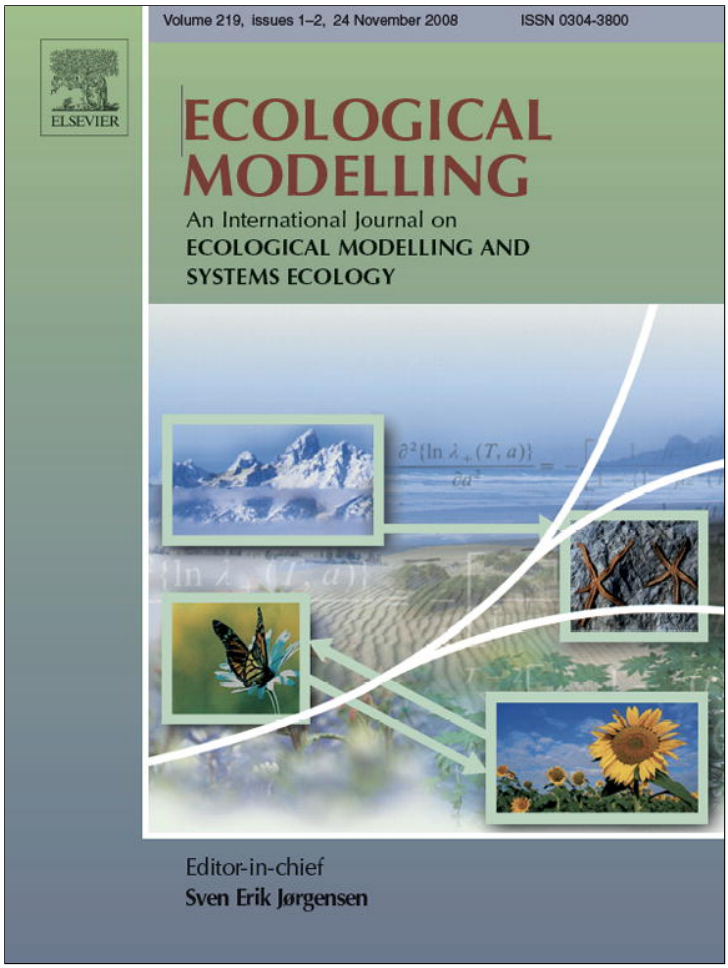
Comparison of stepwise selection and Bayesian model averaging for yield gap analysis

Lorène Prost, David Makowski, Marie-Helene Jeuffroy



ECOLOGICAL MODELLING 219 (2008) 66–76

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel

Comparison of stepwise selection and Bayesian model averaging for yield gap analysis

Lorène Prost, David Makowski, Marie-Hélène Jeuffroy*

INRA, UMR 211 INRA AgroParisTech, 78 850 Thiverval-Grignon, France

ARTICLE INFO

Article history:
Received 30 November 2007
Received in revised form
16 July 2008
Accepted 28 July 2008
Published on line 4 September 2008

Keywords:
Bootstrap
Diagnosis
Limiting factor
Model mixing
Model selection
Stepwise
Parameter estimation
Wheat

ABSTRACT

Stepwise selection is frequently used in ecology and agronomy. In the yield gap analysis approach, linear regression and stepwise selection are used to identify and rank the limiting factors of crop yield. The main value of stepwise selection is that it can be used to select a subset of explanatory variables by using statistical criteria. The number of parameters in the final model obtained by using such a procedure is expected to be less than in the complete model, and the variance of the estimated parameters can be reduced. Nonetheless, several recent studies have emphasized the limitations of stepwise selection, such as the lack of stability of the set of selected variables and bias in the parameter estimates. Model mixing methods like Bayesian model averaging (BMA) have been proposed as an alternative, but these methods have never been used for yield gap analysis. The objective of this paper was to compare stepwise selection methods and BMA for yield gap analysis. Our comparison was based on 10 000 bootstrap samples drawn from a dataset of 160 plots including 8 years of winter wheat (*Triticum aestivum* L.) experiments. Parameter estimates obtained after stepwise selection were compared to the estimated values obtained without any selection and to the estimated values obtained with BMA. The results showed that these statistical methods led to contrasted frequencies of variables selections and to different estimated parameter values. The frequencies of selection were greater with BMA than with stepwise selection. BMA also gave smaller standard deviations for parameter estimates in many cases, but this was not always the case. Compared to the stepwise selection methods, the parameter estimates obtained with BMA were closer to zero. Our results showed that the bootstrap approach can efficiently allow agronomists to compare various statistical methods for selecting explanatory variables and for estimating the effects of limiting factors.

1. Introduction

Multiple regressions with stepwise selection techniques are often used in ecology and in crop science for studying the effects of limiting factors on plant or animal characteristics such as plant biomass, species richness, or crop yield. Whittingham et al. (2006) have reviewed 508 papers published in 2004 in three leading journals of ecology (Journal of Applied Ecology, Animal Behaviour, Ecology Letters) and have shown that, out of 65 papers using a multiple regression approach, 57% used a stepwise procedure. Stepwise selection is also frequently used in agronomy, for example in the yield gap analysis approach. Yield gap analysis is used to identify and rank the factors that can explain the low yields observed in a

* *Corresponding author*. Tel.: +33 1 30815295; fax: +33 1 30815425.
E-mail address: prost@grignon.inra.fr (L. Prost).
0304-3800/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2008.07.026

range of farmers' fields. This method has been widely used in many countries (e.g. Casanova et al., 1999; Bindraban et al., 2000; Verdoodt et al., 2003; De Bie, 2004; Mussgnug et al., 2006), and also in France where it is called agronomic diagnosis (e.g. Leterme et al., 1994; Doré et al., 1997; Brancourt-Hulmel et al., 1999; Le Bail and Meynard, 2003; David et al., 2005; Doré et al., 2008). Lecomte (2005) has automated this method by using a multiple stepwise regression analysis. The same approach has been applied by Brancourt-Hulmel et al. (1999) and Barbottin et al. (2005). This method has been developed with the idea of being easily implemented by various stakeholders such as plant breeders, extension services, or local advisors.

The main value of stepwise selection is that it can be used to select a subset of explanatory variables by using statistical criterion computed from a dataset, like the Akaïke information criteria, the bayesian information criterion, or statistical tests (e.g. Miller, 2002). The number of parameters in the final model obtained with this procedure is expected to be less than in the full model, and the variance of the estimated parameters can also be reduced.

Nonetheless, several papers in medical science (Steyerberg et al., 1999) and in ecology (Burnham and Anderson, 2002; Whittingham et al., 2006) have emphasized the limitations of stepwise selection. A first problem is that the results of stepwise selection can depend on the procedure used for selecting the variables. Different selection procedures can lead to different sets of selected variables. This is an important issue for yield gap analysis, because different procedures may lead to the identification of different sets of limiting factors. A second problem is that the uncertainty of the results of the selection method is generally ignored. All inferences are usually performed using the selected model only, although the selected set of explanatory variables may be highly sensitive to the dataset used to perform the selection. A small change in the dataset may lead to a different set of selected variables. A third problem is that the estimated parameter values obtained after stepwise selection are likely to be biased due to the omission of some important factors and to the use of the same dataset for both variable selection and parameter estimation (Miller, 2002).

Several statisticians have emphasised that, in some cases, it is better to mix all models than to use the single selected model. The basic idea is to use a weighted mean of the individual model predictions instead of the prediction derived from the single 'best' model. Several model mixing methods were recently developed to estimate the weight associated with each model from a training dataset (Buckland et al., 1997; Hoeting et al., 1999; Yang, 2003; Raftery et al., 2005; Yuan and Yang, 2005). Model-mixing can improve the accuracy of model predictions and of parameter estimation, and give more realistic confidence intervals (Chatfield, 1995; Draper, 1995). According to a recent statistical study (Yuan and Yang, 2005), model-mixing is better than selection when the model errors are large. The consequences of using stepwise selection analysis for yield gap analysis have never been studied, and stepwise selection methods have never been compared to the model mixing approach in this context.

The objective of this paper is to compare stepwise selection methods and model mixing for yield gap analysis. Our comparison is based on a large number of bootstrap samples drawn from a dataset including 8 years of winter wheat (*Triticum aestivum* L.) experiments. Parameter estimates obtained after stepwise selection are compared to the estimated values obtained without any selection and to the estimated values obtained with a model mixing approach. The differences are discussed, and the practical value of using bootstrap sampling in yield gap analysis studies is emphasized.

2. Materials and methods

2.1. Data

2.1.1. Trial characteristics

We gathered data from many winter wheat trials carried out for cultivar assessment. These cultivar trials were composed of numerous new cultivars always compared to one control cultivar, Soissons, a mid-early cultivar, widely grown in France. We used Soissons data in this study.

The trials were carried out for 8 years, from 1995 to 2003, in France. Five to sixty-five sites were experimented on each year, representing a wide range of soil weather conditions (Fig. 1). Five experiments were carried out in 1995, five in 1996, four in 1997, six in 1998, four in 1999, 32 in 2003, 65 in 2004 and 39 in 2005. Non-limiting crop management strategies were applied on all trials with high yield targets (9–10 t ha^{-1}) and with full herbicide, fungicide, and insecticide controls. Management was then comparable in all the plots. The total number of plots (site × years) used in this study was 160.

2.1.2. Yield measurements

Wheat yield was measured in each plot from the average of two micro-plot measurements. A yield loss (YL) was calculated for each plot as $YL = (Y_{pot} - Y)/Y_{pot}$ where Y_{pot} is the potential yield of Soissons and Y is the measured yield. Y_{pot} is the yield that the crop would have reached without any environmental limiting factors. It was determined with the procedure described by Brancourt-Hulmel et al. (1999) and was set at 11.4 t ha^{-1} (0% moisture content).

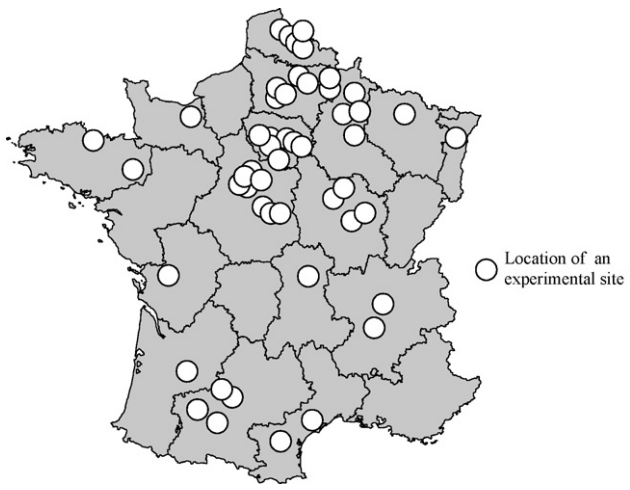


Fig. 1 – Locations of the experimental sites.

Table 1 – Characteristics of the 22 explanatory variables

Variable	Definition	Units	Mean	S.D.	Min	Max
ndfr	Number of days when the temperature is lower than the frost resistance of the genotype from sowing to 1 cm ear (Lecomte et al., 2003)	days	0.34	1.18	0	6
ndefr	Number of days of ear frost (minimal temperature $\leq -4\,^{\circ}C$) from 1 cm ear to flowering (Gate, 1995)	days	0.34	0.87	0	5
stmpw	Sum of the daily average temperatures above 0 by development periods	°C	824.8	104.6	418.6	1097.5
stmpef			759.8	50.4	566.5	889.1
sraw	Sum of the daily radiation by development periods (Monteith, 1972; Gallagher and Biscoe, 1978)	J/cm^2	69 333	13 880	30 498	118 561
sraef			106 986	12 883	73 838	139 686
srafm			75 628	7011	59 872	98 265
ratw	Ratio srad/stmp by development periods (Fischer, 1985)		85.6	22.1	38.9	184.9
raef			283.6	36.4	200.8	403.6
spetpw	Sum of the daily differences rainfall-ETP < 0 by development stages	mm	0.35	2.57	0	23
sdfef	Sum of the daily water deficits ETR-ETM when ETR < ETM, by development stages (Brancourt-Hulmel et al., 1999)	mm	16.47	29.72	0	143.94
sdffm			68.92	43.23	0	163.69
sri1045m	Sum of the daily radiation <1045 J/cm^2 from meiosis-5d to meiosis + 5d	J/cm^2	767	942	0	8992
st25ef	Sum of the daily maximal temperatures >25 °C by development stages	°C	3.8	5.0	0	26.2
st25fm			41.0	27.3	0.9	125.5
lomax	Lodging	Score	1.7	1.7	1	9
fomax	Diseases on foot	% area	0.4	3.5	0	46.75
pwl	Powdery mildew on leaves	Score	1.3	0.8	1	6
br	Brown rust on leaves	Score	1.9	2.0	1	9
sl	Septoria on leaves	Score	2.8	2.4	1	9
nni	Nitrogen nutrition index (Justes et al., 1997) at anthesis		0.98	0.10	0.32	1
npm2	Nb of plants after winter when nb < 200 plants/m^2	nb/m^2	198	9.43	121	200

Phases of development: w = from sowing to 1 cm ear, ef = from 1 cm ear to anthesis, fm = from anthesis to maturity. Bold characters: variables of the reduced set of explanatory variables.

2.1.3. *Explanatory variables*

Three kinds of variables were considered (Table 1): weather variables, diseases, and nitrogen. Fifteen weather variables were defined from five daily weather measurements taken at each plot location (minimal and maximal temperatures, rainfall, Penman potential evapo-transpiration, and global radiation) and for three winter wheat development periods (a winter period from sowing to the beginning of stem elongation, a stem elongation period from 1 cm ear to anthesis, and a grain filling period from anthesis to maturity) (Table 1). Brancourt-Hulmel et al. (1999) and Lecomte (2005) showed that these variables can affect crop yield.

Four variables were defined to describe the levels of infestation of each plot with Brown Rust (*Puccinia graminis*), Septoria (*Mycosphaerella graminicola* and *Phaesphaeria nodorum*), Powdery Mildew (*Erysiphe graminis*), and foot diseases, including foot fusarium (*Fusarium roseum*, var *culmorum*) and foot rot (*Pseudocercosporella herpotrichoides*). These variables were visual scores of infection from 1 (no symptoms) to 9 (organ completely covered by the disease in the whole field) (Godin and Soyer, 2006). Lodging was also scored in each plot from 1 (no lodging) to 9 (completely lodged) using the method described by Godin and Soyer (2006). To describe crop nitrogen status we used the nitrogen nutrition index (Justes et al., 1997), which is less than 1.0 when the nitrogen supply is below the crop's requirement and thus limits crop growth. The last variable represents the number of plants at the end of winter. It can affect grain yield when it is below 200 plants per m^2 (Lecomte personal communication). The total number of limiting factors tested was 22.

2.2. *Linear regression models*

The yield loss was related to the candidate explanatory variables using linear regression models defined by $YL = \theta_0 + \theta_1 x_1 + \cdots + \theta_p x_p + \varepsilon$ where x_1, …, x_p are the explanatory variables, θ_0, …, θ_p are the model parameters, and ε is the residual error term. In yield gap analysis, the explanatory variables correspond to limiting factors and the parameters represent the effects of a unit change of these limiting factors on crop yield. Model parameters are unknown and must be estimated from experimental data.

Two sets of explanatory variables were considered in turn: the full set of 22, and a reduced set of 5. The latter were selected by expertise based on the fact that the field experiments were carried out using intensive cropping systems i.e. with full protection against pests and diseases, and using optimal nitrogen fertilization. To define the reduced set, we assumed that the most important limiting factors were those related to weather rather than to diseases, nitrogen nutrition, or lodging. Five weather variables were thus selected, one for each type of weather variable: winter cold (*stmpw*), low radiation in winter (*sradw*), ear frost (*ndefr*), dryness during grain filling (*sdffm*) and high temperatures during grain filling (*st25fm*) (Table 1).

Four statistical methods were considered successively for selecting the explanatory variables x_1, …, x_p and estimating the model parameters θ_0, …, θ_p:

- no variable selection (estimation of all the model parameters by least squares),
- stepwise selection based on the Akaïke information criterion (AIC) and parameter estimation by least squares,
- stepwise selection based on the Bayesian information criterion (BIC) and parameter estimation by least squares,
- Bayesian model averaging (BMA).

Each method was applied using the full and reduced sets of explanatory variables in turn.

The first three methods were implemented using the glm function of the R statistical software and the parameters θ_0, …, θ_p were estimated by ordinary least squares (www.cran-r.org; Venables and Ripley, 2002).

BMA is a method for model mixing (Raftery et al., 1997). Its main principles are briefly described below. Suppose that θ is a parameter representing the effect of a limiting factor. BMA aims at computing the posterior distribution of θ expressed as

$$P(\theta|D) = \sum_{i=1}^{N} P(\theta|D, M_i) P(M_i|D) \quad (1)$$

where M_1, …, M_N is the set of available models, D is the data set, and $P(.|.)$ is a conditional density probability function. The posterior mean is defined as follows:

$$E(\theta|D) = \sum_{i=1}^{N} w_i \hat{\theta}_i \quad (2)$$

where $\hat{\theta}_i = E(\theta|D, M_i)$ (estimation of the parameter using only model M_i) and $w_i = P(M_i|D)$ (posterior probability of M_i given the data D, used as a weight associated to model M_i). Eq. (2) shows that the estimation of θ obtained using a BMA method is a weighted sum of individual estimations. The use of BMA instead of a single selected model is thus likely to change the practical conclusions derived from the simulated values of any quantities of interest.

Algorithms were developed to implement BMA with linear models (Raftery et al., 1997). In this study, the bicreg function of the BMA library of the R software was used to compute the posterior parameter means defined by Eq. (2). With bicreg, models are excluded when their posterior model probabilities are 20 times lower than the posterior probability of the best model (Raftery et al., 1997). Thus, all the explanatory variables are not necessarily selected with BMA.

2.3. *Bootstrap sampling*

Bootstrap can be used to study the uncertainty in the results of selection methods (Buckland et al., 1997; Chatfield, 1995; Miller, 2002). The principle is to generate a large number of new datasets from the initial dataset by randomly sampling data with replacement (Efron and Tibshirani, 1993). Two sizes of dataset ($N = 40, 160$) were considered successively and 10 000 bootstrap samples of size N were generated from the initial dataset. The statistical methods described in Section 2.2 were applied to each sample in turn, and parameter values were estimated from the same samples. In several datasets, all the sampled values of some explanatory variables were identical. In such situations, it was impossible to estimate the corresponding parameters, and their values were set to zero.

The results were used to compute the following terms for each explanatory variable and each statistical method:

- frequency of selection of each variable across the bootstrap samples,
- mean of the estimated parameter values across the bootstrap samples,
- standard deviation of the estimated parameter values across the bootstrap samples.

The frequency of selection of a given explanatory variable corresponds to the number of bootstrap samples where this variable was selected (i.e. where the corresponding parameter was not set equal to zero) divided by the total number of bootstrap samples (10 000). Mean and standard deviation were computed as $(1/K)\sum_{k=1}^{K}\hat{\theta}_j^k$ and $\left(1/(K-1)\sum_{k=1}^{K}(\hat{\theta}_j^k-(1/K)\sum_{k=1}^{K}\hat{\theta}_j^k)^2\right)^{1/2}$ respectively, where $\hat{\theta}_j^k$ is the estimated value of the jth model parameter obtained with the kth bootstrap sample and K is the number of bootstrap samples where the jth variable was selected. Only the non-zero estimated variables were considered for computing the means and standard deviations of the estimated parameter values. Estimated parameter values were obtained from the R function glm for the 'no selection' method and for the two stepwise methods. For the BMA method, the estimated parameter values correspond to the posterior means computed by the R function bicreg (Eq. (2)). Note that the standard deviations could have been directly derived from glm but the standard deviations provided by glm are calculated just for the selected model and thus ignore the uncertainty induced by the selection procedures. They are thus likely to underestimate the true standard deviations of the parameter estimators (Burnham and Anderson, 2002). For this reason, we decided to compute the standard deviations from the estimated parameter values obtained with the bootstrap samples.

3. Results

3.1. *Full set of explanatory variables*

3.1.1. *Frequency of selection*

The selection frequencies are shown in Tables 2 and 3 for each explanatory variable, the four statistical methods, and two sizes of dataset (40 and 160 plots). Selection frequencies of the explanatory variables are good indicators of the stability of the selection method. A selection frequency close to 0 or 1 indicates that the corresponding variable was almost never or always (respectively) selected in the 10 000 bootstrap samples. They thus reveal that the results provided by the statistical method are stable; there is not much variation across samples. On the contrary, selection frequencies close to 0.5 reveal that the results of the statistical method are not stable and are sensitive to variations in the sample of data.

Table 2 shows that the frequencies of selection obtained with samples of 40 plots ranged from 0.13 to 0.57 with the stepwise selection based on AIC, from 0.05 to 0.50 with the stepwise selection based on BIC, from 0.47 to 0.97 with the Bayesian model averaging technique and from 0.53 to 1 when no selection was performed,. With the 'no selection' method, most of the frequencies were 1, but some were lower. This is because no variation was generated for some of the 22 variables in a few bootstrap samples. This phenomenon can occur in reality because some events (e.g. diseases) are rare and cannot be observed when the sample size is small.

The selection frequencies tended to be higher with 160 plots (Table 3); from 0.95 to 1 when no selection was performed, from 0.14 to 0.89 with the stepwise selection based on AIC, from 0.02 to 0.86 with the stepwise selection based on BIC, from 0.48 to 0.98 with BMA.

Tables 2–3 show that the frequencies of selection were invariably lower with the stepwise method based on BIC than with the stepwise method based on AIC. This result is due to the penalty term used in the BIC criterion, which makes this criterion more conservative (e.g. Burnham and Anderson, 2002). It is interesting to note that the differences between the two stepwise methods were very large for several variables. For example, Table 3 shows that the frequency obtained for the variable *npm2* was 0.81 with AIC but only 0.20 with BIC. The probability of selecting *npm2* is thus much higher with AIC than with BIC.

The frequencies of selection obtained with BMA were systematically higher than the frequencies obtained with the two stepwise methods, but were lower than the frequencies obtained with the 'no selection' method (Tables 2–3). As explained above, this result is due to the fact that the total number of possible models was very large (2^{22}) and that some models were excluded when their posterior probabilities were much lower than the posterior probability of the best model.

The stepwise method based on AIC often led to intermediate values of selection frequencies. With this method, the number of frequencies falling in the range 0.3–0.7 was 7 (out of 22 values) with 40 plots (Table 2) and 13 (out of 22 values) with 160 plots (Table 3). Thus, the results of the selection were not clear-cut for many of the candidate variables with the stepwise method based on AIC. It can be concluded that, for this selection procedure, the sets of selected variables were not stable when the dataset was changed.

The number of selection frequencies in the range 0.3–0.7 was lower with stepwise regression based on BIC and with BMA. The results were thus more stable with these methods. Most of the frequencies were below 0.3 with the stepwise BIC method whereas most of the frequencies were above 0.7 with BMA. The probabilities of selecting the explanatory variables were thus high with BMA and low with the stepwise method based on BIC.

3.1.2. *Average of parameter estimates and standard deviations*

The distributions of estimated parameter values were summarized by their average values and standard deviations (Tables 2–3). Examples of distributions of estimated parameter values are given in Fig. 2.

Fig. 2 and the average values reported in Tables 2–3 show that the estimated parameter values obtained with BMA (i.e.

Table 2 – Frequencies of selection, average estimated parameter values, and standard deviation of the parameter estimates obtained using four statistical methods

Variable	Frequency of selection				Average estimated parameter value				Standard deviation			
	No selection	AIC	BIC	BMA	No selection	AIC	BIC	BMA	No selection	AIC	BIC	BMA
ndfr	0.98	0.29	0.10	0.87	2.12	1.86	0.84	1.11	7.47	7.25	7.50	4.58
ndefr	1.00	0.34	0.15	0.90	−6.09	−6.98	−7.28	−3.33	8.35	5.91	4.69	5.18
stpmw	1.00	0.24	0.11	0.92	0.04	−0.02	−0.03	0.01	0.16	0.07	0.04	0.08
stmpef	1.00	0.21	0.07	0.89	0.05	0.08	0.08	0.03	0.26	0.13	0.09	0.12
sraw	1.00	0.13	0.05	0.93	−6.28E−04	−3.53E−04	9.00E−05	−2.59E−04	1.86E−03	1.10E−03	5.90E−04	9.75E−04
sraef	1.00	0.19	0.06	0.91	−2.71E−05	1.75E−04	2.08E−04	−7.60E−06	1.70E−03	8.40E−04	4.49E−04	7.71E−04
srafm	1.00	0.27	0.10	0.89	−1.97E−04	−4.52E−04	−4.48E−04	−1.29E−04	6.77E−04	5.59E−04	5.18E−04	3.85E−04
ratw	1.00	0.36	0.25	0.97	0.60	0.34	0.26	0.29	1.49	0.52	0.18	0.78
raef	1.00	0.15	0.05	0.92	0.045	0.051	0.009	0.025	0.64	0.36	0.19	0.29
spetpw	0.53	0.20	0.05	0.47	−1.04	−1.00	−0.98	−0.39	4.45	0.88	0.89	2.09
sdfef	1.00	0.23	0.07	0.86	−0.02	−0.01	0.01	−0.01	0.16	0.17	0.17	0.09
sdffm	1.00	0.27	0.10	0.85	−2.21E−03	−3.95E−02	−5.17E−02	−7.58E−03	8.18E−02	9.16E−02	9.34E−02	4.85E−02
sri1045m	1.00	0.26	0.11	0.83	−3.90E−04	1.15E−03	2.11E−03	1.18E−04	4.64E−03	5.84E−03	6.43E−03	3.02E−03
st25ef	1.00	0.42	0.21	0.91	0.63	0.87	0.94	0.40	0.76	0.51	0.40	0.54
st25fm	1.00	0.57	0.50	0.96	0.07	0.17	0.19	0.09	0.17	0.09	0.06	0.11
lomax	1.00	0.27	0.08	0.81	−1.30	−2.73	−3.45	−0.83	3.90	3.69	3.19	2.32
fomax	0.78	0.23	0.10	0.70	−26.52	−14.29	−12.17	−10.17	230.10	53.87	19.74	88.52
pwl	0.98	0.21	0.07	0.81	249.70	−7.78	−17.20	−1.08	24 438.46	22.32	22.05	27.49
br	0.80	0.31	0.19	0.75	2.06	5.18	1.31	1.15	270.10	32.94	27.13	59.60
sl	0.99	0.29	0.10	0.88	−364.94	−1.54	−1.54	0.02	36 378.37	6.29	5.62	7.09
nni	0.93	0.32	0.15	0.80	43.97	50.15	56.84	−2.45	4622.70	116.79	74.19	879.54
npm2	0.88	0.31	0.13	0.84	−0.65	−0.88	−0.90	−0.32	2.16	1.24	1.10	1.16

Results were obtained from 10 000 bootstrap samples of 40 plots. Twenty-two candidate explanatory variables were considered. No selection, AIC = stepwise with AIC, BIC = stepwise with BIC, BMA = Bayesian model averaging.

Table 3 – Frequencies of selection, average estimated parameter values, and standard deviation of the parameter estimates obtained using four statistical methods

Variable	Frequency of selection				Average estimated parameter value				Standard deviation			
	No selection	AIC	BIC	BMA	No selection	AIC	BIC	BMA	No selection	AIC	BIC	BMA
ndfr	1.00	0.44	0.05	0.70	2.25	2.57	2.04	0.84	1.45	1.42	2.19	1.35
ndefr	1.00	0.57	0.10	0.80	−4.66	−4.80	−4.42	−2.05	2.05	1.98	1.84	2.39
stpmw	1.00	0.39	0.16	0.84	0.03	0.01	−0.02	0.01	0.03	0.05	0.02	0.03
stmpef	1.00	0.26	0.05	0.70	0.07	0.07	0.06	0.02	0.08	0.06	0.03	0.04
sraw	1.00	0.32	0.02	0.83	−5.54E−04	−5.61E−04	−3.95E−04	−1.78E−04	3.95E−04	4.51E−04	4.72E−04	3.77E−04
sraef	1.00	0.34	0.07	0.75	−2.30E−04	8.60E−05	2.22E−04	2.06E−05	5.15E−04	4.16E−04	1.74E−04	1.98E−04
srafm	1.00	0.55	0.20	0.82	−3.08E−04	−4.07E−04	−4.24E−04	−1.93E−04	2.00E−04	1.54E−04	1.15E−04	2.08E−04
ratw	1.00	0.69	0.42	0.97	0.52	0.33	0.18	0.21	0.30	0.30	0.11	0.28
raef	1.00	0.18	0.04	0.72	0.14	0.14	0.07	0.02	0.20	0.19	0.10	0.08
spetpw	0.95	0.65	0.15	0.85	−0.63	−0.77	−0.81	−0.22	0.29	0.22	0.11	0.24
sdfef	1.00	0.22	0.03	0.56	−0.01	−0.02	0.01	0.00	0.05	0.08	0.09	0.03
sdffm	1.00	0.33	0.07	0.58	−1.65E−02	−4.43E−02	−5.80E−02	−9.61E−03	2.58E−02	2.76E−02	2.18E−02	1.93E−02
sri1045m	1.00	0.22	0.06	0.48	4.40E−04	1.94E−03	3.55E−03	5.13E−04	1.30E−03	1.89E−03	1.56E−03	1.19E−03
st25ef	1.00	0.76	0.26	0.89	0.50	0.53	0.59	0.27	0.22	0.18	0.15	0.27
st25fm	1.00	0.89	0.86	0.98	0.11	0.13	0.15	0.12	0.04	0.04	0.03	0.06
lomax	1.00	0.39	0.06	0.64	−0.76	−1.29	−1.71	−0.28	0.56	0.45	0.42	0.45
fomax	1.00	0.41	0.09	0.78	−3.44	−6.12	−8.66	−1.47	8.42	8.14	4.56	3.49
pwl	1.00	0.14	0.02	0.50	−0.40	−3.37	−7.08	−0.50	1.94	3.43	3.15	1.43
br	1.00	0.28	0.08	0.58	1.82	5.65	5.24	1.66	6.70	11.59	16.31	6.99
sl	1.00	0.32	0.07	0.61	−0.62	−1.03	−1.30	−0.23	1.58	2.06	2.33	1.01
nni	1.00	0.40	0.17	0.65	9.79	25.73	34.19	9.57	18.79	17.56	10.09	13.75
npm2	1.00	0.81	0.20	0.96	−0.25	−0.25	−0.31	−0.10	0.14	0.15	0.20	0.12

Results were obtained from 10 000 bootstrap samples of 160 plots. Twenty-two candidate explanatory variables were considered. No selection, AIC = stepwise with AIC, BIC = stepwise with BIC, BMA = Bayesian model averaging.

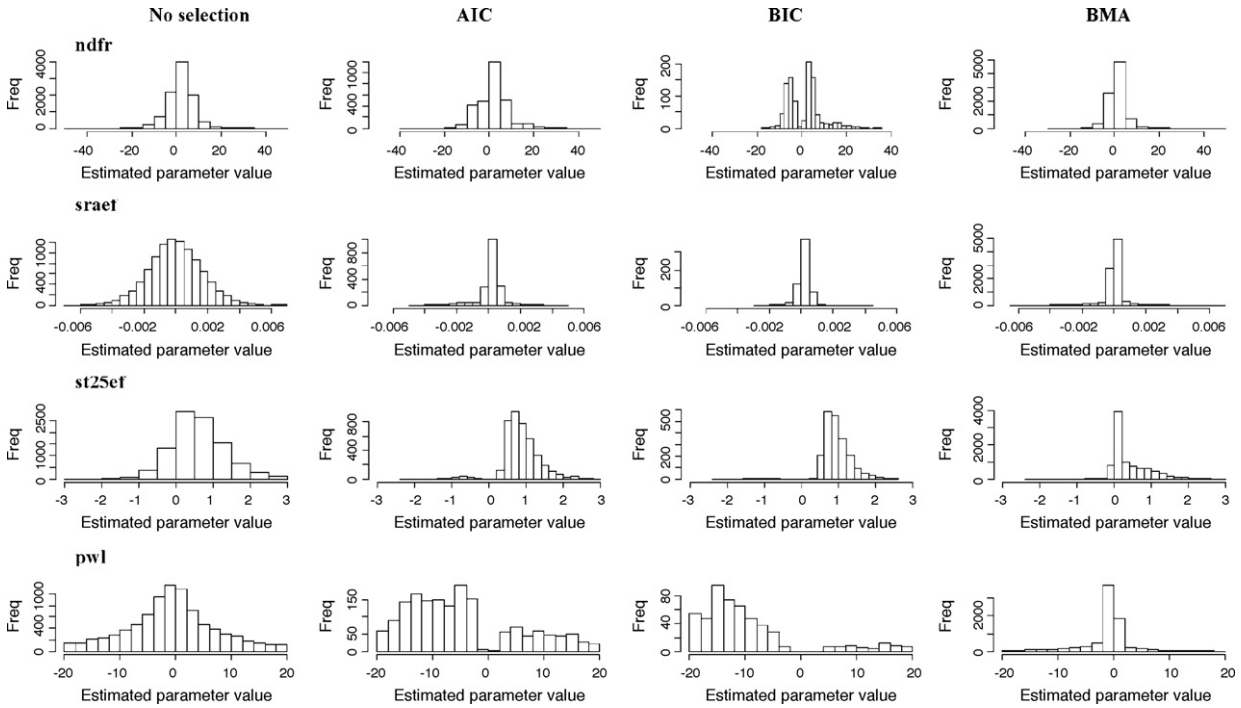


Fig. 2 – Distributions of the estimated parameter values obtained with the four statistical methods. Row = parameters. Column = methods (no selection, AIC = stepwise with AIC, BIC = stepwise with BIC, BMA = Bayesian model averaging). Results obtained from 10 000 bootstrap samples of 40 plots. 22 candidate explanatory variables were considered, only four of them are represented, as an illustration. Extreme values are not shown on this figure.

the posterior means) were closer to zero. The distributions of the estimated parameter values were more peaked around zero with BMA than with the three other statistical methods (Fig. 2). This result is confirmed by the average parameter values reported in Tables 2–3. The absolute values of the averages tended to be closer to zero with BMA than with the other methods. This was the case for 16 out of 22 parameters with 40 plots (Table 2) and for 18 out of 22 parameters with 160 plots (Table 3). For example, when the bootstrap samples include 160 plots, the average value of the parameter estimates associated with the variable *br* (score for brown rust) was 1.66 with BMA and to 5.24 with the stepwise method based on BIC (Table 3).

The standard deviations obtained with the four statistical methods were also very different (Tables 2–3). The standard deviations were larger with the 'no selection' method for most of the parameters, especially when the sample size was small (40 plots). With 40 plots, the 'no selection' method gave the highest standard deviations for 18 out of 22 parameters (Table 2). For example, the standard deviation of the parameter estimator associated with the variable *pwl* was 24 438.5 and about 97 times higher than the corresponding average estimated parameter values. The standard deviations obtained for the same parameter were much lower with the other statistical methods: 22.32, 22.05, and 27.49 with stepwise AIC, stepwise BIC, and BMA respectively (Table 2). Standard deviations were more similar when the parameters were estimated from 160 plots (Table 3). The lowest standard deviations were obtained with stepwise BIC or with BMA for 21 out of 22 parameters (Table 3).

3.2. *Reduced set of explanatory variables*

3.2.1. *Frequency of variable selection*

The selection frequencies are shown in Table 4 for each of the five explanatory variables, the four statistical methods, and two sizes of dataset (40 and 160 plots).

Table 4 shows that the frequencies of selection obtained with the 'no selection' method and with BMA were 1 or 0.99 for all parameters and both sample sizes. This shows that, for these two statistical methods, the five candidate variables were almost invariably selected in the 10 000 samples. As the total number of possible models was 2^5, all models were computed by the bicreg function of the BMA method.

The selection frequencies of the stepwise method based on AIC were in the ranges 0.20–0.79 and 0.22–1.0 with 40 and 160 plots, respectively. The frequencies obtained with BIC were always lower.

Several of the selection frequencies obtained with the two stepwise methods were close to 0.5. For example, when the sample size was fixed at 160 plots, the selection frequency of the variable *stpmw* was 0.68 with AIC and 0.33 with BIC. These intermediate frequency values indicate that the result of the stepwise selection methods is not always stable if the dataset changes.

3.2.2. *Average of parameter estimates and standard deviations*

The distributions of estimated parameter values were summarized by their average values and standard deviations (Table 4). As already noted with the full set of explanatory variables,

Table 4 – Frequencies of selection, average estimated parameter values, and standard deviation of the parameter estimates obtained using four statistical methods

Variable	Frequency of selection				Average estimated parameter value				Standard deviation			
	No selection	AIC	BIC	BMA	No selection	AIC	BIC	BMA	No selection	AIC	BIC	BMA
Sample size = 40 plots												
ndefr	0.99	0.24	0.12	0.99	−1.39	−4.16	−6.00	−0.75	3.17	4.53	4.70	2.17
stpmw	1	0.32	0.16	1	−0.014	−0.029	−0.034	−0.007	1.99E−02	2.06E−02	1.94E−02	1.22E−02
sraw	1	0.25	0.13	1	7.9E−05	2.1E−04	3.0E−04	5.1E−05	1.61E−04	2.19E−04	1.99E−04	1.04E−04
sdffm	1	0.20	0.08	1	−0.011	−0.031	−0.040	−0.004	0.046	0.077	0.090	0.024
st25fm	1	0.79	0.67	1	0.14	0.18	0.19	0.12	0.073	0.049	0.043	0.080
Sample size = 160 plots												
ndefr	1	0.22	0.06	1	−0.75	−2.00	−2.98	−0.21	1.16	1.44	1.87	0.70
stpmw	1	0.68	0.33	1	−0.016	−0.020	−0.023	−0.008	8.32E−03	5.60E−03	4.23E−03	9.01E−03
sraw	1	0.33	0.08	1	8.0E−05	1.4E−04	1.9E−04	2.4E−05	7.30E−05	6.12E−05	7.03E−05	4.86E−05
sdffm	1	0.23	0.05	1	−0.015	−0.035	−0.050	−0.004	0.021	0.025	0.025	0.0108
st25fm	1	1.00	0.99	1	0.14	0.15	0.16	0.16	0.032	0.032	0.028	0.0338

Results were obtained from 10 000 bootstrap samples of either 40 or 160 plots. Five candidate explanatory variables were considered. No selection, AIC = stepwise with AIC, BIC = stepwise with BIC, BMA = Bayesian model averaging.

the estimated parameter values obtained with BMA (i.e. the posterior means) were closer to zero. The absolute values of the average parameter estimates tended to be closer to zero with BMA than with the other methods. This was the case for all the five parameters with 40 plots and for four parameters with 160 plots (Table 4). For example, when the bootstrap samples include 160 plots, the average value of the parameter estimates associated with the variable *stpmw* was −0.008 with BMA, −0.023 with the stepwise method based on BIC, −0.020 with the stepwise method based on AIC, and −0.016 with the 'no selection' method (Table 4). Thus, in most cases, the estimated effects of explanatory variables on yield losses were smaller with BMA than with the three other statistical methods.

The smallest standard deviations were obtained with BMA in most situations. This was the case for four out of five parameters with a sample size of 40, and for three out of five parameters when the sample size was 160 (Table 4). For example, when the sample size was fixed at 160, the standard deviation of the parameter estimator associated with the variable *sdffm* was 0.011 with BMA, 0.021 with the 'no selection' method, and 0.025 with the two stepwise methods. For both sample sizes, the standard deviations obtained with the reduced set of explanatory variables were almost all smaller than the values obtained with the full set of explanatory variables. For example, the standard deviation associated with *stpmw* was 0.16 with the method 'no selection', $N = 40$, and the full set of explanatory variables (Table 2), but was 0.02 with the same method and the reduced set of variables (Table 4). The only exceptions were the standard deviations of the parameters associated to *ndefr* and *sdffm* with the stepwise BIC method.

3.2.3. *Discussion and conclusion*

Our results illustrate the practical value of the bootstrap resampling technique to assess selection procedures used in yield gap analysis. This technique allowed us to assess the stability of the selected sets of explanatory variables to variations in the dataset and to the statistical methods used to perform the selection. Although the bootstrap technique was developed in the 1980s, it is not frequently used in crop science. As computer power now permits intensive calculations, we recommend agronomists to implement the bootstrap approach to assess the results of their yield gap analysis studies. Bootstrap methods can be used to complement an independent evaluation of the models, to make a preliminary assessment of the stability of the model (Guisan and Zimmermann, 2000). The bootstrap technique can efficiently allow insight into model uncertainty and, consequently, it allows agronomists to compare various statistical methods for selecting explanatory variables and for estimating the effects of limiting factors. Note that, for datasets with significant year effect, bootstrap must be adapted and all the data collected a given year must be drawn simultaneously. This sampling method was not implemented in this study because the year effect was not significant at 5%.

Our results show that the frequency of selection of the explanatory variables and the estimated parameter values were dependent on the selection method, on the number of candidate explanatory variables, and on the size of the

datasets. Our case study showed that selection frequencies obtained with a stepwise method based on the AIC criterion often took intermediate values, in the range 0.3–0.7. This indicates that the results obtained with this selection procedure were not stable and that the set of selected variables was highly dependent on the sample used for the analysis. The selection frequencies obtained with the three other statistical methods were more extreme, either closer to zero or closer to one. The results obtained with these methods were thus more stable. These methods, however, behaved differently. The frequencies of selection obtained with the stepwise method based on BIC were below 0.3 for most of the explanatory variables. This selection method based on BIC is thus conservative as it does not easily select the candidate explanatory variables. This can be problematic when one needs to characterize most limiting factors of the yield, as it is the case in a yield gap analysis study. On the other hand, the frequencies of selection obtained with the 'no selection' method and with BMA were above 0.7 for most of the explanatory variables, especially when the sample size was set to a high value and when the number of candidate variables was small. This shows the relevance of BMA method to conduct a yield gap analysis.

The standard deviations of the parameter estimates were very large with the 'no selection' method when the number of candidate explanatory variables was 22 and the sample size was 40. This is logical because the number of parameters was very large compared to the number of available data in this case, and the estimation of 22 parameters led to inaccurate results. The standard deviation values were more even when the sample size was higher and/or the number of candidate variables was set to five. The smallest standard deviations were obtained with BMA for many parameters which confirms the interest of this method, but this was not always so. Small standard deviations were also obtained with the stepwise selection method based on BIC.

Our case study also shows that the parameter estimates obtained with BMA were closer to zero compared to the values obtained with the other methods. This is because each parameter estimate corresponds to a posterior mean computed from a large number of models and the parameter is set at zero in several of these models. Another explanation is that, according to Steyerberg et al. (1999), Miller (2002), and Burnham and Anderson (2002), the parameter estimates obtained with stepwise selection methods tend to be biased away from zero when the same dataset is used for both selection and estimation and, so, tend to be too extreme. The parameter estimates obtained with BMA may be too small compared to the true parameter values. This aspect needs further investigations.

The sensitivity of the results of stepwise methods illustrated in this paper show that these selection techniques can lead to inaccurate diagnosis of the main limiting factors, especially when the number of explanatory variables is large compared with the size of the dataset used to estimate their effects. Nevertheless, stepwise methods can be useful when the ratio of the number of observations to the number of explanatory variables is high, since these methods lead to a reduction in the number of parameters and in the standard deviations of the estimators. Our results show that BMA represents a useful alternative. The stability of a set of selected variables was higher with BMA than with stepwise methods. BMA could allow agronomists to analyse the effects of many limiting factors and obtain estimators with small standard deviations.

When the sample size was low ($N = 40$), the estimation of the full set of parameters was problematic with all statistical methods; the standard deviations of the parameter estimators were high compared to the estimated values. The value of reducing the set of explanatory variables by expertise was studied in our paper. The reduced set of variables included five variables related to weather factors which were supposed to have a predominant effect because high fertilizer rates and full pesticide treatments were applied in our experimental plots. The use of a reduced set of explanatory variables decreased the standard deviations of the parameter estimators. It is, however, important to note that the use of a reduced set of explanatory variables can induce an omission bias when important variables are omitted (Miller, 2002). As omission bias cannot be easily detected, expertise must be used with care for reducing the complexity of models.

Acknowledgements

We are grateful for the financial support from the FSOV (Fonds de soutien à l'obtention végétale- French fund for plant breeding) and the ANR (French Research Agency) under the program JCJC. The field trials used in this study were carried out by the plant breeders of the "Club des 5" group, by GEVES (Groupement d'Etudes des Variétés et Semences - Group of cultivars and seeds), by ARVALIS Institut du Végétal and by INRA (French National Institute for Agricultural Research). We also thank C. Cadet for technical assistance and A. Scaife for correcting the English language.

REFERENCES

Barbottin, A., Lecomte, C., Bouchard, C., Jeuffroy, M.H., 2005. Nitrogen remobilization during grain filling in wheat: genotypic and environmental effects. Crop Sci. 45, 1141–1150.

Bindraban, P.S., Stoorvogel, J.J., Jansen, D.M., Vlaming, J., Groot, J.J.R., 2000. Land quality indicators for sustainable land management: proposed method for yield gap and soil nutrient balance. Agric. Ecosyst. Environ. 81 (2), 103–112.

Brancourt-Hulmel, M., Lecomte, C., Meynard, J.M., 1999. A diagnosis of yield-limiting factors on probe genotypes for characterizing environments in winter wheat trials. Crop Sci. 39, 1798–1808.

Buckland, S.T., Burnham, K.P., Augustin, N.H., 1997. Model selection: an integral part of inference. Biometrics 53, 603–618.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference. Springer, New York.

Casanova, D., Goudriaan, J., Bourma, J., Epema, G.F., 1999. Yield gap analysis in relation to soil properties in direct-seeded flooded rice. Geoderma 91, 191–216.

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. J. R. Statist. Soc. A 158, 419–466.

David, C., Jeuffroy, M.H., Henning, J., Meynard, J.M., 2005. Yield variation in organic winter wheat: a diagnostic study in the Southeast of France. Agron. Sustain. Dev. 25, 213–223.

De Bie, C.A.J.M., 2004. The yield gap of mango in Phrao, Thailand, as investigated through comparative performance evaluation. Sci. Hortic. 102, 37–52.
Doré, T., Sebillotte, M., Meynard, J.M., 1997. A diagnostic method for assessing regional variations in crop yield. Agr. Syst. 54, 169–188.
Doré, T., Clermont-Dauphin, C., Crozat, Y., David, C., Jeuffroy, M.H., Loyce, C., Makowski, D., Malézieux, E., Meynard, J.M., Valantin-Morison, M., 2008. Methodological progress in on-farm regional agronomic diagnosis. A review. Agron. Sustain. Dev. 28 (1), 151–161.
Draper, D., 1995. Assessment and propagation of model uncertainty. J. R. Statist. Soc. A. 57, 45–97.
Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.
Fischer, R.A., 1985. Number of kernels in wheat crops and the influence of solar radiation and temperature. J. Agr. Sci. 105 (2), 447–461.
Gallagher, J.N., Biscoe, P.V., 1978. Radiation absorption, growth and yield of cereals. J. Agr. Sci. 91 (1), 47–60.
Gate, P., 1995. Ecophysiologie du blé. De la plante à la culture. Lavoisier. Tec and Doc, Paris, 430 pp.
Godin, C., Soyer, J., 2006. Protocole d'expérimentation: céréales à paille. Essais de valeur agronomique et technologique. GEVES, Guyancourt, France.
Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 14 (4), 382–417.
Justes, E., Jeuffroy, M.H., Mary, B., 1997. Wheat, barley and durum wheat. In: Lemaire, G. (Ed.), Diagnosis of the Nitrogen Status in Crops, vol. 4. Springer, New York, pp. 73–92.
Le Bail, M., Meynard, J.M., 2003. Yield and protein concentration in spring malting barley: the effects of cropping systems in the Paris Basin (France). Agronomie 23, 13–27.
Lecomte, C., 2005. L'évaluation expérimentale des innovations variétales. Proposition d'outils d'analyse de l'interaction génotype-milieu adaptés à la diversité des besoins et des contraintes des acteurs de la filière semences. Doctoral dissertation. AgroParisTech, Paris, France.
Lecomte, C., Giraud, A., Aubert, V., 2003. Testing a predicting model for frost resistance of winter wheat in natural conditions. Agronomie 23, 51–66.
Leterme, P., Manichon, H., Roger-Estrade, J.R., 1994. Analyse intégrée des rendements de blé tendre et de leurs causes de variation dans un réseau de parcelles d'agriculteurs du Thymerais. Agronomie 14, 341–361.
Miller, A., 2002. Subset Selection in Regression, 2nd ed. Chapman & Hall/CRC, New York.
Monteith, J.L., 1972. Solar radiation and productivity in tropical ecosystems. J. Appl. Ecol. 9 (3), 747–766.
Mussgnug, F., Becker, M., Son, T.T., Buresh, R.J., Vlek, P.L.G., 2006. Yield gaps and nutrient balances in intensive, rice-based cropping systems on degraded soils in the Red River Delta of Vietnam. Field Crop. Res. 98, 127–140.
Raftery, A.E., Madigan, D., Hoeting, J.A., 1997. Bayesian model averaging for linear regression models. J. Am. Stat. Assoc. 92, 179–191.
Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133, 1155–1174.
Steyerberg, E.W., Eijkemans, M.J.C., Habbema, J.D.F., 1999. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J. Clin. Epidemiol. 52, 935–942.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Springer, New York.
Verdoodt, A., Van Ranst, E., Van Averbeke, W., 2003. Modelling crop production potentials for yield gap analysis under semiarid conditions in Guquka, South Africa. Soil Use Manage. 19 (4), 372–380.
Whittingham, M.J., Stephens, P.A., Bradbury, R.B., Freckleton, R.P., 2006. Why do we still use stepwise modelling in ecology and behaviour? J. Anim. Ecol. 75, 1182–1189.
Yang, Y., 2003. Regression with multiple candidate models: selecting or mixing? Stat. Sin. 13, 783–809.
Yuan, Z., Yang, Y., 2005. Combining linear regression models: when and how? J. Am. Stat. Assoc. 100, 1202–1214.