

A new method to analyse relationships between yield components with boundary lines

David Makowski, Thierry Doré, Hervé Monod

▶ **To cite this version:**

David Makowski, Thierry Doré, Hervé Monod. A new method to analyse relationships between yield components with boundary lines. Agronomy for Sustainable Development, 2007, 27 (2), pp.119-128. 10.1051/agro:2006029 . hal-00886363

HAL Id: hal-00886363
https://hal.science/hal-00886363

Submitted on 11 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Agron. Sustain. Dev. 27 (2007) 119–128
© INRA, EDP Sciences, 2007
DOI: 10.1051/agro:2006029

Original article

A new method to analyse relationships between yield components with boundary lines

David MAKOWSKI[a,b]*, Thierry DORÉ[c], Hervé MONOD[b]

[a] INRA, UMR Agronomie INRA-INA P-G, BP 01, 78850 Thiverval-Grignon, France
[b] Unité Mathématiques et Informatique Appliquées INRA, domaine de Vilvert, 78352 Jouy-en-Josas Cedex, France
[c] INA-P-G, UMR Agronomie INRA-INA P-G, BP 01, 78850 Thiverval-Grignon, France

(Accepted 2 October 2006)

Abstract – Crop yield can be decreased by many limiting factors such as water stress, nitrogen stress and disease. The agronomic diagnosis method was developed by agronomists to understand the origin of crop yield variability, identify important limiting factors, and define new cropping systems. The rigorous implementation of this method requires the determination of boundary lines giving the maximum value of a yield component in relation to the value of another yield component; for example, grain weight versus grain number per square metre. Such boundary lines are used by agronomists to adjust cropping practices to environmental characteristics and, thus, to reduce the risk of pollution due to agricultural activities. We describe here a new method based on quantile regression to estimate boundary line parameters from experimental data. First, quantile values were computed from models describing the effect of limiting factors on yield components and accounting for measurement errors. Then, boundary line parameters were estimated by quantile regression, with observations weighted according to the quantile values. This approach was applied to two case studies. The quality of the parameter estimator derived by quantile regression was analysed in relation to the size of the dataset and the practical consequences of a misspecification of the quantile value was studied. Our findings show that quantile regression gives more accurate parameter estimators than the methods currently used by agronomists. Nonetheless, the bias and variances of these estimators highly depend on the chosen quantile value. The use of quantile regression should thus help agronomists to analyse crop yield variability from yield component measurements.

boundary line / model / parameter estimation / quantile regression / yield components / yield gap analysis

1. INTRODUCTION

Several methods have been defined by agronomists for developing sustainable cropping systems. One of these methods is called 'agronomic diagnosis' (Doré et al., 1997). Its objective is to discover which cropping techniques and environmental conditions are responsible for yield variations in a given area. The results can then be used to adjust cropping techniques to economic and environmental targets. An original feature of this method is that it does not directly relate yield to limiting factors. The principle is to measure several yield components (e.g. grain weight, grain number per m^2, stem number per m^2) in a large number of farmers' fields located in the area of interest and to conduct a yield gap analysis taking into account these measurements. For the most relevant pairs of consecutive yield components X and Y, measurements are used to define the functions $Y_{MAX} = f(X;\theta)$ giving the optimal value Y can take for the observed value of X, with θ a set of parameters (Doré et al., 1997; Wey et al., 1998; Brancourt-Hulmel et al., 1999). For example, $f(X;\theta)$ may represent a function relating maximum crop grain number m^{-2} to stem number m^{-2} or a function relating maximum grain weight to grain number m^{-2} (Doré et al., 1998; Brancourt-Hulmel et al., 1999). These functions are then used to identify fields where yield components did not reach their optimal values and, consequently, to determine the most important limiting factors in the area of interest as well as the periods when their effects took place.

The functions $Y_{MAX} = f(X;\theta)$ are called boundary lines (Webb, 1972; Fleury, 1991). They play a key role when performing agronomic diagnosis as explained above, but also when studying the effect of environmental variables on crop characteristics (Casanova et al., 1999) or when developing crop models to predict yield or grain quality (e.g. Gonzalez Montaner et al., 1997; David et al., 2005). A boundary line $f(X;\theta)$ can be defined as follows. When Y is measured without error, all the measurements of Y obtained for a given value of X are lower than or equal to $f(X;\theta)$ and the difference between Y and $f(X;\theta)$ is due to one or several limiting factors such as low water and nutrient soil content, pest attack, frost incidence or weed competition. When Y is measured with error, some measurements may be higher than $f(X;\theta)$ but only due to measurement errors. Thus, when the measurement errors are small, a value of Y lower than $f(X;\theta)$ indicates that the crop was probably affected by one or several limiting factors (Brancourt-Hulmel et al., 1999).

The development of a boundary line $f(X;\theta)$ involves two steps. The first step is the definition of a mathematical function

* Corresponding author: david.makowski@jouy.inra.fr
and makowski@grignon.inra.fr

Article published by EDP Sciences and available at http://www.edpsciences.org/agro or http://dx.doi.org/10.1051/agro:2006029

$f(X;\theta)$ that expresses Y as a function of X and of a set of unknown parameters θ. The second step is the estimation of the value of θ from a set of N measurements of Y and X obtained in N different fields. The first step is based either on an experimental dataset or on some physiological traits of the crop. For example, Fleury (1991) defined several mathematical functions relating several maize yield components from knowledge about organogenesis, radiation use and partitioning of assimilates.

In this paper, we consider the second step. Parameter estimation of boundary lines is not straightforward because the data used for parameter estimation are usually collected in experimental or farmers' fields where one or several limiting factors may affect yield components and where, typically, at least some of these limiting factors are unknown. Given a dataset, it is thus difficult to know which fields were affected by limiting factors and which fields were not.

Various estimation methods have been applied to boundary lines in the past. An early method consists of estimating the value of some parameters by eye or by using only the most extreme measured values of Y (Webb, 1972). Consider, for example, a plateau-plus-linear function relating grain weight to grain number m^{-2} (Gonzalez Montaner et al., 1997). One parameter of this function represents the plateau, i.e. the maximal grain weight value obtained when the grain number is lower than a threshold. A naïve estimator of this parameter is the maximum measured grain weight value. This estimator was combined with a resampling procedure by Lecomte (2005). But this approach does not account for measurement error and so it is likely to overestimate the true parameter value. In addition, its variance may be very high due to sampling variability.

Another method consists of selecting a subset of data and fitting a boundary line $f(X;\theta)$ on this subset (Webb, 1972; Casanova et al., 1999; Johnson et al., 2003). The principle is to divide the domain of variation of X into Q intervals and, for each X interval, to calculate the value of Y corresponding to a high quantile value (for example, the 90th percentile). The resulting dataset is then used for estimating the boundary line parameters by least squares. The drawbacks of this method are, firstly, that the parameters are not estimated from the original dataset and, secondly, that the method is based on an arbitrary number of X intervals and an arbitrary quantile value. Thus, there is a need for methods supported by a more explicit modelling of the whole dataset.

In this paper, we study the practical interest of estimating boundary line parameters by quantile regression. The principle is to define $f(X;\theta)$ as the function of X satisfying $P[Y < f(X;\theta)] = \tau$ for all values of X. With this definition, $f(X;\theta)$ represents the τth quantile of the response variable Y so that a proportion τ of the measurements of Y are below and a proportion $1-\tau$ of the measurements are above the boundary line. Parameters θ can then be estimated by using quantile regression techniques (Koenker and Basset, 1978; Koenker and d'Orey, 1987; Koenker and Park, 1996; Koenker and Machado, 1999). These techniques make use of the whole dataset, with observations weighted according to the chosen quantile value. This approach was already applied by Cade et al. (1999, 2005) to a variety of ecological phenomena but, as far as we know, it has never been applied to estimating the parameters of boundary lines. An important preliminary step is to determine the quantile value to be used for quantile regression. This is a difficult problem because the exact value of τ depends on the probability distribution of the limiting factor effect and on the probability distribution of the measurement errors. No method has been defined for determining this value.

In the next section, we describe a model-based approach for estimating parameters of boundary lines by quantile regression. This approach makes it possible to compute relevant quantile values before parameter estimation. The interests and limitations of our approach are discussed in two case studies in Section 3. The first case study presents an application to a real dataset including measurements collected in 71 pea fields (*Pisum sativum*). The purpose of the second case study is to evaluate the quality of the parameter estimators derived by quantile regression using simulated values of wheat (*Triticum aestivum* L.) yield components.

2. A METHOD FOR ESTIMATING BOUNDARY LINE PARAMETERS

The method includes three steps. In the first step, one or several quantile values are calculated by using a model with random effects. In the second step, estimation of the boundary line parameters is performed by quantile regression for each quantile value specified in step 1. Finally, the last step consists of a numerical assessment of the quality of the parameter estimates.

2.1. Computation of quantile value

We present two models relating the response variable Y to a random variable, Z, representing the overall effect of unmeasured limiting factors and to a measurement error term noted ε. The limiting factor effect is additive in the first model and multiplicative in the second model. We show how these models can be used to compute relevant quantile values.

2.1.1. Model 1

Consider the following model:

$$Y = f(X;\theta) - Z + \varepsilon \tag{1}$$

where Y is the measured value of the response variable of interest, Z is a random variable representing the effect of one or several unmeasured limiting factors ($Z \geq 0$), and ε is a random variable representing a measurement error.

For any given value of X, $f(X;\theta)$ is equal to the τth quantile of Y if

$$\tau = P[Y < f(X;\theta)]. \tag{2}$$

From equation (1), we see that equation (2) is equivalent to

$$eq3\tau = P(\varepsilon < Z). \tag{3}$$

If the probability distributions of Z and ε are independent of X, equation (3) shows that $f(X;\theta)$ corresponds to the same τth quantile for all values of X. Equation (3) also shows that τ is equal to the probability that the error of measurement is lower than the effect of the limiting factors. The quantile τ is thus equal to 0.5 if Y is not affected by any limiting factors ($Z = 0$) and if the measurement errors have a symmetric distribution with zero mean, whereas it is equal to $P(Z > 0)$ if Y is measured without error ($\varepsilon = 0$). Otherwise, the value of τ can be calculated from pre-defined probability densities of Z and ε using, for example, the following equation:

$$\tau = \int_{D_Z} h(z) P(\varepsilon < z)\, dz \tag{4}$$

where h is the density of Z and $P(\varepsilon < z)$ is the probability that ε is lower than z. Alternatively, τ can be approximated by generating a high number of values of Z and ε and by calculating the proportion of values of Z higher than ε.

2.1.2. Model 2

We now consider the following model:

$$Y = Z \times f(X;\theta) + \varepsilon. \tag{5}$$

Here, Z is assumed to be lower than or equal to 1, and higher than or equal to zero.

For any given value of X, $f(X;\theta)$ is equal to the τth quantile of Y if

$$\tau = P\left[Z \times f(X;\theta) + \varepsilon < f(X;\theta)\right]. \tag{6}$$

The value of τ is independent of X in some cases; for example, if $\varepsilon = e \times Z \times f(X;\theta)$ and if e and Z are independent of $f(X;\theta)$, where e denotes, typically, a centred random variable independent of Z. In this case,

$$\tau = P\left[Z(1+e) \times f(X;\theta) < f(X;\theta)\right]$$

and so

$$\tau = P\left[Z(1+e) < 1\right]. \tag{7}$$

If Y is measured without error, $e = 0$ and $\tau = P[Z < 1]$. Otherwise, the quantile can be calculated from equation (7) using samples of Z and e randomly drawn from some pre-defined probability distributions of Z and e.

2.1.3. Probability distributions for Z and ε

The computation of τ requires the specification of probability distributions for ε and for Z. When replicates of Y are available, it is possible to assume that $\varepsilon \sim N\left(0, \sigma^2\right)$ or $e \sim N\left(0, \sigma^2\right)$, and to estimate σ^2 from the replicates. When replicates are not available, σ^2 must be specified from external sources.

With model 2, as Z is in the range (0–1), a good option is to assume that Z follows a beta distribution, $Z \sim \text{Beta}(\alpha, \beta)$. The beta is a flexible probability distribution defined on the unit interval. This distribution is symmetric when $\alpha = \beta$, it is flat when $\alpha = \beta = 1$, and it is dissymmetric when $\alpha \neq \beta$. Examples of beta distributions are shown in Figure 1.

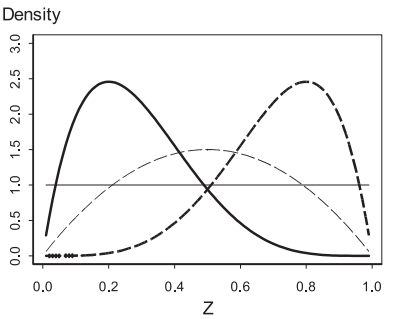


Figure 1. Probability distributions of Z when $Z \sim$ Beta(1, 1) (thin continuous line), $Z \sim$ Beta(2, 2) (thin dashed line), $Z \sim$ Beta(2, 5) (bold continuous line) and $Z \sim$ Beta(5, 2) (bold dashed line).

For model 1, Z can be expressed as $Z = Z_{MIN} + (Z_{MAX} - Z_{MIN}) \times B$ where Z_{MIN} and Z_{MAX} are the minimum and maximum values of Z, respectively, and where B is a random variable in the range (0–1). It is then possible to assume that B follows a beta distribution, $B \sim \text{Beta}(\alpha, \beta)$. In this case, Z is in the range (Z_{MIN}, Z_{MAX}) and has a symmetric or dissymmetric distribution depending on α and β.

In practice, the values of α, β, Z_{MIN}, and Z_{MAX} are unknown, but we assume that the agronomist or the statistician can choose or estimate reasonable values from data or expert knowledge. Because such a choice can only be approximate, it is recommended to study the sensitivity of the quantile to the parameters of the probability distribution of Z. To do that, it is necessary to define different values for α, β, Z_{MIN}, and Z_{MAX} and to compute the corresponding quantiles using model 1 or 2.

2.2. Parameter estimation

The first step leads to one or several estimated quantile values denoted by $\hat{\tau}$. The next step is the estimation of θ from data by quantile regression. This method is nonparametric and consists of minimising a sum of weighted absolute differences between observations and predictions. For a given estimated quantile $\hat{\tau}$, an estimator of θ is a vector $\hat{\theta}$ minimising (Koenker and Basset, 1978)

$$L(\theta) = \sum_{i=1}^{N} \rho_{\hat{\tau}}\left[Y_i - f(X_i;\theta)\right] \tag{8}$$

where Y_i and X_i are the yield components measured in the ith field, $i = 1, \ldots, N$, and $\rho_{\hat{\tau}}[.]$ is a function defined by

$$\rho_{\hat{\tau}}\left[Y_i - f(X_i;\theta)\right] = \hat{\tau} \times \left[Y_i - f(X_i;\theta)\right] \times 1_{\{Y_i \geq f(Xi;\theta))\}} - (1-\hat{\tau}) \times \left[Y_i - f(X_i;\theta)\right] \times 1_{\{Y_i < f(Xi;\theta))\}} \tag{9}$$

and $1_{\{v\}} = 1$ if the condition v is true and zero otherwise.

For the practical problem considered in this paper, the function $f(X;\theta)$ is usually a nonlinear function of the parameters. When the function is nonlinear, equation (8) can be minimised by using an interior point algorithm (Koenker and Park, 1996) or the MM algorithm developed by Hunter and Lange (2000). The former is implemented in the function nlrq of the package quantreg that can be freely downloaded from CRAN: http://cran.r-project.org.

2.3. Numerical assessment of the parameter estimators

2.3.1. Sensitivity analysis

The exact value of τ is not perfectly known due to uncertainty in the distributions of Z and ε. It is therefore useful to compute a series of quantiles $\hat{\tau}_1, \ldots, \hat{\tau}_m$, to determine the corresponding parameter estimates $\hat{\theta}_{\tau_1}, \ldots, \hat{\theta}_{\tau_m}$, and to analyse the sensitivity of the parameter estimates to the quantile value. Parameter estimates can be displayed in a table and/or can be plotted. Simple sensitivity indices can also be computed.

It is also useful to display the fit of the function graphically. To do that, it is necessary to compute $f(X;\hat{\theta})$ for a series of values of X using each parameter estimate $\hat{\theta}_{\tau_1}, \ldots, \hat{\theta}_{\tau_m}$ in turn. A graphical presentation of the fitted function can be used to study the sensitivity of the boundary line to the parameter estimates computed with different quantile values. See case study 1 for an illustration of all these simple methods.

2.3.2. Confidence interval

It is useful to compute the confidence intervals of the parameter estimators to see whether the parameters were accurately estimated or not. As suggested by Cade et al. (1999), confidence intervals can be used to find the most extreme quantile that could be estimated with a reasonable precision. Various methods were defined for estimating confidence intervals of parameter estimators in linear quantile regression models (e.g. Koenker and Basset, 1978; Cade and Richards, 2006). When the model is nonlinear, confidence intervals of parameter estimators can be estimated by nonparametric bootstrap (Efron and Tibshirani, 1986). This method is applied in case study 1.

2.3.3. Root mean squared error

A classical assessment method is to define a true model associated with true parameter values and to use this model for generating a series of datasets and for computing the standard errors, bias and root mean squared errors of parameter estimators derived by quantile regression. The first step is to define a model for generating data. This model could be the model 1 or the model 2 defined above associated with some probability distributions for Z and ε, and a particular value θ that will be considered as the true parameter value. Once the model is defined, the quantile $\tau = P[Y < f(X;\theta)]$ is computed and the model is used to randomly generate K datasets including N values of X and Y. The value of θ is then estimated by quantile regression for each dataset leading to K parameter estimates noted $\hat{\theta}_\tau^k$, $k = 1, \ldots, K$. Finally, the expected value, bias, standard deviation and root mean squared error (RMSE) of the estimator are computed for each quantile as follows:

$$\text{expectation}_\tau = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_\tau^k \tag{10}$$

$$\text{bias}_\tau = \theta - \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_\tau^k \tag{11}$$

$$\text{sd}_\tau = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{\theta}_\tau^k - \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_\tau^k\right)^2} \tag{12}$$

$$\text{rmse}_\tau = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\hat{\theta}_\tau^k - \theta\right)^2}. \tag{13}$$

Note that these criteria are not independent; it is known that $\text{rmse}_\tau \approx \sqrt{\text{bias}_\tau^2 + \text{sd}_\tau^2}$. Bias, standard deviation and RMSE can be computed for different numbers of measurements N and thus provide information on the consequences of increasing or decreasing the size of the dataset. For example, one can generate K datasets with $N = 50$ measurements and K other datasets with $N = 100$ measurements. The parameters are then estimated for each series of datasets successively. A similar approach can also be used to study a consequence of a misspecification of the quantile value used for parameter estimation. An application is presented in case study 2.

3. APPLICATIONS

3.1. Case study 1: grain number vs. stem number

3.1.1. Data

The dataset used in this case study was described in detail by Doré et al. (1998). Data were collected from 71 pea crop fields located in the Paris basin (France) in 1988, 1989 and 1990. Various yield components were recorded from six replicates of 0.5 m^2 in each field. For illustration, average values of grain numbers and stem numbers are presented in Figure 2.

The purpose of this case study is to analyse the relationship between grain number m^{-2} and reproductive stem number m^{-2} by using quantile regression. The question is of practical importance because, in pea crops, branching is known to compensate for low plant numbers but may not be sufficient to give the maximum grain number. It is thus interesting to determine the stem number threshold above which the maximum grain number can be reached. It is also interesting to determine, for each field, if the maximum grain number value was reached or not.

Table I. Case study 1. Quantile values computed with model 1.

Z_{MAX}	α	β	$\hat{\tau}$
500	1	1	0.88
500	2	2	0.90
500	2	5	0.80
500	5	2	0.98
1000	1	1	0.94
1000	2	2	0.97
1000	2	5	0.91
1000	5	2	0.99

Table II. Case study 1. Quantile values computed with model 2.

α	β	$\hat{\tau}$
1	1	0.96
2	2	0.99
2	5	0.99
5	2	0.96

Table III. Case study 1. Parameter estimates obtained with the pea dataset.

$\hat{\tau}$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
0.5	2456.23	5.07	125.98
0.80	2631.53	4.84	140.0
0.88	2725.0	6.61	133.97
0.90	2715.43	10.20	100.11
0.91	2771.0	10.58	117.05
0.94	2750.01	11.0	96.52
0.96	2792.0	10.42	118.13
0.97	2935.2	10.04	134.07
0.98	2970.0	10.42	135.22
0.99	2970.0	10.42	135.22

3.1.2. Probability distributions and quantile values

Quantiles were computed using models 1 and 2 and different probability distributions for ε and Z. Individual replicate grain number measurements were available for 12 fields (six replicates per field) and these measurements were used to define the probability distribution of the measurement errors (ε). A Kolmogorov-Smirnov test was performed on the grain number residuals, $\varepsilon_{ij} = y_{ij} - y_{i.}$ where y_{ij} is the grain number measurement in the jth replicate of the ith field and $y_{i.}$ is the average grain number in the ith field, $i = 1, \ldots, 12$ (field index), $j = 1, \ldots, 6$ (replicate index). The result showed that the residual distribution can be assumed to be normally distributed. The standard errors of the residuals ε_{ij} and of the normalised residuals $\frac{\varepsilon_{ij}}{y_{i.}}$ were then estimated, $\sqrt{\text{var}\left(\varepsilon_{ij}\right)} = 154.3$ grains m^{-2} and $\sqrt{\text{var}\left(\varepsilon_{ij}/y_{i.}\right)} = 0.10$. Finally, the measurement errors were assumed to be distributed as $\varepsilon \sim N\left(0, 154.3^2\right)$ and $\varepsilon \sim N\left\{0, \left[0.10 \times Z \times f\left(X;\theta\right)\right]^2\right\}$ for models 1 and 2, respectively.

The distribution of Z was defined as explained in Section 2.1.3. For model 1, we assumed that $Z = Z_{MIN} + (Z_{MAX} - Z_{MIN}) \times B$ and $B \sim \text{Beta}(\alpha, \beta)$. Z_{MIN} was set equal to zero and two values were considered for Z_{MAX}, 500 and 1000 grains m^{-2}. Four different beta distributions were successively considered for B, Beta(1, 1), Beta(2, 2), Beta(2, 5) and Beta(5, 2) (see Fig. 1). The two Z_{MAX} values and the four beta distributions led to eight different quantiles (Tab. I). The quantile values were computed from 100 000 values of Z and ε randomly drawn from their probability distributions.

For model 2, Z was assumed to be beta distributed, and four beta distributions were considered Beta(1, 1), Beta(2, 2), Beta(2, 5) and Beta(5, 2) (Fig. 1). These distributions were used to compute four quantile values from 100 000 values of Z and ε randomly drawn from their probability distributions (Tab. II).

Tables I and II show that the quantile value $\tau = P\left[Y < f\left(X;\theta\right)\right]$ is sensitive to the model type and to the assumptions made on the probability distribution of the limiting factor effect Z. The value of τ is higher with the multiplicative model (model 2) than with the additive model (model 1). But high quantile values are also obtained with model 1 when the lower bound of Z is set equal to a high value and when the distribution of Z is dissymmetric.

3.1.3. Parameter estimation

A linear-plus-plateau function was defined as follows:

$$f\left(X;\theta\right) = \theta_1 + \theta_2\left(X - \theta_3\right) \text{ if } X \leq \theta_3 \tag{14}$$

$$f\left(X;\theta\right) = \theta_1 \text{ if } X > \theta_3 \tag{15}$$

where $f\left(X;\theta\right)$ is the grain number m^{-2} value obtained when the reproductive stem number m^{-2} is equal to X, and $\theta = (\theta_1, \theta_2, \theta_3)^T$. θ_1 is the maximum grain number, θ_2 is the slope of the linear part of the function, and θ_3 is the stem number threshold.

The three parameters were estimated from the 71 measurements with the function nlrq. This function was run for each of the nine different quantiles reported in Tables I, II, and also for $\hat{\tau} = 0.5$ (as said above, $P\left[Y < f\left(X;\theta\right)\right] = 0.5$ if Y is not affected by any limiting factors and if the measurement errors have a symmetric distribution with zero mean). The resulting ten series of parameter estimates are reported in Table III.

The parameter values obtained for quantiles 0.98 and 0.99 are identical due to lack of data, but all the other estimates are different (Tab. III). Values of $\hat{\theta}_1$ increase in function of $\hat{\tau}$. $\hat{\theta}_1 = 2631.53$ grains m^{-2} when $\hat{\tau} = 0.80$ but $\hat{\theta}_1 = 2970.0$ when $\hat{\tau} = 0.98$. $\hat{\theta}_1$ is equal to the maximum measured grain number (2970) for $\hat{\tau} = 0.98$ and $\hat{\tau} = 0.99$, but is lower than this value for all other quantiles (Tab. III).

The slope $\hat{\theta}_2$ tends to be much higher for high quantiles than for low quantiles. For example, $\hat{\theta}_2 = 4.84$ when $\hat{\tau} = 0.80$ but $\hat{\theta}_2 = 10.42$ when $\hat{\tau} = 0.96$. The variation in the grain number threshold $\hat{\theta}_3$ is erratic and the parameter estimates are in the range 96.52–140 stems m^{-2} (Tab. III).

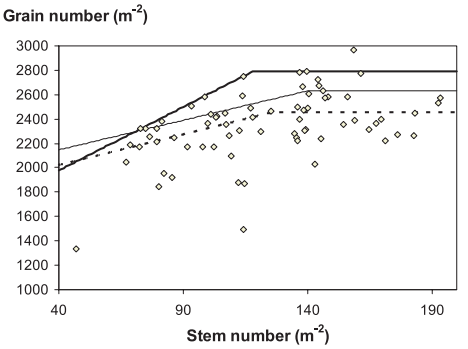


Figure 2. Case study 1. Fit of the linear-plus-plateau function obtained with $\hat{\tau}$ equal to 0.5 (dotted line), 0.80 (thin continuous line) and 0.96 (bold continuous line). Pea data from Doré et al. (1998).

Table IV. Case study 1. Parameter estimates obtained using categories of stem number when $\hat{\tau} = 0.96$.

Number of categories	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
5	2675.64	7.15	110.63
6	2743.61	9.47	117.71
7	2737.44	12.61	112.42
8	2729.23	33.56	93.59
9	2690.83	35.05	89.79
10	2624.88	35.25	86.39

3.1.4. Assessment of the estimators

The computation of the sensitivity index $\frac{\max(\hat{\theta})-\min(\hat{\theta})}{\max(\hat{\theta})}$ for each parameter is one way to analyse the sensitivity of the parameter estimates to the quantile value. This index is equal to 0.17, 0.56 and 0.31 for $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$, respectively. Thus, $\hat{\theta}_2$ is more sensitive to $\hat{\tau}$ than the other two parameters. These results emphasise that it is important to perform a sensitivity analysis of the boundary line parameter estimates with respect to the quantile value.

Three fitted boundary lines are shown in Figure 2. These lines were drawn using the parameter estimates reported in Table III for the quantiles 0.5, 0.80 and 0.96. The lines show significant differences in terms of maximum grain number, slope and stem number threshold.

The results obtained by quantile regression were compared with the results obtained by using the method defined by Webb (1972) and applied by Casanova et al. (1999) and Johnson et al. (2003). The principle is to categorise the variable X (here the stem number) in Q categories and, for each category, to determine the value of Y (grain number) corresponding to the quantile $\hat{\tau}$. The parameters θ are then estimated from the resulting Y values by least squares. Like quantile regression, this method requires the specification of a quantile value. For illustration, $\hat{\tau}$ was set equal to 0.96. Q was set equal to 5, 6, 7, 8, 9 and 10 successively and the three parameters were estimated for each value. A value Y was determined for each category by using the quantile function of R (http://cran.r-project.org).

Table IV shows that, for a given quantile, the results obtained with the method of Webb (1972) depend highly on the number of categories. For example, when $Q = 5$, $\hat{\theta}_2 = 7.15$ but, when $Q = 10$, $\hat{\theta}_2 = 35.25$. The parameter estimates obtained for $Q = 6$ are similar to those obtained by quantile regression (Tab. III) but the estimates obtained for other values of Q are quite different. No clear method has been defined to determine the optimal number of categories and it is thus more convenient to estimate boundary line parameters by quantile regression.

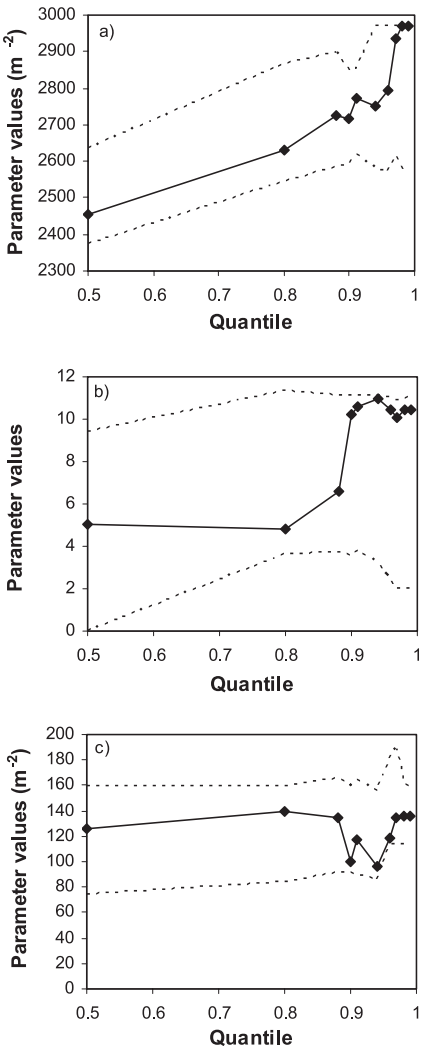


Figure 3. Case study 1. Parameter estimates (continuous lines) and 90% confidence intervals (dotted lines) for the parameters θ_1 (a), θ_2 (b) and θ_3 (c). Confidence intervals were computed by nonparametric bootstrap resampling.

Figure 3 shows the 90% confidence intervals for the parameter estimators. Confidence intervals indicate whether the parameters were accurately estimated or not. A narrow confidence interval indicates an accurate estimation. Figure 3 shows that the confidence intervals are dissymmetric, like those reported by Cade and Richards (2006). Compared with the estimated values, the lengths of the confidence intervals are small for $\hat{\theta}_1$, are higher for $\hat{\theta}_3$, and are very high for $\hat{\theta}_2$. Clearly, the parameter θ_2 is not accurately estimated but its estimators are

significantly different from zero. For this parameter, the narrowest confidence interval (and so the best estimation) is obtained with $\hat{\tau} = 0.91$ (90%CI = 3.76–11.13). This is also the case for the parameter θ_1 (90%CI = 2619.6–2854.7 for $\hat{\theta}_1$ and $\hat{\tau} = 0.91$). It is thus reasonable to select the parameter values estimated for the quantile 0.91.

3.2. Case study 2: grain weight vs. grain number

3.2.1. Model for generating data

The purpose of this second case study is to show how simulated datasets can be used to study the performance of the quantile regression method in relation to the number of observations and to the chosen quantile value.

As explained in Section 2.3.3, the first step is to define a model for generating data. We consider here the boundary line established by Gonzalez-Montaner et al. (1997) relating grain weight to grain number m^{-2} for wheat crops in the Argentinean Southern Pampa. This is a plateau-plus-linear function defined by

$$f(X;\theta) = \theta_1 \text{ if } X \leq \theta_3 \quad (16)$$

$$f(X;\theta) = \theta_1 + \theta_2 (X - \theta_3) \text{ if } X > \theta_3 \quad (17)$$

where $f(X;\theta)$ is the grain weight value (mg) obtained when the grain number m^{-2} is equal to X, and $\theta = (\theta_1, \theta_2, \theta_3)^T$. We assume that the true parameter values are those given by Gonzales-Montaner et al. (1997), $\theta_1 = 44$ mg, $\theta_2 = -0.0025$ mg per grain m^{-2}, $\theta_3 = 14400$ grain number m^{-2}. The plateau-plus-linear function (16, 17) and the chosen parameter values are only used here to demonstrate the potential of the method. Other parameter values or other boundary lines could have been considered.

According to Gonzalez-Montaner et al. (1997), water balance and temperature can have a negative effect on grain weight and this effect is additive. It is thus realistic to use the additive model (1) defined as $Y = f(X;\theta) - Z + \varepsilon$, where X is the number of grains m^{-2}, Y is the grain weight, Z is the overall limiting factor effect and ε is a measurement error. We assume that $\varepsilon \sim N\left(0, 3^2\right)$. The value chosen for the standard deviation of ε (3 mg) falls within the range of values reported by Brancourt-Hulmel et al. (1999). The variable Z was defined as $Z_{MIN} + (Z_{MAX} - Z_{MIN}) \times B$ with $B \sim$ Beta(1, 1), $Z_{MIN} = 0$, and $Z_{MAX} = 17$ mg. The value of Z_{MAX} was chosen according to the results shown in Gonzales-Montaner et al. (1997). Under these assumptions, the value of $\tau = P[Y < f(X;\theta)]$ is equal to 0.93.

3.2.2. Effect of the number of observations

We assume here that the quantile is set equal to its correct value (0.93). Five hundred datasets of N observations were generated as follows:

- X was assumed to be uniformly distributed, $X \sim$ *Uniform*(5000, 22000) (Gonzales-Montaner et al., 1997), and N values were randomly generated, $X_1, \ldots, X_N$.
- N values of maximum grain weight were calculated as $f(X_1;\theta), \ldots, f(X_N;\theta)$ where $\theta = (\theta_1, \theta_2, \theta_3)^T$ is the vector including the true parameter values defined above.
- N observed grain weight values were calculated as $f(X_1;\theta) - Z_1 + \varepsilon_1, \ldots, f(X_N;\theta) - Z_N + \varepsilon_N$, where $\varepsilon_1, \ldots, \varepsilon_N$ were randomly drawn from $\varepsilon \sim N\left(0, 3^2\right)$ and $Z_1, \ldots, Z_N$ were N values randomly drawn from the distribution of Z.

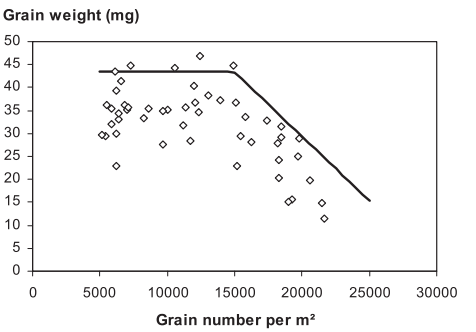


Figure 4. Case study 2. Fit of the plateau-plus-linear function to 50 simulated observations of wheat grain number and grain weight ($\hat{\tau} = 0.93$).

This procedure was implemented with six different N values, 50, 75, 100, 150, 200 and 300. The value of $\theta = (\theta_1, \theta_2, \theta_3)^T$ was estimated by quantile regression with $\hat{\tau} = 0.93$ for each generated dataset with the function nlrq (initial value equal to 35, –0.0015, and 15000 for θ_1, θ_2, θ_3, respectively). An example of a generated dataset and of a fitted curve is presented in Figure 4 for $N = 50$.

Bias, standard deviations and RMSE were computed for each N value as explained in Section 2.3.3. Bias and standard deviations are presented in Figure 5 in function of the size of the dataset. RMSE values were very close to standard deviations and are not shown. Compared with standard deviations, the bias is very small for all parameters and for all sizes of dataset (Fig. 5). This result shows that the quantile regression method does not make any systematic estimation error when τ is set equal to its correct value.

The standard deviations of the parameter estimators decrease in function of the number of observations (Fig. 5). For example, for parameter θ_1, the standard error is equal to 1.95 mg when $N = 50$ but it is equal to 0.74 when $N = 300$. However, Figure 5 shows that it is not useful to estimate the parameter values with more than 200 observations. This result is of practical interest because it indicates that the size of the dataset (50 observations) used by Gonzalez Montaner et al. (1997) was probably too small for estimating the boundary line parameters with a good accuracy.

3.2.3. Consequences of quantile misspecification

The procedure described in the previous section was repeated to study the consequence of using a wrong quantile value ($\hat{\tau} \neq 0.93$). The quantile was fixed to eight different values (0.7, 0.75, 0.8, 0.85, 0.93, 0.95, 0.97 and 0.99) and, for

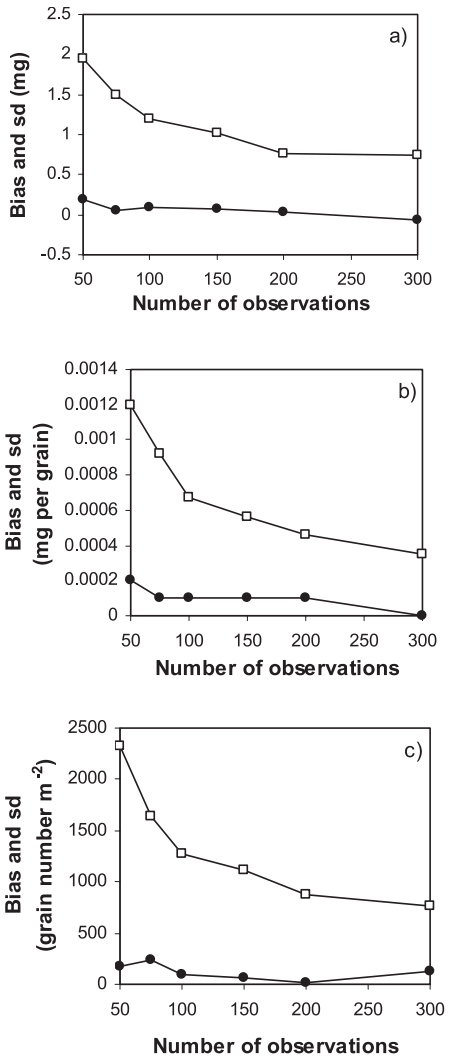


Figure 5. Case study 2. Bias (black circles) and standard deviations (white squares) of the estimator of θ_1 (a), θ_2 (b) and θ_3 (c) in function of the number of observations ($\hat{\tau} = 0.93$).

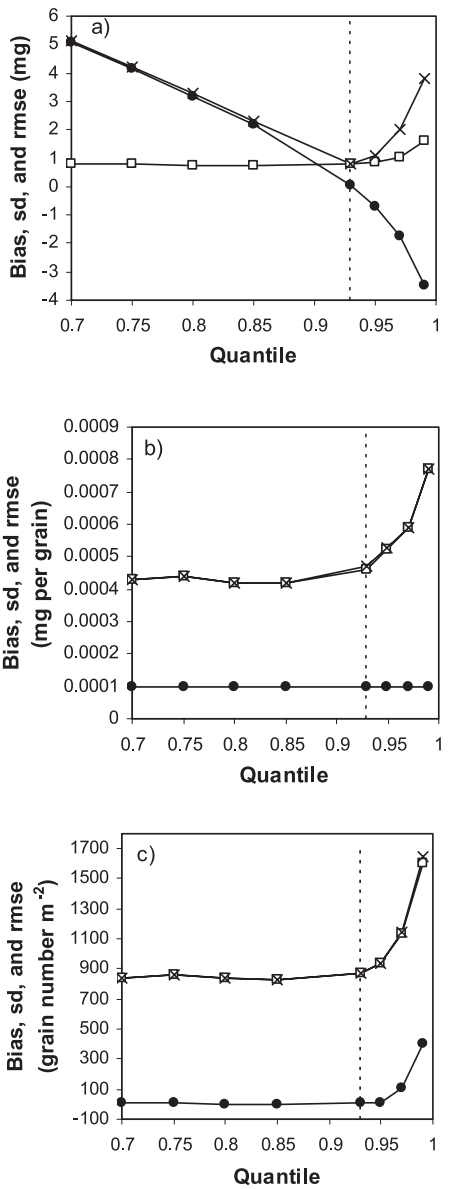


Figure 6. Case study 2. Bias (black circles), standard deviations (white squares) and RMSE (cross) of the estimator of θ_1 (a), θ_2 (b) and θ_3 (c) in function of the quantile value when the number of observations is equal to 200. The vertical bars indicate the true quantile value.

each value, the parameters were estimated with 500 datasets of $N = 200$ observations simulated from the model described in Section 3.2.1. Bias, standard deviation and RMSE of the estimators were calculated for each quantile in turn.

Figure 6a shows that the absolute value of the bias of the estimator of θ_1 increases when $\hat{\tau} \neq 0.93$. For this parameter, the bias is equal to 5.06 when $\hat{\tau} = 0.7$, is equal to 0.03 when $\hat{\tau} = 0.93$, and is equal to –3.49 when $\hat{\tau} = 0.99$. This result shows that an underestimation of the quantile leads to an underestimation of the parameter θ_1, whereas an overestimation of τ leads to an overestimation of the parameter value.

Compared with the bias, the standard deviation of the estimator of θ_1 is small. Its value tends to increase in function of the quantile value; the standard deviation is equal to 0.82 when $\hat{\tau} = 0.7$ and is equal to 1.6 when $\hat{\tau} = 0.99$ (Fig. 6a). The RMSE is almost equal to the absolute value of the bias and takes its minimum value when $\hat{\tau} = 0.93$ (Fig. 6a). This is logical. As said above, we know that $\text{rmse}_\tau \approx \sqrt{\text{bias}_\tau^2 + \text{sd}_\tau^2}$. As the standard deviation is small compared with the bias, we have $\text{rmse}_\tau \approx |\text{bias}_\tau|$.

A naïve estimator of θ_1 is the maximum measured grain weight value among the N measurements of the dataset. The RMSE of this estimator (computed from 500 simulated datasets of 200 observations) is equal to 4.64. Lecomte (2005) suggested estimating the maximum grain weight by using a bootstrap method. The principle is to generate M samples of observations from the original dataset by random sampling. θ_1 is then estimated by the average of the M maximum values

calculated from the M samples. We applied this method to our model and found that the RMSE of the estimator (computed with $M = 500$ and 200 observations) is equal to 4. The RMSE value is even higher when θ_1 is estimated from 300 observations; RMSE = 5.28 if θ_1 is estimated by the maximum observed value and RMSE = 4.14 if θ_1 is estimated by using Lecomte's method. All these values of RMSE are higher than the value obtained by quantile regression for θ_1, $\hat{\tau} = 0.99$, and 200 observations (RMSE = 3.83). Thus, it seems better to estimate θ_1 by quantile regression even if $\hat{\tau}$ is set equal to a very high (and wrong) value. This is logical because the estimator 'maximum measured grain weight' does not account for measurement error and, so, tends to overestimate the true parameter value. Another advantage of using quantile regression is that this method allows one to estimate the three model parameters simultaneously and not θ_1 alone.

For the other two parameters, θ_2 and θ_3, the results obtained by quantile regression are quite different. For these parameters, the bias is small compared with the standard deviation whatever quantile value is used (Fig. 6 b,c). The bias is close to zero and constant for θ_2. For θ_3, it tends to be higher when the quantile is overestimated.

The standard deviations of the estimators of θ_2 and θ_3 are much higher than the bias (Figs. 6 a,b). For both parameters, the standard deviation is almost constant from $\hat{\tau} = 0.7$ to $\hat{\tau} = 0.93$ (the correct value) but increases when the quantile value is overestimated (higher than 0.93). For example, the standard deviation of the estimator of θ_3 is equal to 841.54 grain number m^{-2} when $\hat{\tau} = 0.7$, is equal to 875.56 grain number m^{-2} when $\hat{\tau} = 0.93$ and to 1602.25 grain number m^{-2} when $\hat{\tau} = 0.99$. As the bias is always small, RMSE and standard deviations are very similar for θ_2 and θ_3 (Fig. 6 b,c). Like standard deviations, RMSE values increase in function of the quantile but the increase is very small from $\hat{\tau} = 0.7$ to $\hat{\tau} = 0.93$ and much more significant from $\hat{\tau} = 0.93$ to $\hat{\tau} = 0.99$.

The consequences of a misspecification of the quantile value are thus different depending on the considered parameter. For some parameters, an overestimation of the quantile leads to both an increase in the bias and an increase in the standard deviation. For others, only the standard deviation is increased

These results show that the consequences of a small overestimation of the quantile value are worse than those resulting from a small underestimation. The use of a quantile higher than 0.93 strongly increases the bias of the estimator of θ_1, and increases the variances of the estimators of all parameters. As a result, the RMSE values of the parameter estimators are much higher when the quantile is set equal to 0.99 than when $\hat{\tau} = 0.93$, although the overestimation of the quantile is small.

The use of a quantile lower than 0.93 also increases the bias of the estimator of θ_1 but not the bias of the estimators of θ_2 and θ_3. Moreover, the use of a low quantile value does not increase the standard deviations of any parameter estimator. As a result, an underestimation of τ increases the RMSE only for θ_1. For the other two parameters, the RMSE obtained with low quantile values are close to the RMSE obtained with the correct quantile value. The consequences of a misspecification of τ were studied here with a particular model and it would be interesting to perform new simulations with other models.

4. CONCLUSION

Compared with other estimation methods, quantile regression has two advantages for estimating boundary line parameters. First, quantile regression can be directly applied to the original dataset. This is not the case with the estimation method used by Webb (1972), Casanova et al. (1999) and Johnson et al. (2003). With this method, the dataset is split into Q categories and we showed that the parameter estimates are highly sensitive to the number of categories. This is a problem because no method has been proposed to choose this number. Note that, like quantile regression, this method requires the definition of a quantile value. Second, we showed that, when the correct quantile value is used, the bias and variance of the estimator obtained by quantile regression are small. The effect of an overestimation of τ and the effect of an underestimation of τ are different and, overall, the consequences of an overestimation are worse. However, the results obtained by quantile regression were better than those obtained by using the naïve estimator 'maximum measured value' even when the quantile was set equal to a very high value. We thus advise agronomists to use quantile regression to estimate the parameters of boundary lines.

This study clearly demonstrates that it is not possible to obtain accurate estimators of boundary line parameters without some knowledge of the quantile value and, so, without some information on the distributions of the model errors and of the limiting factor effects. The model error distribution can be defined from replicates obtained in experimental or farmers' fields. The definition of a realistic distribution for the limiting factor effect is more difficult and further research is needed to develop methods which can solve this problem, and be adequate from both the agronomical and statistical points of view. Meanwhile, we advise the user to compute confidence intervals for the parameter estimators in order to assess whether high quantiles are estimated with a reasonable precision.

Acknowledgements: The authors are grateful to Marie-Hélène Jeuffroy who read an earlier version of this paper and provided useful comments.

REFERENCES

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.

Cade B.S., Richards J.D. (2006) A permutation test for quantile regression, JABES 11, 106–126.

Cade B.S., Terrel J.W., Schroeder R.L. (1999) Estimating the effects of limiting factors with regression quantiles, Ecology 80, 311–323.

Cade B.S., Noon B.R., Flather C.H. (2005) Quantile regression reveals hidden bias and uncertainty in habitat models, Ecology 86, 786–800.

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

David C., Jeuffroy M.-H., Mangin M., Meynard J.-M. (2005) The assessment of Azodyn-Org model for managing nitrogen fertilization of organic winter wheat, Eur. J. Agron. 23, 225–242.

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., Meynard J.-M., Sebillotte M. (1998) The role of grain number, nitrogen nutrition and stem number in limiting pea crop (Pisum sativum) yields under agricultural conditions, Eur. J. Agron. 8, 29–37.

Efron B., Tibshirani R.J. (1986) Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy, Stat. Sci. 1, 54–77.

Fleury A. (1991) Méthodologies de l'analyse de l'élaboration du rendement, in: Picard D. (Ed.), Physiologie et production du maïs, INRA, Paris, pp. 279–290.

Gonzales Montaner J.H., Maddoni G.A., DiNapoli M.R. (1997) Modeling grain yield and grain yield response to nitrogen in spring wheat crops in the Argentinean Southern Pampa, Field Crops Res. 51, 241–252.

Hunter D.R., Lange K. (2000) Quantile regression via an MM algorithm, J. Comput. Graph. Stat. 9, 60–77.

Johnson C.K., Mortensen D.A., Wienhold B.J., Shanahan J.F., Doran J.W. (2003) Site-specific management zones based on soil electrical conductivity in a semiarid cropping system, Agron. J. 95, 303–315.

Koenker R., Basset G. (1978) Regression quantiles, Econometrica 46, 33–50.

Koenker R., d'Orey V. (1987) Computing regression quantiles, Appl. Stat. 36, 383–393.

Koenker R., Park B.J. (1996) An interior point algorithm for nonlinear quantile regression, J. Econometrics 71, 265–283.

Koenker R., Machado J.A. (1999) Goodness of fit and related inference processes for quantile regression, J. Am. Stat. Assoc. 94, 1296–1310.

Lecomte C. (2005) L'évaluation expérimentale des innovations variétales. Proposition d'outils de l'analyse de l'interaction génotype-milieu adaptés à la diversité des besoins et des contraintes des acteurs de la filière semence, Ph.D. thesis, INA P-G, Paris.

Webb R.A. (1972) Use of boundary line in the analysis of biological data, J. Hort. Sci. 47, 309–319.

Wey J., Oliver R., Manichon H., Siband P. (1998) Analysis of local limitations to maize yield under tropical conditions, Agronomie 18, 545–561.

To access this journal online:
www.edpsciences.org