Prediction of sunflower grain oil concentration as a function of variety, crop management and environment using statistical models

Fety Nambinina Andrianasolo [a,c,d,*], Pierre Casadebaig [a,d], Elie Maza [b,d], Luc Champolivier [c], Pierre Maury [a,d,1], Philippe Debaeke [a,d,1]

[a] INRA, UMR AGIR, BP 52627, 31326 Castanet-Tolosan Cedex, France
[b] INRA, UMR GBF, BP 52627, 31326 Castanet-Tolosan Cedex, France
[c] CETIOM, Centre INRA de Toulouse, BP 52627, 31326 Castanet-Tolosan Cedex, France
[d] Université de Toulouse, INP, ENSAT, BP 52627, 31326 Castanet-Tolosan Cedex, France

Keywords:
GAM
Genotype by environment interaction
Regression model
Sunflower oil concentration

A B S T R A C T

Sunflower (*Helianthus annuus* L.) raises as a competitive oilseed crop in the current environmentally friendly context. To help targeting adequate management strategies, we explored statistical models as tools to understand and predict sunflower oil concentration. A trials database was built upon experiments carried out on a total of 61 varieties over the 2000–2011 period, grown in different locations in France under contrasting management conditions (nitrogen fertilization, water regime, plant density). 25 literature-based predictors of seed oil concentration were used to build 3 statistical models (multiple linear regression, generalized additive model (GAM), regression tree (RT)) and compared to the reference simple one of Pereyra-Irujo and Aguirrezábal (2007) based on 3 variables. Performance of models was assessed by means of statistical indicators, including root mean squared error of prediction (RMSEP) and model efficiency (EF). GAM-based model performed best (RMSEP = 1.95%; EF = 0.71) while the simple model led to poor results in our database (RMSEP = 3.33%; EF = 0.09). We computed hierarchical contribution of predictors in each model by means of R^2 and concluded to the leading determination of potential oil concentration (OC), followed by post-flowering canopy functioning indicators (LAD2 and MRUE2), plant nitrogen and water status and high temperatures effect. Diagnosis of error in the 4 statistical models and their domains of applicability are discussed. An improved statistical model (GAM-based) was proposed for sunflower oil prediction on a large panel of genotypes grown in contrasting environments.

1. Introduction

Worldwide vegetable oil consumption is expected to grow by 2% per year as a result of increasing edible oil and renewable energy demands (FAO, 2012). In the 2011/2012 campaign however, oilseed grains production was greatly reduced because of adverse cropping conditions, then leading to a negative balance between supply and demand. The use of deemed tolerant oilseed crops, such as sunflower (*Helianthus annuus* L.), should be thus given consideration. The latter shows some agronomic and industrial potentialities (Ayerdi-Gotor et al., 2008; Aguirrezábal et al., 2009; Pilorgé, 2010) as a promising competitive oilseed crop.

Sunflower cultivation could be particularly improved in France, where it is often grown in limited, shallow soils, non-irrigated and poor-nutrient sites (Debaeke et al., 2006; Casadebaig, 2008). In those situations, genotype × environment × management interactions were evidenced (Grieu et al., 2008) since genotypes do not exhibit the same strategies to cope with stress in restrictive conditions (Gallais, 1992; Denis and Vear, 1994).

Obtaining higher-oil concentration varieties appeared to be an alternative track for enhancing sunflower production, and could become a plus-value for French producers (Vear et al., 2003; Roche, 2005). Sunflower oil concentration was reported to be a conservative genetic component (Fick and Miller, 1997; Ruiz and Maddonni, 2006); however, recent studies highlighted differential responses of sunflower genotypes in contrasting cropping conditions; greater variability of oil concentration was whether linked to management and environmental conditions (Champolivier et al., 2011), or to genotypic and environment interactions (Andrianasolo et al., 2012). In both cases, a good understanding of oil concentration elaboration and effects of genotype and environmental factors raised to be essential for proposing convenient management strategies targeting both grain yield and oil content.

Sunflower oil is composed of 98% fatty acids (Berger et al., 2010; Echarte et al., 2010), which are produced from two potential sources; main originates from post-flowering photosynthetic

* Corresponding author at: INRA, UMR AGIR, BP 52627, 31326 Castanet-Tolosan Cedex, France. Tel.: +33 681354270; fax: +33 561735537.
E-mail address: fandrian@toulouse.inra.fr (F.N. Andrianasolo).
[1] Co-advisors of the first author Ph.D. thesis.

http://dx.doi.org/10.1016/j.eja.2013.12.002

carbon (Merrien, 1992), supplemented with carbon assimilates stored in vegetative parts before flowering that will be remobilized thereafter (Hall et al., 1990; Merrien, 1992). Plant parts that provide carbon after flowering are considered as "source" (source pool: leaves, stems) whereas those requiring carbon at this period are denoted "sink", namely grains. Reported determinants of sunflower oil concentration are genotype and environmental factors (Connor and Hall, 1997; Champolivier et al., 2011), among which intercepted radiation, nitrogen availability, high temperatures and water stress are often cited. These factors could play on both source and sink components, though only few studies explicitly separate effects on source and sink or make the link with oil concentration.

Genotype effect – *i.e.* genotypes with intrinsic high or low-oil concentration – was described to play through kernel to hull proportion (López Pereira et al., 2000; Izquierdo et al., 2008). At source level, genotype effect could play through contrasting strategies in mobilizing pre-flowering and post-flowering available carbon (Sadras et al., 1993).

Cumulative intercepted radiation between 250 and 450 degrees days after flowering was found to be the main determinant of oil concentration ($R^2 \sim 80\%$) among sunflower hybrids in Argentine (Aguirrezábal et al., 2003). Higher plant densities could have a positive effect on source before flowering (Ferreira and Abreu, 2001) and on sink after flowering; Diepenbrock et al. (2001) suggested that the variation of oil concentration could be partly linked to negative impact of higher plant densities on final grain weights. However, Rizzardi et al. (1992) observed genotype × plant density interactive effects on final oil concentrations when comparing two contrasting genotypes.

Nitrogen effect is often described through the negative relationship between oil and protein concentration (Connor and Sadras, 1992); highest oil concentrations were met in non-fertilized treatments. Nitrogen doses that are brought during vegetative period permit to optimize dry matter at flowering (Hocking and Steer, 1983; Debaeke et al., 2012) thus potential quantity of mobilized pre-flowering assimilates during grain-filling.

High temperatures after flowering were reported to shorten grain filling duration; depending on authors, we identified various temperature thresholds: 30 °C (Aguirrezábal et al., 2009), 34 °C maximum temperatures (Chimenti et al., 2001; Rondanini et al., 2003) or 17 °C mean temperature (Angeloni et al., 2012).

Little evidence exists about the effect of water availability on oil concentration; Santonoceto et al. (2003) observed significant differences in oil concentration in the final phase of oil accumulation under water stress, with an obvious lower rate of grain oil accumulation for non-irrigated modality. Before flowering, water stress could affect leaf expansion (Casadebaig et al., 2008), while it could limit green leaves photosynthesis and duration in post-flowering period (Aguirrezábal et al., 2009).

Literature-based knowledge about sunflower oil concentration determination is illustrated in a schematic conceptual framework (Fig. 1).

To help understanding crop physiology and yield determinism, crop models are tools that are increasingly developed. These can be used for multiple purposes, either to describing complex biological systems, or to interpreting experimental results, making a diagnosis of limiting factors and providing advices and predictions toward farmers for better crop and policy management (Boote et al., 1996). Statistical/empirical models, particularly, have been of great use in the history of science. Their easiness of computing and usability enhanced their attractiveness among decision-makers and practitioners (Razi and Athappilly, 2005), while they allow highlighting relative importance of variables when much is uncertain (Lobell et al., 2005; Tittonell et al., 2008; Tulbure et al., 2012). Statistical models could be divided into two main subgroups: parametric and non-parametric. Parametric models (*e.g.* simple or multiple linear

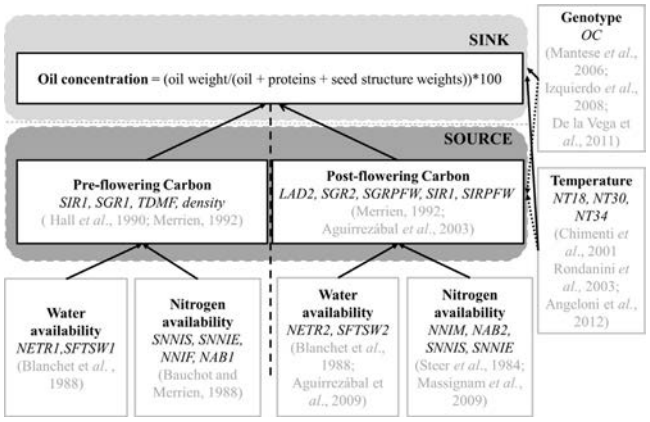

Fig. 1. Schematic framework of sunflower oil concentration elaboration as described in section 1 and relative selected predictors used for statistical modeling. Meanings of abbreviations are given in Table 2. Continuous arrows indicate literature reported relationships which were used to compute the selected predictors. Dotted arrows indicate known relationships that were not used in this study.

regression) have the advantage to be quantifiable, and assessable, but the form of the relationship between dependent and independent variable(s) should be known *a priori* to avoid misleading results; non-parametric ones (*e.g.* GAM, regression trees and neural networks) do not assume neither any *a priori* model structure nor any formal distribution of the data. They permit to bring out non-linear relationships but often lead to heavy parameterized models. Wullschleger et al. (2010) used non-parametric models to establish equations of parametric ones for switchgrass yield prediction. Other non-parametric models (regression trees, Breiman et al., 1984) were utilized to analyze yield variability in maize (Tittonell et al., 2008), wheat (Lobell et al., 2005), soybean (Zheng et al., 2009), sugarcane (Ferraro et al., 2009) or switchgrass (Wullschleger et al., 2010; Tulbure et al., 2012).

Few statistical models exist for seed oil prediction; those existing are mostly parametric. For instance, multiple linear regressions were used to model palm oil (Khamis et al., 2006; Keong and Keng, 2012), though their predictive performances were not assessed. For sunflower, a non-linear empirical model was established by Pereyra-Irujo and Aguirrezábal (2007) relating actual oil concentration to genotypic oil concentration, radiation cumulated during the post-flowering specific period (Aguirrezábal et al., 2003) and plant density. However, the model was parameterized in sites where nitrogen was non-limiting, and where water stress could be likely moderate or non-existing.

For specifically predicting oil concentration, the crop model SUNFLO (Casadebaig et al., 2011) uses a multiple linear regression model linking oil concentration with some simulated genotype, environmental stress and post-flowering canopy functioning indicators. Following oil model evaluation on an independent dataset, it was hypothesized that the acceptable though improvable RMSEP (predictive root mean squared error: ~4 oil points) was due to the narrowness of ranges of situations represented in the database, and the choice of predictors that failed to take into account physiologically-based responses of sunflower.

Therefore, the objectives of this paper are the following: (1) build statistical models based on physiologically-sound predictors and compare their predictive performance for sunflower grain oil concentration on a large dataset; (2) highlight essential features of grain oil elaboration by assessing variable importance and unraveling interactions; (3) compare the performance of these statistical models with the reference one from Pereyra-Irujo and Aguirrezábal (2007). The latter was chosen as reference model given its simplicity (low number of variables, simple equation), easiness of use

(variables that can be simulated by pre-existing model SUNFLO) and physiological-basis relevance of variables.

We proceeded similarly to Casadebaig et al. (2011) by providing model inputs to obtain simulated predictors, and include the latter into different regression models of sunflower oil concentration, while following principle of parsimony simplification approach (Crawley, 2012).

2. Materials and methods

2.1. Dataset collection

We collected sunflower oil concentration data from various French experiments conducted from 2000 to 2011 by CETIOM and INRA institutes, covering South-West to Middle-East French regions with 18 experimental sites and 61 commercial varieties in total. The whole dataset comprised 418 units of simulation (USM): each USM corresponds to one plot describing a site (soil), a growing season, a crop management and a genotype. Based on the factors studied, we established 6 categories of trials: nitrogen fertilization trials (N.trials), water regime trials (W.trials), plant density trials (D.trials), variety assessment trials (V.trials) and trials where factors were combined: nitrogen and water (N × W.trials) and nitrogen × plant density (D × N.trials). A trial was considered as a combination of experimental treatments in a given site × year. In all trials, sunflower oil concentration was measured by MNR (Magnetic Nuclear Resonance) on a subsample of seeds and expressed at equivalent 0% moisture. Information about trials, number of USM and agronomic factors is summarized in Table 1.

2.2. Simulation of oil concentration predictors

SUNFLO model was used to simulate indicators that constituted our putative predictors for modeling. These predictors were simulated by using the previous database as input data. Requested inputs for running SUNFLO dynamic model were available in most of the trials; where appropriate, experts' advice was followed when missing data. These concerned less than 10 USM.

2.2.1. Climate, soil, genotype and crop management characteristics

Climatic weather stations located within a 15 km distance from trials, provided the following meteorological data: rainfall (mm), minimum and maximum temperatures (T, °C), evapotranspiration (ET, mm) and global radiation (GR, MJ m^{-2}). Soil water availability, as a function of soil deepness and stoniness, and residual nitrogen amounts at sowing were measured by experimenters in fields. Genotypic information in SUNFLO included phenology, canopy architecture, water stress response, potential harvest index and potential oil concentration (Debaeke et al., 2010). Particularly, potential oil concentration was measured in an independent set of trials from CETIOM (Champolivier, personal communication) and computed as the maximum observed oil concentration for a given variety on a range of sites and years. Dates and rates of N fertilization and irrigation were provided for all trials, as well as planting density at emergence. Pests and diseases were adequately controlled in all experiments.

We illustrated the variability of observed oil concentration as related to genotypes, environments and management practices diversity (Fig. 2).

2.3. Choice of putative predictors for oil concentration model

Based on physiological processes and determining factors identified in literature, we chose indicators describing pre- and post-flowering periods that were related to environmental resources, canopy general functioning, nitrogen and water-linked indicators of plant state, and specific genotype characteristics (Fig. 1). Most of them were simulated by SUNFLO model since they were not measured in past field experiments. Assuming that intermediate and final variables simulated by SUNFLO have been already evaluated and considered as acceptable (Debaeke et al., 2010; Casadebaig et al., 2011), we used our 25 indicators as putative predictors for sunflower oil concentration statistical model. Information about indicators is provided in Table 2.

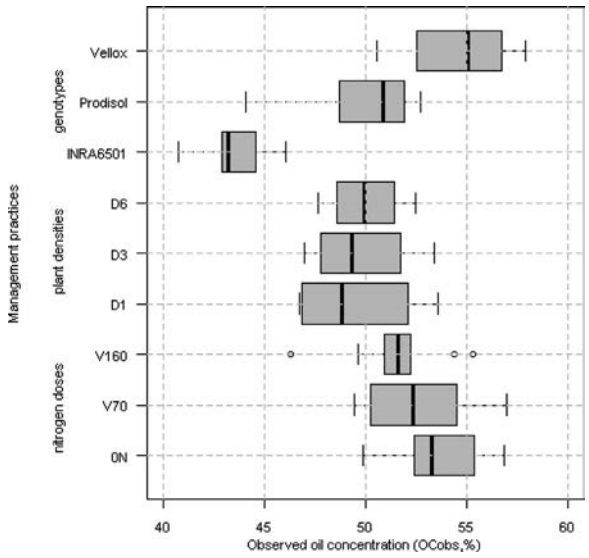

Fig. 2. Variability of observed oil concentration (OCobs) related to variability of management practices (choice of genotypes, plant densities and nitrogen doses) in dataset. 3 most contrasted modalities were picked in each management practice for illustration purposes. Nitrogen doses varied from 0N (no fertilization) to V160 (160 kg N per ha brought during vegetative stage). Plant densities range was 3 (D1) to 8 plants per m^2 (D6). We selected genotypes (INRA6501, Prodisol and Vellox) based on their contrasted potential oil concentration (OC).

2.4. Filtering USM and predictors

2.4.1. Yield difference threshold USM filtering

Before starting oil concentration modeling, we checked the goodness of fit between simulated and observed grain yields in our dataset. We assumed that in situations where SUNFLO model lacked precision to simulate yield, indicators would suffer the same imprecisions. Therefore, we decided to exclude units of simulation where the difference between observed and simulated yields was beyond a given threshold. This threshold (10 quintals per ha) was set according to the observed variability in yield differences in the dataset, and described in results section (Fig. 2).

2.4.2. Reducing multi-collinearity by deleting some putative predictors

We drew particular attention in detecting possible multicollinear variables among our predictors, which would impact the reliability of our statistical models (Dormann et al., 2013). Following the method suggested by Zuur et al. (2010), we computed the variation inflation factor (VIF) and applied a stepwise deletion of predictors according to decreasing VIF values, until a threshold of 2. We assumed that remaining predictors contained essential information so that the dropped ones were only redundant predictors.

2.5. Statistical models building

For practical purposes, we numbered the statistical models that were progressively built from 1 to 4: Pereyra-Irujo and

Table 1
Summary table of dataset trials types and corresponding number of units of simulation (USM), genotypes and sites, within the dataset. N., W., D., and V. correspond to nitrogen, water, density and variety respectively.

Trials type	Number of USM	% of whole dataset	Modalities range and number	Number of genotypes	Number of sites
N.trials	63	15	From 0 to 160 kg N per ha 7 modalities	6	7
W.trials	6	1	Rainfed and irrigated (160 and 200 mm) 3 modalities	1	1
D.trials	24	6	From 3 to 8 plants per m^2 D1–D6: 6 modalities	2	2
V.trials	273	65	From 8 to 20 varieties per site	61	8
N × W.trials	24	6	From rainfed × 0N to irrigated × 160N 12 combinations modalities	2	2
D × N.trials	28	7	From 4.8 to 6.8 plants per m^2 and 0N to 160N 10 combinations modalities	2–8	1

Aguirrezábal (2007) adjusted model, multiple linear regression using BIC-stepwise selection, GAM-wised transformed and regression tree model.

2.5.1. Model 1: Pereyra-Irujo and Aguirrezábal (2007)-adjusted model

We used the equation provided in Pereyra-Irujo and Aguirrezábal (2007) for simulating sunflower oil concentration, and applied their formula to our dataset. 3 predictors were used: SIRPFW (sum of intercepted radiation between 250 and 450 °Cd after flowering), OC (potential oil concentration), and plant density:

$$\mathrm{OCobs} = \min\left\{a + b \times \left(\frac{\mathrm{SIRPFW}}{\mathrm{density}}\right), \mathrm{OC}\right\},$$

where a and b are model parameters and correspond to intercept and slope of the linear part of the equation, respectively. Pereyra-Irujo and Aguirrezábal (2007) potential oil concentration was set at a maximum value of 50%, which was shown to be valid for many sunflower hybrids. However, 50% was quite low regarding

Table 2
List of predictor variables used to build statistical models for sunflower oil concentration, selected according to their literature-relevance characteristics, and simulated by SUNFLO model. Ranges of variation in the dataset and variables units are provided.

Categories of predictors	Predictors	Meaning	Range	Units
Environmental resources				
	SGR1	Sum of global radiation during vegetative period	686–1027	MJ/m^2
	SGR2	Sum of global radiation during reproductive period	509–916	MJ/m^2
	SGRPFW	Sum of global radiation between 250 and 450 degree days after flowering	202–313	MJ/m^2
Environmental constraints				
Water stress	NETR1	Number of days with water stress (real to maximum evapotranspiration ratio lower than 0.6) during vegetative period	0–27	days
	NETR2	Number of days with water stress (real to maximum evapotranspiration ratio lower than 0.6) during reproductive period	0–38	days
	SFTSW1	Sum of 1 – (fraction of transpirable soil water) during vegetative period	7–34	–
	SFTSW2	Sum of 1 – (fraction of transpirable soil water) during reproductive period	20–39	–
Nitrogen stress	NAB1	Sum of nitrogen quantities absorbed by plant in vegetative period	20–172	kg/ha
	NAB2	Sum of nitrogen quantities absorbed by plant in reproductive period	7–60	kg/ha
	NNIF	Nitrogen nutrition index at flowering	0.39–1.39	–
	NNIM	Nitrogen nutrition index at the beginning of grain filling	0.39–1.42	–
	SNNIE	Integration of nitrogen nutrition index when the latter exceeds the value of 1, computed on the whole crop cycle	0–34	–
	SNNIS	Integration of nitrogen nutrition index when the latter is lower than 1, computed on the whole crop cycle	0–39	–
Thermal stress	NT18	Number of days during which seed filling period mean air temperature is higher than 18 °C	21–39	days
	NT30	Number of days during which seed filling period maximum air temperature is higher than 30 °C	0–26	days
	NT34	Number of days during which seed filling period maximum air temperature is higher than 34 °C	0–13	days
Canopy functioning				
	LAD2	Leaf area duration in reproductive period	24–122	m^2 days/m^2
	SIR1	Sum of intercepted radiation during vegetative period	244–454	MJ/m^2
	SIR2	Sum of intercepted radiation during reproductive period	149–352	MJ/m^2
	SIRPFW	Sum of intercepted radiation between 250 and 450 degree days after flowering	62–139	MJ/m^2
	MRUE2	Mean radiation use efficiency during reproductive phase	0.02–0.34	g/MJ
	MRUEPFW	Mean radiation use efficiency during 250–450 degree days postflowering window	0.06–0.73	g/MJ
	TDMF	Total aerial dry matter at flowering	311–713	g/m^2
Management				
	Density	Plant density at emergence	3–8.2	plants/m^2
Genotype				
	OC	Potential oil concentration	47.7–60.8	%

our potential oil concentrations range, so we re-estimated model parameters by the use of *nls* (non-linear least squares) function of basic R to adjust to our data.

2.5.2. Model 2: multiple linear regression (MLR) and stepwise selection by BIC

Following the method of Casadebaig et al. (2011), we built an additive multiple linear regression model (MLR) with the non-dropped predictors. We then carried out a stepwise forward variables selection based on BIC (Burnham and Anderson, 2002) with the help of *stepAIC* function from "MASS" package in basic R (R Development Core Team, 2013).

2.5.3. Model 3: generalized additive model (GAM) and predictors transformations

Generalized additive models (GAM) are non-parametric models that fit to data by means of smoothing functions based on local regression splines (Wood, 2003). They are generally used to visualize possible non-linear relationships between dependent and independent variables and to check the improvement in predictive performance in case non-linear relationships were detected (Wood, 2004; Wullschleger et al., 2010). We fitted our smallest current statistical model with the *gam* function of R "mgcv" package (Wood, 2004).

We went further into investigation by checking possible equations that matched the transformations of predictors suggested by GAM – *i.e.* parameterizing the model. For this, we used Formulize Eureqa version 0.98 Beta software (Schmidt and Lipson, 2009, 2013). Various possible fitting curves were obtained; we chose equations with goodness of fit (R^2) to data higher than 98%. In any case of having several possible equations with $R^2 > 98\%$, we chose the one with the less parameters. Parameters values were proposed by the software, which we used as initial starting guesses parameters for *nls* regression in R.

2.5.4. Model 4: regression tree model

Regression tree (RT) is a non-parametric model that splits hierarchically continuous dependent variable into nodes in a binary way (Breiman et al., 1984). Splits are obtained using a recursive partitioning algorithm, where predictors appear from the one most contributing to the variance of the response variable to the least contributing. We used regression tree in order to (1) assess relative importance of variables with no assumption of linearity, (2) identify possible interactions, which we did not willingly include in our previous statistical models. *rpart* function or R "rpart" package was used for fitting RT (Breiman et al., 1984).

2.6. Statistical models evaluation and diagnosis

Performances of models were evaluated and compared according to their goodness of fit to data, predictive quality and adequacy in simulated patterns for some agronomic trials. We also computed relative variable (predictor) contribution to simulated oil concentration in each model.

2.6.1. Goodness of fit: R^2, EF

All statistical models were first evaluated for their goodness of fit to data, by computing coefficient of determination (R^2) and model efficiency (EF).

2.6.2. Predictive performance and error diagnosis: RMSEP, SDSD, SB and LCS

Then, statistical models were evaluated for their predictive performance (RMSEP) by launching leave-one-out cross-validation (LOOCV) for linear models, using *cv.lm* function of "DAAG" package of R (Maindonald and Braun, 2010). LOOCV involves using a single observation from the whole dataset as the validation set, and the remaining observations as the training set; the process is repeated such that each observation in the dataset is used once as a validation set. For GAM, ML (maximum-likelihood) method was used for model fitting, and GCV (Generalized Cross Validation) for model evaluation (Wood, 2003). For regression tree, cross-validation was used as a standard method for evaluating predictive performance with the help of *xpred.rpart* function of "rpart" package (Breiman et al., 1984). For the non-linear model adjusted from Pereyra-Irujo and Aguirrezábal (2007), we ran a LOOCV with the help of *cross.val* function of "R330" package (Lee and Robertson, 2012).

We then split global MSE into components (Kobayashi and Salam, 2000) that could bring more information on understanding of the model type of error. Components were SDSD (squared difference between standard deviations), SB (squared bias) and LCS (lack of correlation between standard deviations). High SDSD (magnitude) and LCS (pattern) values would suggest that a given statistical model fails to simulate the variability of measurements around the mean. High SB values originate from systematic behavior of the model errors.

2.6.3. Comparing response patterns to varying management practices

Simulated patterns of oil concentrations responses to varying management practices (such as nitrogen fertilization and plant density) were compared to observed ones for each statistical model. These concerned D. and N.trials.

2.6.4. Variable importance computation

We assessed relative variable importance in each statistical model by using *calc.relimp* function of R "relaimpo" package (Grömping, 2006). This function computes coefficient of determination of each variable by partitioning total model R^2 while averaging over orders. For regression tree, variable importance was automatically computed by decomposing variance and then scaled to 100%. Model 1 variable importance was assessed by calculating Sobol indice (Sobol', 2001) with Monte Carlo Sobol sensitivity analysis (*sobol* function in sensitivity package; Saltelli et al., 2000).

3. Results

3.1. Dataset diversity: cropping conditions and observed oil concentrations variability

We illustrated dataset richness and diversity by computing ranges of variations of predictors describing cropping conditions and crop states (Table 2), as well as observed oil concentrations (OCobs) variations (Fig. 2). Sum of global radiation during post-flowering period varied from 509 to 916 MJ/m^2 in the dataset. Regarding environmental constraints, water stress days indicator (NETR2) ranged from 0 to 38 days, while nitrogen nutrition index at flowering (NNIF) varied from 0.4 to 1.4. High temperatures stressing days (NT34) reached up to 13 days in some trials.

Minimum and maximum values of observed oil concentrations (OCobs) for the modalities we selected were 40.8 and 56.9% respectively, but reached up to 59.4% in a non-illustrated modality (Fig. 2). For nitrogen, density and genotype modalities illustrated here, OCobs amplitudes were 7.1, 6.8 and 10.9 oil points respectively. Per variety, OCobs range was from 0.32 to 3.72 oil points. Potential oil concentrations (OC) varied from 47.7 to 60.8% (Table 2).

3.2. Yield threshold

Differences between SUNFLO simulated and observed grain yields varied from 0.002 to 23.58 quintals per ha (Fig. 3). Though, there were only few USM that were concerned by high differences

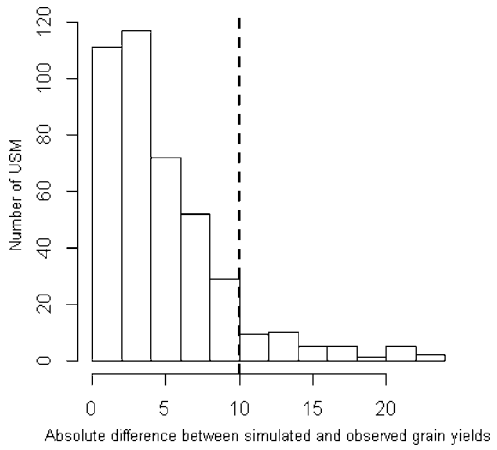

Fig. 3. Histogram of number of units of simulations (USM) as a function of absolute differences between SUNFLO simulated and observed grain yields. Dashed vertical line corresponded to the threshold (10 quintals per ha) chosen for excluding some USM from the dataset.

(higher than 10 quintals per ha). These corresponded to about 10% of total dataset. We then decided to exclude all USM which yield difference was equal or higher than this threshold. Remaining USM totalized 374.

3.3. Statistical models building

3.3.1. Model 1: Pereyra-Irujo and Aguirrezábal (2007) adjusted model

Using Pereyra-Irujo and Aguirrezábal (2007) equation, we re-estimated the parameters a and b which initial values were 36.4 and 0.5 respectively. This led to the following adjusted values: $a = 48.06$ and $b = 0.17$.

3.3.2. Model 2: multiple linear regression (MLR) and stepwise selection by BIC

There remained 12 predictors (out of 25) after VIF stepwise method for deleting multi-collinear variables. After BIC stepwise model selection, 9 predictors were retained. These were potential oil concentration (OC), water stress indicators (SFTW1 and SFTSW2), nitrogen status indicators (SNNIE and NAB2), thermal stress (NT34), canopy functioning after flowering (LAD2 and MRUE2) and management practice (density) predictors. Coefficients values were 0.08, 0.11, −0.05, −0.19, 0.03, 27.6, 0.65 and 0.97 for SFTSW1, SFTSW2, NAB2, SNNIE, NT34, LAD2, MRUE2, density and OC respectively, and −17.96 for intercept.

3.3.3. Model 3: generalized additive model (GAM) and predictors transformation

The previous 9 predictors-model, being the smallest one we got from a stepwise deletion process, was used in GAM to be compared to the linear one. Notation "s( )" corresponds to the transformed values of each predictor (Fig. 4). We first extracted transformation equations by the help of Formulize Eureqa software before plotting observed oil concentration (OCobs) with each predictor and their corresponding transformed values.

3.3.4. Model 4: regression tree model

Regression tree is illustrated in Fig. 5. The tree was highly branched (up to 8 splitting nodes) and demonstrated relatively high levels of predictors' interactions in explaining observed oil concentrations in our dataset.

The main splitting knot was linked to potential oil concentration (OC); OCobs variability of varieties having their OC lower than 54.4% (left part of the tree) was mostly linked to OC, MRUE2, SGR1, SFTSW1 and density. For those displaying higher OC and low values of SGR1 (<763.2 MJ/m^2), OCobs depended on OC, SGR1 and LAD2. If else, OCobs depended on interactions between cited predictors and MRUE2, SGRPFW, SNNIE, NAB2, and SFTSW1. 19 groups of OCobs dependencies were obtained at lower branches of the regression tree.

Table 3
Fit and prediction performances indicators of built sunflower oil concentration statistical models, averaged across all trials. Bias was measured from differences between observed and simulated oil concentrations. Coefficients of determination (R^2) and model efficiency (EF) gave equal values and are expressed on the 0–1 scale. RMSEP (root mean squared error of prediction) was computed using k-fold cross-validation. For error diagnosis, we decomposed mean squared error into SB (squared bias), SDSD (squared difference of standards deviations) and LCS (lack of correlation between standard deviations). Models are numbered from 1 to 4 and correspond to Pereyra-Irujo and Aguirrezábal (2007)-adjusted, BIC-stepwise selected, GAM-based and regression tree models, respectively.

Model	Bias	EF/R^2	RMSEP	SB	SDSD	LCS
Model 1	−0.16	0.09	3.33	0.03	7.96	3.16
Model 2	0.00	0.53	2.41	0.00	0.80	4.96
Model 3	0.01	0.71	1.95	0.10	0.32	3.48
Model 4	0.06	0.70	2.54	0.20	0.28	6.61

3.4. Comparative performances of statistical models

3.4.1. Goodness of fit, predictive performances and error diagnosis

Table 3 is a summary table of fits and predictive performances. Best fits and predictive performances were obtained with Model 3 ($R^2 = 71\%$; RMSEP = 1.95 oil points). Model 1 was the less efficient regarding its EF value (10%) and highest RMSEP (3.33 oil points). Multiple linear regression (Model 2) performed worse than Model 4 for goodness of fit to data ($R^2 = 53\%$ against 70% respectively), but better than regression tree for predictive performance (2.41 and 2.54 for RMSEP values respectively). Models biases values were all close to 0, despite being negative in Model 1 (Bias = −0.16). Graphical illustrations of simulated and observed oil concentrations relationships are provided in Fig. 6. Referring to first bisector, points of Model 1 were located on an horizontal line, while those of Models 2 and 4 were scattered along the bisector line. Model 3 displayed closest scatterplot to the 1:1 line.

Error was found to be linked to LCS component in all models (contribution varying from 86 to 95% of total mean squared error), except in Model 1 where it was rather linked to SDSD (contribution of 71%), and little to LCS (28%). Highest LCS was obtained in Model 4 (regression tree). SDSD contribution to error was relatively low in other models, but its contribution increased from Models 4, 3 to 2.

3.4.2. Variable importance comparison

We computed relative variable importance for each statistical model and compared rankings (Table 4). In all models, the most contributing variable to observed oil concentration was the potential one (OC): from 25 to 56% in Models 4 to 2, except in Model 1 where SIRPFW ranked first (88%). In this model, density and potential oil concentration had similar weights (7 and 5% respectively).

In the other models, ranking differed from second place. MRUE2 ranked second (∼12%) in Models 2 and 3, while it was SNNIE (15%) in regression tree (Model 4). LAD2 had similar relative importance as MRUE2 (∼11%) in the Models 2 and 3; SNNIE indicator was followed by SGR1 (10%) in Model 4.

4th most important variable was found to be water stress (SFTSW1 ∼5%) in Model 2, while it was density (8%) in Model 3 and post-flowering global radiation in Model 4 (SGRPFW = 8%). Thermal stress was also accounted for in the Models 2 and 3 (4 and 7%). There were lower contributions of other predictors in the Models 2 and 3 (from 2 to 5%) and in Model 4 (from 4 to 7%).

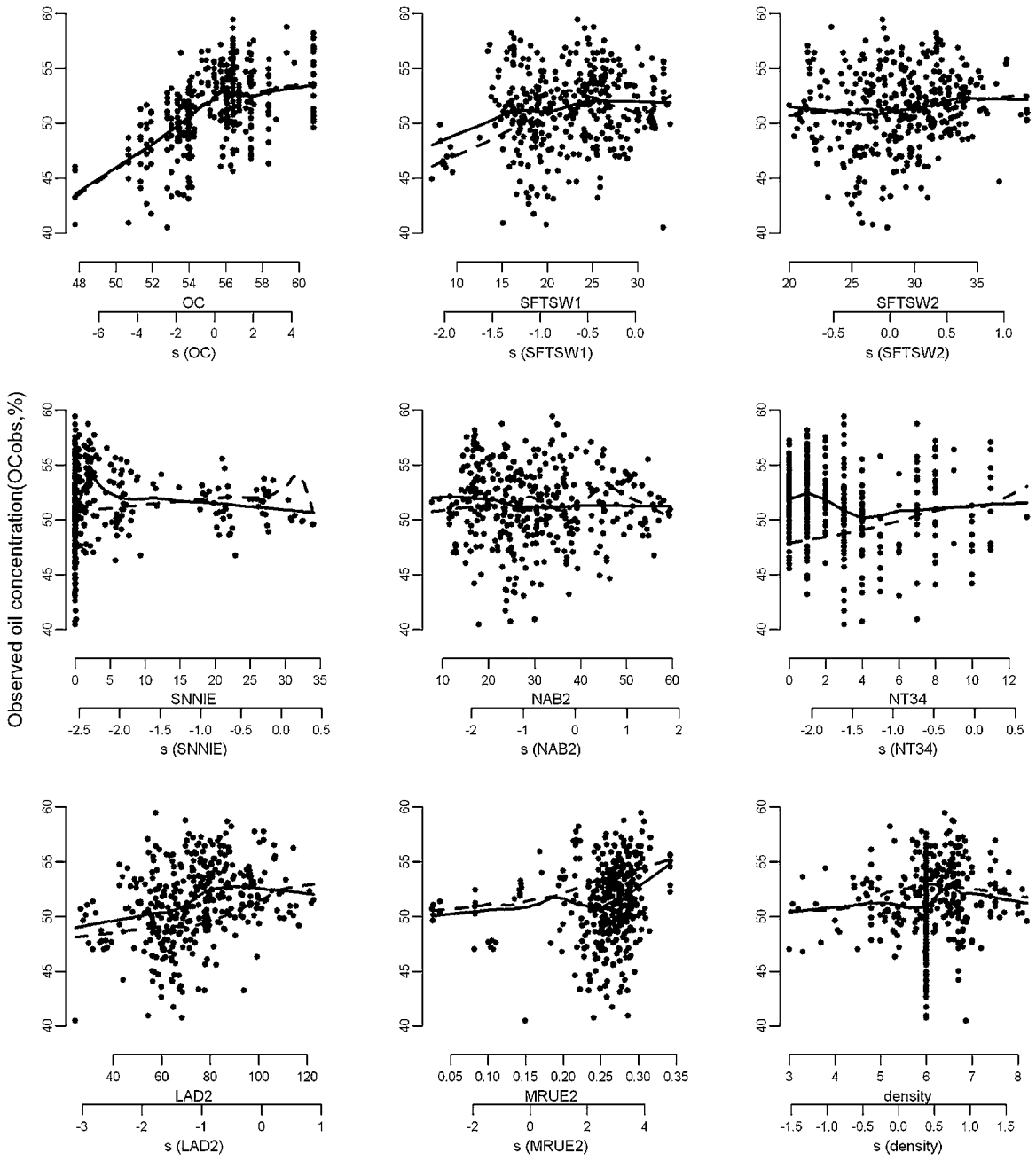

Fig. 4. Relationships between observed oil concentrations (OCobs) with each raw and GAM-transformed (prefixed "s( )") predictor respectively. Predictors were those that were selected by stepwise process in the multiple linear regression model (Model 2). Dots correspond to raw data. Continuous black (raw data smoothing) and dashed gray (transformed data smoothing) lines were obtained using *lowess* functions of R. Upper (raw data) and lower scales (transformed data) are indicated.

3.5. Patterns in response to management practices

We proposed to compare patterns of simulated (OCsim) and observed (OCobs) oil concentrations in some agronomic trials, *e.g.* D. and N.trials. We computed mean values of observed and simulated oil concentrations per modality of each agronomic factor (Fig. 6), and plotted dynamics of OCobs and OCsim against growing levels (amounts) for each model.

3.5.1. N.trials oil concentration patterns

There were three phases in observed oil concentrations patterns in response to growing nitrogen fertilization doses; OCobs stagnated between 0N to V40 (~53.5%), then slightly decreased between V40 to V70 (by 1 oil point) and sharply decreased thereafter (from 52.5 to 51%). Model 1 showed no response to growing nitrogen doses (stagnating 51.5% value). OCsim by Models 2 and 3 displayed very close patterns; those models were able to simulate only a slight oil concentration decrease at highest dose (less than 0.5 oil points); their global behavior in response to nitrogen was a stagnating oil concentration. Model 4 described a sharper decrease of OCsim starting from V70 (from 53.5 to 51.5%) compared to OCobs, but the other nitrogen modalities were badly simulated (sharp increase of 2 oil points between V20 and V70 and stagnation from 0 to V20).

3.5.2. D.trials oil concentration patterns

Mean OCobs increased slightly from D1 to D2 (from 3 to 4 plants per m^2, it varied by 0.5 oil points) and reached up to 50.5% at D4

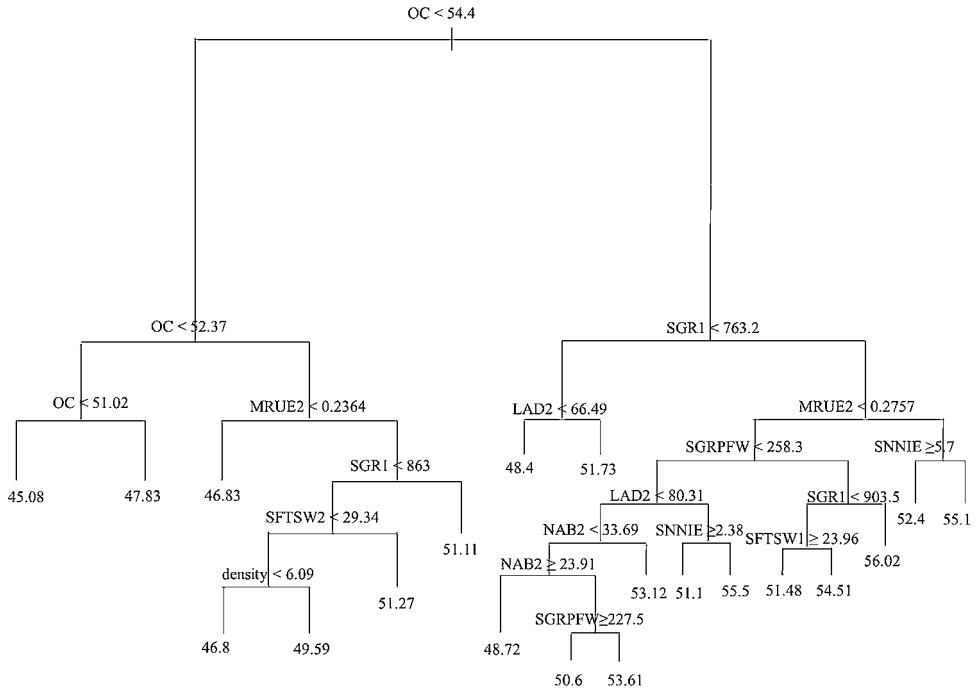

Fig. 5. Regression tree model of observed oil concentration (OCobs) as related to its most contributing predictors from the non-stepwise BIC selected initial model (12 predictors). Mean values of OCobs are represented at final ends of lower-branches. Predictors are hierarchically positioned along the branches; nodes correspond to thresholds splitting values in binary way.

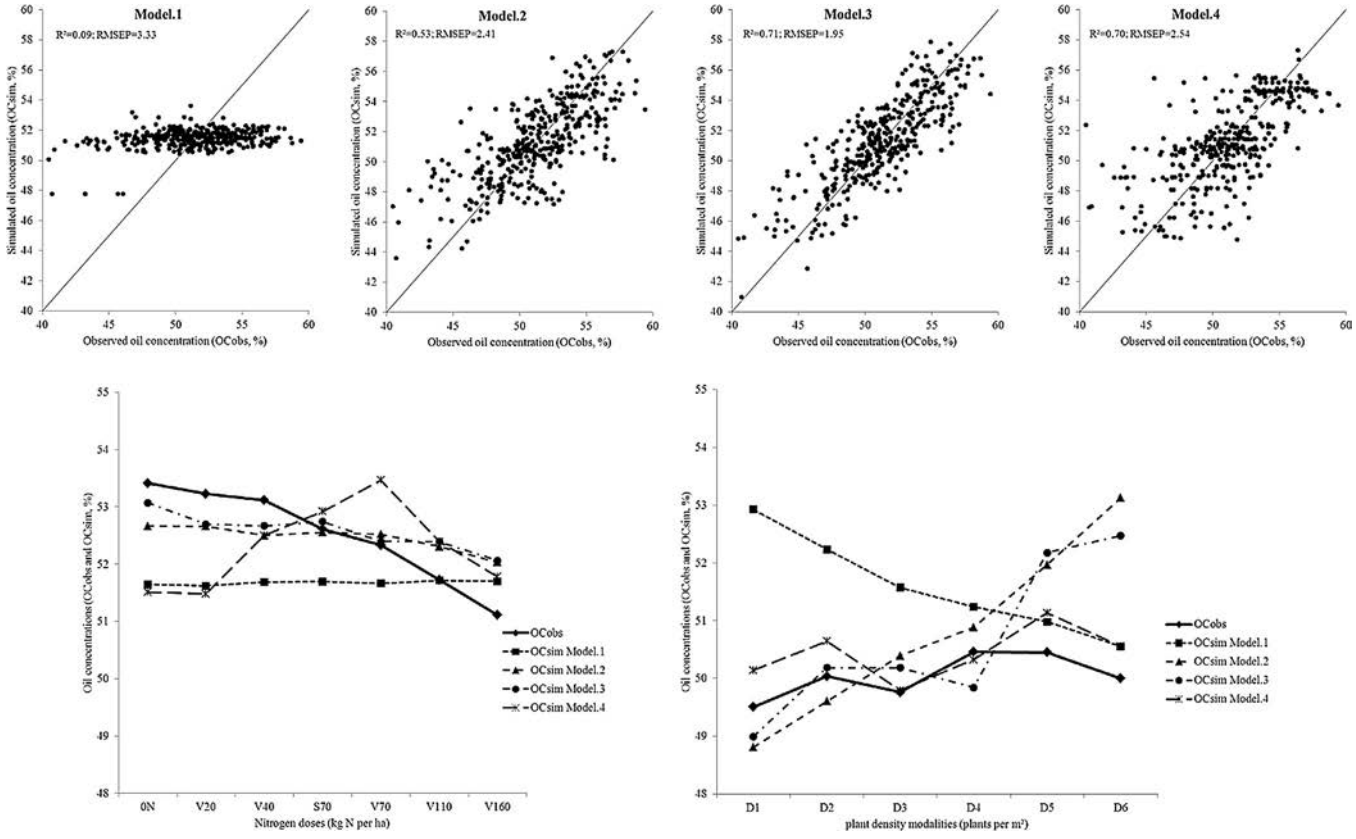

Fig. 6. Graphical patterns of simulated oil concentrations (OCsim) plotted against observed ones (OCobs). Upper line displays global scatterplot of each model, while lower line focuses on oil concentrations patterns in some agronomic trials (N. and D.trials from left to right). R^2 and RMSEP are provided as indicators of global performances of each model. Dynamics are plotted such that, for growing amounts of each factor, we could easily visualize oil concentrations patterns. Models are numbered from 1 to 4 and correspond to adjusted Pereyra-Irujo and Aguirrezábal (2007), BIC-stepwise selected, GAM-based and regression tree models respectively.

Table 4
Relative variables contribution (in % R^2 of total oil concentration variation) of retained predictors in each of the statistical models. Models are numbered from 1 to 4 and correspond to Pereyra-Irujo and Aguirrezábal (2007)-adjusted, BIC-stepwise selected, GAM-based and regression tree models, respectively.

Categories of predictors	Predictors	Model 1	Model 2	Model 3	Model 4
Environmental resources	SGR1	–	–	–	10.0
	SGRPFW	–	–	–	8.0
Water stress	SFTSW1	–	5.0	3.5	7.0
	SFTSW2	–	2.0	4.5	7.0
Nitrogen stress	NAB2	–	2.0	3.0	6.0
	SNNIE	–	3.5	2.0	15.0
Thermal stress	NT18	–	–	–	4.0
	NT34	–	4.0	7.0	1.0
Canopy functioning	LAD2	–	11.5	12.0	12.0
	SIRPFW	88.0	–	–	–
	MRUEPFW	–	–	–	–
	MRUE2	–	12.0	13.0	6.0
Management	Density	7.0	4.0	8.0	4.0
Genotype	OC	5.0	56.0	47.0	25.0

modality. Its mean value stagnated between D4 and D5 modalities, then decreased by 0.5 oil points at highest plant density (D6). There was a sharp decrease of OCsim from 53 to 51% with increasing plant density in Model 1, while it was the opposite trend in Model 2 (from 49 to 53%). Model 3 had similar patterns as OCobs but only between D1 and D2 modalities. Model 4 displayed the closest pattern to observed data, though values differed up to 0.5 oil points.

4. Discussion

4.1. Models building and methods of evaluation

This study aimed at building and comparing statistical models for predicting sunflower oil concentration in contrasting French conditions. While the statistical approaches we proposed are common in literature (Landau et al., 2000), we took particular care of avoiding statistical modeling pitfalls, especially when working with linear models (Dormann et al., 2013).

Stepwise methods for variables selection in linear modeling are widely used in science, but highly criticized for their instability, uncertainty and biased parameters (Whittingham et al., 2006). Prost et al. (2008) suggested using instead Bayesian model averaging (BMA) for selecting variables. The latter authors also evidenced that stepwise selection by BIC led to a reliable selection of predictors when the ratio between number of situations to the number of putative predictors was high, which is the case here (374 situations for 25 variables). Plus, we checked the most probable variables to be included in the linear model by the use of BMA (data not shown), from which we confirmed that the 9 predictors selected by BIC had the highest probabilities of being selected in BMA procedures as well.

We dealt with multi-collinearity by computing stepwise VIF-based indices till a threshold value of 2. The remaining predictors were considered as non-redundant; we assumed that deleted ones did not contribute essentially to oil concentration. However, it is worth noting that other VIF thresholds have been established in the literature: 5 as a common rule of thumb, or even 10 (Kutner et al., 2004). Though lower than usual approaches, chosen VIF threshold seemed adapted to the highly correlated predictors that we used in this study.

Cross-validation method was found to be a reasonable way of evaluating our models given the relatively low number of units of simulation; this technique is recommended when dataset is small in order to avoid model over-fitting (Utz et al., 2000; Hawkins et al., 2003). We used comparable method to compute models performance indicators, though the value of K differed between linear model and the non-parametric and non-linear ones (K = number of USM, K = 10 respectively). However, a 10-fold cross-validation is considered to be the minimum reliable number of sampling for minimizing bias and variance (Fushiki, 2011). Regression tree gave very good fits but very bad predictions, similarly to the study of Borra and Di Ciaccio (2010) where regression tree model over-fitted data. Rao et al. (2008) stated that the probability of under-estimating model error of prediction increased with increasing complexity of functions and decreasing number of situations; RMSEP is then probably under-estimated in the case of GAM-based model. Repeated bootstrapping methods should be used in order to obtain reliable predictive error (Efron and Tibshirani, 1997; Jiang and Simon, 2007). However, we could still compare models performances relative to each other.

4.2. Models performances

The best model was the one obtained from GAM curves and further formulized into parametric equations. This is not surprising since GAM fits closer to real data so that we can deduce simple to complex relationships depending on the structure of the data (Shatar and McBratney, 1999; Wullschleger et al., 2010). We decided to parameterize our GAM in order to obtain quantifiable indicators and compare it to other models. Its performance was equal to that of the non-parametric version (R^2 = 0.71; RMSEP = 1.95). We did not perform any model selection with GAM since we did not have *a priori* known forms of non-linear relationships (Marra and Wood, 2011).

Regression tree fitted well to data (R^2 = 0.70) and performed as well as the GAM-based one, but predicted badly (RMSEP = 2.54). Multiple linear regression with 9 predictors displayed intermediate performance (RMSEP = 2.41, EF = 0.53). Compared to existing RMSEP value in literature (1.4 oil points error for Pereyra-Irujo and Aguirrezábal (2007)), we obtained higher prediction error

values. This could be explained by the wider range of cropping situations and varieties that were used to calibrate and evaluate models; also, method of validation differed (made on an independent dataset in the case of Pereyra-Irujo and Aguirrezábal (2007), cross-validated in our study). This makes the use of RMSEP as the only method of comparing model performances questionable; we however proposed complementary indicators to evaluate our models.

Despite our willing to adapt Pereyra-Irujo and Aguirrezábal (2007) model to our dataset, the model poorly performed in our situations (RMSEP = 3.33%; EF = 0.09). Re-parameterization was justified by the fact that initial model value of potential oil concentration was set to 50%, whereas our dataset displayed a wider and higher range of OC (47.7 to 60.8%) as well as OCobs (~40.7 to 59.4%). The parameter "a" (intercept) differed by 12 oil points with the non-adjusted model, and "b" (slope) was lower in the new model (0.5 and 0.17% oil accumulation rate per MJ per m^2 respectively). Our oil concentrations were less responsive to SIRPFW/density ratios than in the Pereyra-Irujo and Aguirrezábal (2007) model. Uncertainties about SIRPFW, as part of simulated predictors, are discussed in next section. For most USM (359 out of 374), simulated oil concentration did not reach their corresponding potential value. Knowing that this potential is defined by genotype in our adjustment, it is not surprising that the new adjusted model was not able to reproduce different varietal behavior. In most cases, oil concentration was governed by factors other than potential, *i.e.* environmental factors. Only slight variations of oil concentrations were obtained in response to intercepted radiation and density effects, but Model 1 did not take into account nitrogen or water stress factors. We could add, though, that the concept of "potential" differed slightly in Model 1 compared to other models. For all cases, potential is a maximum value to be reached. In the Models 2–4, genotypic potential was included in the calculation of a "real" oil concentration just as other factors, whereas for Model 1, genotypic potential played only when it was equal to the maximum, otherwise oil concentration was determined by intercepted radiation and density. For adding "power" to genotypic determinism, varietal diversity should be included in the linear equation part.

Anyhow, this means that oil concentrations variability in our dataset could not be explained only by sum of intercepted radiation, density and potential oil concentration, so that other factors should be included in the model. Plus, the initial model was constructed such that radiation had higher importance than potential oil concentration, which contrasted greatly to other models obtained from this study.

Most of our predictors values were simulated by SUNFLO, therefore tainted with uncertainty though we took particular care of selecting USM that were acceptably predicted for their grain yields. Excluded USM (~10% of total database) displayed a mean prediction error of 40% (RMSEP = 13.8 quintals per ha). Prediction error higher than 5 quintals per ha was partly linked to soil characteristics for some trials (Middle-West region of France) where soil stoniness and shallowness limited input data accuracy and reliability. SUNFLO could not correctly simulate some extreme situations (very intense water stress, and/or over N fertilization); these might deserve a deeper physiological analysis of water × nitrogen interactions, producing specific effects probably not well reproduced by SUNFLO yet. These limitations have already been mentioned in Debaeke et al. (2010) and deserve further attention. When compared to these authors yield evaluation, we obtained similar or better mean performances, suggesting that threshold of 10 quintals per ha was comfortably acceptable. For the 374 USM remaining, mean prediction error of a given variety in a given environment was 3.88 quintals per ha (against 5 for cited authors), while mean prediction error of a given variety over all its environments equaled 4.46 quintals per ha (against 3.5 for cited authors). We could neither detect any genotype nor climatic year effects that could be linked to poorer performances of the model.

4.3. Predictors' hierarchy and contribution to oil concentration

In the models we built, potential oil concentration was considered to be the main determinant of final oil concentration (from 25 to 56% depending on the model used). This is in line with Borredon et al. (2011) and Andrianasolo et al. (2012) conclusions, who observed that genotype effect on sunflower oil concentration led to three times more oil variability than other factors (nitrogen, density). Regression tree suggested differences of functioning depending on a given threshold of potential oil concentration (54.4%); older varieties would depend on less factors than newer ones (particularly less environmental factors), thus confirming the higher sensitivity of kernel oil concentration toward environment (Aguirrezábal et al., 2009). The genetic determinism of potential oil content is complex; though QTLs for this trait have been identified in sunflower, the phenotypic variance explained by these QTLs remains relatively low (Ebrahimi et al., 2008). Leaf area duration and mean radiation use efficiency during post-flowering period were found to have similar contributions to final oil concentration and rank second after potential oil concentration (~12%), corroborating their places as the main source of photosynthetic carbon after flowering (Merrien, 1992). Differences between hierarchies given by our models were inner linked to each model own method of variance partitioning. It is reinsuring though, to obtain similar hierarchies for the top-determinant factors – OC, MRUE2/LAD2, SGR1/density; contributions of temperature and water stress deserve to be further investigated. Nitrogen was found to be as important as radiation until flowering period (regression tree); this goes in line with the observation that leaf nitrogen profile is determined by light profile in the canopy, at least until flowering and under non-limiting water conditions (Archontoulis et al., 2011). Neither radiation (SGR) nor intercepted radiations (SIR) were retained in the Models 2 and 3. SIR and SGR were in fact dropped from potential predictors in the BIC stepwise procedure. We believe that radiation effects, especially intercepted radiation ones on oil concentration, have been mitigated by the higher contributions of genotypic and stress factors effects to oil concentration variability. LAD2 (green leaf area duration after flowering) and MRUE2 (mean radiation use efficiency after flowering) behaved as better indicators of canopy functioning diversity than sum of radiation/intercepted radiation in this study, maybe because of the narrower range of variation of SGRPFW and SIRPFW indicators (coefficients of variations: SGRPFW: 10%; SIRPFW: 13%; LAD2: 25%; MRUE2: 21%, respectively). SGR indicators were though retained in Model 4 and contributed up to 10% of oil concentration variations. This can be explained by the fact that tree model helped to unravel meaningful interactions and identify important variables for contrasting situations, typically limiting/non limiting conditions. Radiation effects might be mitigated a bit less in situations where nitrogen and water were not limiting. Anyhow, this reinforces the necessity to include radiation/intercepted radiation effects in oil concentration mechanistic modeling processes. Density had a similar contribution to oil concentration in GAM-based model (8%); we could suppose that density took into account part of radiation effects though not explicitly expressed in the model. Models 2 and 4 also highlighted the importance of water availability before (5–7%) and after flowering (2 and 7%), but NT34 contribution was as high in Model 3 (7%).

4.4. Toward a better understanding of sunflower oil concentration elaboration

On a physiological point of view, genotypic effect could play whether through hull content for older varieties (López Pereira et al., 2000) or through oil concentration in kernel for more recent ones (Izquierdo et al., 2008; Aguirrezábal et al., 2009). Mantese et al. (2006) demonstrated that contrasting oil-potential cultivars differed in initial pericarp and embryo weights and dynamics, as well as oil deposition duration.

Canopy functioning indicators ranked second: as for other yield components (grain number and grain weight), sunflower oil concentration elaboration was largely source-dependent (Andrade and Ferreiro, 1996; Alonso et al., 2007). Source could be modulated by genotype, as illustrated for stay-green varieties able to maintain longer functioning leaves (De la Vega et al., 2011) or by varieties more efficient to remobilize pre-flowering assimilates after flowering (Sadras et al., 1993; López Pereira et al., 2008).

Considering intercepted radiation/density effects, lower radiation/higher plant density effects could result in lower pericarp weights as observed in Lindström et al. (2006), but could also play at source level through the relationship between radiation use efficiency and SLN (specific leaf nitrogen) for maintaining photosynthesis capability (Steer et al., 1984; Massignam et al., 2009). All things being equal, higher nitrogen doses favor higher duration of green leaf area since the onset of senescence is linked to the achievement of a minimal value of SLN in leaves (De la Vega et al., 2011); at sink level, nitrogen would enhance protein and other seed components accumulation relative to oil, leading to what Connor and Sadras (1992) call "dilution" of oil concentration.

High temperature and water stress effects deserver further investigation, especially since they could be confounded; drying could be triggered by temperature and/or water deficit, which would lead to shorter grain filling duration at sink level (Chimenti et al., 2001) and/or sooner leaf senescence at source level (Aguirrezábal et al., 2009). Higher hull weights were measured in water-stress conditions (Denis and Vear, 1994); some authors demonstrated that remobilization of pre-flowering assimilates was triggered when water was limiting (Blanchet et al., 1988; Hall et al., 1990). Others evidenced specific genotype behavior of source regulation in response to water stress (Maury et al., 2000; Casadebaig et al., 2008).

A step further in oil physiology understanding would be the calculation of source-sink indicators (Ruiz and Maddonni, 2006; Izquierdo et al., 2008) that could help to decorrelate effects of factors (genotype and environment) specifically impacting sink, source, or both.

4.5. Models error diagnosis

Diagnosis per trial type helped to highlight problems of lack of correlation (LCS) – *i.e.* faithfulness to patterns – of simulated oil concentration for all statistical models; problems of differences in magnitude (SDSD) were found in Pereyra-Irujo and Aguirrezábal (2007) adjusted-model only, but this could be explained by the fact that it could not reproduce nitrogen and water stress effects. For comparison, model error for oil concentration prediction also originated mainly from lack of correlation component (82%) in Pereyra-Irujo and Aguirrezábal (2007) paper.

We observed an average decreasing pattern of oil concentration in response to growing nitrogen fertilization amounts. Merrien (1992) stated that such depressive effect of nitrogen highly depended on water availability and water × nitrogen interaction. However, models displayed differential patterns, and none of the models could closely describe negative effect of growing nitrogen amounts. Density effect also highly depended on the model considered; each one of them revealed different thresholds at which density effect was positive or negative. For Pereyra-Irujo and Aguirrezábal (2007) adjusted-model, effect of density was observed to systematically be negative on oil concentration, though it was sometimes stagnating (from 3 to 4 plants per m^2), positive (between 4 and 6 plants per m^2) or negative (highest density) in observed oil concentration values. There were actually contrasted patterns depending on the variety × site interaction (data not shown); Vellox variety displayed decreasing OCobs values in En Crambade, while they stagnated in Montmaur. The OC values of LG5450_HO variety also stagnated in Montmaur, while increasing in En Crambade.

Assuming that each model establishes mean threshold effect of a given factor, it is not surprising that they displayed differential OCsim patterns and could not take into account individual specific pattern. This is thus the limitation of statistical models: generic relationships (or patterns) are computed, and specific genotype behavior that deviates from this generic relationship could not be correctly predicted (Shatar and McBratney, 1999; Ferraro et al., 2009). The use of more process-based models could help to unravel such specific genotype × environment × management interactions, and greatly reduce lack of correlation model error component. Before moving to more complex process-based models, correct hypotheses about oil concentration elaboration should be validated by field experiments, otherwise only the choice of process-based indicators in statistical models should be preferred (Landau et al., 2000).

With our best minimum adequate model, we were able to explain up to 70% of sunflower oil concentration variability. The general performances of our models can be considered as satisfactory when compared to other existing statistical models involving a wide range of varieties/cropping conditions (R^2: 45–61% for GAM in Tulbure et al., 2012; 51–56% for regression tree in Ferraro et al., 2009; 36–43% for multiple linear regression in Khamis et al. (2006)). The remaining unknown 30% might be linked to several causes. Though we considered simulated predictors as reliable, we could not ignore SUNFLO model uncertainties; if predictors were measured/measurable, this could have generated a much wider range of variability for some predictors, which in turn could potentially increase their contribution to oil concentration variability while reducing final prediction error (RMSEP). Also, the explicit inclusion of interacting terms might improve R^2, although regression tree highlighted simple to complex interactions but performed equally to the best minimum adequate model (GAM-based). It is not excluded that some predictors we dropped by stepwise VIF procedure could have added some explanatory power, suggesting that a less "severe" threshold could have been chosen for dealing with multi-collinearity.

We have established a comprehensive list of putative predictors by the help of our conceptual framework, but we may have missed other possible variables that could have been relevant if expressed in a different way. For instance, identifying periods of thermal time-based sensitivity to stress factors (water stress, high/low amounts of nitrogen), as done for intercepted radiation (Aguirrezábal et al., 2003), could lighten the weight of complex interactions and establish a strong common physiological basis for oil concentration response to water or nitrogen factors, regardless of genotype or other environmental conditions.

In this study, we aimed at building the most parsimonious minimum adequate model and particularly focused on the trade-off between low number of variables, predictive and explanatory power. Model 1 was not enough explanatory nor predictive, though it was totally the contrary in Argentine experiments (Aguirrezábal et al., 2003; Pereyra-Irujo and Aguirrezábal, 2007). Indeed, this model was initially calibrated on mainly one variety and in non-limited conditions. Our attempt to re-parameterize the model did

not give satisfactory results; there is rather a need to include more than 3 variables for describing oil concentration variability in response to contrasting environmental and management effects.

Model 2 (MLR) better fitted to our data; we gained in satisfactory predictive power with 6 more variables, and those that were selected have a legitimate physiologically-sound basis, assuming a linear relationship between each predictor and oil concentration which might be a too simple way to model reality. Though, most important contributors have been identified and confirmed in more complex equations (Model 3). Then, Model 2 could be used by agronomists if this is about identifying determining factors and bringing more information about sunflower grain oil physiology.

Model 3 (GAM) added more predictive power to Model 2 with the same number of variables, but the transformed relationships deserve to be assessed on other datasets. The aim would be to dissociate relationships artificially generated by the structure of our dataset and "real" relationships having sound physiological explanation. Anyhow, Model 3 could be used by both physiologists and crop modelers, for understanding and predicting sunflower oil concentration.

Finally, Model 4 (regression tree) retained more or less the same number of variables as Models 2 and 3; most contributing variables were identified but types of relationships remained unknown. This model could be more useful to an agronomist or a crop modeler, who wants to be routed for identifying main trends or possibly for decision support tool.

Decomposition of processes by source and sink and effects of determining factor on respective components appear to be essential for better understanding final oil concentration elaboration regarding genotype × environment × management interactions and leading toward a more mechanistic model.

5. Conclusion

This study aimed at building and comparing statistical models for sunflower oil concentration prediction. The GAM-based model performed best whereas the Pereyra-Irujo and Aguirrezábal (2007) adjusted one was not adapted to our data. Though displaying differential patterns in response to agronomic practices, the models helped to establish an hierarchy among determining factors of observed oil concentration; varietal potential oil concentration ranked first, and depending on oil percent amounts, interacted differently with environmental (radiation, nitrogen, water, temperature) and management practices (density) factors. This helped us to better understand source and sink relationships and order of priority for oil elaboration, which could be of valuable interest when moving to more mechanistic models.

References

Aguirrezábal, L., Martre, P., Pereyra-Irujo, G., Izquierdo, N., Allard, V., 2009. Management and breeding strategies for the improvement of grain and oil quality. In: Crop Physiology. Academic Press, San Diego, pp. 387–421 (Chapter 16).

Aguirrezábal, L.A., Lavaud, Y., Dosio, G.A., Izquierdo, N.G., Andrade, F.H., González, L.M., 2003. Intercepted solar radiation during seed filling determines sunflower weight per seed and oil concentration. Crop Science 43, 152–161.

Alonso, A.P., Goffman, F.D., Ohlrogge, J.B., Shachar-Hill, Y., 2007. Carbon conversion efficiency and central metabolic fluxes in developing sunflower (*Helianthus annuus* L.) embryos. Plant Journal 52, 296–308.

Andrade, F.H., Ferreiro, M.A., 1996. Reproductive growth of maize, sunflower and soybean at different source levels during grain filling. Field Crops Research 48, 155–165.

Andrianasolo, F.N., Champolivier, L., Maury, P., Debaeke, P.,2012. Plant density contribution to seed oil content the responses of contrasting sunflower genotypes grown in multi-environmental network. In: Proceedings of the 18th International Sunflower Conference. Mar del Plata and Balcarce, Argentina, pp. 724–729.

Angeloni, P., Echarte, M.M., Aguirrezábal, L.A.N.,2012. Temperature during grain filling affects grain weight and oil concentration in sunflower hybrid both directly and through the reduction of radiation interception. In: Proceedings of the 18th International Sunflower Conference. Mar del Plata and Balcarce, Argentina, pp. 354–359.

Archontoulis, S.V., Vos, J., Yin, X., Bastiaans, L., Danalatos, N.G., Struik, P.C., 2011. Temporal dynamics of light and nitrogen vertical distributions in canopies of sunflower, kenaf and cynara. Field Crops Research 122, 186–198.

Ayerdi-Gotor, A., Berger, M., Labalette, F., Centis, S., Eychenne, V., Daydé, J., Calmon, A., 2008. Variabilité des teneurs et compositions des composés mineurs dans l'huile de tournesol au cours du développement du capitule. Oléagineux, Corps Gras, Lipides 15, 400–406.

Berger, M., Ayerdi-Gotor, A., Sarrafi, A., Maury, P., Daydé, J., Calmon, A., 2010. Compréhension du determinisme de la qualité des huiles du tournesol face aux nouvelles attentes. Oléagineux, Corps Gras, Lipides 17, 171–184.

Blanchet, R., Piquemal, M., Cavalié, G., Hernandez, M., Quinones, H.,1988. Influence de contraintes hydriques sur la répartition des assimilats entre les organes du tournesol. In: Proceedings of the 12th International Sunflower Conference. Novi-Sad, Yugoslavia, pp. 124–129.

Boote, K.J., Jones, J.W., Pickering, N.B., 1996. Potential uses and limitations of crop models. Agronomy Journal 88, 704–716.

Borra, S., Di Ciaccio, A., 2010. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Computational Statistics & Data Analysis 54, 2976–2989.

Borredon, M.E., Berger, M., Dauguet, S., Labalette, F., Merrien, A., Mouloungui, Z., Raoul, Y., 2011. Débouchés actuels et futurs du tournesol produit en France – Critères de qualité. Innovations Agronomiques 14, 19–38.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and regression trees. Belmont, Chapman and Hall (Wadsworth, Inc.), New York, USA.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach. Springer Verlag, New York.

Casadebaig, P., (Ph.D. thesis) 2008. Analyse et modélisation de l'interaction génotype-environnement-conduite de culture: application au tournesol (*Helianthus annuus* L.). Institut National Polytechnique, Toulouse.

Casadebaig, P., Debaeke, P., Lecoeur, J., 2008. Thresholds for leaf expansion and transpiration response to soil water deficit in a range of sunflower genotypes. European Journal of Agronomy 28, 646–654.

Casadebaig, P., Guilioni, L., Lecoeur, J., Christophe, A., Champolivier, L., Debaeke, P., 2011. SUNFLO, a model to simulate genotype-specific performance of the sunflower crop in contrasting environments. Agricultural and Forest Meteorology 151, 163–178.

CETIOM – Centre technique des oléagineux [WWW Document], http://www.cetiom.fr/ (accessed 7.12.13).

Champolivier, L., Debaeke, P., Thibierge, J., Dejoux, J.F., Ledoux, S., Ludot, M., Berger, F., Casadebaig, P., Jouffret, P., Vogrincic, C., 2011. Construire des stratégies de production adaptées aux débouchés à l'échelle du bassin de collecte. Innovations Agronomiques 14, 39–57.

Chimenti, C.A., Hall, A.J., Sol López, M., 2001. Embryo-growth rate and duration in sunflower as affected by temperature. Field Crops Research 69, 81–88.

Connor, D.J., Hall, A.J., 1997. Sunflower physiology. In: Schneiter, A.A. (Ed.), Sunflower Technology and Production. ASA, Madison, WI, USA, pp. 113–182.

Connor, D.J., Sadras, V.O., 1992. Physiology of yield expression in sunflower. Field Crops Research 30, 333–389.

Crawley, M.J., 2012. The R Book, 2nd ed. John Wiley & Sons, Chichester, West Sussex, United Kingdom.

De la Vega, A.J., Cantore, M.A., Sposaro, M.M., Trápani, N., López Pereira, M., Hall, A.J., 2011. Canopy stay-green and yield in non-stressed sunflower. Field Crops Research 121, 175–185.

Debaeke, P., Mailhol, J.-C., Bergez, J.-E., 2006. Adaptations agronomiques à la sécheresse. Systèmes de grande culture. In: Sécheresse et agriculture. Réduire la vulnérabilité de l'agriculture à un risque accru de manque d'eau. INRA, France, pp. 258–360.

Debaeke, P., Casadebaig, P., Haquin, B., Mestries, E., Palleau, J.-P., Salvi, F., 2010. Simulation de la réponse variétale du tournesol à l'environnement à l'aide du modèle SUNFLO. Oléagineux, Corps Gras, Lipides 17, 143–151.

Debaeke, P., van Oosterom, E.J., Justes, E., Champolivier, L., Merrien, A., Aguirrezabal, L.A.N., González-Dugo, V., Massignam, A.M., Montemurro, F., 2012. A species-specific critical nitrogen dilution curve for sunflower (*Helianthus annuus* L.). Field Crops Research 136, 76–84.

Denis, L., Vear, F., 1994. Environmental effects on hullability of sunflower hybrids. Agronomie 14, 589–597.

Diepenbrock, W., Long, M., Feil, B., 2001. Yield and quality of sunflower as affected by row orientation, row spacing and plant density. Bodenkultur-Wien and Munchen 52, 29–36.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 027–046.

Ebrahimi, A., Maury, P., Berger, M., Poormohammad Kiani, S., Nabipour, A., Shariati, F., Grieu, P., Sarrafi, A., 2008. QTL mapping of seed-quality traits in sunflower recombinant inbred lines under different water regimes. Genome 51, 599–615.

Echarte, M.M., Pereyra-Irujo, P.-I., Covi, M., Izquierdo, N.G., Aguirrezábal, L.A.N., 2010. Producing better sunflower oils in a changing environment. In: Advances in Fats and Oil Research, Transworld Research Network. Mabel Cristina Tomás, Argentina, pp. 1–23.

Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: the 632+ bootstrap method. Journal of the American Statistical Association 92, 548–560.

Ferraro, D.O., Rivero, D.E., Ghersa, C.M., 2009. An analysis of the factors that influence sugarcane yield in Northern Argentina using classification and regression trees. Field Crops Research 112, 149–157.

Ferreira, A.M., Abreu, F.G., 2001. Description of development, light interception and growth of sunflower at two sowing dates and two densities. Mathematics and Computers in Simulation 56, 369–384.

Fick, G.N., Miller, J.F., 1997. Sunflower breeding. In: Schneiter, A.A. (Ed.), Sunflower Technology and Production. ASA, Madison, WI, USA, pp. 395–439.

Food and Agriculture Organization of the United Nations [WWW Document], 2012. http://www.fao.org/home/en/ (accessed 7.12.13).

Fushiki, T., 2011. Estimation of prediction error by using *K*-fold cross-validation. Statistics and Computing 21, 137–146.

Gallais, A., 1992. Bases génétiques et stratégie de sélection de l'adaptation générale. Le Sélectionneur Français 42, 59–78.

Grieu, P., Maury, P., Debaeke, P., Sarrafi, A., 2008. Améliorer la tolérance à la sécheresse du tournesol: apports de l'écophysiologie et de la génétique. Innovations Agronomiques 2, 37–51.

Grömping, U., 2006. Relative importance for linear regression in R: the package relaimpo. Journal of Statistical Software 17, 1–27.

Hall, A.J., Whitfield, D.M., Connor, D.J., 1990. Contribution of pre-anthesis assimilates to grain-filling in irrigated and water-stressed sunflower crops II. Estimates from a carbon budget. Field Crops Research 24, 273–294.

Hawkins, D.M., Basak, S.C., Mills, D., 2003. Assessing model fit by cross-validation. Journal of Chemical Information and Computer Science 43, 579–586.

Hocking, P.J., Steer, B.T., 1983. Distribution of nitrogen during growth of sunflower (*Helianthus annuus* L.). Annals of Botany 51, 787–799.

Izquierdo, N.G., Dosio, G.A.A., Cantarero, M., Luján, J., Aguirrezábal, L.A.N., 2008. Weight per grain, oil concentration, and solar radiation intercepted during grain filling in black hull and striped hull sunflower hybrids. Crop Science 48, 688–699.

Jiang, W., Simon, R., 2007. A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. Statistics in Medicine 26, 5320–5334.

Keong, Y.K., Keng, W.M., 2012. Statistical modeling of weather-based yield forecasting for young mature oil palm. APCBEE Procedia 4, 58–65.

Khamis, A., Ismail, Z., Haron, K., Mohammed, A.T., 2006. Modeling oil palm yield using multiple linear regression and robust M-regression. Journal of Agronomy 5, 32–36.

Kobayashi, K., Salam, M.U., 2000. Comparing simulated and measured values using mean squared deviation and its components. Agronomy Journal 92, 345–352.

Kutner, M.H., Nachtsheim, C., Neter, J., 2004. Applied Linear Regression Models, 5th ed. McGraw-Hill/Irwin, New York.

Landau, S., Mitchell, R.A.C., Barnett, V., Colls, J.J., Craigon, J., Payne, R.W., 2000. A parsimonious, multiple-regression model of wheat yield response to environment. Agricultural and Forest Meteorology 101, 151–166.

Lee, A., Robertson, B., 2012. R330 Package [WWW Document]. http://cran.r-project.org/web/packages/R330/R330.pdf

Lindström, L.I., Pellegrini, C.N., Aguirrezábal, L.A.N., Hernández, L.F., 2006. Growth and development of sunflower fruits under shade during pre and early post-anthesis period. Field Crops Research 96, 151–159.

Lobell, D.B., Ortiz-Monasterio, J.I., Asner, G.P., Naylor, R.L., Falcon, W.P., 2005. Combining field surveys, remote sensing, and regression trees to understand yield variations in an irrigated wheat landscape. Agronomy Journal 97, 241–249.

López Pereira, M., Trapani, N., Sadras, V.O., 2000. Genetic improvement of sunflower in Argentina between 1930 and 1995. Part III: Dry matter partitioning and grain composition. Field Crops Research 67, 215–221.

López Pereira, M., Berney, A., Hall, A.J., Trápani, N., 2008. Contribution of pre-anthesis photoassimilates to grain yield: Its relationship with yield in Argentine sunflower cultivars released between 1930 and 1995. Field Crops Research 105, 88–96.

Maindonald, J., Braun, W.J., 2010. Data Analysis and Graphics Using R: An Example-Based Approach. Cambridge University Press, Cambridge, United Kingdom.

Mantese, A.I., Medan, D., Hall, A.J., 2006. Achene structure, development and lipid accumulation in sunflower cultivars differing in oil content at maturity. Annals of Botany 97, 999–1010.

Marra, G., Wood, S.N., 2011. Practical variable selection for generalized additive models. Computational Statistics & Data Analysis 55, 2372–2387.

Massignam, A.M., Chapman, S.C., Hammer, G.L., Fukai, S., 2009. Physiological determinants of maize and sunflower grain yield as affected by nitrogen supply. Field Crops Research 113, 256–267.

Maury, P., Berger, M., Mojayad, F., Planchon, C., 2000. Leaf water characteristics and drought acclimation in sunflower genotypes. Plant and Soil 223, 155–162.

Merrien, A., 1992. Physiologie du tournesol. Centre Technique Interprofessionnel des Oléagineux Métropolitain (CETIOM), Paris, France.

Pereyra-Irujo, G.A., Aguirrezábal, L.A.N., 2007. Sunflower yield and oil quality interactions and variability: analysis through a simple simulation model. Agricultural and Forest Meteorology 143, 252–265.

Pilorgé, É., 2010. Nouveau contexte environnemental et réglementaire: quel impact pour la culture du tournesol? Oléagineux, Corps Gras, Lipides 17, 136–138.

Prost, L., Makowski, D., Jeuffroy, M.-H., 2008. Comparison of stepwise selection and Bayesian model averaging for yield gap analysis. Ecological Modelling 219, 66–76.

R Development Core Team, 2013. R: a language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Rao, R.B., Fung, G., Rosales, R., 2008. On the Dangers of Cross-Validation. An Experimental Evaluation. SDM, pp. 588–596.

Razi, M., Athappilly, K., 2005. A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. Expert Systems with Applications 29, 65–74.

Rizzardi, M.A., da Silva, P.R.F., da Rocha, A.B., 1992. Dry matter and oil partitioning in sunflower achenes as a function of cultivar and plant density. In: Proceedings of the 13th International Sunflower Conference, Pisa, Italy, pp. 7–11.

Roche, J., (Ph.D. thesis) 2005. Composition de la graine de tournesol (*Helianthus annuus* L.) sous l'effet conjugué des contraintes agri-environnementales et des potentiels varietaux. Institut National Polytechnique, Toulouse.

Rondanini, D., Savin, R., Hall, A.J., 2003. Dynamics of fruit growth and oil quality of sunflower (*Helianthus annuus* L.) exposed to brief intervals of high temperature during grain filling. Field Crops Research 83, 79–90.

Ruiz, R.A., Maddonni, G.A., 2006. Sunflower seed weight and oil concentration under different post-flowering source-sink ratios. Crop Science 46, 671–680.

Sadras, V.O., Connor, D.J., Whitfield, D.M., 1993. Yield, yield components and source-sink relationships in water-stressed sunflower. Field Crops Research 31, 27–39.

Saltelli, A., Chan, K., Scott, E.M., 2000. Sensitivity analysis. Wiley, New York.

Santonoceto, C., Anastasi, U., Riggi, E., Abbate, V., 2003. Accumulation dynamics of dry matter, oil and major fatty acids in sunflower seeds in relation to genotype and water regime. Italian Journal of Agronomy 7, 3–14.

Schmidt, M., Lipson, H., 2009. Distilling free-form natural laws from experimental data. Science 324, 81–85.

Schmidt, M., Lipson, H., 2013. Eureqa (Version 0.98 beta) [Software], Available from http://www.eureqa/com// (accessed 7.12.13).

Shatar, T.M., McBratney, A.B., 1999. Empirical modeling of relationships between sorghum yield and soil properties. Precision Agriculture 1, 249–276.

Sobol', I.M., 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation 55, 271–280.

Steer, B.T., Hocking, P.J., Kortt, A.A., Roxburgh, C.M., 1984. Nitrogen nutrition of sunflower (*Helianthus annuus* L.): yield components, the timing of their establishment and seed characteristics in response to nitrogen supply. Field Crops Research 9, 219–236.

Tittonell, P., Shepherd, K., Vanlauwe, B., Giller, K., 2008. Unravelling the effects of soil and crop management on maize productivity in smallholder agricultural systems of western Kenya—an application of classification and regression tree analysis. Agriculture, Ecosystems & Environment 123, 137–150.

Tulbure, M.G., Wimberly, M.C., Boe, A., Owens, V.N., 2012. Climatic and genetic controls of yields of switchgrass, a model bioenergy species. Agriculture, Ecosystems & Environment 146, 121–129.

Utz, H.F., Melchinger, A.E., Schön, C.C., 2000. Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. Genetics 154, 1839–1849.

Vear, F., Bony, H., Joubert, G., Tourvieille de Labrouhe, D.T., Pauchet, I., Pinochet, X., 2003. 30 years of sunflower breeding in France. Oléagineux, Corps Gras, Lipides 10, 66–73.

Whittingham, M.J., Stephens, P.A., Bradbury, R.B., Freckleton, R.P., 2006. Why do we still use stepwise modelling in ecology and behavior? Journal of Animal Ecology 75, 1182–1189.

Wood, S.N., 2003. Thin plate regression splines. Journal of the Royal Statistical Society. Series B, Statistical Methodology 65, 95–114.

Wood, S.N., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association 99, 673–686.

Wullschleger, S.D., Davis, E.B., Borsuk, M.E., Gunderson, C.A., Lynd, L.R., 2010. Biomass production in switchgrass across the United States: database description and determinants of yield. Agronomy Journal 102, 1158–1168.

Zheng, H., Chen, L., Han, X., Zhao, X., Ma, Y., 2009. Classification and regression tree (CART) for analysis of soybean yield variability among fields in Northeast China: the importance of phosphorus application rates under drought conditions. Agriculture, Ecosystems & Environment 132, 98–105.

Zuur, A.F., Ieno, E.N., Elphick, C.S., 2010. A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution 1, 3–14.