HAL
open science

Statistical Learning Methodology to Leverage the Diversity of Environmental Scenarios in Crop Data : Application to the prediction of crop production at large-scale

Xiangtuo Chen

To cite this version:

Xiangtuo Chen. Statistical Learning Methodology to Leverage the Diversity of Environmental Scenarios in Crop Data : Application to the prediction of crop production at large-scale. Statistics [math.ST]. Université Paris Saclay (COmUE), 2019. English. NNT : 2019SACLC055 . tel-03164008

HAL Id: tel-03164008

https://theses.hal.science/tel-03164008v1

Submitted on 9 Mar 2021

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.

université PARIS-SACLAY

CentraleSupélec

NNT : 2019SACLC055

Statistical Learning Methodology to Leverage the Diversity of Environmental Scenarios in Crop Data. Application to the Prediction of Crop Production at Large-Scale

Thèse de doctorat de l'Université Paris-Saclay
préparée à CentraleSupélec

Ecole doctorale n°573 Interface:
Approches interdisciplinaires / fondements, applications et innovation
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Gif-sur-Yvette, le 4 Juillet 2019, par

XIANGTUO CHEN

Composition du Jury :

Prof. Céline Hudelot Laboratoire MICS, CentraleSupélec	Président
Prof. Marc Jaeger UMR AMAP, CIRAD	Rapporteur
Prof. Baogang Hu Institute of Automation, Chinese Academy of Sciences	Rapporteur
Prof. Samis Trevezas Department of Math, National and Kapodistrian University of Athens	Examinateur
Prof. Paul-Henry Cournède Laboratoire MICS, CentraleSupélec	Directeur de thèse

Thèse de doctorat

Statistical Learning Methodology to Leverage the Diversity of Environmental Scenarios in Crop Data. Application to the Prediction of Crop Production at Large-Scale

CentraleSupélec

Xiangtuo Chen

MICS Laboratory, CentraleSupélec
University of Paris Saclay

This dissertation is submitted for the degree of
Doctor of Philosophy

École doctorale Interface

March 2020

Acknowledgements

Ph.D. thesis is always a long journey, and finally, my journey comes to an end. Now, it is my turn to give my thanks to those who have been along with me for this journey and those who will still be important for the rest of my life.

First of all, I need to thank my supervisor, Prof. Paul-Henry Cournède. Thank you for trusting me and choosing me to join your research team at the beginning, even if we haven't seen each other before. I'm so impressed by your professional skills and knowledge as a researcher and thesis supervisor. I still keep all the emails from you late in the evening. Your ideas and experience helped me a lot during this period. I also felt so grateful for your patience with me when my research came across some trouble. Thanks for your encouragement. Of course, as a new husband and a new father, I appreciate it a lot for your advice in my daily life.

I would also like to express my gratitude to Prof. Baogang Hu and Prof. Marc Jaeger for accepting to report my thesis. Your insightful suggestions and helpful advice give me a lot of new research perspectives. Thanks also to Prof. Céline Hudelot for your interest in my work, and for being the president of the juries. Of course, I should give my special appreciation to Prof. Samis Trevezas, not only for being a member of the juries but also for his suggestions and advice during the redaction of my thesis.

My following gratitude goes to members of my research team, including those who have left. Benoit, thank you for your solutions to every coding problem. You are precisely a "geek" as what I have imagined in my mind. Remi, Guillaume, and Laurent, my thesis cannot finish either without your help. You, mesocenter, did great things for the researchers. My deskmate, Antonin, Brice, and Mahmoud, it is a lovely experience to stay so close to you for the joke and discussion. I like the French humour, even if it comes from only two films, and I hope that something could also remind you about mine. Gautier, we have been in the laboratory for about the same time. You are the best baby foot player. But the Chinese wall is not easy to go through. Jean-Christophe, it was a pity that France did not win the Europe Cup, but you have got the world cup instead. Let's yell "Allez O.M." next time when I go back to Marseille. Thanks to Sarah, Véronique, Dany, and all the doctoral students: Jun, Mathilde, Chloé, Andreas, Sylvain, Gurvan, Adrien, Alexandre, Erwan. Nothing could be achieved without you being on my side.

Thank you to the secretariats for your help, suggestions, and advice: Annie, Sylvie, Emmanuelle, Fabienne, and Suzanne. Your administrative support gave me the first impression of France and

made my life so much easier! Thank you for all these questions that you have never hesitated to answer.

My final thanks go to my family, my parents, and my brother. My deepest gratitude owns to you, Lin. Thanks for your understanding and patience. Thank you for giving me such a lovely daughter. Nothing could fully show my greatness but to love you more.

I would like you to conclude by addressing those who are brave enough to fight against coronavirus all over the world. You are doing great things! Thanks a lot for your contribution.

Xiangtuo

Abstract

Crop yield prediction is a paramount issue in agriculture. Considerable research has been performed with this objective, relying on various methodologies. Generally, they can be classified into knowledge-driven approaches and data-driven approaches.

The knowledge-driven approaches are based on crop mechanistic modelling. They describe crop growth in interaction with their environment as dynamic systems. Since these models are based on the mechanical description of biophysical processes, they potentially imply a large number of state variables and parameters, whose estimation is not straightforward. In particular, the resulting parameter estimation problems are typically non-linear, leading to non-convex optimisation issues in multi-dimensional space. Moreover, data acquisition is very challenging and necessitates heavy specific experimental work to obtain the appropriate data for model identification.

On the other hand, the data-driven approaches for yield prediction necessitate data from a large number of environmental scenarios, but with data far from straightforward to obtain: climatic data and final yield. However, the perspectives of this type of model are mostly limited to prediction purposes.

An original contribution of this thesis consists in proposing a statistical methodology for the parameterisation of potentially complex mechanistic models, when datasets with different environmental scenarios and large-scale production records are available, named Multi-scenario Parameter Estimation Methodology (MuScPE). The main steps are the following:

- First, we take advantage of prior knowledge on the parameters to assign them relevant prior distributions and perform a global sensitivity analysis of the model parameters to screen the most important ones that will be estimated in priority;
- Then, we implement an efficient non-convex optimisation method, the parallel particle swarm optimisation, to search for the MAP (maximum a posterior) estimator of the parameters;
- Finally, we choose the best configuration by taking into account specific criteria, which account for the predictive capacity, the model complexity, computational limitation, and some other essential benchmarks in modelling.

This methodology is firstly tested with the CORNFLO model, a functional crop model for the corn.

A second contribution of the thesis is the comparison of this knowledge-driven method with classical data-driven methods. For this purpose, according to their different methodology in fitting the model complexity, we consider two classes of regression methods: the Statistical methods derived from generalised linear regression that are good at simplifying the model by dimensional reduction, such as Ridge and Lasso Regression, Principal Components Regression or Partial Least Squares Regression; the Machine Learning Regression based on re-sampling techniques like Random Forest, k-Nearest Neighbour, Artificial Neural Network and Support Vector Machine (SVM) regression.

At last, a weighted regression is applied to predict crops' production on large-scale. Significant economic crop production in France, soft wheat production, is taken as an example. Knowledge-driven and data-driven approaches have also been compared for their performance in achieving this goal, which could be recognised as the third contribution of this thesis.

Key words: Crop yield prediction, knowledge-driven approaches, data-driven approaches, sensitivity analysis, MuScPE, environmental diversity, large scale

Table of contents

List of figures xi

List of tables xv

Introduction 1

Context . . . 1

CYP Methodologies and the related data format . . . 2

Objectives of the thesis . . . 7

Organisation of the manuscript . . . 9

I Mechanic Crop Models for Crop Yield Prediction 11

1 Dynamic model of plant growth 13

1.1 General context of plant growth modelling . . . 14

1.2 Crop model of corn: CORNFLO . . . 15

1.2.1 Experimental data for corn crop . . . 16

1.2.2 First model evaluation . . . 16

1.3 Problems description . . . 16

1.3.1 Model complexity and uncertainty . . . 18

1.3.2 Optimization problem in biological engineering . . . 18

2 Plant Model Analysis 21

2.1 Sensitivity Analysis . . . 21

2.1.1 Basic Notations . . . 22

2.1.2 Sensitivity analysis process . . . 23

2.1.3 Sensitivity analysis method on platform . . . 23

2.1.4 Sensitivity analysis results . . . 24

2.2 Identifiability, Continuity and Convexity Analysis . . . 30

2.3 Conclusion . . . 31

3 MuScPE-PSO: MuScPE methodology with PSO optimisation **33**
3.1 Multi-Scenarios parameters estimating methodology 33
3.1.1 Mathematical interpretation of the problem 34
3.2 Basic Particle Swarm Optimisation 35
3.3 Variants of PSO 37
3.3.1 Partitioning particles 37
3.3.2 Constriction factor 38
3.3.3 Local PSO with Neighbourhood Topology 38
3.3.4 PSO and hybridisation 39
3.4 A well set PSO and evaluation with simulated dataset 39
3.5 Improved PSO with parallelism 40
3.5.1 Programming of massively parallel machines 41
3.5.2 Proposition of a parallelisation approach of the PSO method 42
3.5.3 Comparison of OpenMP and MPI 43
3.5.4 Results 43
3.6 Estimation results with MuScPE-PSO 46
3.6.1 Estimation with all observations 46
3.6.2 Boosting estimation with MuScPE 47
3.6.3 Fitness and Prediction evaluation 48
3.7 Conclusion 49

4 Study of Climatic Variability **51**
4.1 Basic descriptive statistics on Meteorological records 51
4.1.1 Spatial Distribution 51
4.2 Inter-regional variability analysis on Meteorological records 52
4.2.1 Inter-regional variability on rainy days 53
4.2.2 Inter-regional variability on rainy sequences 53
4.3 Inter-annual variability analysis 54
4.4 Unsupervised Clustering of Meteorological Records 56
4.4.1 K-means Algorithm 56
4.4.2 Clustering validation 57
4.4.3 Clustering result of meteorological records 58
4.5 Clustering-based Cross-validation 60
4.6 Conclusion 61

II Data-driven Models for Crop Yield Prediction **63**

5 Crop Yield Production with Statistical Learning Methods **65**

5.1 Context of data science . . . 65
5.1.1 A brief history of data analysis . . . 65
5.1.2 Basics of Learning . . . 66
5.1.3 Important notions in statistical learning . . . 68
5.1.4 Data Analysis Framework . . . 69
5.2 Methods Description . . . 70
5.3 Choice of method . . . 73
5.3.1 Criteria of penalised likelihood . . . 74
5.3.2 Empirical approach . . . 76
5.4 Results and conclusion . . . 79

6 Influence of Meteorological Variability on the Predictive Capacity 85
6.1 Reducing the inter-annual variability by regrouping the meteorological data . . . 86
6.2 Conclusion . . . 90

III Large-Scale Crop Production Prediction 91

7 Weighted regression for Large-Scale Production Prediction 93
7.1 Agriculture and Soft Wheat in France . . . 93
7.2 Data description . . . 95
7.3 Basic statistical analysis of soft wheat production in France. . . . 95
7.3.1 National level analysis . . . 96
7.3.2 Departmental level analysis . . . 97
7.4 Modelling the production of soft wheat . . . 98
7.4.1 Crop Production modelling . . . 98
7.4.2 Data preprocessing . . . 99
7.5 Weight regression with Crop model . . . 101
7.5.1 Sensitivity Analysis with Sobol indices . . . 101
7.5.2 Smoothness properties of the Loss function . . . 102
7.5.3 Prediction results for the French soft wheat production . . . 103
7.6 Weighted regression with the Statistical Learning model . . . 105
7.6.1 Prediction results with the initial dataset . . . 105
7.6.2 Inter-annual variability analysis . . . 106
7.6.3 Reducing the inter-annual variability by regrouping the meteorological data 107
7.7 Conclusion . . . 113

IV Conclusion **115**

Appendix A Synthèse en Français **123**

Appendix B CORNFLO: A crop model of corn **125**

B.1 Phenology module 125

B.2 Morphogenesis and Photosynthesis Module 126

B.3 Biomass Production and Biomass Distribution Module 127

B.4 Genotype parameters 128

Appendix C Clustering Result of 720 scenarios **131**

References **143**

List of figures

1 Number of phytomers along plant main axes with respect to the accumulated sum of temperature . . . 4
2 Experimental data versus simulation curve . . . 5
3 Philosophy of Multiple-Scenarios Parameter Estimation . . . 8

1.1 Real and simulated observations Distributions . . . 17
1.2 Relative error distributions . . . 17

2.1 Sensitivity analysis procedure . . . 23
2.2 A result of sensitivity analysis result with SRC method . . . 25
2.3 A result of sensitivity analysis with the Sobol method . . . 26
2.4 Parameter with smooth relation to the objective function . . . 30
2.5 Parameter with irregular relation (non-convex) to the objective function . . . 31

3.1 3-D presentation of the optimisation surface . . . 34
3.2 Movement of particle in the searching space . . . 35
3.3 Some of the neighbourhood topology used for the PSO technique: (a) Fully connected structure, (b) Ring structure, and (c) Von Neumann. . . . 39
3.4 The structure of parallelised PSO . . . 42
3.5 First comparison result between MPI and OPENMP . . . 44
3.6 Second comparison result between MPI and OPENMP . . . 45

4.1 Spatial Distribution . . . 52
4.2 Rainy sequences duration in the four counties selected during 2001-2007 . . . 54
4.3 Rainfall index in the four counties selected during 2001-2007 . . . 55
4.4 Elbow criteria in meteorological records clustering analysis . . . 58
4.5 Silhouette criteria in meteorological records clustering analysis . . . 59
4.6 Distribution of five parameters based on the clustering analysis with 50 samples from each cluster . . . 62

5.1 Illustration of supervised learning process . . . 68

5.2 An example of regression tree 71
5.3 An example of Support Vector Machine 73
5.4 Neural network structure 74
5.5 The bias-variance compromise 77
5.6 Root mean square error of fitness and prediction for different statistical learning approaches: Ridge, Lasso, PLS (partial least squares regression), PCA (principal component regression), DT (Decision Tree regression), BT (Boosting), BG (Bagging), RF (random forest), SVM (support vector machine), ANN (artificial Neural Network) 80
5.7 Mean absolute relative error of fitness and prediction for different statistical learning approaches: Ridge, Lasso, PLS (partial least squares regression), PCA (principal component regression), DT (Decision Tree regression), BT (Boosting), BG (Bagging), RF (random forest), SVM (support vector machine), ANN (artificial Neutral Network) 81
5.8 Predicted v.s. Observed values for different statistical learning approaches: Ridge, Lasso, PLS (partial least square regression), PCA (principal component regression), DT (Decision Tree regression), BT (Boosting), BG (Bagging), RF (random forest), SVM (support vector machine), ANN (artificial Neutral Network) 83
5.9 Residuals v.s. Predicted values for different statistical learning approaches:Ridge, Lasso, PLS(partial least square regression), PCA(principal component regression), DT(Decision Tree regression), BT(Boosting), BG(Bagging), RF(random forest), SVM(support vector machine), ANN(artificial Neutral Network) 84

6.1 Initial TMIN daily records in Cochise County 87
6.2 TMIN records regrouped by 5 days in Cochise County 87
6.3 TMIN records regrouped by 10 days in Cochise County 87
6.4 cluster-cross-validation-04 88
6.5 Evaluation of machine learning methods with regrouped meteorological records: DT (Decision tree), BT (Boosting), BG (Bagging), RF (Random Forest), SVM, ANN, KNN regression 89

7.1 A photo of wheat in field 94
7.2 Average wheat yield distribution all over the world 95
7.3 National Production of soft wheat from 1990 to 2010 96
7.4 National farmland of soft wheat from 1990 to 2010 96
7.5 Average yield of soft wheat from 1990 to 2010 97
7.6 Soft wheat yields distribution in France in 1999 (left) and 2009 (right) 97
7.7 The cultivated area for soft wheat in France in 2009 99

7.8 Geographical centres for each department in France 99
7.9 Estimation with simple interpolation in rectangle 100
7.10 Parameters' relation to the objective function 102
7.11 Production prediction results with 2 parameters calibrated 103
7.12 Production prediction results with 3 parameters calibrated 103
7.13 Production prediction results with 4 parameters calibrated 104
7.14 Production prediction results with 5 parameters calibrated 104
7.15 Application of weighted regression on data regrouped every 5 days 105
7.16 An example of instability caused by inter-annual variability 106
7.17 Initial TMIN daily records at point (5.35°E, 46.10°N) in Ain 107
7.18 TMIN records regrouped by 5 at point (5.35°E, 46.10°N) in Ain 108
7.19 TMIN records regrouped by 10 at point (5.35°E, 46.10°N) in Ain 108
7.20 Application of weighted regression on initial data 109
7.21 Application of weighted regression on data regrouped every 5 days 109
7.22 Application of weighted regression on data regrouped every 10 days 110
7.23 Application of weighted regression on data regrouped every 30 days 110
7.24 Application of a weighted regression on data regrouped every 80 days 111
7.25 Application of weighted regression on data regrouped every 280 days 111
7.26 Absolute relative error statistic of Random-Forest 112
7.27 Absolute relative error statistic of Ridge model 112
7.28 Sentinel images(left), the related form of NDVI which combines the VV and VH polarisations, acquired at 2016.06.11 in South Dakota (USA) 121

B.1 Maximum surface of the leaf $Ae(i)$ at each rank for three corn genotypes 126

List of tables

2.1 Variation intervals and the recommended value for the model parameters 25
2.2 Total first order Sobol indices for parameters in the CORNFLO model 29

3.1 Parameterisation results with an increasing number of parameters. 40
3.2 Best parameters setting of CORNFLO model that minimize the fitness with actual dataset. 46
3.3 Estimated values for the five parameters with different sample size 47
3.4 Associated standard deviation 47
3.5 Performance metrics of machine learning methods. 48
3.6 Fitness and prediction capacity evaluation 49

4.1 Recorded frequencies at state level 52
4.2 Rainy day frequency and cumulative precipitation during 2001-2007 53
4.3 Estimation of five parameters with different sample size based on the clustering analysis 60
4.4 Standard deviation of the five parameters with different sample sizes 60
4.5 Cluster-based Fitness and prediction capacity evaluation 60

5.1 Evaluation of statistical learning for CYP 79

6.1 Evaluation of goodness-of-fit and prediction with clustering-based cross-validation 86

7.1 Sobol Indexes of LNAS model 102
7.2 Absolute relative error 105
7.3 Statistics of the absolute relative error 106

B.1 The parameter values of *CORNFLO* model for genotype studies. 129

C.1 Scenarios in cluster 02 131
C.2 Scenarios in cluster 01 132
C.3 Scenarios in cluster 03 133

C.4 Scenarios in cluster 04 134
C.5 Scenarios in cluster 05 135
C.6 Scenarios in cluster 06 136
C.7 Scenarios in cluster 07 137
C.8 Scenarios in cluster 08 138
C.9 Scenarios in cluster 09 139
C.10 Scenarios in cluster 10 140
C.11 Scenarios in cluster 11 141

Introduction

Context

Agriculture is always one of the human beings' essential activities concerning soil cultivation and all other work on the natural environment. It is still a vital economic sector and today remains the leading sector of business in many countries. It provides a large part of our food consumed every day in our life. However, with climate changes and population increase, food production is faced with significant challenges. It is reported that the global population is rising from one billion in the early nineteenth century to more than seven billion today, which leads to a steady increase in food demand. A considerable rise in global food production is thus required each year in an increasingly severe climate (Change, 2014). The production of qualified food in sufficient quantity is essential for the well-being of people all over the world. Due to the limitation of arable land, yield improvement is a means of meeting a growing demand for agricultural commodities.

During the twentieth century, crop yields were improved thanks to ever more efficient farming techniques and the progress of the varietal selection. However, it is reported that the technical capacities and genetics no longer allow a significant increase in yields, which have stagnated since the 1990s (Brisson et al., 2010). In faced with these challenges, the accuracy of crop production prediction is a fundamental requirement of the managers or governments to formulate their agricultural policy to adopt strategies for food security (Change, 2014).

For a long time, Crop Yield Prediction (CYP) is a significant topic of interest in agriculture research, and farmers have been doing it roughly for hundreds of years. In modern times, CYP is requested to be carried out as early and accurately as possible. However, it is made very difficult by the variety of agricultural systems, the diversity of biophysical processes implied in plant growth, and the complexity of crop responses to stress (Liu et al., 2001). The non-linear behaviour of crops' reaction to the environment introduces large deviations from year to year and makes the traditional method inaccurate (Liu et al., 2001), (Drummond et al., 2003). What's more, farmers' management, such as land preparation, irrigation, sowing date, or fertiliser applications, also have a significant influence on crop yield. Sometimes, even the agricultural market can also have a substantial impact on the farmers' decision. Thus, more efficient methods should be developed in faced with these challenges.

The advances in mathematics and information technology lead to an explosive growth of new methods in plant crop yield modelling. According to their different modelling methodologies and different format of numerical observation, they could be classified into two categories as following (Tarsha-Kurdi et al., 2007):

- Knowledge-driven approaches: Mechanistic parametric models are created to describe crop's behaviour and growth according to environmental conditions
- Data-driven approaches: Empirical relations between the crop yield and the related environmental condition is constructed with statistical regression algorithms.

CYP modelling methodologies and data format

Knowledge-driven approaches with longitudinal data

For crop yield prediction, knowledge-driven approaches greatly depend on the construction of crop models. These models try to mimic the functioning of the plant system in interaction with climatic, soil, and other agricultural conditions. They make a useful abstraction of the dynamics of the plant's physiological development into mathematical equations (Safa et al., 2004). Since the 1970s, several families of models have been created, aiming at different objectives: to understand the eco-physiological functioning, such as SUCROS model (De Wit, 1978) or AFRCWHEAT (Porter, 1993) model; to analyze the implications of agricultural practices like CERES-Maize model (Jones et al., 1986) or CROPGRO model (Boote et al., 1998); to study the influence of environmental issues, such as EPIC model (Williams et al., 1984). At that time, most of them were crop-specific models.

Later in the 1990s, thanks to the general formalism, the agronomic and environmental objectives were integrated, which makes it possible to analyze different crops under the same formalism, such as SUCROS2 (Goudriaan and van Laar, 1994), Greenlab (Hu et al., 2003), STICS (Brisson et al., 2003) and APSIM (Keating et al., 2003). Sometimes, for a specific plant, there exist several different models for different research objectives. (Baey et al., 2014) has evaluated five typical models for the sugar beet crop: GreenLab, a generic functional-structural plant model dealing with the architecture description and physiological functioning (De Reffye and Hu, 2003); LNAS, a functional-structural plant model coping with the biomass allocation and the whole leaves compartment (Cournède et al., 2013); STICS, for the research of biomass production and intercepted radiation (Brisson et al., 2003); Pilote, a model for the study of crop–soil interaction (Khaledian et al., 2009) and CERES, in which the irrigation or nitrogen uptake can be integrated (Godwin and Jones, 1991). All these five models are proved to have a reasonable accuracy of CYP for sugar beet in (Baey et al., 2014), and behave very similarly despite their apparent differences.

In this thesis, for the crop models under study, the CORNFLO model for crop yield prediction in Part I and the LNAS-wheat model for large-scale production prediction in Part III, it is assumed that the dynamics of the involved biological processes can be commonly captured by a discrete dynamic system of the following form:

$$X_{t+1} = F_t(X_t, U_t, \theta), \tag{1}$$

where X_t and U_t represent respectively the state and the environmental variables of the system at time t, θ is the parameter vector, and the function F_t (model dependent) specifies the functional form of the dynamic system at time t. This function reflects the effects of the involved eco-physiological processes to the state variables of the system at time t, usually in the form of an empirical law. The dynamic system is assumed here to be deterministic. Nevertheless, it is also possible to make it stochastic by introducing modelling noises (Cournède et al., 2013). Ideally, as it is typically done in discrete dynamic systems, such crop models could be described in the form of a full longitudinal record of the objective variable at modelling time scale, in the following form:

$$\begin{aligned} \mathbf{X} &= (X_1, X_2, \ldots, X_N), \\ X_i &= (y_i, t_i), \quad i \in [1:N], \end{aligned} \tag{2}$$

where y_i represents the i-th possible record of the objective variable y at time t_i.

However, because of some constraints, not all the state variables can be directly observed, and not all possible records of the objective variable are available. Unfortunately, this is a typical situation in this application context leading to incomplete, asynchronous, and unbalanced (different lengths) data for different individuals (Newlands and Townley-Smith, 2010). In Figure 1, an example is presented with a dataset of 567 observation points to illustrate all the above characteristics. The dataset contains accumulated temperatures and associated numbers of phytomers corresponding to the first five months of development of 122 Acacia erioloba plants (Della Noce et al., 2016).

A significant amount of work is needed to develop and test these models in a real application context. The confrontation with the real dataset is the ultimate goal, and a successful model should satisfactorily reproduce the quantities of interest. A significant part of this process is related to a successful parameterisation and the development of efficient parameter estimation techniques. Several classical parameter estimation methods for dynamic crop models, such as Maximum Likelihood estimation and Ordinary Least Squares are listed in (Brun et al., 2006). Besides, the generalised least squares (GLS) is simple, powerful and particularly adapted for discrete dynamic models as described in (Goodwin and Payne, 1977), and has been widely used in Greenlab (Zhan et al., 2003), (Guo et al., 2006), (Ma et al., 2007) and Digiplant (Letort et al., 2008), (Christophe et al., 2008) and (Ma et al., 2010).

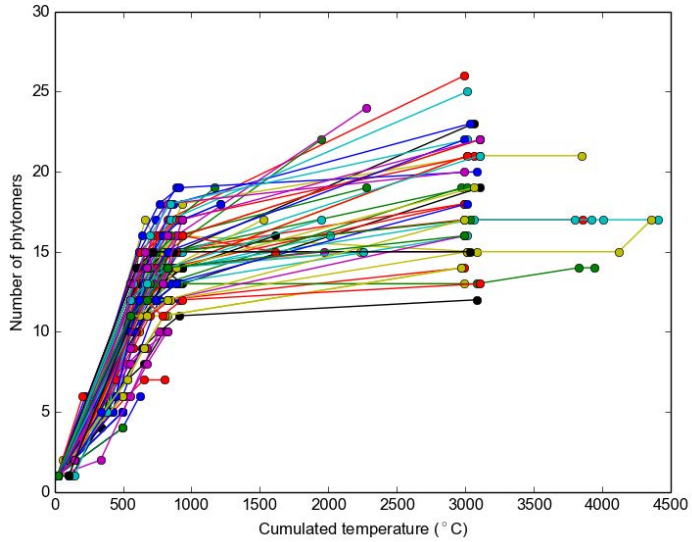


Fig. 1 Number of phytomers along plant main axes with respect to the accumulated sum of temperature

Let us now give a small description of the GLS model as it appears in our cases of interest. Let $(t_k)_{1\leq k\leq n}$ denote the sequence of times at which the crop was observed, and $y_k \in \mathbb{R}^p$ $(p \geq 1)$ the observation vector at time t_k. Since we assumed that no modelling noise is present, then it is typically assumed that observations are only subject to measurement errors. If the measurements were perfect, then the total observation vector $y = (y_1, y_2, \ldots, y_n)^t \in \mathbb{R}^{np}$ could be written in the form $y = f(\theta)$, where f succinctly represents the functional dependence of y on the unknown parameter vector θ, specific to the model under study. By adding a deviation ε from these perfect measurements, then the following representation holds:

$$y = f(\theta) + \varepsilon \tag{3}$$

where $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^t \in \mathbb{R}^{np}$, and represents the random deviation from the perfect model $f(\theta)$. In GLS it is assumed that $E(\varepsilon) = 0$ and $V(\varepsilon) = \Sigma$, where Σ is an unknown covariance matrix. It is also often assumed that $\varepsilon \sim \mathcal{N}(0, \Sigma)$. In any case, the GLS estimator is given by $\hat{\theta} = \arg\min(Y - f(\theta))^t \Sigma^{-1} (Y - f(\theta))$, that is by minimising the weighted norm (with respect to the precision matix) of the difference between the observed vector y and its expected one. An example of the application of this method for estimating biomass production as a function of thermal time is given in Figure 2 (Bayol, 2016). Notice that the estimated growth curve (red line) fits really well the experimental data (blue points). (Cournède et al., 2011) gives a detailed description of the estimation algorithm and propose ways to model the covariance matrix of the error vector ε.

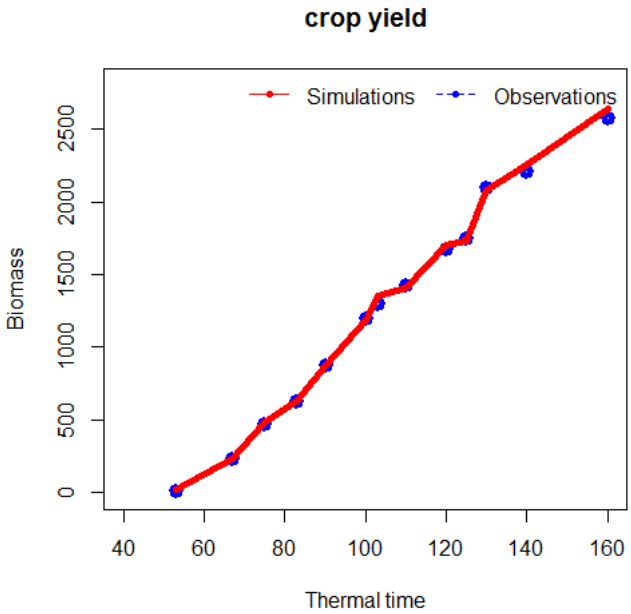


Fig. 2 Experimental data versus simulation curve

However, this methodology is considered to be impractical for massive application in agriculture for several reasons: firstly, this methodology is only efficient with micro-scale data obtained from the measurements along the crop's growth, for which the experiment is expensive in terms of time and money as stated in (Varcoe, 1990) and (Drummond et al., 2003), so that the sample size is usually quite small; secondly, the experiments are conducted in the same environment, which makes the universality of the calibrated model questionable; thirdly, it seems unnecessary to target the coincidence of the crop growth curve for the whole life if the only interesting state is the yield, the final state. It will be discussed in Part I.

Data-driven approaches cross-section data

Data-driven approaches have many synonyms, such as "statistical learning", "data mining", "machine learning" or "data science". It is a subject that deals with data changes with the explosion of data volume and data diversity. Their names updated with the advances in analytical and information technology. The basic idea is to figure out in the best possible way the relationship between the explanatory variables and the objective variable. This relation is always drawn from historical observations. The only difference is that the models and algorithms become more and more complex (O'Neil and Schutt, 2013). Today, they are used in almost all sectors of human activity and are part of the basic knowledge of the engineer, the manager, the economist, the biologist, the computer scientist. Innumerable applications are cited in the industrial field: reliability of equipment, quality control, analysis of measurement results and their planning, forecasting, and in

the field of economics and human sciences: econometric and agricultural models, surveys, opinion surveys, quantitative market studies (Saporta, 2006).

Common statistical practice includes data collection, processing, and interpretation (Montgomery and Runger, 2010).

Generally, the data collected on n individuals are usually presented in tabular form with n rows and the term individual is relative to the context. In general, a plant model is calibrated with micro-scale observations for each individual plant, but in our case, we want to use large-scale observations, which correspond to the harvest data for a certain region or the general crop yield for this region (county scale in USA, or province scale in France). When only numerical variables are observed, the array has the form of a matrix with n rows and p columns of general term $x_{i,j}$, and the related vector of the objective variable of of general term y_i:

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots & \ddots & \\ x_{i,1} & \cdots & x_{i,j} & \cdots & x_{i,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & & \cdots & x_{n,p} \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \tag{4}$$

As for CYP, the main idea is to build empirically the relationship between the crop yield and the living conditions in which the crop is cultivated by taking into account historical observations. In most cases, it doesn't require a deep knowledge of biological mechanisms that produced the plant. Such techniques are inexpensive, relatively easy to apply, and do not need a predefined structure of the model (Lobell and Burke, 2010). Consequently, data-driven approaches have been widely applied in recent years with classical statistical methods (Dixon et al., 1994), (Sudduth et al., 1996) and machine learning methods (Drummond et al., 2003), (Roel and Plant, 2004), (Irmak et al., 2006).

In this research, the dataset, on which the statistical learning approaches will be applied, is in the form $\{U_i, y_i\}$, with U_i the meteorological records and y_i the crop yield at harvest. As introduced above, the available dataset for this research is in the form $\{U_i, y_i\}$. The objective of statistical learning approaches is to build a regression model of the form:

$$y = g(U) + \varepsilon, \tag{5}$$

where $g(\cdot)$ represents the complex relationship between the yield and its relative meteorological condition. In most cases, such a model is considered to be a "black box", and it will be trained with available datasets through a learning process before coming into use. As stated in (Von Storch, 1999), the meteorological records always introduce a high dimension and a strong correlation. When dealing with these difficulties, solutions can be divided into two parts according to their modelling

discipline: the statistical methods by dimension reduction and penalisation, like Ridge and Lasso regressions, principal component regression and partial least squares regression; the machine learning methods depending on similarity and re-sampling technique, such as regression trees, random forest, k-nearest neighbours (KNN), Artificial neural network (ANN), SVM regression, etc. More details will be discussed in Part II.

Objectives of the thesis

As described in the previous Section, eco-physiological crop models have been widely used to analyse genotype-by-environment interactions. But the estimation of their parameters remains a crucial issue. In parameter estimation, an estimation process is carried out by taking into account the measured data as input and evaluating the uncertainty on the parameter estimation. In the process of parameterisation of crop models, an important issue is related to the generally increased number of parameters compared to a few field data (Makowski et al., 2006). On the other hand, to allow precise discrimination between genotypes, model calibration should be accurate enough, which necessitates the availability of a sufficient number of experimental data (Reymond et al., 2003), generally difficult to obtain. Such a contradiction in breeding programs seems to be unsolvable.

The first objective for this thesis is to propose a solution that could meet the demand for cost reduction and precise calibration for crop models at the same time. While the traditional longitudinal data records used to calibrate dynamic systems are costly, another large-scale production data, which could be regarded as a special case of longitudinal data with only one observation, the final harvest, are available and exist in a large amount in the government's dataset. As suggested by (Jeuffroy et al., 2006), it would be beneficial if a methodology could be devised to take advantage of these official data (that are classically available at a reduced cost) for the parameterisation of crop models. Normally, a well-chosen panel of environmental conditions in which a few plant traits are measured should mathematically provide enough information for model identification. It is a matter of investigation if the poor information of one single record for each longitudinal data could be hedged by the rich knowledge of diverse environments where the crop grows. Under such circumstances, a novel methodology, called multi-scenario parameter estimation methodology (MuScPE), is designed to verify this hypothesis. The main idea of this methodology is to estimate parameters by taking into account large scenarios' simple data, instead of using an extensive collection of detailed plant growth data, as shown in Figure 3.

In particular, MuScPE will be firstly tested on CORNFLO, a crop model of maize. Virtual experiments, comprising real weather data and model simulation-generated virtual yield data, are used to prove MuScPE's technical feasibility. MuScPE's implementation with real field experiments data encounters much higher difficulty than virtual experiments. Practical issues for successfully carrying out MuScPE on real data are required to be discussed. The first issue is the determination

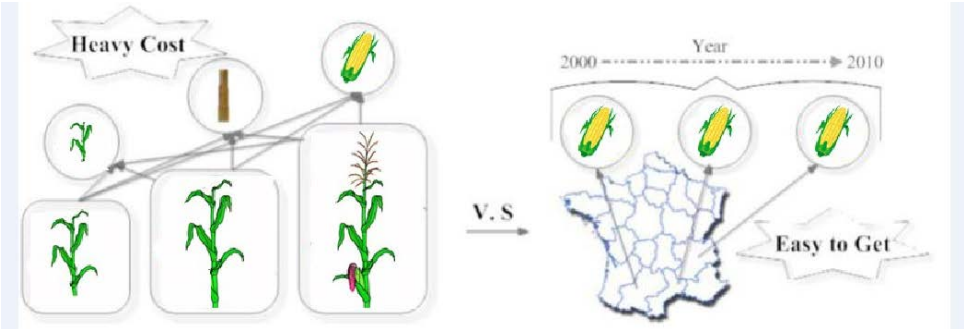


Fig. 3 Philosophy of Multiple-Scenarios Parameter Estimation

of parameters' estimation priorities with sensitivity analysis. The second includes ensuring the most appropriate numerical optimisation methods for the model. We will mainly focus on Particle Swarm Optimisation and its variants, to figure out the best algorithm to deal with the corresponding optimisation problems. The use of the high-performance computing machine "Mesocentre" of Paris Saclay University helps to enhance the efficiency of the computation for this purpose. Two assumptions about MuScPE's estimation and prediction performance are tested: "the increase of scenarios decrease the variance in the distribution of the estimated parameters" and "the increase of scenarios makes estimated parameters possess better prediction ability". Hence, the link between scenario amount and the MuScPE estimate's accuracy and precision will be investigated.

As for data-driven approaches, since they are good at dealing with the regression problem with cross-sectional data, various algorithms have been proposed and applied to solve the agricultural issues. However, less work has been carried out to compare the efficiency of different approaches under different assumptions. Thus, the second objective for this thesis is to compare different statistical learning approaches for the crop yield prediction and find out the best algorithm to explain the relationship between the crop yield and its environments.

When dealing with models that take environmental data as inputs, the diversity of ecological variables is an unavoidable issue. Furthermore, the idea of MuScPE is based on the hypothesis that the loose information caused by the few records in longitudinal data could be hedged by the diversity of scenarios information. This methodology aims at finding out the best configuration of the parameter for a generic environment, which requires that the diversity of backgrounds should be kept as much as possible. Therefore, the third objective is to improve the knowledge-driven and data-driven approaches by adopting a clustering analysis. In particular, non-supervised clustering analysis is firstly conducted to divide the dataset into different subgroups according to their variety. Then, samples from different subsets should be equally drawn to ensure that varied environmental information is taken into consideration. The influence of the meteorological variety will also be studied for both methodologies.

According to the definition of "crop yield", which is "the measure of grains or seeds generated from a unit of land expressed as kilograms per hectare", it remains an academic or a scientific

subject. But, if we talk about "crop production", it becomes more realistic. Finally, an ambitious objective is to build a stable predictive framework for large-scale crop production. In the future, it is supposed that, by taking into consideration more environmental and economic constraints, large-scale crop management could also be made with this platform.

Organisation of the manuscript

The first part of the thesis (Chapter 1 - 4) deals with crop yield prediction via plant growth modelling.

Chapter 1 clarifies the context of dynamic modelling in agronomy and makes a brief introduction to plant growth modelling for the corn crop. After the first evaluation of the model, the recommended parameter settings provided by the modeller seems to be incorrect. The results lead to a discussion about the difficulties in dealing with the parameterisation of a dynamic model.

Chapter 2 attempts to deal with the problem proposed in Section 1.3 about model complexity. A sensitivity analysis based on SRC and Sobol indices is carried out to study the interaction of parameters along the life cycle of corn. It also helps to rank the parameters according to their importance for parameterisation. A brief study of continuity, convexity, and identifiability is also performed to evaluate the difficulties in the parameterisation process. Finally, a subset of 5 identifiable parameters is chosen to be estimated at the first stage, by ensuring that the objective function is continuous. The absence of convexity is also addressed.

In Chapter 3, a methodology named "multi-scenarios parameter estimation", is introduced for the plant growth model with an available dataset of the form $\{U_i, y_i\}$, where U_i are the meteorological records of a specific environment and y_i its corresponding yield. Since the parameterisation of the plant model turns out to be a single-objective non-convex optimisation problem, the Particle Swarm Optimisation(PSO) algorithm will be integrated into the **MuScPE** methodology. A detailed introduction of the PSO and its variants is given in Section 3.2 and 3.3. The first essay of **MuScPE-PSO** will be accomplished to verify its capacity in global optimisation in the plant growth model. Since PSO is a population-based algorithm, which is costly in computation, a parallel PSO combined with **MPI** and **OpenMP** will be presented in Section 3.5. A parallelised PSO with proper settings is finally chosen to be a robust optimisation algorithm for the **MuScPE** methodology. In Section 3.6, the parametrisation results with **MuScPE-PSO** will be presented. The results can be divided into two parts: the estimation result with simulated data and that with real data. In the simulated case, since we know precisely well the settings used to generate the simulated data. The results in this part can be used to test some properties of **MuScPE-PSO**, like the stop criteria of the PSO algorithm, the uncertainty of the estimated parameters relative to the numbers of scenarios, The stop criteria condition that we get from the last part will be used in parametrisation with real data. The CORNFLO model will be calibrated, and its prediction capacity will be evaluated.

In Chapter 4, the study of knowledge-driven methods is completed with another important subject concerning models that take into account meteorological records, the inter-annual variability. Descriptive statistics on the evolution of meteorological variables, the rainfall, for example, will be discussed in Section 4.2 and 4.3. In Section 4.4, a non-supervised clustering method, the k-means clustering algorithm, will be applied to the meteorological records. According to the clustering results, almost the records of the same year are classified into the same group, which means that the inter-annual variance of the meteorological records is more important than the inter-regions variance. Finally, cross-validation that takes into account the inter-annual variance will be implemented to study its influence on the CYP prediction.

From Chapter 5, the topic will change to the study of a data-driven approach for the CYP. First of all, a simple introduction to data-driven methods will be made in this Chapter. The first difficulty in dealing with meteorological data is the collinearity between the adjacent daily records of the meteorological variable and its high dimension. The solutions to deal with these two difficulties can be divided into statistical regression methods and machine learning regression methods. The former depends on the dimension reduction while the latter takes advantage of re-sampling technique and similarity. Some conventional techniques will be introduced, including the statistical methods like Ridge and Lasso regressions, principal component regression and partial least squares regression, and the machine learning methods like regression trees, random forest, k-nearest neighbours (KNN), Artificial neural network (ANN), SVM regression.

In Chapter 6, the methods presented in Chapter 5 will be tested in terms of their yield prediction capacity. The inter-annual variance, previously discussed in Chapter 4 is proved to have a significant influence on the predictive ability of data-driven methods. A straightforward but effective solution is to regroup the meteorological records. In particular, for specific days, the variables are averaged out to produce the new explanatory variable. It is important to consider different periods since the sensitivity of the performance to the weather varies over time.

In Chapter 7, a case study is presented for the prediction of large-scale crop production. The national French soft wheat production is taken as an example. The dataset consists of the crop harvest and cultivated surface at the departmental level and the related environmental information from 1990 to 2010. A crop model, LNAS-wheat, is compared with other data-driven approaches under the weighted regression framework. Moreover, the results with an average relative and absolute error less than 5% prove the accuracy achieved by Random-Forest.

Finally, for the conclusion part, a discussion of the proposed methodology developed in this thesis and a discussion of the primary results will be carried out. Some perspectives for future work are also presented.

Part I

Mechanic Crop Models for Crop Yield Prediction

Chapter 1

Dynamic model of plant growth

In the history of humanity, we human beings face different difficulties from nature or our creations, such as natural disasters, diseases, food shortage, pollution, and even war. A vital member of our planet, the plants, on the contrary, plays an essential role in the balance. For resources, such as oxygen, energy, the emergence of agro-fuels, the basis of our diets, even the origin of many medicines, we depend entirely on the plant and the eco-system for our survival. Also, a global context of improvement in living conditions but also global climate change, considerable effort should be made to reposition the scientific research to satisfy quality requirements, environmental concerns, and "low input" specifications at the same time. In the field of ecological modelling, this proposes a new evaluation principle for plant growth models, that includes not only their excellent average performance and their stability in different scenarios but also the diagnosis and the control of the genotype $\times$ environment interactions ($G \times E$). In other words, this new stance requires and incorporates better understanding and predictions of the $G \times E$ interactions.

Dynamic simulation models have been developed since the 1980s to account for the response of a crop to extreme environmental conditions (temperature, radiation, water.) (Hammer et al., 2002). For most of the plant growth models, the genetic variability is reflected by different settings for the same parameters (Boote et al., 2001), and direct or inverse measurements, (Jeuffroy et al., 2006) could estimate these values. Several recent examples show that dynamic models can be successfully applied to understand and predict the $G \times E$ interactions. (Agüera et al., 1997), (Brun et al., 2006). These models can also be used to inform some environmental covariates used in the $G \times E$ interactions. It is supposed that better control of $G \times E$ interactions is likely to increase the competitiveness of the crop by reducing the gap between the yield collected by farmers and that allowed by the environment (Champolivier et al., 2011).

This chapter begins with some prerequisites and a general description of plant growth models in Section 1.1; then, the CORNFLO model for the corn crop is presented, and a first predictive evaluation is given in Section 1.2; the chapter ends with a discussion of some difficulties encountered in a dynamic crop model in Section 1.3.

1.1 General context of plant growth modelling

Modelling is one of the most important scientific research activities in modern science. The modelling of environmental, ecological, economic, agronomic, physical, and chemical systems has developed considerably in recent decades as a result of advances in information technology and the development of powerful computational tools. Modelling consists of integrating the knowledge acquired through experimentation, experience, and theory into mathematical equations and computation codes. It provides a more accessible and efficient way to understand the phenomenon under study as well as the simulation and prediction of the impacts caused by some extreme environmental conditions to the crops. Usually, models are developed for a specific purpose, mostly for supporting applied research and providing decision-making tools for economic or political decision-makers.

Many models are developed at appropriate scales to have a better representation of the phenomenon under study. Parameter determination is a critical aspect of the modelling practice, and many times, their interpretation may be complicated and subtle. The difficulty increases when multiple interactions, typical in biological applications, for example, are present in a dynamically evolving environment. In particular, in agronomy, models are very often developed at daily time steps to simulate the effects of agricultural practices on crops (quality and yields), on the environment (pollution and emission of greenhouse gases). Some of these models are used to guide farmers in their agricultural practices and policymakers in management and regulation (Brisson et al., 1998), (Meynard et al., 2002).

The modern plant growth modelling can date back to the early 1970s. Since then, with the explosion of information technology, this domain has advanced in multiple directions, which can broadly be classified as geometric models and the agronomic model (De Reffye et al., 2008). The recent focus on the visualisation of plant growth in space leverages numerical simulation and computer vision, while the latter is designed for the internal biological processes and their interaction with the environment.

The agronomic models are also called "process-based" models in the literature since their objective is to study processes such as the interception of solar radiation, the change of absorbed solar radiation into biomass, and biomass allocation. The production of biomass is then obtained with the help of a system of equations involving the biological processes of photosynthesis, respiration, and distribution. It has been shown that agronomic models allow reasonable estimation of crop yields as in (Oteng-Darko et al., 2013).

A dynamic system formulation has been recently proposed to model the plant growth process (Cournède et al., 2013). It is supposed to be an advantageous framework because, under this formulation, the mathematical equations that describe the biological processes could be easily translated into the programming language as the simulation tools require. In the case of plant growth modelling, a dynamic model can be expressed as follows:

$$y(t) = f(U, \theta, t), \qquad t \in \{1, 2, ..., T\} \tag{1.1}$$

where U is the vector of the input variables of the model containing the environmental information, θ is the vector of uncertain parameters with genotypic information, $y(t)$ is the output of the model at day t and T corresponds to the total observation time. The function $f(\cdot)$ is model specific and reflects the way that all the previous quantities are linked. It could be either deterministic or stochastic, depending on the assumptions. Since U contains the environmental information and θ incorporates the genetic information, the above equation illustrates the way that the output $y(t)$ results from the $G \times E$ interaction.

1.2 Crop model of corn: CORNFLO

Several simulation models of corn have been proposed in the literature, such as STICS in (Brisson et al., 2003). Some of them are even specific to maize, like CERES-Maize in (Fang et al., 2011). However, the parameterisation of these models is not directly linked with the relation between the phenotype and the genotype. The absence of a model which directly assesses this type of variability gave the motivation for the CORNFLO model to be developed, which is a new simulation model that meets the requirements of varietal evaluation of the $G \times E$ interaction. In the sequel, we describe its essential characteristics, and for a more detailed description of the model, the reader is referred to the Appendix B.

CORNFLO is a functional plant growth model for the corn crop (Zea mays L.) (Kang, 2013). This model has been coded in C++ in the modelling platform Pygmalion (Cournède et al., 2013). It simulates the daily progress of rooting, the development of the leaf surface, and the above-ground biomass of corn according to the constraints of temperature, radiation, and water. Biomass production is a function of the energy intercepted by the canopy. This model is more or less a "structural" model because several notions related to the structure of the plant, such as "leaf placement", "leaf surface distribution" and "leaf senescence" are introduced instead of a "big-leaf". Environmental constraints interact and therefore influence the potential production allowed by radiation and temperature.

The life cycle of corn is separated into 4-phases according to the accumulated thermal time ($°C \cdot day$): (i) the flowering bud appearance stage (TTE1); (ii) the early flowering stage (TTF1); (iii) early seed filling stage (TTM0); (iv) the physiological maturity stage (TTM3). Each stage change induces differentiated physiological processes.

Water is evaluated daily, and stress indices are calculated to reflect the multiplicative effect of the constraints on leaf expansion and biomass accumulation. Yield is estimated using a harvest index (HIgraine) that applies to the total dry biomass produced at physiological maturity. The

daily climate used for the simulation includes five standard variables: maximum and minimum temperatures, precipitation, potential evapo-transpiration, and global radiation.

1.2.1 Experimental data for corn crop

The experimental data are essential for the modelling process, especially for plant modelling. Normally, before building a conceptual model, the first step is to analyze the data. The experimental data should be used to calibrate the model to enhance its prediction ability. For the CORNFLO model, two kinds of data are taken into consideration: the environmental data and the plant data. The environmental ones consist of meteorological and plant management data. Among the meteorological data, usually, we use the daily maximum and minimum temperatures, precipitation, relative humidity, and solar radiation. The crop management data include the date of sowing, irrigation, plant density, and harvest date. The plant data consist of measurements related to leaf area and seed biomass.

In this research, the experimental database consists of 720 experimental 100% irrigated scenarios in the form $\{U_i, y_i\}$. Each situation consists of the environmental data U_i, and the crop yield y_i. The environmental ones include the daily meteorological records and crop management data such as density, sowing date, and the date of harvest. Among them, daily meteorological data between 2001-2010 are obtained from the primary database Syngenta Corporation. The yield data are given by the National Agricultural Statistical Service (NASS) of the United States Department of Agriculture (USDA).

1.2.2 First model evaluation

In this section, we assume that the parameter θ_0 given in Table B.1 corresponds to the "true" setting and will be used to simulate the maize yields in 720 different scenarios. Then, θ_0 will be evaluated by the comparison of the simulated corn yield with the real observations. In Figure 1.1 the comparison of the two histograms corresponding to the actual (in red) and the simulated (in blue) observations from the recommended θ_0 in Table B.1 indicates the inappropriateness of the recommended values since the real observations are over-estimated. A more detailed analysis of the distribution of the relative errors (see Figure 1.2) reveals that in most cases, the error exceeds 50% and, in some cases, even 150%.

Better parameter calibration is needed for the model to demonstrate its full capacity for CYP.

1.3 Problems description

The identification of model parameters on the real plant is an important issue since these parameters are components of yield that can subsequently be used in genetics, or in optimising crop management. However, parameterisation remains challenging due to several reasons.

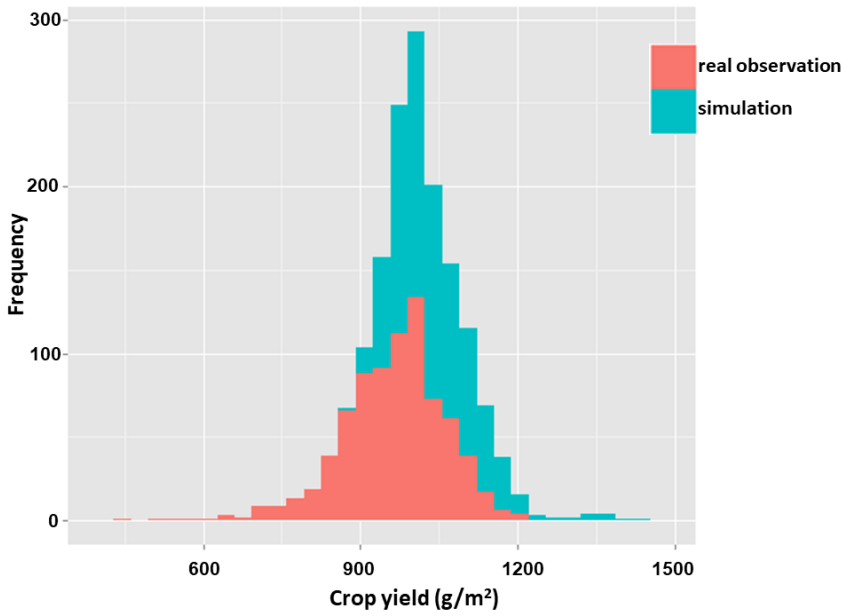


Fig. 1.1 Real and simulated observations Distributions

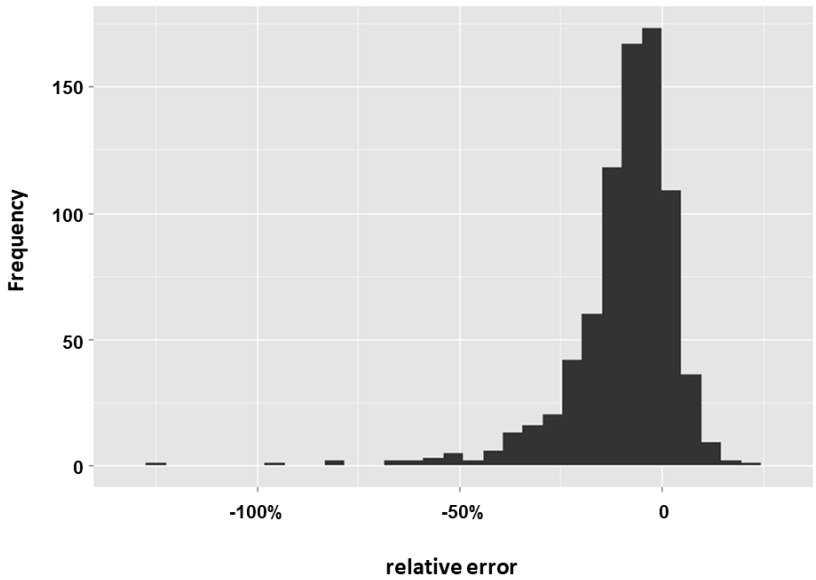


Fig. 1.2 Relative error distributions

1.3.1 Model complexity and uncertainty

The increasing role of dynamic models and their complexity make it essential to measure and incorporate different sources of uncertainty of these models. For the sake of a better description of the observed phenomenon, deterministic dynamic models generally introduce many parameters, which constitute one of the primary sources of uncertainty in model outputs. The estimation of model parameters is a crucial step in the modelling process. Generally, the model's performance is largely dependent on the accuracy of the estimates (Lehuger et al., 2009). The estimation of model parameters for dynamic models (mostly non-linear) is a problem both in theory and in practice for various reasons, such as a lack of observations to estimate all parameters, the presence of non-identifiable parameters (Brun et al., 2001), or the structure of the model (essentially the non-regularity) (Bechini et al., 2006).

In general, a large number of uncertain parameters in a model is unfortunately accompanied by few observations for reasons of cost and difficulty of measurement. Parameter estimates with fewer observations are less accurate and increase the risk of prediction errors when using this model. A natural approach would stem from Bayesian Statistics. A good demonstration of Bayesian Statistics in plant growth models can be found in (Chen, 2014). Although this field is evolving rapidly, it is known that Bayesian estimation in the presence of a small number of observations may provide less accurate posterior distributions (Lehuger et al., 2009). Sensitivity analysis in Section 2.1 offers a way around this problem in the case of a scalar output model (Makowski et al., 2006), (Brun et al., 2006). Sensitivity analysis makes it possible to identify the most critical parameters using virtual experimentation of the model without using observations. The use of sensitivity analysis to choose the parameters to estimate under generalised least squares criteria is presented in (Cournède et al., 2013).

1.3.2 Optimization problem in biological engineering

Solving optimisation problems has become a central topic in biological research since a biological engineering problem is typically formalised in the form of an optimisation problem. It includes problems such as learning neural networks, task planning or identification.

According to the number of objective functions, the optimisation problems can be grouped into multi-objective problems and single-objective problems. In this thesis, the only objective is to make an accurate prediction of crop yield. Hence, It is considered as a single-objective optimisation problem.

Single-objective optimisation

An optimisation problem, in general, is defined by a search space Θ and an objective function f. The goal is to find the best quality solution in Θ. According to the problem, one seeks either the

minimum or the maximum of the function f. In the remainder of this document, we only deal with minimisation, that is, we search for

$$\tilde{\theta} = \arg\min_{\theta \in \Theta} f(\theta), \tag{1.2}$$

since the maximisation of a function f is exactly equivalent to the minimisation of $-f$.

Moreover, an optimisation problem may have equality and inequality constraints on the candidate solutions in Θ. In this thesis, we will only study single-objective issues. Many deterministic methods can solve certain types of optimisation problems quickly. However, these methods require that the objective function has a certain number of characteristics, such as convexity, continuity or differentiability. Among the best-known techniques, there are linear programming methods, quadratic or dynamic, Newton's method, the "simplex" method, or the gradient-based method (Wright and Nocedal, 1999).

Some problems, however, remain too complicated to be solved with deterministic methods. Some features may cause a discrepancy in these methods. Among these, one can mention discontinuity, non-differentiability or the presence of noise.

Meta-heuristics for difficult mono-objective optimisation

Meta-heuristics are a family of stochastic algorithms for solving optimisation problems. Their particularity lies in the fact that they are adaptable to a large number of problems without major changes in their algorithms, hence the meta qualifier. Their ability to optimise a problem from a minimal amount of information is counterbalanced by the fact that they offer no guarantee as to the optimality of the best solution found. Only an approximation of the global optimum is obtained. However, for operational research purposes, this attribute is not necessarily a disadvantage, since one can always prefer a good approximation of the global optimum found swiftly than an exact value found at an inefficient and long time.

Meta-heuristics are methods that have, in general, an iterative behaviour, that is to say, that the same pattern is reproduced repeatedly during the optimisation, and that is "direct", in the sense that they do not use the calculation of the gradient of the function. These methods are useful. After all, they are less easily trapped in local optima, because they accept, during the treatment, impairments of the objective function and the research is often conducted by a population of points and not a single location.

To overcome the difficulty during the modelling process, we are particularly interested in one of the effective Meta-heuristics, the Particle Swarm Optimisation (PSO), which constitutes the primary optimisation method for the estimation of the plant growth model in this thesis. It is a new class of meta-heuristics proposed in 1995 by Kennedy and Eberhart (Eberhart and Kennedy, 1995). The social behaviour of swarming animals inspires this algorithm. The various particles of a swarm

communicate directly with their neighbours and thus construct a solution to a problem, based on their collective experience. This detailed description and some tests of important priority will be discussed in Section 3.2.

Chapter 2

Plant Model Analysis

After the first evaluation of the prediction efficiency of the plant growth model, the CORNFLO model, with recommended parameters as in Table B.1, there is still a big difference between predictions and real observations. We must have recourse to mathematical and statistical tools to have a better approximation to the observed phenomenon. Much progress has been made in this area. In (Jeuffroy et al., 2006), several axes of plant model research have been proposed, including sensitivity analysis, parameter estimation and model evaluation.

Ideally, all the parameters of the CORNFLO model should be estimated. However, in reality, it is impractical for several reasons. Firstly, plant models like CORNFLO have already been studied extensively in other contexts. Consequently, it is an excellent choice and highly advisable to take some of the consistent results regarding parameter estimates into account. Secondly, some parameters are genotype-independent, and thus their values could be estimated by conducting preliminary studies. This practice could result in a significant reduction in the number of parameters that have to be determined. Since in this research project, we only focus on genotype-dependent parameters, we can indeed take advantage of this principle. Thirdly, the objective function of the fitness is not always a convex function of the parameters, which makes somewhat risky the use of a descent-based method (like Gauss-Newton), since it could increase the cases where convergence takes place to a local minimum, which is the reason why, in this chapter, the analysis of parameters is first carried out by conducting a sensitivity analysis in Section 2.1, and then by proceeding to an identifiability, continuity and convexity analysis in Section 2.2. Finally, only a subset of the initial parameters, the ones which are identifiable and the most important, will be selected for calibration in the parameter estimation step. This step will be described in the next chapter.

2.1 Sensitivity Analysis

Sensitivity analysis consists of a mathematical tool which determines how uncertainty in the output of a system can be attributed to the uncertainty in its inputs (Saltelli, 2002). By estimating sensitivity

indices that quantify the influence of inputs on the output, it can be beneficial for many engineering applications:

- Testing the robustness of a mathematical model with uncertainty to have a better understanding of the relationships between input and output variables in a system (Iooss and Lemaître, 2015);
- Simplifying the model by setting to constant values the inputs that have less effect on the output and simplifying the model calibration process by adjusting parameters that have the most important sensitivity indices. It could be very beneficial since a lot of time and effort might be wasted to adjust parameters which are not very sensitive, (Bahremand and De Smedt, 2008).
- Leading to a better understanding of the parameters' interaction with observations, model inputs and predictions (Hill and Tiedeman, 2006).

2.1.1 Basic Notations

The goal of sensitivity analysis is to explain how the uncertainty in the output Y is attributed to different sources of uncertainty in its inputs $X = (X_1, ..., X_p)$. The output and the inputs are linked through a model with a generic mathematical formulation of the form

$$Y = f(X), \tag{2.1}$$

where f is considered here to be a black box.

There are many constraints imposed by the function f that should be taken into consideration, mainly as follows:

- Computational cost: In most cases, performing a sensitivity analysis requires evaluating the model f a large number of times (Helton et al., 2006). It becomes even worse when the model has a large number of inputs.
- Correlated inputs: the independence assumption among model's inputs greatly simplifies sensitivity analysis, but sometimes it can not be overlooked that inputs are highly correlated. The correlations between the inputs must then be taken into consideration in the analysis (Sainte-Marie et al., 2017).
- Non-linearity: the linearity or not of the modelling function f concerning the inputs will be decisive in the choice of different techniques to deal with sensitivity. The sensitivity indices based on standardised regression coefficients (SRC) are usually applied to a linear model while in the case of a non-linear model, a variance decomposition method is taken advantage of (Jacques, 2005).
- Interactions: the interaction among the inputs will complexify the sensitivity analysis. The simple and total Sobol indices are effective tools to quantify the interaction effect (Nossent et al., 2011).

2.1.2 Sensitivity analysis process

Many approaches for sensitivity analysis have been developed to deal with one or more of the constraints discussed above. However, most procedures are carried out with the 4 steps as shown in Figure 2.1

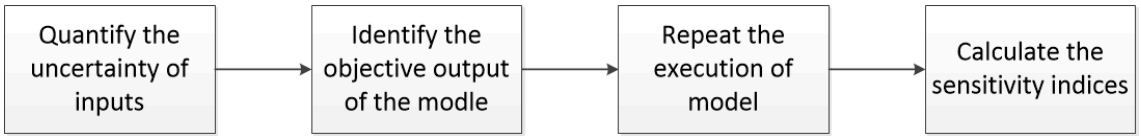


Fig. 2.1 Sensitivity analysis procedure

1. Quantify the uncertainty for each input, for example, by determining the range of possible values or/and by assigning relative probability distributions.
2. Identify the outputs of the model that will be analysed.
3. Run the model some times following a plan of experiments.
4. Compute the sensitivity measurement chosen for the problem.

2.1.3 Sensitivity analysis method on platform

On the Platform PyGMAlion (Cournède et al., 2013), two types of sensitivity measures have been implemented: the SRC indices and the Sobol indices.

SRC: the index based on regression analysis

In the context of sensitivity analysis, the SRC method analysis uses standardised regression coefficients as sensitivity index (Hamby, 1994). This method works best in the case that the response variable (the output) is indeed linear to the inputs. The coefficient of determination could be used as an indicator, but it should be used with caution since this index corresponds to a measure of global model fitting. Besides, the SRC is well known for its simplicity and low cost in computations. Despite its simplicity, in ecological models, the model's non-linearity and the strong interactions among the parameters cause many difficulties in the use of the SRC index (Cariboni et al., 2007).

Sobol method: the indices based on variance decomposition

The Sobol index is one of the sensitivity analysis methods based on variance decomposition, where the model inputs are considered as random variables, and the variance of the output is focused (Sobol, 1993). The Sobol index represents the proportion of variance explained by an input or a group of inputs.

In the crops models framework, (Wu, 2012) proposes an improvement of the Homma-Saltelli method for the calculation of the Sobol sensitivity indices (Saltelli et al., 2004). The authors manage to calculate higher-order sensitivity indices without additional simulations. These methods have been implemented on the PyGMAlion platform, and they are easy to call when analysing the plant growth model.

Let $\mathbb{E}(\cdot)$ and $\mathbb{V}(\cdot)$ stand for the expectation and the variance operator. Since we also refer to conditional expectations and variances standard simplified notations will be used, and the context will understand the underlying measures.

The Sobol method is based on the following decomposition of the variance:

$$\mathbb{V}(Y) = \sum_{i=1}^{p} V_i + \sum_{1 \le i < j \le p} V_{ij} + \ldots + V_{1\ldots p}, \tag{2.2}$$

where

$$\begin{aligned} V_i &= \mathbb{V}(\,\mathbb{E}[Y|X_i]\,) \\ V_{ij} &= \mathbb{V}(\,\mathbb{E}[Y|X_i,X_j]\,) - V_i - V_j \\ &\vdots \\ V_{1\ldots p} &= \mathbb{V}(Y) - \sum_{i=1}^{p} V_i - \sum_{1 \le i < j \le p} V_{ij} - \ldots - \sum_{1 \le i_1 < \cdots < i_{p-1} \le p} V_{i_1 \ldots i_{p-1}}. \end{aligned} \tag{2.3}$$

The Sobol index of first order for the input X_i is defined as:

$$S_i = \frac{\mathbb{V}(\,\mathbb{E}[Y|X_i]\,)}{\mathbb{V}(Y)} = \frac{V_i}{\mathbb{V}(Y)}. \tag{2.4}$$

The first order Sobol index S_i does not take into account the uncertainty caused by the interactions of X_i with the other variables. To include all the interactions in which X_i is involved, we use the total Sobol index:

$$S_i^T = 1 - \frac{\mathbb{V}(\mathbb{E}[Y|X_{\sim i}])}{\mathbb{V}(Y)}, \tag{2.5}$$

where $X_{\sim i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p)$. Variance decomposition-based methods allow a full explanation of the input space. It makes it possible to analyse the parameters' interactions and model's non-linearity. For these reasons, Sobol indices are widely used in solving complex engineering problems.

2.1.4 Sensitivity analysis results

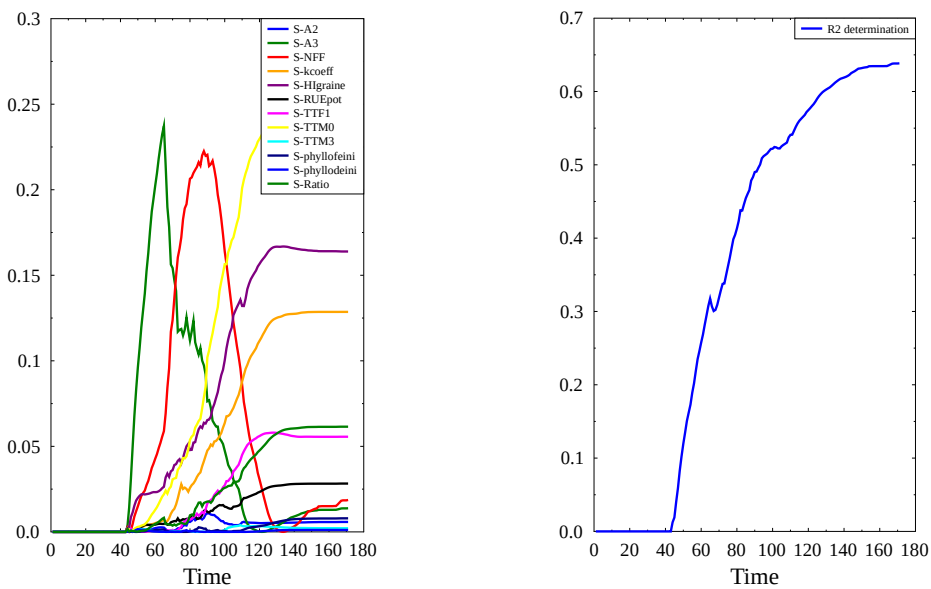
Since the model analysis is a complicated process, the sensitivity analysis will be carried out in a different context (Bayol, 2016). Normally, an SRC based analysis will be carried out first to verify the linearity of the objective function relative to the inputs. If the linearity is not well verified, Sobol

indices will be applied. It allows a better understanding of the interactions among the parameters during a plant's growth cycle. In terms of parameter estimation, a "Sobol-gls" algorithm, which is integrated by the generalised least squares (gls) criterion, makes it possible to obtain the ranking of the most important parameters for estimation.

Table 2.1 Variation intervals and the recommended value for the model parameters

Parameter	Interval	Recommended Value
A2	$[7, \quad 19]$	14.07
A3	$[400, 720]$	645
phyllodeini	$[22, \quad 42]$	32
Ratio-phyllodephyllofe	$[0.5, 0.9]$	0.7
TTF1	$[410, 890]$	723
TTM3	$[950, 1750]$	1477
kcoef	$[0.4, 0.75]$	0.53
phyllofeini	$[30, \quad 50]$	40
NFF	$[12, \quad 26]$	21
HI	$[0.3, \quad 0.8]$	0.5
RUE	$[0.5, \quad 0.9]$	3.5
TTM0	$[550, 1060]$	884

General sensitivity analysis



(a) SRC indices for parameters relative to the "yield"

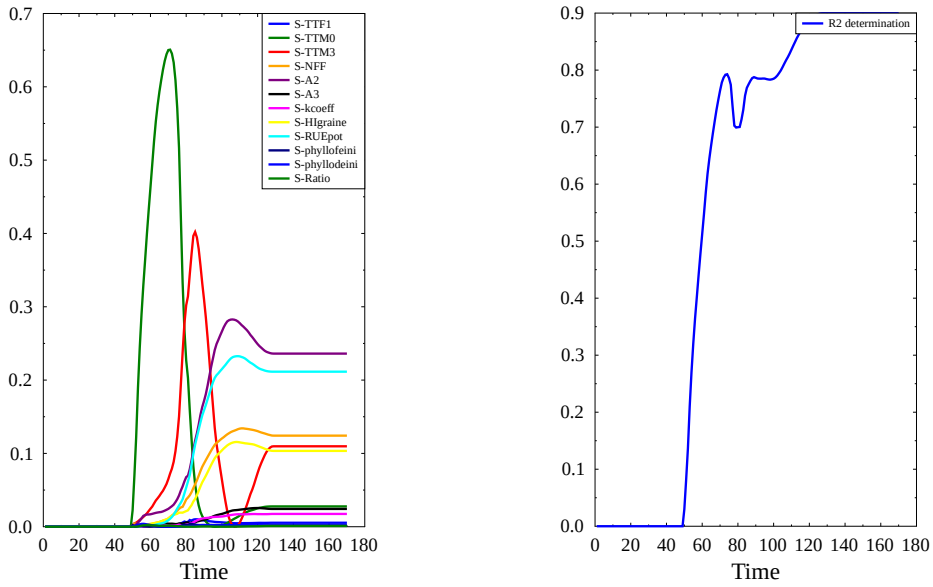
(b) R^2 determination

Fig. 2.2 A result of sensitivity analysis result with SRC method

As a first step, a general sensitivity analysis with the SRC index is applied to verify the linearity between the output (yield) and the inputs. The parameters are supposed to be random, independently and uniformly chosen from their uncertain intervals as listed in Table 2.1. These intervals are suggested by the plant modeller, after a large number of statistical measurements. As described in Section 1.2.1, the database consists of 720 yields observations and the relevant meteorological records. Since plant's growth has a lot of interactions with their environment, the SRC method is repeated with ten different backgrounds. The SRC indices of the parameters relative to the "yield" and the R^2 determination for a specific environment are shown in Figure 2.2.

The results seem to be reasonable because the variable like the A3 (the potential leaf surface) and NFF (number of leaves) are outstanding after the appearance of the leaves. However, according to the R^2 determination coefficient (around 0.64), there is still a significant amount of unexplained variance under the linear assumption.

Thus, a Sobol analysis, with the indices that depend on the variance decomposition, is carried out under the same configuration with the SRC in 10 environments. One of the results are shown in Figure 2.3a and Figure 2.3b.



(a) Sobol first order indices for parameters

(b) R^2 determination

Fig. 2.3 A result of sensitivity analysis with the Sobol method

The Sobol sensitivity indices for each parameter are slightly different from the SRC indices. The parameters like RUEpot (the potential radiation use efficiency), and the ones relative to the leaves like A2 and NFF stand out from the others in the biomass accumulation process. What

is important is that the R^2 determination increases to 0.90, which means that nearly 90% of the variance relative to the yield could be well explained by the variance decomposition method. Also, the rest of the variances is supposed to be caused by the correlation among the parameters. Maybe better exploitation could be obtained with the work in (Sainte-Marie et al., 2017).

Sensitivity analysis for parameter estimation

A good comprehension of the behaviour of the crop yield relative to each parameter and their interactions could be obtained with the general sensitivity analysis. As far as parameter estimation is concerned, more constraints should be taken into account, like the observation frequency and the level of observation noise. For this, on the platform, a "Sobol-gls" algorithm, which integrates the minimisation criterion for estimation, has been implemented to select the most critical parameters for estimation concerning experimental data (Bayol, 2016). This criterion corresponds to the generalised least squares criterion used in the Aitken estimator (Cournède et al., 2011). The parameterisation in discrete dynamic models is described in (Goodwin and Payne, 1977), and (Zhan et al., 2003) presented an application in the case of the Greenlab model.

The generic formulation of the model in our case has been described as Equation 1.1, where the model output is considered to be a function of environmental variables and a parameter vector. When individual measurements y_i concern only the final yield at a given time t_i, counted from the sowing day, in a specific environment U_i, then assuming an additive error term, we get the following form

$$y_i = f(U_i, \theta, t_i) + \varepsilon_i, \qquad i \in [1:N], \tag{2.6}$$

where y_i are uni-dimensional and $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ is usually assumed to follow a multivariate normal distribution $N(0, \Sigma)$.

But in our case, for each meteorological record, we have only the final observation, the yield for the corn in their whole life. It means that y consist only a single component. The estimation of the parameter with the observation y_i in relative environment U_i can be obtained easily by:

$$\tilde{\theta} = \arg\min_{\theta} \ (f(\theta, U_i) - y_i)^2. \tag{2.7}$$

In order to have a common estimation of parameter for a certain genotype in different environment $(U_1, U_2, ...U_n)$, then the common estimation can be obtained as following:

$$\tilde{\theta} = \arg\min_{\theta} \ \sum_{i=1}^{n} (f(\theta, U_i) - y_i)^2. \tag{2.8}$$

Ideally, a sensitivity analysis integrated with the least-squares error criteria helps to rank the parameters' importance for estimation. It is also a rank that has considered the climatic variance. However, the calculation cost doesn't allow taking account of all the environments at the same time.

An alternative is to calculate the sensitivity indices in each context and make the average value at last. The summary of Sobol indices in terms of parametrisation is listed in the following table.

The subset of parameters {NFF, RUEpot, A2, HIgraine, TTF1, TTM0 and TTM3} is firstly chosen to be the critical parameters for calibration for the CORNFLO model according to the importance of total first-order Sobol indices.

Table 2.2 Total first order Sobol indices for parameters in the CORNFLO model

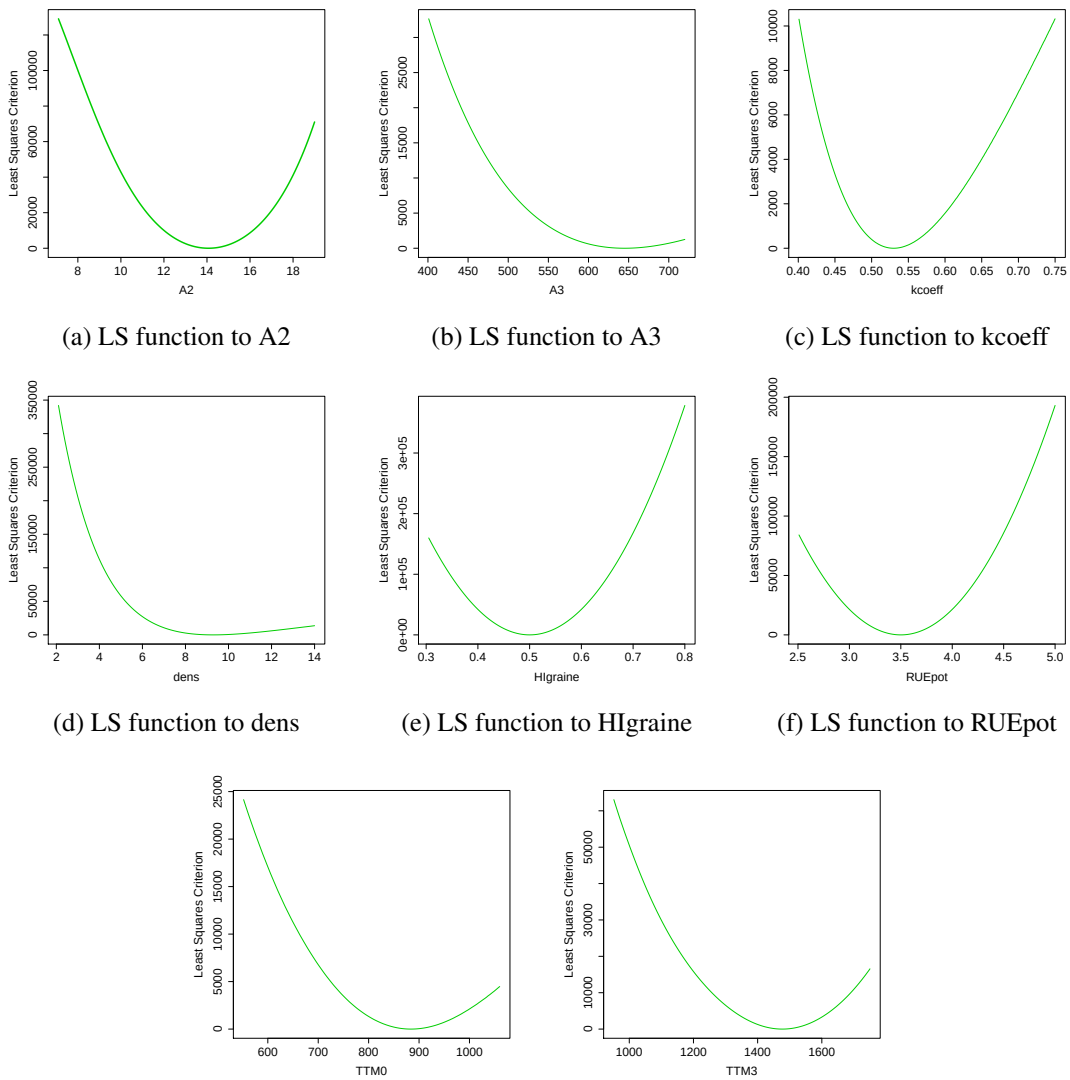
envID	TTF1	TTM0	TTM3	NFF	A2	A3	kcoeff	HIgraine	RUEpot	phyllofeini	phyllodeini	Ratio
1	0.0778	0.1851	0.1232	0.2917	0.4319	0.0987	0.0216	0.3145	0.4867	0.0124	0.0142	0.0160
2	0.0697	0.1513	0.1105	0.2860	0.4222	0.0887	0.0211	0.3091	0.4673	0.0145	0.0086	0.0112
3	0.0987	0.1504	0.0999	0.3492	0.3949	0.0757	0.0171	0.2663	0.4378	0.0376	0.0140	0.0333
4	0.1425	0.1582	0.1175	0.3718	0.3442	0.0923	0.0245	0.2685	0.4156	0.0942	0.0322	0.0885
5	0.2071	0.1166	0.0448	0.4566	0.2117	0.0332	0.0055	0.1418	0.2258	0.0687	0.0882	0.0319
6	0.1867	0.1053	0.0153	0.3486	0.1140	0.0005	0.0145	0.0642	0.1160	0.0923	0.1560	0.4668
7	0.1648	0.1232	0.0312	0.2694	0.1117	0.0320	0.0171	0.0643	0.0922	0.0173	0.2013	0.0347
8	0.2928	0.1277	0.0691	0.4745	0.3000	0.0524	0.0096	0.1718	0.2695	0.0622	0.0176	0.0357
9	0.1722	0.1244	0.0566	0.4656	0.1634	0.0408	0.0056	0.1374	0.2224	0.0234	0.0238	0.0421
10	0.0896	0.1585	0.1226	0.3006	0.4047	0.0877	0.0199	0.3048	0.4589	0.0168	0.0577	0.0169
Average	0.1502	0.1401	0.0791	0.3614	0.2899	0.0602	0.0116	0.2043	0.3192	0.0439	0.0614	0.0777

2.2 Identifiability, Continuity and Convexity Analysis

The identifiability of the parameters of a given model is a crucial step (Brun et al., 2001) insofar as it is one of the assumptions in the statistical modelling that ensures the consistency of the parameter estimators. In mathematical statistics, a statistical model is identifiable if:

$$f(\theta) = f(\theta)' \Rightarrow \theta = \theta' \tag{2.9}$$

with $f(\theta)$ and $f(\theta^{'})$ the distribution of the response function when the parameter vector is θ and $\theta^{'}$. Let us note that a model is said to be identifiable if for any pair of vectors different from values of the parameters lead to different outputs. Referring to the model's description in Appendix B, it is a problem of identifiability will appear when RUE and HI are estimated at the same time. It happens again when A3 and kcoef are estimated at the same time.



(a) LS function to A2 (b) LS function to A3 (c) LS function to kcoeff

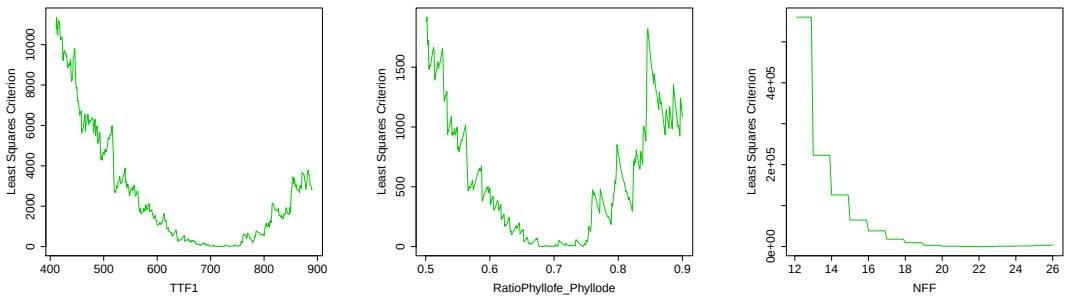
(d) LS function to dens (e) LS function to HIgraine (f) LS function to RUEpot

(g) LS function to TTM0 (h) LS function to TTM3

Fig. 2.4 Parameter with smooth relation to the objective function

The function to be minimised concerning parameters can serve as a tool for the continuity and convexity analysis. The objective function related to each parameter of CORNFLO model is analysed. The associated results are shown in Figure 2.4 and Figure 2.5. The parameter {RT, RE and RO} related to the water stress is not analysed in this part since the observations are from 100% irrigated case.

Finally, the parameters can be divided into three categories: 1. The parameters with smooth curves, like {dens, A2, A3, kcoef, HI, RUE, TTM0 and TTM3}; 2. Parameters with irregular curves as {TTF1, phyllofeini, phyllodeini and Ratio}; 3. The parameters that only take integer values, like {NFF}.



(a) LS function to TTF1 (b) LS function to Ratio (c) LS function to NFF

Fig. 2.5 Parameter with irregular relation (non-convex) to the objective function

2.3 Conclusion

In summary, considering the result of sensibility analysis in Section 2.1 and the other analysis in Section 2.2, a subset of five important variables with continuity {TTF1, RUE, TTA2, TTM3 and TTM0} will be estimated in the parametric stage in the following chapters.

Chapter 3

MuScPE-PSO: MuScPE methodology with PSO optimisation

In the last chapter, sensibility analysis has been presented as a mathematical tool for the determination of crucial parameters in crop models. And a subset of the sensitive parameters has been chosen, and their parametrisation remains as a difficult task since it turns out to carry out an optimisation project in high dimension, where the optimal solution is challenging to settle. What's more, in this research, we want to take advantage of a dataset with easy access. And they are generally agricultural record in large-scale and from a large number of diverse environmental scenarios. Thus, one of the most significant contributions of this thesis is to propose a methodology, named "the Multi-Scenario Parameter Estimation methodology (MuScPE)", for such a specific problem. In Section 3.1, we are going to present the general idea of MuScPE; then , a basic introduction of the implemented optimisation algorithm, the particle swarm optimisation, will be made in Section 3.2; some improvement for PSO will be presented in Section 3.3-3.5; the objective model will be calibrated in Section 3.6

3.1 Multi-Scenarios parameters estimating methodology

In (Kang, 2013), an innovative methodology, based on inverse methods, is firstly mentioned for plant model parameters estimation: the Multi-Scenario Parameter Estimation methodology (MuScPE). This methodology takes advantage of the large-scale data in the form $\{U_i, y_i\}$, with U_i the daily meteorological records and y_i the average yield observation at county level (usually large amounts of environmental conditions U_i are available, but for each scenario only a single experimental observation, the final crop yield y_i is known). It gives an alternative for parametrisation in crop modelling.

3.1.1 Mathematical interpretation of the problem

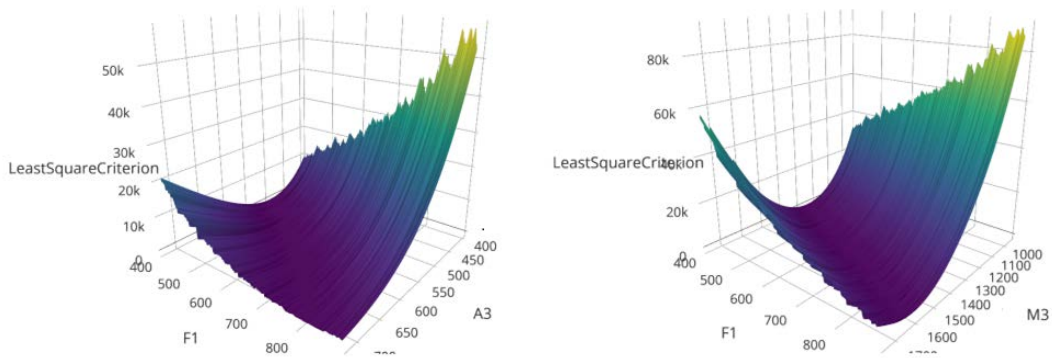
From a mathematical perspective, the MuScPE methodology is a parameter estimation method based on the inverse method. The i-th output of a plant growth model y_i can be expressed as a result of biological interaction between a specific genotype θ and its relative environment U_i as shown in Equation 2.6.

During the parameterisation process, a classical LSE estimation method is mostly applied to calibrate the unknown parameter vector, i.e. the optimal parameter setting is obtained via the minimisation of the squared loss function

$$L(\theta) = \sum_{i=1}^{N} (y_i - f(\theta, U_i))^2. \tag{3.1}$$

so as to obtain $\tilde{\theta}$.

It is a typical inverse problem. As already mentioned in Section 1.3.2, it turns out to be a continuous optimisation problem with a single objective function of a multi-dimensional parameter. A 3-D representation of $L(\theta)$ as a function of two different choices of parameter components is shown in Figure 3.1:



(a) Optimisation surface with F1 and A3 (b) Optimisation surface with F1 and M3

Fig. 3.1 3-D presentation of the optimisation surface

The non-convexity demonstrated in the above Figure increases the difficulty of optimisation. Moreover, the choice of the optimisation algorithm has an essential effect on the results. As already mentioned in Section 1.3.2, the ordinary gradient-based algorithm cannot deal efficiently with non-convexity. In the following section, a robust global optimisation algorithm, the Particle Swarm Optimisation, and its variants will be introduced. Moreover, a dynamic configuration of the PSO algorithm will be further developed and then integrated into the MuScPE methodology with parallelism.

3.2 Basic Particle Swarm Optimisation

Kennedy and Eberhart proposed Particle Swarm Optimisation (PSO) algorithm in 1995 (Eberhart and Kennedy, 1995). The social behaviour of swarming animals inspires this method. The most commonly used example is the behaviour of fishes and birds (Reynolds, 1987). We can observe from these animals some complex moving dynamics, whereas each individual has limited intelligence and local knowledge of his situation in the swarm. An individual of the swarm is only aware of the position and speed of his nearest neighbours. Therefore, each one uses both his memory and local information about his nearest neighbours to decide on his move. Thus, the moving rules can be quickly determined, such as "maintain the same speed with the others", "move in the same direction" or "stay near to the neighbours". Consequently, the "group's intelligence" is the integration of local interactions among the different particles of the swarm. As the saying goes, "many heads are better than one".

Kennedy and Eberhart were inspired by these socio-psychological behaviours to create the PSO. A swarm of particles, which are potential solutions to the problem of optimisation, "flies over" the search space to search for the global optimum. The following three components influence the movement of a particle:

- A physical component: the particle tends to follow its current direction of movement;
- A cognitive component: the particle tends to move towards the best site by which it has already passed;
- A social component: the particle tends to become the experience of its neighbours and, thus, to move towards the best site already reached by its neighbours.

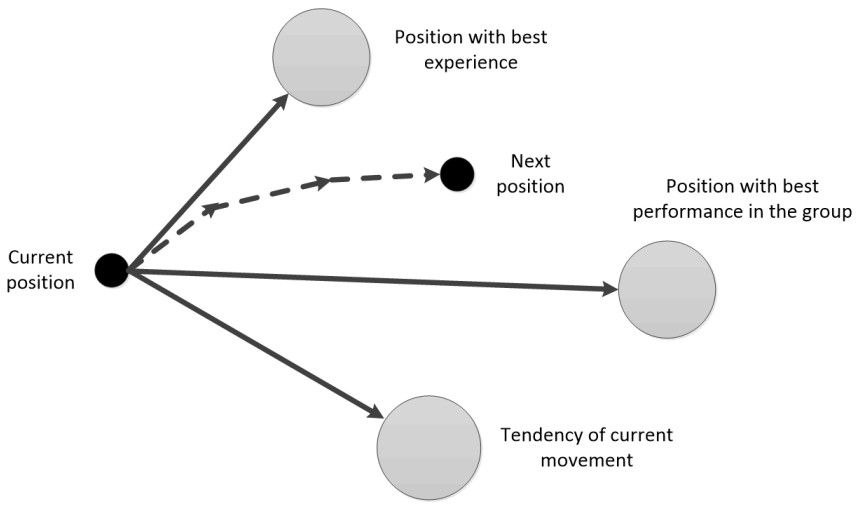


Fig. 3.2 Movement of particle in the searching space

In a D-dimensional research space, the i-th particle of the swarm is modelled by its position vector $\overrightarrow{X}_i = (x_{i,1}, x_{i,2}, ..., x_{i,D})$ and its velocity vector $\overrightarrow{V}_i = (v_{i,1}, v_{i,2}, ..., v_{i,D})$. It keeps in memory

the best position by which it has already passed, denoted by $\overrightarrow{p_i} = (p_{i,1}, p_{i,2}, ..., p_{i,D})$. The best position reached by all the particles of the swarm is denoted by $\overrightarrow{g} = (g_{i,1}, g_{i,2}, ..., g_{i,D})$.

The position $x_i(t)$ of the i-th particle at time t is assumed to be the vector sum of its previous position $x_i(t-1)$ and a velocity vector $v_i(t)$ which is determined by the influence of the three components described above on its previous state ($x_i(t-1)$,$v_i(t-1)$). In particular, it is assumed that the time step is sufficiently small to describe well the state change of each component $x_{i,k}(t)$ of the position vector $x_i(t)$ by a locally linear approximation of the form

$$x_{i,k}(t) = x_{i,k}(t-1) + v_{i,k}(t), \quad k \in \{1, 2, \ldots, D\}, \tag{3.2}$$

where $v_{i,k}(t)$ is given by

$$v_{i,k}(t) = v^p_{i,k}(t) + v^c_{i,k}(t) + v^s_{i,k}(t), \quad k \in \{1, 2, \ldots, D\}, \tag{3.3}$$

and each component vector of the vector sum by

$$v^p_{i,k}(t) = \omega * v_{i,k}(t-1), \tag{3.4}$$

$$v^c_{i,k}(t) = c_1 * r^c_{i,k}(t) * (p_{i,k}(t-1) - x_{i,k}(t-1)), \tag{3.5}$$

$$v^s_{i,k}(t) = c_2 * r^s_{i,k}(t) * (g_k(t-1) - x_{i,k}(t-1)), \tag{3.6}$$

where ω is called coefficient of inertia and is generally treated as a parameter which is constant in time and common to all particles, c_1 and c_2 are two constants, called acceleration coefficients, $r^c_{i,k}(t)$ and $r^s_{i,k}(t)$ correspond to uniformly distributed random numbers independently drawn from the unit interval $[0, 1]$ for the k-th component of each particle i at time t. The following interpretation holds:

- $v^p_{i,k}$ corresponds to the physical component of the displacement. The parameter ω controls the influence of the direction of movement on the future displacement. It should be noted that, in some applications, the parameter ω can be variable (Dréo et al., 2006).
- $v^c_{i,k}$ corresponds to the cognitive component of the displacement. Actually, c_1 controls the cognitive behaviour of the particle, since it acts as a common scaling factor of the effect given by the cognitive process specifically to the i-th particle.
- $v^s_{i,k}$ corresponds to the social component of displacement and this time the factor c_2 scales the effect given by the social fitness of the i-th particle.

The combination of the parameters ω, c_1 and c_2 makes it possible to adjust the balance between the diversification and intensification phases of the research process (Ishibuchi and Murata, 1998) (Kennedy et al., 2001). Now, we give a short description of the PSO algorithm.

The PSO is a population-based algorithm. It starts with random initialisation of the swarm in the search space S_D. At each iteration (time), every particle decides its velocity according to

Algorithm 1 Basic Particle Swarm Optimisation Algorithm

```
Initialise a population of particles with random values positions and velocities from D dimensions in the search space
while Termination condition not reached do
    for Each particle i in 1 : N do
        Adapt velocity of the particle using Equation 3.3
        Update the position of the particle using Equation 3.2
        Evaluate the fitness f(X_i)
        if f(X_i) < f(P_i) then
            P_i ← X_i
        end if
        if f(X_i) < f(P_g) then
            P_g ← X_i
        end if
    end for
end while
```

Equation 3.3 and updates its position by Equation 3.2. Then, the objective function is evaluated by taking into account the new positions and the best particle-specific vector $\overrightarrow{p_i}$, as well as the best vector, $\overrightarrow{g}$ are updated. This procedure is summarised by Algorithm 1, where N is the total number of particles in the swarm.

This process will be terminated if one of the stopping criteria is satisfied. Typically, the choice of stopping criteria depends on the actual problem. If the global optimum is known a priori, we can define an acceptable error as a stopping criterion. In most cases, it is common to set a maximum number of iterations to avoid endless loops. This topic will be discussed in the next chapter.

3.3 Variants of PSO

The basic PSO, as described in Algorithm 1, when applied in practice, can easily be encountered with some difficulties, such as the explosion of particles or the fast convergence to local minimum. To overcome these problems and make the particles behave finer in the search space, some variants of the original PSO algorithms have been proposed. These proposals can be classified mainly into four aspects, as follows:

3.3.1 Partitioning particles

An unpleasant situation can arise if a particle moves out of the search space, thus leading to the system's divergence. To overcome this problem, it is possible to introduce a new parameter of V_{max}, which makes it possible to control the explosion of the velocity (Shi and Eberhart, 1998). A study on the behaviour of the PSO algorithm according to the values of V_{max} is available in (Fan, 2001).

Besides, other strategies could be introduced to bring back a particle, which has moved out of the search space. Different treatments include exclusion of the particle, repetition of the randomly selected velocity until acceptance of the new position or repositioning in the space centre. In fact, no overall best strategy exists (Zielinski and Laur, 2007).

3.3.2 Constriction factor

The previously mentioned strategies which aim at bringing back to the search space divergent particles or delaying this effect by the introduction of a parameter bound V_{max} can only be considered as a partial solution to a persistent problem. The "explosion" could still happen. Several other studies, including (Kennedy, 1998), (Clerc and Kennedy, 2002) and (Van Den Bergh, 2007), attempt to study on the swarm's dynamics so that a convergence of the swarm is essentially ensured.

In (Clerc and Kennedy, 2002), it has been shown that good convergence can be achieved by making the parameters ω, c_1 and c_2 dependent. The use of a constriction factor ϕ makes it possible to prevent the explosion of the swarm. In particular, this modification can be described by the following equation

$$v^*_{i,k}(t) = \chi(\phi) * \left(v^p_{i,k}(t) + \phi_1 * v^c_{i,k}(t) + \phi_2 * v^s_{i,k}(t) \right), \quad k \in \{1, 2, \ldots, D\}, \tag{3.7}$$

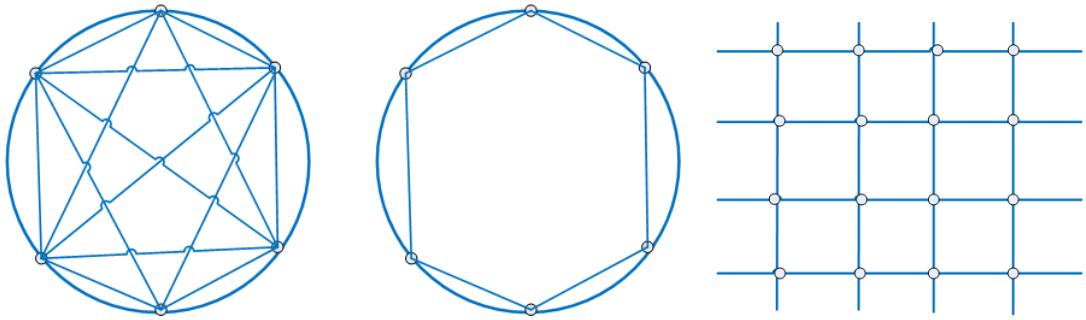
where $\chi(\phi) = \frac{2}{\phi - 2 + \sqrt{\phi^2 - 4\phi}}$, $\phi = \phi_1 + \phi_2$, $\phi > 4$. Note also that the SPO with constriction coefficient is equivalent to the basic PSO with $\omega = \chi(\phi)$, $c_1 = \chi(\phi) * \phi_1$ and $c_2 = \chi(\phi) * \phi_2$.

In (Clerc and Kennedy, 2002), numerous tests are conducted to determine the optimal values of ϕ_1 and ϕ_2. In the majority of cases, we use $\phi = 4.1$ and $\phi_1 = \phi_2 = 2.05$, thus using as a multiplicative coefficient $\chi = 0.7298$.

3.3.3 Local PSO with Neighbourhood Topology

To overcome the fast convergence into a local minimum, Eberhart has introduced a so-called local version of the PSO, which also uses a static graphic information (Eberhart and Shi, 1998). This version uses a circular information graph as in Figure 3.3 to represent the neighbourhood relation. The particles of the swarm are virtually arranged in a circle and numbered sequentially from 1 through the loop. The particle is therefore no longer informed by all the particles of the swarm, but by itself and its two neighbours. If we refer to the basic version of the PSO summarised in Algorithm 1, the best particle $\overrightarrow{g}$ is chosen from the entire population. From the view of graphics information, the swarm is therefore wholly-connected (or path-connected).

Although it converges less rapidly than the global version, the local version of the PSO gives better results because it is less subject to attraction by local minimum (Kennedy, 1999). Figure 3.3 illustrates the difference between a wholly-connected and a circular graph.



(a) Fully connected structure (b) Ring structure (c) Von Neumann structure

Fig. 3.3 Some of the neighbourhood topology used for the PSO technique: (a) Fully connected structure, (b) Ring structure, and (c) Von Neumann.

Many other topology have been tested. The results using different topology are available in (Kennedy, 1999), (Clerc, 2005). The important information that emerges from these results is that no topology is globally better than the others since no topology dominates in a wide range of problems.

3.3.4 PSO and hybridisation

Hybridisation involves combining the characteristics of two different methods to derive the benefits of both ways (Talbi, 2002). This topic will not be discussed here since, in this research, there was no application of hybridisation procedures. For some reviews on this point, the interested reader is referred to (Thangaraj et al., 2011) and (Xin et al., 2012).

3.4 A well set PSO and evaluation with simulated dataset

Numerous tests were performed to determine the best version of the PSO algorithm corresponding to the optimisation problem of the CORNFLO model. The best performance was exhibited by the local version of the PSO with the Von Neumann structure, as shown in Figure 3.3c and the incorporation of a constriction factor, as explained in Section 3.3.2.

To prove its efficiency and robustness in global optimisation, this well-set PSO algorithm, in which the particles are controlled by constriction coefficients and a Von Neumann neighbourhood structure, is firstly applied in the parameter estimation with simulated data. In the sequel, we describe in detail the tests:

- With a certain reasonable parameter setting, denoted by θ_0, the virtual crop yields $\{y_1, y_2, ..., y_n\}$ are simulated under certain environments $\{U_1, U_2, ..., U_n\}$ as follows:

$$y_i = f(\theta_0, U_i) \quad \text{for} \quad i \in 1:n \tag{3.8}$$

- Assuming that the parameter setting is unknown, the estimation of the common parameters $\tilde{\theta}$ could be performed by model inversion as in Equation.1.2, with a subset of virtual yields and their corresponding environments.

A series of tests following the above process with an increasing number of parameters were conducted, where $\theta_1 = \{A2\}$, $\theta_2 = \{A2, RUE\}$ and $\theta_3 = \{A2, RUE, TTF1\}$. The process with 100 randomly chosen scenarios was repeated for 100 times for each test, and the resulting mean parameter estimates, together with the bias and the standard deviation are recorded in Table 3.1

θ_1:			θ_2:					
$A2 = 14.07$			$A2 = 14.07$			$RUE = 3.5$		
mean	*bias*	*sd*	*mean*	*bias*	*sd*	*mean*	*bias*	*sd*
14.096	0.026	0.118	14.125	0.055	0.45	3.531	0.031	0.156

θ_3:								
$A2 = 14.07$			$RUE = 3.5$			$TTF1 = 723$		
mean	*bias*	*sd*	*mean*	*bias*	*sd*	*mean*	*bias*	*sd*
13.857	-0.214	3.36	3.745	0.245	0.612	726.32	3.32	3.51

Table 3.1 Parameterisation results with an increasing number of parameters.

The results showed that MuScPE-PSO works well for parameter calibration in this complex system, since the $\tilde{\theta}$ is close to the initial settings with small bias and standard derivation, especially in the first two cases. It could also be observed that the bias and the uncertainty increase when more parameters are accounted for calibration. The increase is more significant when $TTF1$ is introduced since non-convexity is caused by $TTF1$, which make the surface of optimisation surface more complicated, as shown in Figure 3.1. These results could also serve as a way to support the "feasibility" of MuScPE methodology.

3.5 Improved PSO with parallelism

In the above section, good performance of parameter estimation was illustrated with the use of particle swarm optimisation "PSO". Indeed, this global optimisation algorithm is a powerful meta-heuristic in the resolution of difficult optimisation problems. Nevertheless, some drawbacks also exist. The most important one concerns the fact that this algorithm is very computationally expensive.

On the other hand, due to the rapid evolution of hardware resources in the computer field, computers have seen their number of processors/cores significantly increased in recent years. It makes it possible to compensate for the limits of power increase of a single processor and to obtain a factor of acceleration. To fully exploit this computing power, it is necessary to realise applications capable of performing several tasks in parallel (Kumar et al., 1994). In this section, the parallelism

will be introduced as a means to decrease the computational cost significantly. This principle will be applied to a specific version of the PSO algorithm, and its performance will be compared with that of the sequential one.

3.5.1 Programming of massively parallel machines

There are different models of programming adapted to different hardware architectures of massively parallel devices. Threads are the technology that makes applications multitasking. Most of the parallel technology takes advantage of the threads to reduce computing time (Kumar et al., 1994). However, when talking about memory architectures, some are more appropriate for distributed memory architectures like MPI (Message Passing Interface), while others for shared memory architectures such as OpenMP (Open Multi-Processing).

MPI with allocated memory

The message-passing model is one of the massive parallel models that exist. The parallelisation of a problem is carried out with the use of some processes running concurrently. Each of them having only access to a private area of memory. When they have to share information, they exchange messages. Thus, the processes do not need to share memory for communication, and therefore, message programming is particularly suited to machines where memory is distributed (Gropp et al., 1996).

MPI is one of the most used standards of Message Passing paradigm, developed in the 1990s. It defines a programming interface for implementing point-to-point, corporate communications and synchronisations between different processes. This interface can be either synchronous or asynchronous. Building an MPI parallel application involves writing a single program that will be duplicated n times. Each program instance is assigned a unique identification number (or rank) so that it can select which processes communicate. A concept of a communicator is defined, allowing to create groups of operations and thus to limit the range of the messages to the members of the same group.

OpenMP with shared memory

Parallel shared memory architectures allow different execution streams to use the same memory area to communicate with each other (Dagum and Menon, 1998). The most natural way of programming these architectures is the use of lightweight processes (or threads). Thus, each instruction flow is encapsulated in different light processes sharing the same memory space between them. The use of threads allows a great deal of flexibility but requires the programmer to take many precautions, mainly when competing for access to shared memory areas. Thus, a whole set of synchronisation primitives (condition variables, semaphores) is available in all thread implementations.

OpenMP is a specification developed in 1997 by a set of actors involved in parallel computing (Chapman et al., 2008). The specification defines a set of annotations, a collection of functions, and a set of environment variables. The significant advantage of OpenMP over the others is its ability to hide programmers from thread management details. Indeed, to parallelise a loop, a loop, for example, indicate before it that we want to parallelise it. The compiler will then automatically cut the data and create threads (Dagum and Menon, 1998).

3.5.2 Proposition of a parallelisation approach of the PSO method

In the implementation of the classical algorithm of the PSO method, all calculations are done sequentially. Nonetheless, the evaluation of each particle (candidate solution is independent). That is where the idea of parallelisation comes from to improve the performance of this algorithm. Several proposals are made with different principals (Chang et al., 2005), (Zhou and Tan, 2009). The one we adopted for our implementation, is to parallelise the calculations by launching a set of threads on batches of particles based on CPU.

Threads, a kind of process, run in parallel for each iteration of the algorithm. Each thread executes one iteration of its batch of particles and waits for the other threads to finish processing to update the neighbourhoods and start a new iteration. This process is repeated until a satisfactory solution is obtained: "achievement of the stopping criterion". Figure 3.4 is a representation of the proposed approach.

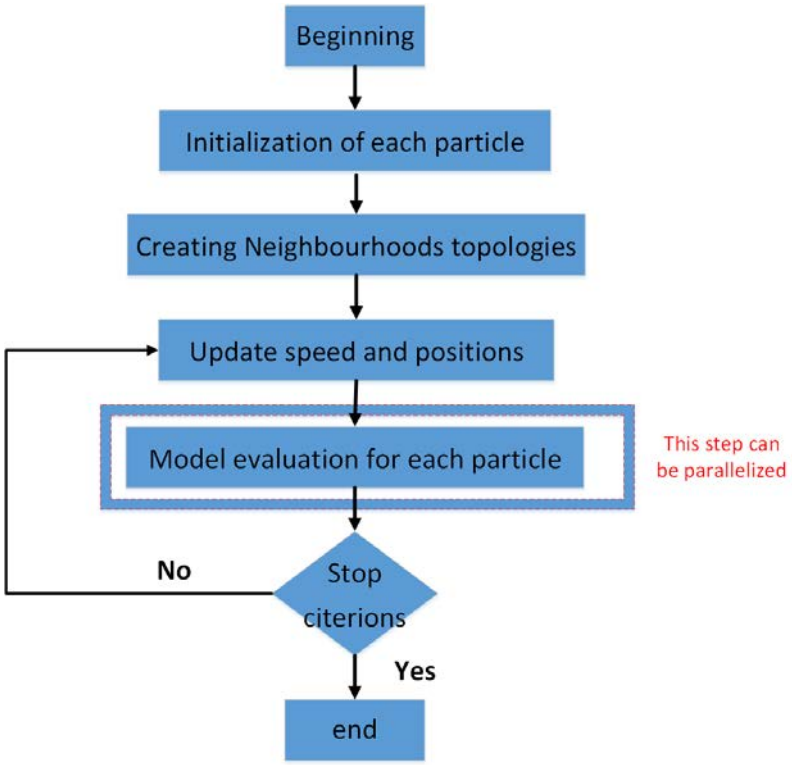


Fig. 3.4 The structure of parallelised PSO

3.5.3 Comparison of OpenMP and MPI

In the following, a comparison of the OpenMP and MPI based parallelised mechanism will be made in terms of the acceleration efficiency. The calculation example will be one of the parametrisation examples of the plant growth model, where the MuScPE-PSO is used to parametrise the 5-dimension parameter vector of the dynamic model in 300 scenarios.

The comparison can be divided into two part: in the first part, the number of particles is used in the PSO swarm is set to be 1000, while the process number is changed from 1 to 10 to test the different speed-up efficiency; in the second part, the number of processors is fixed to be 10, while the number of particle changes from 100 to 1000 for testing the influence of calculation scale on the speed-up efficiency. The time used for 100 iterations is considered as the criteria. Each calculation is repeated for ten times, and the average time is taken to exclude the effect of randomness. The notion "speed-up" is defined as Equation 3.9.

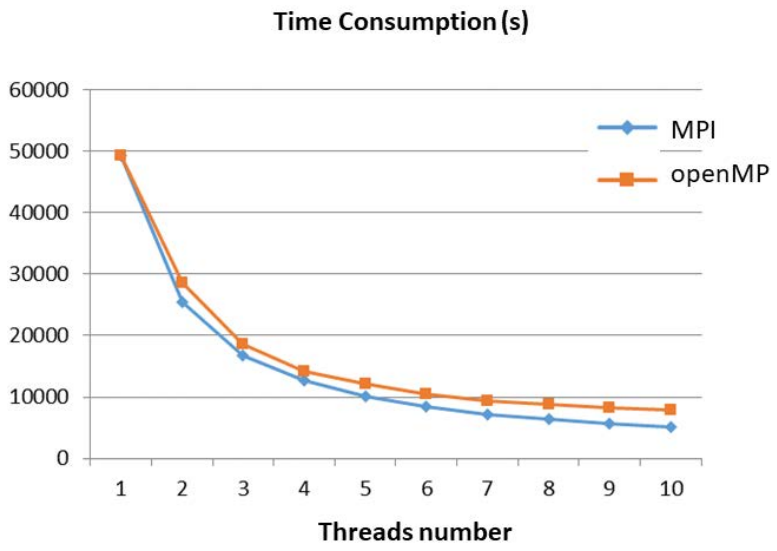
$$\text{speed} - \text{up} = \frac{T_{seq}}{T_{par}} \tag{3.9}$$

where T_{seq} stands for the calculation time consumed in sequential PSO algorithm, and T_{par} for the calculation time consumed in parallelised PSO algorithm.

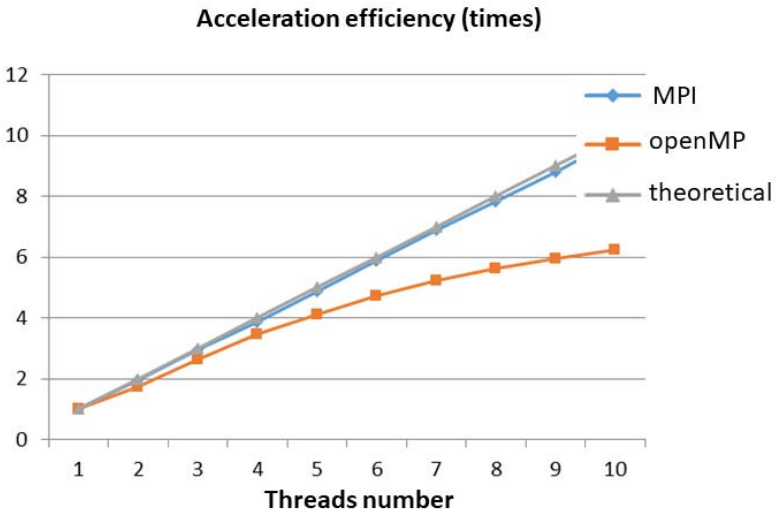
3.5.4 Results

The results for the first part of testing the different speed-up efficiency of MPI and OpenMP based parallelised PSO algorithm is shown in Figure 3.5. It can be concluded as follows:

- when a total calculation is fixed, adding threads can significantly improve the calculation efficiency in both mechanism;
- MPI based parallelism can achieve nearly the theoretical speed-up when a thread is added into to structure;
- OpenMP based parallelism cannot make fully used of the added thread as MPI. For example, when the threads are 10, the speed-up is just about 6.23.



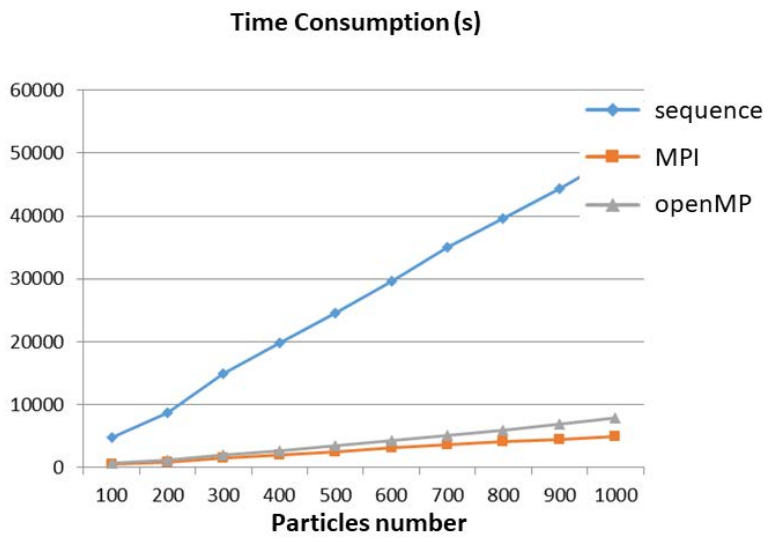
(a) Time consumption in the first comparison



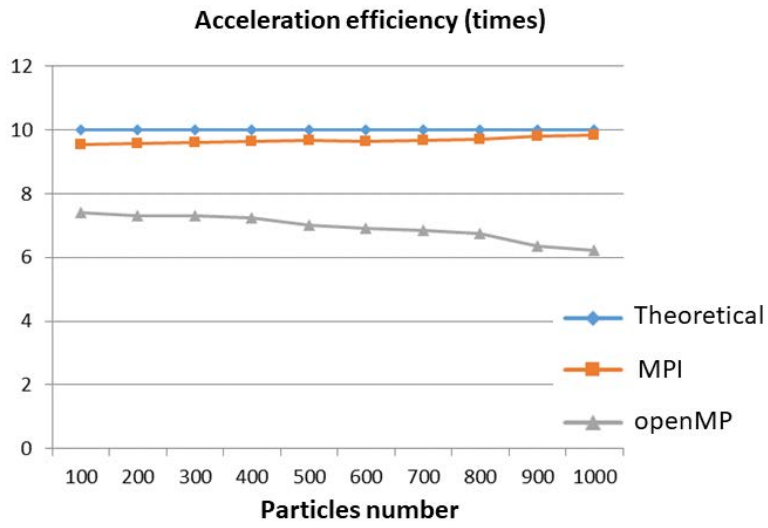
(b) Speed up efficiency in the first comparison

Fig. 3.5 First comparison result between MPI and OPENMP

The results of the second comparison for comparing the influence of calculation amount, when more and more particles are introduced in PSO algorithms, are shown in Figure 3.6.



(a) Time consumption in the second comparison



(b) Speed up efficiency in the second comparison

Fig. 3.6 Second comparison result between MPI and OPENMP

We can see that for the MPI-based PSO, the speed-up efficiency gets closer and closer to the theoretical value when the calculation becomes more substantial and more substantial (more and more particles are introduced). It can be explained that, when the estimate becomes more massive, the time used for data transmission becomes ignorable in comparing with the total running time. On the contrary, the speed-up efficiency for OpenMP-based PSO becomes lower and lower. As explained in Section 3.5.1, OpenMP is based on shared memory. The heavier the calculation

becomes, the more time will be consumed in arranging the shared memory for the different thread. That's also why the OpenMP is recommended in lightweight processes.

In conclusion, according to the performance analysis above, the MPI mechanism could make full use of the computing resources. Even though the programming of MPI-based PSO is more complicated, codes can be done once for all the time. The parallelism architecture adopted in MuScPE-PSO will be MPI.

3.6 Estimation results with MuScPE-PSO

By performing sensitivity analysis and by dealing with identifiability issues as explained in Section 2, some important parameters have been selected to be estimated in the first step. Besides, a well-configured paralleled PSO has been implemented to the MuScPE methodology in the previous section. In this section, the calibration process for the CORNFLO model will be explained. The process is divided into two steps for parametrisation with MuScPE: (1). Estimation with all observations to find out the best setting for the data set; (2). "Boosting estimation", where only a subset of the training set takes part in the estimation process. But the estimation process is repeated for several times, and the subset is always re-sampled from the training set.

3.6.1 Estimation with all observations

In this part, all the 720 observations are used to estimate the best combination of the 5 most important parameters :$\{A2, RUE, TTM3, TTF1, TTM0\}$. According to the stopping condition of the PSO, a local PSO with four neighbours is used to perform the optimisation. Moreover, the total number of iterations is fixed at 300. To ensure the convergence to the global minimum, this process is repeated for ten times.

Results

First of all, let's have a look at the descent process of the objective function when the algorithm is running. Even though these ten repetitions don't take the same trajectory because of the randomness, but they finally converge to the same level. By combining the estimated values of these parameters, a conclusion can be drawn that PSO based MuScPE methodology has found out the best parameters setting in Table 3.2 for the CORNFLO model.

Parameter	A2	RUE	TTM3	TTF1	TTM0
Value	13.95	3.33	1453.7	707.35	869.45

Table 3.2 Best parameters setting of CORNFLO model that minimize the fitness with actual dataset.

3.6.2 Boosting estimation with MuScPE

Several times in modelling, the volume of the dataset is so large that the estimation results could not be obtained without a computer-cluster. Numerous solutions have been proposed to solve this problem, among which the boosting idea outperforms the others. The main purpose lies in the fact that, instead of using the whole dataset, only a subset is used in the boosting estimation process. But the process is repeated for several times and the subsets are randomly re-sampled for each repetition. By increasing the size of the subgroup, it has been proved that the estimated value converges to the results obtained with the whole dataset.

It is already mentioned in the previous section that the MuScPE-PSO algorithm induces high-cost in computation. The computation time can take hours, in the case that numerous scenarios are accounted for model calibration with the MuScPE-PSO. This is the reason why the idea of boosting estimation is tested in this section.

In this test, model calibration is performed with different sample sizes, such as 300, 400 and 500. Moreover, we take the same combination of parameters as in the previous section. Furthermore, each test is repeated for 100 times with re-sampling. Some commonly used features of these distributions will be analysed in the end.

Results

Table 3.3 Estimated values for the five parameters with different sample size

nbr_scenarios	RUE	TTM0	TTF1	TTM3	A2
300	3.1315	886.3476	718.1324	1498.169	14.3746
400	3.6684	882.1347	724.3489	1475.134	14.1476
500	3.3987	874.3564	726.3984	1468.324	13.7648

Table 3.4 Associated standard deviation

nbr_scenarios	RUE	TTM0	TTF1	TTM3	A2
300	0.1256	57.0013	42.2358	67.3258	0.4348
400	0.0814	29.1241	29.1345	45.1564	0.3781
500	0.0734	25.1327	24.2534	35.1214	0.1698

The mean estimated values of the parameters and their associated standard deviations for different numbers of scenarios are recorded in Table 3.3 and 3.4. According to the optimisation results, the difference in the sample size is reflected by a difference in the results. As expected, the increase in the number of scenarios induces less variability in parameter estimates as indicated by the smaller standard deviation. In this way, this increase reduces, in general, the resulting bias. This information could be necessary for the modeller, to have an idea of the number of scenarios needed

approximately to obtain a prespecified precision in parameter estimates and the relative importance of this information to each parameter.

3.6.3 Fitness and Prediction evaluation

To compare the performance of different parameter settings obtained with different strategies, two of the most common accuracy metrics of regression models were used: root mean squared error (RMSE) and the mean absolute relative error (MARE).

Root Mean Square Error

If $\hat{Y}$ is a vector of n predictions associated to n specific input vectors, and Y is the corresponding vector of observed values, then the RMSE of the predictor is defined by:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{3.10}$$

Mean absolute relative error

The mean absolute relative deviation (MARE), is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation. It usually expresses accuracy as a percentage, and is defined by:

$$\text{MARE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \times 100\% \tag{3.11}$$

where Y_i is the actual value and $\hat{Y}_i$ is forecast value.

The performance metrics of machine learning methods are summarised in the following table.

Metric	Expression
RMSE	$\sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n}$
MARE(%)	$\frac{1}{n} \times \left(\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid /\bar{y}\right) \cdot 100$

Table 3.5 Performance metrics of machine learning methods.

Results

The results for fitness and prediction evaluation is shown as in Table 3.6. It can be observed that better fitness is achieved with parameters settings obtained by MuScPE with a large dataset in the training process.

Table 3.6 Fitness and prediction capacity evaluation

sample size	RMSE0	RMSE	MARE0	MARE
300	71.7	73.1	6.21%	6.43%
400	71.9	72.9	6.22%	6.31%
500	71.2	71.8	6.18%	6.22%

3.7 Conclusion

The parameterisation of a dynamic plant growth model is an optimisation problem and in most cases, non-convex. Recognised for many years for their efficiency, meta-heuristics have stood out from the other algorithm in global optimisation. The Particle Swarm Optimisation (PSO) is a metaheuristic inspired by the group-living animals which provide optimal solutions based on population cooperation. It is well popular in optimisation research for its efficiency and simplicity in implementation.

It is essential to have a proper set of parameters that lead to the optimal performance of the algorithm. However, this work is always tedious and time-consuming. According to the studies have been conducted before and the actual tests with the real data for CORNFLO model, an instrumental version of the PSO algorithm is decided in Section 3.4. MPI-based parallelism is also applied to the PSO algorithm, and the computational capacity is well increased as in Section 3.5. And the evaluation for fitness and prediction has been carried out in Section 3.6 as well as in the research about the influence of datasets' size.

Chapter 4

Study of Climatic Variability

4.1 Basic descriptive statistics on Meteorological records

As discussed in Section 1.2.1, the dataset for corn crops analysis consists of 720 observations, each one corresponding to a different environmental scenario named by its county code and the relative year. For example, the first scenario's name is "04003-2001", where "04003" corresponds to the Cochise's County code in Arizona and "2001" to the year of meteorological records. Each yearly record consists of daily measurements of 5 environmental variables relative to the plant modelling process. These variables are listed below:

- Tmax (°C): daily maximal temperature,
- Tmin (°C): daily minimal temperature
- RG (MJ/m^2·day): daily radiation per square meter
- Prate (mm/day): daily precipitation
- ETP (mm/day): daily potential evapo-transpiration

The current research aims at studying the influence of the scenarios diversity on the parameter estimation process of the plant model. This diversity is dependent on the spatial distribution of the 720 scenarios.

4.1.1 Spatial Distribution

As stated in (Ackerly et al., 2010), the climate depends greatly on the relative geography. For this reason, a first insight into the variability present in the dataset can be gained by determining the spatial distribution of the 720 scenarios, as shown in Figure 4.1.

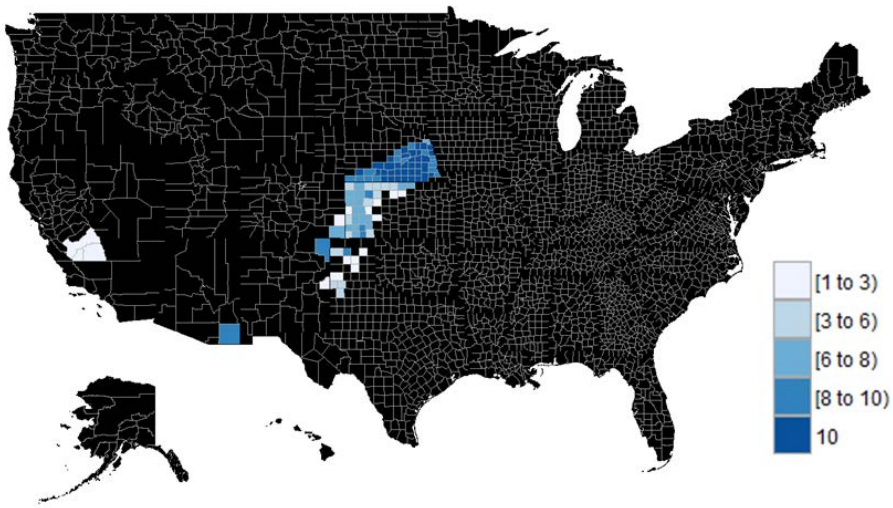


Fig. 4.1 Spatial Distribution

Notice that the available scenarios are far from being uniformly distributed in the United States. Most of them originate from the American centre land, like Colorado, Kansas, Nebraska, New Mexico, and Texas, while some others are isolated in regions like California near to the west coast and like Arizona near to Mexico. Detailed information about observations frequency at the state level is listed in Table 4.1. Further information concerning the inter-regional and inter-annual variability will be given in Section 4.2 and 4.3.

Table 4.1 Recorded frequencies at state level

Arizona	California	Colorado	Kansas	Nebraska	New Mexico	Texas
8	5	8	172	488	9	30

4.2 Inter-regional variability analysis on Meteorological records

We start this study by giving a description of the frequency of rainy days in four selected counties. These counties were selected since full meteorological records from 2001 to 2007 are available, as shown in Figure 4.1. In particular, Cochise County (04003) in Arizona, Sheridan County (20179) in Kansas, Union County (35059) in New Mexico and Franklin County (31061) in Nebraska. Additionally, these four counties are located nearly straightly on the direction going from the south-west to the centre of the United States. The basic descriptive statistics will be carried out with daily precipitation records since it is the typical variable that introduces more variability than the other meteorological variables. The variability will be analysed in two aspects: the inter-regional variability on the frequency of rainy days and rainy sequences.

4.2.1 Inter-regional variability on rainy days

This analysis is carried out in terms of frequency of rainy days(days) and their cumulative precipitation (mm) during 2001-2007. The results are listed in Table 4.2.

Table 4.2 Rainy day frequency and cumulative precipitation during 2001-2007

County	rainy days(days)	cumulative P(mm)	Rainfall intensity(mm/day)
Cochise (04003)	833	2204.4	2.64
Sheridan (20176)	1006	2770.5	2.78
Franklin (31061)	1060	3377	3.18
Union (35059)	1000	2128.6	2.12

The following conclusions can be drawn from these results:

- The states in the centre of America, like Nebraska and Kansas record more rainy days than the states in the south-west, such as Arizona;
- Although the Union County in New Mexico enjoys a higher frequency of rainy days than the Cochise County in Arizona and almost comparable to the others, it has the lowest cumulative precipitation. As a consequence, Union County has the lowest average precipitation per rainy day.

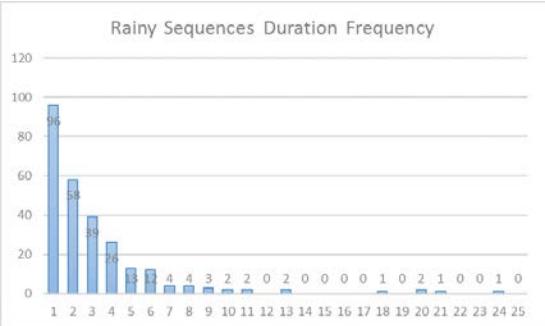
4.2.2 Inter-regional variability on rainy sequences

A rainy sequence is distinguished from a downpour by its long duration and its discontinuity; it can last several days, and it includes a series of deluges. Due to its potentially serious consequences, it attracts more attention to the scientific community than the simple frequency of rainy days. The duration of the sequences can provide information on regions where the short-term rainfall prevails and where the long rainfall periods out-stands. For this reason, we have calculated the rainy courses of different duration for the four counties individually during the year 2001-2007. The distribution of the duration of consecutive rainy days is shown in Figure 4.2.

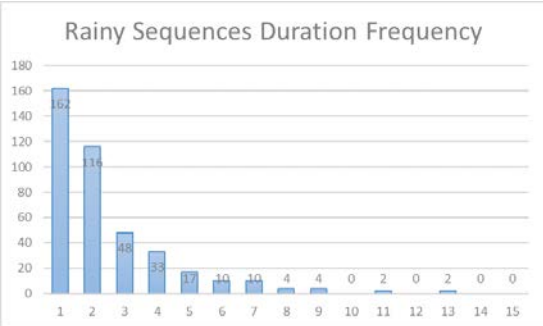
Looking at Figure 4.2, we find that:

- Sequences of short duration are the most frequent, especially the 1-day rainy sequence. Globally, the number of 1-day rainy sequences accounts for 40.2% of the total number of rainy sequences;
- The long sequences with duration >7 days accounts for only 3% of the total number of sequences;
- The counties in the centre of the United States like Sheridan and Franklin have similar rain sequence distributions with possible values of consecutive rainy days varying from 1 to 13 during 2001-2007;

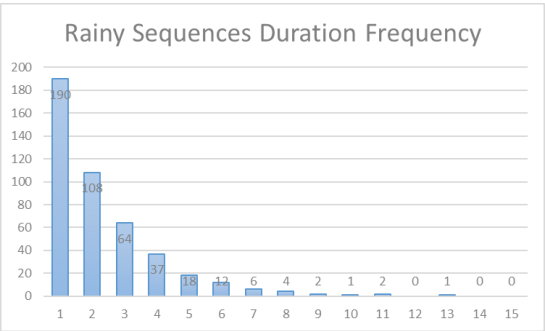
- Sequences with extremely long duration (>20 days) are easier to be found in the south-west counties, such as 24 days in Cochise and 19 days in Union.



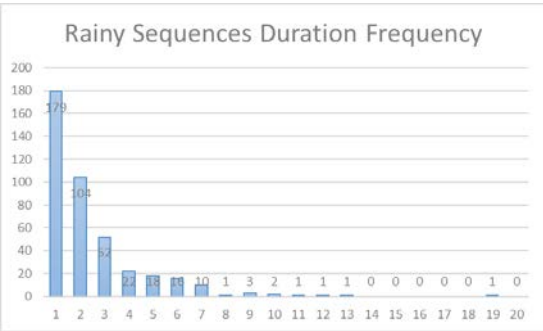
(a) Rainy sequences duration in Cochise



(b) Rainy sequences duration in Sheridan



(c) Rainy sequences duration in Franklin



(d) Rainy sequences duration in Union

Fig. 4.2 Rainy sequences duration in the four counties selected during 2001-2007

We can not draw definite conclusions, given such a small number of counties at our disposal. However, we can still advance the following hypothesis, which must be further supported by a study based on a more dense network of observation stations: by moving away from the sea, the frequency of rainy sequences with short duration becomes higher.

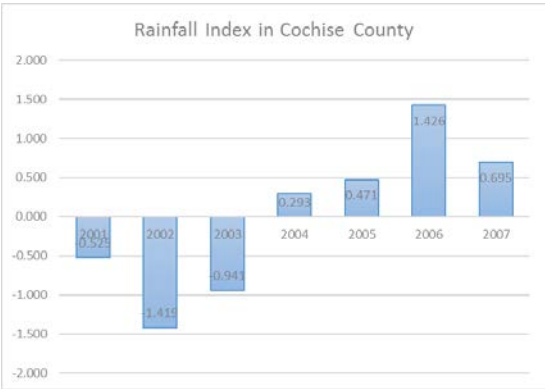
4.3 Inter-annual variability analysis

In this section, we intend to study the variability of the cumulative precipitation per year. The Rainfall Index (R_{ij}) is a commonly used criterion to present the precipitation variability. It corresponds to the ratio of the difference between the annual precipitation height at station i and the average annual precipitation height with the standard deviation. In particular, the annual rainfall index for station i at year j is defined by the following formula proposed by Lamb (1982):

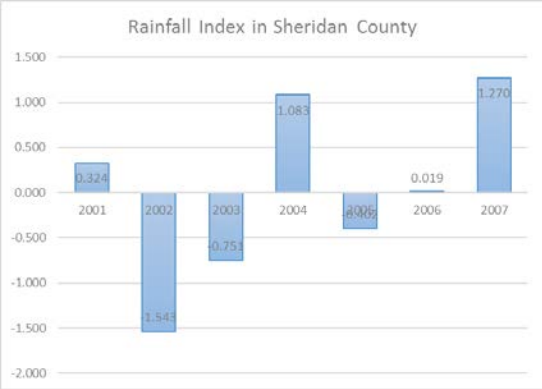
$$R_{ij} = \frac{(x_{ij} - \bar{x}_i)}{\sigma_i}, \tag{4.1}$$

where x_{ij} is the total rainfall height for station i at year j; $\bar{x}_i$ is the annual average rainfall at station i during the period of registration, and σ_i corresponds to the standard deviation of the annual rainfall

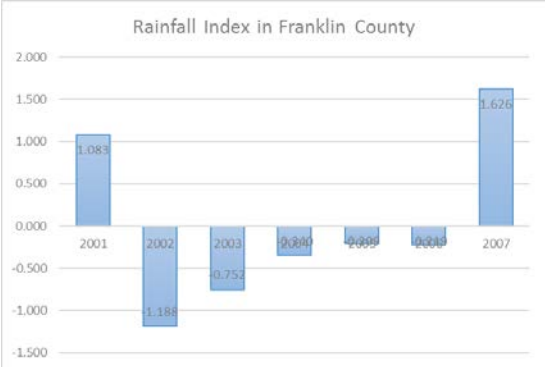
among different stations. The values of the rainfall index for the selected four counties during 2001-2007 are shown in Figure 4.3.



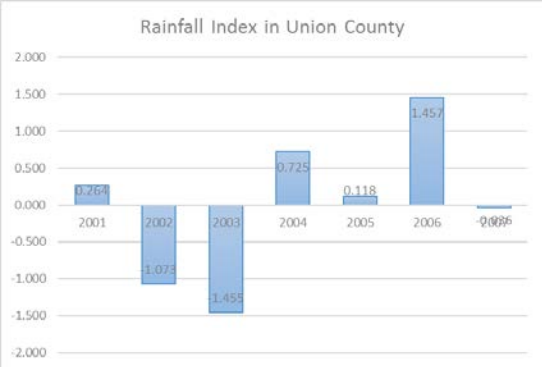
(a) Rainfall index in Cochise



(b) Rainfall index in Sheridan



(c) Rainfall index in Franklin



(d) Rainfall index in Union

Fig. 4.3 Rainfall index in the four counties selected during 2001-2007

The application of the rainfall index to our daily precipitation gives the results indicated on the Figure 4.3. Some remarks could be made as follows:

- The inter-annual variability in the cumulative precipitation per year throughout the selected region is considerable: in some years, the cumulative precipitation exceeds 140% of standard deviation of the cumulative precipitation during this period (2002 and 2006 in Cochise County; 2002 in Sheridan County; 2007 in Franklin County; 2003 and 2006 in Union County);

- The inter-annual variability in the cumulative precipitation per year throughout selected region does not always move in the same direction; in 2006, the Cochise and the Union County recorded much more cumulative precipitation than their average while the Sheridan and Franklin County had no great change.

- In 2002, all four Counties suffered a great decline in cumulative precipitation.

The previous analysis is based on the fact that the daily precipitation is one of the most important meteorological variables which introduce the highest amount of variability. Some basic descriptive statistics have also been computed to indicate the existence of inter-regional and inter-annual variability. However, when faced with questions of the type "How about the variability of the other climatic variables?", "How does it work when taking all the variables into consideration?", "The inter-regional and inter-annual variability, which one is more important?", It is difficult to find out a solution with traditional statistics based on a relatively limited amount of data. That's why in the following section, an unsupervised clustering method will be introduced to study the whole meteorological records with all the variables included.

4.4 Unsupervised Clustering of Meteorological Records

In a nutshell, clustering is the task of grouping together a set of objects or more largely data in an unsupervised manner, so that objects of the same group (called a cluster) are closer together (in the sense of a selected criterion of similarity) to each other than those of the other groups (clusters). This is the main task in the exploratory data mining, and a statistical data analysis technique widely used in many fields, including machine learning, pattern recognition, signal and image processing, etc. The idea is to discover groups within the data, automatically (Jain et al., 1999). In the following, one of the most popular clustering techniques, the K-means clustering, will be introduced for its simplicity and its ability to handle large datasets (Kogan, 2007).

4.4.1 K-means Algorithm

The k-means algorithm is one of the simplest unsupervised learning algorithms developed by McQueen in 1967 (MacQueen et al., 1967). It is also called the mobile centre algorithm because it assigns each point in a cluster whose centre (centroid) is the closest. Generally, the centre is taken by the average of all the points in the group, which means that their coordinates are the arithmetic mean for each dimension separately from all the points in the group. As a result, each group is represented by its centroids (Likas et al., 2003). The basic k-means algorithm is listed as Algorithm 2.

As an effective clustering algorithm, k-means has been widely studied and applied in many academic and industrial areas, such as segmentation of the market by discovering distinct customer groups from purchasing databases in marketing research (Punj and Stewart, 1983); identification of similar terrestrial areas with geographical and climatic database in environmental research; identification of separate insured groups associated with a large number of returns in finance (Huang, 1998).

The main limitation of this method is the dependence of the results on the initial values (initial centres). Each initialisation corresponds to a different solution (local optimum) which may in some

Algorithm 2 Basic Kmeans Algorithm

```
Randomly initialize k point as k clusters D = {D_1, D_2, ..., D_k} with centers C = {C_1, C_2, ..., C_k}
while Termination condition not reached do
    for Each point X_i do
        (Re) calculate its distances to the cluster centers {dist(X_i,C_1), dist(X_i,C_2),..., dist(X_i,C_k)}
        (Re) assign the point X_i to new cluster D_j when dist(X_i,C_j) is minimal
    end for
    for Each each cluster D_i do
        Update the coordinates of the cluster center C_i
    end for
end while
```

cases be very far from the optimal solution (global optimum). Nevertheless, many improvements have been proposed in the literature, including convergence issues (Li et al., 2015). The rate of convergence is a critical issue, especially with large datasets, and the k-means clustering is particularly adapted to such cases.

However, in the context of unsupervised clustering, it is natural to question the validity of the obtained results. What is the optimal number of clusters? Do the discovered groups correspond to our a priori knowledge? Do they match the set of objects that we have? Which two classifications are the most relevant?

4.4.2 Clustering validation

The k-means is a partitioning-based clustering method which depends crucially on the number of clusters. This number is generally left to the user, but a clever choice could significantly improve the performance of the algorithm. Several methods have been proposed for determining an appropriate number, mostly based on two types of selection criteria, called elbow and silhouette.

The elbow criteria

The elbow method is one of the oldest methods for determining the optimal number of clusters in a dataset. In this method, the sum of inter-cluster variance is written as a function of the number of clusters: Some clusters must be chosen so that the addition of another group does not give much better modelling of the data. More specifically, if we plot the sum of inter-cluster variances against the number of clusters, the first added cluster will generally decrease the sum of inter-cluster variances by a large amount, but at a certain point, the marginal gain will slow down, which gives a graphical angle. The number of clusters is chosen at this stage, hence the "elbow criterion" (Peeples, 2011). The total variance explained by inter-cluster variability is given as a function of the total number of clusters k in the form

$$was(k) = \sum_{i=1}^{k} \frac{1}{n_i} D_i, \tag{4.2}$$

where n_i is the size of the i-th group and D_i the sum of distances between the points of the group i.

The silhouette criteria

The silhouette criteria correspond to indices which take into account both inter-cluster and intra-cluster variability. Each point i in the dataset is associated with a silhouette value given by

$$S(i) = \frac{b_i - a_i}{\max\{a_i, b_i\}}, \tag{4.3}$$

where a_i represents the average distance between i and all the other points within the same cluster of i; b_i represents the minimum average distance between i and all the other points of a specific cluster different to that of i. That is why the silhouette criteria are also considered to be a comparison of compactness and separation of a clustering result. Finally, the average width of the silhouette will be used to identify the optimal cluster number, that is

$$S = \frac{1}{k}\sum_{i=1}^{k} S(i), \tag{4.4}$$

where k is the number of clusters. It can easily be deduced that $S \in [-1, 1]$ and the nearer to 1, the better clustering is achieved.

4.4.3 Clustering result of meteorological records

As presented in the chapter, the dataset for the research of corn crop is a table with 720 rows and 1641 columns, containing all the climatic variable. Each column has been normalised to standard normal distribution to eliminate the influence of scale. The clustering results are as follows.

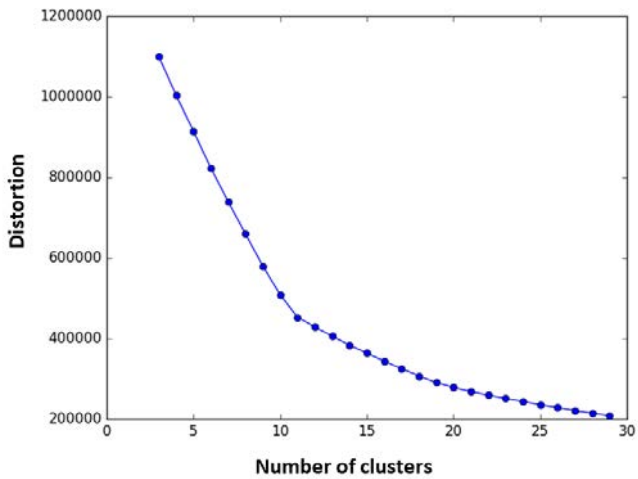


Fig. 4.4 Elbow criteria in meteorological records clustering analysis

Results of the elbow method

We describe here the application of the previous methods to our dataset consisting of the climatic data described above. In particular, we applied the k-means clustering method to determine the optimal number of clusters, with k varying from 2 to 30. The results that we obtained with the elbow criteria are summarised in Figure 4.4. Notice that in this case, the optimal number of clusters is k = 11.

Results of the silhouette method

Additionally, to the previous study, we applied the k-means clustering with silhouette criteria. The results are shown in Figure 4.5.

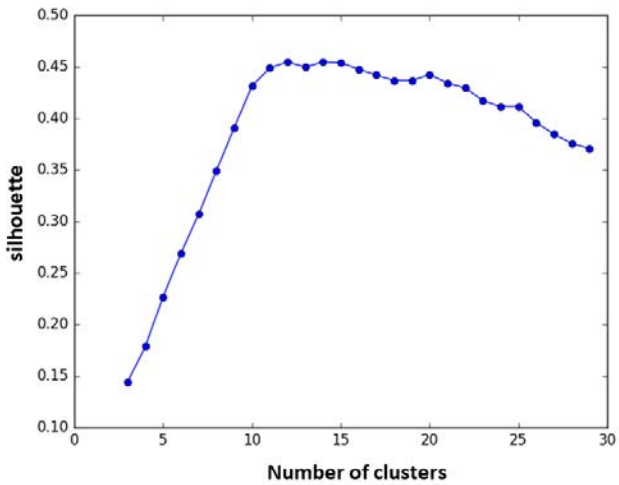


Fig. 4.5 Silhouette criteria in meteorological records clustering analysis

Notice that with the silhouette method the silhouette values obtained from k between 11 and 16 differ only slightly and seem to be good solutions for the number of clusters, so $k = 11$ appears to be good with both criteria.

Clustering result explication

The detailed clustering results for each scenario are listed in Appendix C. The scenarios which originate from the centre of the United States are classified into ten different clusters corresponding to different years. The other scenarios from California and Arizona, far from the centre, are classified as a single cluster.

With these impressive results, it can be concluded that the k-means works pretty well in clustering the meteorological records. It distinguishes very well the inter-regional variance and inter-annual variance of the meteorological records. In this research, the inter-annual variation is more important than the inter-regional variation. This clustering result makes it possible to carry

out cluster-based cross-validation, to verify the influence of inter-annual variation on the prediction capacity of the plant growth model in the next section.

4.5 Clustering-based Cross-validation

In this part, the parametrisation of CORNFLO model based on the results of k-means clustering, which means that the samples are equally drawn from different clusters. To compare with the results obtained in Section 3.6, each estimation takes 27 (300/11), 36 (400/11), 45 (500/11) from each cluster. This process is repeated for 100 times, and the results are listed in the following tables.

According to the results presented in Table 4.3 and Table 4.4, if the MAP estimators are selected, then the results obtained with MSPE based on well-chosen samples are closer to their "true value". Their distributions are more likely to be uni-modal with a single peak, as shown in Figure 4.6.

On the other hand, as for the result of fitness and prediction, fitness has been decreased while the forecast doesn't change a lot. It could be explained that the training set becomes more similar when a cluster-wise strategy is applied. And the testing scenarios have never appeared during the parametrisation process. In other words, they are the results when the model is used for an entirely new scene. It could be considered as the lower limit for this methodology.

Table 4.3 Estimation of five parameters with different sample size based on the clustering analysis

sample size	A2	F1	M0	M3	RUE
300	13.9157	708.1574	876.6731	1453.391	3.3143
400	14.0123	707.9132	868.7222	1462.214	3.3242
500	13.9189	707.3595	869.3464	1453.617	3.3253

Table 4.4 Standard deviation of the five parameters with different sample sizes

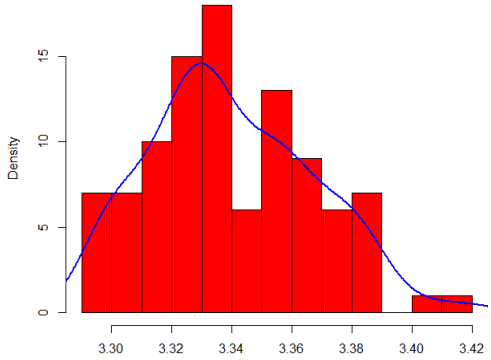
sample size	A2	F1	M0	M3	RUE
300	0.6172	20.4081	37.0013	57.7638	0.0556
400	0.3662	18.5054	19.084	35.0282	0.0346
500	0.3572	14.7435	15.3157	33.4155	0.0292

Table 4.5 Cluster-based Fitness and prediction capacity evaluation

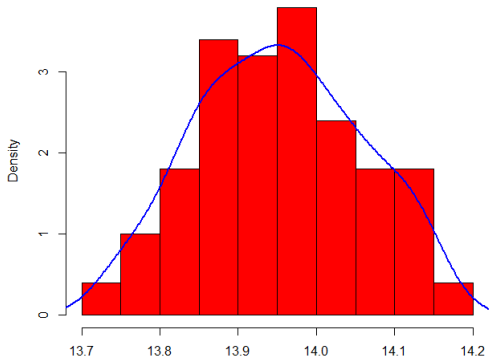
sample size	RMSE0	RMSE	MARE0	MARE
300	70.4	72.6	6.01%	6.23%
400	70.3	72.5	5.93%	6.18%
500	70.2	72.4	5.87%	6.12%

4.6 Conclusion

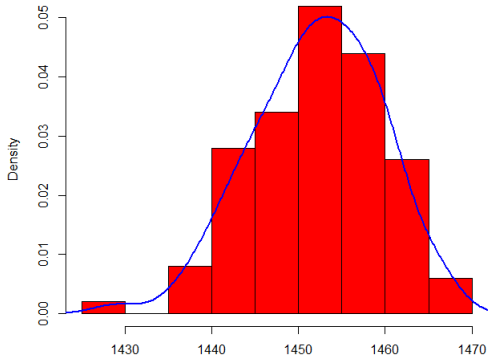
In this chapter, another important subject when dealing with models taking into account the meteorological records, the inter-annual variability, has been discussed. Descriptive statistics on the evolution of meteorological variables, the rainfall, for example, has been discussed in Section 4.2 and 4.3. In Section 4.4, an unsupervised clustering method, the k-means clustering algorithm, is applied to the meteorological records. According to the clustering results, almost all the records of the same year are classified into the same group, which shows that the interannual variance of the meteorological records is more important than the inter-regional variance. Finally, cross-validation taking into account the inter-annual variance has been implemented to study its influence on the CYP prediction. And the MuScPE based on clustering analysis outperforms both on the accuracy and the uncertainty in comparison with the initial version in the last chapter for the parameter estimation. The evaluation for fitness and prediction has also been more or less improved as in Section 4.5.



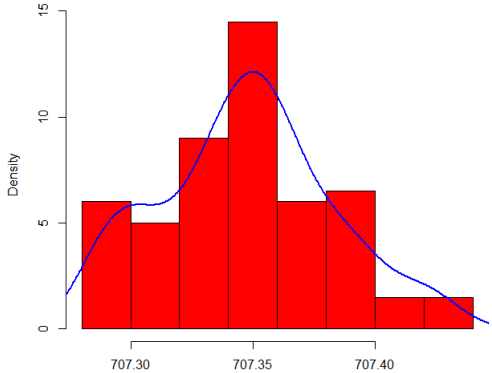
(a) RUE



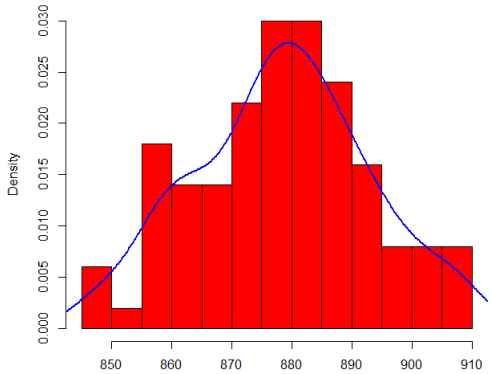
(b) A2



(c) M3



(d) F1



(e) M0

Fig. 4.6 Distribution of five parameters based on the clustering analysis with 50 samples from each cluster

Part II

Data-driven Models for Crop Yield Prediction

Chapter 5

Crop Yield Production with Statistical Learning Methods

As stated in Chapter 1, dataset in the form of cross-sectional data $\{U_i, y_i\}$, where $\{U_i\}$ refers to the meteorological records and $\{y_i\}$ to the corn yield observations, makes it possible to construct a predictive model with data-driven methods for CYP. Another term usually employed to describe data-driven methods is "statistical learning methods", as in (Friedman et al., 2001). Over the last twenty years, with the developments in computer science and statistics, statistical learning has been developed greatly and has become an important subject of modern data analysis. It has been applied to solve problems in many fields, such as physics, chemistry, biology, economics, etc. Before carrying out the modelling process, the context of statistical learning will be introduced in Section 5.1; some particular statistical learning methods for regression will be detailed in Section 5.2; to choose a good setting for each method, model selection criteria will be presented in Section 5.3, as well as the criteria, to evaluate their predictive efficiency; finally, the prediction results for each method are presented in Section 5.4 followed by a brief conclusion.

5.1 Context of data science

5.1.1 A brief history of data analysis

Nowadays, it is said that data analysis plays a vital role prevalent in almost every scientific area. It aims mainly at concluding the dependencies among the observed variables by introducing models with predictive capacity. "Statistics", "data mining", "statistical learning", and nowadays "data science", are not distinct scientific domains, but only variants of the same methodology which evolves with the explosion of data volume and diversity by demanding a combination of skills, such as computer science and mathematics (John Walker, 2014).

From the 1930s to 1970s, a statistical question is associated with an experimentally refutable hypothesis H_0. A planned experiment is typically carried out with a representative sample of hundreds of individuals observed on p variables with $p \approx 10$. Then, with a Gaussian linear model assumed to be accurate, the hypothetical test will be performed so that the statistician can decide if H_0 should be rejected or not according to a controlled risk (usually 5%) (Durbin and Watson, 1950). When it comes to 1970s, the first computer tools became widespread. A systematic data analysis methodology was proposed to deal with more complex problems than the linear model. The scale of datasets became more massive thanks to the information technology. In the 1980s, the notion of Artificial Intelligence (AI) supplanted by the learning of neural networks was first proposed (Newell, 1982). Some new models, like non-parametric or functional models, are introduced into the traditional statistics. A decade later, the notion " data mining" brought significant change to the data analysis. The data is previously acquired and based in warehouses for the general purposes of the users (He, 2009). The software integrated with different modular in the same environment, such as database management, exploratory techniques, and statistical modelling, became popular in use, such as Matlab and SAS. At the same time, the emergence of machine learning makes AI a subset of statistical learning methods (Vapnik, 2013).

At the beginning of the new century, the development of biotechnology has facilitated and popularised the production of big data, particularly with recent sequencing techniques (Baldi and Brunak, 2001). As a result, a considerable increase in the variables number p is introduced while the sample size n for each biological sample remains modest. To analyse a problem with millions of variables for a few individuals is more indeterminate. The correction of the multiple tests made in (Benjamini and Hochberg, 1995) makes the statistical methods adapted to such kind of situations by the variable selection method. For example, in (Lê Cao et al., 2011), a penalty constraint in l_1-norm is taken into account to select the variable. In recent years, industrial applications, e-commerce, and other tools record everyday's life. In this stage, it is the number of individuals n that explodes. The usual test statistics lose their usefulness in favour of unsupervised or supervised learning methods. In the stage, the storage and computational efficiency become the main challenges in data science (John Walker, 2014).

5.1.2 Basics of Learning

What is a learning process ?

The term "learning", like that of a "neural network", naturally evokes the functioning of the brain. However, we should not expect to find here explanations on the mechanisms of information processing in the nervous systems; these are of high complexity, resulting from mental electrical and chemical processes, still poorly understood despite a large amount of experimental data available Zhang et al. (2017). While statistical learning methods can be beneficial for creating empirical

models of a particular function performed by the nervous system, the statistical learning methods described in this work deny any pretension to imitate, even vaguely, the functioning of the brain (Russell et al., 2003).

Nevertheless, from the point of view in "result", these two functions seem to coincide in some way. One of the essential tasks of the brain is to transform information into knowledge: identifying the letters that constitute a text, assembling them into words and sentences, extracting meaning from them, are activities that seem natural to us once the necessary learning has been accomplished. One of the essential tasks of the brain is to transform information into knowledge, such as identifying the letters that constitute a text, assembling them into words and sentences, extracting meaning from them. These activities can be easily achieved if the "learning" of the brain is accomplished (Kandel et al., 2000). Similarly, in statistical learning, with the algorithms implemented, the goal is to imitate the capacity of living beings to learn by example so that it can reproduce a similar response as the learned examples but also be able to generalise a correct result in a new situation (Vapnik, 2013).

Application of learning

In data science, it is easy to come across problems, such as: identify handwritten code numbers from digitised images (LeCun et al., 1995); identify the aggravating factors of certain types of cancer according to clinical and demographic variables (Cruz and Wishart, 2006); search for genes potentially involved in a disease from sequencing data; more generally, biomarkers for early diagnosis; predict an air pollution rate according to meteorological conditions; establish appetence or attrition scores in customer relationship management; build meta-models or models to replace a numerical code that is too complex to analyse the sensitivity to the parameters; to detect or better predict the failures of a process, etc. They are all examples from different areas of our daily life. However, their solutions have something in common, that is, to minimise a forecast or a risk.

At the same time, the methods and algorithms derived from Artificial Intelligence also become a part of statistical learning methods. A more detailed and systematic introduction of statistical learning can be found in (Vapnik, 2013) and (Friedman et al., 2001).

These objectives can be classified into four axes:

- to explore or verify, represent, describe, variables, their relationships and position sample observations,
- to explain or test the influence of a variable or factor in a model assumed to be known a priori,
- to predict and select a better set of predictors, for example in the search for biomarkers,
- to provide a possible better "black box" without the need for explicit interpretation.

Off course, it is not forced that all the analysis process should follow this framework. Some more requirements would be made, such as variable selection, model explication, calculation efficiency, etc.

5.1.3 Important notions in statistical learning

Supervised and Unsupervised learning problems

In statistical learning, supervised and unsupervised learning problems are distinguished by the presence of an objective variable Y to be explained together with X, the corresponding explainable variables that will be used to infer the objective variable (Friedman et al., 2001). In the former case, it is called a supervised learning problem, to which the objective is to find a function $\hat{f}$, according to a predefined criterion, to reproduce Y from the observed matrix of features X:

$$Y = \hat{f}(X) + \varepsilon, \tag{5.1}$$

where $\varepsilon \sim N(0, \sigma^2)$ represents the noise or measurement error.

In the opposite case, in the absence of an objective variable Y to be explained, it is then called unsupervised learning, just as in the application of the k-means with meteorological records in Section 4.4. The general objective is to search for a typology, according to which the observations are grouped into homogeneous but most dissimilar classes.

Supervised Problem

Normally, in a supervised problem, two types of variables $\mathbf{X}$ and $\mathbf{Y}$ are provided, with X the explanatory variables (features) and Y the objective variables. Each example of observation is in the form of a pair $(\mathbf{x}_i, \mathrm{y}_i)$. A sample is a finite set of examples $D = \{(\mathbf{x}_i, \mathrm{y}_i) \in \mathbf{X} \otimes \mathbf{Y}\}_{i=1}^{n}$.

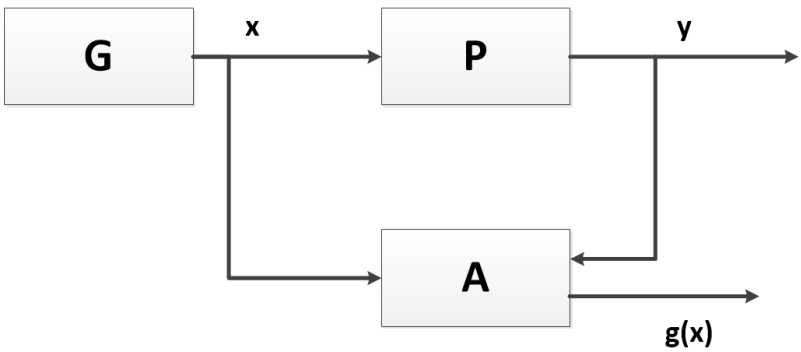


Fig. 5.1 Illustration of supervised learning process

As stated in (Vapnik, 2013), a supervised learning process can be illustrated in Figure 5.1. During the learning process, the learning machine supervises the pairs (x, y). After learning, it should be able to reproduce a value $\hat{y} = g(x, \theta)$ for any x. The goal is not only to produce the

response as close as possible as the "true" model $f(\cdot)$ does in D, but also generalise the response to other different D. For all the learning problem, the component S generates independent and identically distributed (i.i.d.) random vectors $\mathbf{x} \in \mathbb{R}^p$ according to the distribution function $F(\mathbf{x})$; then a "true" model $f(\cdot)$ produces the output $\mathrm{y} = f(\mathbf{x})$; under the supervision of $f(\cdot)$, a learning machine $g(\cdot)$ tries to produce an output $g(\mathrm{x}, \theta)$, where θ represents the parameter vector of the model $g(\cdot)$.

Discrimination or regression

In a supervised problem, the objective is to build, from this learning sample, a model f, which will allow us to predict the output Y associated with a new input (or predictor) X. The output Y can be quantitative (the price of a stock, electrical consumption, pollution map, ...) or qualitative (occurrence of cancer, recognition of figures, ...). Some methods of learning or modelling adapt to all types of explanatory variables, while others are specialised. If the objective variable Y is qualitative or categorical, the problem is qualified as a classification one. Otherwise, it is called a regression problem (Vapnik, 2013).

Estimation and Learning

In some cases, the terms "estimation" and "learning" are used as synonyms, but there are still some nuances. In a traditional statistical problem, the central objective is the construction of a model with a good explanatory capacity. This is judged by the degree to which the model fits the data (model fitting) and interpreted as an effort to approximate the real underlying mechanisms.

On the other hand, when the objective concerns the predictive capacity, it appears that the best model is not necessarily the one that best fits real observations. The main idea is to have a model that can balance well fitness and prediction, as explained in (Vapnik, 2013).

5.1.4 Data Analysis Framework

The main motivation of data analysis is to evaluate the data by searching for relevant information that helps in making a decision. In (Frawley et al., 1992), the data analysis process is divided into different stages, as follows:

1. Understand the application context, the research objectives and take into account the a priori knowledge;
2. Create a targeted subset of the data (matrix) from different data resources;
3. Clean the errors and treat missing data, outliers, etc.
4. Transform the data with "normalisation", linearisation, etc.

5. Explain the purpose and strategy of data analysis, such as exploration, association, classification, discrimination, etc.
6. Choose the right methods and algorithms according to their interpretation or predictability;
7. Make tests on the testing set with appropriately defined criteria such as quality of adjustment, prediction, simplicity, etc.
8. Dissemination of results for decision making.

5.2 Methods Description

Today, it is said that data analysis plays a vital role that wins out in almost every science area. It makes it possible to conclude the dependencies among the observed variables, thus introducing a model with predictive capacity (John Walker, 2014). The problem is particularly crucial in biology where data are expensive. One of the objectives of this thesis is to construct models for CYP with meteorological records, which turns out to be a multivariate regression problem. In dealing with the complex relationship between the explanatory variables and the objective variable, even among the explanatory variables, several regression methods are recommended in (Ryan, 2008). According to their different modelling discipline, the statistical learning methods can be classified into traditional statistical methods and machine learning methods (Iniesta et al., 2016). With the traditional methods, such as lasso or ridge regression, PLS or PCA regression, the complex relationship is simplified by dimension reduction and penalisation; while the machine learning methods are based on the individuals' similarity and the resampling technique. In this research, regression tree, bagging, boosting, random forest, k-NN, and neural network regression, some of which are considered to be ten of the most important data mining algorithms (Wu et al., 2008), are selected as candidate methods to realise the CYP. Some of the modelling and forecasting methods used in this work are well known and have been widely used. The traditional methods are not recalled here. However, machine learning methods result from the interface between statistics and the theory of learning. They deserve some introductory words.

Regression Tree

It has been a long time since the proposition of a regression tree, but it is still trendy, especially in marketing applications (Friedman et al., 2001). It leads to the construction of binary decision trees straightforward to interpret. On the other hand, it is often the basic model of aggregation algorithms. As shown in Figure 5.2, a classification tree from an example of (Therneau et al., 2015) is built recursively, where each node is defined by a quantitative explanatory variable and a threshold value in regression. This choice is made by optimising a criterion that aims to generate the most homogeneous leaves in terms of the external variable, such as the inter variance of the

objective variable. Cross-Validation is always carried out the overcome the instability of a single tree model to ensure its predictive capacity (Prasad et al., 2006b).

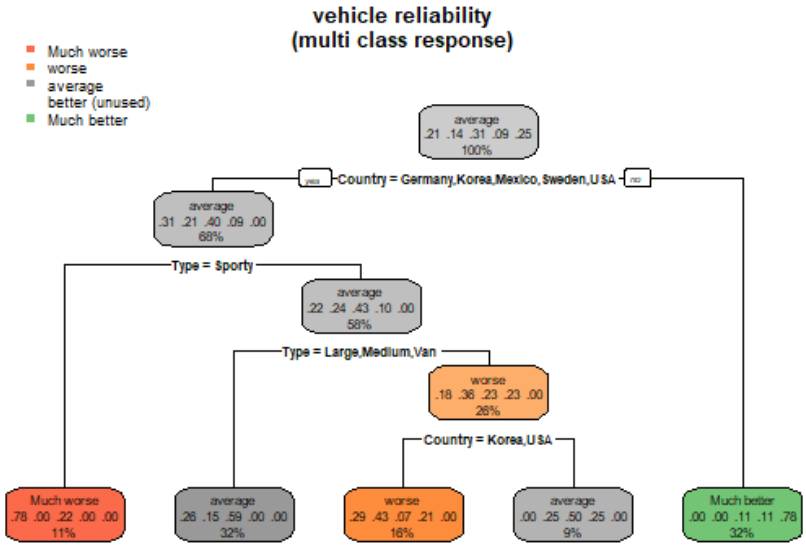


Fig. 5.2 An example of regression tree

Bagging regression

Ensemble learning algorithms is a simple and efficient solution to solve the instability problem caused by a single tree model, as shown in Figure 5.2 (Dietterich et al., 2002). The Bagging technique, which is one of the most used ensembles learning algorithms, is proposed in (Breiman, 1996). This algorithm consists of two crucial steps: bootstrap and aggregating. A sampling process with putting-back is firstly conducted to generates the subsets of the dataset for model training at the "bootstrap" step. Then, the prediction results should be aggregated at the "aggregating", where a quantitative result is often averaged while it is decided by voting for the qualitative variable. Ideally, if we have m independent samples, according to the law of large numbers, the variance of the averaged model should be divided by $\sqrt{m}$. Thus, the instability of the trees can make the whole quite efficient: each tree is of significant bias while their average is of low variance. In practice, the large number of observations is generated by bootstrap, even though "new" samples are not independent enough.

Random forest and Boosting

It has been mentioned in the above section that the instability of a single tree could be overcome with ensemble learning and bootstrap. However, since the "new" samples depend on each other,

which results in the dependency of the generated trees. Several solutions have been proposed subsequently for the improvement of bagging algorithm (Liaw et al., 2002).

Random forest is one of the most famous data mining algorithm introduced in (Breiman, 2001). The main idea is to add additional randomness to allow differentiation and dependence reduction among tree model estimates. To be more in details, in the construction of a regression tree with ”new" samples, the optimal explainable variables, and their associated thresholds are not selected among all the explanatory variables, but on a randomly chosen subset. Even though this results in a suboptimal tree, in practice, model aggregation leads to better overall results.

Another solution to the similarity of generated trees is named Boosting (Freund et al., 1999). It is considered as an enhanced algorithm, where a weak classifier trained into a robust classifier iteratively. For this algorithm, the observations have gained weight to reflect their contribution to the bias of the generated trees. At each iteration, each new model gives more substantial weight to the poorly predicted observations at the previous iteration.

It is important to note that both methods in this section can lead to the estimation of a large number of parameters without leading to over-fitting. The main disadvantage lies in the computational efficiency since a significant amount of trees should be generated to make a "forest".

SVM regression

Support Vector Machine (SVM) is a crucial algorithm that could be applied in the regression and classification problems. More and more attention is paid to this algorithm, since its success in multiple research fields and its strong theoretical background. (Cortes and Vapnik, 1995) and (Cristianini et al., 2000) are two important references for the basic theory of SVM. (Schölkopf et al., 2002) and (Smola and Schölkopf, 2004) could serve as a guideline that outlines the basic idea of support vector machines for regression.

The first SVM was firstly applied in a binary classification problem where it is required to separate observations with p quantitative variables into two groups. To map the original data into a new high-dimensional space, where it is possible to apply a linear model to obtain a hyperplane for separation new high-dimensional space as in Figure 5.3.

The basic idea is to search for a linear hyperplane separating the two classes if it exists. The optimal hyperplane is the one that maximises the margin so that the individuals from different classes are as distant as possible from each other.

However, in practice, such hyperplane separation is difficult or even impossible when it doesn’t exist. Consequently, a transformed problem is created by adding a penalty, which allows the misclassification of observations depending on the value of the parameter. Moreover, the search for a non-linear operator F can be made linear by embedding the problem into a space of larger dimension H, equipped with a scalar product which is defined by a positive bilinear kernel function.

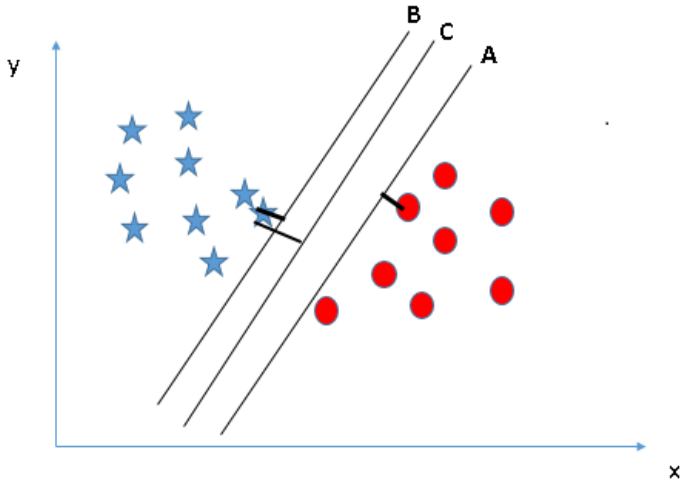


Fig. 5.3 An example of Support Vector Machine

It leads to the fact that the non-linear function F is not required to be explained explicitly, which could be regarded as the smartest part of this approach (Vapnik, 2013).

Neuronal models

A neural network (Hagan et al., 1996) is a parallel and distributed information processing structure, consisting of computing units (neurons) connected by valued links as in Figure 5.4. The state of a neuron depends on the value of the signals arriving at the unit and the contents of the local memory attached to that unit. Since it is collective, the behaviour of the network is largely regulated by its connectivity. It is also by the non-linearity of the interactions rather than by the individual properties of the neurons. However, as it is said in (Friedman et al., 2001), neural computation is historically inspired by the observation of natural systems, but it is not a biologically plausible model.

The modes of representation and information processing, as well as the topology, vary significantly from one neuronal model to another. The common denominator of these models is the process of determining synaptic weights, verbally termed learning. The learning process refers to the gradual, iterative process by which network weights are adjusted.

5.3 Choice of method

As presented in the above section, the CYP is carried out with several parametric approaches from both classical models and machine learning methods. An important topic that can not be ignored in modelling is the number of explanatory variables used for each method, such as the number of trees in random-forest or number of neurons in a hidden layer in the neural network. It is clear that

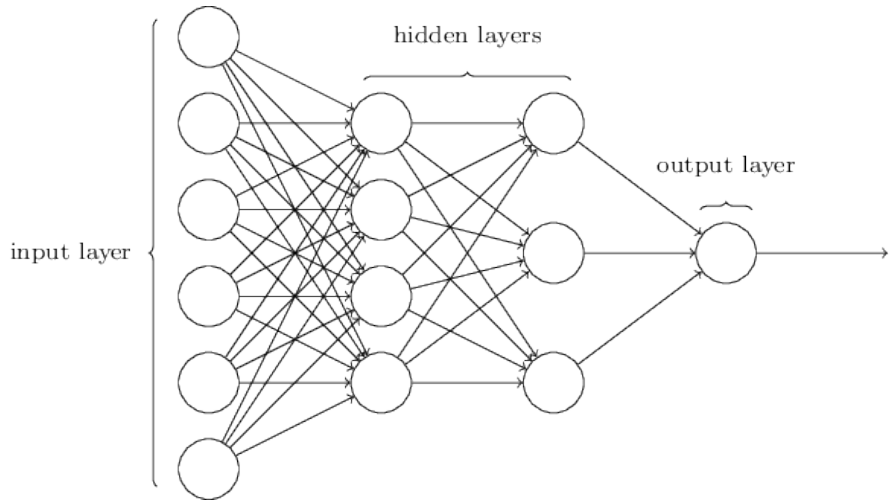


Fig. 5.4 Neural network structure

the more parameters are integrated, the more complex a model will be and the more flexible it can adjust to the observed data, thus the smaller the generated error of fitness will be. On the contrary, such a model may prove to be flawed when it comes to predicting with "new" data that did not participate in the estimation with so many parameters.

According to the variance decomposition in Equation 5.7, it is important to optimise the balance between bias and variance by controlling the number of variables in the model (its complexity) when minimising the risk. These remarks lead to the definition of criteria for model selection, such as **AIC**, **BIC**, **AICc**. These criteria were proposed in the 1970s and are widely used in model selection. Some sampling techniques, such as cross-validation can also help to balance the compromise.

5.3.1 Criteria of penalised likelihood

Let us assume that the likelihood function $L(\theta)$ of a parametric model resulted from a density function $g(y|x,\theta)$. Then, in a predictive problem, it corresponds to the conditional density function of y given x and under the parameter value θ. In this case, the parameter vector θ could be estimated by the maximum likelihood method.

The likelihood $L(\theta)$ can be interpreted as a way of measuring the fit of a model to the observed data since it gives a measure of how plausible the given data y is in terms of the parameter vector of the model. However, in most cases, the log-likelihood $\ln L(\theta)$ is used instead. In a linear regression problem, a list of possible models is available with $i = 1,2,\ldots,p$ explanatory variables (for a fixed i several combinations of competing covariates are possible). The relative log-likelihood can also be obtained easily $\ln L(\theta_i)$, by assuming that the remaining covariates are associated with null coefficients. However, the likelihood or the log-likelihood function could not serve as model

selection criterion since their values increase with i, which means that the "best" model is the one that has the most parameters.

Therefore, the AIC and BIC criteria will penalise the log-likelihood by taking into consideration the number of parameters. They appear to be very similar, but in fact, their objective is different.

The AIC criterion

The classical AIC criterion is expressed as follows:

$$AIC = -2\ln L(\hat{\theta}) + 2k, \tag{5.2}$$

where k is the number of model parameters and $\hat{\theta}$ the maximum likelihood estimator. The best model is, therefore, the one that minimises AIC.

The AIC (Sakamoto et al., 1986) was originally proposed from a Kullback-Leibler divergence perspective. Let f, g be two probability density functions, and assume that f is the "true" unknown density and g an approximate density of an observed vector y. Then, the loss function from using g instead of f, is defined as

$$l(f,g) = \int \ln\left(\frac{f(y)}{g(y)}\right) f(y)dy.$$

Notice that

$$l(f,g) = \mathbb{E}_f[\ln(f(Y))] - \mathbb{E}_f[\ln(g(Y))], \tag{5.3}$$

that is, the loss function corresponds to the expected difference of the log-likelihoods between the true and the approximate model, when computed under the true model. Since f is unknown, it is impossible to compute (5.3) exactly. Instead, approximations are possible and the most popular way results from maximum likelihood. In fact, since the minimisation of the loss function is equivalent to the maximisation of the term $\mathbb{E}_f[\ln(g(Y))]$ the approximations need to be done for this quantity.

The BIC criterion of Schwartz

The BIC criterion (Chen et al., 1998) is defined as follows:

$$BIC = -2\ln L(\hat{\theta}) + \ln(n)\,k. \tag{5.4}$$

The penalty is much higher than that of the AIC since the sample size is also taken into account. Consequently, when large samples are available, the BIC criterion will favour models with fewer parameters than the AIC.

The BIC criterion was inspired by a Bayesian point of view. In particular, let us assume that a finite list of models, denoted by M_i, and depending on a parameter vector θ_i are available. In a Bayesian perspective, a priori probabilities $P(M_i)$ for each model M_i, as well as a distribution for θ

given the model M_i are assigned. Thus, the posterior probability of the model M_i given the data y is proportional to $P(M_i)f(y|M_i)$.

If the prior distribution is uniform over all possible models, then the posterior probability of the model M_i is proportional to the conditional density:

$$f(y|M_i) = \int f(y|M_i, \theta_i) f(\theta_i|M_i) d\theta_i \tag{5.5}$$

Under some mild regularity conditions, it can be shown that the approximation could be obtained as follows (Friedman et al., 2001) :

$$\ln(f(y|M_i)) \sim \ln(f(y|\hat{\theta}, M_i)) - \frac{k}{2}\ln(n), \tag{5.6}$$

where $\ln(f(y|M_i))$ is the log-likelihood of the model M_i and the objective is to choose the model M_i that minimises the BIC criterion.

5.3.2 Empirical approach

The bias-variance compromise

A regression problem is generally represented in the form $y = f(x) + \varepsilon$ and the prediction problem can be solved via the response function $\hat{f}$. In particular, for a given condition $\mathbf{x}_0$ the prediction corresponds to $\hat{y_0} = \hat{f}(\mathbf{x}_0)$. The prediction error e_0 at $\mathbf{x}_0$ can thus be expressed by

$$e_0 = y_0 - \hat{y_0} = f(\mathbf{x}_0) - \hat{f}(\mathbf{x}_0) + \varepsilon_0$$

The above decomposition indicates that the prediction error has two random components, the one which is related to the estimation error of $f(\mathbf{x}_0)$ by $\hat{f}(\mathbf{x}_0)$ and the other which is related to the noise ε_0 associated with a new observation y_0. By assuming that the errors from different observations are uncorrelated (at least valid in a least squares approach), then the mean squared prediction error can thus be decomposed as

$$\mathbb{E}(y_0 - \hat{y}_0)^2 = \mathbb{E}[f(x_0) - \hat{f}(x_0)]^2 + \sigma^2 = \left(\mathbb{E}[\hat{f}(x_0) - f(x_0)]\right)^2 + \mathbb{V}(\hat{f}(x_0)) + \sigma^2, \tag{5.7}$$

where the first term of the right member in the above equation corresponds to the square of the bias related to model prediction, the second term to its associated variance and the last term to noise variance, which is irreducible.

The term bias is related here to the adjustment of the model to the training set, while the variance corresponds to the prediction of new data. The more complex the model is, the lower the bias will be, but the higher the variation will become.

Evaluation and model selection

Figure 5.5 shows that there is an optimal way to compromise between bias and variance if we are ready to assign equal importance to each one of them. A practical way to perform this task is via estimation of the prediction error. A classical way to estimate unbiased prediction errors consists in using different datasets to estimate the model and its prediction error. In particular, the prediction error is estimated with the dataset that does not participate in model training. Thus, when a large number of observations are available, the data will be divided into two subsets:

- the learning set which serves to estimate each model in competition;
- the test set which serves to estimate the predictive performance;

The error measurement with the training set cannot be used since it is biased. Nevertheless, it is essential to keep data that serve no other purpose than to evaluate the error. Thus a "good" model can be obtained according to the bias-variance compromise.

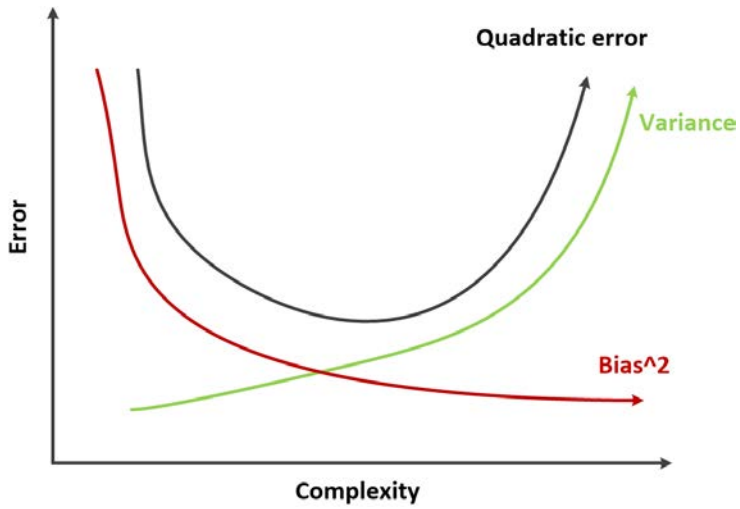


Fig. 5.5 The bias-variance compromise

Cross-validation

If the available dataset is not large enough, then cross-validation is a standard way to assess the prediction error as described in Algorithm 3. Firstly, the whole dataset is divided into k disjoint subsets of the same volume. Then the prediction error is evaluated on each subset after the model is calibrated with the other $k-1$ subsets. Finally, the predictive error is taken by averaging the errors for all the subsets. The choice of k is still a bias-variance trade-off: a large k will produce results with great variance and low bias, while a small k will increase the bias. In most cases, k is set between 5 and 10.

Generally, when dealing with problems in data science, the choice of methods and models, are always complicated to carry out.

Algorithm 3 k-folds cross-validation algorithm

```
Randomly distribute the dataset into k parts (k-fold) with approximately equal volume;
for i = 1 to k do
    Leave out one the i-th part;
    Estimate the model on the k - 1 remaining part;
    Calculate the error on each of the observations of the i-th part
end for
Take the average error of the prediction errors obtained with the cross-validation.
```

As stated in (Hall et al., 2009), the choice of an algorithm (model) is initially decided by the practical problems, which means the real data. In data science, the dataset under investigation is almost in a table structure, of which the critical parameters are the dimensions, that is, the sample size n and the number of variables p. The traditional statistical methods are typically designed for the case where $n > p$; the statistical learning offers a set of processes and algorithms that are effective when $n < p$. The strategies of model selection to deal with the crop yield prediction with meteorological records are the primary objective in this part. (Friedman et al., 2001) offer a rather exhaustive overview of the choice of algorithms in data science.

In this section, a wide variety of criteria and methods are proposed, and their implementation conditions are discussed. The attention is particularly drawn to the optimisation of model complexity. It is also an opportunity to recall that some consideration should still be paid to some robust and linear methods as well as some "old" strategies (descending, ascending, stepwise) or more recent (lasso) selection of linear models in academic or industrial practices.

Let $D_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a set of training data. The X_i are considered to be input variables taking values in a set X and Y_i are the output variables, taking values in a set Y. We call the $\{X_i\}$ features and the $\{Y_i\}$ labels. For simplicity, let us also assume that these data are independent and identically distributed (i.i.d.). Nevertheless, note that in practice, these assumptions are often violated. A learning rule corresponds to a function f that will be trained with the dataset D_n. Moreover, the constructed function $\hat{f}$ will be used to predict X, and Y where (X, Y) is a pair of test data. Some attention should be paid to distinguish the learning phase and the test phase.

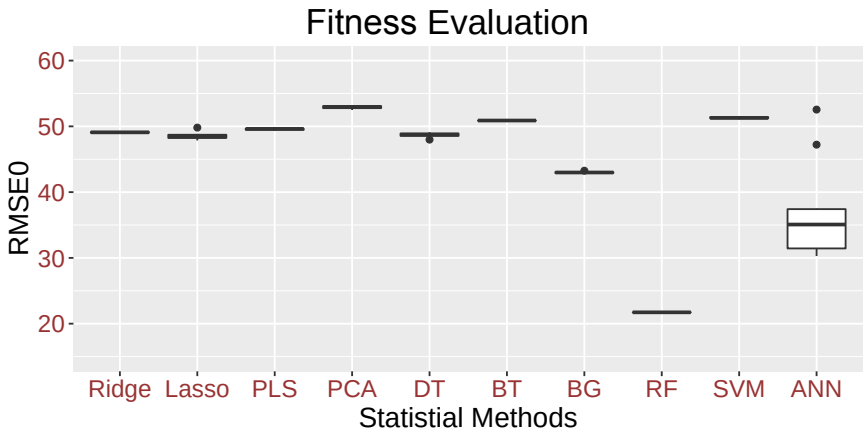
With the development of data mining, numerous articles compare and contrast techniques on public datasets and propose incremental improvements of specific algorithms. After a feverish period, when everyone tried to display the supremacy of their method, a conclusion is made that there is no "best" way. Each method is more or less well adapted to a specific problem, the nature of the data or/and the properties of the function f to be estimated. However, it is crucial to know how to compare methods to choose the most relevant in each situation.

5.4 Results and conclusion

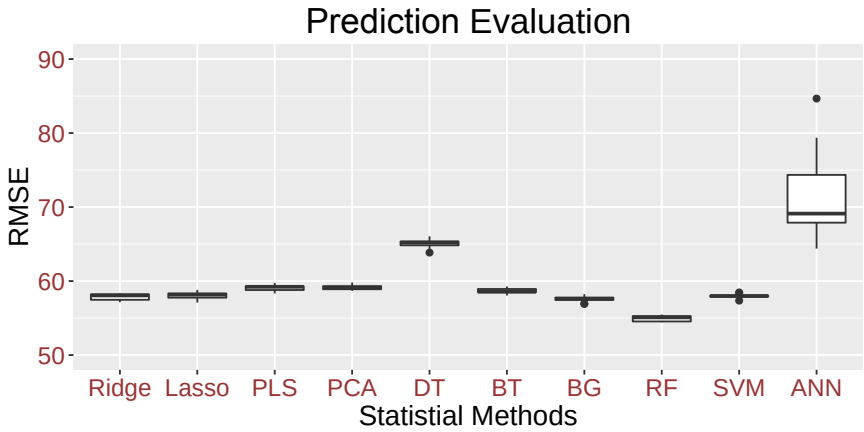
To compare the performance of different statistical learning approaches for CYP, an 11-folds cross-validation is applied to balance the variance-bias compromise. This process is repeated for ten times to test their robustness to the randomness in resampling. The average RMSE and MARE for fitness (RMSE0, MRE0) and prediction (RMSE0, MRE0) are listed in Table 5.1. The corresponding boxplots are shown in Figure 5.6 (RMSE0, RMSE) and Figure 5.7 (MRE0, MRE).

Table 5.1 Evaluation of statistical learning for CYP

	RMSE0	MRE0	RMSE	MRE
Ridge	49.09± 0.12	0.0406± 0.0001	57.84± 0.46	0.0475± 0.0003
Lasso	48.54± 0.54	0.0399± 0.0004	58.03± 0.52	0.0473± 0.0003
PLS	49.58± 0.14	0.0407± 0.0001	59.09± 0.48	0.0483± 0.0003
PCA	52.90± 0.25	0.0437± 0.0002	59.13± 0.33	0.0487± 0.0002
DT	48.68± 0.37	0.0401± 0.0004	65.09± 0.07	0.0528± 0.0006
BT	50.88± 0.02	0.0426± 0.0001	58.64± 0.39	0.0488± 0.0003
BG	42.98± 0.14	0.0354± 0.0001	57.59± 0.44	0.0471± 0.0003
RF	21.72± 0.02	0.0172± 0.0001	54.94± 0.40	0.0442± 0.0003
SVM	51.30± 0.14	0.0402± 0.0001	57.98± 0.30	0.0475± 0.0002
ANN	36.83± 7.50	0.0276± 0.0062	71.69± 6.37	0.0558± 0.0046

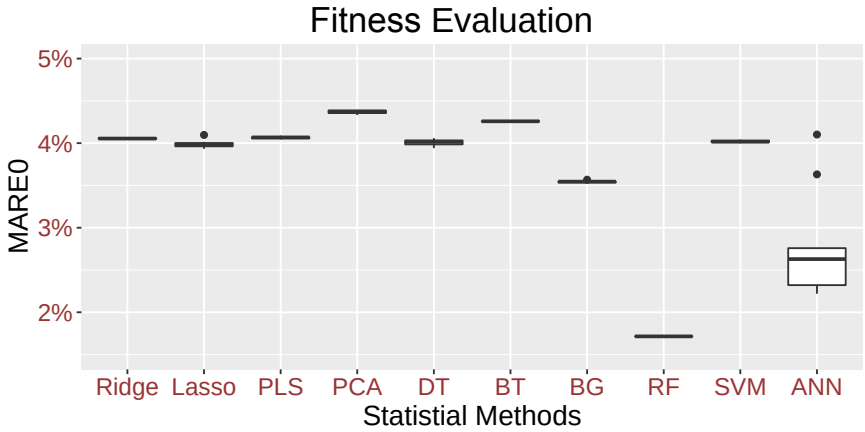


(a) RMSE0 for different statistical learning approaches

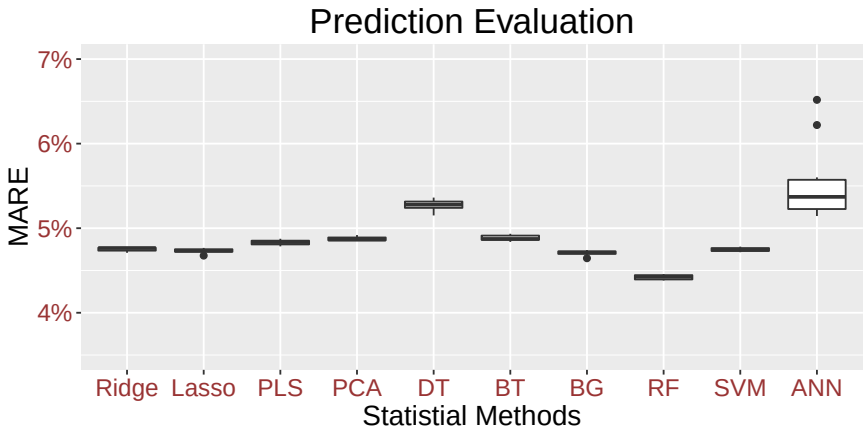


(b) RMSE for different statistical learning approaches

Fig. 5.6 Root mean square error of fitness and prediction for different statistical learning approaches: Ridge, Lasso, PLS (partial least squares regression), PCA (principal component regression), DT (Decision Tree regression), BT (Boosting), BG (Bagging), RF (random forest), SVM (support vector machine), ANN (artificial Neural Network)



(a) MARE0 for different statistical learning approaches



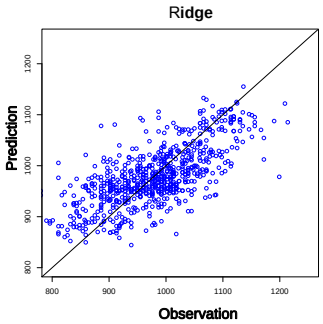
(b) MARE for different statistical learning approaches

Fig. 5.7 Mean absolute relative error of fitness and prediction for different statistical learning approaches: Ridge, Lasso, PLS (partial least squares regression), PCA (principal component regression), DT (Decision Tree regression), BT (Boosting), BG (Bagging), RF (random forest), SVM (support vector machine), ANN (artificial Neutral Network)

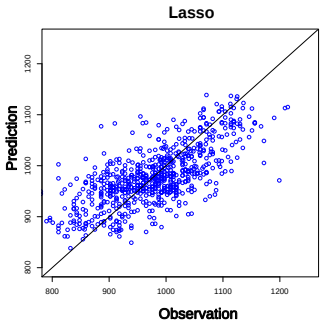
Note that all the statistical approaches have excellent performance both in fitness and prediction of CYP. The best method both in fitting and prediction is Random Forest. In particular, the best three results in fitness are Random Forest, Artificial Neural network and Bagging regression. As for the prediction results, the best ones are Random Forest, Bagging and SVM. The neural network does not bring significant improvements over the other methods, and its unique character lies in the fact that it exhibits the most significant variation in its evaluations. This result indicates that some extra effort is needed to find the best setting for a neural network to improve its capacity in CYP.

The results of the predicted yield as compared to the real observations are shown in Figure 5.8. The simple DT (Decision Tree) model provides only a finite set of values as described in Section 5.2. An improvement can be observed with a group of trees obtained by "Boosting", "Bagging" and "Random Forest". The results of the random forest in Figure 5.6 and Figure 5.7 illustrate this fact more clearly.

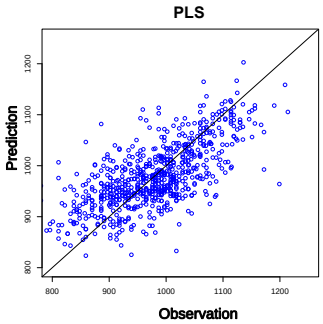
As far as the residuals of different models are concerned, the differences between the predicted and the observed yield, are plotted against the predicted values in Figure 5.9. The results indicate that the homoscedasticity of the errors is confirmed by all the algorithms, except for the neural network one.



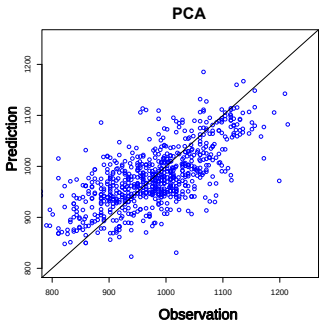
(a) Prediction with Ridge



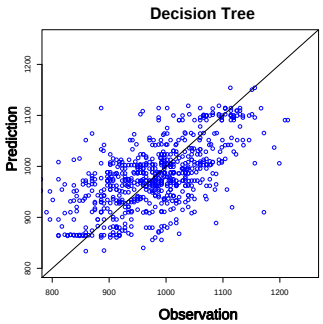
(b) Prediction with Lasso



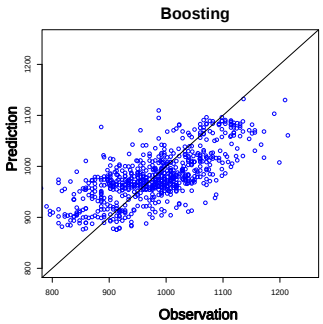
(c) Prediction with PLS



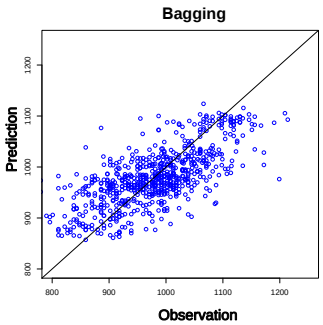
(d) Prediction with PCA



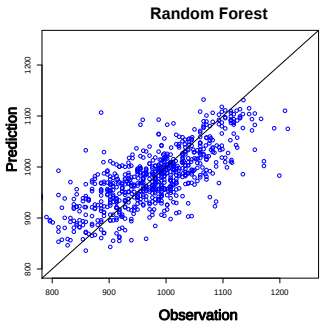
(e) Prediction with DT



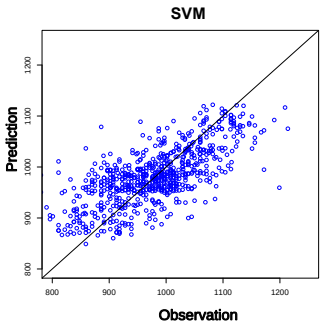
(f) Prediction with BT



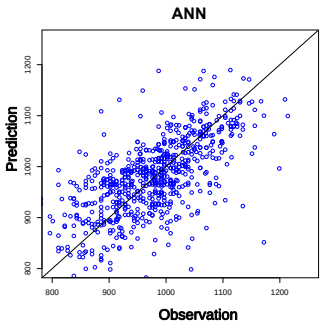
(g) Prediction with BG



(h) Prediction with RF

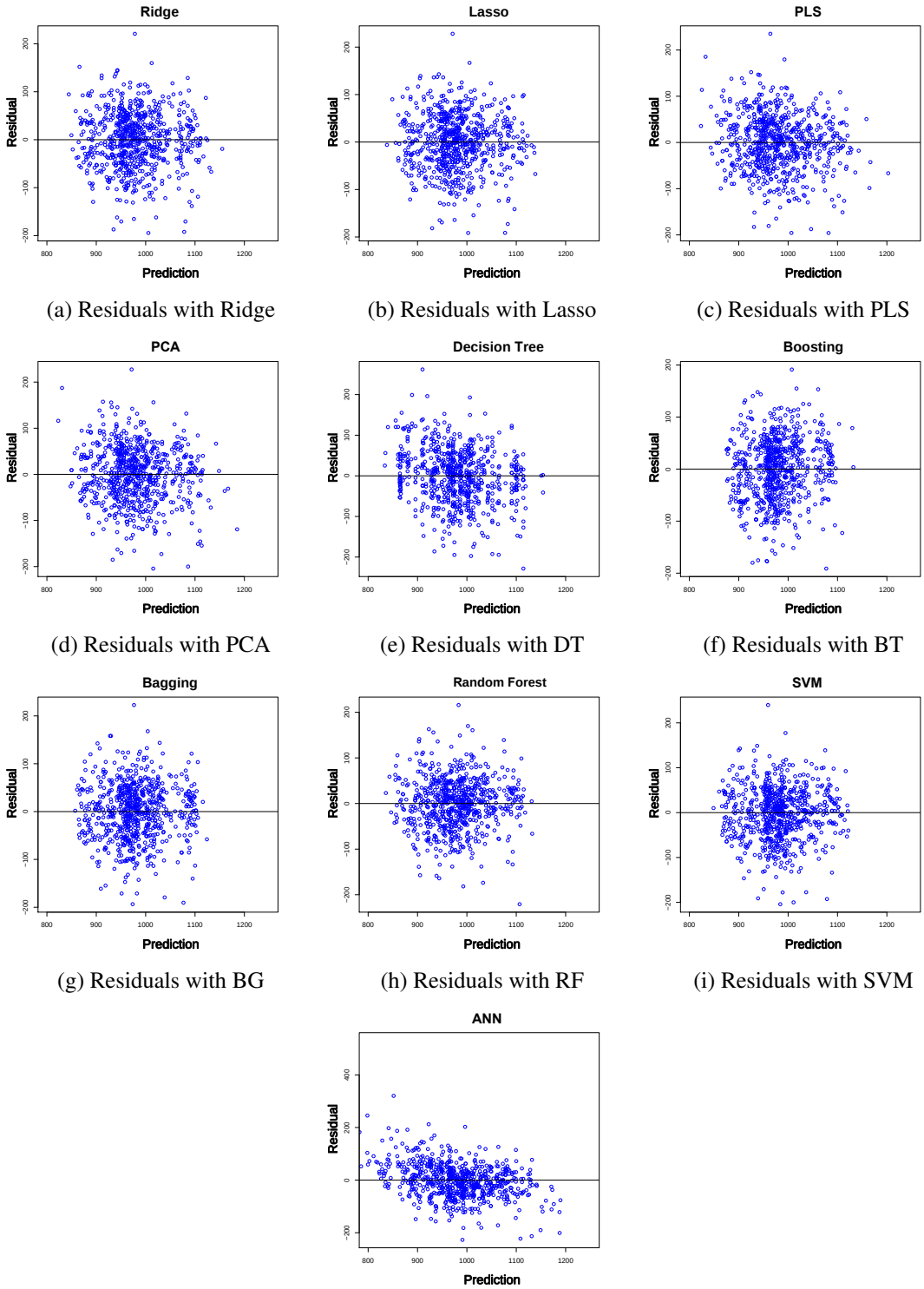


(i) Prediction with SVM



(j) Prediction with ANN

Fig. 5.8 Predicted v.s. Observed values for different statistical learning approaches: Ridge, Lasso, PLS (partial least square regression), PCA (principal component regression), DT (Decision Tree regression), BT (Boosting), BG (Bagging), RF (random forest), SVM (support vector machine), ANN (artificial Neutral Network)



(a) Residuals with Ridge (b) Residuals with Lasso (c) Residuals with PLS

(d) Residuals with PCA (e) Residuals with DT (f) Residuals with BT

(g) Residuals with BG (h) Residuals with RF (i) Residuals with SVM

(j) Residuals with ANN

Fig. 5.9 Residuals v.s. Predicted values for different statistical learning approaches:Ridge, Lasso, PLS(partial least square regression), PCA(principal component regression), DT(Decision Tree regression), BT(Boosting), BG(Bagging), RF(random forest), SVM(support vector machine), ANN(artificial Neutral Network)

Chapter 6

Influence of Meteorological Variability on the Predictive Capacity

One of the main objectives of modelling is to approximate the real phenomenon by model simulation as we have done in the previous chapters. Another more critical and challenging goal is to ensure high predictive power, that is, the excellent performance of the assumed model in a completely "new" scenario that has never been met before.

What has been accomplished in the "prediction" stage in Section 5.4 is considered to be a "fake" prediction in the sense that in Section 4.4 the observations from the same year were shown to be much more similar as compared to observations from different years. Note that the previous observations participate in both the "fitness" and the "prediction" steps. That is why a new strategy named "clustering result-based cross-validation" is proposed to have a more "strict" evaluation of the predictive capacity. The detailed process is described in Algorithm 4.

Algorithm 4 Clustering-based Cross-Validation Algorithm

1: An unsupervised clustering algorithm is called to divide the datasets into K ideal clusters $D_n = \{D_{\{1,n_1\}}, D_{\{2,n_2\}}, \ldots, D_{\{K,n_K\}}\}$
2: **for** i = 1 to K **do**
3: Leave out one of the clusters $D_{\{i,n_i\}}$;
4: Estimate the model on the K - 1 remaining parts $D_n/D_{\{i,n_i\}}$;
5: Evaluate the predictive capacity of the testing set $D_{\{i,n_i\}}$;
6: **end for**
7: Take the weighted average error as the prediction error obtained with the cross-validation since the sample volumes for different clusters may differ.

Table 6.1 Evaluation of goodness-of-fit and prediction with clustering-based cross-validation

	RMSE0	MRE0	RMSE	MRE
Ridge	48.28	0.0398	85.71	0.0722
Lasso	47.92	0.0394	96.89	0.0821
PLS	49.69	0.0407	92.29	0.0770
PCA	53.19	0.0440	86.98	0.0739
DT	48.38	0.0395	100.32	0.0860
BT	50.69	0.0425	75.89	0.0645
BG	42.90	0.0352	79.34	0.0682
RF	21.60	0.0170	75.32	0.0643
SVM	50.72	0.0396	76.01	0.0649
ANN	33.65	0.0251	119.98	0.1057

The results of this strategy are listed in Table 6.1. In comparison with the results that don't take into account the climatic variability as illustrated in Table 5.1, the fit for all the methods is slightly improved. It can be explained by the fact that the training set is more similar since it contains only ten different clusters and one cluster has been omitted, therefore reducing its overall variability. However, as for their predictive capacity, the predictive error becomes more significant as compared to the results shown in Table 5.1. For example, the RMSE for the ridge regression is changed from 57.84 g/m^2 to 85.71 g/m^2, etc.

6.1 Reducing the inter-annual variability by regrouping the meteorological data

The modelling process with the meteorological records from different years is usually associated with a certain instability caused by the inter-annual variability. To deal with this variability and reduce the uncertainty, to group the daily meteorological records in a certain period is supposed to be a relatively easy and effective way.

From Figure 6.1 to Figure 6.3 the 5 meteorological records of TMIN in Cochise County from 2001 - 2007 are presented. In Figure 6.1 the original daily records are given, while in Figure 6.2 these records are regrouped in 5-days recordings. Finally, in Figure 6.3, a 10-day regrouping is performed. It is obvious, even graphically, that the inter-annual variability is reduced.

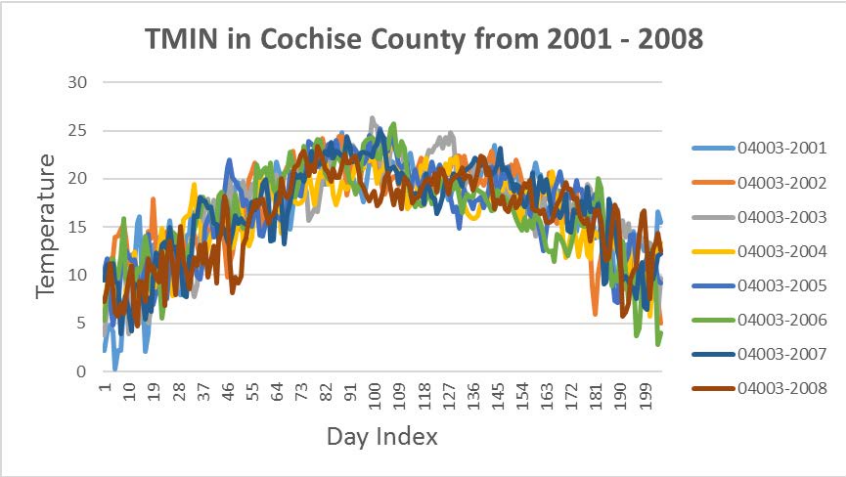


Fig. 6.1 Initial TMIN daily records in Cochise County

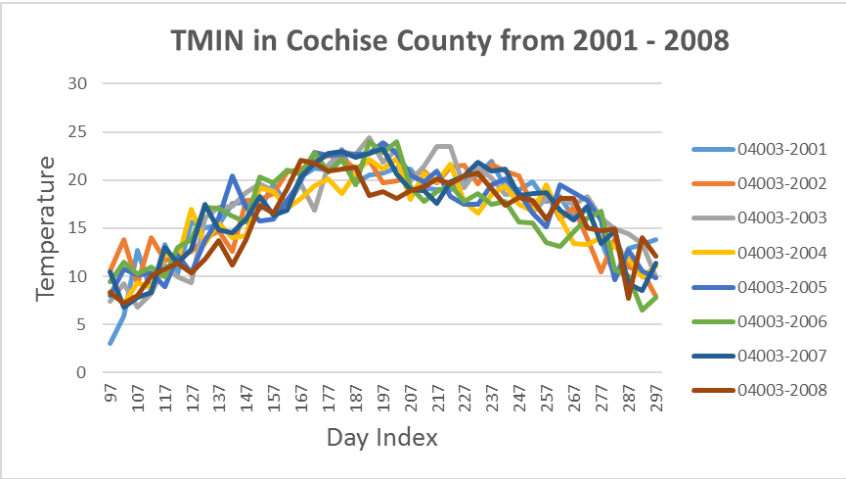


Fig. 6.2 TMIN records regrouped by 5 days in Cochise County

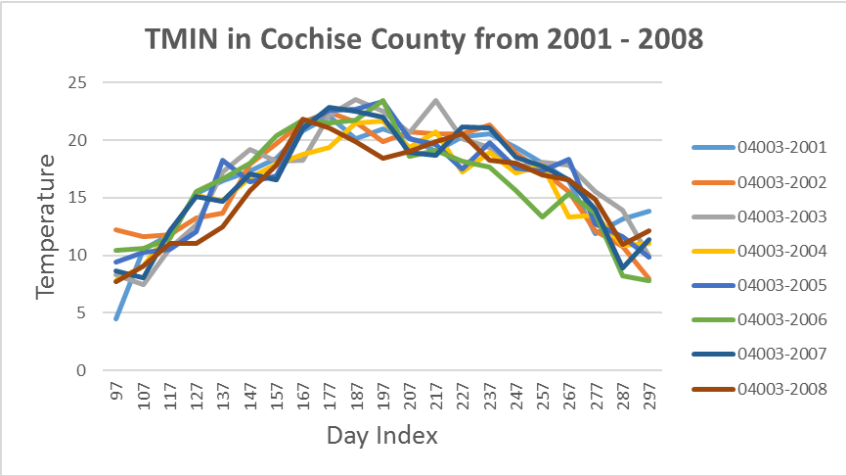


Fig. 6.3 TMIN records regrouped by 10 days in Cochise County

However, how will it influence the production model? Moreover, do the classical statistical methods, and machine learning methods are similarly influenced? To answer to these two questions, the weighted regression processes will be carried out with a typical statistical regression method, like the "ridge regression" for example. Moreover, for the machine learning methods, boosting regression is selected. Moreover, these two weighted regressions will be applied to the initial data, and those regrouped by every t-days for different values of t. Their performances are shown in Figure 6.4 for the classical statistical methods and Figure 6.5 for the machine learning methods.

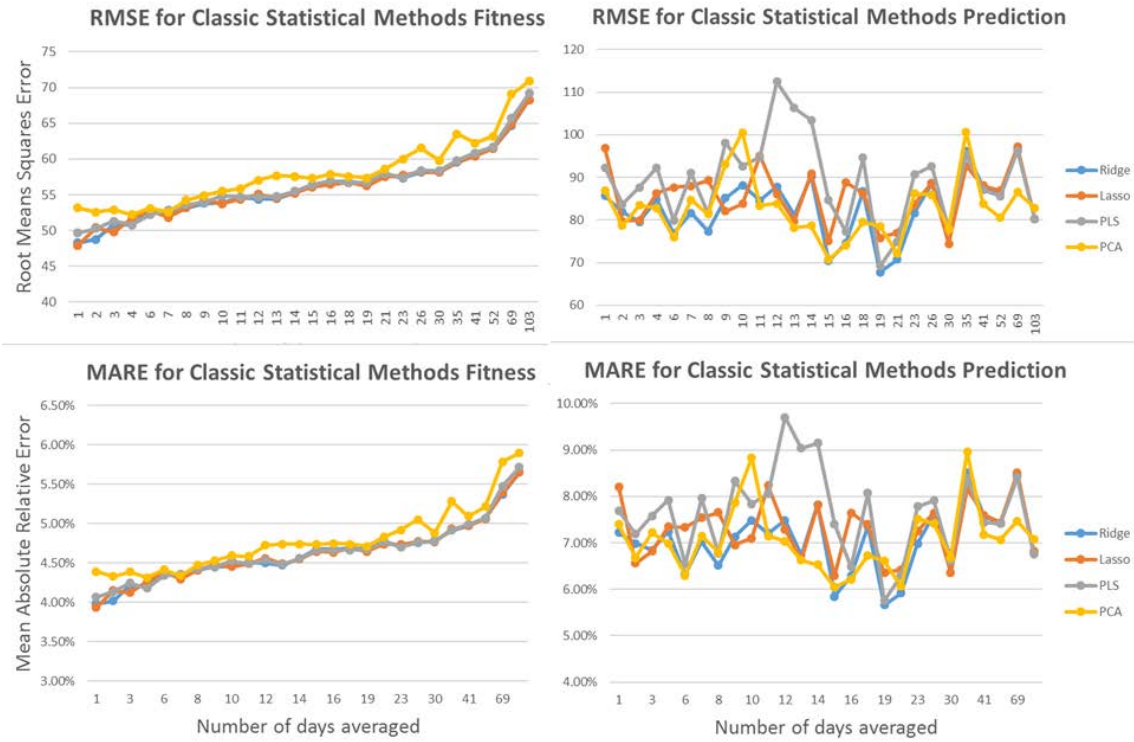


Fig. 6.4 cluster-cross-validation-04

According to the results, the evaluation of fitness for the traditional statistical methods are getting worse when the size t of the groups increases. This conclusion reflects the fact that averaging induces a loss of information and this results in a decrease in the fitting quality. However, on the contrary, information loss is associated here with a reduction in inter-annual variability, since the similarity between the training and the testing sets increases. The new compromise concerns “information loss” and “similarity gain”. According to Figures (b) and (d), Ridge regression can obtain a functional prediction capacity with RMSE = 67.79 g/m^2 with records regrouped by 19 days. There are also some other good sets, such as PCA with the regrouped records by every 15 days, PLS with the regrouped records by every 19 days and Lasso regrouped records by every 30 days.

As far as machine learning methods are concerned, most of them perform similarly, or even worse, both for fitness and prediction except for ANN, as shown in Figure 6.5. However, the

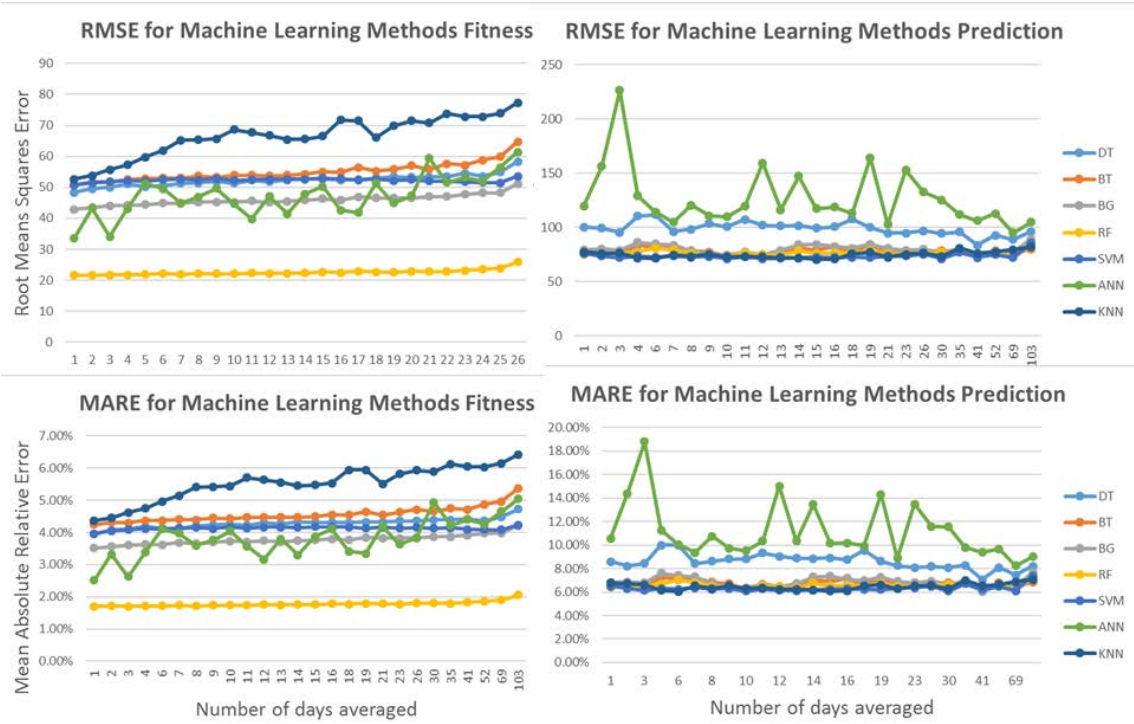


Fig. 6.5 Evaluation of machine learning methods with regrouped meteorological records: DT (Decision tree), BT (Boosting), BG (Bagging), RF (Random Forest), SVM, ANN, KNN regression

Random Forest regression has always the best performance for functional fitness, while the Bagging, Random Forest, SVM, and KNN have still 6% absolute relative error.

The different effect of the regrouping strategy of meteorological records on classical statistical methods and machine learning regression methods comes from their different methodology for fitness in dealing with collinearity. For the statistical regression methods, the most critical challenge consists in the reduction of dimensionality. The answer could be found in (Indyk and Motwani, 1998), where model variance reduces exponentially along with the dimension reduction, a phenomenon which in the opposite direction is named as the "curse of dimensionality" in (Friedman, 1997).

On the other hand, the machine learning methods try to fit the observations usually by the resampling technique. However, the sample size does not increase with the regrouping strategy. That is why the regrouped data has a different effect on these two methodologies.

6.2 Conclusion

When dealing with a learning problem, attention should be paid to prevent samples from appearing in both the training (for fitting) and the testing sets (for prediction) to avoid overfitting issues. An advisable strategy for this type of partition, or with cross-validation is detailed in Section.5.3. However, when dealing with meteorological data, the records of the positions geographically near to each other in the same year are strongly correlated. Some new strategy should be devised to forbid these similar samples appearing both in the learning and the prediction step. In this direction, a new methodology named "clustering-based cross-validation" is suggested in this chapter.

In this methodology, the dataset is first subdivided into several subsets by an unsupervised clustering technique. The "leave one out" principle is then adapted to the simulation of "new" scenarios for evaluation.

The results in Table 6.1 indicate that, under this "strict" predictive evaluation, models cannot perform as well as before (compare with Table 5.1 in Section 5.4). However, the models are proved to have robustness in their predictive capacity, when coming across completely new situations as illustrated in Figure 6.4 and 6.5. Additionally, a simple suggestion is made to increase the prediction performance by regrouping the daily records, and this approach is tested with different methods, including classical statistical and machine learning ones. According to their performance in both fitness and prediction, ridge regression stands out from the traditional statistical methods, while Random Forest regression from the Machine Learning methods.

Part III

Large-Scale Crop Production Prediction

Chapter 7

Weighted regression for Large-Scale Production Prediction

The scientific definition of "crop yield" corresponds to "the measure of grains or seeds generated from a unit of land, often expressed as (kg/H)". It is a term mostly used in scientific laboratories and rather unpopular. On the contrary, "crop production" is a large-scale subject, which greatly interests the government, the farmers, the market participants, etc. In Part I and Part II, knowledge-driven and data-driven approaches, which take into account the meteorological records as inputs, have shown their potential capacity for crop yield prediction. It is natural to ask, if model calibration with large-scale data, could contribute to the large-scale prediction problem.

In this chapter, the challenges and difficulties of large-scale crop production prediction with meteorological records will be discussed. A short introduction of French agriculture and soft wheat production will be made in Section 7.1, to illustrate the importance of soft wheat production to France, but also worldwide. In Section 7.2, potential data resources are indicated that could be used to deal with large-scale modelling. Some primary statistical analysis will be given in Section 7.3, including the methodological data and the harvest data of soft wheat in France. In Section 7.4, the large-scale crop modelling process will be described. The prediction results obtained with knowledge-driven approaches and data-driven approaches will be compared and analysed in Section 7.5 and 7.6. Finally, a conclusion and some perspectives will be made in Section 7.7.

7.1 Agriculture and Soft Wheat in France

France benefits a lot from a significant agricultural area, a favourable geographical and climatic situation, and also, of course, the agrarian policy, the Common Agricultural Policy. It has become the world's 6th largest and the leading agricultural country in the European Union, with about 1/3 of all the farmland and approximately 18% agricultural product in the European Union(IBP, 2015). It is also the second-largest exporter in the world of both services and farm products.

Agreste, the statistical department of the Ministry of Agriculture, is the primary department responsible for agricultural data in France. Every year, they conduct several surveys on agricultural production methods (SAPM) to estimate areas and yields of the main crops (Eurostat, 2017). Each study is conducted through telephone interviews with farmers. Some critical statistic analysis for the crop sector and the government decision-makers will be carried out, among which the most interesting one could be the production level for certain individual crops. From the perspective of the global market, these statistics are needed to make accurate price predictions, which in turn serve to make business decisions. However, the statistics are dependent on the phone survey. Thus an acceptable accuracy cannot be reached until the harvest date, which leads to a lack of predictability. Consequently, it is essential to have an effective alternative to replace SAPM with a quick and accurate prediction for the agricultural products in France.

Fig. 7.1 A photo of wheat in field

"Wheat" is a generic term for several grains belonging to the genus **Triticum**, grown in many countries. It is the third most essential grains, judged by the importance of the world harvest after rice and corn, with about 700 million tons annually. According to their endurance to low temperatures, it can be divided into winter wheat and spring wheat. Moreover, the soft wheat accounts approximately for 87% of the wheat production in France for the year 2015/2016 (Agreste, 2016).

In terms of production efficiency, according to Figure 7.2, South America has stable yields at 20 q/ha, Africa 10 q/ha, Egypt about 35 to 40 q/ha. At the same time, in Europe, remarkably high yields can be obtained in intensive cultivation. In France, the gains are even more remarkable: the current production amounts to 100 q/ha for the most successful farmers. In 2016, French wheat production was about 29.3 million tons, which account fort 20.3% of the EU production (144.5 million tons). That is why we choose to model soft wheat production in France.

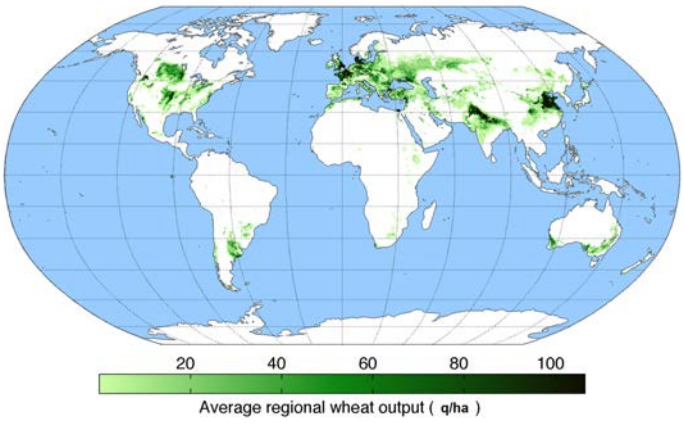


Fig. 7.2 Average wheat yield distribution all over the world

7.2 Data description

The data used in this research come from two main parts:

- Agricultural harvest data: including the crop and their surface of farmland, with which we introduce the yield data, as the model in this research is used for yield production (harvest/surface). The data are obtained from Agreste on their website (DISAR, 2018). The harvest data are at the departmental level. Finally, the soft wheat harvest data at the departmental level from 1989 to 2010 will be used in this research.

- Meteorological data along with the crops' growth: These data have been used in many fields in agricultural research. Their usefulness depends on the level of accuracy with which they represent the atmospheric conditions in the fields where the crops grow. For now, AgMIP Climate Forcing Datasets is one of the most complete and accurate meteorological datasets in the world. It is based on the NASA Modern-Era Retrospective Analysis for Research and Applications (MERRA) (Ruane et al., 2015). AgMERRA is the newly updated dataset that corrects to gridded temperature and precipitation, incorporates satellite precipitation, and replaces solar radiation with NASA/GEWEX SRB to cover the 1980-2010 periods. It is recommended to go to their website for more detailed information (GISS, 2014).

The preprocessing of these two datasets will be detailed later in Section 7.4.2 before the modelling process.

7.3 Basic statistical analysis of soft wheat production in France.

In this section, the primary statistical analysis of the soft wheat will be carried out in two levels: the national and the departmental level.

7.3.1 National level analysis

According to Figures 7.3 to 7.5, the national production of soft wheat varies between 28.85 million tons (lowest record in 1993) and 38.1 million tons (highest record in 1998). However, note also that the cultivated farmland with soft wheat in 1993 was the lowest one (4.26 million ha), thus explaining its low production at that year. Moreover, in 2008, the cultivated area in France reached a peak of about 5.04 million ha. As a result, the yields have been kept between 64.27 q/ha (in 2003) and 77.81 q/ha (in 2004, just one year after the lowest.).

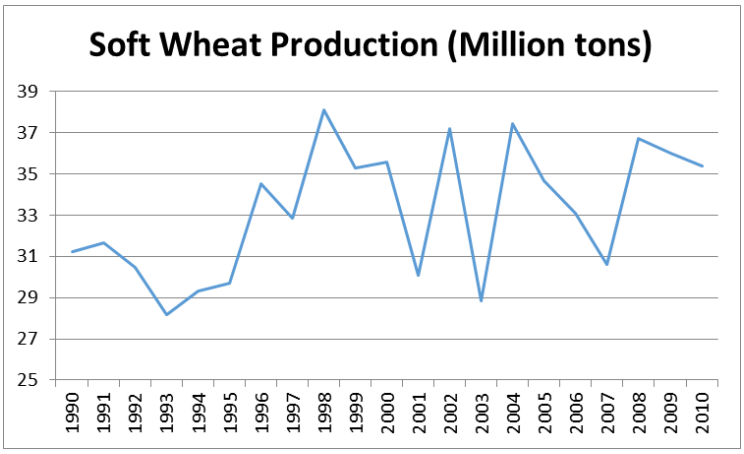


Fig. 7.3 National Production of soft wheat from 1990 to 2010

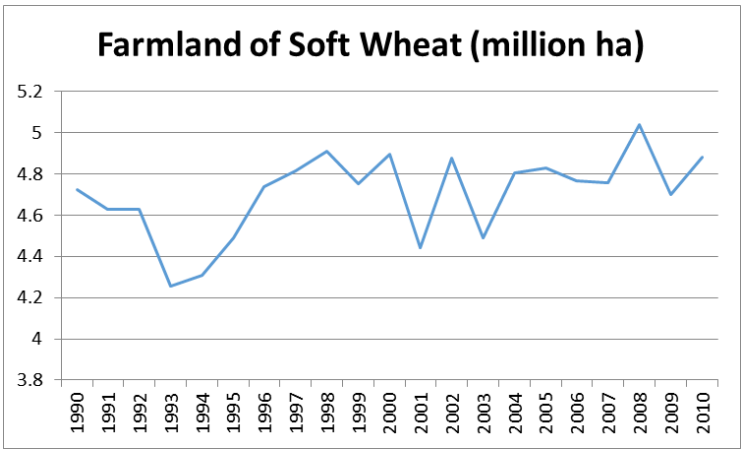


Fig. 7.4 National farmland of soft wheat from 1990 to 2010

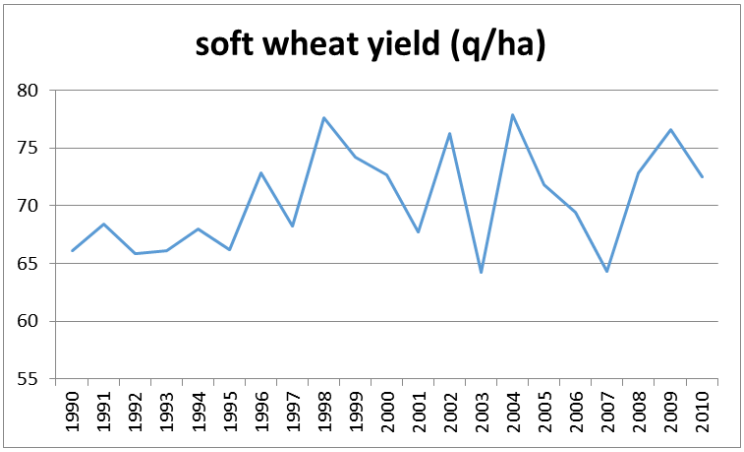
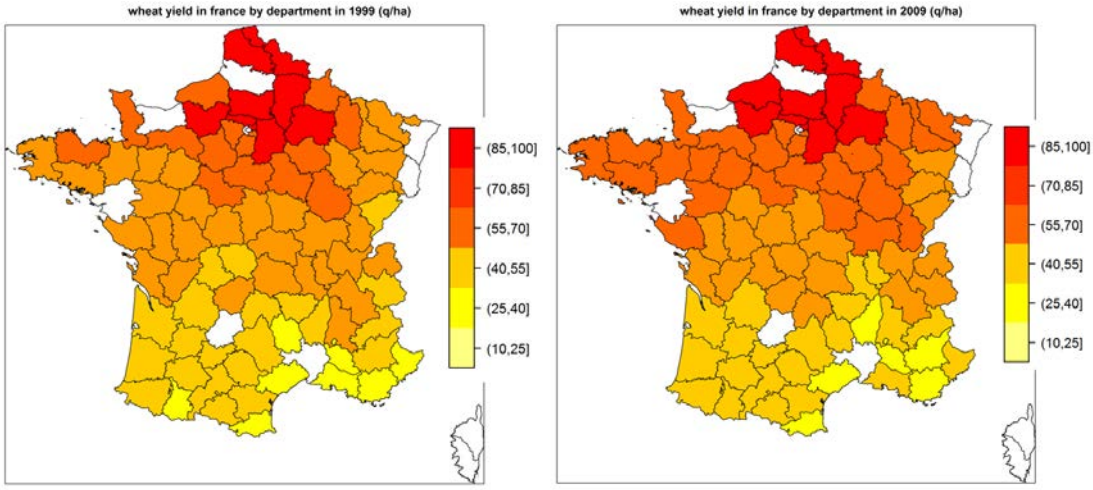


Fig. 7.5 Average yield of soft wheat from 1990 to 2010

7.3.2 Departmental level analysis

France is a country with rich geographical diversity, which makes the crop's performance in this country vary a lot from place to place. In terms of soft wheat's yield in France, it can be observed that the soft wheat yield varies from department to department and from year to year, as shown in Figure 7.6a and Figure 7.6b. However, a geographical correlation between the departments in terms of their soft wheat yield can be easily found. The lowest yield for each year is always observed in the south, close to the Mediterranean Sea and the Alps like Gard (33th department), Pyrénées-Orientales (66th department), Var (83th department), Vaucluse (84th department). On the other hand, the highest yield always occurs in the north, like Somme (80th department), OISE (60th department) and AISNE (2th department).



(a) Soft wheat yields by the department in 1999 (b) Soft wheat yields by the department in 2009

Fig. 7.6 Soft wheat yields distribution in France in 1999 (left) and 2009 (right)

7.4 Modelling the production of soft wheat

7.4.1 Crop Production modelling

As demonstrated in Part I and Part II, knowledge-driven approaches and data-driven approaches work well for corn yield prediction in America with a dataset of the form $\{U_i, y_i\}$, with U_i the meteorological record and y_i the corn yield at the county level. And for this case study, the available dataset is in the form $\{U_i, S_i, P_i\}$ with U_i the available climate condition, S_i and P_i the cultivated surface and production at the departmental level in France. It is possible that the crop production model could be derived from the crop yield model.

Let us denote by f, a general expression for the yield model, including knowledge-driven or data-driven models as expressed in Equation 1.1 and , for the production model, the quadratic loss function could be expressed as follows:

$$Loss = \sum_{i,j} (P_{ij} - \hat{P}_{ij})^2, \qquad i \in \mathbb{Z}[1990, 2010],\ j \in \mathbb{Z}[1, 89], \tag{7.1}$$

where P_{ij} and $\hat{P}_{ij}$ correspond to the observation and the estimation of the crop production for the year i in the department j respectively. Since the production could be decomposed into the crop yield and the cultivated surface in the form $P_{ij} = S_{ij} * y_{ij}$, the loss function given above can be written as follows:

$$Loss = \sum_{i,j} (P_{ij} - \hat{P}_{ij})^2 = \sum_{i,j} S_{ij}^2 * (y_{ij} - \hat{y}_{ij})^2. \tag{7.2}$$

Then, the production of a certain year i, denoted by Q_i can be easily obtained by:

$$Q_i = \sum_j \hat{P}_{ij} = \sum_j S_{ij} * \hat{y}_{ij}. \tag{7.3}$$

The above expression of the loss function indicates that the large-scale production prediction result is a weighted sum of the loss function in the departmental level. Consequently, attention should be paid more to those departments, where their contributions to total production are more significant. In Figure 7.7, it can be easily seen that the cultivated surfaces differ greatly from department to department, and this should be taken seriously into account. Such a regression model, where observations are treated differently, with unequal weights, correspond to the problem of "weighted least squares estimates".

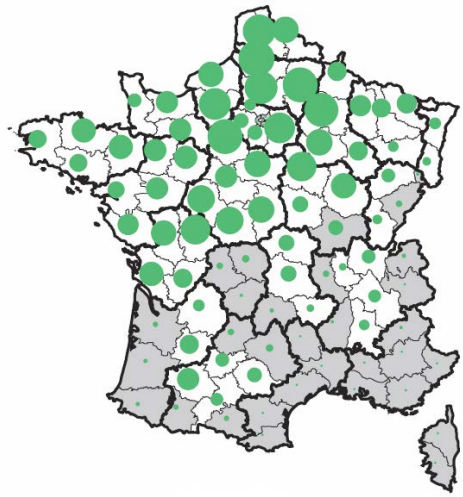


Fig. 7.7 The cultivated area for soft wheat in France in 2009

7.4.2 Data preprocessing

As expressed in Equation 7.2, the crop yield model plays a key role in the production model. The first step consists in generating the same data structure $\{U_i, y_i\}$ as we have already done in Part I and Part II, from the available data resource. The soft wheat yield observation can be easily obtained by Equation 7.2. However, it comes with some difficulties in generating the "correct" meteorological data for each department. The first effort is to take the geographical centres as the representative points for each department as in Figure 7.8.

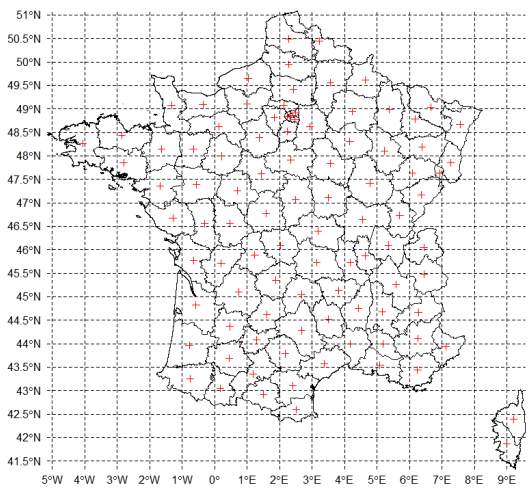


Fig. 7.8 Geographical centres for each department in France

It is stated in (GISS, 2014) that the meteorological variables are provided with different geographical scales. For Mean, Min and Max Temperature (°C) is by 0.5° × 0.5° while 0.25° × 0.25° for Precipitation (mm/day) and 1° × 1° for Solar Radiation ($MJ/(m^2 * day)$). The value of the meteorological variable for each representative point can be estimated by simple interpolation with the smallest rectangle around it.

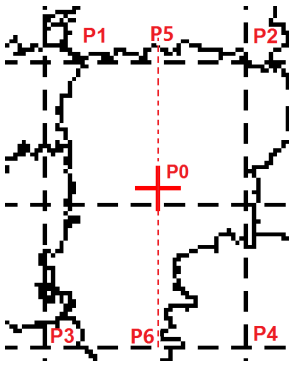


Fig. 7.9 Estimation with simple interpolation in rectangle

We take the Solar Radiation (SR) for the representative point of AISNE (2nd department) as an example. As shown in Figure 7.9, the representative point P_0 is surrounded by the rectangle represented by $\{P_1, P_2, P_3, P_4\}$(P_5 and P_6 are auxiliary points for interpolation). Each point P_i is characterised by $\{lon_i, lat_i, SR_i\}$ with lon_i, lat_i the longitude and latitude of point P_i and SR_i the value of solar radiation. It is obvious that $lat_2 = lat_1$,$lon_3 = lon_1$, $lat_4 = lat_3$, $lon_4 = lon_2$, $lon_5 = lon_6 = lon_0$, $lat_5 = lat_1$, $lat_6 = lat_3$.

The Solar Radiation of point P_5, SR_5, can be easily obtained by linear interpolation:

$$\begin{aligned} SR_5 &= SR_1 + (SR_2 - SR_1) \times \frac{lon_5 - lon_1}{lon_2 - lon_1} \\ &= SR_1 + (SR_2 - SR_1) \times \frac{lon_0 - lon_1}{lon_2 - lon_1} \end{aligned} \tag{7.4}$$

SR_6 for point P_6, can be obtained in the same way with SR_3 and SR_4:

$$\begin{aligned} SR_6 &= SR_3 + (SR_4 - SR_3) \times \frac{lon_6 - lon_3}{lon_4 - lon_3} \\ &= SR_3 + (SR_4 - SR_3) \times \frac{lon_0 - lon_1}{lon_2 - lon_1} \end{aligned} \tag{7.5}$$

Finally, the estimated value for the representative point P_0, SR_0, can be obtained by SR_5 and SR_6 by the following equation:

$$\begin{aligned} \mathrm{SR}_0 &= \mathrm{SR}_5 + (\mathrm{SR}_6 - \mathrm{SR}_5) \times \frac{\mathrm{lat}_0 - \mathrm{lat}_5}{\mathrm{lat}_6 - \mathrm{lat}_5} \\ &= \mathrm{SR}_5 + (\mathrm{SR}_6 - \mathrm{SR}_5) \times \frac{\mathrm{lat}_0 - \mathrm{lat}_1}{\mathrm{lat}_3 - \mathrm{lat}_1} \end{aligned} \tag{7.6}$$

For the other variables and the other representative points, the values can be obtained in the same way. The only difference is the grid-scale for different meteorological variables.

The final task in data preprocessing is to determine an appropriate interval for the life cycle of the soft wheat. A commonly used range for soft winter wheat, where the wheat semis 15 October and harvested on 23 July of the following year, is also used for this research. It accounts for 280 days in the whole life cycle for the soft winter wheat.

Since we take the daily record of Tmin, Tmax, Tavg, Prate and RG into consideration, the final dataset will contain 1 column of objective variable and 280 $\times$ 5=1400 columns of explanatory variables. As for the number of rows, since we have 21 years production record for 96 departments in metropolitan France, normally we could get 96 $\times$ 21 = 2016 records. However, there is a lack of meteorological information for 14th,44th ,67th ,68th department, and also a lack of harvest data for 75th, 92th, and 2B. Finally, the dataset turns out to be a table of (96-7) $\times$ 21 =1689 rows and 1401 columns. Each observation can be expressed as $\{U_{i,j}, y_{i,j}\}$, with $i \in \mathbb{Z}[1990, 2010]$, $j \in \mathbb{Z}[1, 89]$. This dataset will be applied in the following section to predict the national soft wheat production in France with both knowledge-driven and data-driven approaches.

7.5 Weight regression with Crop model

The Log-Normal Allocation and Senescence (LNAS) crop model is a functional, structural crop model (FSPMs) that is firstly proposed and applied to sugar beet in (Cournède et al., 2013). Some important works such as (Viaud, 2018), (Chen, 2014) have been carried out, and the LNAS model is proved to have some impressive and stable property in crop modelling for different crops. And in this section, the crop model LNAS-wheat will be applied to study the soft wheat production. Before model calibration, sensitivity analysis and study of parameters' properties should be carried out as in Part I.

7.5.1 Sensitivity Analysis with Sobol indices

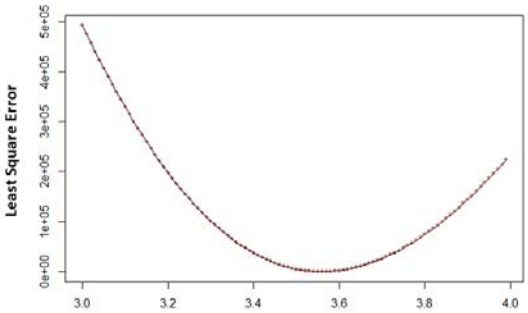
The results from the first order Sobol index for the LNAS model is listed in Table 7.1. The subgroup {rue, kB, e, sinit, muAlloc} are considered as important parameters. Theoretically, the predictive capacity of the crop model wouldn't be achieved without the calibration.

Table 7.1 Sobol Indexes of LNAS model

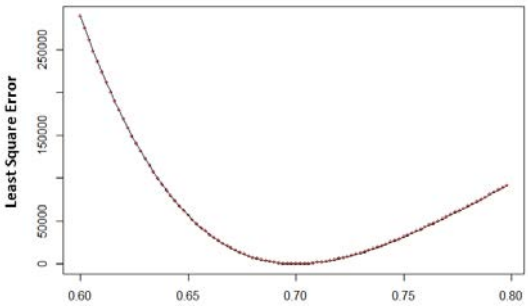
S_rue	S_kB	S_e	S_sinit	S_muAlloc	S_TTinit
0.3341	0.2180	0.2167	0.1495	0.0330	0.0086

7.5.2 Smoothness properties of the Loss function

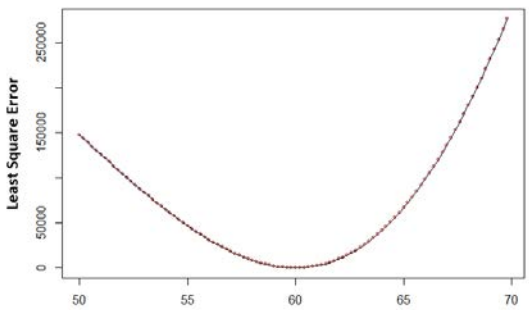
In this part, we are going to analyse some smoothness properties of the loss function concerning each parameter, including continuity and convexity. This kind of study enables a better evaluation of parameters uncertainty and also helps to decide the most appropriate optimisation algorithms.



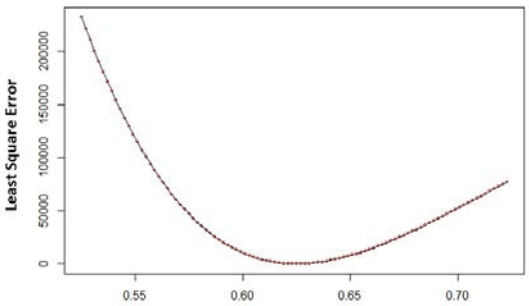
(a) LS function to parameter "rue"



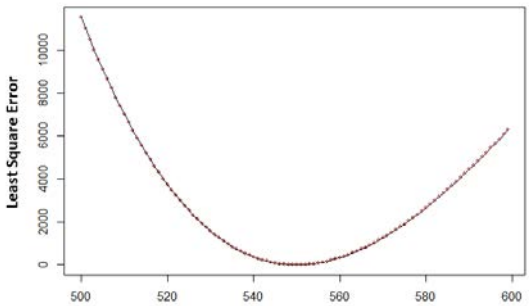
(b) LS function to parameter "kB"



(c) LS function to parameter "e"



(d) LS function to parameter "Sinit"



(e) LS function to parameter "muAlloc"

Fig. 7.10 Parameters' relation to the objective function

As in the Figure 7.10, the least squared error about these five parameters seems to be continue and may be convex (convex individual not convex globally), Which reduce the difficulties in optimisation process significantly.

7.5.3 Prediction results for the French soft wheat production

The prediction of soft wheat production is carried out with a different number of calibrated parameters. The different cases correspond to {case 01: rue, kB}, {case 02: rue, kB, e},{case 03: rue; kB; e; sinit} and {case 04 : rue; kB; e; sinit; muAlloc}. For each case, the process is repeated for 10 times and the results of production prediction from 1995 to 2010 are shown in Figures 7.11 to Figure 7.14.

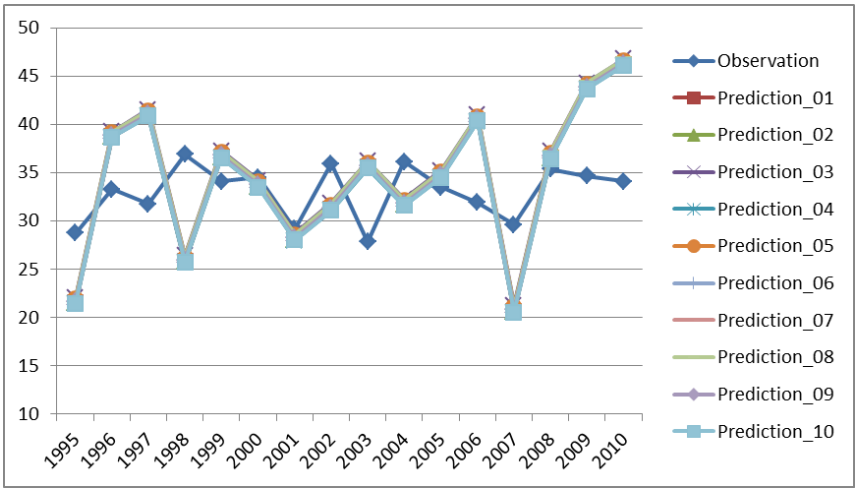


Fig. 7.11 Production prediction results with 2 parameters calibrated

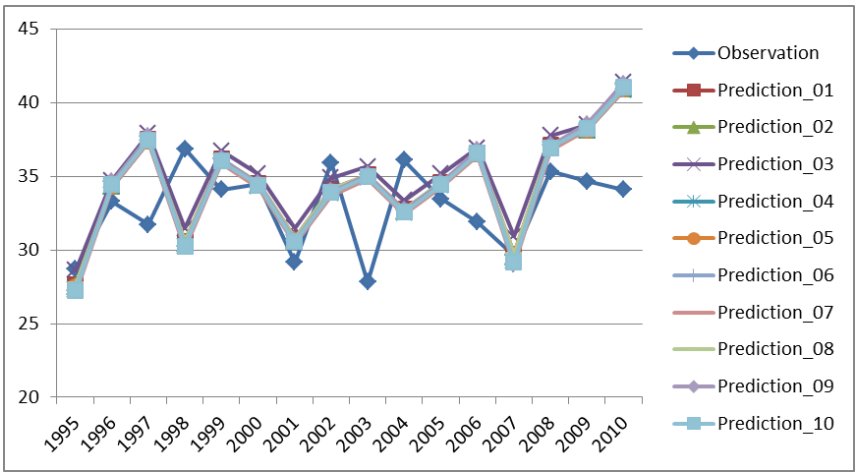


Fig. 7.12 Production prediction results with 3 parameters calibrated

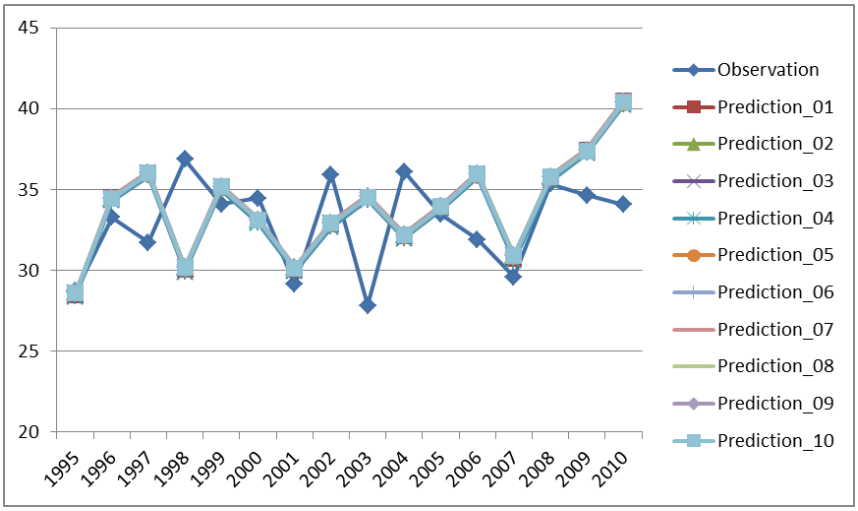


Fig. 7.13 Production prediction results with 4 parameters calibrated

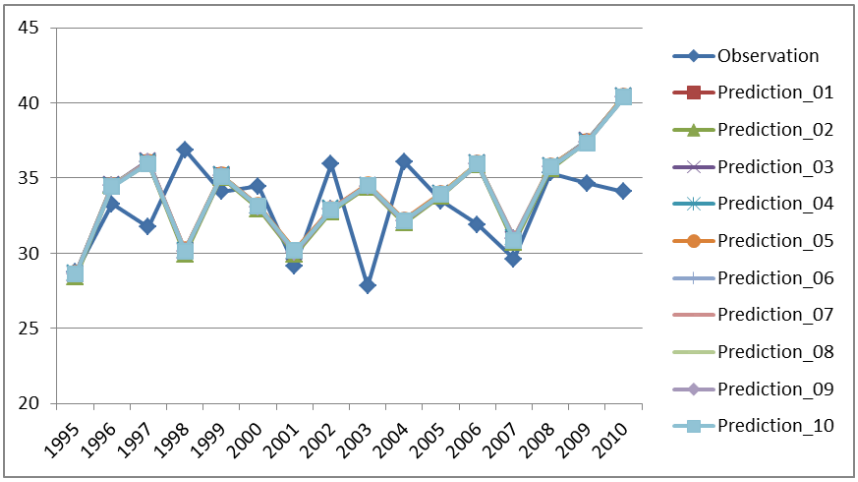


Fig. 7.14 Production prediction results with 5 parameters calibrated

The predicted production curve in Figure 7.11 to Figure 7.14 simulate more the less the tendency to the real sequences. For each setting, the repeated prediction coincide each other, which demonstrated the convergence of the MuScPE-PSO with good settings as in Section 3.4. By comparing the results of Figure 7.11 and Figure 7.12, it is clear that by increasing the number of calibrated parameters, a remarkable improvement in the prediction is achieved for the years 1995, 1996, 1998 and 2007. Nevertheless, when the number increases from 4 to 5, then the improvement is not so clear.

The results from the absolute relative error are listed in Table 7.2. The LNAS model could achieve better prediction when more parameters are calibrated. But the difficulty in optimisation increases at the same time. Thus, the LNAS model with four parameters calibrated could achieve both computational efficiency and accurate prediction.

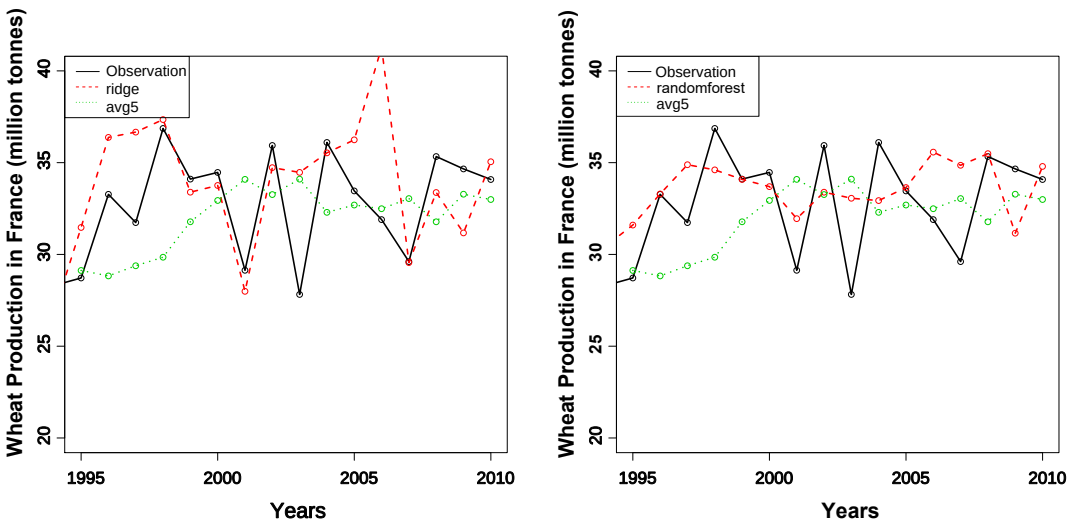
Table 7.2 Absolute relative error

models	average	std
LNAS case 01	0.183980	0.117976
LNAS case 02	0.092873	0.077370
LNAS case 03	0.084680	0.071953
LNAS case 04	0.084986	0.071843
avg-5	0.089406	0.064341

7.6 Weighted regression with the Statistical Learning model

7.6.1 Prediction results with the initial dataset

Two statistical learning methods will be applied in the weighted regression framework to predict the French soft wheat production. This choice is justified by the the ranking of machine learning and classical statistical methods with respect to their performance in crop yield prediction, as illustrated in Part II. In particular, the best performance was attained by the Random-Forest approach, and then by the ridge regression. A "year-wise" strategy, where the observations of the i-th objective year will not participate in the training step, is applied. Their prediction sequence from 1995-2010 are given in Figure 7.15, and the related statistics are shown in Table 7.3.



(a) Prediction with ridge regression (b) Prediction with Random-Forest regression

Fig. 7.15 Application of weighted regression on data regrouped every 5 days

It can be observed that, for the ridge model, the sub-period 1995-2002 are well predicted, while the predictions for 2005 and 2006 are far from reality. As for the Random-Forest model, the tendency is well caught by this model overall. On average, these two methods have superior

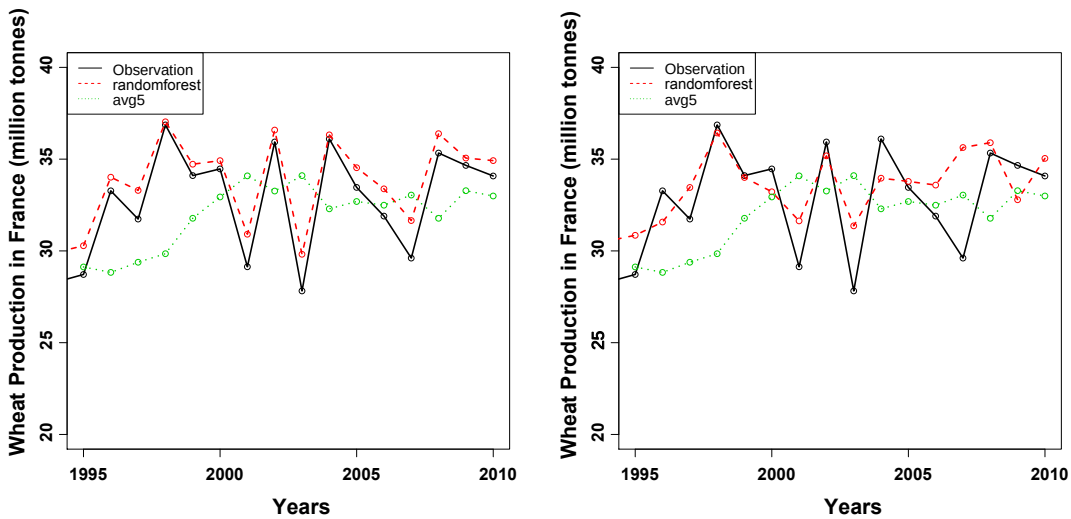
Table 7.3 Statistics of the absolute relative error

	average	std
Random-Forest	0.055524	0.050997
Ridge	0.074126	0.113421
avg-5	0.089406	0.064341

performance to the commonly used 5-year moving average. However, it is evident that the traditional statistical method, the ridge, exhibits much higher standard derivation.

7.6.2 Inter-annual variability analysis

The inter-annual variability of meteorological data is illustrated by a large variance of meteorological records among different years. Since this variance is much greater than the one resulting from year specific inter-departmental recordings, it is clear that including records from different years will increase instability. In the following example, we illustrate this problem in two cases. In case 1, the observations from the different year are mixed together. And the prediction results as in Figure 7.16a are based on normal cross-validation. As for the second case, a "year-wise" strategy is applied in the training-test process. That is, the observations for the year for prediction should not be used in the training process. The prediction results are shown in Figure 7.16b.



(a) Result without consideration of inter-annual variability

(b) Results with consideration of inter-annual variability

Fig. 7.16 An example of instability caused by inter-annual variability

As shown in the above figure, the predictions without consideration of inter-annual variability have a very similar prediction curve with the real observation curve as in Figure 7.16a, while the "year-wise" strategy gives a not so good curve in Figure 7.16b. However, the right one is more

close to a "real" prediction, since the inter-annual meteorological records vary so much that no similar records of the objective year should appear in the training set.

7.6.3 Reducing the inter-annual variability by regrouping the meteorological data

Modelling a process with meteorological records from different years usually comes up with instability caused by inter-annual variability. To deal with the inter-annual variability and reduce the uncertainty, it produced to the model, grouping the daily records of meteorological records in a certain period is supposed to be a relatively easy and effective way.

Figure 7.17 to Figure 7.19 are the 5 meteorological records of TMIN at the representative point (5.35°E, 46.10°N) in Ain (department 01 in France). The only difference in Figure 7.17 is the initial daily record, while in Figure 7.18 records are regrouped every 5 days, while in Figure 7.19 every 10 days. The inter-annual variability is reduced clearly by visualisation.

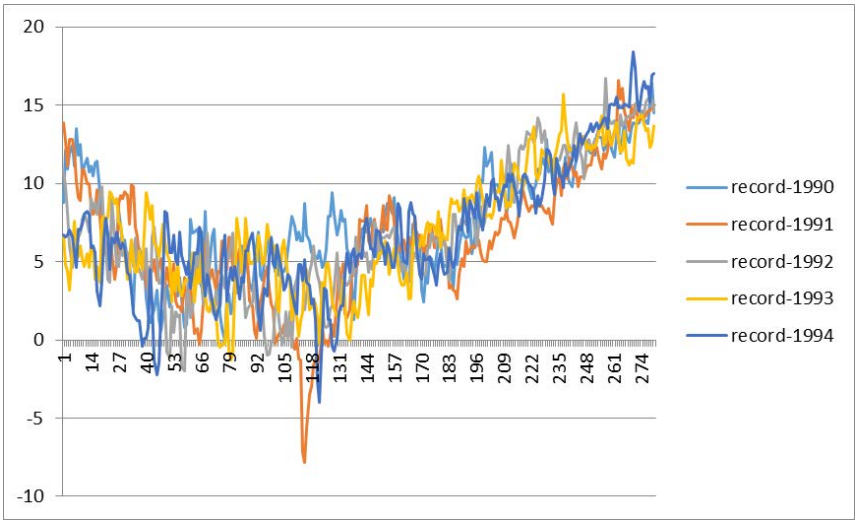


Fig. 7.17 Initial TMIN daily records at point (5.35°E, 46.10°N) in Ain

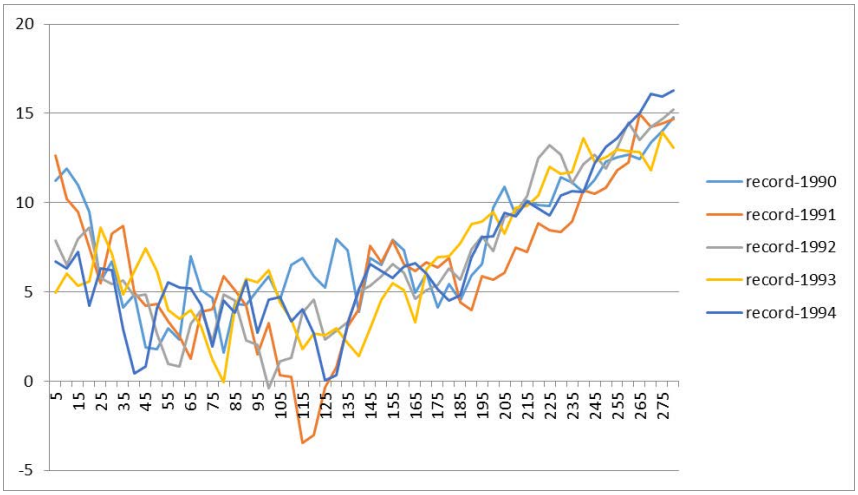


Fig. 7.18 TMIN records regrouped by 5 at point (5.35°E, 46.10°N) in Ain

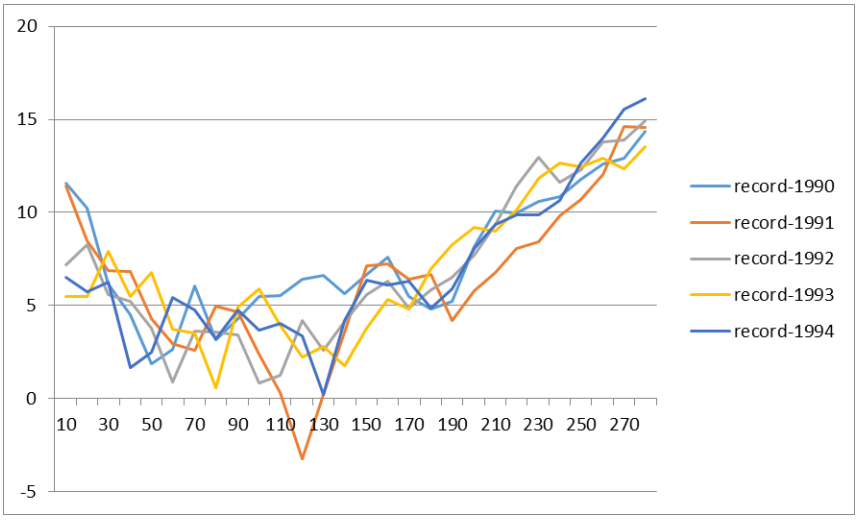
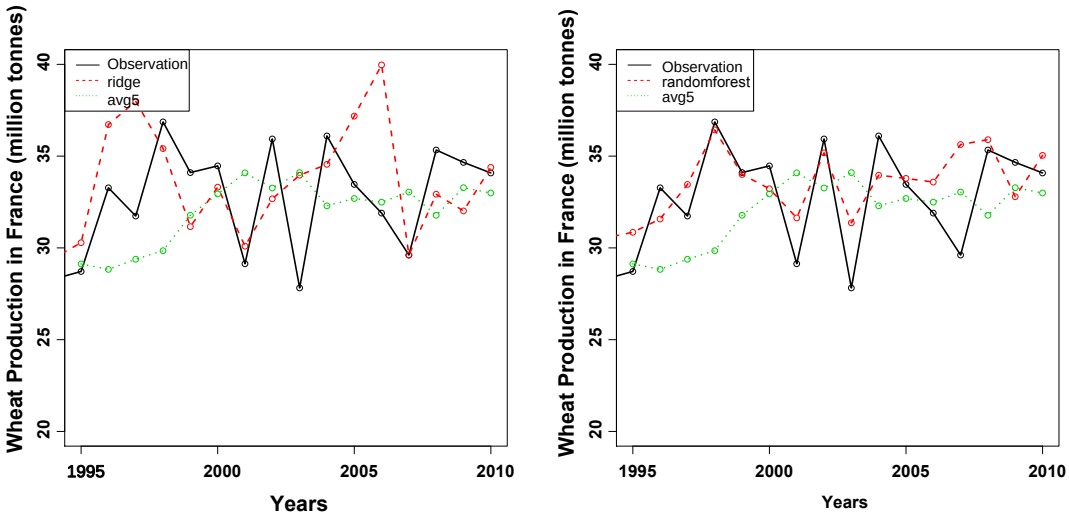


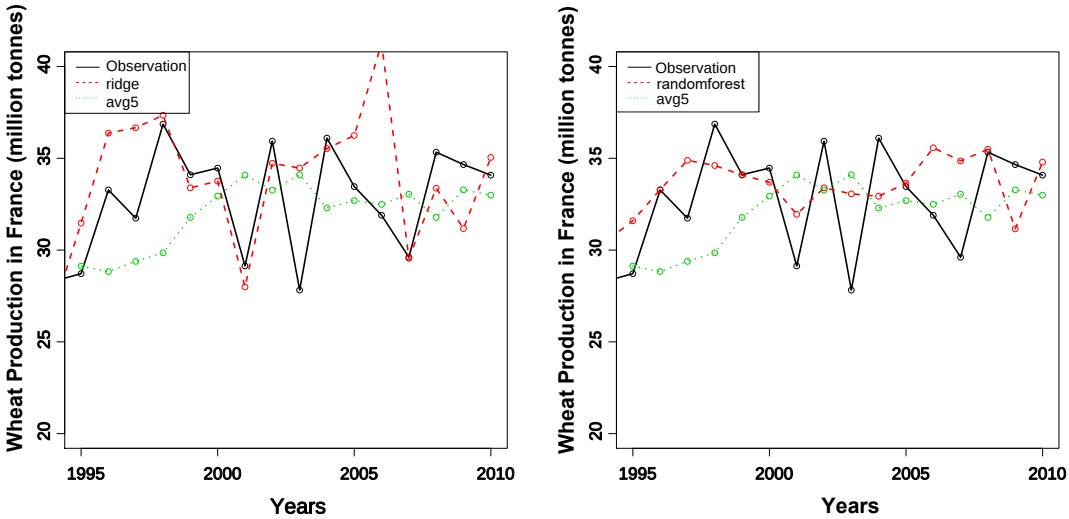
Fig. 7.19 TMIN records regrouped by 10 at point (5.35°E, 46.10°N) in Ain

However, how will it influence the production model? Moreover, does the influence of statistical and machine learning methods is of the same level? To answer these two questions in the framework of weighted regression, we opted for the “ridge regression” and the boosting regression respectively. Furthermore, these two weighted regressions will be applied to the initial data, and to those regrouped by a different number of consecutive days as illustrated in the following figures.



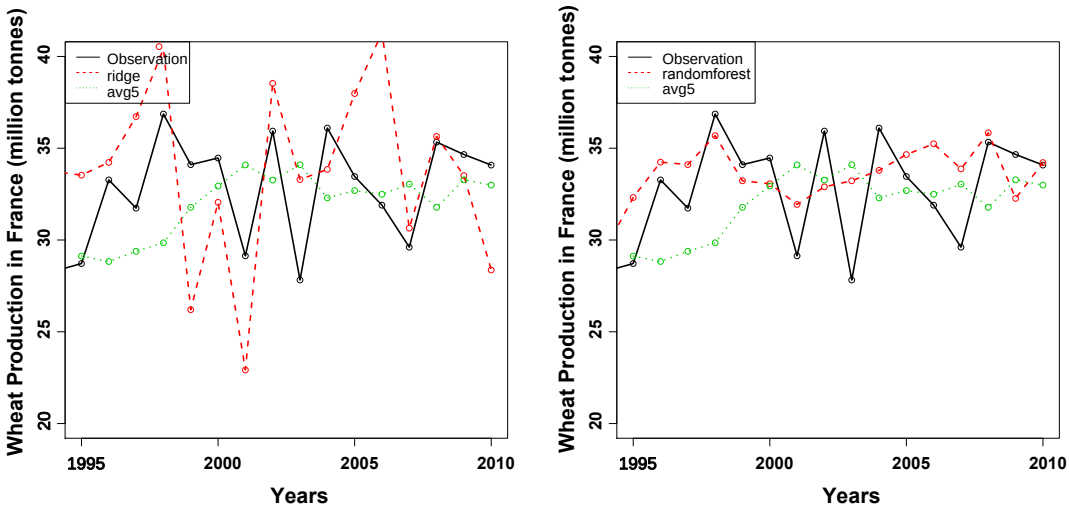
(a) Prediction with ridge regression (b) Prediction with randomforest regression

Fig. 7.20 Application of weighted regression on initial data



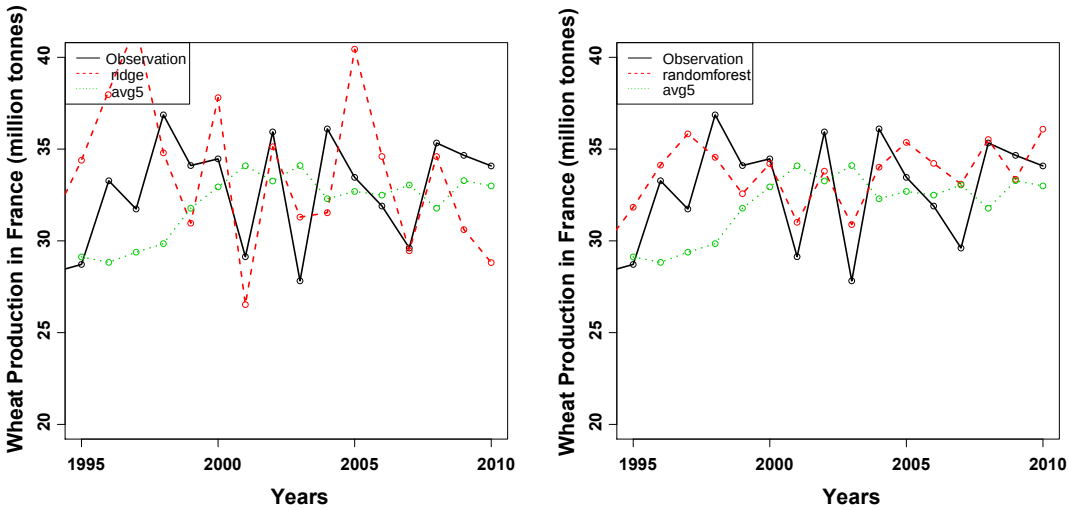
(a) Prediction with ridge regression (b) Prediction with randomforest regression

Fig. 7.21 Application of weighted regression on data regrouped every 5 days



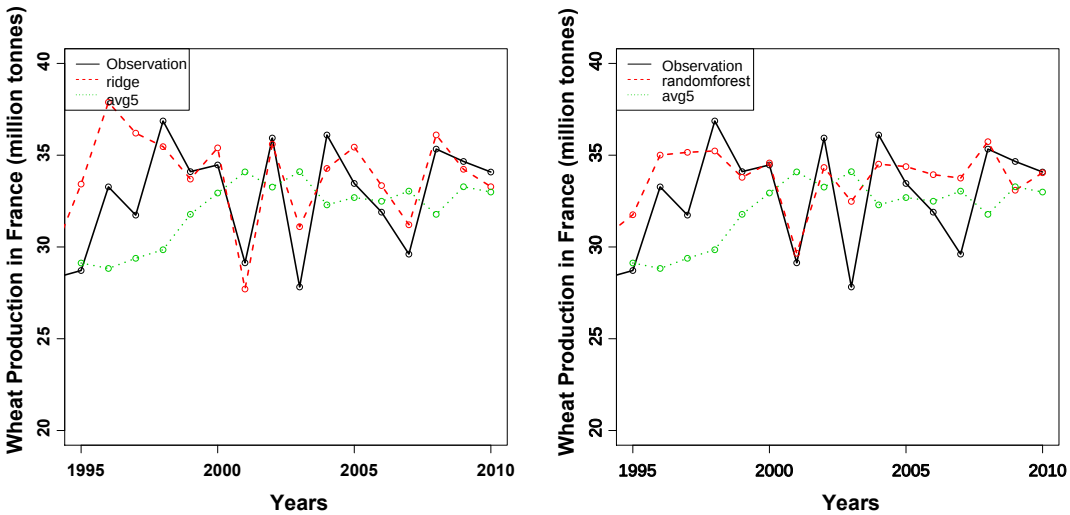
(a) Prediction with ridge regression (b) Prediction with randomforest regression

Fig. 7.22 Application of weighted regression on data regrouped every 10 days



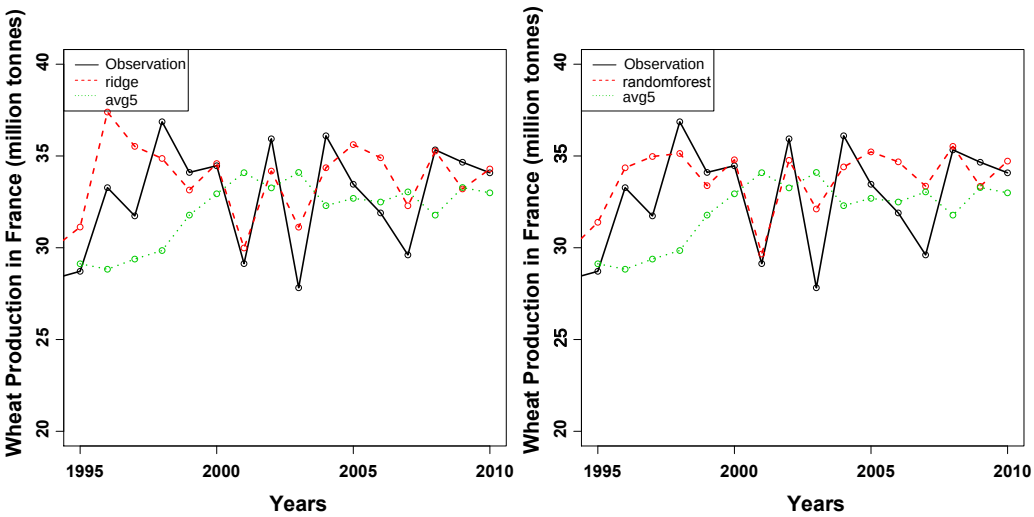
(a) Prediction with ridge regression (b) Prediction with randomforest regression

Fig. 7.23 Application of weighted regression on data regrouped every 30 days



(a) Prediction with ridge regression (b) Prediction with randomforest regression

Fig. 7.24 Application of a weighted regression on data regrouped every 80 days



(a) Prediction with ridge regression (b) Prediction with random forest regression

Fig. 7.25 Application of weighted regression on data regrouped every 280 days

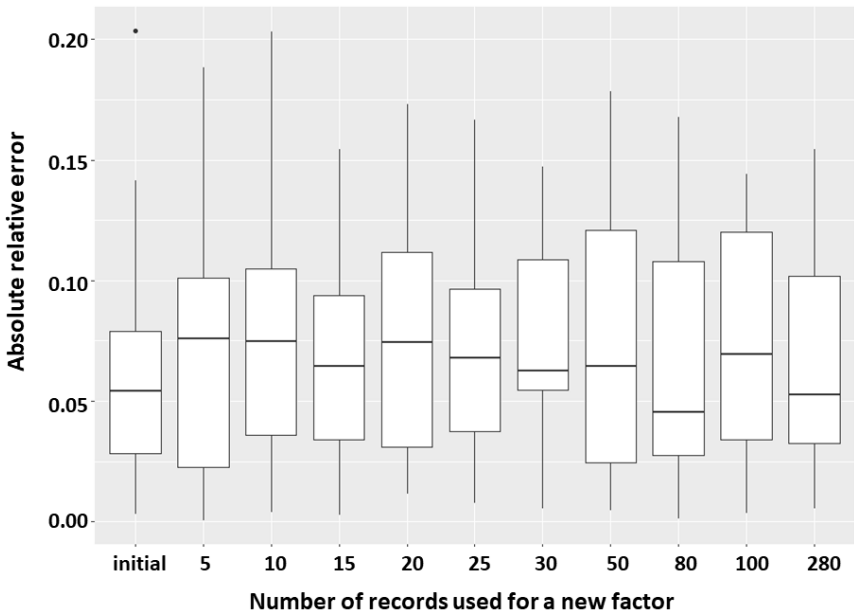


Fig. 7.26 Absolute relative error statistic of Random-Forest

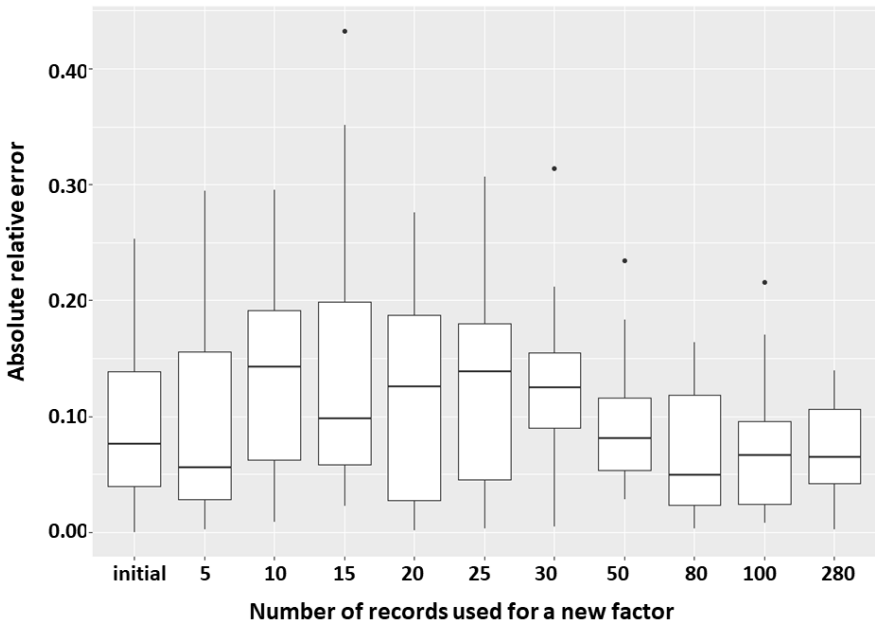


Fig. 7.27 Absolute relative error statistic of Ridge model

The reduction of inter-annual variability by regrouping the data helps a lot to improve the weighted regression accuracy with the ridge method. Moreover, the ridge regression method works best when the data are regrouped every 80 days and reduces the median of absolute relative error

from 7.5% to around 5.0%. As for the results of the weighted regression based on random-forest, the performance is not so significantly changed. The only remarkable finding is that the median of the absolute relative error is also obtained when the data are regrouped every 80 days. (Maybe 80 day is close to 90 days = 1 season, more filed experiments should be done to explain it).

7.7 Conclusion

In this chapter, to predict large-scale crop production, a methodology of using remote sensing and historical harvest data as an alternative option to replace the field crop surveys is proposed and examined. Under the weighted regression framework, knowledge-driven and data-driven approaches have been analysed and compared. Both are proved to be efficient when they are integrated into the framework of large-scale production prediction. As for data-driven strategies, the inter-annual variability is also discussed. It helps to improve efficiency by fairly regrouping the meteorological records.

Part IV

Conclusion

Discussion and Perspectives

Discussion

In this manuscript, the following issues were addressed:

- Analysis and comparison of knowledge-driven and data-driven approaches for the crop yield prediction;
- Influence of inter-annual variability of the meteorological data for both approaches;
- Application to the prediction of crop production at large-scale.

Methodology

One of the main contributions of this thesis is the proposition of an original parameter estimation methodology MuScPE (Multi-scenario Parameter Estimation Methodology) for datasets of the form $\{U_i, y_i\}$, which can be readily available. The key idea is to take advantage of crop's performance in a different scenario to hedge the lack of detailed description in a single environment. The parameters of a crop model, which are meant to describe the biological processes, are supposed to be genotype-specific. The performance of parameter estimation was shown to vary greatly when the environmental scenarios change, see Section 1.2.2. A sensitivity analysis is first carried out in Chapter 3 to study parameters' interactions during the crop's life and to evaluate their importance for parametrisation. By taking into account the continuity and the convexity of the objective function, together with the identifiability of the model, five critical parameters have been chosen to be estimated at first. The Particle Swarm Optimisation algorithm is then integrated into the MuScPE methodology due to the non-convexity of the loss function concerning specific parameters. Some improvements have been made by adding some tools, such as a controlling factor, a neighbourhood topology, as well as a parallelised framework, to make this optimisation algorithm more accurate and more efficient (Chapter 4. Good estimation results were achieved with the MuScPE methodology in a real case study. Some other characteristics related to the number of calibrated parameters and the number of scenarios used in the parametrisation process were also studied: 1. An accurate crop prediction was achieved when five parameters were calibrated for the

corn crop model; 2. An increase in the number of environmental scenarios resulted in a decrease in the lengths of the marginal confidence intervals of the parameters, thus reducing the uncertainty related to the prediction problem.

The second contribution is detailed in Part II. In this part, a systematic analysis of the prediction capacity with different statistical learning methods was conducted. The discussion is mainly focused on their capacity to deal with the problem of high-dimensional feature vectors. Traditional statistical regression methods derived from the linear model, are good at reducing the data dimension, such as the Ridge regression and the PCA regression. As for the machine learning methods, they outperformed the previous ones by adopting some sampling and resampling techniques.

Finally, these approaches have been integrated into a "weighted regression" framework to predict the crop production at large-scale. The French national soft wheat production prediction is the first application, and an accurate result was obtained compared to the traditional moving average.

In agro-environmental modelling, the variability of environmental scenarios is a crucial subject and was a central problem that we dealt with in this thesis. For example, in Part I, to obtain a parameter setting for the general environmental conditions, a k-means clustering analysis was carried out to divide the training set into different sub-groups, and the training examples were equally chosen from different sub-groups; In Part II, the influence of scenarios' variability was also studied. A simple solution by regrouping the daily records were also tested; in the case study concerning the large-scale crop production prediction, a year-wise training strategy is also proposed to make the forecast close to the "real" prediction.

Datasets

In this study, the datasets come from different resources:

- For the yield prediction, the dataset consists of 720 records of corn yield at county scale and the associated climatic data are provided by the United States Department of Agriculture (USDA). The five important daily climatic variables are daily maximum and minimum temperatures, radiation, precipitation, and potential evapotranspiration.

- For the production of soft wheat production in France, 1890 records of soft wheat yield at departmental scale and their daily meteorological records from 1990 to 2010. The yield data is collected from Agreste, the statistical department of the Ministry of Agriculture in France, and the database AgMERRA, the newly updated dataset that corrects to gridded temperature and precipitation, incorporates satellite precipitation and replaces solar radiation with NASA/GEWEX SRB to cover the 1980-2010 periods. The spatial resolution scale remains a problem, notably for a key variable like rainfall which is known to vary a lot at tiny scales.

With the government's efforts towards open data, we may expect to have in the future a better availability of data. Likewise, the rapid development of climatic sensors should help refine the resolutions at which data are available. Similarly, with the recording yield maps during harvest at very fine scales, which is more and more frequent with the very sophisticated modern combine harvester, we can expect that our method will be able to refine the analysis of the climatic variability and its impact on crop yield, thus potentially improving the accuracy of model prediction. Of course, more and finer data implies an increase in the computational needs of the learning process. It will remain an issue to consider.

Results

In terms of corn yield prediction, the results showed that among the data-driven approaches, Random Forest was the most robust and generally achieved the best prediction error (MAEP 4.27%). It also outperformed our knowledge-driven approach (MAEP 6.11%). However, the method to calibrate the mechanical model from an easily available dataset offers several side-perspectives. The mechanical model can potentially help to underline the stresses suffered by the crop or to identify the biological parameters of interest for breeding purposes. For this reason, an interesting perspective is to combine these two types of approaches.

As for robustness of prediction related to the inter-annual variations of meteorological records, it remains to be completed with some more tests. The inter-annual varieties have a significant influence on the data-driven model. However, a simple technique to regroup the meteorological records can help to reduce the impact. The knowledge-driven approach is proved to be more stable in the prediction than the data-driven methods. A possible reason to explain this result may be the relationship between the final yield and the cumulative solar energy. Crop modelling is based on cumulative thermal time. The variation of meteorological records in certain days could be hedged by the differences in another day so that the cumulative thermal time could be more stable. However, in the data-driven methods, all the individual records are considered to have the same importance to the objective variable, the crops yield. A variation for a particular day may lead to a pretty different result.

Finally, in terms of the soft wheat production in France, the random forest outplays the others with an average absolute, and relative error equal to 5.1% from 1995 to 2010. Another noteworthy observation is under the weighted regression framework, and a well-calibrated LNAS-wheat model outperformed some statistical learning methods, such as ridge and lasso regressions. This methodology could be considered as a general framework in the prediction of variables of other agricultural products. More effort could be devoted to making this framework useful in deciding crop policy and planning.

Perspectives

Potential of crop models coupled with MuScPE for crop breeding

One of the challenges of modern plant breeding is to provide genetic solutions to increase plant productivity. A breeding program can be considered as the process of developing improved cultivars by manipulating available genetic variability to create new allelic combinations best adapted to target environments and applications (Messina et al., 2006). Traditionally, breeding is based only on phenotypic observations, which makes the work costly, long, and highly based on breeder's experience. However, the breeding of higher-yielding crop plants would be greatly accelerated if the phenotypic consequences of changes at some genetic markers of an organism could be reliably predicted (Messina et al., 2006). Crop models, which aim to simulate the genotype $\times$ environment interactions to predict the corresponding phenotypes, are born in such circumstances. They can be used to assist genetic improvement in four main ways: environmental characterisation for testing genotypes, assessment of specific putative traits for designing improved plant types, analysis of responses of probe genotypes for improved interpretation of multi-environment trials, and optimising combinations of genotypes and management for target environment (Messina et al., 2006). It has also been demonstrated how plant growth models could be used for a breeding program in (Letort et al., 2008), by showing that crop model parameter is compelling traits to analyse since models help deconvolve the Gene by Environment interaction.

Well constructed crop models should be able to simulate phenotypic traits of various genotypes in diverse environments: it can predict crop performance over a range of environmental conditions and help explain the principle causes of phenotypic features from the environment and genotypic factors. Many models on various crops have demonstrated their prediction ability in diverse environments (Kang, 2013), (Guo et al., 2006), (Chen, 2014), (Viaud, 2018). However, the range of environmental variations that can realistically be explored in the model calibration process of these models is usually pretty limited, so that the performed statistical analysis generally lacks some predictive capacity in a wide range of environmental conditions. MuScPE Methodology was therefore designed to break the dilemma. It can deal with limited or aggregated types of data as soon as they are collected under a large number of diverse scenarios. In the case of lacking sufficient complex experimental data, MuScPE takes advantage of a large amount of simple trait data, e.g., crop yield data, that is generally recorded at large-scales across thousands of counties and over tens of years (even though on such longer periods, genotypes or cultivars used by farmers usually change). This methodology brings the potential to enlarge many existing excellent crop models' prediction ability to large scales, which meanwhile will increase the utility scope of crop models and their added-value for crop breeding and other crop commercial purposes.

Statistical learning method combined with remote-sensing dataset

The word "remote sensing" refers to all the techniques that make it possible to study objects or phenomena remotely. It has been taken advantage of in many areas, such as meteorology, climatology, oceanography, cartography, and geography. However, no matter what the application field is, a proper interpretation of remote sensing data requires an understanding of the physical principles upon which the remote sensing technique is based. In crop modelling, remote sensing technology is mainly used for yield estimation and crop identification (Atzberger, 2013).

Chlorophyll is an essential component of the photosynthesis process. It is the pigment that absorbs solar energy and produces sufficient power for the photosynthetic reaction to take place. As it consumes heat, it has a significant impact on the amount of energy that will be reflected. This character makes each plant have a specific response to electromagnetic radiation, the "spectral signature" of vegetation (Huete and Jackson, 1987). Many factors could influence the reflectance of specific vegetation, such as the crop type or the growth stage.

The advances in remote sensing and information technology make it possible to catch and translate the useful information about plants from the reflectance and can be taken advantage of in different studies, such as crop yield prediction (Prasad et al., 2006a), crop identification (Lillesand et al., 2014), monitoring the state of crops and smart agriculture (Wang et al., 2006).

Many relationships between vegetation indices and yields have been highlighted in this way. Figure 7.28a and 7.28b are the initial satellite image and its related representation of Normalized Difference Vegetation Index (NVDI) (Walid Hammache and Cournède, 2018). This information can potentially be used to increase the learning datasets for yield prediction models.

(a) The initial image captured by satellite (b) The presentation in the form of NDVI

Fig. 7.28 Sentinel images(left), the related form of NDVI which combines the VV and VH polarisations, acquired at 2016.06.11 in South Dakota (USA)

Also, many factors could influence directly or indirectly the content of chlorophyll, thus affecting the spectral signatures of cultures caught by the satellite, such as the lack of nutrient deficiencies (nitrogen deficiency or manganese deficiency), lack of water, attack of diseases or pests. (Hatfield and Pinter Jr, 1993) gives a practical way to monitor crop disease and pests attack based on remote sensing. Since the traditional agricultural irrigation system is not highly efficient

in terms of water use, (Kim et al., 2008) and (Bausch, 1995) give an interesting method to locate the regions with water deficit by analysing the remote sensing images, which can help to provide water more accurately to crops according to their real needs. It goes the same with pesticides or fertilisers that are applied to a field. Coupled with remote sensing technologies, crop/plant models can help to determine the exact amount of the chemical product, which helps to protect the environment and help the farmers to save money. As a result, the advances of technology leads to the concept of "smart agriculture", which refers to a set of agricultural practices carried out in well-targeted areas of a field and at specific times (Abbasi et al., 2014). In brief, remote sensing data will play a more and more essential role in agriculture and will serve the purpose of the statistical methodologies we developed in this thesis.

To sum up, crop modelling is a crucial research topic, which involves biology, mathematics, information technology, and even economy or social sciences, when it comes to farmer practices. It paves the way to new methodologies and new technologies in agriculture, with the potential to change farming practices, to improve the product, thus ensuring humankind's subsistence, while preserving the environment for the sustainability. Such significant changes generally take time and will only be possible by proving the benefits to all stakeholders: farmers, governments, consumers. Quantitative methods are necessary tools for this purpose.

Appendix A

Synthèse en Français

La prévision du rendement des cultures est toujours une question primordiale. De nombreuses recherches ont été menées avec cet objectif en utilisant diverses méthodologies. Généralement, les méthodes peuvent être classées en approches basées sur les modèles et en approches basées sur les données.

Les approches basées sur les modèles reposent sur la modélisation mécaniste des cultures. Ils décrivent la croissance des cultures en interaction avec leur environnement comme systèmes dynamiques. Comme ces modèles sont basés sur la description mécanique des processus biophysiques, ils impliquent potentiellement un grand nombre de variables d'état et de paramètres, dont l'estimation n'est pas simple. En particulier, les problèmes d'estimation des paramètres résultant sont généralement non linéaires et conduisent à des problèmes d'optimisation non-convexes dans un espace multidimensionnel. De plus, l'acquisition de données est très difficile et nécessite un travail expérimental lourd afin d'obtenir les données appropriées pour l'identification du modèle.

D'un autre côté, les approches basées sur les données pour la prévision du rendement nécessitent des données provenant d'un grand nombre de scénarios environnementaux, mais les données sont plus simples à obtenir: (données climatiques et rendement final). Cependant, les perspectives de ce type de modèles se limitent principalement à la prévision de rendement.

La première contribution originale de cette thèse consiste à proposer une méthodologie statistique pour calibrer les modèles mécanistes potentiellement complexes, lorsque des ensembles de données avec différents scénarios environnementaux et rendements sont disponibles à grande échelle. Nous l'appellerons Méthodologie d'estimation de paramètres multi-scénarios (MuScPE). Les principales étapes sont les suivantes:

- Premièrement, nous tirons parti des connaissances préalables sur les paramètres pour leur attribuer des distributions a priori pertinentes et effectuons une analyse de sensibilité globale sur les paramètres du modèle afin de sélectionner les paramètres les plus importants à estimer en priorité.

- Ensuite, nous mettons en œuvre une méthode d'optimisation efficace non convexe, l'optimisation parallèle des essaims de particules, pour rechercher l'estimateur MAP (maximum a posteriori) des paramètres;
- Enfin, nous choisissons la meilleure configuration en ce qui concerne le nombre de paramètres estimés par les critères de sélection de modèles. Il y a en effet un compromis à trouver entre d'un côté l'ajustement aux données, et d'un autre côté la variance du modèle et la complexité du problème d'optimisation à résoudre.

Cette méthodologie est d'abord testée avec le modèle CORNFLO, un modèle de culture fonctionnel pour le maïs.

La seconde contribution de la thèse est la comparaison de cette méthode basée sur un modèle mécaniste avec des méthodes classiques d'apprentissage statistique basées sur les données. Nous considérons deux classes de méthodes de régression: d'une part, les méthodes statistiques dérivées de la régression linéaire généralisée qui permettent de simplifier le modèle par réduction dimensionnelle (régressions Ridge et Lasso, Régression par composantes principales ou régression partielle des moindres carrés) et d'autre part les méthode de régression de machine learning basée sur des modèles non-linéaires ou des techniques de ré-échantillonnage comme la forêt aléatoire, le réseau de neurones et la régression SVM.

Enfin, une régression pondérée est appliquée pour prédire la production à grande échelle. La production de blé tendre, une culture de grande importance économique en France, est prise en exemple. Les approches basées sur les modèles et sur les données ont également été comparées pour déterminer leur performance dans la réalisation de cet objectif, ce qui est finalement la troisième contribution de cette thèse.

Mots clés:Prédiction du rendement des cultures, approches basées sur les connaissances, approches basées sur les données, analyse de sensibilité, MuScPE, diversité environnementale, grande échelle

Appendix B

CORNFLO: A crop model of corn

The CORNFLO model is a plant growth model that simulates the growth and yield of the corn crop. It is inspired by the SUNFLO model for sunflower (Lecoeur et al., 2011). It consists of two versions: a potential model and a stress model. In this thesis, we consider only the potential model. The potential model is used to do the simulation of the potential corn yield without environmental stress. This model has been well described in (Kang, 2013). It includes the following three modules: crop phenology module, morphogenesis and photosynthesis module, biomass production and biomass distribution module. First of all, let's have a review of this model.

B.1 Phenology module

Normally, the initiation and the development of an organ depends on the cumulative time and also their environmental temperature. So does the development of the plant from that stage to another. To combine the influence of these two factors and to simplify the model complexity, a new notion named "thermal time" (cumulative heat) (Midmore et al., 2015) has been introduced into the plant modelling. It has been proved that the cumulative thermal time has a significant advantage over the use of calendar time for phrenology. In the CORNFLO model, the development of the plant is characterized by a succession of four physiological stages according to the cumulative thermal time: the flowering bud appearance time TTE1 ($^\circ C \cdot \text{days}$), the beginning of flowering TTF1 ($^\circ C \cdot \text{days}$), the early maturation TTM0 ($^\circ C \cdot \text{days}$) and the physiological maturity TTM3 ($^\circ C \cdot \text{days}$). The daily efficient temperature $T_{eff}(d)(^\circ C)$ at day d is calculated by Equation B.1:

$$T_{eff}(d) = T_{moy}(d) - T_{base}. \tag{B.1}$$

with $T_{moy}(d)$ the daily average temperature at day d and T_{base} the phenology base temperature.

According to (Williams, 2012), it is generally equal to 10 for maize. The thermal time $TT(d)(^{\circ}C \cdot \text{days})$ at day d is calculated as the accumulation of $T_{eff}(d)$. Then it is used to determine at what stage the plant.

B.2 Morphogenesis and Photosynthesis Module

A3 (cm^3) is the parameter giving the potential surface of the larger leaf of the simulated plant and A2 is the rank of the leaf which has the largest leaf surface in the entire plant growth period. $Ae(i)$ (cm^2) is the largest leaf surface for each leaf rank i. It is calculated with A2 and A3 in Equation B.2:

$$Ae(i) = A3 * e^{-0.0344*(i-A2)^2+0.000731*(i-A2)^3} \tag{B.2}$$

Figure B.1 illustrates the different values of $Ae(i)$ for three maize genotypes.

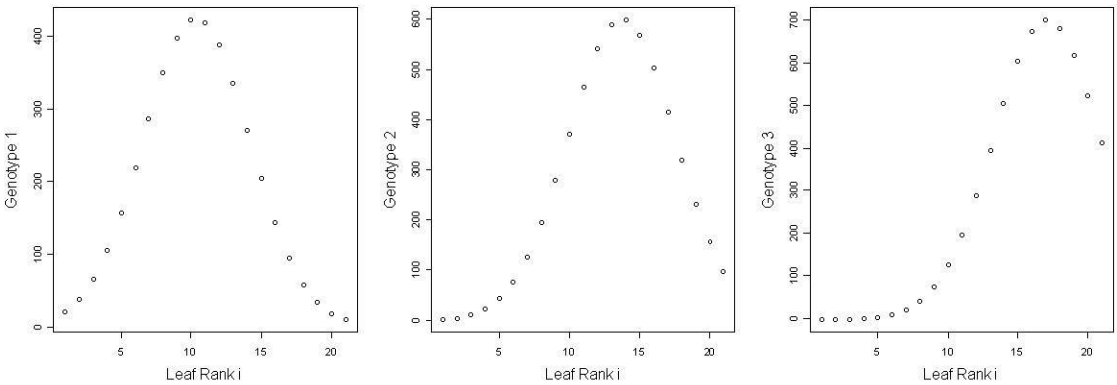


Fig. B.1 Maximum surface of the leaf $Ae(i)$ at each rank for three corn genotypes

The thermal time of appearance and death for each leaf i are designated as $TT_{de_pot}(i)(^{\circ}C \cdot \text{days})$ and $TT_{fe_pot}(i)(^{\circ}C \cdot \text{days})$. The duration of leaf expansion denoted as $TT_{exp_pot}(i)(^{\circ}C \cdot \text{days})$, expressed in thermal time is as following:

$$TT_{de_pot}(i) = \begin{cases} 1, & for \quad 0 \leq i \leq 2 \\ TT_{de_pot}(i-2) * phyllo_de_ini, & for \quad 2 \leq i \leq 5 \\ TT_{de_pot}(4) + (i-5) * phyllo_de_pot, & for \quad 5 \leq i \end{cases} \tag{B.3}$$

$$TT_{fe_pot}(i) = \begin{cases} TT_{fe_pot}(i-1) + (i-8) * phyllo_fe_pot, & for \quad i > 8 \\ TT_{fe_pot}(i-1) * phyllo_fe_ini, & for \quad i \leq 8 \end{cases} \tag{B.4}$$

$$TT_{exp_pot}(i) = TT_{fe_pot}(i) - TT_{de_pot}(i) \tag{B.5}$$

Where phyllodeini($^\circ C \cdot$ days) and phyllofeini($^\circ C \cdot$ days) are parameters of phyllochrone for leaf ranking below 8. For the other leaf, the beginning and ending time are noted as phyllodepot($^\circ C \cdot$ days) and phyllofepot($^\circ C \cdot$ days).They are the variables that depend on the number of leaves NFF, the stage $TTF1(^\circ C\ days)$, and the parameter $Ratio_phyllo_fe_phyllo_de$ defined as Equation B.6 and Equation B.7:

$$phyllofepot = \frac{TTF1 - 7 * phyllofeini}{NFF - 8} \tag{B.6}$$

$$phyllo_de_pot = phyllo_fe_pot * Ratio_phyllo_fe_phyllo_de \tag{B.7}$$

The expansion speed of the leaf i is calculated by the largest surface of each leaf i and its thermal expansion time period:

$$V_{exp_pot}(i) = \frac{Ae(i)}{TT_{exp_pot}(i)} \tag{B.8}$$

Therefore, the surface of the leaf i on day d, $SF_{i_pot}(d,i)(cm^2)$ is calculated as Equation B.9:

$$SF_{pot}(d,i) = SF_{pot}(d-1,i) + V_{exp_pot}(d,i) * T_{eff}(d) \tag{B.9}$$

It is initialised by $SF_{pot}(0,i) = 0, \forall i$.Then, the total leaf area $SFP_{pot}(d)(cm^2)$ at day d is given by Equation B.10:

$$SFP_{pot}(d) = \sum_{i=1}^{n} SF_{pot}(d,i) \tag{B.10}$$

The ratio of the green portion of all the leaf surface is noted as $Frac_g$. This coefficient will be used to calculate the index of leaf area $LAI_pot(cm^2/m^2)$ as in Equation B.12:

$$Frac_verte(d) = 1 - \frac{TT(d) - F1}{M3 - F1} \tag{B.11}$$

$$LAI_pot(d) = dens * SFP_pot(d) * Frac_verte(d)/10000 \tag{B.12}$$

where $dens(m^{-2})$ is the planting density of maize.

B.3 Biomass Production and Biomass Distribution Module

In order to calculate the biomass , another two parameters should be introduced: the radiation absorption efficiency $E_{i_pot}(d)$ and the radiation use efficiency $E_{b_bot}(d)(g.MJ^{-1})$. They are defined as in Equation B.13 and Equation B.14:

$$E_{i_pot}(d) = 0.95 * (1 - e^{-k_coeff * LAI_pot(d)}) \tag{B.13}$$

$$E_{b_pot}(d) = \begin{cases} RUE_pot & for \quad M0 \geq TT(d) \\ RUE_pot * (1 - \frac{TT(d)-M0}{M3-M0}) & for \quad M3 \geq TT(d) > M0 \\ 0 & for \quad TT(d) > M3 \end{cases} \tag{B.14}$$

with the extinction coefficient k_coeff and the maximum radiation use efficiency RUEpot$(g.MJ^{-1})$. Both are genotype parameters.

According to the energetic approach of Monteith, the daily biomass production $dMS_pot(g.m^{-2})$ should be calculate by the energy transferred from the solar energy (Monteith, 1977). In this model, the solar energy is represented by the radiation $RG(d)(MJ.m^{-2})$. Finally, a climate efficiency coefficient which is relatively constant at 0.48 will be used to adjust this equation:

$$dMS_pot(d) = 0.48 * RG(d) * E_{b_pot} * E_{i_pot}(d) \tag{B.15}$$

So the total biomass at day d,$MS_{tot_pot}(d)(g.m^{-2})$ results from the accumulation of the daily biomass production as in Equation B.16.

$$MS_{tot_pot}(d) = \sum_{t=1}^{d} dMS_{pot}(t) \tag{B.16}$$

In order to determine the performance $MS_{grain_pot}(d)$, a constant proportion of biomass (harvest index, HI) is assigned to the grain compartment:

$$MS_{grain_pot}(d) = MS_{tot_pot}(d) * HI \tag{B.17}$$

B.4 Genotype parameters

The CORNFLO model settings for ten genotypes were estimated by Syngenta using direct experimental measurements and statistical analysis. Our studies in the following Chapter are based on one of these genotypes. The recommended values and units of its parameters are shown in Table B.1.

Parameter	value	unit	Parameter	value	unit
$A2$	14.07	$\sharp$	$k_c oeff$	0.53	$\sharp$
$A3$	645	cm^2	$phyllo_fe_ini$	40	$^\circ C\ days$
$phyllo_de_ini$	32	$^\circ C\ days$	NFF	21	$\sharp$
$Ratio_phyllo_fe_de$	0.7	$\sharp$	HI	0.5	$\sharp$
$dens$	7	m^{-2}	RUE_{pot}	3.5	$g.MJ^{-1}$
$F1$	723	$^\circ C\ days$	$M0$	884	$^\circ C\ days$
$M3$	1477	$^\circ C\ days$	RT	0.36	$\sharp$
RE	0.37	$\sharp$	RO	0.2	$\sharp$

Table B.1 The parameter values of *CORNFLO* model for genotype studies.

Appendix C

Clustering Result of 720 scenarios

Table C.1 Scenarios in cluster 02

senario_name	Year	StateName	senario_name	Year	StateName
20089_2002	2002	Kansas	31085_2002	2002	Nebraska
20157_2002	2002	Kansas	31095_2002	2002	Nebraska
20179_2002	2002	Kansas	31097_2002	2002	Nebraska
31001_2002	2002	Nebraska	31099_2002	2002	Nebraska
31011_2002	2002	Nebraska	31109_2002	2002	Nebraska
31019_2002	2002	Nebraska	31119_2002	2002	Nebraska
31021_2002	2002	Nebraska	31121_2002	2002	Nebraska
31023_2002	2002	Nebraska	31125_2002	2002	Nebraska
31025_2002	2002	Nebraska	31127_2002	2002	Nebraska
31035_2002	2002	Nebraska	31129_2002	2002	Nebraska
31037_2002	2002	Nebraska	31137_2002	2002	Nebraska
31039_2002	2002	Nebraska	31141_2002	2002	Nebraska
31047_2002	2002	Nebraska	31143_2002	2002	Nebraska
31053_2002	2002	Nebraska	31151_2002	2002	Nebraska
31055_2002	2002	Nebraska	31153_2002	2002	Nebraska
31057_2002	2002	Nebraska	31155_2002	2002	Nebraska
31059_2002	2002	Nebraska	31159_2002	2002	Nebraska
31061_2002	2002	Nebraska	31163_2002	2002	Nebraska
31067_2002	2002	Nebraska	31167_2002	2002	Nebraska
31073_2002	2002	Nebraska	31169_2002	2002	Nebraska
31077_2002	2002	Nebraska	31173_2002	2002	Nebraska
31079_2002	2002	Nebraska	31177_2002	2002	Nebraska
31081_2002	2002	Nebraska	31181_2002	2002	Nebraska
31083_2002	2002	Nebraska	31185_2002	2002	Nebraska

Table C.2 Scenarios in cluster 01

senario_name	Year	StateName	senario_name	Year	StateName
8009_2001	2001	Colorado	31057_2001	2001	Nebraska
8099_2001	2001	Colorado	31059_2001	2001	Nebraska
20023_2001	2001	Kansas	31061_2001	2001	Nebraska
20039_2001	2001	Kansas	31063_2001	2001	Nebraska
20055_2001	2001	Kansas	31065_2001	2001	Nebraska
20063_2001	2001	Kansas	31067_2001	2001	Nebraska
20065_2001	2001	Kansas	31073_2001	2001	Nebraska
20067_2001	2001	Kansas	31077_2001	2001	Nebraska
20069_2001	2001	Kansas	31079_2001	2001	Nebraska
20081_2001	2001	Kansas	31081_2001	2001	Nebraska
20089_2001	2001	Kansas	31083_2001	2001	Nebraska
20093_2001	2001	Kansas	31085_2001	2001	Nebraska
20109_2001	2001	Kansas	31087_2001	2001	Nebraska
20129_2001	2001	Kansas	31093_2001	2001	Nebraska
20137_2001	2001	Kansas	31095_2001	2001	Nebraska
20147_2001	2001	Kansas	31097_2001	2001	Nebraska
20153_2001	2001	Kansas	31099_2001	2001	Nebraska
20157_2001	2001	Kansas	31109_2001	2001	Nebraska
20171_2001	2001	Kansas	31121_2001	2001	Nebraska
20175_2001	2001	Kansas	31125_2001	2001	Nebraska
20179_2001	2001	Kansas	31127_2001	2001	Nebraska
20181_2001	2001	Kansas	31129_2001	2001	Nebraska
20183_2001	2001	Kansas	31131_2001	2001	Nebraska
20187_2001	2001	Kansas	31137_2001	2001	Nebraska
20189_2001	2001	Kansas	31141_2001	2001	Nebraska
20193_2001	2001	Kansas	31143_2001	2001	Nebraska
20195_2001	2001	Kansas	31145_2001	2001	Nebraska
20199_2001	2001	Kansas	31151_2001	2001	Nebraska
20203_2001	2001	Kansas	31155_2001	2001	Nebraska
31001_2001	2001	Nebraska	31159_2001	2001	Nebraska
31011_2001	2001	Nebraska	31163_2001	2001	Nebraska
31019_2001	2001	Nebraska	31167_2001	2001	Nebraska
31021_2001	2001	Nebraska	31169_2001	2001	Nebraska
31023_2001	2001	Nebraska	31173_2001	2001	Nebraska
31025_2001	2001	Nebraska	31177_2001	2001	Nebraska
31035_2001	2001	Nebraska	31181_2001	2001	Nebraska
31037_2001	2001	Nebraska	31185_2001	2001	Nebraska
31039_2001	2001	Nebraska	35059_2001	2001	New Mexico
31047_2001	2001	Nebraska	6107_2006	2006	California
31055_2001	2001	Nebraska			

Table C.3 Scenarios in cluster 03

senario_name	Year	StateName	senario_name	Year	StateName
20067_2003	2003	Kansas	31081_2003	2003	Nebraska
20069_2003	2003	Kansas	31083_2003	2003	Nebraska
20081_2003	2003	Kansas	31085_2003	2003	Nebraska
20153_2003	2003	Kansas	31087_2003	2003	Nebraska
20157_2003	2003	Kansas	31093_2003	2003	Nebraska
20171_2003	2003	Kansas	31095_2003	2003	Nebraska
20175_2003	2003	Kansas	31097_2003	2003	Nebraska
20179_2003	2003	Kansas	31099_2003	2003	Nebraska
20199_2003	2003	Kansas	31109_2003	2003	Nebraska
31001_2003	2003	Nebraska	31119_2003	2003	Nebraska
31011_2003	2003	Nebraska	31121_2003	2003	Nebraska
31019_2003	2003	Nebraska	31125_2003	2003	Nebraska
31021_2003	2003	Nebraska	31129_2003	2003	Nebraska
31023_2003	2003	Nebraska	31131_2003	2003	Nebraska
31025_2003	2003	Nebraska	31137_2003	2003	Nebraska
31035_2003	2003	Nebraska	31141_2003	2003	Nebraska
31037_2003	2003	Nebraska	31143_2003	2003	Nebraska
31039_2003	2003	Nebraska	31145_2003	2003	Nebraska
31047_2003	2003	Nebraska	31151_2003	2003	Nebraska
31053_2003	2003	Nebraska	31153_2003	2003	Nebraska
31055_2003	2003	Nebraska	31155_2003	2003	Nebraska
31057_2003	2003	Nebraska	31159_2003	2003	Nebraska
31059_2003	2003	Nebraska	31163_2003	2003	Nebraska
31061_2003	2003	Nebraska	31167_2003	2003	Nebraska
31063_2003	2003	Nebraska	31169_2003	2003	Nebraska
31065_2003	2003	Nebraska	31173_2003	2003	Nebraska
31067_2003	2003	Nebraska	31177_2003	2003	Nebraska
31073_2003	2003	Nebraska	31181_2003	2003	Nebraska
31077_2003	2003	Nebraska	31185_2003	2003	Nebraska
31079_2003	2003	Nebraska			

Table C.4 Scenarios in cluster 04

senario_name	Year	StateName	senario_name	Year	StateName
8009_2004	2004	Colorado	31079_2004	2004	Nebraska
20039_2004	2004	Kansas	31081_2004	2004	Nebraska
20069_2004	2004	Kansas	31083_2004	2004	Nebraska
20081_2004	2004	Kansas	31085_2004	2004	Nebraska
20093_2004	2004	Kansas	31087_2004	2004	Nebraska
20109_2004	2004	Kansas	31095_2004	2004	Nebraska
20147_2004	2004	Kansas	31097_2004	2004	Nebraska
20157_2004	2004	Kansas	31099_2004	2004	Nebraska
20171_2004	2004	Kansas	31109_2004	2004	Nebraska
20175_2004	2004	Kansas	31119_2004	2004	Nebraska
20179_2004	2004	Kansas	31121_2004	2004	Nebraska
20187_2004	2004	Kansas	31125_2004	2004	Nebraska
20189_2004	2004	Kansas	31127_2004	2004	Nebraska
20193_2004	2004	Kansas	31129_2004	2004	Nebraska
20203_2004	2004	Kansas	31131_2004	2004	Nebraska
31001_2004	2004	Nebraska	31137_2004	2004	Nebraska
31011_2004	2004	Nebraska	31141_2004	2004	Nebraska
31019_2004	2004	Nebraska	31143_2004	2004	Nebraska
31021_2004	2004	Nebraska	31145_2004	2004	Nebraska
31023_2004	2004	Nebraska	31151_2004	2004	Nebraska
31025_2004	2004	Nebraska	31153_2004	2004	Nebraska
31035_2004	2004	Nebraska	31155_2004	2004	Nebraska
31037_2004	2004	Nebraska	31159_2004	2004	Nebraska
31039_2004	2004	Nebraska	31163_2004	2004	Nebraska
31047_2004	2004	Nebraska	31167_2004	2004	Nebraska
31053_2004	2004	Nebraska	31169_2004	2004	Nebraska
31055_2004	2004	Nebraska	31173_2004	2004	Nebraska
31057_2004	2004	Nebraska	31177_2004	2004	Nebraska
31059_2004	2004	Nebraska	31181_2004	2004	Nebraska
31061_2004	2004	Nebraska	31185_2004	2004	Nebraska
31065_2004	2004	Nebraska	35059_2004	2004	New Mexico
31067_2004	2004	Nebraska	48421_2004	2004	Texas
31073_2004	2004	Nebraska			

Table C.5 Scenarios in cluster 05

senario_name	Year	StateName	senario_name	Year	StateName
8009_2005	2005	Colorado	31057_2005	2005	Nebraska
20023_2005	2005	Kansas	31059_2005	2005	Nebraska
20039_2005	2005	Kansas	31061_2005	2005	Nebraska
20055_2005	2005	Kansas	31063_2005	2005	Nebraska
20063_2005	2005	Kansas	31065_2005	2005	Nebraska
20065_2005	2005	Kansas	31067_2005	2005	Nebraska
20067_2005	2005	Kansas	31073_2005	2005	Nebraska
20069_2005	2005	Kansas	31077_2005	2005	Nebraska
20075_2005	2005	Kansas	31079_2005	2005	Nebraska
20081_2005	2005	Kansas	31081_2005	2005	Nebraska
20089_2005	2005	Kansas	31083_2005	2005	Nebraska
20093_2005	2005	Kansas	31085_2005	2005	Nebraska
20109_2005	2005	Kansas	31087_2005	2005	Nebraska
20119_2005	2005	Kansas	31093_2005	2005	Nebraska
20129_2005	2005	Kansas	31095_2005	2005	Nebraska
20137_2005	2005	Kansas	31097_2005	2005	Nebraska
20147_2005	2005	Kansas	31099_2005	2005	Nebraska
20153_2005	2005	Kansas	31109_2005	2005	Nebraska
20157_2005	2005	Kansas	31119_2005	2005	Nebraska
20171_2005	2005	Kansas	31121_2005	2005	Nebraska
20175_2005	2005	Kansas	31125_2005	2005	Nebraska
20179_2005	2005	Kansas	31127_2005	2005	Nebraska
20181_2005	2005	Kansas	31129_2005	2005	Nebraska
20183_2005	2005	Kansas	31131_2005	2005	Nebraska
20187_2005	2005	Kansas	31137_2005	2005	Nebraska
20189_2005	2005	Kansas	31141_2005	2005	Nebraska
20193_2005	2005	Kansas	31143_2005	2005	Nebraska
20199_2005	2005	Kansas	31145_2005	2005	Nebraska
20203_2005	2005	Kansas	31151_2005	2005	Nebraska
31001_2005	2005	Nebraska	31153_2005	2005	Nebraska
31011_2005	2005	Nebraska	31155_2005	2005	Nebraska
31019_2005	2005	Nebraska	31159_2005	2005	Nebraska
31021_2005	2005	Nebraska	31163_2005	2005	Nebraska
31023_2005	2005	Nebraska	31167_2005	2005	Nebraska
31025_2005	2005	Nebraska	31169_2005	2005	Nebraska
31035_2005	2005	Nebraska	31173_2005	2005	Nebraska
31037_2005	2005	Nebraska	31177_2005	2005	Nebraska
31039_2005	2005	Nebraska	31181_2005	2005	Nebraska
31047_2005	2005	Nebraska	31185_2005	2005	Nebraska
31053_2005	2005	Nebraska	35059_2005	2005	New Mexico
31055_2005	2005	Nebraska	48421_2005	2005	Texas

Table C.6 Scenarios in cluster 06

senario_name	Year	StateName	senario_name	Year	StateName
8009_2006	2006	Colorado	31061_2006	2006	Nebraska
20023_2006	2006	Kansas	31063_2006	2006	Nebraska
20039_2006	2006	Kansas	31065_2006	2006	Nebraska
20055_2006	2006	Kansas	31067_2006	2006	Nebraska
20067_2006	2006	Kansas	31073_2006	2006	Nebraska
20069_2006	2006	Kansas	31077_2006	2006	Nebraska
20081_2006	2006	Kansas	31079_2006	2006	Nebraska
20089_2006	2006	Kansas	31081_2006	2006	Nebraska
20093_2006	2006	Kansas	31083_2006	2006	Nebraska
20109_2006	2006	Kansas	31085_2006	2006	Nebraska
20119_2006	2006	Kansas	31087_2006	2006	Nebraska
20129_2006	2006	Kansas	31093_2006	2006	Nebraska
20137_2006	2006	Kansas	31095_2006	2006	Nebraska
20153_2006	2006	Kansas	31097_2006	2006	Nebraska
20157_2006	2006	Kansas	31099_2006	2006	Nebraska
20171_2006	2006	Kansas	31109_2006	2006	Nebraska
20175_2006	2006	Kansas	31119_2006	2006	Nebraska
20179_2006	2006	Kansas	31121_2006	2006	Nebraska
20181_2006	2006	Kansas	31125_2006	2006	Nebraska
20183_2006	2006	Kansas	31127_2006	2006	Nebraska
20187_2006	2006	Kansas	31129_2006	2006	Nebraska
20189_2006	2006	Kansas	31131_2006	2006	Nebraska
20193_2006	2006	Kansas	31137_2006	2006	Nebraska
20199_2006	2006	Kansas	31141_2006	2006	Nebraska
20203_2006	2006	Kansas	31143_2006	2006	Nebraska
31001_2006	2006	Nebraska	31145_2006	2006	Nebraska
31011_2006	2006	Nebraska	31151_2006	2006	Nebraska
31019_2006	2006	Nebraska	31153_2006	2006	Nebraska
31021_2006	2006	Nebraska	31155_2006	2006	Nebraska
31023_2006	2006	Nebraska	31159_2006	2006	Nebraska
31025_2006	2006	Nebraska	31163_2006	2006	Nebraska
31035_2006	2006	Nebraska	31167_2006	2006	Nebraska
31037_2006	2006	Nebraska	31169_2006	2006	Nebraska
31039_2006	2006	Nebraska	31173_2006	2006	Nebraska
31047_2006	2006	Nebraska	31177_2006	2006	Nebraska
31053_2006	2006	Nebraska	31181_2006	2006	Nebraska
31055_2006	2006	Nebraska	31185_2006	2006	Nebraska
31057_2006	2006	Nebraska	35059_2006	2006	New Mexico
31059_2006	2006	Nebraska	48421_2006	2006	Texas

Table C.7 Scenarios in cluster 07

senario_name	Year	StateName	senario_name	Year	StateName
8009_2007	2007	Colorado	31059_2007	2007	Nebraska
8099_2007	2007	Colorado	31061_2007	2007	Nebraska
20023_2007	2007	Kansas	31063_2007	2007	Nebraska
20039_2007	2007	Kansas	31065_2007	2007	Nebraska
20055_2007	2007	Kansas	31067_2007	2007	Nebraska
20063_2007	2007	Kansas	31073_2007	2007	Nebraska
20065_2007	2007	Kansas	31077_2007	2007	Nebraska
20067_2007	2007	Kansas	31079_2007	2007	Nebraska
20069_2007	2007	Kansas	31081_2007	2007	Nebraska
20071_2007	2007	Kansas	31083_2007	2007	Nebraska
20081_2007	2007	Kansas	31085_2007	2007	Nebraska
20089_2007	2007	Kansas	31087_2007	2007	Nebraska
20093_2007	2007	Kansas	31093_2007	2007	Nebraska
20109_2007	2007	Kansas	31095_2007	2007	Nebraska
20119_2007	2007	Kansas	31097_2007	2007	Nebraska
20129_2007	2007	Kansas	31099_2007	2007	Nebraska
20137_2007	2007	Kansas	31109_2007	2007	Nebraska
20153_2007	2007	Kansas	31119_2007	2007	Nebraska
20157_2007	2007	Kansas	31121_2007	2007	Nebraska
20171_2007	2007	Kansas	31125_2007	2007	Nebraska
20175_2007	2007	Kansas	31127_2007	2007	Nebraska
20179_2007	2007	Kansas	31129_2007	2007	Nebraska
20181_2007	2007	Kansas	31131_2007	2007	Nebraska
20183_2007	2007	Kansas	31137_2007	2007	Nebraska
20187_2007	2007	Kansas	31141_2007	2007	Nebraska
20189_2007	2007	Kansas	31143_2007	2007	Nebraska
20193_2007	2007	Kansas	31145_2007	2007	Nebraska
20199_2007	2007	Kansas	31151_2007	2007	Nebraska
20203_2007	2007	Kansas	31153_2007	2007	Nebraska
31001_2007	2007	Nebraska	31155_2007	2007	Nebraska
31011_2007	2007	Nebraska	31159_2007	2007	Nebraska
31019_2007	2007	Nebraska	31163_2007	2007	Nebraska
31021_2007	2007	Nebraska	31167_2007	2007	Nebraska
31023_2007	2007	Nebraska	31169_2007	2007	Nebraska
31025_2007	2007	Nebraska	31173_2007	2007	Nebraska
31035_2007	2007	Nebraska	31177_2007	2007	Nebraska
31037_2007	2007	Nebraska	31181_2007	2007	Nebraska
31039_2007	2007	Nebraska	31185_2007	2007	Nebraska
31047_2007	2007	Nebraska	35059_2007	2007	New Mexico
31053_2007	2007	Nebraska	48421_2007	2007	Texas
31055_2007	2007	Nebraska	20083_2007	2007	Kansas
31057_2007	2007	Nebraska	20101_2007	2007	Kansas

Table C.8 Scenarios in cluster 08

senario_name	Year	StateName	senario_name	Year	StateName
31001_2008	2008	Nebraska	31095_2008	2008	Nebraska
31011_2008	2008	Nebraska	31097_2008	2008	Nebraska
31019_2008	2008	Nebraska	31099_2008	2008	Nebraska
31021_2008	2008	Nebraska	31109_2008	2008	Nebraska
31023_2008	2008	Nebraska	31119_2008	2008	Nebraska
31035_2008	2008	Nebraska	31125_2008	2008	Nebraska
31037_2008	2008	Nebraska	31127_2008	2008	Nebraska
31039_2008	2008	Nebraska	31129_2008	2008	Nebraska
31047_2008	2008	Nebraska	31131_2008	2008	Nebraska
31053_2008	2008	Nebraska	31137_2008	2008	Nebraska
31057_2008	2008	Nebraska	31141_2008	2008	Nebraska
31059_2008	2008	Nebraska	31143_2008	2008	Nebraska
31061_2008	2008	Nebraska	31145_2008	2008	Nebraska
31063_2008	2008	Nebraska	31151_2008	2008	Nebraska
31065_2008	2008	Nebraska	31153_2008	2008	Nebraska
31067_2008	2008	Nebraska	31155_2008	2008	Nebraska
31073_2008	2008	Nebraska	31159_2008	2008	Nebraska
31077_2008	2008	Nebraska	31163_2008	2008	Nebraska
31079_2008	2008	Nebraska	31167_2008	2008	Nebraska
31081_2008	2008	Nebraska	31169_2008	2008	Nebraska
31083_2008	2008	Nebraska	31177_2008	2008	Nebraska
31085_2008	2008	Nebraska	31181_2008	2008	Nebraska
31087_2008	2008	Nebraska	31185_2008	2008	Nebraska
31093_2008	2008	Nebraska			

Table C.9 Scenarios in cluster 09

senario_name	Year	StateName	senario_name	Year	StateName
20055_2009	2009	Kansas	31083_2009	2009	Nebraska
20075_2009	2009	Kansas	31085_2009	2009	Nebraska
20109_2009	2009	Kansas	31087_2009	2009	Nebraska
20137_2009	2009	Kansas	31093_2009	2009	Nebraska
20171_2009	2009	Kansas	31095_2009	2009	Nebraska
20179_2009	2009	Kansas	31097_2009	2009	Nebraska
20181_2009	2009	Kansas	31099_2009	2009	Nebraska
20195_2009	2009	Kansas	31109_2009	2009	Nebraska
20199_2009	2009	Kansas	31119_2009	2009	Nebraska
31001_2009	2009	Nebraska	31121_2009	2009	Nebraska
31011_2009	2009	Nebraska	31125_2009	2009	Nebraska
31019_2009	2009	Nebraska	31127_2009	2009	Nebraska
31021_2009	2009	Nebraska	31129_2009	2009	Nebraska
31023_2009	2009	Nebraska	31131_2009	2009	Nebraska
31025_2009	2009	Nebraska	31137_2009	2009	Nebraska
31035_2009	2009	Nebraska	31141_2009	2009	Nebraska
31037_2009	2009	Nebraska	31143_2009	2009	Nebraska
31039_2009	2009	Nebraska	31145_2009	2009	Nebraska
31047_2009	2009	Nebraska	31151_2009	2009	Nebraska
31053_2009	2009	Nebraska	31153_2009	2009	Nebraska
31055_2009	2009	Nebraska	31155_2009	2009	Nebraska
31059_2009	2009	Nebraska	31159_2009	2009	Nebraska
31061_2009	2009	Nebraska	31163_2009	2009	Nebraska
31063_2009	2009	Nebraska	31167_2009	2009	Nebraska
31065_2009	2009	Nebraska	31169_2009	2009	Nebraska
31067_2009	2009	Nebraska	31177_2009	2009	Nebraska
31073_2009	2009	Nebraska	31181_2009	2009	Nebraska
31077_2009	2009	Nebraska	31185_2009	2009	Nebraska
31079_2009	2009	Nebraska	20141_2009	2009	Kansas
31081_2009	2009	Nebraska			

Table C.10 Scenarios in cluster 10

senario_name	Year	StateName	senario_name	Year	StateName
20023_2010	2010	Kansas	31073_2010	2010	Nebraska
20055_2010	2010	Kansas	31077_2010	2010	Nebraska
20063_2010	2010	Kansas	31079_2010	2010	Nebraska
20071_2010	2010	Kansas	31081_2010	2010	Nebraska
20075_2010	2010	Kansas	31083_2010	2010	Nebraska
20081_2010	2010	Kansas	31085_2010	2010	Nebraska
20089_2010	2010	Kansas	31087_2010	2010	Nebraska
20109_2010	2010	Kansas	31093_2010	2010	Nebraska
20153_2010	2010	Kansas	31095_2010	2010	Nebraska
20175_2010	2010	Kansas	31097_2010	2010	Nebraska
20179_2010	2010	Kansas	31099_2010	2010	Nebraska
20181_2010	2010	Kansas	31109_2010	2010	Nebraska
20193_2010	2010	Kansas	31119_2010	2010	Nebraska
20199_2010	2010	Kansas	31121_2010	2010	Nebraska
20203_2010	2010	Kansas	31125_2010	2010	Nebraska
31001_2010	2010	Nebraska	31127_2010	2010	Nebraska
31011_2010	2010	Nebraska	31129_2010	2010	Nebraska
31019_2010	2010	Nebraska	31131_2010	2010	Nebraska
31021_2010	2010	Nebraska	31137_2010	2010	Nebraska
31023_2010	2010	Nebraska	31141_2010	2010	Nebraska
31035_2010	2010	Nebraska	31143_2010	2010	Nebraska
31037_2010	2010	Nebraska	31145_2010	2010	Nebraska
31039_2010	2010	Nebraska	31151_2010	2010	Nebraska
31047_2010	2010	Nebraska	31155_2010	2010	Nebraska
31053_2010	2010	Nebraska	31159_2010	2010	Nebraska
31055_2010	2010	Nebraska	31163_2010	2010	Nebraska
31057_2010	2010	Nebraska	31167_2010	2010	Nebraska
31059_2010	2010	Nebraska	31169_2010	2010	Nebraska
31061_2010	2010	Nebraska	31177_2010	2010	Nebraska
31063_2010	2010	Nebraska	31181_2010	2010	Nebraska
31065_2010	2010	Nebraska	31185_2010	2010	Nebraska
31067_2010	2010	Nebraska	20123_2010	2010	Kansas

Table C.11 Scenarios in cluster 11

senario_name	Year	StateName	senario_name	Year	StateName
4003_2001	2001	Arizona	48069_2003	2003	Texas
4003_2002	2002	Arizona	48069_2010	2010	Texas
4003_2003	2003	Arizona	48111_2001	2001	Texas
4003_2004	2004	Arizona	48111_2002	2002	Texas
4003_2005	2005	Arizona	48111_2003	2003	Texas
4003_2006	2006	Arizona	48117_2010	2010	Texas
4003_2007	2007	Arizona	48233_2002	2002	Texas
4003_2008	2008	Arizona	48233_2003	2003	Texas
8009_2002	2002	Colorado	48279_2007	2007	Texas
20055_2002	2002	Kansas	48279_2008	2008	Texas
20067_2002	2002	Kansas	48279_2009	2009	Texas
20069_2002	2002	Kansas	48341_2001	2001	Texas
20081_2002	2002	Kansas	48341_2002	2002	Texas
20093_2002	2002	Kansas	48357_2010	2010	Texas
20119_2002	2002	Kansas	48369_2001	2001	Texas
20119_2009	2009	Kansas	48369_2002	2002	Texas
20175_2002	2002	Kansas	48369_2003	2003	Texas
20175_2009	2009	Kansas	48369_2010	2010	Texas
20187_2002	2002	Kansas	48421_2002	2002	Texas
20187_2010	2010	Kansas	48421_2003	2003	Texas
20189_2002	2002	Kansas	48421_2008	2008	Texas
35009_2001	2001	New Mexico	48421_2009	2009	Texas
35059_2002	2002	New Mexico	48421_2010	2010	Texas
35059_2003	2003	New Mexico	6107_2007	2007	California
35059_2009	2009	New Mexico	6019_2008	2008	California
48065_2002	2002	Texas	6031_2007	2007	California
48069_2001	2001	Texas	6031_2008	2008	California
48069_2002	2002	Texas			

References

Abbasi, A. Z., Islam, N., Shaikh, Z. A., et al. (2014). A review of wireless sensors and networks' applications in agriculture. *Computer Standards & Interfaces*, 36(2):263–270.

Ackerly, D., Loarie, S., Cornwell, W., Weiss, S., Hamilton, H., Branciforte, R., and Kraft, N. (2010). The geography of climate change: implications for conservation biogeography. *Diversity and Distributions*, 16(3):476–487.

Agreste, F. (2016). L'offre mondiale excédentaire en 2015/2016 et la perspective de bonnes récoltes en 2016 continuent de faire pression sur les prix des grains. *Agreste Synthèses - Grandes cultures*.

Agüera, F., Villalobos, F., and Orgaz, F. (1997). Evaluation of sunflower (helianthus annuus, l.) genotypes differing in early vigour using a simulation model. *European Journal of Agronomy*, 7(1-3):109–118.

Atzberger, C. (2013). Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. *Remote Sensing*, 5(2):949–981.

Baey, C., Didier, A., Lemaire, S., Maupas, F., and Cournède, P.-H. (2014). Parametrization of five classical plant growth models applied to sugar beet and comparison of their predictive capacity on root yield and total biomass. *Ecological modelling*, 290:11–20.

Bahremand, A. and De Smedt, F. (2008). Distributed hydrological modeling and sensitivity analysis in torysa watershed, slovakia. *Water Resources Management*, 22(3):393–408.

Baldi, P. and Brunak, S. (2001). *Bioinformatics: the machine learning approach*. MIT press.

Bausch, W. C. (1995). Remote sensing of crop coefficients for improving the irrigation scheduling of corn. *Agricultural Water Management*, 27(1):55–68.

Bayol, B. (2016). *Système informatique d'aide à la modélisation mathématique basé sur un langage de programmation dédié pour les systèmes dynamiques discrets stochastiques. Application aux modèles de croissance de plantes.* PhD thesis, Université Paris-Saclay.

Bechini, L., Bocchi, S., Maggiore, T., and Confalonieri, R. (2006). Parameterization of a crop growth and development simulation model at sub-model components level. an example for winter wheat (triticum aestivum l.). *Environmental Modelling & Software*, 21(7):1042–1054.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Boote, K., Kropff, M., and Bindraban, P. (2001). Physiology and modelling of traits in crop plants: implications for genetic improvement. *Agricultural Systems*, 70(2-3):395–420.

Boote, K. J., Jones, J. W., and Hoogenboom, G. (1998). Simulation of crop growth: Cropgro model.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brisson, N., Gary, C., Justes, E., Roche, R., Mary, B., Ripoche, D., Zimmer, D., Sierra, J., Bertuzzi, P., Burger, P., et al. (2003). An overview of the crop model stics. *European Journal of agronomy*, 18(3):309–332.

Brisson, N., Gate, P., Gouache, D., Charmet, G., Oury, F.-X., and Huard, F. (2010). Why are wheat yields stagnating in europe? a comprehensive data analysis for france. *Field Crops Research*, 119(1):201–212.

Brisson, N., Mary, B., Ripoche, D., Jeuffroy, M. H., Ruget, F., Nicoullaud, B., Gate, P., Devienne-Barret, F., Antonioletti, R., Durr, C., et al. (1998). Stics: a generic model for the simulation of crops and their water and nitrogen balances. i. theory and parameterization applied to wheat and corn. *Agronomie*, 18(5-6):311–346.

Brun, F., Wallach, D., Makowski, D., and Jones, J. W. (2006). *Working with dynamic crop models: evaluation, analysis, parameterization, and applications*. Elsevier.

Brun, R., Reichert, P., and Künsch, H. R. (2001). Practical identifiability analysis of large environmental simulation models. *Water Resources Research*, 37(4):1015–1030.

Cariboni, J., Gatelli, D., Liska, R., and Saltelli, A. (2007). The role of sensitivity analysis in ecological modelling. *Ecological modelling*, 203(1):167–182.

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

Chang, J.-F., Roddick, J. F., Pan, J.-S., and Chu, S. (2005). A parallel particle swarm optimization algorithm with communication strategies. *Journal of Information Science and Engineering*, (21):809–818.

Change, I. C. (2014). Synthesis report summary for policymakers. 2014. *URl: https://www. ipcc. ch/pdf/assessment-report/ar5/syr/AR5_SYR_FINAL_SPM. pdf.*

Chapman, B., Jost, G., and Van Der Pas, R. (2008). *Using OpenMP: portable shared memory parallel programming*, volume 10. MIT press.

Chen, S., Gopalakrishnan, P., et al. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. darpa broadcast news transcription and understanding workshop*, volume 8, pages 127–132. Virginia, USA.

Chen, Y. (2014). *Inférence bayésienne dans les modèles de croissance de plantes pour la prévision et la caractérisation des incertitudes*. PhD thesis, Châtenay-Malabry, Ecole centrale de Paris.

Christophe, A., Letort, V., Hummel, I., Cournède, P.-H., de Reffye, P., and Lecœur, J. (2008). A model-based analysis of the dynamics of carbon balance at the whole-plant level in arabidopsis thaliana. *Functional Plant Biology*, 35(11):1147–1162.

Clerc, M. (2005). *L'optimisation par essaims particulaires*. Hermes Science Publications.

Clerc, M. and Kennedy, J. (2002). The particle swarm-explosion, stability, and convergence in a multidimensional complex space. *IEEE transactions on Evolutionary Computation*, 6(1):58–73.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cournède, P.-H., Chen, Y., Wu, Q., Baey, C., and Bayol, B. (2013). Development and evaluation of plant growth models: Methodology and implementation in the pygmalion platform. *Mathematical Modelling of Natural Phenomena*, 8(4):112–130.

Cournède, P.-H., Letort, V., Mathieu, A., Kang, M. Z., Lemaire, S., Trevezas, S., Houllier, F., and De Reffye, P. (2011). Some parameter estimation issues in functional-structural plant modelling. *Mathematical Modelling of Natural Phenomena*, 6(2):133–159.

Cristianini, N., Shawe-Taylor, J., et al. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Cruz, J. A. and Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2:117693510600200030.

Dagum, L. and Menon, R. (1998). Openmp: an industry standard api for shared-memory programming. *IEEE computational science and engineering*, 5(1):46–55.

De Reffye, P., Heuvelink, E., Barthémémy, D., and Cournède, P.-H. (2008). Plant growth models. In Jorgensen, S. and Fath, B., editors, *Ecological Models. Vol. 4 of Encyclopedia of Ecology (5 volumes)*, pages 2824–2837. Elsevier (Oxford).

De Reffye, P. and Hu, B.-G. (2003). Relevant qualitative and quantitative choices for building an efficient dynamic plant growth model: Greenlab case. In *International Symposium on Plant Growth Modeling, Simulation, Visualization and their Applications-PMA'03*, pages 87–107. Springer and Tsinghua University Press.

De Wit, C. T. (1978). Simulation of assimilation, respiration and transpiration of crops.

Della Noce, A., Letort, V., Hansart, A., Baey, C., Viaud, G., Barot, S., Lata, J.-C., Raynaud, X., Cournède, P.-H., and Gignoux, J. (2016). Modeling the inter-individual variability of single-stemmed plant development. In *Functional-Structural Plant Growth Modeling, Simulation, Visualization and Applications (FSPMA), International Conference on*, pages 44–51. IEEE.

Dietterich, T. G. et al. (2002). Ensemble learning. *The handbook of brain theory and neural networks*, 2:110–125.

DISAR, D. I. d. S. A. d. R. (2018). Chiffres et analyses.

Dixon, B. L., Hollinger, S. E., Garcia, P., and Tirupattur, V. (1994). Estimating corn yield response models to predict impacts of climate change. *Journal of Agricultural and resource economics*, pages 58–68.

Dréo, J., Pétrowski, A., Siarry, P., and Taillard, E. (2006). *Metaheuristics for hard optimization: methods and case studies*. Springer Science & Business Media.

Drummond, S. T., Sudduth, K. A., Joshi, A., Birrell, S. J., and Kitchen, N. R. (2003). Statistical and neural methods for site–specific yield prediction. *Transactions of the ASAE*, 46(1):5.

Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least squares regression: I. *Biometrika*, 37(3/4):409–428.

Eberhart, R. and Kennedy, J. (1995). A new optimizer using particle swarm theory. In *Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on*, pages 39–43. IEEE.

Eberhart, R. C. and Shi, Y. (1998). Comparison between genetic algorithms and particle swarm optimization. In *International conference on evolutionary programming*, pages 611–616. Springer.

Eurostat, E. U. (2017). Survey on agricultural production methods.

Fan, H. (2001). Study on vmax of particle swarm optimization. In *Proceedings of the Workshop on Particle Swarm Optimization, Indianapolis, 2001*.

Fang, H., Liang, S., and Hoogenboom, G. (2011). Integration of modis lai and vegetation index products with the csm–ceres–maize model for corn yield estimation. *International Journal of Remote Sensing*, 32(4):1039–1065.

Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI magazine*, 13(3):57–57.

Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, 1(1):55–77.

GISS, Goddard Space Flight Center, N. (2014). Agmerra and agcfsr climate forcing datasets for agricultural modeling. *See Also: https://stats.agriculture.gouv.fr/disar-web/*.

Godwin, D. and Jones, C. A. (1991). Nitrogen dynamics in soil-plant systems. *Modeling plant and soil systems*, (modelingplantan):287–321.

Goodwin, G. C. and Payne, R. L. (1977). *Dynamic system identification: experiment design and data analysis*, volume 26. Academic press New York.

Goudriaan, J. and van Laar, H. (1994). Modelling potential crop growth processes kluwer academic publishers.

Gropp, W., Lusk, E., Doss, N., and Skjellum, A. (1996). A high-performance, portable implementation of the mpi message passing interface standard. *Parallel computing*, 22(6):789–828.

Guo, Y., Ma, Y., Zhan, Z., Li, B., Dingkuhn, M., Luquet, D., and De Reffye, P. (2006). Parameter optimization and field validation of the functional–structural model greenlab for maize. *Annals of botany*, 97(2):217–230.

Hagan, M. T., Demuth, H. B., Beale, M. H., and De Jesús, O. (1996). *Neural network design*, volume 20. Pws Pub. Boston.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hamby, D. (1994). A review of techniques for parameter sensitivity analysis of environmental models. *Environmental monitoring and assessment*, 32(2):135–154.

Hammer, G., Kropff, M., Sinclair, T., and Porter, J. (2002). Future contributions of crop modelling—from heuristics and supporting decision making to understanding genetic regulation and aiding crop improvement. *European Journal of Agronomy*, 18(1-2):15–31.

Hatfield, P. and Pinter Jr, P. (1993). Remote sensing for crop protection. *Crop protection*, 12(6):403–413.

He, J. (2009). Advances in data mining: History and future. In *2009 Third International Symposium on Intelligent Information Technology Application*, volume 1, pages 634–636. IEEE.

Helton, J. C., Johnson, J. D., Sallaberry, C. J., and Storlie, C. B. (2006). Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering & System Safety*, 91(10):1175–1209.

Hill, M. C. and Tiedeman, C. R. (2006). *Effective groundwater model calibration: with analysis of data, sensitivities, predictions, and uncertainty*. John Wiley & Sons.

Hu, B.-G., De Reffye, P., Zhao, X., Yan, H.-P., and Kang, M. Z. (2003). Greenlab: A new methodology towards plant functional-structural model–structural part. In *Plant growth modelling and applications*, pages 21–35. TsingHua University Press and Springer.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304.

Huete, A. and Jackson, R. (1987). Suitability of spectral indices for evaluating vegetation characteristics on arid rangelands. *Remote sensing of environment*, 23(2):213–IN8.

IBP, I. (2015). *France Investment and Business Guide*, volume 1. lulu.com.

Indyk, P. and Motwani, R. (1998). Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM.

Iniesta, R., Stahl, D., and McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological medicine*, 46(12):2455–2465.

Iooss, B. and Lemaître, P. (2015). A review on global sensitivity analysis methods. In *Uncertainty Management in Simulation-Optimization of Complex Systems*, pages 101–122. Springer.

Irmak, A., Jones, J., Batchelor, W., Irmak, S., Boote, K., and Paz, J. (2006). Artificial neural network model as a data analysis tool in precision farming. *Transactions of the ASABE*, 49(6):2027–2037.

Ishibuchi, H. and Murata, T. (1998). A multi-objective genetic local search algorithm and its application to flowshop scheduling. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 28(3):392–403.

Jacques, J. (2005). *Contributions à l'analyse de sensibilité et à l'analyse discriminante généralisée*. PhD thesis, Université Joseph-Fourier-Grenoble I.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Jeuffroy, M., Barbottin, A., Jones, J., and Lecoeur, J. (2006). Crop models with genotype parameters. *Working with crop models'.(Eds D Wallach, D Makowski, JW Jones) pp*, pages 281–308.

John Walker, S. (2014). Big data: A revolution that will transform how we live, work, and think.

Jones, C. A., Kiniry, J. R., and Dyke, P. (1986). *CERES-Maize: A simulation model of maize growth and development*. Texas A and M University Press.

Kandel, E. R., Schwartz, J. H., Jessell, T. M., of Biochemistry, D., Jessell, M. B. T., Siegelbaum, S., and Hudspeth, A. (2000). *Principles of neural science*, volume 4. McGraw-hill New York.

Kang, F. (2013). *Modèles de croissance de plantes et méthodologies adaptées à leur paramétrisation pour l'analyse des phénotypes*. PhD thesis, Châtenay-Malabry, Ecole centrale de Paris.

Keating, B. A., Carberry, P. S., Hammer, G. L., Probert, M. E., Robertson, M. J., Holzworth, D., Huth, N. I., Hargreaves, J. N., Meinke, H., Hochman, Z., et al. (2003). An overview of apsim, a model designed for farming systems simulation. *European journal of agronomy*, 18(3-4):267–288.

Kennedy, J. (1998). The behavior of particles. In *Evolutionary programming VII*, pages 579–589. Springer.

Kennedy, J. (1999). Small worlds and mega-minds: effects of neighborhood topology on particle swarm performance. In *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, volume 3, pages 1931–1938. IEEE.

Kennedy, J., Kennedy, J. F., Eberhart, R. C., and Shi, Y. (2001). *Swarm intelligence*. Morgan Kaufmann.

Khaledian, M., Mailhol, J., Ruelle, P., and Rosique, P. (2009). Adapting pilote model for water and yield management under direct seeding system: The case of corn and durum wheat in a mediterranean context. *Agricultural Water Management*, 96(5):757–770.

Kim, Y., Evans, R. G., and Iversen, W. M. (2008). Remote sensing and control of an irrigation system using a distributed wireless sensor network. *IEEE transactions on instrumentation and measurement*, 57(7):1379–1387.

Kogan, J. (2007). *Introduction to clustering large and high-dimensional data*. Cambridge University Press.

Kumar, V., Grama, A., Gupta, A., and Karypis, G. (1994). *Introduction to parallel computing: design and analysis of algorithms*, volume 400. Benjamin/Cummings Redwood City.

Lê Cao, K.-A., Boitard, S., and Besse, P. (2011). Sparse pls discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC bioinformatics*, 12(1):253.

Lecoeur, J., Poir-Lassus, R., Christophe, A., Pallas, B., Casadebaig, P., Debaeke, P., Vear, F., and Guilioni, L. (2011). Quantifying physiological determinants of genetic variation for yield potential in sunflower. sunflo: a model-based analysis. *Functional Plant Biology*, 38:246–259.

LeCun, Y., Jackel, L., Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Muller, U., Sackinger, E., Simard, P., et al. (1995). Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective*, 261:276.

Lehuger, S., Gabrielle, B., Van Oijen, M., Makowski, D., Germon, J.-C., Morvan, T., and Hénault, C. (2009). Bayesian calibration of the nitrous oxide emission module of an agro-ecosystem model. *Agriculture, Ecosystems & Environment*, 133(3-4):208–222.

Letort, V., Cournède, P.-H., Mathieu, A., De Reffye, P., and Constant, T. (2008). Parametric identification of a functional–structural tree growth model and application to beech trees (fagus sylvatica). *Functional plant biology*, 35(10):951–963.

Li, H., He, H., and Wen, Y. (2015). Dynamic particle swarm optimization and k-means clustering algorithm for image segmentation. *Optik*, 126(24):4817–4822.

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

Likas, A., Vlassis, N., and Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461.

Lillesand, T., Kiefer, R. W., and Chipman, J. (2014). *Remote sensing and image interpretation*. John Wiley & Sons.

Liu, J., Goering, C., and Tian, L. (2001). A neural network for setting target corn yields. *Transactions of the ASAE*, 44(3):705.

Lobell, D. B. and Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, 150(11):1443–1452.

Ma, Y., Wen, M., Guo, Y., Li, B., Cournède, P.-H., and De Reffye, P. (2007). Parameter optimization and field validation of the functional–structural model greenlab for maize at different population densities. *Annals of botany*, 101(8):1185–1194.

Ma, Y., Wubs, A. M., Mathieu, A., Heuvelink, E., Zhu, J., Hu, B.-G., Cournède, P.-H., and De Reffye, P. (2010). Simulation of fruit-set and trophic competition and optimization of yield advantages in six capsicum cultivars using functional–structural plant modelling. *Annals of botany*, 107(5):793–803.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Makowski, D., Naud, C., Jeuffroy, M.-H., Barbottin, A., and Monod, H. (2006). Global sensitivity analysis for calculating the contribution of genetic parameters to the variance of crop model prediction. *Reliability Engineering & System Safety*, 91(10-11):1142–1147.

Messina, C., Boote, K., Löffler, C., Jones, J., and Vallejos, C. (2006). Model-assisted genetic improvement of crops. *Working with Dynamic Crop Models: Evaluation, Analysis, Parameterization, and Applications*, pages 309–335.

Meynard, J.-M., Cerf, M., Guichard, L., Jeuffroy, M.-H., and Makowski, D. (2002). Which decision support tools for the environmental management of nitrogen? *Agronomie*, 22(7-8):817–829.

Midmore, E. K., McCartan, S. A., Jinks, R. L., Cahalan, C. M., et al. (2015). Using thermal time models to predict germination of five provenances of silver birch (betula pendula roth) in southern england. *SILVA FENNICA*, 49(2).

Monteith, J. (1977). Climate and the efficiency of crop production in britain. *Proceedings of the Royal Society of London B*, 281:277–294.

Montgomery, D. C. and Runger, G. C. (2010). *Applied statistics and probability for engineers*. John Wiley & Sons.

Newell, A. (1982). Intellectual issues in the history of artificial intelligence. Technical report.

Newlands, N. K. and Townley-Smith, L. (2010). Predicting energy crop yield using bayesian networks. In *Proceedings of the Fifth IASTED International Conference*, volume 711, pages 014–106.

Nossent, J., Elsen, P., and Bauwens, W. (2011). Sobol'sensitivity analysis of a complex environmental model. *Environmental Modelling & Software*, 26(12):1515–1525.

O'Neil, C. and Schutt, R. (2013). *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc.".

Oteng-Darko, P., Yeboah, S., Addy, S., Amponsah, S., and Danquah, E. O. (2013). Crop modeling: A tool for agricultural research. *J. Agricultural Res. Develop.*, 2(1):001–006.

Peeples, M. A. (2011). R script for k-means cluster analysis. *Retrieved May*, 13:2016.

Porter, J. R. (1993). Afrcwheat2: a model of the growth and development of wheat incorporating responses to water and nitrogen. *European Journal of Agronomy*, 2(2):69–82.

Prasad, A. K., Chai, L., Singh, R. P., and Kafatos, M. (2006a). Crop yield estimation model for iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation*, 8(1):26–33.

Prasad, A. M., Iverson, L. R., and Liaw, A. (2006b). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9(2):181–199.

Punj, G. and Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of marketing research*, pages 134–148.

Reymond, M., Muller, B., Leonardi, A., Charcosset, A., and Tardieu, F. (2003). Combining quantitative trait loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. *Plant Physiology*, 131:664–675.

Reynolds, C. W. (1987). Flocks, herds and schools: A distributed behavioral model. *ACM SIGGRAPH computer graphics*, 21(4):25–34.

Roel, A. and Plant, R. E. (2004). Factors underlying yield variability in two california rice fields. *Agronomy Journal*, 96(5):1481–1494.

Ruane, A. C., Goldberg, R., and Chryssanthacopoulos, J. (2015). Climate forcing datasets for agricultural modeling: Merged products for gap-filling and historical climate series estimation. *Agricultural and Forest Meteorology*, 200:233–248.

Russell, S. J., Norvig, P., Canny, J. F., Malik, J. M., and Edwards, D. D. (2003). *Artificial intelligence: a modern approach*, volume 2. Prentice hall Upper Saddle River.

Ryan, T. P. (2008). *Modern regression methods*, volume 655. John Wiley & Sons.

Safa, B., Khalili, A., Teshnehlab, I. M., and Liaghat, A. (2004). Artificial neural networks application to predict wheat yield using climatic data. *parameters*, 7:8.

Sainte-Marie, J., Viaud, G., and Cournède, P.-H. (2017). Indices de sobol généralisés aux variables dépendantes: tests de performance de l'algorithme hogs couplé à plusieurs estimateurs paramétriques. *Journal de la Société Française de Statistique*, 158(1):68–89.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, 81.

Saltelli, A. (2002). Sensitivity analysis for importance assessment. *Risk analysis*, 22(3):579–590.

Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. (2004). *Sensitivity analysis in practice: a guide to assessing scientific models*. John Wiley & Sons.

Saporta, G. (2006). *Probabilités, analyse des données et statistique*. Editions Technip.

Schölkopf, B., Smola, A. J., Bach, F., et al. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Shi, Y. and Eberhart, R. (1998). A modified particle swarm optimizer. In *Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on*, pages 69–73. IEEE.

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222.

Sobol, I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical modelling and computational experiments*, 1(4):407–414.

Sudduth, K., Drummond, S., Birrell, S. J., and Kitchen, N. (1996). Analysis of spatial factors influencing crop yield. *Precision Agriculture*, (precisionagricu3):129–139.

Talbi, E.-G. (2002). A taxonomy of hybrid metaheuristics. *Journal of heuristics*, 8(5):541–564.

Tarsha-Kurdi, F., Landes, T., Grussenmeyer, P., and Koehl, M. (2007). Model-driven and data-driven approaches using lidar data: Analysis and comparison. In *ISPRS Workshop, Photogrammetric Image Analysis (PIA07)*, pages 87–92.

Thangaraj, R., Pant, M., Abraham, A., and Bouvry, P. (2011). Particle swarm optimization: hybridization perspectives and experimental illustrations. *Applied Mathematics and Computation*, 217(12):5208–5226.

Therneau, T., Atkinson, B., Ripley, B., and Ripley, M. B. (2015). Package 'rpart'. *Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016).*

Van Den Bergh, F. (2007). *An analysis of particle swarm optimizers*. PhD thesis, University of Pretoria.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Varcoe, V. (1990). A note on the computer simulation of crop growth in agricultural land evaluation. *Soil Use and Management*, 6(3):157–160.

Viaud, G. (2018). *Méthodes statistiques pour la différenciation génotypique des plantes à l'aide des modèles de croissance*. PhD thesis, Université Paris-Saclay.

Von Storch, H. (1999). Misuses of statistical analysis in climate research. In *Analysis of Climate Variability*, pages 11–26. Springer.

Walid Hammache, T. L. and Cournède, P.-H. (2018). Crop recognition from sequential sentinel-1 images with lstm recurrent networks. *proceeding of PMA2018: 6th International Symposium on Plant Growth Modeling, Simulation, Visualization and Application*.

Wang, N., Zhang, N., and Wang, M. (2006). Wireless sensors in agriculture and food industry—recent development and future perspective. *Computers and electronics in agriculture*, 50(1):1–14.

Williams, J., Jones, C., and Dyke, P. T. (1984). A modeling approach to determining the relationship between erosion and soil productivity. *Transactions of the ASAE*, 27(1):129–0144.

Williams, M. M. (2012). Agronomics and economics of plant population density on processing sweet corn. *Field Crops Research*, 128:55–61.

Wright, S. J. and Nocedal, J. (1999). Numerical optimization. *Springer Science*, 35(67-68):7.

Wu, Q. (2012). *Sensitivity Analysis for Functional Structural Plant Modelling*. PhD thesis, Ecole Centrale Paris.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37.

Xin, B., Chen, J., Zhang, J., Fang, H., and Peng, Z.-H. (2012). Hybridizing differential evolution and particle swarm optimization to design powerful optimizers: a review and taxonomy. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(5):744–767.

Zhan, Z.-G., De Reffye, P., Houllier, F., and Hu, B.-G. (2003). Fitting a functional-structural growth model with plant architectural data. In *International Symposium on Plant Growth Modeling, Simulation, Visualization and their Applications-PMA'03*, pages 108–117. Springer and Tsinghua University Press.

Zhang, L., Tan, J., Han, D., and Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug discovery today*, 22(11):1680–1685.

Zhou, Y. and Tan, Y. (2009). Gpu-based parallel particle swarm optimization. In *Evolutionary Computation, 2009. CEC'09. IEEE Congress on*, pages 1493–1500. IEEE.

Zielinski, K. and Laur, R. (2007). Stopping criteria for a constrained single-objective particle swarm optimization algorithm. *Informatica*, 31(1).

ÉCOLE DOCTORALE
INTERFACES
Approches Interdisciplinaires : Fondements, Applications et Innovation

Titre : Méthodologie d'Apprentissage Statistique Profitant de la Diversité des Scénarios Environnementaux pour les Modèles de Cultures. Application à la Prévision de la Production Végétale à Grande-échelle

Mots clés : Prédiction du rendement des cultures, approches basées sur les connaissances, approches basées sur les données, MuScPE, diversité environnementale, grande échelle

Résumé : La prévision du rendement des cultures est une question primordiale en agriculture. Des recherches considérables ont été menées dans ce but, en s'appuyant sur diverses méthodologies. Généralement, ils peuvent être classés en approches basées sur les connaissances et approches basées sur les données. Tous les deux ont leurs avantages et leurs inconvénients.
Pour les méthodes basées sur la connaissance, elles sont basées sur la description mécanique des processus biophysiques et impliquent potentiellement un grand nombre de variables et de paramètres d'état, dont l'estimation n'est pas simple. Une nouvelle stratégie de modélisation statistique d'estimation de paramètres à scénarios multiples (MuScPE) est proposée pour profiter de données avec un accès facile et de la diversité des scénarios environnementaux. Il est testé avec un ensemble de données sur le maïs au milieu des États-Unis. Un résultat satisfaisant est obtenu avec un modèle mécanique nommé CORNFLO.
De l'autre, tant de données différentes ont été proposées avec une philosophie variée. Une comparaison systémique de cette méthode guidée par le modèle a été effectuée pour satisfaire les données sous divers formats. Les personnes choisies ayant les meilleures aptitudes et prévisions ont été comparées à des modèles axés sur les connaissances.
Enfin, une régression pondérée est appliquée à la prévision du rendement à grande échelle. La production de blé tendre en France est prise comme un exemple. Les approches axées sur les connaissances et sur les données ont également été comparées pour leurs performances.

Title : Statistical Learning Methodology to Leverage the Diversity of Environmental Scenarios in Crop Data. Application to the Prediction of Crop Production at Large-Scale

Keywords : Crop yield prediction, knowledge-driven approaches, data-driven approaches, MuScPE, environmental diversity, large scale

Abstract : Crop yield prediction is a paramount issue in agriculture. Considerable research has been performed with this objective, relying on various methodologies. Generally, they can be classified into knowledge-driven approaches and data-driven approaches. Both have their advantages and shortcomings.
For knowledge-driven methods, they are based on the mechanical description of biophysical processes, and they potentially imply a large number of state variables and parameters, whose estimation is not straightforward. A new statistical modelling strategy Multiple-Scenarios Parameter Estimation (MuScPE) is proposed to take advantage of the dataset with easy access and leverage the diversity of environmental scenarios. It is tested with a dataset about the corn in the middle of the United States. A satisfactory result is achieved with a mechanical crop model named CORNFLO.
On the other, so many different data-driven have been proposed with variate philosophy. A systemic comparison of this model-driven method has been made to satisfy data in diverse format. The chosen ones with best fitness and prediction have been compared with knowledge-driven models.
At last, a weighted regression is applied to large-scale yield prediction. Soft wheat production in France is taken as an example. Model-driven and data-driven approaches have also been compared for their performances in achieving this goal.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France