The advent of Big Data in Agriculture

Salima TAÏBI [1,2] & Jérôme DANTAN [2]

[1] *Head of Department of Physics, Mathematics and Computer Sciences*
E-mail : *staibi@esitpa.fr*

[2] *Institut Polytechnique LaSalle Beauvais-Esitpa*
E-mail : *jdantan@esitpa.fr*

Abstract
Nowadays the number of data is huge! Companies and research centers are flooded by the amount of data. Agriculture is also affected and especially since the use of connected objects. So how research in statistic could develop a methodology to manage, organize and analyze the data, particularly when data are heterogeneous, collinear, mixed between qualitative and quantitative data?

Keywords: Big Data, Random Forests, prediction models, qualitative variables, multiple correspondence analysis.

Résumé
Aujourd'hui, le nombre de données ne cesse de croître! Les entreprises et les centres de recherche sont inondés par la quantité de données. L'agriculture est également affectée et ce surtout depuis l'utilisation d'objets connectés. Comment la recherche en statistique pourrait offrir une méthodologie pour gérer, organiser et analyser les données, en particulier lorsque les données sont hétérogènes, colinéaires, ou encore lorsqu'il y a présence de données qualitatives et quantitatives?

Mots-clés : Big Data, Forêts aléatoires, modèles de prédiction, variables qualitatives, analyse des correspondances multiples.

1. Introduction

Agriculture is affected by big data, especially since the use of connected objects; this raises many questions on the part of researchers. How to structure data for better management? How to check the accuracy of such data? How to take into account their heterogeneity? How to treat a mix of qualitative and quantitative data? How to explore step by step and automatically analyze these data? How to transform the results into decision tools? How can we develop a collaborative platform for a research project?
Through two research projects and by using Random Forests Method (Breiman 2001) we try to answer to these questions. In fact Random Forest is powerful in the case of Big Data. In addition, Random Forests can overcome the correlation problem. We will use the extension of this method to the case of qualitative data (Laroutis and Taibi 2011). Indeed the extension of the Random Forest method in case of qualitative data is an innovation and can offer many prospects for applications in agriculture, biology food science, socio-economic surveys,… We compare Random Forest to others predictive methods.

2. Random forest and qualitative data

2.1 Impact of erosion: a regional project

2.1.1 Context

In France, 4 million hectares are now affected by erosion. In front of these inconveniences, policies evolve and communities are seeking preventive and curative solutions. We analyze the behavior of people living in the "Valley of Commerce" (Normandie) in relation to risk of runoff and flooding and identify the level of risk perception by people and their awareness to reduce these risks.

2.1.2 The survey

The purpose of the survey is to judge whether it is possible or not to implement this program against the erosive runoff in the valley of Commerce.
This program would be over twenty years and involve an additional fee for the inhabitants of this valley. This survey is interesting for the benefit that might have the inhabitants through such a program. We are in the general framework of survey data analysis; we wish on the one hand to predict individuals "buy-in" to the Seine Estuary wetlands conservation program, and on the other hand to be able to use the same methodology for successive waves of the survey. We have limited our analysis to the case of a binary dependant variable (participation or not in the conservation program) – but the procedure can be extended to cases of multinomial qualitative variables.
In this study, we used two different classification methods – Discriminant Analysis and Random Forests (RF) – with the aim of comparing their performance.
Peters et al. (2008) underline the absence of studies using Random Forests in some domains, a fact which makes comparisons and analyses difficult. For example, very few studies have been done in the area of ecological modeling.

2.2 Random Forests

Leo Breiman (2001) proposed a family of classification methods called Random Forests (RF).
based on the concept of randomization.
The RF method proposes an improvement to bagging for the case of binary trees, making the trees more independent by adding randomness to the process of selecting the variables in the model. This approach has been successful in problems with multiple variables and Random forest performed better than regression tree (Lieb and al. 2012).To address this issue, "bootstrap" samples are used, obtained from the initial sample population by random selection with replacement, instead of independent samples.
Let *n* be the size of the training sample. A decision tree is built according to the following algorithm.
1. Select a set of *n* individuals (with replacement) which will be used for the tree.
2. For *p* variables, select a sub-set of K variables. The best subset is used for partitioning.
3. The tree is built in this way until it reaches its maximum size.
This process implies choosing two parameters: K and the number of trees m. The "hyperparameter" K is taken to be $\sqrt{p}$ or log (p+1).
The results below were generated using the *Random Forests* algorithm in the Data Miner module of STATISTICA.

2.3 Treatment of the qualitative variables

As the predictors are mostly qualitative variables, we use a procedure based on the Disqual method (Saporta, 1977) to transform them into quantitative data. We carried out Multiple Correspondence Analysis of the predictors. The p variables X_1 , X_2 , …, X_p are replaced by n individuals on the q factorial axes (q< p) with a weighting process which allows us to preserve the inertia of each component.

2.4. Results

As we see below (Table 1), classification by RF (70%) is less than FDA (76%). But by Mc Nemar test we show that there is no significant difference between FDA and RF.

Models	Kappa	Pcc
Random Forests	0,48	70%
Discriminant Analysis	0,61	76%

Table 1: accuracy and performance of the models

3. A statistical approach using Random Forest to assess biodiversity

3.1 Context

The objectives of the national Bioindicators Program (2005-2012) were to provide a base of referential values for various land-use contexts and to measure relevant biological tools in agricultural but also polluted contexts in the aim to transfer them to stakeholders.
Our aim is to present how statisticians, computer scientists have worked together with ecologists, biologists, chemical researchers in a multidisciplinary project. Soil is a dynamic and complex system. Information emanating from compartments “Microflora” “Flora” or “Fauna” are abundant and in interaction with the environment, climatic conditions and time data.
Given that more than 200,000 observations were obtained, it was necessary to design a database and some analytical process.

3.2 Global methodology

First we plan a methodology to assess the quality of soil. Each researcher or responsible of the biologic indicator had to send data to a data manager. To validate the accuracy and the quality of data, programs are implemented to calculate basic statistics. The approach we have developed and applied in the context of this project is below:

1. Database design
2. Data quality
3. Development of an “interface user”
4. Automation software data extraction
5. Homogenization of data processing
6. Definition of the baseline
7. Outline of the sensitivity of the indicator
8. Selection of sets of indicators
9. Indicator support tools for risk assessment
10. Global analysis

3.3 Selection biologic indicators

We use Random Forest method to select the smallest set of discriminating variables in each compartment. We choose Gini Index and we split the sample into 2/3 for the learning sample and 1/3 for the test sample. We compare the performance between Logistic Model and Random Forests method. We conclude by Mc Nemar test that there is no significant difference between the two models.

Models	Kappa	Pcc
Random Forests	0,78	77,3%
Logistic	0,73	75,2%

Table 2: accuracy and performance of the models

3.4 Decision tool for soil quality

We have built an approach to design a functional index of a soil (Taibi et al., 2011, 2015) taking into account the almost information of the data base.
This index quality soil is defined as the smallest set of parameters which gives information on the capacity of a complex dynamic system such as the soil.
The results of the Random Forests returned by the Gini index allow us to determine a score of sensitivity for each biologic indicator.
As this score is the most important, the global score can be determined as follows:

$$Global\ Score = Sensitivity\ Score\ \times (aScore\ _{Tech} + bScore\ _{socio-eco})$$

where $Score\ _{Tech}$ et $Score\ _{socio-eco}$ are determined by surveys data analysis (experts).

4. Conclusion

The data mining and statistical tools used here allow us to take into account the fair-recurring problems in biology studies such as the heterogeneity, "collinearity" and presence of missing data.
Nonparametric models and particularly Random Forests give benefits in such environmental and biological studies (Cutler and al. 2007).
The volume of data, and the large number of biological variables to be tested (one hundred), require analytical techniques, such as Random Forests, which can overcome the problem of multicollinearity to select indicators sensitive to various factors.
Random Forests is more appropriate to discriminate a lot of variables. Random Forests are a solution for biological data mining and classification; in fact distributions of observations are often unknown. The Random Forests allow us to select the smallest set of discriminating variables and our statistical approach will lead to the development of prediction tools in socio-economic, biologic, surveys or in other contexts.

Acknowledgement

The authors are grateful to Jeanne-Chantal Dur, Laurence Rougé, Antonio Bispo and Guenola Perez and wish to thank the Region Haute Normandie and the ADEME (French national agency of environment and energy management.

References

Breiman, L., (2001). Random Forests. *Machine Learning*, 45, 5-32.

Cutler D.R., Edwards T.C., Beard K.H., Cutler A., Hess K.T., Gibson J., Lawler J., (2007). Random Forests for classification in ecology. *Ecology*, 88,2783-2792.

Laroutis, D., Taibi, S., (2011). Discriminant analysis versus Random Forests on qualitative data: Contingent valuation method applied to the Seine estuary wetlands. *International Journal of Ecological Economics & Statistics*, 20, 1-13.

Lieb M., Glaser B., Huwe B., (2012). Uncertainty in the spatial prediction of soil texture. Comparison of regression tree and Random Forest models. *Geoderma*, 170, 70-79.

Peters J., Verhoest N.E.C., Samson R., Boeckx P., De Baets B., (2007). Wetlands vegetation distribution modeling for the identification of constraining environmental
variables. *Landscape Ecology*, 23, 1049-1065.

Saporta G.,(1977). Une méthode et un programme d'analyse discriminante pas à pas sur variables qualitatives, INRIA *Analyse des données et informatique*, 1, 201-210.

Taïbi-Hassani S., Lepelletier P., Blot A., Thoisy-Dur J-C. , (2015). A statistical approach to the evaluation and modeling of contamination in an agro-ecosystem. International Journal of Ecological Economics & Statistics. 36 (1), 83-97, 2015.

Taïbi-Hassani, S., Thoisy-Dur, J-C., Lepelletier, P., Bodin, J., Bennegadi-Laurent, N., Bessoule, J-J., Bispo, A.,Bodilis, J., Chaussod, R., Cheviron, N., Cortet, J., Criquet, S., Dantan, J Dequiedt, , A., Faure, O., Gangneux, C., Harris-Hellal J., Hedde, M., Hitmi, A., Le Guedard, M., Legras, M., Pérès, G., Repinçay, Rougé, L., C., Ruiz, N., Trinsoutrot-Gattin, I., Villenave, C., (2013). Démarche statistique pour la sélection des indicateurs par Random Forests pour la surveillance de la qualité des sols. *Etude et Gestion des Sols*, 20 (2), 127-136.