AGROSTAT 2016 CONGRESS, MARCH 21-24 2016
LAUSANNE SWITZERLAND

The advent of Big Data in Agriculture

Salima TAÏBI-HASSANI & Jérôme DANTAN

Department of Mathematics and Computer Sciences
Institut Polytechnique LaSalle Beauvais -Esitpa

Plan

1. Introduction
2. Context
3. Multidisciplinary research projects
4. Random forest and qualitative data
5. A statistical approach using Random Forest to assess biodiversity
6. Conclusion
7. Perspective

INTRODUCTION

Nowadays the number of data is huge !

Agriculture is also impacted and especially since the use of connected objects, drones, information systems, social networks,…

So how researchers in statistic could develop a methodology to :

- Manage data
- Organize data
- Analyze data
- Take into account heterogeneity, colinearity, missing values
- Deal with mixed data qualitative and quantitative data ?

Two Multidisciplinary Projects

- EMIRE : project *financed by GRR VASI (Normandy Great NetWork Research)*
- Bioindicators of Soil Quality BIO2 : project financed by ADEME

EMIRE : TEAM PROJECT

Coordinator: Esitpa

Academic Partners

- LMRS UMR 6085
- Rouen University
- UMR IDEES-Geosyscom
- UMR MA-Granem
- SGGW Poland

Professional Partners :

- AESN
- Comcom of Lillebonne

Multidisciplinary projects

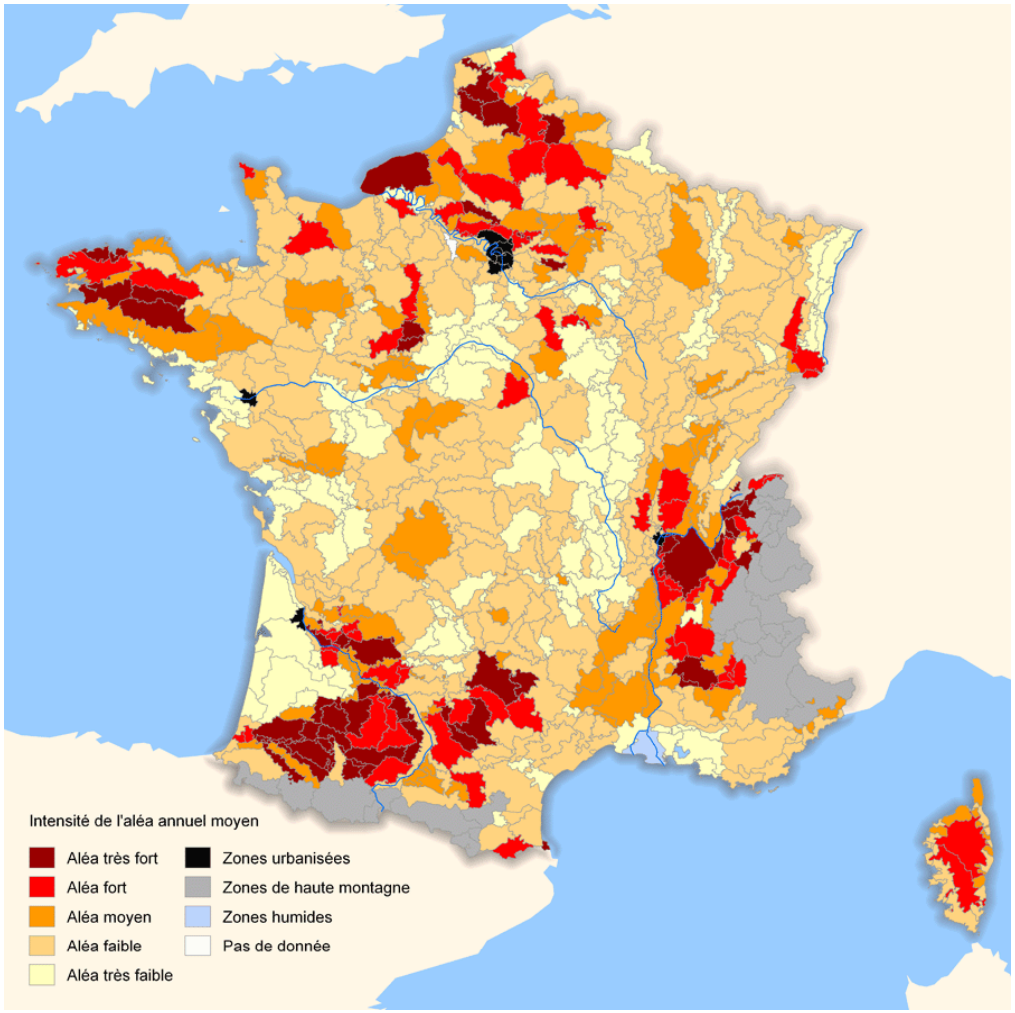


Fig. 1. Map of studied region

Source : Gis Sol – Inra – SOeS,.

In France 4 million hectares of 56 are affected by erosion

We estimate the willingness to pay (WTP) for protection from erosive runoffs among the people in area of Haute-Normandy region in France, which is presented on this map in figure 1

This area was chosen because it is highly impacted by erosive runoffs.

The survey took place in the Vallée du Commerce, a 48,000 hectares widewatershed located 40 kilometers away from Le Havre (Normandy).

This area is composed of 47 municipalities, for a total of 67,000 inhabitants.

The Project

- The purpose of this program 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 last over twenty years and involve an additional fee for the inhabitants of this valley.
- Inhabitants might benefit from this survey through such a program.

PROBLEMATIC

General framework of survey data analysis and we would like :

1. to predict individuals ”buy-in” to the Seine Estuary wetlands conservation program
2. 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 process can be extended to cases of multinomial qualitative variables.

RANDOM FOREST METHOD
A MACHINE LEARNING METHOD

A random forest is a set of m classification trees or regression trees constructed from the available data, together with *n 'bootstrap' samples*.

- *For each sample i,* we construct the *ith* tree by choosing the best partition from *k* variables chosen randomly from the entry variables (with replacement)
- *The* resulting entry vector *(Yi, Xi1 , Xi2 ,…., Xip)* is then the most popular class among the *m trees (when* classifying) or the mean obtained (regression).
- The result from each tree depends on the subset of predictors chosen independently (with replacement) and with the same distribution for all the trees in the forest.

RANDOM FOREST

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 observations (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 mawimum size

This process implies two parameters : K and the number of trees m.

The « hyperparameter » K can be chosen $K=\sqrt{p}$ ou $K = Pe(ln(p+1)$

RANDOM FOREST METHOD

The number of trees *m* must also be fixed, in general this number is between 100 and 500.

Breiman (2001) : when *m is large, we have no problems related to* overfitting large quantities of data.

Generalization error converges almost surely, it's estimated by using the *out-of-bag error (OOB),* which is calculated at each iteration of the algorithm.

The OOB error corresponds to the fact of predicting data outside the training sample which had been used to build the tree.

It is also useful when selecting importance variables and to understand the interactions between the observed data.

In fact, if two variables contain identical information, only one of them is useful and the second will have no influence in reducing the error.

RANDOM FOREST ON QUALITATIVE VARIABLES

- The predictors, (sex, geographical zone, family status, education, opinions of the program, etc.) are mostly qualitative variables (nominal or ordinal).
- We transformed them into quantitative variables, using Multiple Correspondence Analysis (MCA), in order to avoid the problem of multi co linearity while keeping the structure of the original data table.
- All factors resulting from the MCA were used, in order to preserve all the information from the initial data set. The quantification of variables *Xj is that which gives the largest Mahalanobis distance between the two* groups.
- The coordinates of each individual were transformed by weighting them so that the inertia of each factor (MCA) is conserved.

Random Forests on qualitative data
Vs
Discriminant Method

We compare RF and Discriminant Analysis on qualitative data (all observations).

Models	Kappa	Pcc
Random Forests	0,48	70%
Disqual Method	0,61	76%

Table 1: Performance of the models

Disqual Method : Discriminant Analysis on Qualitative data (Saporta, 1977)

Random Forests on Qualitative data

Context

- Big data
- Qualitative and quantitative data
- Correlation
- Missing values
- Total information must be preserved

Methodology

Multiple composant Analyzis

MCA on TCD

New coordinates

Results

- Automation process for qualitative variables
- Agregated variables
- Predictive models
- Development index
- Tools for decision making(

Bio 2 Team Project
Ademe

INRA

1. Champenoux
2. Dijon
3. Grignon
4. Gotheron
5. Versailles

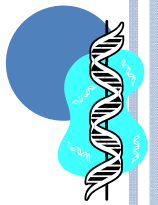
Research Inst./Universities

1. LIMOS
2. Rennes
3. Univ Besançon
4. Uni. Bordeaux
5. Univ. Marseille
6. Univ. Clermont
7. Univ. Rouen
8. Univ. Lille

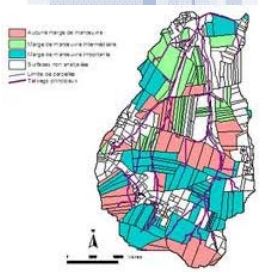
1. BRGM
2. Ecole Centrale
3. ENSAIA
4. ENM St Etienne
5. Esitpa
6. ISA
7. ISARA
8. IRD Bondy
9. IRD Montpellier





BIO II Project

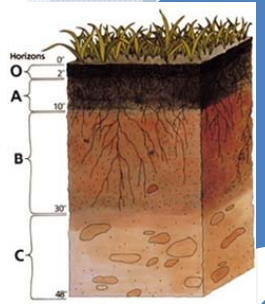


Landscape

Perspective of global analysis

Land Use

Biological activity of soil organisms



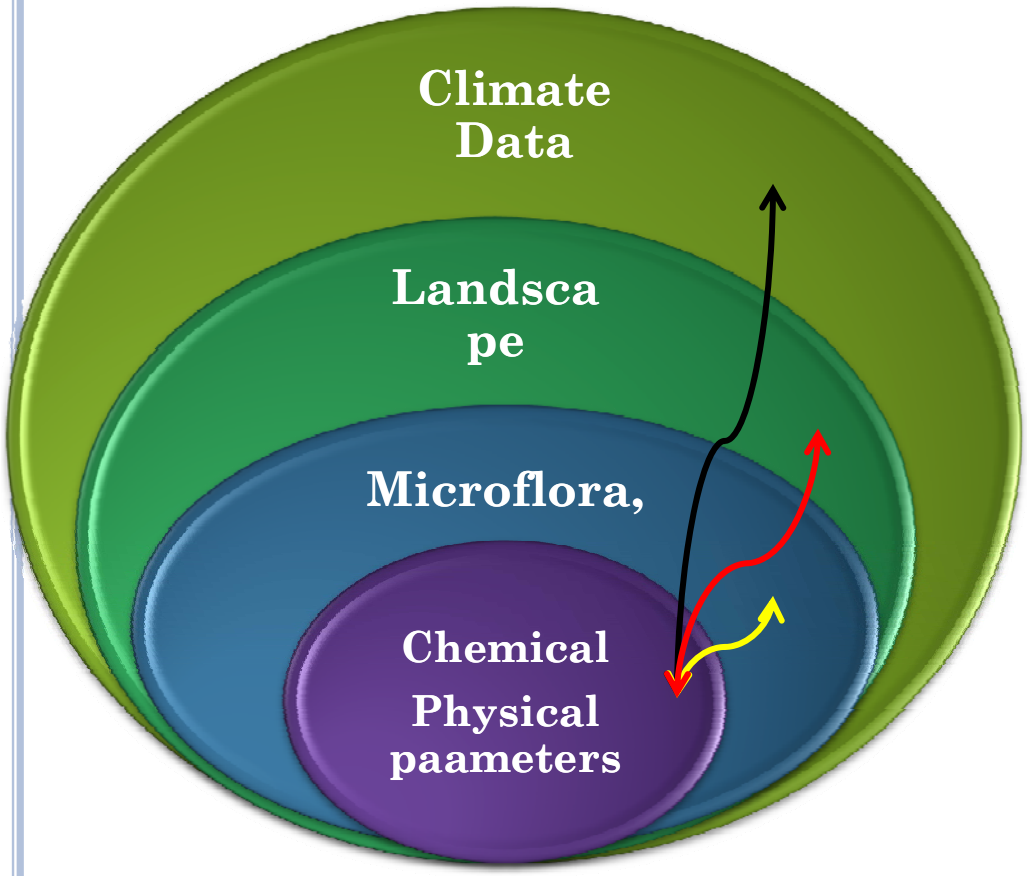
Pedology

Contamination

Environmental gradients

Stress

THE CONTEXT : MULTISCALE



47 plots characterized by a set of data
=> more than 200,000 observations.

How to manage them?
How to treat them ?
How to analyze relationships?
How to find the smallest set of explanatory indicators of soil ?

CONTEXT

Soil is a dynamic and complex system.

Information from "Microflora" "Flora" or "Fauna" are abundant and in interaction with the environment, climatic conditions ,…

Variables must be managed and analyzed together for a better understanding of soil system.

Given that the volume of data generated is more than 200,000 elements, it was necessary to design a database and some analytical process.

MULTIDISCIPLINARY PROJECTS "BIO-INDICATORS II"

BIOII : Random Forests are used to select variables

1) Develop a **roadmap**

2) A methodology to built a predictive model of soil quality given several pedo-climatic situations in France : "land use" , " organic or metallic contaminations (Taibi et al., 2011, 2012).

The group Biomath was created to manage and analyze data issued from "Bioindicators" project (Ademe, French Environment and Energy Management Agency).

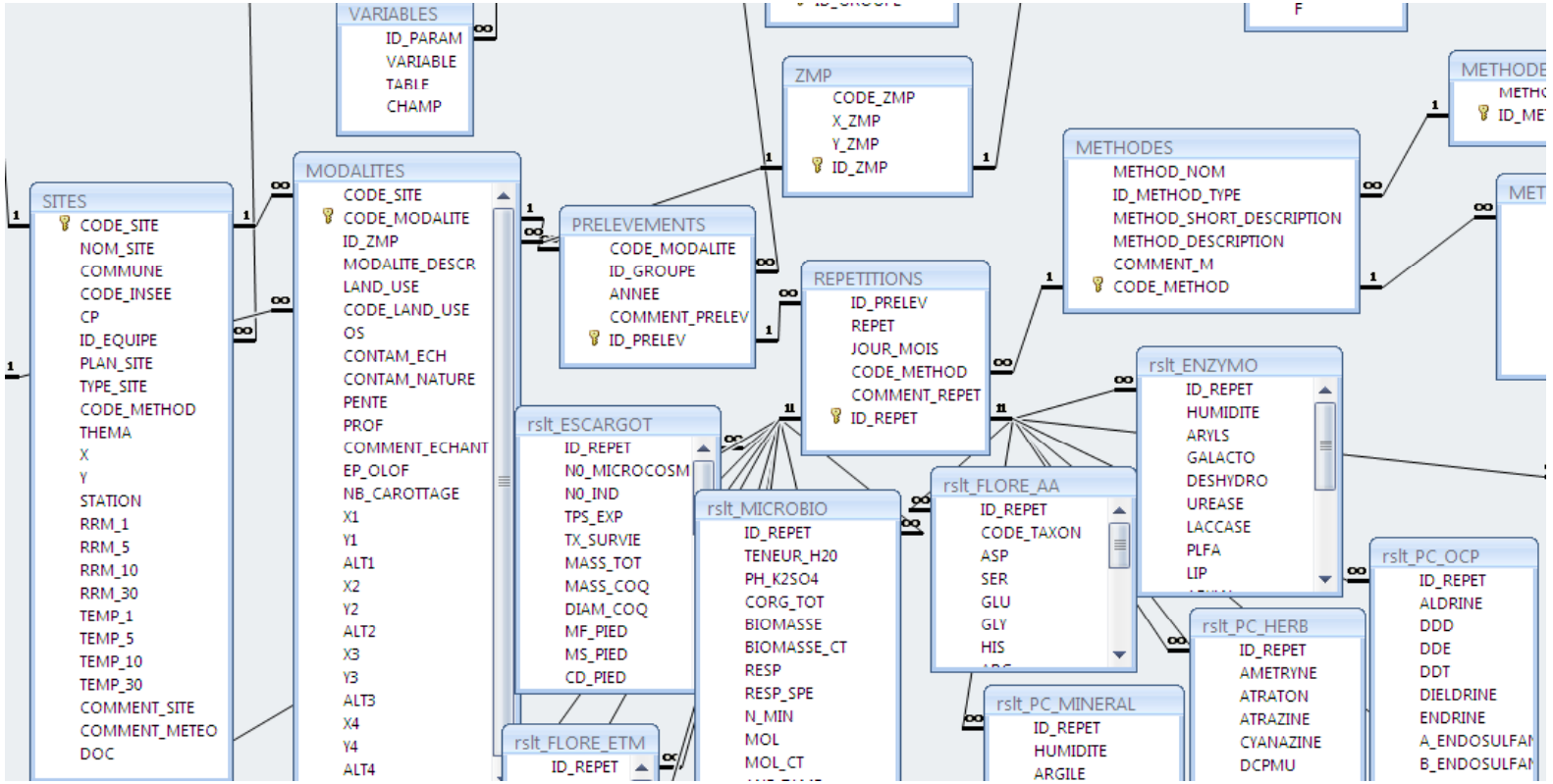
Biomath was composed of statisticians and computer scientists

Team project has worked together from **the start of the project**.

Methodology

- Database design
- Data quality
- Development of a user interface
- Software automatic data extraction
- Homogenization of data processing
- Definition of the baseline
- Outline of the sensitivity of the indicator
- Selection sets of indicators
- Indicator support tools for risk assessment
- Global analysis

DATA BASE…


VARIABLES
ID_PARAM
VARIABLE
TABLE
CHAMP
ZMP
CODE_ZMP
X_ZMP
Y_ZMP
ID_ZMP
METHODES
METHOD_NOM
ID_METHOD_TYPE
METHOD_SHORT_DESCRIPTION
METHOD_DESCRIPTION
COMMENT_M
CODE_METHOD
SITES
CODE_SITE
NOM_SITE
COMMUNE
CODE_INSEE
CP
ID_EQUIPE
PLAN_SITE
TYPE_SITE
CODE_METHOD
THEMA
X
Y
STATION
RRM_1
RRM_5
RRM_10
RRM_30
TEMP_1
TEMP_5
TEMP_10
TEMP_30
COMMENT_SITE
COMMENT_METEO
DOC
MODALITES
CODE_SITE
CODE_MODALITE
ID_ZMP
MODALITE_DESCR
LAND_USE
CODE_LAND_USE
OS
CONTAM_ECH
CONTAM_NATURE
PENTE
PROF
COMMENT_ECHANT
EP_OLOF
NB_CAROTTAGE
X1
Y1
ALT1
X2
Y2
ALT2
X3
Y3
ALT3
X4
Y4
ALT4
PRELEVEMENTS
CODE_MODALITE
ID_GROUPE
ANNEE
COMMENT_PRELEV
ID_PRELEV
REPETITIONS
ID_PRELEV
REPET
JOUR_MOIS
CODE_METHOD
COMMENT_REPET
ID_REPET
rslt_ESCARGOT
ID_REPET
N0_MICROCOSM
N0_IND
TPS_EXP
TX_SURVIE
MASS_TOT
MASS_COQ
DIAM_COQ
MF_PIED
MS_PIED
CD_PIED
rslt_MICROBIO
ID_REPET
TENEUR_H20
PH_K2SO4
CORG_TOT
BIOMASSE
BIOMASSE_CT
RESP
RESP_SPE
N_MIN
MOL
MOL_CT
rslt_FLORE_ETM
ID_REPET
rslt_FLORE_AA
ID_REPET
CODE_TAXON
ASP
SER
GLU
GLY
HIS
rslt_ENZYMO
ID_REPET
HUMIDITE
ARYLS
GALACTO
DESHYDRO
UREASE
LACCASE
PLFA
LIP
rslt_PC_MINERAL
ID_REPET
HUMIDITE
ARGILE
rslt_PC_HERB
ID_REPET
AMETRYNE
ATRATON
ATRAZINE
CYANAZINE
DCPMU
rslt_PC_OCP
ID_REPET
ALDRINE
DDD
DDE
DDT
DIELDRINE
ENDRINE
A_ENDOSULFAN
B_ENDOSULFAN

SELECTION OF A SET OF INDICATORS BY RANDOM FORESTS

To highlight the significant responses of more than 200 quantitative and qualitative variables to a disturbance related to organic or metallic contamination or related to land use, we use Random Forests.
Ranking factors: land use, heavy metal elements and organic pollutants.

1) The tests are performed on a **training set**,

2) **Validation set**,

3) **Test set** and we estimate the accuracy of this select approach.

Indeed, RF allow us to reduce the nb of variables to less than 30

Random Forets on qualitative data Vs Logistic Method

A statistical approach using Random Forest to assess biodiversity

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

Table 2: accuracy and performance of the models

EXEMPLE OF PREDICTIVE MODEL

-0,106635*AS_VISC_T28 - 0,0785029*Abond_Aneciques + 0,967628*Biom_Endoges + 2,04216*PHYTOPARA + 1,06283*Biom_Aneciques - 0,0669448*MASS_TOT_T28 - 0,340986*CD_VISC_T28 + 0,0258859*MICROARTHR_TOT + 0,058547*Abond_Endoges - 0,293435*CD_PIED_T28 - 0,641227*COLL_EQUI - 0,132356*COLL_DIV + 0,259975*AS_PIED_T28 - 0,0578636*EU_Meso + 1,22523*EI - 1,47886*PHYTO + 0,171285*TOT_ENTOM + 0,31032*PB_VISC_T28 - 0,13839*PPI - 1,3127*EI_DL

BIOINDICATORS PROJECT

Our approach allows us to :

1) Develop a roadmap and respond to various questions and queries.
2) Select discriminant and explicative indicators : fauna, flora, microbiological parameters, physical and chemical parameters
3) Built a predictive model of a given situation like "land use" , " organic or metallic contamination (Taibi et al., 2011, 2012, 2015).

Feasibility and operability indicators
Decision making Tools

The aim is to develop a general index of soil quality

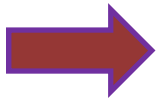
The previous results need to be technical feasible and public understanding

We decided to establish scores to assess the importance of these criteria in the choice of a decision tool.

Experts Survey

Responsible indicators

Questionnaire in three parts : scientific, technical and socio-economic criteria



Survey for Users,

Two groups of users: "Agronomists" and managers of "Polluted Sites and Soils"

The same questionnaire was administered to two groups, technical and socio-economic

$$\text{Score (final)} = S_{scien} \left(S_{tech}\ \text{final} + S_{eco\text{-}socio}\ \text{final} \right)$$

where S_{tech} et $S_{eco\text{-}socio}$ are determined by surveys data analysis(experts).

CONCLUSION

Data mining using RF allows us to take into account the problems occurs such as the heterogeneity, the "colinearity" and the presence of missing data.

In fact Random Forest is powerful in the case of Big Data.

Random Forests give benefits in socio economic, environmental, consumers studies, biological studies and in other contexts.

Random Forest on qualitative variables allows us :

1. To develop a methodology to select a set of discriminant variables.
2. To elaborate a predictive model mixing qualitative and quantitative.
3. To be able to built an approach to design a composite biological index of a soil (Taibi *et al.*, 2013 & 2015)

PERSPECTIVES

- Implementation of the algorithm of Random Forests on qualitative data in R software
- Conventional methods of Small Data are not always appropriate for "Big Data.
- Sampling , Inference Tests, confidence bands , … useless in the case of Big Data ?
- Most of problems : several variables and the nature of variables.
- Nonparametric methods (kernel method, K NN method, ...) are typically methods well appropriate in case of Big Data.
- Nonparametric methods in case of **measurable spaces** is a perspective for modeling big data (Taibi *and al.* 2015), we can used in many other domains such as food science, geophysics, bioclimatology, agroforestry…. using a certain metric and the smoothness parameter can be estimated by cross validation method.

REFERENCES

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

2. Crastes R., Beaumais O., Arkoun O., Laroutis D., Mahieu P.A, Rulleau B., **Hassani-Taïbi S**. Barbu V., Gaillard D., Erosive runoff events in the European Union: using discrete choice experiment to assess the benefits of integrated management policies when preferences are heterogeneous, *Ecological Economics*, **102**, 105-112, (2014.)

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. **Taïbi-Hassani, S**. et Laroutis, D., S.L Adigaw-E-Touck . 2015.Pointwise convergence of a nonparametric estimator of regression in a measurable space used in Contingent Valuation Method. Journal of Mathematics and System Science 5, 188-195.

THANKS FOR YOUR ATTENTION

Acknowledgement

The authors are grateful to Dimitri Laroutis, Jeanne-Chantal Dur (INRA), Laurence Rougé, Jeanne Bodin, Antonio Bispo, Cecile Grand , Guenola Perez and all participants of the research projects. and we wish to thank the Region Normandie and Ademe.