Collect and analysis of agro-biodiversity data in a participative context: A Business Intelligence Framework

Sandro Bimonte, University Clermont Auvergne, TSCF, INRAE
Sandro.bimonte@inrae.fr

Olivier Billaud
CESCO, Muséum National d'Histoire naturelle
olivier.billaud@edu.mnhn.fr

Benoît FONTAINE
CESCO, Muséum National d'Histoire naturelle
benoit.fontaine@mnhn.fr

Thomy Martin, Frédéric Flouvat
University New Caledonia
frederic.flouvat@univ-nc.nc

Ali Hassan
CESCO, Muséum National d'Histoire naturelle
ali.hassan@mnhn.fr

Nora Rouillier
CESCO, Muséum National d'Histoire naturelle
nora.rouillier@mnhn.fr

Lucile Sautot
AgroparisTech, UMR TETIS
lucile.sautot@agroparistech.fr

Collect and analysis of agro-biodiversity data in a participative context: A Business Intelligence Framework

November 10, 2020

Abstract

In France and Europe, farmland represents a large fraction of land cover. The study and assessment of biodiversity in farmland is therefore a major challenge. To monitor biodiversity across wide areas, citizen science programs have demonstrated their effectiveness and relevance. The involvement of citizens in data collection offers a great opportunity to deploy extensive networks for biodiversity monitoring. But citizen science programs come with two issues: large amounts of data to manage and large numbers of participants with heterogeneous skills, needs and expectations about these data. In this article, we offer a solution to these issues, concretized by an information system. The study is based on a real life citizen science program tailored for farmers. This information system provides data and tools at several levels of complexity, to fit the needs and the skills of several users, from citizens with basic IT knowledge to scientists with strong statistical background. The proposed system is designed as follows. First, a data warehouse stores the data collected by citizens. This data warehouse is modeled depending on future data analysis. Secondly, associated with the data warehouse, a standard OLAP tool enables citizens and scientists to explore data. To complete the OLAP tool, we implement and compare four feature selection methods, in order to rank explanatory factors according to their relevance. Finally, for users with extended statistical skills, we use Generalized Linear Mixed Models to explore the temporal dynamics of invertebrate diversity in farmland ecosystems. The proposed system, a combination of business intelligence tools, data mining methods and advanced statistics, offers an example of complete exploitation of data by several user profiles. The proposition is supported by a real life citizen science program, and can be used as a guideline to design information systems in the same field.

1 Introduction

The European Parliament highlights that the current loss of biodiversity has economic costs devastating for the society but that, so far, have not been sufficiently integrated into economic policies. Indeed, the economy and society are highly dependent on ecosystems and biodiversity. The monitoring and conservation of biodiversity in farmland currently represent major challenges [1].

Firstly, farmland is the dominant land-use in many regions of the world, such as in Europe. Secondly, biodiversity is rapidly being eroded by intensive agricultural practices. Finally, many promising alternatives to improve the sustainability of agriculture rely on the ecosystem services provided by biodiversity. However, financial and human resources may be limited to collect the data needed to measure impacts, assess effectiveness of conservation policies or changes in agricultural practices. Observation data at large spatial and temporal extents are needed to define biodiversity indices used in these assessments. These data are usually provided through standardized monitoring schemes, where a large number of observers must be mobilized, at a cost that would be prohibitive, unless when they are volunteers (such as in citizen science programs [2]). Due to the complexity of agro-biodiversity dynamics, a systemic approach must be deployed to highlight, discover and explain knowledge. Agro-ecosystems are complex systems whose dynamics are controlled by multiple elements, including biophysical variables and human activities. Therefore, understanding agro-biodiversity requires a systemic approach using large volumes of data, which describe agricultural practices over broad spatial and temporal extents (for example, collecting farms' data at national scale over several years). The deployment of such kinds of observatories implies huge and time-consuming collecting data activities. Moreover, the impact of agriculture and biodiversity over social and economic domains implies the involvement of various kinds of stakeholders, which can have different needs and skills in terms of analysis. For example, farmers are interested in simple indicators of biodiversity (for example, the average abundance of pests or beneficial organisms in their field), while ecologists need more detailed models and indices (for example, community specialization index in small agricultural regions).

To the best of our knowledge, no work provides a global framework to face these analysis challenges. Indeed, existing work exclusively focuses on a particular analysis goal and data collection, without providing a unified and integrated decision-support system for supporting multiple decision-makers to analyze several variables over a huge volume of data.

Therefore, in the context of the French ANR project VGI4Bio (vgi4bio.fr), we propose a complete and integrated Business Intelligence (BI) system to overcome the above described limitations. The BI system aims at analyzing agro-biodiversity data collected by a participatory observatory at the national scale over several years. In particular, we have studied the impact of the usage of chemical treatments on the abundance of solitary bees. BI systems allow for the exploration, visualization and analysis of huge volumes of data. They refer to different analysis tools, such as simple visualization/exploration tools (i.e. Data Warehouse (DW) and Online Analytical Processing (OLAP)), and advanced statistical and data mining methods [3]. Each BI tool is conceived for a particular kind of analysis and decision-makers.

To achieve the collection of data at large scale (i.e. France country scale), a citizen science program has been designed with and for the agricultural activity area: the Farmland Biodiversity Observatory (FBO). In this paper, we detail the collection data protocol of the FBO (Section 2), and the database allowing for the storage of FBO data. We also present the data quality cleaning methods used to handle the volunteer character of data collection of FBO (Section 5). A complete BI suite has been set up for the analysis of FBO data (Figure 1). DW and OLAP systems are used by non-skilled OLAP users (such as farmers, public

stakeholders) to analyze FBO data. They also allow data scientists to extract and visualize data needed for further analytical studies (Section 6). Faced with an important number of attributes, the visualization and knowledge extraction from OLAP systems can be difficult for non-skilled OLAP users. At the same time, many attributes can be correlated (or not) to the studied phenomenon. Therefore, we have used data science (machine learning and statistical methods) to highlight underlying interactions, and to model temporal trends with respect to biodiversity (section 7). More precisely, we have used feature selection and feature extraction methods to analyze the importance of the attributes with respect to biodiversity, and to analyze correlations among them. Based on extracted knowledge, a generalized linear mixed model has been developed to show the temporal trends in biodiversity, and to explain how it is related with the agricultural practices and the surrounding landscape.

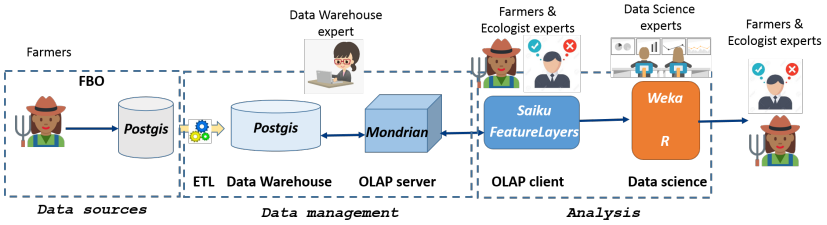

Figure 1: Our Business Intelligence system

2 Related works

In this section, we present related works about DW and Data Science for biodiversity. To the best of our knowledge, the usage of DW and OLAP in the context of biodiversity has been little investigated. [4] propose the joint use of unsupervised classification and OLAP to build information systems about bird biodiversity, with an emphasis on relationships between bird abundance and landscape structure. [5] provide a survey of existing work of DW applications in the agricultural and environmental contexts. The author highlights the usage of DW and OLAP for different kinds of applications ranging from water and air pollution to dairy farming and cotton production. However, no work studies the relationships among agricultural practices and biodiversity. Some recent works investigate the interest of Big Data and DW in the agricultural domain. [6] propose a big data warehouse architecture for the analysis of sales and stock of farms. [7] also present a DW for the management of stocks. The authors introduce some data mining-based algorithms to clean data sources. Finally, [8] and [9] propose the exploitation of semantic web technologies to build DW for dairy farming and drainage applications, respectively. It is also important to note that some works propose design approaches for participative BI (such as [10] and [11]). However, they provide solutions for collaborative design and analysis of warehoused data, but they do not integrate data science methods.

Data science has been also applied in the context of environmental monitoring [12]. Indeed, environmental data are complex (large, multidimensional, heterogeneous, etc.) and associated with complex phenomena. They require advanced data analysis methods. Several works have highlighted new environ-

mental knowledge, or proposed new monitoring tools, thanks to data science. [13] describe the potential of existing techniques and tools to analyze environmental data with several example applications. The authors explain that isolated techniques applied by data scientists are not sufficient to answer the complex questions raised by environmental phenomena, an interdisciplinary approach is necessary. They also present the interest of semantic networks to facilitate understanding of extracted knowledge. In [12], the authors focus on pre-processing techniques and present a review for non-experts of data science methods that can be used to deal with the issues of environmental data (noise, errors, redundancies or irrelevant). The same authors extend their analysis in [12] to the whole data science process, and provide references of applications for various types of environmental systems (water, air, land, forest vegetation, fauna/wildlife, fossil fuels/energy, climate and climate change). [14] present how ecology has joined a world of big data, emphasizing that variety is the main problem in ecoinformatics. They highlight how technical solutions must evolve to efficiently process such data. Apart from technical solutions, they also discuss the importance of integrating users, ecologists and data scientists together, with early training of non-experts and strengthened collaborations.

Among environmental topics, using data science to study biodiversity has also been a topic of interest for several years. For example, [15] introduce a data-intensive workflow for identifying factors influencing biodiversity (e.g. abundance of birds in North America). They propose a framework based on a data warehouse and exploratory analyses. The DW is used to structure and store collected data. Exploratory analysis aims at highlighting interaction patterns. For this, the authors propose to train a model (decision tree) and to analyse it using feature ranking and partial dependencies functions (a summarization method). In [7], the authors experiment several regression techniques to predict six biodiversity indices based on physical scans of a forest (acquired by a terrestrial laser). One objective is to study the impact of harvesting trees. [16] use text mining methods and ontology to extract a taxonomy and to identify species from scientific papers. They use semantic graphs to represent the concepts, relations and vocabulary in input texts. Then, they perform lexical, syntactic and semantic analysis of these semantic graphs.

To conclude, no work experiments an integrated framework using DW, OLAP and data science methods for the analysis of agro-biodiversity.

3 FBO Pollinators data collection protocol

Farmland Biodiversity Observatory (FBO) is a long-term and national observatory program launched in 2011 by the French ministry of agriculture. FBO has two main objectives: i) monitoring the biodiversity in farms, and ii) to sensitizing and empowering farmers through the direct observation of biodiversity on their farms.

FBO is a standardized monitoring protocol for biodiversity in farmland, in which farmers collect data. The database holds data collected over 2000 sites all over France. FBO stores more than one million observations about mollusks, ground beetles, butterflies, solitary bees and earthworms. This program is efficiently deployed at the national scale by means of local agricultural development organizations that recruit farmers, help monitoring and report feedback about

observations. The diversity of involved actors allows a wide scope of actions, and creates good conditions for dialog among them.

The FBO data set used in this project goes from 2011 to 2017. In this period, 1,216 farmers monitored biodiversity in 2,382 fields, covering the metropolitan France territory. Field crops (1,515 fields), meadows (705 fields), vineyards (538 fields) and orchards (240 fields) are monitored. Conventional and organic farming are both monitored. The data depends on the volunteer commitment over time, thereby the proportions of each crop's type and farming conduct are not representative of the French agriculture. Most observers are non-experts naturalists. They are volunteers who received a small training, from the FBO coordination team or from the local contacts. The protocols are simplified in order to be used by non-experts. The collected biodiversity data are also simplified (for example, species group instead of species) in order to be accessible for all participants, regardless of their existing knowledge. A notable effort has been made to standardize observations for each protocol through training, materials, and formats. In this paper, we focus on the pollinator trap nest protocol. Monitoring solitary bees is essential because of the major utility of pollinators for agriculture. It has been estimated that crops dependent on pollinators contribute to the volume of global food production by up to 35 percent. Solitary bee monitoring uses two trap nests, located 5m apart in the field edge, facing south at, 1-meter above ground. Each nest is composed of 32 paper tubes of 7mm diameter (Figure 2). 1,345 fields were monitored with the bee protocol by the FBO. Observers monthly count and identify sealed tubes. Seven types of sealing materials are listed: wax, dirt and mud, pieces of leaves, chewed leaves, grass, petals and cotton. The data are expressed in abundance and diversity of wild bees (i.e. the number of filled tubes). This protocol takes place from February (installation of the nests) to October of each year. Moreover, observers also provide data about the landscape in a radius of 200m around the plot: this includes the type of field edge, neighboring land use, and the presence of flowers in the crop (called Neighboring land use type). Agricultural practices are also described, such as tillage or use of synthetic inputs.

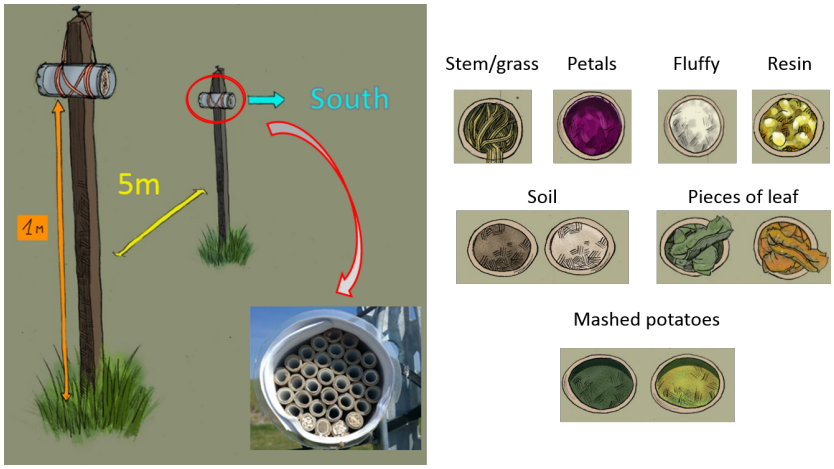

Figure 2: Schema of the pollinator trap nest protocol and the 7 type of sealing materials

4 Business Intelligence Framework

In this section, we present the main concepts and systems underlying our Business Intelligence system for the analysis of FBO data highlighting the different kinds of users involved. Our Business Intelligence system is depicted in Figure 1. Data sources are represented by the database containing the FBO data. These data are then moved in the Data Warehouse (DW) using an Extraction-Transformation-Loading tool (ETL). The ETL tool allows extracting data from the FBO database, transforming them using some cleaning operators, and loading them into the DW. Then, warehoused data are extracted from the DW and analyzed using Data Science methods. Different levels of analysis are considered to target the different types of users. Warehoused data is usually stored in classical relational Database Management Systems (DBMSs) such as Postgres, Oracle, etc. Warehoused data is modelled according to the multidimensional model that stores data according to dimensions and facts [3]. Dimensions represent analysis axes and they are organized in hierarchies representing different spatial, temporal and thematic granularities. Facts represent the analysis subjects and they are described by numerical attributes, called measures. Measures are then visualized according to levels of dimensions' hierarchies. Simple aggregation functions, implemented in the underlying DBMS, are used to aggregate measures values over hierarchies. This relational storage comes with an OLAP server that implements OLAP operators that can be triggered by means of OLAP clients. Common OLAP operators are: Roll-up and Drill-down, which allow to climb and go down into dimensions hierarchies aggregating and disaggregating data, respectively; Dice and Slice that permit to select a subset of the warehoused data. OLAP clients also provide an interactive visualization of OLAP queries results by means of user-friendly pivot tables, and graphical displays (such as bar charts, pie charts, etc.). OLAP systems permit end-users to explore huge data sets in a simple and intuitive way, and they provide some basic statistical analyses by means of SQL aggregation functions such as sum, minimum, average, etc. Usually, OLAP analysis represents a first step towards more complex ones to explain and understand phenomena. Indeed, they allow decision-makers to identify interesting data sets that "could" reveal interesting patterns and/or trends using data science methods. Therefore, in our BI framework, OLAP systems play two main roles: (i) allow decision-makers without data analysis skills to conduct basic agro-biodiversity analysis, and (ii) permit to simply explore warehoused data in order to identify (aggregated) data subsets that require further complex investigations using data science methods. Data science (DS) [17] is a domain regrouping all the data analysis methods. Historically, these methods came from different communities such as machine learning, database, statistics, data mining, or artificial intelligence. All these works have been regrouped under the appellation "data science" because problems and approaches are common. The main challenge of data science is to analyze, model and extract knowledge from more and more complex data (heterogeneous, voluminous, noisy, multidimensional, etc.). A wide variety of problems has been studied such as pattern extraction, supervised or unsupervised learning. For each issue, an important number of approaches has been proposed. Faced with complex data, several methods have generally to be combined to extract knowledge. Consequently, advanced skills are required to use, set up and combine these data science methods. Moreover, a blind analysis of data is rarely a suc-

cess. Data scientists need to strongly interact with domain experts (for example, ecologists in our case study) to really extract useful knowledge. At the end, data science results in a complex iterative and interactive process [18][19].

This BI framework is illustrated by a use case studying the impact of agricultural treatments on biodiversity (Figure 3). This analytical process is composed of the following steps:

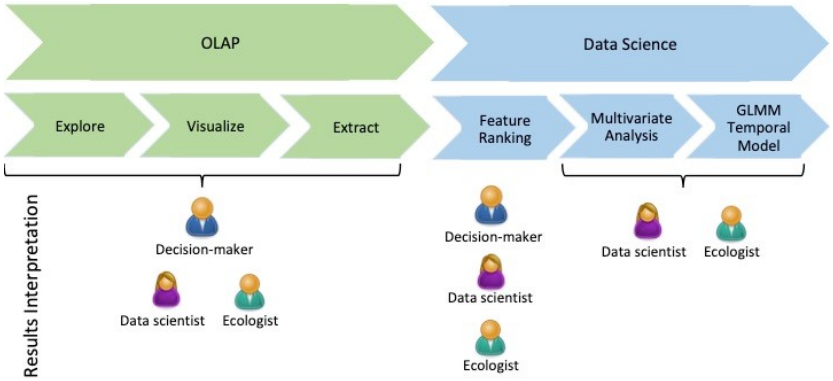

Figure 3: Steps of our analytical process

1. Firstly, the decision-maker explores the warehoused data visualizing the several dimensions of the DW. Very quickly, he/she finds out that crops under organic farming have higher abundances for some taxonomic groups. However, data can contain several biases and it is difficult to trigger a multi-attribute analysis in a visual way. Thus, some more advanced analysis must be done.

2. Therefore, data science methods are applied.

   (a) Based on the data extracted from the DW, a DS expert with feature selection methods is able to rank attributes according to their impact on biodiversity. As a result, treatments are confirmed as the factors having the most important influence on biodiversity. Feature selection methods can provide important information and they do not require any skills in ecology, but they only answer the question "which dimension influences biodiversity ?". They are not able to explain "how" these data characterize biodiversity.

   (b) Thus, a deeper analysis has to be done in interaction with ecologists. In our framework, feature extraction methods (principal component analysis and multiple correspondence analysis) are used to study more in detail the impact of treatments over biodiversity. Then, extracted knowledge is used to construct a temporal model of biodiversity dynamics over time by crop type.

All the steps of this scenario are detailed in the next sections.

5 FBO information system

In this section, we present the database we have developed to store the FBO data. Then, we discuss some rules that we have implemented to clean data, and in particular to handle quality problems associated with the volunteer nature of the data collecting protocols. We have used PostgreSQL as DBMS, since it natively supports spatial data, and it is open source and freeware. This database contains 30 tables. They can be grouped into five main sets:

1. *Users*: this group includes the tables that contain the information concerning the farmers, their networks, their types and their profiles.
2. *Plots*: this group contains the geographic representation of farms (municipalities, departments, and regions) and their fields (plots).
3. *Plot description*: this group represents the description of the different agricultural activities in the plots (cultivated areas, applied treatments, and type of crops).
4. *FBO site*: this group contains the descriptions specific to each biodiversity monitoring protocol (for example the height of the trap nest, etc.).
5. *Observations*: this group represents the weather conditions at the time of carrying out each protocol, and the abundance of different species observed within each protocol.

The quality rules that we have implemented can be classified in two main groups: (i) duplicated data, and (ii) missing data.

Duplicated data may come from several types of data, as for instance different registrations of a same user or a plot/farm created multiple times. We needed to add a new identifier to represent identical data. Since duplications were mostly due to human errors and do not follow a standard structure, their identification is difficult and the procedure could not be automated. This is why this procedure is carried out in several stages with manual validation. The first step is to extract data using ETL tools (such as Talend Open Studio for Data Quality). Then, the FBO data project manager manually provides a second identification step. The third step concerns integration (update) of these new identifiers in the database.

Missing data is due to the fact that when an observation is uploaded, a user can use an empty value, instead of zero value, for the abundance of a bee that has not been seen during the observation. To solve this problem, we automatically replace empty values with zero values. In addition, for agricultural practices where treatments and fertilizations are not used, the user may not add information concerning them. To facilitate data processing, we complete this data by adding tuples that clearly indicate that this information is not provided.

6 Data warehouse and OLAP

In this section, we present the DW implemented for the analysis of the FBO data (Section 6.1), its implementation (Section 6.2), and some examples of analysis (Section 6.3).

6.1 Multidimensional model

Using the FBO database, and applying the participative DW design methodology described in [11], we have defined the multidimensional model described in the next of the section. It presents the following dimensions (French terms are also presented since they are used in the real current OLAP implementation of the VGI4bio project):

1. Agricultural system (*Conduite*): organic, conventional or other kind of agricultural system.
2. Crop (*Culture*): type of crop cultivated (corn, barley, etc.).
3. Observation date (*Date pollinisateurs*): year of observation.
4. Fertilization (*Fertilisation*): input to amend the soil (organic fertilizer, nitrogen, etc.).
5. Cover crop (*Interculture*): crop cultivated among the major crop rotations (clover, alfalfa, etc.).
6. Interrow management (*Interrang*): management of the space between two rows in perennial crops (weeds, honey flowers, etc.).
7. Geographic location (*Localisation*): plot, farm, city, department and region.
8. Landscape type (*Paysage*): open field or mosaic of habitats.
9. Compliance with protocol (*Respect Frequence passage*): boolean value stating if temporal observation constraints were observed or not.
10. Tillage (*Travail sol*): usage and intensity of plowing.
11. Field edge type (*Type bordures*): presence and type of hedge, road, etc.
12. Neighboring landuse type (*Type milieu limitrophe*): wood, urban area, or other cultivated area.
13. Presence of landscape elements (*Type presences significatives*): significant landscape element relative to the studied taxa (honey flowers, hedge, wooded area, etc.).
14. Pesticide use (*Type traitement*): type of pesticide used (glyphosate, copper, etc.).
15. Meadow (*Usage*): type of meadow.

The multidimensional model is composed of three groups of measures that are aggregated using average, median, and first and last quartile:

1. comparing diversity and abundance at departmental, regional and national scale,
2. analyzing the trend of the abundance and the diversity over time and space, and
3. analyzing the observations behavior.

6.2 Implementation

To implement our OLAP system, we have used the OLAP Server Mondrian, which is an open source tool that is able to connect to most of the relational DBMSs via a JDBC connection, and the OLAP client Saiku.

6.2.1 Relational DBMS and OLAP server

The Data Warehouse is implemented in the Postgres DBMS. Its logical design is very complex since the multidimensional model presents complex hierarchies, facts, and aggregations [20].

Firstly, some dimensions, such as Pesticide use, Neighboring land use type, etc are non-strict hierarchies with two levels. Non-strict hierarchies are hierarchies where some levels members have more than one parent member. In order to avoid the "double counting" problem, we have modelled them using the snowflake design pattern, which uses one table for each level, and we have forced the OLAP server Mondrian to use fact table data when aggregate data at the "all" level of these hierarchies. The Crop dimension is composed by a non-onto hierarchy, which is a hierarchy where some levels members do not have children. To solve this issue, we have used some dummy members to fulfill the missing members. However, since we have noticed that dummy members led to confusion for decision-makers, we have created, in the OLAP server, on top of these dummy members, some calculated members that replace them by an advertising message in the OLAP client.

Aggregations used on the abundance and diversity measures are quite complex. Indeed, decision-makers need to aggregate these measures using different aggregation functions according to the different dimensions. For example, from the observation site level to the plot level, the sum is applied. Then, since these measures are semi-additive, the maximum is applied on the temporal dimension. Finally, the average, median, first and last quartile are used. Commonly, the aggregation of the measures is provided by the DBMS tier, which can apply only one aggregation function over all the dimensions, and the other aggregation functions are applied by the OLAP server, which computes them in the main memory. This approach does not work with our dataset, since the first aggregation is the sum and all the other aggregations are computed on too much factual data. To solve this issue, we have used a materialized view for time and location dimensions, with pre-calculated aggregations, and we have forced Mondrian to aggregate them in the DBMS tier along the other dimensions. It is also important to note that Mondrian does not support median, first quartile as SQL aggregation functions. Then, we have updated Mondrian code to deal with user-defined aggregation functions.

The Data Warehouse is loaded with data from the FBO database using the Talend ETL tool. The refresh of the data warehouse with new data is done at the end of each year. This yearly update is sufficient in our case study, since biodiversity evolution is a long term phenomenon. There is no need for real time or fast data loading and analysis.

To conclude, the Data Warehouse represents an important component of the BI framework, since it allows to hide the complexity of the warehoused data and to facilitate their aggregation for the data science experts. Indeed, without the OLAP client functionalities to explore and extract warehoused data, application

of advanced analysis methods by data scientists will be very difficult and time consuming.

6.2.2 OLAP client

OLAP systems allow the exploration of warehoused data with a simple click, and "drag & drop" actions, by any users [21]. In the following, we describe the main functionalities of the Saiku OLAP client that we have used in our implementation. Warehoused data is organized in a set of dimensions/levels and measures (the left panel of Figure 4). Navigation in dimensional data is simply provided by clicking over the particular dimension folder. After that, a hierarchical tree structure with data is shown. For example, Figure 4 shows the location dimension that is spanned. In order to trigger a query over the warehoused data, the decision-maker has simply to drag and drop dimensions levels and measures as columns and rows in the top panel of the OLAP client. An example is shown in Figure 4.

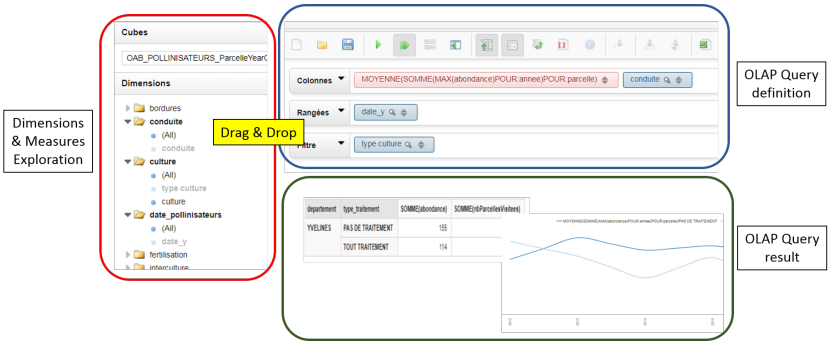

Figure 4: Saiku OLAP client user interface

The client automatically triggers the query and shows the results in the form of a pivot table (Figure 5). The pivot table is interactive. Indeed, by selecting a dimension value, it is possible to trigger a new OLAP query over the children of the selected member. Finally, according to the visual analytical paradigm [22], the OLAP client also proposes a set of visual displays such as bars charts, pie charts, etc. (Figure 5). To provide cartographic visualization of results of OLAP queries, we have extended Saiku with a web mapping tool, which is implemented using Feature Analyzer (Figure 5).

In our project, our tool has been used by five decision-makers (farmers and regional managers of agriculture organizations). These decision-makers had no skills in DW and OLAP, while they have advanced Excel proficiency. We have trained the decision-makers in this way:

1. Firstly, we have trained them for a few hours showing some representative OLAP queries using another case study concerning the birds biodiversity that we have prepared in advance with an ecological expert.

2. Secondly, we asked them to trigger a few OLAP queries and we helped them in this task.

3. Thirdly, we let the decision-makers provide their analysis, and we have accompanied them with on-demand meetings to solve some of their difficulties with the usage of the OLAP client.

After this training step, we have noticed that all the decision-makers have acquired a good proficiency in the OLAP system. In particular, the most recurrent difficulties for decision-makers do not come from the OLAP query composition via the drag and drop functionalities or the graphical displays configuration. This is because they master Excel pivot tables and graphical displays. Most of their difficulties rise from the user interface of Saiku, for example the "hidden empty cells" button that is checked by default, the two icons for graphical displays in and out of the pivot table that are quite similar, etc. All these difficulties have been easily solved during few meetings or email exchanges. Although, the decision-makers involved in this project are not representative of all kinds of possible users of our OLAP system, we can conclude that decision-makers with a similar profile (i.e. Excel proficient) could easily use our BI system. However, some limitations still unsolved. Indeed, the decision-makers have found out some limitations of the OLAP client and the cartographic visualization. In particular, for the OLAP client, they pointed out the lack of some particular graphical displays, the impossibility to modify their appearance, etc. For the cartographic client, they highlighted that the exclusive usage of bar charts is not enough for advanced geovisual analysis.

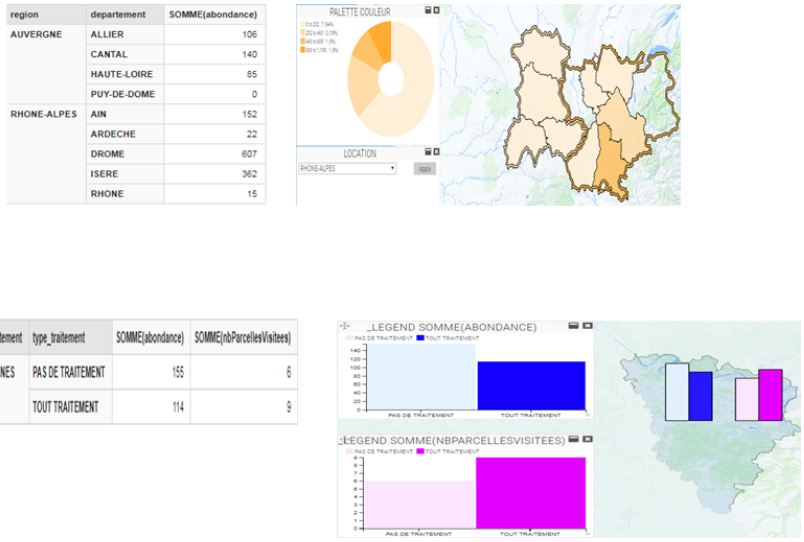

Figure 5: Graphical and cartographic displays of the OLAP client

6.3 Results

In this section, we report two of the most common analyses provided by decision-makers concerning the impact of treatments on biodiversity. Using the OLAP client, it is very simple to visualize the average of abundance at the national scale over several years. Decision-makers can simply compare this average value for biological and conventional wheat as shown in Figure 6.

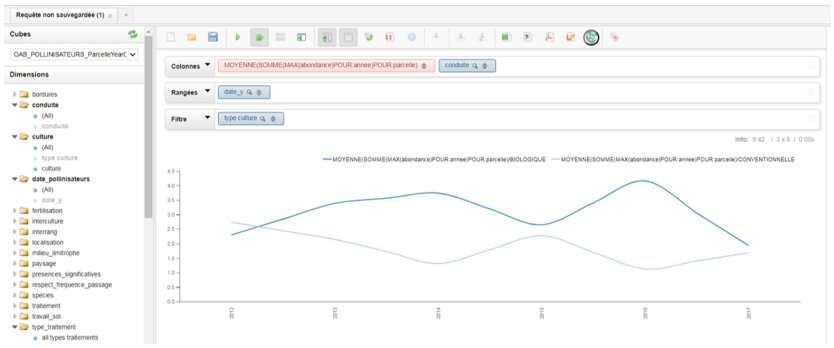

Figure 6: Visualization of the query "What is the trend in biodiversity over time for biological and conventional wheat?"

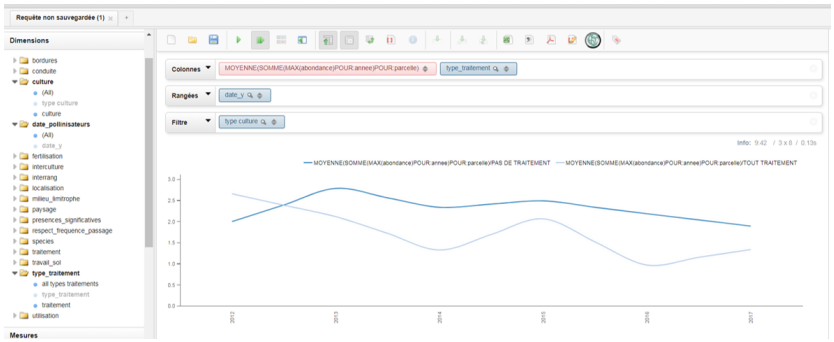

Figure 7: Visualization of the query "What is the trend in biodiversity over time for wheat with and without treatments?"

Figure 6 shows that biological wheat performs better than conventional wheat in terms of biodiversity abundance for each year. However, the gain of biological wheat is not constant each year. Based on this result, decision-makers analyze this data using some other criteria that directly involve treatments. Therefore, they replace the agricultural system dimension with the treatments one. They select the dimension member "*Pas de traitement*" (i.e. No treatments) and "*Tout traitement*" (i.e. All treatments). The result is shown in Figure 7. The lost of biodiversity is progressive over the years. In this way, a simple OLAP query enables to highlight the impact of treatments on biodiversity, and we can observe a long term decline of the biodiversity for wheat. However, Figure 6 and 7 highlight a strange trend in 2012, where this pattern is not respected. Therefore, the decision-makers visualize the number of collected data for 2012, and this year few data have been collected about wheat. Therefore, this year is not representative of biodiversity at the national scale.

Although these results are very interesting and could be achieved without efforts by decision makers, the influence on biodiversity of other parameters (i.e. dimensions) must be also studied. Therefore, the main question for decision-makers is how to proceed for this kind of OLAP analysis that potentially implies to visualize all dimensions at the same time reaching unreadable pivot tables and graphic displays. Therefore, some advanced analysis methods, as described

in the next section, are needed.

7 Data Analysis

In this section, we present how warehoused data are mined by data scientists and ecologists to highlight most important dimensions with respect to biodiversity, and construct a temporal model per taxonomic group.

The analytical process is composed of 3 main steps (Figure 3):

1. Firstly warehoused data are extracted using the OLAP client,
2. Then, data are mined using feature selection. These methods are used to quantify the importance of each dimension to "explain" biodiversity. At the end, non-skilled ecological users obtain a simple ranking of the dimensions.
3. Finally, we analyze more deeply interactions between dimensions (using a multivariate analysis) and construct a temporal model based on this detailed analysis (a Generalized Linear Mixed Model - GLMM). These results may be more difficult to analyze by decision-makers, but they give a much more detailed view of temporal interactions.

7.1 Feature ranking

In this section, we present how feature selection methods have been used to rank the most important attributes with respect to biodiversity, based on data provided by the OLAP system described in Section 6. In particular, we compare four classical feature selection methods, and study the impact of their results on six classification and regression algorithms using cross validation. Results demonstrate the interest of such an approach to hide some irrelevant information to users, and confirm the agricultural attributes affecting biodiversity in our data. Without any prior knowledge in ecology, these methods allow analyzing the complex systemic context of agro-biodiversity by correlating several parameters.

7.1.1 Feature selection methods

Feature selection is a pre-processing step of the knowledge discovery process. It is also referenced as dimensionality reduction. It is often used in conjunction with supervised learning (classification or regression) to remove less relevant dimensions, which reduces noise and improves classification performances [23] [24]. A basic feature selection approach is illustrated in Figure 8. In this example, the dependency of several dimensions against a studied measure is analyzed. The correlation of each dimension with the studied measure is processed, resulting in a dependency score that could be used to rank dimensions with respect to their ability to predict the measure.

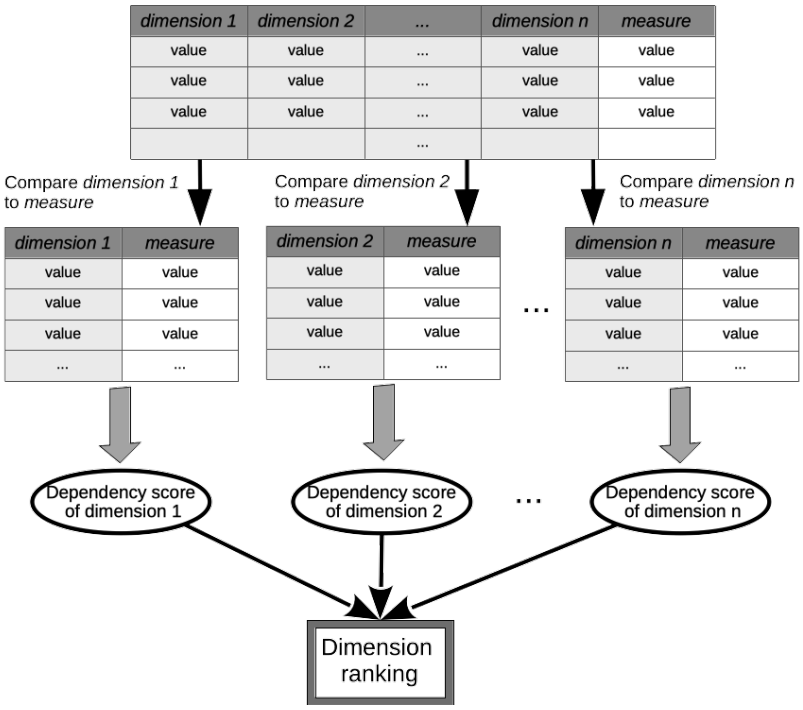

Figure 8: Example of feature selection approach

In our work, we study four feature selection methods: univariate linear regression test, C4.5 decision trees [25], extremely randomized tree regressor (also called extra-trees) [26], and backward recursive feature elimination (RFE) [27]. For this, we use Scikit Learn (functions f-regression, ExtraTreesRegressor and RFE) and Weka (J48 algorithm). Scikit Learn is a popular Python library dedicated to machine learning [28]. Weka is also a popular software, and Java library, including various machine learning and data mining methods [29].

Before discussing results, we briefly introduce the principle of each method used. The univariate linear regression test estimates the degree of linear dependency between a dimension and the studied measure. It computes their correlation (using Pearson correlation coefficient), and converts it to a F-score and a p-value. The F-score captures the accuracy of the supposed linear dependency (from 0 to 1), and the p-value estimates its statistical significance. For example, Figure 9 shows that the "dimension i" (y-axis of the left subfigure) is linearly correlated with the studied measure (x-axis of the left subfigure), while "dimension j" is not (right subfigure).

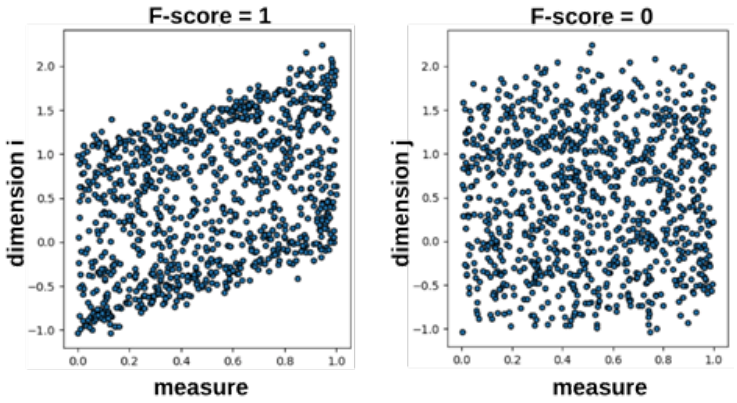

Figure 9: Example of linear regression test

C4.5 (and its Java implementation J48) [25] builds a decision tree (part of a decision tree related to agricultural practices and biodiversity is given in Figure 10). It is a tree-like structure used to visually represent decisions (e.g. pesticide used is herbicide and the agricultural system is organic) and their consequence on a studied attribute (e.g. abundance of biodiversity is between 2 and 3). To construct such a tree, C4.5 algorithm uses a recursive data partitioning schema based on the information gain measure. It does a successive selection of dimensions such that the resulting data partition is the most homogeneous with respect to the studied measure. It is dedicated to categorical data. Thus, numerical data have to be discretized before using this algorithm.

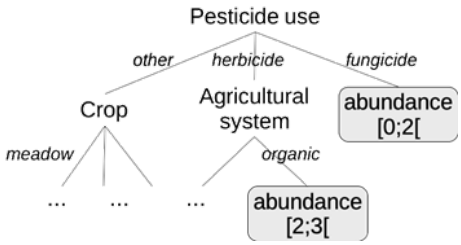

Figure 10: Example of decision tree

Decision trees have two limitations. First, generated models tightly fit training data (overfitting). Second, they are sensitive to dimension order (the optimal partition at each iteration is found based on heuristics). Extremely randomized trees (also called extra-trees) [26] have been proposed to deal with these limitations. This method trains several decision trees on random samples of the data and processes an "average" predictive model. For each tree, data partitioning is random. Feature selection is based on the average variance obtained for all generated trees. Figure 11 illustrates this approach.

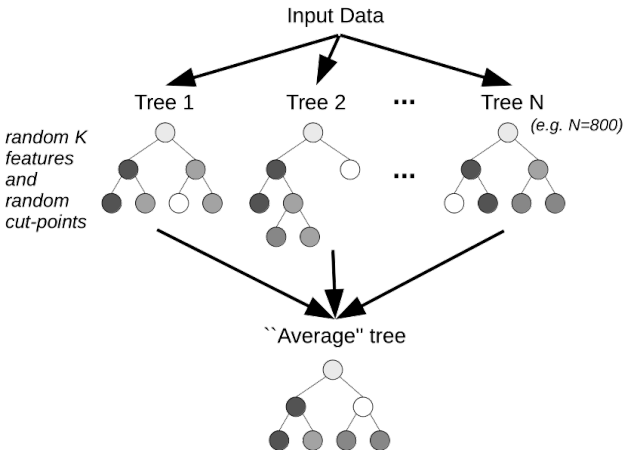

Figure 11: Example of extremely randomized tree generation

The backward recursive feature elimination (RFE) is a process (illustrated by the flowchart in Figure 12) that deletes dimensions recursively based on an external estimator. At each iteration, the estimator is used to evaluate the importance of each remaining dimension. Then, the least important dimension is removed, and the process is repeated. The estimator used in our work is the SVM algorithm (regression) with a linear kernel. Its accuracy is used to rank the dimensions.

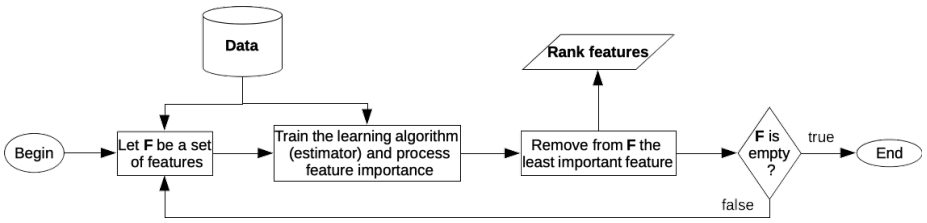

Figure 12: The backward recursive feature elimination (RFE) process

7.1.2 Results

Due to the important volume of data in the data warehouse, we focus our analysis on the eight dimensions that characterize agricultural practices: cover crop, fertilization, crop type, observation year, interrow management, tillage, pesticide use and agricultural system. For these dimensions, we consider the average of abundance as measure. The input consists in 2,509 rows. The results of the selected feature selection methods on this data are displayed in Table 1. Each dimension is ranked according to the results of the feature selection methods.

As shown by Table 1, results vary depending on the method used. For the univariate linear regression test, the top 3 features w.r.t. abundance are agricultural system, interrow management and pesticide use, while the ranking is different for the others. Several factors can explain this difference. Firstly, the linear model hypothesis is not necessarily good in our dataset. A nonlinear model (such as in the other methods) may be more adapted. Secondly, the

Feature rank	Univariate linear regression test	C4.5 decision tree	Extra-tree regressor	RFE with SVR
1	agricultural system	pesticide use	pesticide use	agricultural system
2	interrow management	crop	fertilization	fertilization
3	pesticide use	agricultural system	cover crop	cover crop
4	fertilization	fertilization	tillage	tillage
5	cover crop	interrow management	interrow management	interrow management
6	crop	tillage	crop	crop
7	tillage	cover crop	agricultural system	pesticide use

Table 1: Results of feature selection methods

Pearson correlation coefficient may not be the most appropriate measure. As discussed in [23], similarity based feature selection algorithms fail to tackle feature redundancy, i.e they may find highly correlated features. At the opposite, information theoretical measure (s.t. information gain) considers both "feature relevance" and "feature redundancy". C4.5 decision trees are nonlinear models. Their overfitting is not necessarily a problem in a feature selection process. The main problem is that we have to discretize the abundance measure before using such an approach. In our experiments, it was discretized into five intervals of equal size (i.e. equal intervals). Such empirical discretization may affect results. This could explain the differences with the results of the extra-tree regressor, while it is another decision tree approach. Indeed, crop and agricultural system dimensions are the latest dimensions for the extra-tree regressor (i.e. the less correlated to abundance), while they are in second and third position for C4.5. Interestingly, results of extra-tree regressor and recursive feature elimination (RFE) are quite similar. There is one main difference: pesticide use and agricultural system are ranked first and last by the extra-tree regressor, while it is the opposite for RFE. Pesticide use and agricultural system are correlated dimensions. Thus, once we have selected one of these dimensions, the other one is not necessary to explain abundance, which explains this difference. These feature rankings can be summarized in the following table. It simply sums the rank of each feature for each feature selection method used. As shown by Table 2, the top 3 dimensions with respect to abundance in this dataset are agricultural system, pesticide use and fertilization, which is highly consistent with the scientific literature on the effects of farming on wild bees.

Global ranking	Sum of ranks	Detailed ranks
agricultural system	12	1,3,7,1
pesticide use	12	3,1,1,7
fertilization	12	4,4,2,2
interrow management	17	2,5,5,5
cover crop	18	5,7,3,3
crop	20	6,2,6,6
tillage	21	7,6,4,4

Table 2: Global dimension ranking

To validate these results, we study the impact of these dimensions on performances of a set of supervised learning algorithms (provided by the Weka software). More precisely, we remove one dimension at a time and study performances of the selected algorithms without the removed dimension. These performances are compared with the performances of the algorithms with all the dimensions. For this, we use three regression algorithms: linear regression, Gaussian process and SVR with a polynomial kernel. We also use three classifications algorithms (with a discretization in 10 equal intervals): binary logistic regression, C4.5 decision trees and random forests [30]. To consolidate results, we perform a k-fold cross validation (with k=10) and process average performances. The following classical performance measures are studied: the percentage of instances correctly classified (classification only), the Matthews correlation coefficient (regression only), the relative absolute error and the root relative squared error. The results obtained are summarized in Figure 13 and Figure 14. The y-axis represents performance and the x-axis represents the removed dimension. Table 3 and Table 4 present more in detail the difference in performance obtained when removing one dimension.

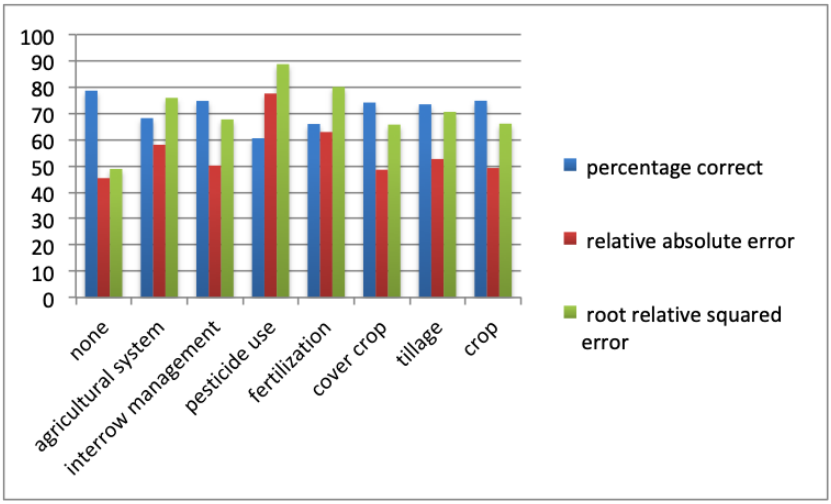

Figure 13: Results of classification algorithms when removing one dimension

removed feature	**percentage correct**	**relative absolute error**	**root relative squared error**
none	0	0	0
pesticide use	-18.10207337	32.18144093	39.84609666
fertilization	-12.67942584	17.53766142	31.18587678
agricultural system	-10.44657097	12.7001049	27.01861189
tillage	-5.19338118	7.282709638	21.68873862
cover crop	-4.515550239	3.187911028	16.84537941
interrow management	-3.897527911	4.772257027	18.82365821
crop	-3.827751196	3.852505026	17.20349211

Table 3: Difference in classification performance when removing one dimension

As shown by these figures and tables, the dimensions that have a stronger impact on the classification of abundance are pesticide use (-18 percent for

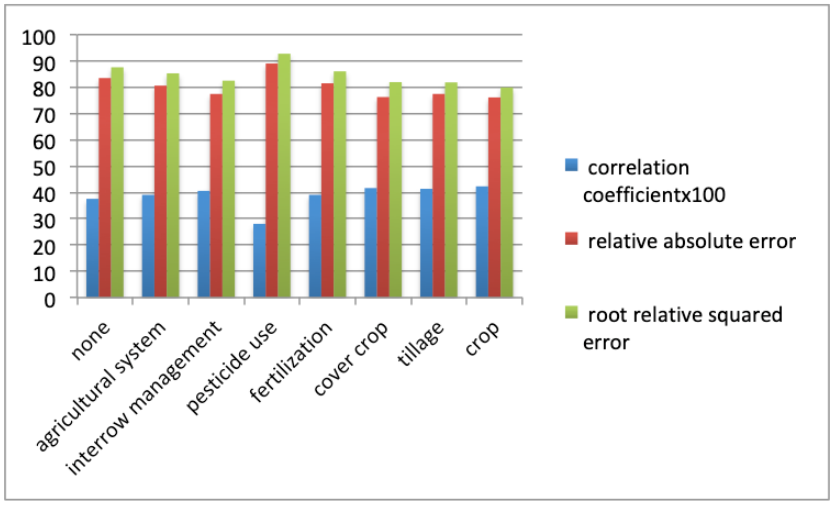

Figure 14: Results of regression algorithms when removing one dimension

removed feature	**correlation coefficientx100**	**relative absolute error**	**root relative squared error**
none	0	0	0
pesticide use	-9.515843223	5.51786819	5.15376433
agricultural system	1.502354085	-2.892439	-2.3040674
fertilization	1.515506421	-2.0095611	-1.5424833
interrow management	3.020521995	-6.0533608	-5.0845484
tillage	3.794020552	-6.0462508	-5.7437386
cover crop	4.100357787	-7.222808	-5.647832
crop	4.743917494	-7.3708131	-7.6727245

Table 4: Difference in regression performance when removing one dimension

classification and -10 percent for regression), fertilization (-13 per cent for classification and +2 per cent for regression) and agricultural system (-10 per cent for classification and +2 per cent for regression). These results confirm the results obtained by feature selection methods. Pesticide use, fertilization and agricultural system seem to be the most influencing dimensions with respect to abundance in our dataset. Surprisingly, removing a dimension often increases performances of regression methods (except for the pesticide use dimension). The crop dimension is the one that increases the most performances when removed from the dataset (+12.6 percent). The cover crop, tillage and interrow management dimensions have a similar impact (+10.9 per cent, +10 per cent and +8 percent). In other words, regression algorithms had better predict abundance when these dimensions are removed.

removed feature(s)	**percentage correct**	**relative absolute error**	**root relative squared error**
none	0	0	0
agricultural system & pesticide use & fertilization	-22.49800638	51.72005295	50.12183005
cover crop & tillage & crop	-14.81259968	24.64752623	35.33312445

Table 5: Difference in classification performance when removing three dimensions together

removed feature(s)	**correlation coefficientx100**	**relative absolute error**	**root relative squared error**
none	0	0	0
agricultural system & pesticide use & fertilization	-20.87080561	12.45668228	10.62297998
cover crop & tillage & crop	-2.155241372	0.89467572	1.572337969

Table 6: Difference in regression performance when removing three dimensions together

Table 5 and Table 6 show the impact on learning performances when removing the three most important features (agricultural system, pesticide use, and fertilization) and the three least important (cover crop, crop and tillage). As expected, performances decrease a lot when removing the three best features, while it decreases much less for the three other ones. This confirms the impact of the three best dimensions on abundance in biodiversity.

7.2 Feature extraction and construction of a temporal model

In this section, we present a statistical approach to investigate, not only how each variable explains the biodiversity dynamics, but also how they interact among them (the effect of one depending on the values of others). Therefore, we show the temporal trends in abundance of solitary bees, and how it is related

with agricultural practices and surrounding landscape. This method allows to study interaction effects between dimensions, but it requires more statistical and programming skills [31].

7.2.1 Methods

Firstly, we visualized data by means of histograms or 2-by-2 plots (a dimension against another one) aiming to find potential correlations between dimensions and visualizing data in the OLAP client. Due to the consistency of agronomic systems, we observed that agricultural practices, as well as landscape variables, were correlated with another one. To circumvent this problem, we applied feature extraction methods such as multivariate analysis to summarize practices and landscape variables. We used a principal component analysis (PCA) on quantitative data such as fertilization and pesticide use and multiple correspondence analysis (MCA) on binary landscape dimensions (presence / absence of elements). PCA is a well-known exploratory data method, dividing data in uncorrelated dimensions (call components) that are linear combinations of the original dimensions (reference). It allows us to reduce the number of dimensions and make the information less redundant. MCA uses a similar method than PCA but adapted for qualitative variables. Regardless of crop type, we observed the same general pattern in the outputs of the PCA, with the two main axes (components) easily understood as a "chemical treatment axis" (mostly pesticides and mineral fertilization variables) and an "organic fertilization" axis. As for MCA, one of the two first axes was understood as the proximity to woodland. The other axis included many different variables and was not easily interpretable.

Finally, to investigate the temporal trends in abundance per taxonomic group and their correlation with farming practices and landscape dimensions, we used generalized linear mixed models (GLMM) [32]. GLMM are regression methods allowing the analysis of non-normal data, such as count data (number of solitary bees) in our study, with a random effect. Random effect aims to quantify variation among units of the study that we cannot explain with our dimensions. In particular, we used a field-specific random effect (variation among the fields). We selected dimensions of the models starting from a complete model with year, practice and landscape dimensions as described by the first axes of the multivariate analyses, and their interactions, plus relevant additional covariates depending on the taxonomic group (like weather or degree-days) and random effects of the field. We selected dimensions thanks to a backward stepwise elimination, removing one by one the non-significant variables. The general structure of the model was the following figure:

$$\log(\mu_{AB}) = \beta_0 + \beta_1 Year + \beta_2 Axis1_{PCA} + \beta_3 Axis2_{PCA} + \beta_4 Axis1_{MCA} + \beta_5 Axis2_{MCA} + \beta_{6x} SpecificPractices + \beta_{7x} Covariates + \beta_8 Year{:}\, Axis1_{PCA} + \beta_9 Year{:}\, Axis2_{PCA} + \beta_{10} Year{:}\, Axis1_{MCA} + \beta_{11} Year{:}\, Axis2_{MCA} + \beta_{12x} Year{:}\, SpecificPractices + Field_i$$

Figure 15: General structure of the model

With β_j the regression coefficients, $Field_i$ the field-specific random effect and ":" the interactions between variables. $Axis_{PCA}$ and $Axis_{MCA}$ stands for positions on the multivariate axis. $SpecificPractices$ (tillage, inter-row...) and $Covariates$ (weather conditions, GPS coordinates...) varied depending on the

protocols and the type of crops. We checked that all the "control" covariates had an ecological consistent relationship with abundance, e.g. more abundant bees in the South. For its implementation, we used the R package Buildmer.

7.2.2 Results

We found significant temporal trends in the four crop types. The abundance of solitary bees appeared to be declining significantly in field crops, with declines stronger in fields with more pesticide use or more mineral fertilization (effects are difficult to separate) or less organic fertilization. Similarly, the decline in orchards was also stronger in fields with more pesticide/mineral fertilization use and in meadows declines were less steep with more organic fertilization. Finally, regarding landscape effect, on the one hand, bee decline was stronger in meadows closer to woodland, and on the other hand, increases were stronger in vineyard fields closer to woodland.

We illustrated these temporal interactions in Figure 16, showing marginal effects of practice/landscape variables with all others at their mean (quantitative variables) or at representative values (qualitative variables). Trends are shown for three contrasting values of the practice/landscape variable, from fields with a high level of use (darker line) through average (medium line) to low level (light line). Levels are given by: mean of the first axis (medium), and mean plus (darker) or minus one (light) standard deviation.

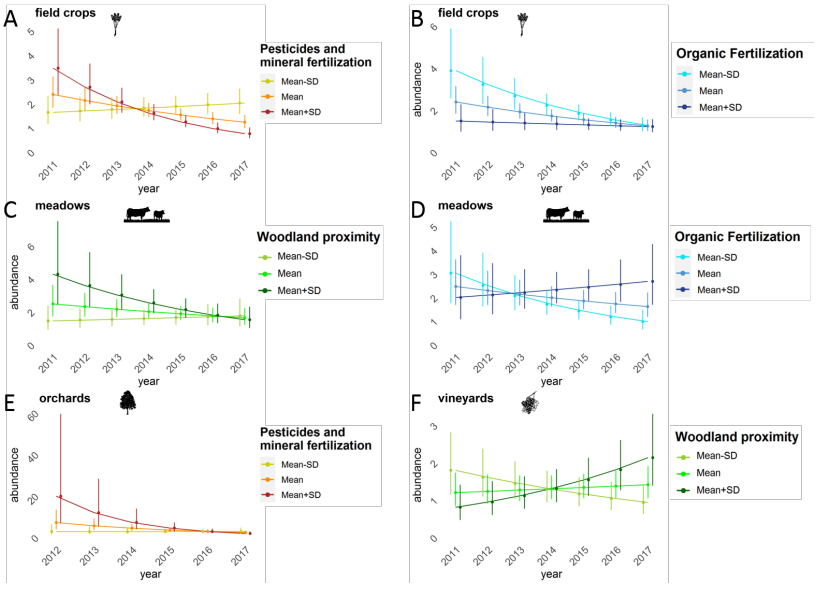

Figure 16: Relationship between temporal trends in bee abundance and agricultural landscape or practices in the different crop types (A-B field crops, C-D meadows, E orchards, F vineyards)

The declining patterns observed are in line with recent studies describing a decline of insects [33] [34] [35].

8 Conclusion and future work

In France and in Europe, farmland represents a large fraction of land use. The study and assessment of farmland biodiversity is therefore a major challenge. To monitor biodiversity across wide areas, citizen science programs have demonstrated their effectiveness and relevance. The involvement of citizens in data collection offers a great opportunity to deploy extensive networks for biodiversity monitoring. However, citizen science programs come with two issues: large amounts of data to manage and large numbers of participants with heterogeneous skills, needs and expectations about these data. In this work, we propose a solution to these issues, concretized by a Business Intelligence (BI) system. The study is based on a real life citizen science program called the Farmland Biodiversity Observatory (FBO). This BI system provides data and tools at several levels of complexity, to fit the needs and the skills of several users, from citizens with basic Information Technology knowledge to scientists with strong statistical background. The proposed system is designed as follows. First, a data warehouse stores the data collected by citizens. This data warehouse is designed according to the decision-makers analysis needs. Secondly, associated to the data warehouse, a standard OLAP tool enables citizens and scientists to explore data. To complete the OLAP tool, we implement and compare four feature selection methods, in order to rank explicative factors according to their relevance. Finally, for users with good statistical skills, we use Generalized Linear Mixed Models to explore the temporal dynamics of invertebrate diversity in farmland ecosystems. The proposed system, a combination of business intelligence tools, data mining methods and advanced statistics, offers an example of complete exploitation of data by several user profiles. We illustrate the advanced complex analysis possibilities offered by our BI framework by demonstrating the impacts of agricultural treatments on biodiversity. In particular, we show how from the simple OLAP analysis it is possible to define a complex temporal evolution model to characterize agro-biodiversity.

Our future work concerns the usability of the OLAP client. Indeed, a complete and formal usability study must be conducted using different kinds of decision-makers having different profiles in order to evaluate whatever the Saiku OLAP client could be an effective solution for a large participative usage and analysis of the FBO data.

Moreover, we also plan to investigate the usage of clustering approaches to identify farmers sharing similar practices. Temporal clustering enables to group farmers having relatively similar data. However, it does not give a description of each group. Moreover, clustering does not target a specific biodiversity outcome such as practices leading to a high biodiversity. For that, subgroup discovery approaches should be used. Their aim is to identify and describe subgroups of the data given a property of interest. However, few works have studied spatio-temporal subgroup discovery. One problem is to efficiently integrate the specificity of the spatial and temporal dimensions such as the existence of hierarchies (e.g. fields can be aggregated in farms, which can be aggregated in regions, etc.).

9 ACKNOWLEDGEMENT

This work has been supported by the French ANR project ANR-17-CE04-0012 VGI4bio, and the ISITE CAP2025 HubInnovergne project.

References

[1] R. Bommarco, D. Kleijn, and S. G. Potts, "Ecological intensification: harnessing ecosystem services for food security," *Trends in Ecology & Evolution*, vol. 28, no. 4, pp. 230 – 238, 2013.

[2] C. Régnier, G. Achaz, A. Lambert, R. H. Cowie, P. Bouchet, and B. Fontaine, "Mass extinction in poorly known taxa," *Proceedings of the National Academy of Sciences*, vol. 112, no. 25, pp. 7761–7766, 2015.

[3] R. Kimball, *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*. USA: John Wiley & Sons, Inc., 1996.

[4] L. Sautot, B. Faivre, L. Journaux, and P. Molin, "The hierarchical agglomerative clustering with gower index: A methodology for automatic design of olap cube in ecological data processing context," *Ecological Informatics*, vol. 26, pp. 217 – 230, 2015. Information and Decision Support Systems for Agriculture and Environment.

[5] S. Bimonte, "Current approaches, challenges, and perspectives on spatial OLAP for agri-environmental analysis," *Int. J. Agric. Environ. Inf. Syst.*, vol. 7, no. 4, pp. 32–49, 2016.

[6] V. M. Ngo and M.-T. Kechadi, "Crop knowledge discovery based on agricultural big data integration," 2020.

[7] A. McCarren, S. McCarthy, C. O. Sullivan, and M. Roantree, "Anomaly detection in agri warehouse construction," in *Proceedings of the Australasian Computer Science Week Multiconference*, ACSW '17, (New York, NY, USA), Association for Computing Machinery, 2017.

[8] C. G. Schuetz, S. Schausberger, and M. Schrefl, "Building an active semantic data warehouse for precision dairy farming," *Journal of Organizational Computing and Electronic Commerce*, vol. 28, no. 2, pp. 122–141, 2018.

[9] N. Gür, K. Hose, T. B. Pedersen, and E. Zimányi, "Enabling spatial olap over environmental and farming data with qb4solap," in *Semantic Technology* (Y.-F. Li, W. Hu, J. S. Dong, G. Antoniou, Z. Wang, J. Sun, and Y. Liu, eds.), (Cham), pp. 287–304, Springer International Publishing, 2016.

[10] M. A. Teruel, A. Maté, E. Navarro, P. González, and J. C. T. Mondéjar, "The new era of business intelligence applications: Building from a collaborative point of view," *Bus. Inf. Syst. Eng.*, vol. 61, no. 5, pp. 615–634, 2019.

[11] A. Sakka, S. Bimonte, L. Sautot, G. Camilleri, P. Zaraté, and A. Besnard, "A volunteer design methodology of data warehouses," in *Conceptual Modeling - 37th International Conference, ER 2018, Xi'an, China, October 22-25, 2018, Proceedings* (J. Trujillo, K. C. Davis, X. Du, Z. Li, T. W. Ling, G. Li, and M. Lee, eds.), vol. 11157 of *Lecture Notes in Computer Science*, pp. 286–300, Springer, 2018.

[12] K. Gibert, J. Izquierdo, M. Sànchez-Marrè, S. H. Hamilton, I. Rodríguez-Roda, and G. Holmes, "Which method to use? an assessment of data mining methods in environmental data science," *Environmental Modelling & Software*, vol. 110, pp. 3 – 27, 2018. Special Issue on Environmental Data Science and Decision Support: Applications in Climate Change and the Ecological Footprint.

[13] A. Lausch, A. Schmidt, and L. Tischendorf, "Data mining and linked open data – new perspectives for data analysis in environmental research," *Ecological Modelling*, vol. 295, pp. 5 – 17, 2015. Use of ecological indicators in models.

[14] S. S. Farley, A. Dawson, S. J. Goring, and J. W. Williams, "Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions," *BioScience*, vol. 68, pp. 563–576, 07 2018.

[15] S. Kelling, W. M. Hochachka, D. Fink, M. Riedewald, R. Caruana, G. Ballard, and G. Hooker, "Data-intensive science: a new paradigm for biodiversity studies," *BioScience*, vol. 59, no. 7, pp. 613–620, 2009.

[16] M. Vilares, M. Fernández, and A. Blanco, "Supporting knowledge discovery for biodiversity," *Data & Knowledge Engineering*, vol. 100, pp. 34 – 53, 2015.

[17] L. Cao, "Data science: A comprehensive overview," *ACM Comput. Surv.*, vol. 50, June 2017.

[18] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The kdd process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, p. 27–34, Nov. 1996.

[19] L. Cao, "Domain-driven data mining: Challenges and prospects," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 6, pp. 755–769, 2010.

[20] T. B. Pedersen, C. S. Jensen, and C. E. Dyreson, "A foundation for capturing and querying complex multidimensional data," *Inf. Syst.*, vol. 26, no. 5, pp. 383–423, 2001.

[21] M. Scotch and B. Parmanto, "Sovat: Spatial olap visualization and analysis tool," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pp. 142b–142b, 2005.

[22] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler, "Challenges in visual data analysis," in *10th International Conference on Information Visualisation, IV 2006, 5-7 July 2006, London, UK*, pp. 9–16, IEEE Computer Society, 2006.

[23] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, Dec. 2017.

[24] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70 – 79, 2018.

[25] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[26] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, p. 3–42, Apr. 2006.

[27] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, p. 389–422, Mar. 2002.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[29] M. A. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[30] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[31] O. Billaud, R.-L. Vermeersch, and E. Porcher, "Citizen science involving farmers as a means to document temporal trends in farmland biodiversity and relate them to agricultural practices," *Journal of Applied Ecology*, vol. n/a, no. n/a.

[32] B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White, "Generalized linear mixed models: a practical guide for ecology and evolution," *Trends in Ecology & Evolution*, vol. 24, no. 3, pp. 127 – 135, 2009.

[33] C. A. Hallmann, T. Zeegers, R. van Klink, R. Vermeulen, P. van Wielink, H. Spijkers, J. van Deijk, W. van Steenis, and E. Jongejans, "Declining abundance of beetles, moths and caddisflies in the netherlands," *Insect Conservation and Diversity*, vol. 13, no. 2, pp. 127–139, 2020.

[34] T. J. M. Van Dooren, "Assessing species richness trends: Declines of bees and bumblebees in the netherlands since 1945," *Ecology and Evolution*, vol. 9, no. 23, pp. 13056–13068, 2019.

[35] R. van Klink, D. E. Bowler, K. B. Gongalsky, A. B. Swengel, A. Gentile, and J. M. Chase, "Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances," *Science*, vol. 368, no. 6489, pp. 417–420, 2020.