A data-driven method for detecting and diagnosing causes of water quality contamination in a dataset with a high rate of missing values

Raymond Houé Ngouna, Romy Ratolojanahary, Kamal Medjaher, Fabien Dauriac, Mathieu Sebilo, Jean Junca-Bourié

A data-driven method for detecting and diagnosing causes of water quality contamination in a dataset with a high rate of missing values

Raymond Houé Ngouna[a,*], Romy Ratolojanahary[a], Kamal Medjaher[a], Fabien Dauriac[b], Mathieu Sebilo[c,d], Jean Junca-Bourié[e]

[a] *Laboratoire Génie de Production, École Nationale d'Ingénieurs de Tarbes, BP1629, 47 avenue d'Azereix 65016 Tarbes Cedex 16, France*
[b] *Chambre d'Agriculture des Hautes-Pyrénées, 20 Place du Foirail, 65000 Tarbes, France*
[c] *Sorbonne Université, CNRS, INRAE, IRD, UPD, UPEC, Institute of Ecology and Environmental Sciences Paris, iEES, 75005 Paris, France*
[d] *Université de Pau et des Pays de l'Adour, E2S UPPA, IPREM (Institut des Sciences Analytiques et de Physico-Chimie pour l'Environnement et les Matériaux), Pau, France*
[e] *Agence de l'eau Adour-Garonne, Tarbes, 7 Passage de l'Europe, 64000 Pau, France*

Abstract

Democratization of sensing devices in industrial systems has made it possible to collect a large amount of data of different types, which has led to the necessity of handling complex analyses for knowledge extraction. The field of water resources is of those areas which has drawn the attention of decision-makers seeking to preserve human health and safety. Recent advances in Artificial Intelligence, particularly in the domain of Machine Learning, have opened the potential to leverage massive data to better address the issue related to the relationship between water quality and human activities. However, high rate of missing data and heterogeneity of the measurements are scientific issues that cannot be solved by standard methods, especially when no prior knowledge on the label of each observation is provided. In this article, Prognostics and Health Management was implemented to detect and diagnose anomalies in water quality

*Corresponding author
Email addresses: raymond.houe-ngouna@enit.fr (Raymond Houé Ngouna), romy-alinoro.ratolojanahary@enit.fr (Romy Ratolojanahary), kamal.medjaher@enit.fr (Kamal Medjaher), fdauriac@hautes-pyrenees.chambagri.fr (Fabien Dauriac), mathieu.sebilo@sorbonne-universite.fr (Mathieu Sebilo), jean.junca-bourie@eau-adour-garonne.fr (Jean Junca-Bourié)

datasets, taking into account the uncertainties induced by the above-mentioned issues. Fuzzy c-means was used to identify the different water quality classes, while Random Forest was applied to determine the most influencing parameters, with respect to potential contamination of water resources in the southwest of France. The results suggest that multiple imputation methods can handle the missingness issue, while the use of decision rules based on well-known water quality standards can solve the problem regarding the lack of labelled observations. In addition, two potential sources of contamination (atrazine and nitrate) were identified and then validated by hydrogeology experts, prior to further online deployment of the proposed model.

Keywords: intelligent fault detection; diagnostics; water quality; uncertainty; unsupervised learning; Fuzzy c-means; Random Forest.

---

1. Introduction

Before technological breakthrough in the field of data acquisition, collecting measurements for systems monitoring was not straightforward (Sen and Bricka, 2009). Nowadays, this is quite the opposite, i.e. a large amount of data is available at a low cost. As a result, decision-makers are facing new issues, which are: how to make sense of the available data, and how to extract knowledge and information in order to effectively monitor a system and suggest appropriate actions depending on its health state. Advances in Artificial Intelligence (AI) have shown their ability to solve such issues, with less expert intervention and less prior knowledge needed on the studied system. AI methods and techniques are indeed appropriate to handle massive data with complex relationships between a high number of variables (Frank and Kppen-Seliger, 1997). However, studies undertaken in the literature are usually conducted with complete, regular, and friendly datasets. But, those conditions are rarely met in real-world contexts, as it is the case in the present research work. Indeed, irregularities, such as outliers and missing values that make data non reliable and therefore unexploitable, can occur; which may affect the performance of the learning algorithms. A general

approach, allowing to take into consideration the whole lifecycle of the data to handle those issues is then required. Prognostics and Health Management (PHM) is such is one of the candidate approaches that has been successfully applied in various industrial areas to monitor complex systems such as bearings, motors or trains (Atamuradov et al., 2017; Gouriveau et al., 2016; Benkedjouh et al., 2013). In those industrial domains, it has allowed effective maintenance policies, since the maintenance is performed only when needed, based on analyses of the health states of the system, resulting in less undesirable interventions to solve failures and reduced costs. The general goal of this research work is then to transpose this approach into the field of water quality monitoring, focusing on detection and diagnostics, and taking into consideration the specificities of the water resource system.

Indeed, it is well known that water is the most important natural resource without which no life can exist. In today's globalized world, where geostrategic issues in relation with the quest for natural resources are gaining in importance, any contamination, whether unintentional (e.g. due to the use of pesticides in intensive farming) or intentional (e.g. in case of a terrorist act), can result in dramatic consequences for humans, fauna and flora. Monitoring its quality has thus became a major concern that has been the subject of global, national or local action plans, aiming at preserving human health and biodiversity. Due to its chemical and physical properties, water is capable of dissolving many substances, which makes the chemistry of natural water very complex. In addition, a simple measurement of the total solids content of a sample is not enough to determine its character (Tebbutt, 1997). So, to fully understand the nature of a particular sample, especially in the context of pollution (which is the main scope of this study), it is generally necessary to measure several different properties by performing analyses under the broad frame of (1) physical characteristics (e.g. turbidity, electrical conductivity), (2) organic contaminants (e.g. pesticides, volatile organic compound), (3) inorganic contaminants (e.g. aluminium, nitrate) and (4) microbiologic contaminants (e.g. coliform bacteria). Within this frame, mutual influences of parameters can occur, while the system itself

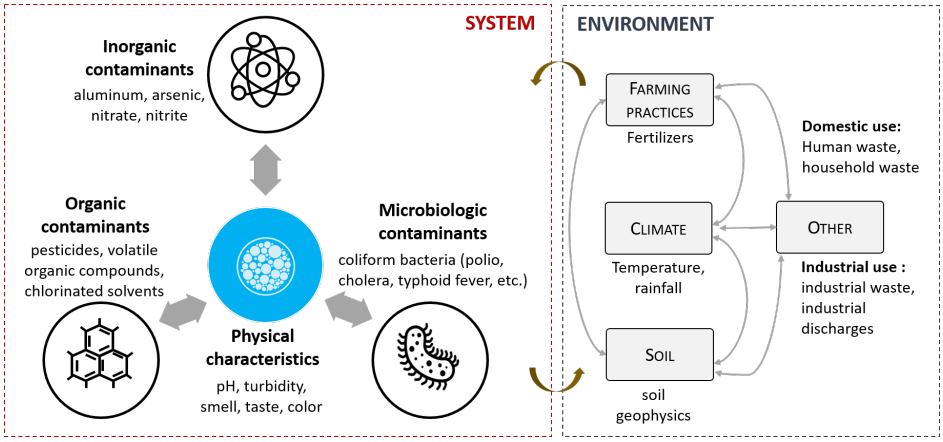

Figure 1: The system "water and its environment".

can interact with its environment, composed of human and industrial activities, climate and soil evolutions (as depicted in Fig. 1). Although these interactions may have positive effect on a contaminated water to recover a healthy state, in the current context of climate change, in conjunction with perpetual conflicts, scarcity of natural resource has now received much public attention, while the growing use of pesticides in intensive farming is putting pressure on politicians and decision-makers in cities. Hence, this situation requires suitable decision tool allowing to early detect risks to humans, fauna and flora and to initiate appropriate remediation actions.

In the specific context of this study, the number of observed parameters is large (around 400), the missingness rate is high (around 84%) and no prior knowledge on the water state (good or bad) for each observation is available. This makes the analysis difficult for providing insightful and reliable knowledge to decision-makers. Therefore, the two first issues require to pre-process the raw data to make them exploitable. Regarding the third issue, unlike the industrial context, there is no universal indicator that can allow to determine the health state of the system, i.e. the water resource. Very often, Water Quality Indices

(WQI) are used to identify different levels of quality, but they are not universal and may not be relevant in some context (Tyagi and Singh, 2013).

Thus, the specific goal of this study is to develop a general and effective model that can make it possible to detect contamination of water resource and then diagnose the associated causes, taking into consideration the complexity and uncertainties induced by the above-mentioned issues. As with any system monitoring, a first major step (concerned with this work) is to validate the proposed model by human experts (in the field of hydrogeology), prior to deployment in production, using an online monitoring process (which is not in the focus of the present study). To mitigate the outlined complexity and uncertainties, the novelty of this paper concerns (i) the development of a method allowing to handle the high rate missingness (that cannot be solved with traditional methods), combined with (ii) the definition of a decision rule based on water quality standards, to allow the implementation of effective detection and diagnostic methods (since no prior knowledge on the state of the collected samples is available). The rest of the article is organized as follows: section 2 reviews studies related to water quality, section 3 describes the proposed methodology and section 4 provides its implementation to a case study, the theoretical background and the main principles of the implemented methods, along with associated discussions on the results. Finally, section 5 concludes the work and gives directions for further improvements.

2. Related work

This section reviews the relevant previous research works in the field of water quality, focusing on the detection and diagnostics issues that are related to the main scope of this study. Following Tubbett (Tebbutt, 1997), it is worth noting that all natural waters contain a variety of contaminants, due to erosion, leaching or weathering processes. In addition to this natural contamination, two other types can be noted: the first one is prompted by domestic and industrial wastes, the second one is due to the use of pesticides in intensive farming (which is the

context of this study). However, any amount of water is capable of assimilating a certain amount of pollution without serious effects, because of dilution and self-purification processes (Tebbutt, 1997); which makes it hard to provide universal method for analyzing water quality. This high variability in the character of water led to diversity in approaches and techniques of analysis, as reflected in the associated research works literature summarized in the following subsections.

2.1. Water quality detection

In this study, as mentioned above, water quality detection relates to the detection of its pollution, with regard to potential groundwater contamination by human activities in intensive farming. In the following, the term anomaly will then refer to a pollution detected in a given sample. In the literature, two main approaches for detecting anomalies in water have been proposed: model-based and data-driven. Model-based approach comprises methods that are established on the basis of mathematical formalization of the physics of the system under study, while data-driven one includes methods that take into consideration parameters of interest that go beyond the physics of the system. In practice, both require data to be implemented. However, while in the first approach few data are needed and used only in an identification purpose to set the physical model, in the second one, a large amount of representative data is required to explore all the possible states of the system, using data analytics.

Traditional model-based techniques include spectrophotometry (Abbas and Mostafa, 2000), electrochemical analyses (Denuault, 2009) or chromatography (Lamb et al., 2006). Unlike those methods, Raman's spectroscopy has high precision and short detection time (Li et al., 2015). The parameters of this method are different, depending on the type of pollutant which can be organic, inorganic or biological. Although there are many applications of Raman's spectroscopy (Li et al., 2014), few challenges have emerged, that require further works, particularly in the area of instrumentation and for determining the right type of spectroscopy technique for each pollutant. In most studies using a model-based approach, a specific pollutant is targeted. For example, in (Eli-

ades et al., 2015) the authors used real-time estimates of chlorine concentration, at the lower-bound and multi-level thresholds, to reduce false positive alarms in an existing event detection system. Although significant number of these studies have shown convincing results in the analysis of the water state, they have been conducted almost exclusively on the basis of a relatively small number of parameters of interest, focusing on the water phenomenology. Likewise, their contexts of application are not representative of all the variations that can characterize the very frequent evolution of the character of water (as mentioned above). In a context of scarce water resources (due in particular to the effects of climate change), combined with an era of intensive farming (to meet human food demand) that is increasingly exposed to health risks that have received public attention, all possible health states of water should be explored, so as to better explain the impact of human activities. In this direction, advances in sensing devices have opened a promising channel to collect large amount of data from high number of parameters of interest, while the development of AI provides suitable models and methods that allow complex analyses of high number of variables, in order to extract insightful knowledge to decision-makers, in a reasonably low processing time.

Data-driven studies undertaken in the field of water pollution can be organized into two main categories: those using a synthetic quality indicator (i.e. WQI), and those related to knowledge extraction from data, i.e. based on Machine Learning (ML). As in the case of model-based methods, studies using a synthetic quality indicator are generally carried out with few parameters, i.e. a real value is determined to summarize the overall water quality, usually based on a weighted combination of parameters of interest. Several WQI have been proposed in such a way (Balan et al., 2012; Lumb et al., 2006; Cude, 2001), each distinguished by the parameters used in the associated formula, the weights assigned to each parameter, and the aggregation method. In (Vasanthavigar et al., 2010), for instance, an index has been proposed to assess groundwater quality in India. This index is based on 12 parameters and uses Indian Water Quality Standards, along with human expertise to determine the weights of each param-

eter. In such a study, an advantage of using human expertise is to solve the issue regarding the lack of knowledge, allowing to label the samples, and to provide a method to take into consideration the variety of the water purposes. However, the choice of the parameters of interest remain questionable and may not be relevant to another country. Besides, data considered in the above-mention study are complete and periodic (namely biannual), which does not meet the requirement of high missingness rate, as it is the case in this research work.

In the field of ML, two main categories of methods have been applied to water pollution to address the detection issue: supervised methods (i.e. samples are labelled according to their quality level), and unsupervised methods (i.e. no prior knowledge is avalaible to assign labels to the data samples). Supervised methods, such as Artificial Neural Networks (ANN) and Support Vector Machine (SVM) (Castillo et al., 2016), have been successfully used to assess water quality, whereas deep learning methods usually provide best results (Dogo et al., 2019). However, the closest data-driven methods that meet the specific context of this study are unsupervised ones (since no prior knowledge on the state of each sample of the dataset is available). Within the frame of these methods, the most common techniques implemented are those aiming at determining water quality classes. The idea behind such techniques is that, by discriminating samples into different classes, one may be able to isolate contaminated ones. But the problem is that (since samples are not labelled), after performing clustering, human expertise is needed to characterise each class, so as to find the contaminated ones. Cluster analysis was, for instance, applied in (Machiwal and Jha, 2015) where the data considered are regular, with the characterization of classes from "poor" to "excellent", manually determined by an expert. On their side, the authors in (Balderas et al., 2017) have performed clustering on a dataset composed of 72 pollutants and biological indicators, with 9.70% of missing data. An imputation-based method, namely Multiple Imputations by Chained Equations (MICE), has been applied to handle the missingness issue, while outliers have been replaced using k-Nearest Neighbours (kNN). Water quality levels (characterized from "clean" to "very polluted") were derived on

the basis of the amount of pollutants present in the targeted site. The authors concluded that the results obtained on the pre-processed datasets were more consistent than those on the non-preprocessed ones. Indeed, an interesting feature of MICE is that missing values are imputed taking into consideration the relationships among the variables of the observations under study, leading to a preservation of the overall behavior of the original dataset. However, when this method is applied in its standard form, linear relationships are used by default, which may not be relevant to any context, especially the one concerned with the present study that is characterized by a high number of variables, in conjunction with a high rate missingness.

To summarize, on one hand, indicator-based methods applied to address water pollution are usually implemented with few paramters, chosen in a specific context, and using specific standards that are relevant only in the country or area for which they are defined. On the other hand, ML-based methods (including unsuperivsed methods that are close to the context of the present research work) are generally performed using friendly datasets and/or focus more on analysis than on data processing. Even when imputation methods are implemented to address the missingness issue, possible complex relationships among variables (rather than linear ones) are not taken into consideration, which is strongly required in case of high number of variables. As a consequence, the main issues addressed in this study, for an efficient detection of water pollution (as defined in the introduction), remain poorly or partially solved by the works reported in the literature.

2.2. Diagnostics of water anomalies

In this study, diagnostics of water anomalies refers to the process of determining the possible causes of water pollution (particularly in relation to the use of pesticides in intensive farming). Similar to the detection issue, two main approaches exist in the literature: model-based and data-driven.

In the first approach, modeling the contaminant's reactivity and transport can be conducted in order to determine the causes of water pollution. In (Al-

masri and Kaluarachchi, 2007), for instance, the authors proposed a modeling framework integrating both point and non-point sources of pollution, by considering soil nitrogen dynamics and groundwater flow, along with nitrate and transport, to study the impact of land use on nitrate pollution. The method allowed a better understanding of the behavior of nitrate in groundwater and provided decision support for water management. In order to identify nitrate sources and transformations, isotopes of nitrogen and oxygen, along with a Bayesian mixing model, were successfully implemented in (Zhao et al., 2019). Once again, although convincing results have been obtained for water pollution diagnostics, it appears that high chemical expertise is strongly required. Moreover, the associated methods are more contaminant-oriented and, therefore, do not take into consideration non-chemical parameters that may be of great interest for the analysis. Besides, these methods are undertaken on the assumption that the possible causes of the pollution are known, at least suspected with regard to the observation of the state of the water resource. This does not correspond to the context of this study, where analyses must be carried out in a blind way and which are not easy to achieve, given the large number of parameters to consider. For all these reasons, data-driven methods are more appropriate.

Data-driven methods for identifying the main causes of water pollution predominantly consist in combining Hierarchical Cluster Analysis (HCA) and Principal Component Analysis (PCA) (Simmonds et al., 2017; Machiwal and Jha, 2015). In (Mastrocicco et al., 2017), for instance, chlorate origin in groundwater is associated to variation in nitrate, volatile fatty acids and oxygen reduction potential. PCA allowed then to significantly reduce the dimension, but was applied only on the parameters that were measured at each campaign (with no missing values). It is widely acknowledged that PCA does not preserve distances between observations in the reduced space, which can result in the discarding of discriminating information. Moreover, PCA is recognised to outperform competing methods only in case of small training datasets (Martinez and Kak, 2001), which is not the case in this study. Despite its unsupervised

nature that meets the requirement regarding the lack of knowledge on each sample, this method is not suitable to answer the pollution diagnostics issue, in the context of the present research work. Indeed, this latter requires a method allowing to strongly discriminate pollutant and non-pollutant samples in order to better diagnose pollution causes.

To sum up, according to this review of the previous studies related to detection and diagnostics of water pollution, the following findings can be noted.

- Data-driven methods, which are suitable for water pollution detection and diagnostics, are generally applied to complete and regular datasets. In addition, studies that have addressed the issue related to missing values have been undertaken on datasets with low rate missingness. Therefore, methods allowing to preserve the behavior of the raw data after imputation, i.e. taking into consideration **complex relationships among a high number of variables**, are required.

- Algorithms performance is improved after pre-processing the raw data. Therefore, a systematic application of the pre-processing step is needed in order to provide, in a **low processing time**, relevant answers to the issues addressed.

- The above discussions have also shown that the main problem addressed is unsupervised in nature (samples are not labelled).

  - As noted in (Tebbutt, 1997), examining the various standards and guidelines that are used to specify water quality for various uses can help characterizing the state of each sample, so as to determine decision rules for labelling them. Since no universal health indicator is available in the field of hydrogeology, expert knowledge is needed to provide relevant water quality sub-classes. This will help **transforming the problem from unsupervised form into supervised**.

– To provide relevant sub-classes of water quality, a method allowing strong and **clear separation between the resulting sub-classes** is required, which will reduce the need of expert intervention for validating the proposal.

In order to meet the above mentioned requirements, a general method allowing to take into consideration the whole lifecycle of the data should be established. For this purpose, the transposition of PHM steps (Fig. 2) to water quality could be one of the suitable candidates. The main phases shown in Fig. 2 are shortly described hereafter.

1. **Phase I: data acquisition.** This phase initializes the method by providing the raw data that will later be subjected to various analyses. In this study, data were collected from a catchment well. A part of these data were obtained from sensors and the other part from lab chemical analyses. Detailed information on the organisation and operation of the considered well is provided in the subsection describing the case study.

2. **Phase II: data pre-processing.** It is concerned with tasks that are performed to clean up the raw data in order to make them reliable and exploitable for relevant analyses. To this end, among others, issues such as heterogeneity of measurement periods, missingness, dimensionality (in case of high number of parameters of interest), followed by feature extraction, are addressed.

3. **Phase III: data analyses.** It is at this phase that water quality analyses, in relation to detection of anomalies and diagnostics of their underlying causes, are undertaken. In addition, PHM provides a specific analysis task, allowing to anticipate failures of the system under study. For this purpose, prognostics methods are developed. However, this prediction step is out of scope of this paper.

4. **Phase IV: decision support.** As its name suggests, this phase aims at providing support to decision-makers in order for them to keep the system

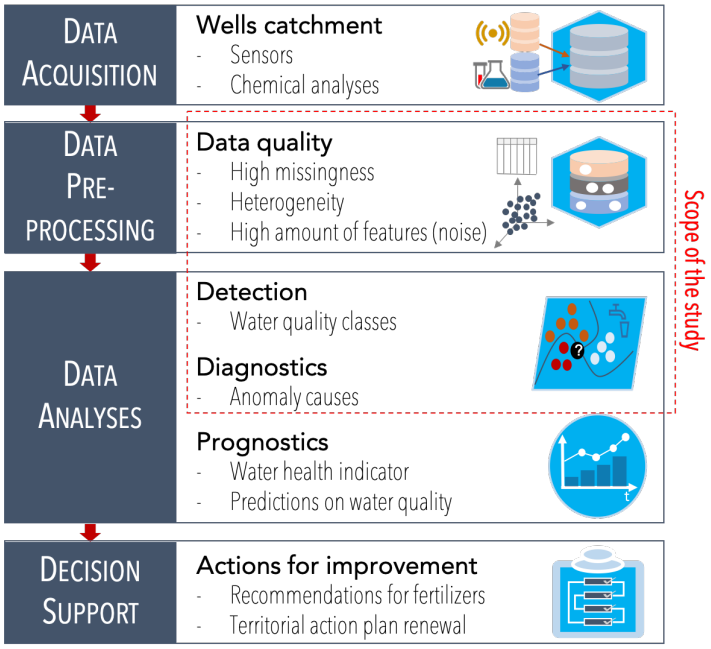

Figure 2: PHM framework applied to water quality analysis.

in a healthy state, usually in the form of recommendations for actions to be taken. Among others, data acquisition strategy, allowing for instance to focus on the most influencing parameters, can be provided.

As shown in Fig. 2, only **pre-processing, detection and diagnostics** are considered in the scope of the present study. The proposed PHM-based methodology, to handle these issues, and its associated results will be presented and discussed in the following sections.

3. The proposed methodology

This section describes the proposed methodology to handle detection and diagnostics of water pollution, in relation to human activities in intensive farming. It is an instanciation of the PHM framework presented above. One aim is to show how it can better address the three main issues specific to the context of this study. These issues are: (1) high rate missingness, (2) high number of

parameters of interest and (3) unlabelled samples (i.e. no prior knowledge on their state). A first sub-section outlines the main steps of the methodology, followed by a sub-section describing the theoretical background and the main principles of the implemented methods.

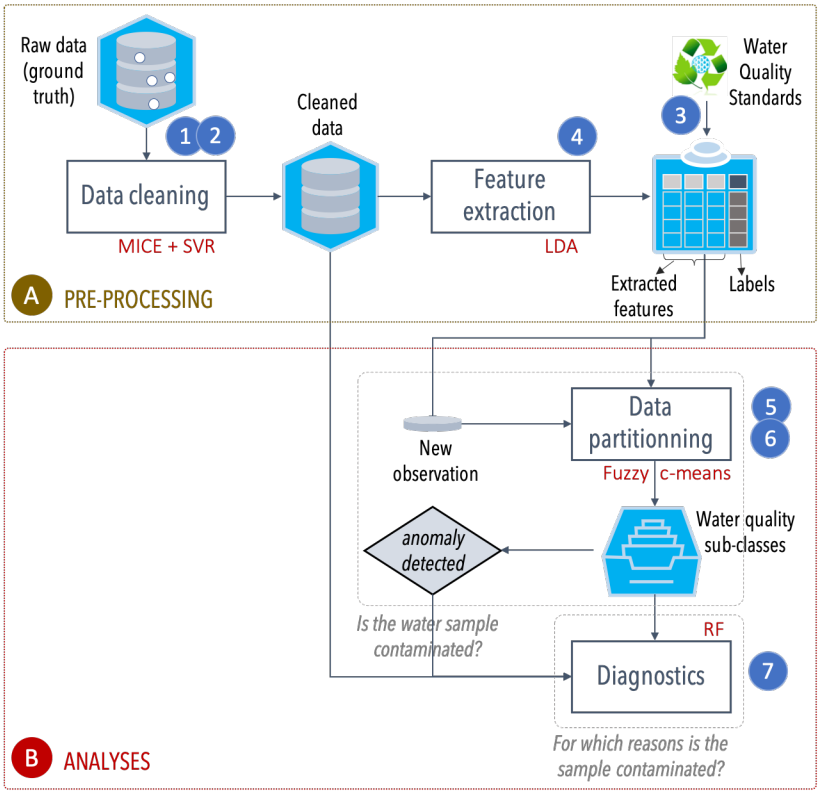

Figure 3: The proposed methodology for detection and diagnostics of water pollution.

3.1. Main steps of the methodology

Fig. 3 highlights the seven main steps implemented and the underlying tasks which are structured into two main phases: pre-processing and analyses.

A. Pre-processing

- **Step 1.** It can be noted that the raw data of the study were provided with some unnecessary measurements to handle the above mentioned issues. Thanks to the intervention of hydrogeology experts, combined with

some exploratory analyses, the observations gathered were then cleaned up. To this end, outliers, observations within a period where measurements techniques were judged not reliable and parameters that were rarely monitored have been removed.

- **Step 2.** The raw data of the study were also provided with a large number of missing values. In fact, some measurements were not conducted because chemical analysts may have found them unnecessary at some times when environmental conditions were quite good. But the problem, for the data analyst, is that this can lead to a loss of information and introduce bias. It was then necessary to fix this issue. Since the rate of missingness of the dataset considered was very high (more than 80%), missing values were filled using an algorithm described in (Ratolojanahary et al., 2019), combining MICE with Support Vector Regression (SVR). Indeed, among other competing methods, SVR gave the best results and was suitable to build non-linear relationships between the variables. A model selection methodology was proposed for that purpose.

- **Step 3.** Following the recommendations resulting from the literature review, a major choice was made at this step to handle the issue regarding the lack of knowledge allowing to label the data samples. As suggested in (Tebbutt, 1997), european Water Quality Standards (WQS) were used, so as to transform the unsupervised nature of the problem into an supervised one. As already discussed, this can refine the results of the detection and diagnostics procedures. To this end, decision rules were defined as follows: if at least one parameter does not comply with the underlying threshold, the associated sample was labelled *non-compliant*, otherwise *compliant*.

- **Step 4.** It is worth noting that dimensionality reduction is important to mitigate uncertainties induced by redundancy, presence of unnecessary parameters, and also to speed up the processsing time. To cope with the necessity to properly discriminate pollutant and non-pollutant samples (as discussed in the literature review), Linear Discriminant Analysis (LDA)

was adopted for feature extraction. Since it is a supervised method, its use in conjunction with the samples labelling of step 3 is therefore suitable to handle the dimensionality reduction issue addressed in this study.

- **Step 5.** Once the samples are labelled, a trick in using LDA is also to provide a mechanism to handle the detection issue. Indeed, since the samples are labelled (step 4), LDA will generate sub-classes of compliant water quality and the same for non-compliant observations. Thus, given a new observation, the underlying model will be able to predict the sub-class to which it belongs. The left side of Fig. 4 outlines this procedure. In order to guide this mechanism, data are divided into compliant and non-compliant categories, using decision rules elaborated from european WQS.

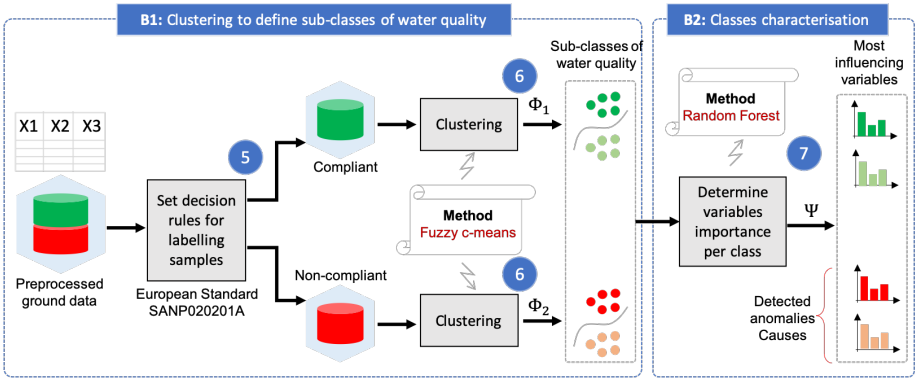

Figure 4: Detailed operation of the data analysis phase.

- **Step 6.** The previous step resulted in two sets of data composed of compliant and non-compliant samples, each used for clustering. Since the corresponding samples are labelled (using WSQ), clustering will actually result in sub-classes of compliant observations and those of non-compliant ones. This will make it possible to precisely determine several sub-categories of water quality, beyond a simple separation of data into compliant and

non-compliant observations. Considering sub-classes obtained with non-compliant data, one can then gather insight on water pollution causes, by characterizing their most influencing parameters. In order to fine-tune this procedure (as expected for relevant diagnostics), a fuzzy-based clustering method was implemented: Fuzzy c-means (FCM). Thus, for each given observation, it was possible to predict to what extent it belongs to a given class, using a membership function that gives the probability of belonging.

B. Analyses

- **Step 7.** The inputs of this step are the different sub-classes of water quality obtained in the previous step. Although the trick introduced in step 6 has provided us with different classes of compliant observations, and others of non-compliant ones, this does not indicate which parameters characterize each of these classes. To cope with this issue, a decision tree-based method was adopted. Among the competing methods of the literature, Random Forest (RF) (Breiman, 2001) is the one which can solve this issue. Among its main features, it should be noted that RF is robust against overfitting. In addition, the test procedure on top of which it is built allows to explore all the classes, preventing the model from bad performance in case of imbalanced classes. Furthermore, in practice, its implementation allows computing variables' importances that are of great interest for diagnostics. The right side of Fig. 4 outlines the underlying process.

Considering the resulting sub-classes within the non-compliant observations and their corresponding variable importances (computed through RF), it becomes possible to define the most influencing parameters of water pollution. These conclusions are then confronted with hydrogeology experts for validation, before deployment in production through online monitoring (which is out the scope of the present study).

3.2. Theoretical background and principles of the implemented methods

Multiple Imputations by Chained Equations (MICE), Linear Discriminant Analysis (LDA), Fuzzy c-means (FCM), Random Forest (RF) and Genetic Algorithm (GA) are the main methods required to implement the proposed methodology. For simplicity, the following presentation focuses on their key features.

3.2.1. MICE

According to the literature, there are 3 different mechanisms to describe what can lead to missing values (van Buuren, 2018): (1) Missing Completely At Random (MCAR), i.e. the probability of missingness is independent of both the observed variables and the variables with missing values; (2) Missing At Random (MAR), i.e. the probability of missingness is due entirely to the observed variables and is independent of the unseen data; (3) Not Missing At Random (NMAR), i.e. the missing value is related to the actual values. As shown in (Ratolojanahary et al., 2019) our study relates to the MAR category. It should be noted that MICE was used to handle the missingness, due to its ability to preserve the overall behavior of the original data, and to take into consideration the relationships between the variables of interest. Indeed, simplistic imputation (such as mean or median) may not be relevant. For instance, imputing the missing value of a temperature with the mean of the samples may not be consistent if the imputed sample in question was collected in winter whereas most existing samples were collected in summer. Such a choice would generate an unfortunate bias that can lead to inappropriate decision making.

3.2.2. LDA

The main principle of LDA is to search for the feature space in which compliant and non-compliant water samples are the best separated (as shown in Fig. 5). To this end, eigenvectors based on covariance matrices of *between* and *within* classes are determined. To choose the number of linear discriminants, the cumulative percentage of explained variance is then calculated.

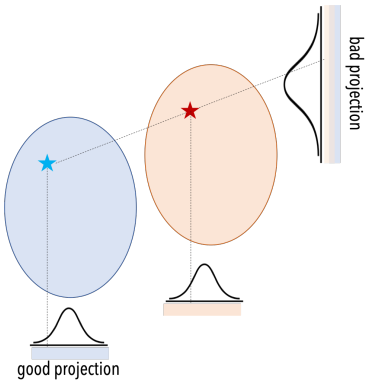

Figure 5: Classes separation in Linear Discriminant Analysis.

3.2.3. FCM

FCM, initially developped by Dunn (Dunn, 1973) and improved by Bezdek (Bezdek, 1981), was adopted for that purpose. Its main principle is to assign each observation to a given cluster with a degree of membership (i.e. a coefficient) (Bezdek, 1981), guided by the follwing procedure:

- Choose a number of clusters c;
- Assign randomly to each point, degrees of membership to each cluster;
- Repeat until convergence, i.e when the coefficients change between two iterations is lower than a predefined threshold ε :
  - Compute the centroid for each cluster using Eq. (1):

$$c_k = \frac{\sum_{i=1}^{c} w_{ij}^m x}{\sum_{i=1}^{c} w_{ij}^m}, \tag{1}$$

    where w_{ij} is the degree of belonging of the observation i to the cluster j, and m is the fuzzifier;
  - Compute the new coefficient for each point using Eq. (2):

$$w_{ij} = \frac{1}{\sum_{i=1}^{c} \frac{||x_i - c_j||}{||x_i - c_k||}^{\frac{2}{m-1}}} \tag{2}$$

3.2.4. RF

The main objective of RF is to build several weak learners (shallow decision trees) in parallel in order to produce a stronger classifier. The key is to choose a large number of uncorrelated trees that will together outperform any individual model. The underlying reason of this powerful feature is that uncorrelated models (i.e. trees) will protect each other from their individual errors. The main steps of RF are the following:

(a) Select a random sample of the observations with replacement;

(b) Select a set of variables randomly;

(c) Choose the variable providing the best split;

(d) Repeat step (c) until all nodes are pure or the maximum depth is reached;

(e) Repeat steps (a)-(d) until the specified number of trees is reached;

(f) Decide upon a majority vote.

For diagnostics purpose, the importance score of each feature is computed by averaging the difference in out-of-bag error, before and after the permutation over all trees. The score is then normalized by the standard deviation of these differences (Zhu et al., 2015). In order to provide good results, RF has to be fine-tuned by setting its hyperparameters. These latter are the parameters that have an influence on the quality of the algorithm (in contrast, parameters refer to those that define the underlying model, used to predict the result for a new observation).

Hyperparameters such as the number of variables (n_var) to consider at each split, the criterion ($crit$) for choosing the best split while constructing the trees, the maximum depth of the trees (max_depth) and the number of trees (n_trees) are those concerned with a tuning procedure. The combinatorial nature of their respective choice makes it hard to provide good performance. To address this issue, Genetic Algorithm (GA) (Ng and Perera, 2003; Goldberg and Holland, 1988) was combined with RF.

Indeed, a well-known issue in supervised learning is the misclassification due to (1) *underfitting*, which reflects the bias issue, or to (2) *overfitting*, which pertains the variance issue. To handle such problems, the choice of the hyperparameters of the implemented methods becomes crucial and falls within the frame of combinatorial optimization. One of the basic methods for hyperparameter tuning is grid search: it consists in building a model for each of the combination of the parameters provided, which makes it simple to implement. However, the size of the search space can increase exponentially, leading to a high processing time. A randomized version of grid search allows exploring larger search space, picking randomly one combination of values in each iteration. But the way it operates cannot guarantee the most appropriate choice of hyperparameters, since it is not an optimization method. In (Dhaenens and Jourdan, 2019), the authors investigated several metaheuristics for data mining, among which, GA, Particle Swarm Optimization (PSO), and their hybridization. Following the authors, and to the best of our knowledge, no single method can outperform the others in all cases: the performance depends on the search space defined to fine tune the hyperparameters, and on the data themselves, regarding their size, their variety, their veracity and their velocity (within the broad frame of big data concerns).

3.2.5. GA

It should be noted that, in the context of the present research work, big data concerns do not matter. Among the metaheuristics provided in the literature, GA, which has shown good performance in many cases, has then be chosen for the present work in order to improve a basic or randomized grid search. Briefly, given an optimization problem, GA operates through these main steps:

(a) Create an initial population, composed of individuals that are potential solutions of the problem (usually, randomly selected);

(b) Compute the fitness score of each individual (which is a criterion allowing to find the "best" individuals for the next generation);

(c) Select the best fitted individuals;

(d) Perform a crossover of each pair of parents from step (b), so as to randomly introduce new features in the next generation;

(e) Perform mutation (with a low probability) to maintain diversity and prevent premature convergence;

(f) Repeat steps (c)-(e) until convergence or until a specified number of generations is reached.

In this study, each individual is set as a vector of size 4, corresponding to the hyperparameters to optimize (n_var, $crit$, max_depth, n_trees), and the fitness score is the AUC (Area Under the ROC Curve) of the derived RF model. The search spaces defined for these hyperparameters are given in the following:

- $n_{vars} = \{\text{“}auto''$, “$sqrt''$, “$log2''$, $None\}$;
- $crit = \{\text{“}gini''$, “$entropy''\}$;
- $max\_depth = \{5, 10, 15, 20\}$;
- $n\_trees = \{100, 101, \ldots 500\}$.

4. Implementation of the proposed methodology

This section presents an implementation of the proposed methodology to address detection and diagnostics of water pollution, in relation to human activities in intensive farming. As noted earlier, the raw data used hereafter were acquired from a catchment well located in the southwest of France.

4.1. Case study : catchment well at Oursbelille (France)

This case study is concerned with a real-world groundwater dataset collected from a catchment well at Oursbelille, in the southwest of France. Fig. 6 shows the corresponding geographical localisation (left side of the figure) and the operation principle of the catchment(right side of the figure).

- (1) Groundwater is first pumped and treated with active charcoal which is the most effective mean for removing chlorine, particles such as sediment, volatile organic compounds (VOCs), taste and odor from water;

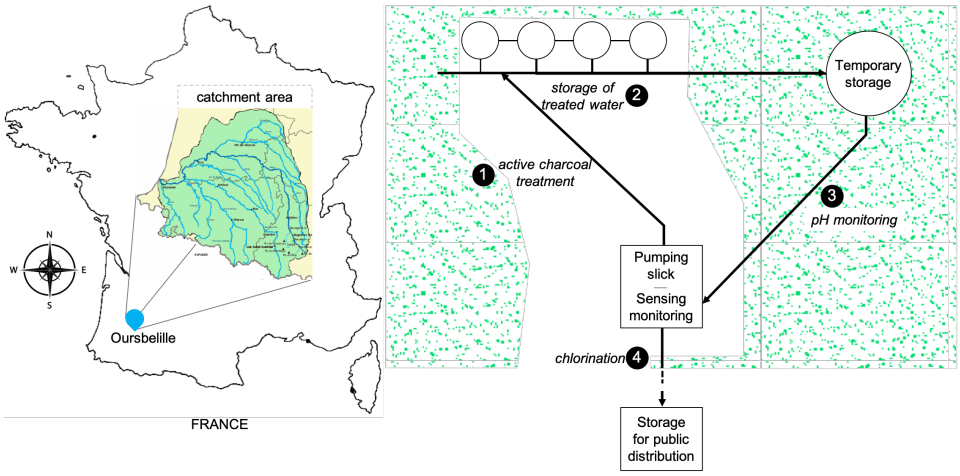

Figure 6: Location of the case study and operation principle of the catchment.

- (2) Pre-treated water is then temporarily kept in storage tanks;
- (3) On demand, the stored pre-treated water is brought back for pH monitoring. Let us recall that pH is the measurement of how acidic or alkaline (basic) a solution is. It corresponds to the balance of the amount of acid and base chemicals in that solution. In drinking water, its values range from 0 to 14: 7 is neutral (i.e. there is a balance between acid and alkalinity), a value below 7 means acid is present and above 7 means the solution is basic (or alkaline). Among the various types of measurements that are available, in this case study, sensing devices were adopted;
- (4) Finally, before sending water to the inhabitants, it is chlorinated and then routed to a distribution center. This final treatment is usually carried out as primary or residual disinfection.

In 2008, farming practices have increased the level of nitrates and pesticides

in that area, which has led to a territorial action plan. To that end, measurements of 414 water quality parameters (including chemical, physical and biological) were acquired from 1991 to 2018. It is worth noting that monitoring is still made at several levels of the studied area, but the data used for this study are those concerned with the measurements at the source (i.e. before sending the water to the inhabitants). In order to deepen the understanding of these data, in relation to their correlation with water pollution, preliminary statistical investigations were conducted.

4.2. Data exploration

The main idea behind the exploration made was to understand the data acquisition strategy and then analyse its consistency. In that purpose, the number of measurements per parameter was considered. As outlined by Fig. 7, only few parameters were frequently measured. In other words, the number of measurements per variable is very unequal, ranging from 1 to 127.

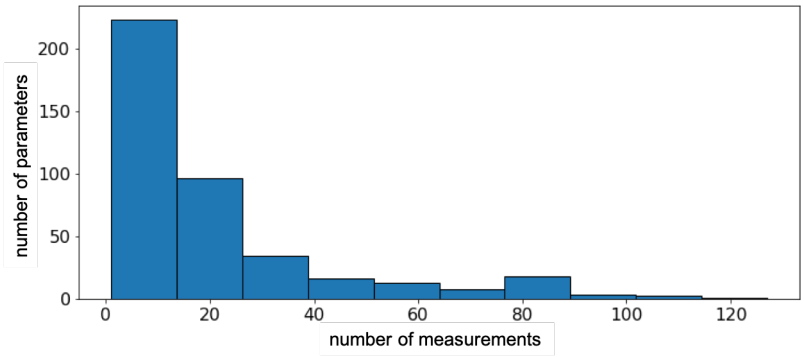

Figure 7: Number of parameters per measurement.

To deepen this first analysis, the number of measurements per year was also considered. As shown by Fig. 8, monitoring has become regular and more consistent from 2002. For example, the statistics of three parameters, namely `nitrate`, `pH` and `sum of pesticides` are given in Table 1: the table shows that, on average, the values of these parameters are out of their corresponding

Table 1: Statistics of 3 parameters.

	Nitrate	pH	Sum of pesticides
unit	mg/L	none	$\mu g/L$
count	127	97	29
mean	43.74	6.73	0.37
std	4.03	0.20	0.19
min	28.70	6.40	0.05
25%	41.00	6.60	0.22
50%	44.00	6.70	0.39
75%	46.00	6.80	0.50
max	55.00	7.70	0.79
tolerance	< 50	$[6.5, 9]$	< 0.5

tolerance ranges (i.e. under 50 mg/L, between 6.5 and 9 and under 0.5 μg/L, respectively).

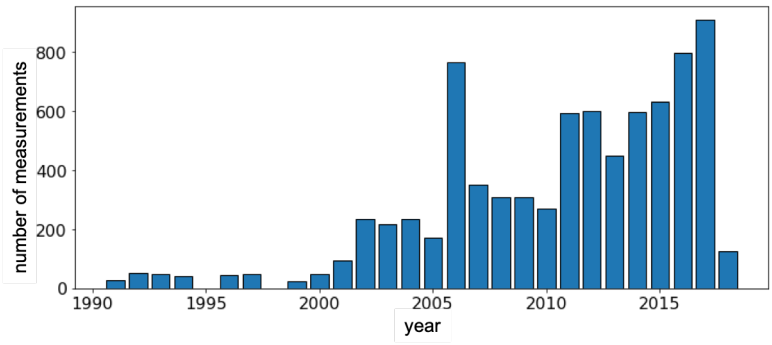

Figure 8: Number of measurements per year.

These exploratory analyses illustrate the complexity for implementing a data-driven approach: it requires studies that go beyond the physics of the system under consideration and a variety of influencing factors should be taken into account. In this case study, more than 400 parameters were identified as potential water pollution factors of contamination, but only few were regularly

Table 2: Basic statistics of datasets resulting from pre-processing.

	Raw data	Cleaned	Imputed	Reduced
Numb. of observations	156	133	133	133
Numb. of features	414	58	58	3
Time span	1991-2018	2002-2018	2002-2018	2002-2018
Missing data	84%	60%	0%	0%

measured, resulting in high rate missingness (within the whole period of measurements). This is why a robust imputation mechanism was required during the pre-processing phase to provide relevant conclusions.

4.3. Implementation

The implementation of the proposed methodology was carried out according to a general procedure summarized in Algorithm 1. Each step of each phase is described in the form of a pseudo-code, including (A) pre-processing and (B) data analysis phases of the PHM general framework.

4.3.1. Implementation of the pre-processing phase

According to the procedure described earlier, the pre-processing consisted in cleaning up the raw data (step 1), imputing the missing values (step 3) and reducing the dimension (steps 3 and 4) in order to make the data exploitable. The main results of the pre-processing are summarized in Table 2. After the first cleaning step, only 58 variables out of the 414 provided were retained.

Fig. 9 illustrates the trend of some parameters after imputation. The blue lines represent the (partial) trend of the raw data, the red ones show the trend of the imputed data, and the dashed horizontal lines indicate the threshold of the corresponding parameter (as defined in WQS). As the figure suggests, the imputed data tend to follow the overall behavior of the original data, and therefore provide visual validation of the consistency of the imputation procedure (step 2). In addition, other analyses (which are out of the scope of this study) have

Algorithm 1 Implementation of the proposed methodology

Input: Raw data (incomplete, no labels).

Output: Data with labels (water quality level).

(A) Pre-processing

(1) Clean up the data using expert knowledge and exploratory analysis

(2) Impute missing values using MICE+SVR

(3) Add labels using WQS

for all $o \in observations$ **do**

  $label(o) = compliant$

  for $p \in parameters$ **do**

    if $o(p) \notin toleratedrange(p)$ **then** ▷ According to european WQS

      $label(o) = non\ compliant$

    end if

  end for

end for

(4) Reduce the dimension using LDA

(B) Analysis

(5) Divide the data into compliant and non compliant samples

(6) Determine water quality sub-classes

for $data \in \{compliant\ samples, non\ compliant\ samples\}$ **do**

  Cluster the data using FCM ▷ Gives model Φ_1 for compliant samples and Φ_2 for non-compliant ones

end for

Given a new sample, say x_{new}, compute $\Phi_2(x_{new})$ ▷ Degrees of membership of x_{new} to non-compliant sub-classes

(7) Identify the cause of anomalies

Perform RF classification ▷ On imputed data (before dimensionality reduction)

for all determined *clusters* **do**

  Find the most influencing parameters through RF

end for

595 shown that the natural cycle of water contamination was preserved; which was also validated by hydrogeology experts involved in the present research work.

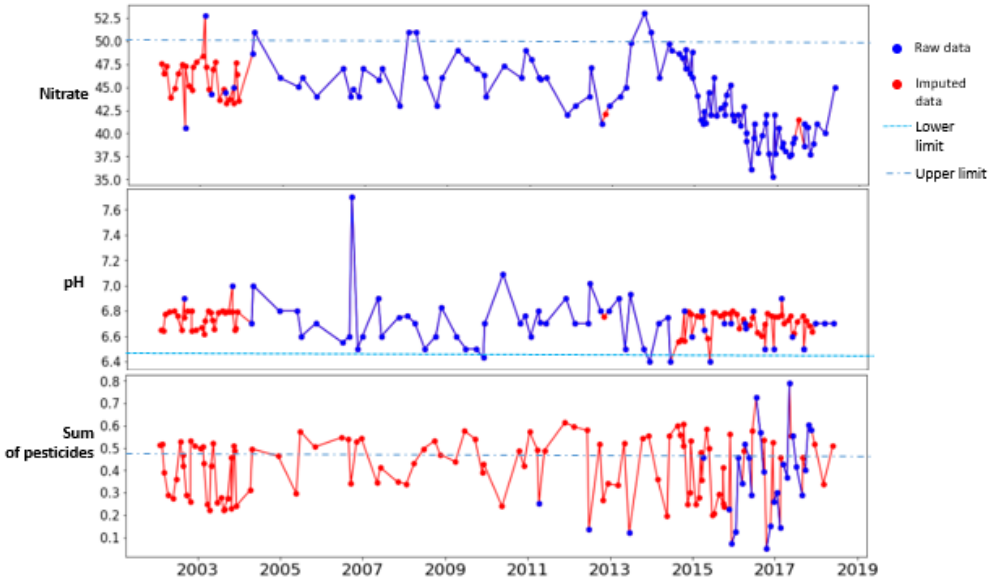

Figure 9: Variation of some parameters over time.

The application of water quality standards gave 60 compliant samples and 73 non-compliant ones (step 3). Using LDA (as required in step 4), the samples were projected into a three-dimension space, showing the good performance of
600 LDA: 100% of the variance explaining the difference between compliant and non-compliant observations was retained (Fig. 10).

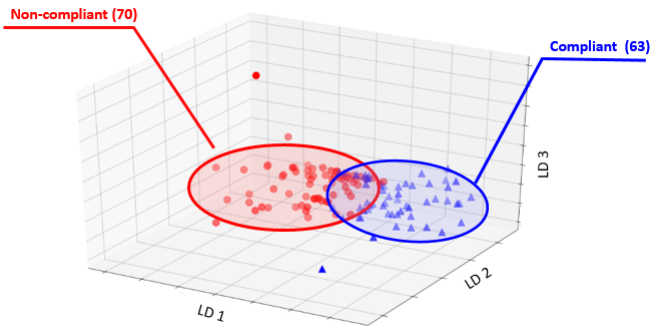

Figure 10: Representation of the data in the linear discriminant space.

4.3.2. Implementation of the detection phase

Once the raw data are pre-processed, it is possible to apply water quality analyses to provide relevant detection results. Detection is first carried out to specify the membership (polluted / non-polluted) of each single data sample. As explained in the presentation of the proposed methodology, the trick was to determine strongly discriminated sub-classes, each characterised by a certain level of contamination, thanks to the LDA implementation (step 4). To this end, the data were divided in two learning sets, compliant and non-compliant samples. In both sets, FCM was applied for clustering, using the following hyperparameters (step 6): the fuzzifier index $m = 2$, which is the recommended value (Bezdek, 1981), and the number of clusters c is determined using the silhouette score which suggests 2 clusters for both compliant and non-compliant samples (see Fig. 11). Indeed, the trick introduced in step 6 for detection provided 2 models (through FCM clustering), Φ_1 and Φ_2, characterising the compliant and the non-compliant samples (respectively). These models can then be used to deeply investigate the characterization of the sub-classes. Given a new observation, say x_{new}, the computation of $\Phi_i(x_{new})$ $(i = 1, 2)$ gives the respective degrees of belonging of x_{new} to each of the 4 sub-classes.

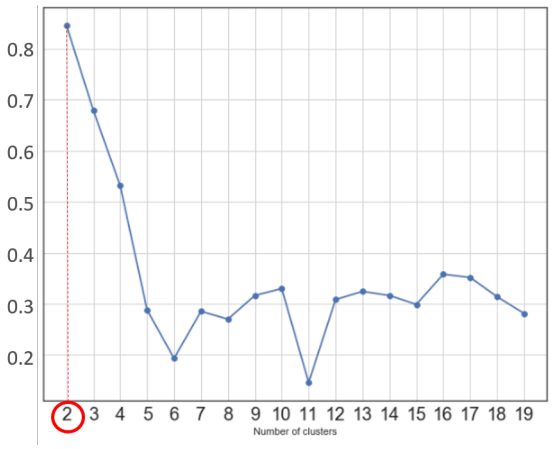

Figure 11: Silhouette score diagram for the non-compliant samples.

The result of the clustering implementation shows 2 sub-classes that characterise the water pollution (non-compliant samples), which is a fine-tuned analysis suggesting 2 possible causes of water pollution. Fig. 12 illustrates the overall 4 resulting sub-classes: the compliant sub-classes are represented in shades of green and the non-compliant ones in shades of red. To obtain these results, an euclidean metric was arbitrary used to compute the distance between observations. The right side of the figure shows the number of observations per cluster, which is overall imbalanced, justifying the appropriate choice of RF for diagnostics (step 7).

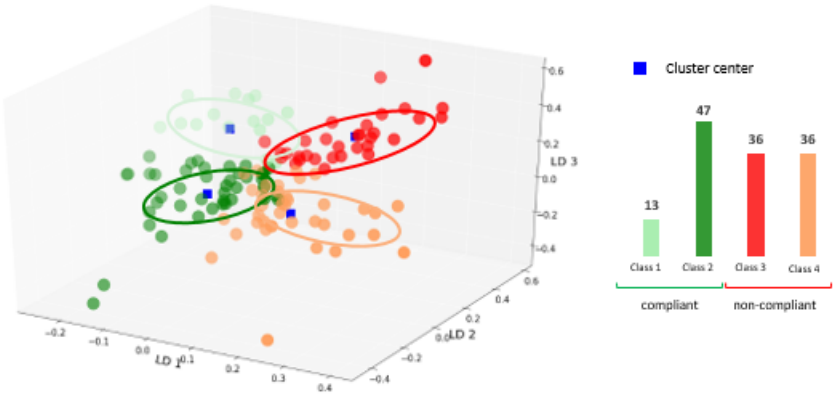

Figure 12: Water quality sub-classes.

4.3.3. Implementation of the diagnostics phase

Following the general rule set to characterize the samples, if the results of step 6 provide strong evidence that a new observation belongs to one of the 2 non-compliant sub-classes, it is declared abnormal. But this does not yet explain the reasons of this anomaly, which is the goal of the diagnostics task. To this end, following step 7, a classification problem using RF and the 4 sub-classes obtained in step 6, is performed. The most influencing parameters of the sub-classes are then computed.

To define a classification problem, a column named *class* was added to the cleaned imputed data: it corresponds to the 4 sub-classes found in step 6. For each class c, the following steps were finally applied to diagnose any detected anomaly:

- Create a new categorical variable equal to 1 if the observation belongs to the class c, 0 otherwise;
- Build a RF to discriminate that new variable;
- Retrieve the most influencing parameters to determine the class membership.

Table 3: Optimal choice of RF hyperparameters per water quality class

	class 1	class 2	class 3	class 4
n_var	log2	sqrt	log2	log2
crit	Gini	entropy	entropy	entropy
max_depth	10	7	4	7
n_trees	287	307	358	471

To optimize the RF operation, an hybridization with GA was performed, resulting in the choice of the hyperparameters values presented in Table 3.

It is worth noting that, while LDA allowed to reduce the dimensionality before applying the FCM for detection, RF was implemented on imputed data (see Table 2). The main reason of this choice was to perform RF on data containing original parameters, so as to be able to characterise each resulting sub-class accordingly.

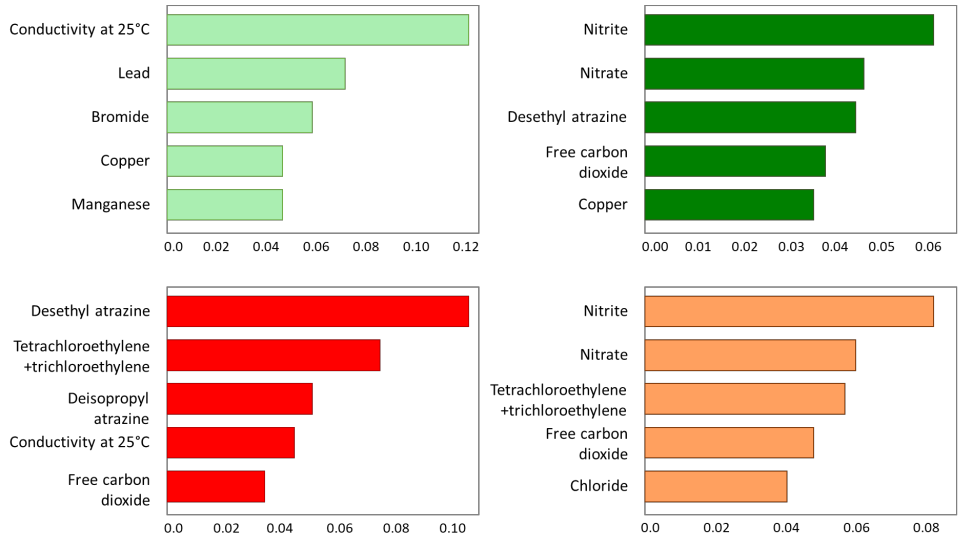

Figure 13: Most influencing variables per water quality class.

The process of retrieving the most influencing parameters in RF consists in

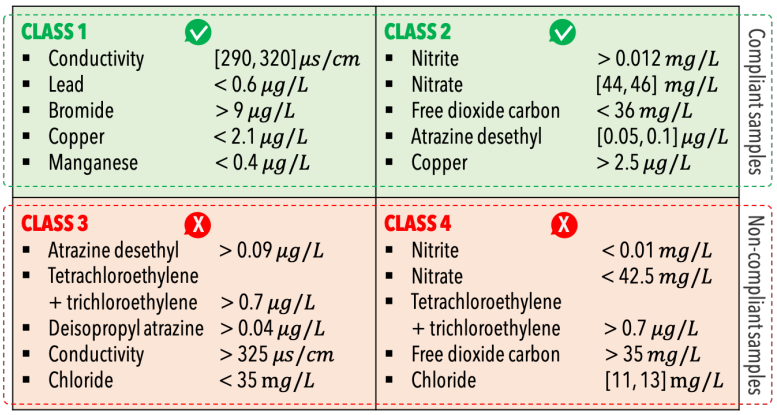

Figure 14: Characterization of the resulting water quality classes.

searching, for each class, those that discriminate most the current class from the others: these are the parameters with the most predictive power. They are the drivers of membership in their corresponding class, since they have significant impact on its characterization. In contrast, the low importance parameters are less significant and can even be omitted from the model without loosing its predictive power, making it more simple and faster.

In the present research work, it has been decided to compute not a single influencing parameter (i.e. the top ranking), but the five first ranking ones (as depicted in Fig. 13). In this figure, the color code remains the same as above (i.e. shade of green for compliant observations and shade of red for non-compliant ones).

To investigate the characterization of the resulting water quality classes, Fig. 14 gives the values of the most influencing parameters per class using those of the average individual (i.e. the centroid of each class), thus providing hydrogeology experts with enough knowledge to validate the diagnostics.

- **Compliant samples**
  - Class 1 includes observations with relatively low concentrations of conductivity, lead, copper and manganese, but higher concentration of bromide than in the other classes.

- Class 2 corresponds to observations with higher concentrations of nitrite, nitrate and atrazine than those in class 1. In addition, the copper concentration is also higher and the amount of free dioxide carbon is lower than that of class 1. Based on this information, class 2 appears to be less compliant than class 1.

- **Non-compliant samples**

  - Observations of class 3 are characterized by high conductivity, with atrazine, tetrachloroethylene and trichloroethylene being more concentrated than the other classes. The chloride concentration is lower than that of classes 2 and 3.

  - Finally, class 4 contains observations with lower levels of nitrite, nitrate, tetrachloroethylene and trichloroethylene than class 3. Free dioxide carbon is higher than that of class 3.

In summary, parameters that are likely to exceed the threshold defined by the Water Quality Standards are atrazine and nitrate in class 3 and exclusively atrazine in class 4. Other parameters, such as conductivity, also indicate water quality deterioration even if their value complies with the standards.

4.3.4. Discussion

In view of the above results, the following observations can be made:

- The results from pre-processing as well as the classes characterization have been validated by hydrogeology experts. Any automation would have been possible to compute the diagnostics. As in any system monitoring study, it is necessary to have the model validated by a human expert from the studied domain, prior to deployment in production.

- The most important variables that characterize the 4 water quality classes were obtained through correlation analyses. But correlation does not necessarily imply causality. A counter-example is given in class 4 where the two most discriminating parameters are nitrites and nitrates. However,

their concentration level is lower in class 4 than in the other classes (as shown in Fig. 15). Thus, a complementary analysis, for instance, a confirmatory analysis using Structural Equation Modeling (SEM), is necessary to validate the causality suspected, unless human expertise is used (if available).

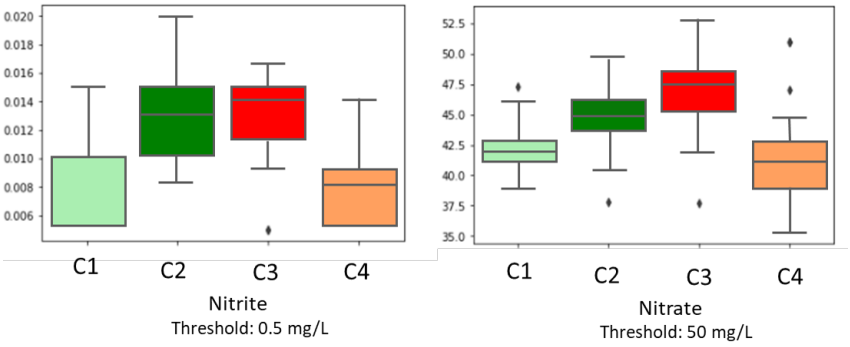

Figure 15: Distribution of nitrites and nitrates per class.

- The number of influencing parameters per class was arbitrarily set to 5. Instead, an automation could be used, such as the elbow method for instance. As illustrated in Fig. 16, only the first 3 parameters show evidence of their influence.

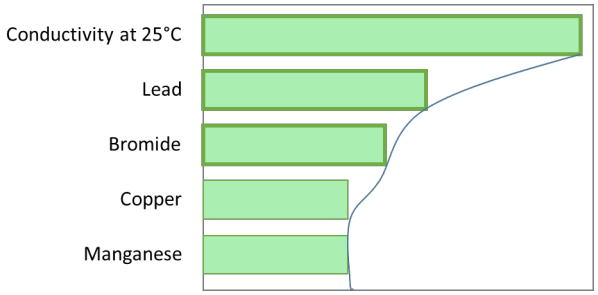

Figure 16: Elbow method applied to the influencing parameters of class 1.

- The most influencing parameters, obtained through diagnostics, are consequently representative of the quality of water resource. Therefore, it

may be recommended to monitor them closely and reduce the effort in collecting unnecessary data, so as to save time and costs of analyses. This may allow developing a generalizable and effective model.

5. Conclusion

In this paper, a PHM-based methodology for detection and diagnostics of water quality anomalies was proposed. This methodology allowed to obtain relevant results (validated by hydrogeology experts), using heterogeneous dataset, with high rate missingness and without prior knowledge on the health state of the samples collected. It performed good results for detection, thanks to the implementation of LDA that perfectly separated compliant and non-compliant observations. Concerning the diagnostics issue, an indirect method was proposed, using european WQS that allowed to switch from an unsupervised problem to a supervised one. Thanks to RF implementation, significant results to determine the 4 resulting water quality classes were provided, so as to allow deep investigation on their characterization. Nitrate and atrazine were finally declared the main sources of water pollution (more precisely, the sources of water quality disqualification), which was validated by hydrogeology experts involved in this research work. Such validation is commonly required in systems monitoring, prior to online deployment, which is planned for a further study. In addition, the characterization of the resulting classes allowed to recommend which parameters should be closely monitored, so as to mitigate the complexity induced by the number of parameters of interest.

Further studies to improve this work include using a fuzzy rule for labelling the samples and taking into consideration expert knowledge, combined with multi-criteria decision method, within the frame of a group decision. A second direction of improvement will consist in considering the time dimension in order to predict the future state of the water, that is, implementing the prognostics phase of PHM. Several methods like neural networks and time series prediction can be applied to this end.

References

Abbas, M., Mostafa, G., 2000. Determination of traces of nitrite and nitrate in water by solid phase spectrophotometry. Analytica Chimica Acta 410, 185–192. doi:10.1016/s0003-2670(00)00736-4.

Almasri, M.N., Kaluarachchi, J.J., 2007. Modeling nitrate contamination of groundwater in agricultural watersheds. Journal of Hydrology 343, 211–229. doi:10.1016/j.jhydrol.2007.06.016.

Atamuradov, V., Medjaher, K., Dersin, P., Lamoureux, B., Zerhouni, N., 2017. Prognostics and health management for maintenance practitioners-review, implementation and tools evaluation. International Journal of Prognostics and Health Management 8, 1–31.

Balan, I., Kumar, P.M., Shivakumar, M., 2012. An assessment of groundwater quality using water quality index in chennai, tamil nadu, india. Chronicles of Young Scientists 3, 146. doi:10.4103/2229-5186.98688.

Balderas, E.C.S., Berti-Equille, L., Hernandez, M.A.A., Grac, C., 2017. Principled data preprocessing: Application to biological aquatic indicators of water pollution, in: 2017 28th International Workshop on Database and Expert Systems Applications (DEXA), IEEE. doi:10.1109/dexa.2017.27.

Benkedjouh, T., Medjaher, K., Zerhouni, N., Rechak, S., 2013. Remaining useful life estimation based on nonlinear feature reduction and support vector regression. Engineering Applications of Artificial Intelligence 26, 1751–1760. doi:10.1016/j.engappai.2013.02.006.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms (Advanced Applications in Pattern Recognition). Plenum Press.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32. doi:10.1023/a:1010933404324.

van Buuren, S., 2018. Flexible Imputation of Missing Data. Chapman and Hall/CRC.

Castillo, E., Corrales, D.C., Lasso, E., Ledezma, A., Corrales, J.C., 2016. Data processing for a water quality detection system on colombian rio piedras basin, in: Computational Science and Its Applications – ICCSA 2016. Springer International Publishing, pp. 665–683. doi:10.1007/978-3-319-42089-9_47.

Cude, C.G., 2001. Oregon water quality index a tool for evaluating water quality management effectiveness 1. Journal of the American Water Resources Association 37, 125–137. doi:10.1111/j.1752-1688.2001.tb05480.x.

Denuault, G., 2009. Electrochemical techniques and sensors for ocean research. Ocean Science Discussions 6, 1857–1893. doi:10.5194/osd-6-1857-2009.

Dhaenens, C., Jourdan, L., 2019. Metaheuristics for data mining. survey and opportunities for big data. 4OR 17, 115–139. doi:10.1007/s10288-019-00402-4.

Dogo, E.M., Nwulu, N.I., Twala, B., Aigbavboa, C., 2019. A survey of machine learning methods applied to anomaly detection on drinking-water quality data. Urban Water Journal 16, 235–248. doi:10.1080/1573062x.2019.1637002.

Dunn, J.C., 1973. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. Journal of Cybernetics 3, 32–57. doi:10.1080/01969727308546046.

Eliades, D.G., Stavrou, D., Vrachimis, S.G., Panayiotou, C.G., Polycarpou, M.M., 2015. Contamination event detection using multi-level thresholds. Procedia Engineering 119, 1429–1438. doi:10.1016/j.proeng.2015.08.1003.

Frank, P.M., Kppen-Seliger, B., 1997. New developments using AI in fault diagnosis. Engineering Applications of Artificial Intelligence 10, 3–14. doi:10.1016/s0952-1976(96)00072-3.

Goldberg, D.E., Holland, J.H., 1988. Genetic algorithms and machine learning. Machine learning 3, 95–99. doi:10.1023/A:1022602019183.

Gouriveau, R., Medjaher, K., Zerhouni, N., 2016. From Prognostics and Health Systems Management to Predictive Maintenance 1: Monitoring and Prognostics. ISTE - Wiley.

Lamb, J.D., Simpson, D., Jensen, B.D., Gardner, J.S., Peterson, Q.P., 2006. Determination of perchlorate in drinking water by ion chromatography using macrocycle-based concentration and separation methods. Journal of Chromatography A 1118, 100–105. doi:10.1016/j.chroma.2006.01.138.

Li, Z., Deen, M., Kumar, S., Selvaganapathy, P., 2014. Raman spectroscopy for in-line water quality monitoring—instrumentation and potential. Sensors 14, 17275–17303. doi:10.3390/s140917275.

Li, Z., Wang, J., Li, D., 2015. Applications of raman spectroscopy in detection of water quality. Applied Spectroscopy Reviews 51, 333–357. doi:10.1080/05704928.2015.1131711.

Lumb, A., Halliwell, D., Sharma, T., 2006. Application of CCME water quality index to monitor water quality: A case study of the mackenzie river basin, canada. Environmental Monitoring and Assessment 113, 411–429. doi:10.1007/s10661-005-9092-6.

Machiwal, D., Jha, M.K., 2015. Identifying sources of groundwater contamination in a hard-rock aquifer system using multivariate statistical analyses and GIS-based geostatistical modeling techniques. Journal of Hydrology: Regional Studies 4, 80–110. doi:10.1016/j.ejrh.2014.11.005.

Martinez, A., Kak, A., 2001. PCA versus LDA. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 228–233. doi:10.1109/34.908974.

Mastrocicco, M., Giuseppe, D.D., Vincenzi, F., Colombani, N., Castaldelli, G., 2017. Chlorate origin and fate in shallow groundwater below agricultural land-

scapes. Environmental Pollution 231, 1453–1462. doi:10.1016/j.envpol.2017.09.007.

Ng, A., Perera, B., 2003. Selection of genetic algorithm operators for river water quality model calibration. Engineering Applications of Artificial Intelligence 16, 529–541. doi:10.1016/j.engappai.2003.09.001.

Ratolojanahary, R., Ngouna, R.H., Medjaher, K., Junca-Bourié, J., Dauriac, F., Sebilo, M., 2019. Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset. Expert Systems with Applications 131, 299–307. doi:10.1016/j.eswa.2019.04.049.

Sen, S.M., Bricka, S.G., 2009. Data collection technologies past, present, and future.

Simmonds, J., Gómez, J.A., Ledezma, A., 2017. Knowledge inference from a small water quality dataset with multivariate statistics and data-mining, in: Advances in Intelligent Systems and Computing. Springer International Publishing, pp. 1–15. doi:10.1007/978-3-319-70187-5_1.

Tebbutt, T.H.Y., 1997. Principles of water quality control. Elsevier. doi:10.1016/B978-0-7506-3658-2.X5000-9.

Tyagi, S., Singh, P., 2013. Water quality assessment in terms of water quality index. American Journal of Water Resources 1, 34–38.

Vasanthavigar, M., Srinivasamoorthy, K., Vijayaragavan, K., Ganthi, R.R., Chidambaram, S., Anandhan, P., Manivannan, R., Vasudevan, S., 2010. Application of water quality index for groundwater quality assessment: Thirumanimuttar sub-basin, tamilnadu, india. Environmental Monitoring and Assessment 171, 595–609. doi:10.1007/s10661-009-1302-1.

Zhao, Y., Zheng, B., Jia, H., Chen, Z., 2019. Determination sources of nitrates into the three gorges reservoir using nitrogen and oxygen isotopes. Science of The Total Environment 687, 128–136. doi:10.1016/j.scitotenv.2019.06.073.

Zhu, R., Zeng, D., Kosorok, M.R., 2015. Reinforcement learning trees. Journal of the American Statistical Association 110, 1770–1784. doi:10.1080/01621459.2015.1036994.

LGP

LABORATOIRE
GÉNIE DE PRODUCTION

Concevons l'avenir
Design the future

Raymond HOUÉ NGOUNA

Maître de Conférences
Associate Professor
Équipe SDC - SDC team

+33 (0)5 62 44 27 45
Poste 2745

ÉCOLE NATIONALE D'INGÉNIEURS DE TARBES
47 AVENUE D'AZEREIX - BP 1629
65016 TARBES CEDEX - FRANCE

www.enit.fr