



**HAL**  
open science

## Predictive maintenance from event logs using wavelet-based features: an industrial application

Stéphane Bonnevey, Jairo Cugliari, Victoria Granger

► **To cite this version:**

Stéphane Bonnevey, Jairo Cugliari, Victoria Granger. Predictive maintenance from event logs using wavelet-based features: an industrial application. 2018. hal-01856309

**HAL Id: hal-01856309**

**<https://hal.science/hal-01856309>**

Preprint submitted on 10 Aug 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Predictive maintenance from event logs using wavelet-based features: an industrial application

Stéphane Bonnevoy<sup>1</sup>

Jairo Cugliari<sup>1</sup>

Victoria Granger<sup>2</sup>

<sup>1</sup>ERIC EA3083, Université de Lyon, 5 av. Pierre Mendès France, 69676 Bron Cedex, France

<sup>2</sup> ENEDIS, 124 boulevard Marius Vivier Merle, Lyon, France

August 10, 2018

## Abstract

In industrial context, event logging is a widely accepted concept supported by most applications, services, network devices, and other IT systems. Event logs usually provide important information about security incidents, system faults or performance issues. In this way, the analysis of data from event logs is essential to extract key informations in order to highlight features and patterns to understand and identify reasons of failures or faults. The objective is to help anticipate equipment failures to allow for advance scheduling of corrective maintenance. In this paper, we address the problem of fault detection from event logs in the electrical industry. We propose a supervised approach to predict faults from an event log data using wavelets features as input of a random forest which is an ensemble learning method. This work was carried out in collaboration with ENEDIS, the distribution operator of the electrical system in France.

## 1 Introduction

Smart electric devices automatically monitor information about energy consumption or production, they are defined by the ability to connect to a network and to operate remotely. They report meaningful and appropriate information to relevant parties (consumers, energy distribution system operators or energy providers) and their systems. Modern electric smart devices produce enormous amount of data. The first one is the inherently primary data associated to the devices' main activity and implemented features. Its exploration and use involve privacy issues which have been largely debated and are beyond the scope of this work. In addition to this, the second category of transmitted information is about events, a relatively new category of data, the value of which has yet not been assessed. Event is basically a notification that originates from a electrical device and contains the information regarding the object, action or process to which the event is related. Events are issued while monitoring different aspects of the system and give an overview about equipment communications, devices' secondary non-core functionalities, network intrusions or activity on the grid.

We believe that event logs could be processed and analysed to unveil useful information, in addition to devices' primary data. More precisely, we assume that these data can be useful to inform about the device's operative state and eventually to predict device failure. However, event logs concern a wide range of uses and the difficulty comes from the volume and variety of logs received. Log events are continuously recorded composing a data streamflow related with high volumes, as being generated not only for irregular functional conditions, but also for normal operative states. The main challenge is to analyse this data and extract useful knowledge from the unremitting flow of notifications. The issue therefore is to identify appropriate events containing helpful information. Furthermore, it is essential to detect a shift or an alteration in the patterns of these specific events which could alert users about a fault occurrence.

In literature, patterns from event logs are defined in various ways, for example as partial orders of a process<sup>1,2,3</sup>, or considered as Petri nets<sup>4</sup>. Also as repeated sequences that capture process models from event logs in order to improve their detection<sup>5</sup>. From these definitions, authors develop some specific pattern detection approaches mainly based on unsupervised or supervised learning

29 techniques. Unsupervised pattern detection approaches take an event log as input and generate  
 30 patterns based on statistical properties<sup>2,3,5</sup>. In unsupervised learning, clustering techniques are  
 31 widely used<sup>6,7</sup>. Supervised pattern detection approaches take patterns and logs as input and detect  
 32 pattern instances as results<sup>4</sup>. Combination of these two approaches into semi-supervised techniques  
 33 have been also studied<sup>8</sup>. From another point of view, visualization and interactive tools have been  
 34 developed to help user observe and analyse both patterns and event sequences, as EventFlow<sup>9</sup>.  
 35 Event logs are frequently composed of event codes and their associated text messages. In that  
 36 case, the use of text parsing or natural language processing techniques is necessary<sup>6,10</sup>.

37 Moreover, some specific works dealing with predictive maintenance based on event logs have  
 38 also been tackled. Let us mention a general classification-based failure prediction method which  
 39 has been tested on real ATM run-time event logs data<sup>11</sup>, or event logs data extracted from medical  
 40 equipments used to treat a multi-instance learning task<sup>12</sup>. Also, a Cox proportional hazard model  
 41 has been used to provide a prediction of system failures based on the time-to-failure data extracted  
 42 from the event sequences<sup>13</sup>.

43 In this work, we consider the event distribution over time as a function of time. Our first  
 44 objective is to extract characteristic features from the time series, which will then be presented to  
 45 a learning algorithm. In order to make this step as automatic as possible, we decided to perform the  
 46 Discrete Wavelet Transform (DWT) which is an appropriate tool for noise filtering, data reduction,  
 47 and singularity detection, and thus it a good choice for time series and signal processing. The  
 48 decomposition coefficients obtained from the DWT are then used as input of a supervised learning  
 49 algorithm. A variety of task can be successfully tackled using this approach<sup>14,15,16,17</sup>. In our case,  
 50 we use a random forest both to predict and to measure variable importance in order to select the  
 51 best features.

52 In this paper, we propose a supervised approach to predict faults from event log data using  
 53 wavelets features. The goal is first to use the Discrete Wavelet Transform to detect and charac-  
 54 terize features of our electric event logs. Then, we use these features as an input of a random  
 55 forest model to predict faults. Next section introduces the information we use from event logs and  
 56 how we transform them into time series trajectories or time functions. To cope with the temporal  
 57 dependence and functional structure of these objects, we introduce in Section 3 the wavelets trans-  
 58 form. The section also includes an overview of random forest. Section 4 describes the experiences  
 59 and presents the results. The work concludes with a discussion on both industrial and modelling  
 60 aspects in Section 5.

## 61 2 From event to time functions

62 Our study is based on events monitored on electrical devices installed on ENEDIS network, the  
 63 French Distribution System Operator. Each electrical device records and transmits real-time event  
 64 data to a centralized information system. We extract and deal with 3 available attributes: the  
 65 event code related to a time-stamp and the id of the source device. An example of our logs is  
 66 displayed in Table 1. We define an additional feature, a group code representing an hierarchical  
 67 level of event codes. These values were agreed upon with domain experts into 13 groups. None of  
 68 these notifications have any level of criticality or priority.

timestamp	deviceId	eventCode	groupCode
2014-01-24 17:49:44.537	001	A3	A
2014-01-24 15:09:35.970	001	A23	A
2014-01-25 03:55:56.872	002	A3	A
2014-01-27 00:14:42.463	002	B8	B
2014-01-27 08:10:25.470	002	A23	A
...	...	...	...

Table 1: Event logs data.

69 Events were aggregated to a daily basis. Figure 1 shows the A23 event distributions for two  
 70 electrical devices from the beginning of the study to the last day of observations. The observation  
 71 period starts on May 01 2013 and stops on November 02 2014 and the total number of events is  
 72 about 1.25 millions recorded on 2623 devices.

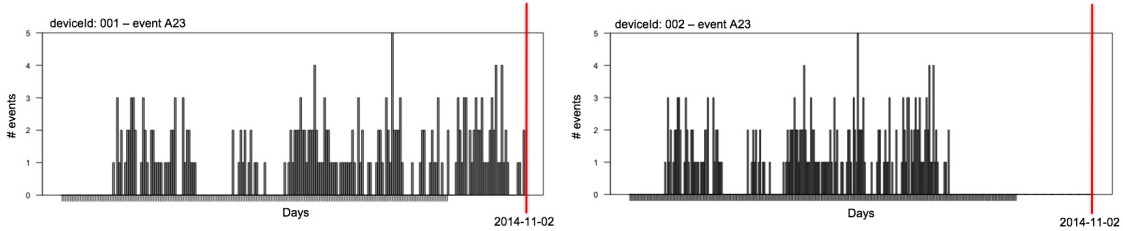


Figure 1: Example of A23 daily event distributions for two electrical devices (`dev001` and `dev002`) from 2013-05-01 to 2014-11-02. The red line shows the end of the observation period.

73 Devices were monitored over a considerable period of time and presented similar settings and  
 74 technical specifications during this period. The devices main activity is monitored throughout  
 75 their lifespan. A fault occurrence is considered when the device fails to provide its main function.

76 Among all, 1858 devices were properly functional and had a functional status over the obser-  
 77 vation period, providing their primary function. All of the device were brought to service *a priori*  
 78 to the beginning of observation period (see `dev003`, Figure 2) and were selected as being operative  
 79 *a posteriori* to the observation period, over a significant interval of time, to ensure their normal  
 80 functioning.

81 A part of devices developed a fault before the end of the observation period, with the lost of  
 82 their primary functions. These devices were brought to service either before or after the beginning  
 83 of the observation time. The devices were withdrawn from the field and a technical diagnosis  
 84 confirmed failure on these equipments. Devices failing to provide their main activity for which  
 85 technical diagnosis did not confirm the failure were not considered in the study. 765 faulty devices  
 86 were considered in this study (see `dev002` or `dev001`, Figure 2).

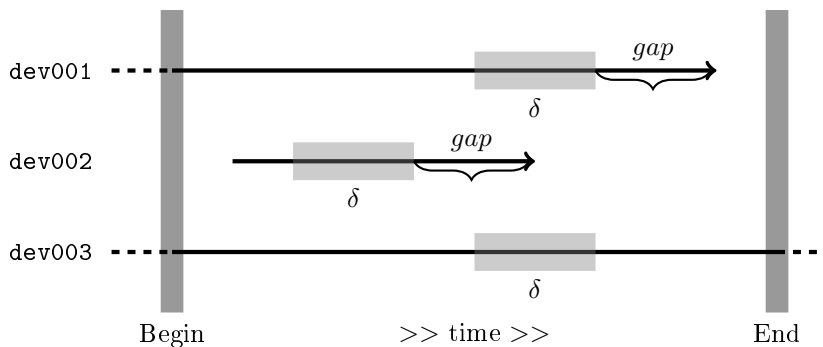


Figure 2: Examples of 3 devices throughout the observation period. Device `dev001` and `dev002` present a fault occurrence. *gap* represents the number of days in advance the fault occurrence is recorded, and therefore predicted. Device `dev003` does not present any fault occurrence.  $\delta$  is the temporal period on which wavelet features decomposition is applied.

87 As stated above, both devices with positive and negative fault occurrences present event profiles  
 88 throughout their lifespan. The purpose of this study is to compare temporal event profile of faulty  
 89 and working devices in order to identify useful events for predictive maintenance. Hence, the first  
 90 goal is to capture events frequencies and dynamics of both devices' health status. The second  
 91 goal is to predict fault occurrence using these summarized temporal profiles while identifying  
 92 meaningful events. In operative conditions, we wish to detect a failure with a delay which needs  
 93 to be sufficiently long to allow the attendance of alarms on devices. In this study we considered a  
 94 predictive gap ranging from zero up to 15 days.

95 From the available data we only use the absolute frequency of events and the event code  
 96 classification. We focus then on the number of logs effectively observed over a reasonable period of  
 97 length  $\delta$  using a given time resolution (e.g. hours, days) for each type of event. We then consider

$$(N(t_1), \dots, N(t_\delta)), \quad (1)$$

98 where  $N(t_j)$  is the number of events at time  $t_j$ . To fix ideas, say that  $\delta$  may span over two weeks

99 and using a daily resolution the vector of counts would have length equal to 14. Figure 3 plots  
 100 four cases of these trajectories.

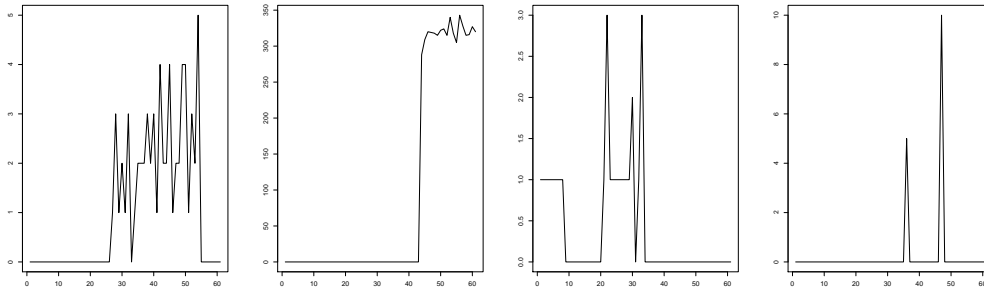


Figure 3: Examples of trajectories from event logs data. Tracking is done daily over 64 days. Cases (a) and (b) are from working devices; and cases (c) and (d) are from faulty ones.

101 This vector constitutes the building block of our approach since we create instances of this  
 102 vector for both normal and abnormal regimes (*cf.* Section 3). Actually, each instance is the  
 103 tracking of a device along a period of length  $\delta$ . Then the couple device  $\times$  time should not be view  
 104 as tracking over contemporaneous instants but as snapshots of the life time of the devices.

105 If we now consider that  $K$  different type of events exists, then we have

$$(N_k(t_1), \dots, N_k(t_\delta)), k = 1, \dots, K,$$

106 that is each device  $\times$  time is a set of  $K$  counting vectors. From the mathematical point of view,  
 107 we may look at vectors (1) as time series trajectories. And since there exists  $K$  of them, we have  
 108 a multivariate time series where there may be some dependence structure between components of  
 109 the vectors as well as time dependence within components.

110 One way to cope with time dependence is to see each trajectory as a discrete sampling, even-  
 111 tually with some noise, of the time function  $z_k(t), t \in [0, \delta], k = 1, \dots, K$ . Notice again that time  
 112 should be consider as relative to period  $\delta$  and not as an absolute quantity.

113 To fix ideas, let us introduce a graphical representations of the some of the data up to this point.  
 114 We follow the construction detailed above considering only events from one code to construct a  
 115 sample of trajectories containing both faulty and working devices. Then, we use a simple metric  
 116 between trajectories based on the euclidean distance on standardized versions of the trajectories.  
 117 The associated distance matrix is then used as input of a multidimensional scaling in order to get  
 118 the a simple planar representation of the observations, represented in Figure 4.

119 Here, each point is a trajectory and its coordinates are chosen to preserve, as well as possible,  
 120 the distances between trajectories. Notice that since no information about the class is used this  
 121 technique is essentially unsupervised. However we add a colour reference (grey: working devices,  
 122 red: faulty devices) to the scatter plot in order to visualize eventual differences. Even if the sample  
 123 is very unbalanced, a clear distinction between both classes is appreciated. Distances between  
 124 working devices are relatively small with reference to distances between faulty devices. Other  
 125 conclusion we can draw is that a (eventually non linear) reduction of the dimension may suffice to  
 126 extract the useful information on the signals. Taking into account the time dependent structure  
 127 of the functions is necessary to obtain an appropriate construction that yields on a dimension  
 128 reduction.

### 129 3 Methods

130 We describe here the methods we use to construct our solution. They are related first on how to  
 131 represent the multivariate time series with a handy set of interpretable features. Then we build  
 132 a 2-class discriminant classifier where we assume that each class represents a logging regime. The  
 133 first one is the normal log regime where the working device has a working status. This regime,  
 134 should be the most frequent one. The second regime, more rare by nature, represent a working  
 135 device that is close to a failure status. The first problem is solved using the wavelet transform  
 136 while the classifier we use is random forest.

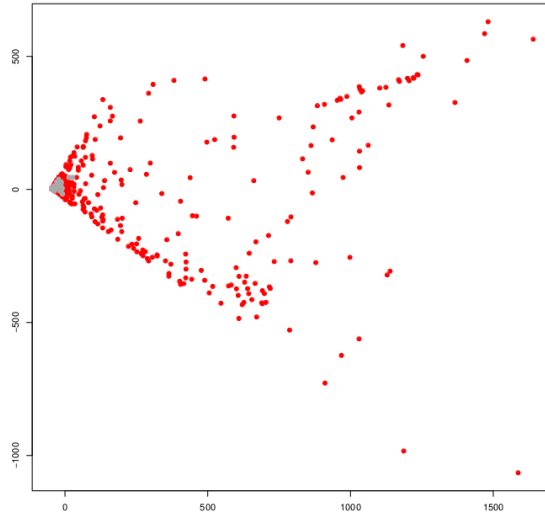


Figure 4: Multi dimensional scaling of trajectories from one event code. Each point represents a trajectory from a working device (in gray) or a faulty device (in red).

### 137 3.1 Wavelets transform

138 Wavelets are a domain transform technique that allows one to represent time domain signals into  
 139 a bivariate domain location-scale<sup>18</sup>. While location in the new domain is connected to the original  
 140 time domain, scales can be associated to Fourier frequencies and both with good localization  
 141 properties. That is, the transform will give information on locations connected to only a time span  
 142 (not the global time) and scales connected to only some frequencies (and not all of them). This  
 143 is in difference with a time domain analysis that has no localization on frequencies or a frequency  
 144 domain analysis that has no localization on time.

145 Moreover we use the Discrete Wavelet Transform (DWT) which provides an orthonormal  
 146 basis of the space, allowing us to encode all the available information on a signal without any loss  
 147 of information<sup>19</sup>. In what follow we explain the necessary material to understand our approach.

148 Consider the signal  $z(t)$  which is an univariate function defined on the time domain  $\mathcal{T}$ , for  
 149 example  $\mathcal{T} = [0, 1]$ . The DWT will provide two terms: a global approximation of the signal  $\mathcal{S}(t)$   
 150 and the ensemble of details  $\mathcal{D}(t)$  well localized both in time and frequency. If  $z \in L_2([0, 1])$ ,  
 151 then the DWT provides us with a basis of the functional space. The basis is created by simple  
 152 transformations of a scaling function  $\phi(t)$  and a wavelet mother  $\psi(t)$  which are associated to  
 153 the orthogonal multi resolution analysis of  $L_2([0, 1])$ . Indeed, we consider the family  $\{\phi_{j,k}(x) =$   
 154  $2^{-j}\phi(2^{-j}x - k)\}_{j,k}$  which is obtained by dilatations of a factor  $2^j$  and by integer translations on  
 155 the new scale. Similar operations are done to get the family  $\{\psi_{j,k}(x)\}_{j,k}$ . Then, a finite energy  
 156 signal  $z$  can be expressed as

$$z(t) = \underbrace{\sum_{k=0}^{2^{j_0}-1} c_{j_0,k} \phi_{j_0,k}(t)}_{\mathcal{S}_{j_0}(t)} + \underbrace{\sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} d_{j,k} \psi_{j,k}(t)}_{\mathcal{D}(t)}, \quad (2)$$

157 where  $c_{j,k} = \langle z, \phi_{j,k} \rangle$ ,  $d_{j,k} = \langle z, \psi_{j,k} \rangle$  are the scale coefficients and wavelet coefficients re-  
 158 spectively. The scale  $j_0$  separates the two terms. The first one, gives a smooth approximation  
 159 at resolution  $2^{j_0}$ . The second one, keeps all the details of the curves on a hierarchical structure  
 160 depending on scales and locations. The approximation coefficients  $c_{j_0,k}$  retains the information  
 161 of the local (at location  $k$ ) mean level of the curve, while the detail coefficients  $d_{j,k}$  codes the  
 162 information of discontinuities and other singularities.

163 With finite data  $\{z(t_i), i = 1, \dots, N\}$ , the signal  $z(t)$  can only be approximated by a truncation  
 164 at some maximum scale level  $J = \log_2(N)$ , that is we approximate (2) by

$$z_J(t) = c_0 \phi_{0,0}(t) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} d_{j,k} \psi_{j,k}(t). \quad (3)$$

165 Notice that we have also fixed the approximation part at the coarsest resolution  $j_0 = 0$  which  
 166 means that only one scaling coefficient is used to approximate this term. For convenience we  
 167 choose the number of sampling points per curve,  $N$ , to be a power of 2. The maximum number  
 168 of scales  $J$  is then an integer. With this, we are in conditions to use the highly efficient Mallat's  
 169 pyramidal algorithm<sup>18</sup> to obtain both the scaling and wavelet coefficients. If the sampling grid  
 170  $\{i/N, i = 1, \dots, N\}$  is not regular or  $N$  is not a power of 2, then one can choose a finer regular  
 171 grid and use any interpolation scheme to meet our choices.

172 Haar wavelet leads to a easy and clear intuition on the wavelets coefficients. The only scaling  
 173 coefficients we retain is proportional to mean level of the whole signal. The approximation term is  
 174 then a constant function  $S_0(t) = c_{0,0}\psi(t)$  proportional to the mean function of the signal.

175 If we increase the resolution of the approximation to the next scale, then the approximation  
 176 part will be a ladder function, that is a piecewise constant function with a jump in the middle point  
 177 of the sampling grid. Aside the jump, the signal is approximated by the mean level of each side.  
 178 A similar reasoning applies to the next scales, at each time cutting into halves and approximating  
 179 each half by a constant function equal to the mean level of the observations on the half.

180 The detail coefficients are the difference on the constant approximations between two juxtaposed  
 181 halves. We interpret them as the change observed at some resolution (related to the scale  $j$ ) and  
 182 at some time (related to the location  $k$ ).

183 In what follows we will need to reduce the number of coefficients we use in order to keep the  
 184 calculations into a reasonable time. With this, we are further truncating the approximation on  
 185 (3) into smaller values of  $J$ . Since finer approximations may capture only the signal's noise, the  
 186 changes on these scales would reflect random fluctuations not necessarily connected to the structure  
 187 of the signal. For this, one should only retain coarsest scales. and the detail coefficients  $d_{0,0}, d_{0,1}$ .

188 In what follows we set  $\psi$  to be

$$\psi(t) = \begin{cases} 1 & \text{if } 0 \leq x < 1/2 \\ -1 & \text{if } 1/2 \leq x < 1 \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

189 which is known as the Haar wavelet. The corresponding scaling function is  $\psi(x) = 1$  if  $0 \leq x \leq 1$   
 190 and 0 otherwise.

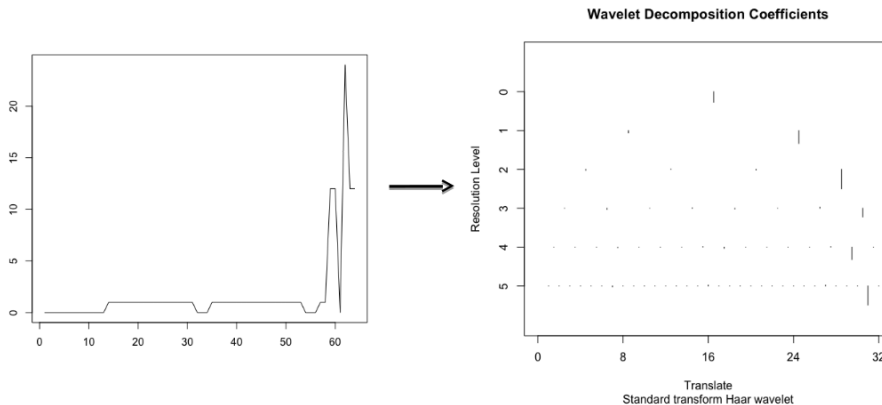


Figure 5: Tracking of one event over time (right) and its DWT (left).

191 To illustrate the kind of results we obtain with the wavelet transform, we show in right hand  
 192 side of Figure 5 the detail coefficients of the signal represented on the right hand side. The signal  
 193 shows a low-activity steady state during almost all the tracking period with an important raise  
 194 on the number of events at the end. The wavelets coefficients (on the left of the figure) show an  
 195 alternative picture of the same phenomena. Ranged by scales, the coefficients are small in absolute  
 196 value almost everywhere but at those location near the end of period. What is near, depends on  
 197 the resolution level at which we look at the signal, for scales close to  $j = 0$  the analysing functions  
 198 are global, while at scale  $j = 5$ , the 32 resulting coefficients gives very localized information. Note  
 199 that this level of detail can be misleading if considered only at one individual scale. For instance  
 200 the last coefficient at scale 5 is large and negative because the last number of logs is lower than the  
 201 precedent one. Moreover, noise is also more important at these high frequencies. One may rely

202 on shrinkage methods to choose which of the estimated coefficients are significantly different from  
203 zero.

204 Our approach is slight different, we choose to work only with the scale coefficient  $c_{0,0}$  and  
205 the detail coefficients  $d_{0,0}, d_{0,1}$ . With this, the number of coefficients retained in what follows  
206 is kept into a reasonable size when multiplied by the number of event codes. Intuitively, these  
207 coefficients allows one to reconstruct the trajectories with the approximation  $\mathcal{S}_1(t)$  which is exactly  
208 the mean level of the function, given by  $c_{0,0}$ , and the detail coefficients  $d_{0,0}, d_{0,1}$ . Notice that this  
209 reconstruction is the best linear approximation one can do with three coefficients. In what follows  
210 we are using the estimated coefficients as features of a random forest predictor.

## 211 3.2 Random forest

212 Very popular in statistical machine learning, random forests (RF) are an ensemble method<sup>20</sup>.  
213 It builds up on specific versions of CART (Classification And Regression Trees)<sup>21</sup>, which is an  
214 algorithm that constructs binary tree-based predictors. With respect to individual predictors, the  
215 aggregate one aims to augment robustness, variance reduction and improve prediction performance.

216 For this, RF add two layers of randomness. First, each tree-based predictors is trained only on  
217 a different bootstrap sample from the data. Second, only a strictly subset of variables are randomly  
218 chosen as candidates at each split of the trees' construction. Note that the trees are constructed  
219 up to its maximal size and they are not pruned. While using a stopping criterion and pruning  
220 are usual in CART, these versions of tree-predictors sacrifices generalization power by a better in-  
221 sample fit – at least on each bootstrapped sample – and introduces bias by considering only partial  
222 information from available variables. With this, individual trees tend to be less dependent between  
223 them which is useful under an aggregation scheme. RF is then the resulting predictor obtained by  
224 some aggregation rule of the individual prediction of the so described trees. Usual choices of the  
225 aggregation rule are majority vote for classification and mean average for regression.

226  
227 We use two intrinsic features of random forest to help the interpretation of the results : a  
228 measure of variable importance and a notion of proximity between observations.

229 **Variable importance measure.** Different approaches can be used to determine the importance  
230 of a feature for the construction of the forest<sup>22,23</sup>. In this work, a variable is considered more  
231 important if it participates more to the decrease of some impurity notion (e.g. the Gini index).  
232 Then, we can track over the individual trees where each variable participates on each node split  
233 and record the decrease on the Gini coefficient. Then a plot like the one in Figure 7 where the  
234 variables are represented in lines sorted decreasingly on the mean Gini reduction. Most important  
235 variables on the construction of the classifier are on top of the plot.

236 **Observations proximity** Two observations are closer if they are classified within the same  
237 terminal node by more and more individuals trees. Then, the proximity is normalized to be between  
238 0 and 1. If we call  $p_{ii'}$  the proximity between observation  $i$  and  $i'$ , then we obtain a dissimilarity  
239 measure  $1 - p_{ii'}$ . While the change is trivial, it allows us now to perform a multidimensional  
240 scaling on the proximity matrix associated to the proximity measurements. This yields on a  
241 natural representation of the individuals that analogously to discriminant analysis represents in a  
242 low dimensional space how the classifier 'sees' the individuals.

## 243 4 Experiments

244 In all our experiments we use the open source R software. DWT is performed using `wavethresh`  
245 and `randomForest` is used to learn random forests. In the experiments to follow we use the default  
246 options of `randomForest` to construct the predictors, *i.e.* the number of trees `ntree` is set at 500  
247 and the number of variables `mtry` chosen randomly at each split is roughly the square root of the  
248 total number of variables.

249 For each gap before fault, we create a dataset of positive and negative failure occurrences as  
250 shown in 2). From faulty devices,  $\delta$  time points (albeit the gap period) before fault occurrence  
251 were considered to compute  $K$  event vectors. Among working devices, a period of length  $\delta$  is  
252 drawn randomly per device to compute  $K$  event vectors per device. Notice that each device at



253 some point of the time is described as a number of 39 features, that is 3 wavelets coefficients per  
 254 group of events' code, with a total of 13 event codes.

## 255 4.1 Predictive performance

256 We apply a random forest classifier for each of 16 datasets composed of 39 wavelets coefficients.  
 257 Two week event profile (for each device) is characterized by 3 coefficients for each of the 13 groups  
 258 of events. We compute both false negative rate (FNR) representing the percentage of faulty devices  
 259 classified as working devices and false positive rate (FPR) as the percentage of working devices  
 260 predicted as prone to failure. We also compute the global model error, summarizing the percentage  
 261 of observations which are classified wrong and resuming model global accuracy.

262 The performance scores of random forest models are displayed in Figure 6, results are presented  
 263 in relation to the predictive gap before failure occurrence. Global model accuracy ranges from 79%,  
 264 when the predictive gap equals to 15 days, up to 89% when detecting fault the day of occurrence.  
 265 This performance, that at first sight appears rather inaccurate in an industrial context, displays  
 266 evidence of meaningful information in the event logs.

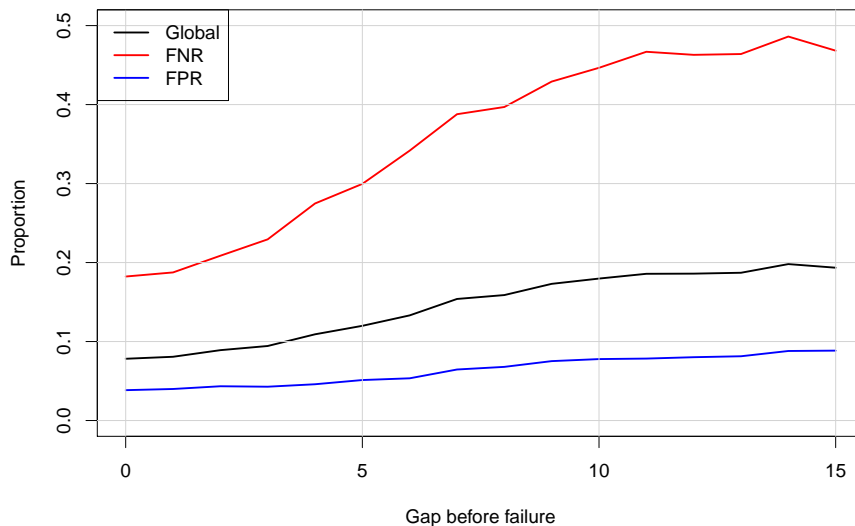


Figure 6: Predictions performance at various gaps, in days, to failure. Red curve shows classification results of failed devices, blue curve shows classification results of working devices and black curve global error.

267 Overall, the predicting error rate is higher for faulty devices and it is easier to decide on a  
 268 working status on the basis of resumed event profile of a device, independently of the prediction  
 269 gap (see Figure 6, the red curve corresponding to the FNR stays above the blue line representing  
 270 the FPR). The error rate is lower when classifying working devices, as observations of negative  
 271 fault occurrence dominate the learning error. The result is consistent with the fact that random  
 272 forests tend to maximize the model global accuracy, keeping a low error rate on larger classes  
 273 (working devices) while allowing the smaller classes have a larger error rate.

274 The smaller the temporal gap is, the more precise it is to predict both fault occurrence or  
 275 devices' normal regime by event data, for example the FPR for working regimes being equal to  
 276 3.74% for 1 day-ahead prediction and 7.15% for a 10-day-ahead horizon, see Figure 6.

## 277 4.2 Variable importance

278 Figure 7 shows importance ranking of attributes in classification for a 0 days predictive gap, variable  
 279 importance displays similar results for all of the 16 models for different values of the predictive gap  
 280 (results not shown). 3 groups of events appear relevant when predicting fault occurrence : A, B and  
 281 J. First and third wavelet components of B group appear to be the red flag for an abnormal regime

282 leading to a failure. A different level of these events for a device and an alteration of the number of  
 283 received events can be seen as an alarm for failure occurrence. Overall, we observe the same pattern  
 284 for all of three groups of events: the faulty devices' average level of events is generally higher than  
 285 working devices' event frequencies. Moreover, there is a substantial gap between the event regime  
 286 7 days before failure occurrence and the week before that. This is particularly interesting, as events  
 287 are related to low level communication on the grid. We suppose that failure affects the ability of  
 288 devices to interact with other devices on the network.

289 More helpful, Figure 7 shows that a considerable amount of information received and processed  
 290 by the system are not relevant for revealing devices' operative status. Independently of their  
 291 number or frequency, events of group C, G, K, D and F, seem to carry very little information about a  
 292 possible failure of an equipment. This is to be expected as these events monitor different software  
 293 activity of various devices of the grid. In a predictive maintenance framework, the monitoring  
 294 and processing of these categories presents no interest, events have no correlation with the fault  
 295 occurrence.

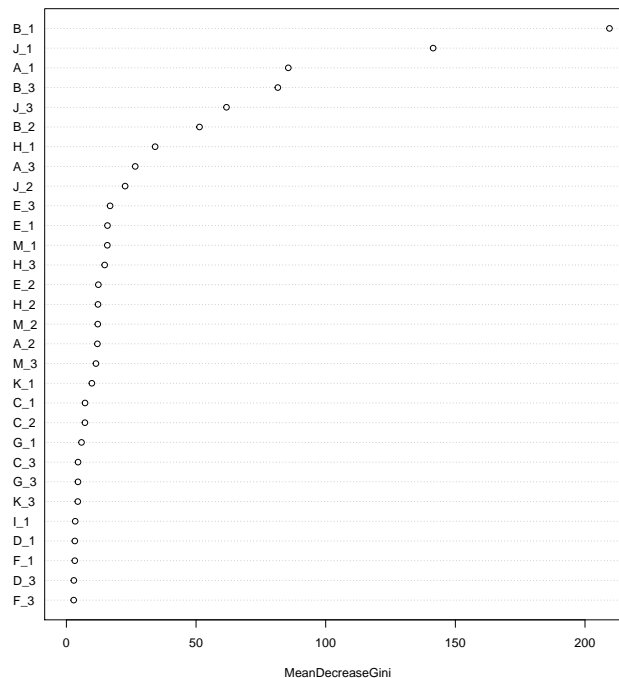


Figure 7: Random forest variable importance output for a 0 day temporal gap

### 296 4.3 Observations proximity

297 As in Figure 4, we use a multidimensional scaling to represent observation proximity in Figure 8.  
 298 Recall that now the distances on plot are the ones implicitly learned by the classifier so it is  
 299 effectively using the information on the labels (coded in colours on the plot). Two important  
 300 differences are to be highlighted. First, the classes are now better clearly separated even if with  
 301 some overlap. The class of faulty devices (smaller, in red) forms now a compact group that aligned  
 302 along a straight pattern. Second, the class of normal operation, that is without fault, presents  
 303 a two arm structure. This means that while connected by some elements that are close to both  
 304 arms in the middle of the plan, the structure suggest that this class is actually formed by two  
 305 subclasses which are homogeneous for each of them. From a technical point of view, this result  
 306 also indicates that working devices present two distinctive event profiles, which shift to a single  
 307 highly abnormal regime when failure emerges. This outcome is of a particular interest, as experts  
 308 do not have any *a priori* knowledge about this singularity. Additional investigations may reveal  
 309 different manufacturer implementations or material configurations having no impact on devices  
 310 primary functionalities.

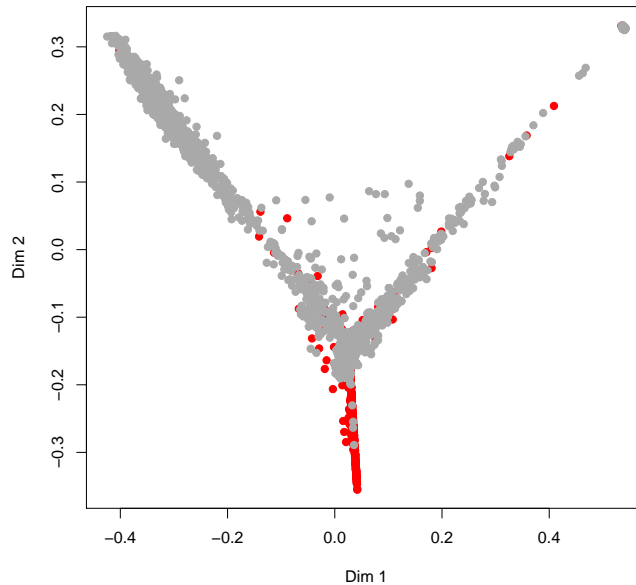


Figure 8: MDS from RF. Each point represents a trajectory from a working device (in gray) or a faulty device (in red).

#### 311 4.4 Additional experiments

312 Events of category B, J and A were the most relevant to describe device operative state and  
 313 to detect a device in a abnormal regime leading to a failure. Using exclusively events of these  
 314 three groups, we performed an additional set of experiments by including non grouped events as  
 315 features and applying same methodological approach as described above on individual events. We  
 316 treated 17 individual code events, therefore 51 wavelets attributes were computed and introduced  
 317 as features to random forests classifier. All other parameters remained unchanged. Performances  
 318 scores of 16 models based on 51 features perform similarly to the global approach (results are not  
 319 shown) which is using all grouped events.

320 Essentially, error rates of non grouped B, J and A events models are lower than error of models  
 321 using all code events.

### 322 5 Discussion and conclusion

323 From the modeling point of view, the use of wavelets and random forests gave several benefits.  
 324 First, the proposed approach is general in the sense that it is not specific for predictive mainte-  
 325 nance. Actually, it may be used on different kind of anomaly detection from event logs such as  
 326 intrusion detection, outage occurrence, *etc.*, as long as one disposes with a way to construct the two  
 327 class learning data set. Second, wavelets allows an important dimension reduction while keeping  
 328 discriminatory power. With this, up-scaling the procedure is feasible since the processing needed  
 329 to pass from functions to wavelet coefficients may be done independently (and so using parallel or  
 330 distributed computing schemes) for each device. Last, random forest gives interest insights through  
 331 feature selection and observations proximity. The former benefits from localized coefficients that  
 332 gives nice interpretation properties to the DWT. The latter can be used together with graphical  
 333 displays to make emerge patterns in data that are otherwise difficult to unveil.

334 A natural question that may arise is about the particular choice of the functional basis (i.e.  
 335 wavelets) and the classifier (i.e. random forest). One may naturally argue that other combination  
 336 like for instance principal components and logistic regression would be equally reasonable. Our  
 337 choices are guided for both interpretability and performance. Besides of having a low error level  
 338 classifier we look for tools that allow the practitioner to better understand the underlying problem.  
 339 The first three principal components would extract a mean behaviour of the curves where what we

340 look for is the specific behaviour of each frequency curve that best explains its evolution. Location  
341 properties of wavelets, discussed in Section 3, induce the nice interpretability we look for while  
342 compressing the information by a handy number of features. Random Forest also contribute to  
343 provide with insights on the fitted model as it was discussed in Section 4.2 and 4.3.

344 In our case, when processing log events we found evidence of untapped information on both  
345 fault occurrence and the normal devices' regime when monitoring electrical devices information  
346 flow. When predicting the fault occurrence the FNR ranges from 18.72% when confirming the fault  
347 by analysing event profile the day of fault occurrence, and goes up to 46.38% to 15 days predictive  
348 gap. In regard to these results, there a number of points that we should comment on.

349 First, the increase of FNR inversely proportional to temporal gap implies that, in some extent,  
350 at least two weeks before failure occurrence a part of the devices have a similar event profile as  
351 working devices and their event regime undergoes a gradual daily alteration until failure. The  
352 degradation seems to accelerate 7 days before failure occurrence, the FNR equals to 40.99% and  
353 we gain several points of precision each day. We suppose that failure firstly affects non essential  
354 functionalities, from which the event logs are issued, and only secondarily it leads to the cession  
355 of the main activity. This progressive shift underlines the fact that a fault occurrence does not  
356 necessarily imply a full and immediate standstill of a device as it continues to provide their primary  
357 function. In regard to these elements, the use of these notifications in a predictive maintenance  
358 tool is of a particular interest to track future devices fault.

359 Second, a part of faulty devices are misclassified and display similar log event profile as working  
360 devices until failure occurrence (FNR equals to 18.44% for a 1 day predictive gap). See also Figure 8,  
361 a part of red observations (faulty devices) are situated among grey observations (working devices).  
362 It is likely that the category of tracked failure is not similar to the previously described case, which  
363 affects primary and secondary functionalities differently. For these devices, functional features  
364 related to the main activity should probably be measured as log event profile do not change priori  
365 to the failure.

366 One last point of a particular interest in the results are the early signs of breakdown affecting  
367 the devices more than two weeks in advance (FNR equals to 46.24% 2 weeks priori to failure). This  
368 result is supported by the high variable importance of the first wavelet component for all three  
369 groups of events, even for a high prediction gap (see 4.2). A really moderate usury of hardware  
370 related to external factors or network overload could affect a part of equipment. To a different  
371 degree, we could also suppose that these devices show an abnormal event profile as soon as they  
372 are installed and a latent defect affects their non-core functionalities.

373 To sum up, classification results show that abnormal dynamics in specific events, can be con-  
374 sidered, to a certain extent, forerunner of a future fault. For a long term preventive strategy,  
375 there is an obvious need to cross the profiles of identified group of events with other sources of  
376 informations to increase model accuracy. Geographical situation the grid, power demand, voltage  
377 quality, or environmental factors could affect gradually devices leading potentially to a failure.  
378 Primary data and the monitoring of information resulting of the implemented features could allow  
379 to enhance the predictive capacity of events. Information on other components of the grid could  
380 offer complementary perspectives on the network activity leading to devices usury. Even if the  
381 predictive performance does not allow to develop an operative tool, this model allows to identify  
382 a high risk population to failure. In a supervision context, the daily processing of ongoing events  
383 could allow for these devices to be prioritized and then acted upon with necessary actions.

## 384 Acknowledgements

385 We would like to acknowledge the ENEDIS company for this collaboration and we especially thank  
386 Pierre Achaichia, Paul Mersy and Thomas Pilaud from ENEDIS for rich discussions.

## 387 References

- 388 [1] Sander J. J. Leemans, Dirk Fahland, and Wil M. P. van der Aalst. Discovering block-structured  
389 process models from event logs - a constructive approach. In *Application and Theory of Petri*  
390 *Nets and Concurrency*. Springer, Berlin, Heidelberg, 2013. doi: 10.1007/978-3-642-38697-8\_  
391 17.

- 392 [2] Sander J. J. Leemans and Wil M. P. van der Aalst. Discovery of frequent episodes in event  
393 logs. In *Data-Driven Process Discovery and Analysis*. Springer, Cham, 2015. doi: 10.1007/  
394 978-3-319-27243-6\_1.
- 395 [3] Claudia Diamantini, Laura Genga, and Domenico Potena. Behavioral process mining for  
396 unstructured processes. *J. Intell. Inf. Syst.*, 47(1):5–32, August 2016. ISSN 0925-9902. doi:  
397 10.1007/s10844-016-0394-7.
- 398 [4] F. Mannhardt, M. de Leoni, H. A. Reijers, Wil M. P. van der Aalst, and P. J. Toussaint. From  
399 low-level events to activities - a pattern-based approach. In *Business Process Management*.  
400 Springer, Cham, 2016. doi: 10.1007/978-3-319-45348-4\_8.
- 401 [5] R. P. Jagadeesh Chandra Bose and Wil M. Aalst. Abstractions in process mining: A tax-  
402 onomy of patterns. In *Proceedings of the 7th International Conference on Business Process*  
403 *Management, BPM '09*, pages 159–175. Springer-Verlag, 2009. ISBN 978-3-642-03847-1. doi:  
404 10.1007/978-3-642-03848-8\_12.
- 405 [6] Risto Vaarandi and Mauno Pihelgas. Logcluster - a data clustering and pattern mining al-  
406 gorithm for event logs. In *Proceedings of the 2015 11th International Conference on Net-*  
407 *work and Service Management (CNSM), CNSM '15*, pages 1–7, Washington, DC, USA, 2015.  
408 IEEE Computer Society. ISBN 978-3-9018-8277-7. doi: 10.1109/CNSM.2015.7367331. URL  
409 <http://dx.doi.org/10.1109/CNSM.2015.7367331>.
- 410 [7] Adetokunbo A.O. Makanju, A. Nur Zincir-Heywood, and Evangelos E. Milios. Clustering  
411 event logs using iterative partitioning. In *Proceedings of the 15th ACM SIGKDD International*  
412 *Conference on Knowledge Discovery and Data Mining, KDD '09*, pages 1255–1264, New York,  
413 NY, USA, 2009. ACM. ISBN 978-1-60558-495-9. doi: 10.1145/1557019.1557154. URL [http://](http://doi.acm.org/10.1145/1557019.1557154)  
414 [doi.acm.org/10.1145/1557019.1557154](http://doi.acm.org/10.1145/1557019.1557154).
- 415 [8] Xixi Lu, Dirk Fahland, Robert Andrews, Suriadi Suriadi, Moe T. Wynn, Arthur H.M. ter  
416 Hofstede, and Wil M.P. van der Aalst. Semi-supervised log pattern detection and exploration  
417 using event concurrence and contextual information. In *25th International Conference on*  
418 *Cooperative Information System*, Rhodes, Greece, 2017. Springer Verlag. URL [https://](https://eprints.qut.edu.au/110716/)  
419 [eprints.qut.edu.au/110716/](https://eprints.qut.edu.au/110716/).
- 420 [9] Megan Monroe, Rongjian Lan, Hanseung Lee, Catherine Plaisant, and Ben Shneiderman.  
421 Temporal event sequence simplification. *IEEE Transactions on Visualization and Computer*  
422 *Graphics*, 19(12):2227–2236, December 2013. ISSN 1077-2626. doi: 10.1109/TVCG.2013.200.
- 423 [10] Wei Xu, Ling Huang, Armando Fox, David Patterson, and Michael Jordan. Using machine  
424 learning techniques in console log analysis. In *the 27th International Conference on Machine*  
425 *Learning, ICML'10*, 2010.
- 426 [11] J. Wang, C. Li, S. Han, S. Sarkar, and X. Zhou. Predictive maintenance based on event-log  
427 analysis: A case study. *IBM Journal of Research and Development*, 61(1):11:121–11:132, Jan  
428 2017. ISSN 0018-8646. doi: 10.1147/JRD.2017.2648298.
- 429 [12] Ruben Sipos, Dmitriy Fradkin, Fabian Moerchen, and Zhuang Wang. Log-based predictive  
430 maintenance. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowl-*  
431 *edge Discovery and Data Mining, KDD '14*, pages 1867–1876. ACM, 2014. ISBN 978-1-4503-  
432 2956-9. doi: 10.1145/2623330.2623340.
- 433 [13] Zhiguo Li, Shiyu Zhou, Suresh Choubey, and Crispian Sievenpiper. Failure event prediction  
434 using the cox proportional hazard model driven by frequent failure signatures. *IIE Transac-*  
435 *tions*, 39(3):303–315, 2007. doi: 10.1080/07408170600847168.
- 436 [14] Skander Soltani. On the use of the wavelet decomposition for time series predic-  
437 tion. *Neurocomputing*, 48(1):267 – 277, 2002. ISSN 0925-2312. doi: [https://doi.org/](https://doi.org/10.1016/S0925-2312(01)00648-8)  
438 [10.1016/S0925-2312\(01\)00648-8](https://doi.org/10.1016/S0925-2312(01)00648-8). URL [http://www.sciencedirect.com/science/article/](http://www.sciencedirect.com/science/article/pii/S0925231201006488)  
439 [pii/S0925231201006488](http://www.sciencedirect.com/science/article/pii/S0925231201006488).
- 440 [15] C. A. G. Santos, P. K. M. M. Freire, G. B. L. Silva, and R. M. Silva. Discrete wavelet transform  
441 coupled with ann for daily discharge forecasting into tres marias reservoir. In *Proceedings of*  
442 *the International Association of Hydrological Sciences*, volume 364, pages 100–105, 2014.

- 443 [16] D. Jothimani, R. Shankar, and S.S. Yadav. Discrete wavelet transform-based prediction of  
444 stock index: A study on national stock exchange fifty index. *Journal of Financial Management*  
445 *and Analysis*, 28(2):35–49, 2015.
- 446 [17] Tianhong Liu, Haikun Wei, Chi Zhang, and Kanjian Zhang. Time series forecasting based  
447 on wavelet decomposition and feature extraction. *Neural Computing and Applications*, 28:  
448 183–195, 2017.
- 449 [18] Stephane Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.
- 450 [19] Guy Nason. *Wavelet methods in statistics with R*. Springer Science & Business Media, 2010.
- 451 [20] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- 452 [21] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and*  
453 *regression trees*. CRC press, 1984.
- 454 [22] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using  
455 random forests. *Pattern Recognition Letters*, 31(14):2225 – 2236, 2010. ISSN 0167-8655.  
456 doi: <https://doi.org/10.1016/j.patrec.2010.03.014>. URL [http://www.sciencedirect.com/  
457 science/article/pii/S0167865510000954](http://www.sciencedirect.com/science/article/pii/S0167865510000954).
- 458 [23] Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. Grouped variable impor-  
459 tance with random forests and application to multiple functional data analysis. *Computational*  
460 *Statistics & Data Analysis*, 90:15–35, 2015.