BIVARIATE ANALYSIS OF COMPLEX VIBRATION DATA: AN APPLICATION TO CONDITION MONITORING OF ROTATING MACHINERY

P. Pennacchi, A. Vania, N. Bachschmid

Dipartimento di Meccanica – Politecnico di Milano, Milano, Italy
paolo.pennacchi@polimi.it, andrea.vania@polimi.it, nicolo.bachschmid@polimi.it

Abstract
The problem of the robust definition of the acceptance regions in condition monitoring of the vibrations of rotating machinery is related to the more wide field of the analysis of bivariate data. Traditional parametric techniques and innovative non-parametric methods based on the statistical concept of the data depth are presented and critically examined in the paper. The performances with respect to the robustness in the estimation of the acceptance regions are analysed by means of experimental cases of real rotating machinery of a power plant.
Keywords: *bivariate analysis of data, rotor dynamics, condition monitoring, acceptance regions, data depth, robust estimation.*

1 INTRODUCTION

One of the most common tasks performed by condition monitoring systems applied to rotating machines, or to vibrating systems in general, is to define acceptance regions for safe operation on the basis of statistical analysis of the monitored data. Producers of condition monitoring systems usually implement algorithms in their systems, which are able to define dynamically over time, on the basis of the observation and the statistical analysis of the monitored vibration vectors, acceptance region boundaries (see for instance [1]). Usually, both pre-alarm, alarm and trip levels can be set by the user and the exceeding of a level can cause different actions. Sometimes this information is used for diagnostic purposes [2][3][4], so the correct definition of the acceptance region boundaries is a very important task.
The errors in the definition of the acceptance regions can be of two types. The first is the defective inclusion of normal operating conditions outside of the acceptance region: this causes false alarm or at least trips with consequent losses of production. The second is the over estimation of the acceptance region so that actual dangerous operating conditions are considered as normal. The capability of avoiding both errors can be seen as robust estimation of the acceptance region, therefore the discussion will be focused on the robustness in the following.
The harmonic components of the vibration signal are complex numbers and among all their possible graphical representation for condition monitoring purpose, a $\mathbb{R}^2$ space with Real and Imaginary axes is usually employed, in which the tip of the vibration vector defines a bivariate data cloud of points (figure 1). The data cloud can be organized as $m \times 2$ matrix $\mathbf{X}$, composed by m row vectors $\mathbf{x}_j$, $j \in (1,...,m)$, of real x_{j1} (1st column) and imaginary x_{j2} (2nd column) part of a data sample. Each vector $\mathbf{x}_j$ has module (amplitude in case of a vibration vector) x_j and phase ϕ_j. $\mathbf{F}_m$ is the empirical distribution of $\mathbf{X}$.
The analytical methods presented in the paper are applied in the case of condition monitoring of rotating machinery, but they are suitable to any kind of bivariate data.
Statistical analysis of complex bivariate data is included in the more comprehensive field of multivariate analysis, which is a field growing in importance in statistics. On one hand, classical multi-variate analysis relies strongly on the hypothesis of normality or near-normality of $\mathbf{F}_m$, which is often difficult to justify in condition monitoring. On the other hand, these methods are parametric and use straightforward extension of the statistical moment approach of the univariate case, so that they are easy to understand and to implement. Thus they are very popular in engineering field.

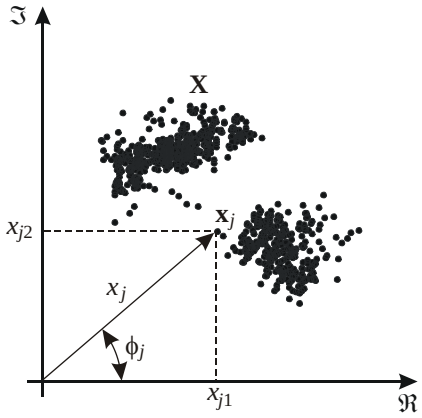

Figure 1. Graphical representation of vibration data in $\mathbb{R}^2$ space.

Historically, first acceptance regions were defined by circles in the $\mathbb{R}^2$ plane, but they were substituted by circle sectors in order to have a better evaluation of the phase changes [2]. Then, in order to define the acceptance region for a normal operating condition of the machine, once the collected data are composed by a sufficient number of measures, the boundaries of the region have to be defined.

This task is usually performed by considering the amplitude and the phase of the vibration vector having separately a univariate normal distribution. The maximum and minimum value of the amplitude and the phase are fixed so that a certain percentile, for instance 95%, of the collected data are included in the amplitude and phase boundary respectively. This methodology is not robust and has four main drawbacks:

i) it is not guaranteed that the required amount of data is actually included in the region;
ii) the assumption of a univariate normal distribution for the amplitude and the phase could not be appropriate since the actual data distribution might not be normal nor phase and amplitude might be correlated;
iii) phase boundaries sometimes could not be accurate when the vibration amplitude is small;
iv) circle sectors have a fixed orientation, i.e. they are generated by circles centred in the polar axes.

These topics are discussed in detail in section 2.

A partial solution to the third and fourth drawbacks is to consider a normal bivariate distribution of the vibration data, as introduced in section 3. Even if this method can give a better evaluation of the acceptance region in case of data having small amplitude, it has its drawback (and lack of robustness) in the assumption of the normal distribution of the data.

The solution, proposed by the authors in the paper, to these problems is to apply a general nonparametric multivariate methodology based on the statistical concept of *data depth*. The main idea is to analyse the region containing a certain amount of observations without inferring its data statistical distribution, but simply *counting* and *ordering* them, or better introducing a *ranking* in the values. In this sense the approach is nonparametric.

This fact is not trivial, since the considered data have a $\mathbb{R}^2$ distribution and ordering is not defined in $\mathbb{R}^2$, but considerable efforts have been made over the years, a survey can be obtained in [5], and different data depth have been developed [6][7] for statistical multivariate analysis. Section 4 presents an overview on data depth and ranking.

All the introduced methods are also analysed under the aspect of descriptive statistics and in particular the location, the dispersion (spread or scale), the correlation (related to the orientation of the sample), the skewness and the kurtosis (related to the presence of tails), accordingly also to the definition of robustness usually adopted by statistical inference, i.e. with respect to gross errors and outliers. It is worth to note that a parametric based method that is not robust in the last meaning, is also not robust in the estimation of the acceptance region, since, for instance, an error in the location

of the acceptance region might cause both the consideration of normal operating conditions outside of the acceptance region or the inclusion of dangerous operating conditions inside it.
Finally, the comparison of the results obtained by means of parametric and nonparametric methods on condition monitoring data of rotating machinery is presented and the different capabilities of the method used to define acceptance regions are evaluated.

2 UNIVARIATE PARAMETRIC METHOD APPLIED TO BIVARIATE DATA

The definition of acceptance regions by means of sectors is based on the hypothesis, questionable, that amplitude and the phase of the vibration vector have uncorrelated univariate normal distribution. Let:

$$\mathbf{X}_A = \{|\mathbf{x}_1|, |\mathbf{x}_2|, \ldots, |\mathbf{x}_m|\} = \{x_1, x_2, \ldots, x_m\} \text{ and } \mathbf{\Phi} = \{\arg(\mathbf{x}_1), \arg(\mathbf{x}_2), \ldots, \arg(\mathbf{x}_m)\} = \{\phi_1, \phi_2, \ldots, \phi_m\} \quad (1)$$

respectively the vectors of amplitude and phase of the vibration vector. It is assumed that the probability density functions of the amplitude and phase of the sample are univariate normal $N_A(\mu_A, \sigma_A)$ and $N_\phi(\mu_\phi, \sigma_\phi)$, so they are parameterised with respect to means and standard deviations. These are estimated by means of the averages and sample standard deviations:

$$\mu_A \leftarrow \overline{x}_A = \frac{1}{m}\sum_{i=1}^{m} x_i, \quad \mu_\phi \leftarrow \overline{\phi} = \frac{1}{m}\sum_{i=1}^{m} \phi_i \quad (2)$$

$$\sigma_A \leftarrow s_A = \left(\frac{1}{m-1}\sum_{i=1}^{m}\left(x_i - \overline{x}_A\right)^2\right)^{1/2}, \quad \sigma_\phi \leftarrow s_\phi = \left(\frac{1}{m-1}\sum_{i=1}^{m}\left(\phi_i - \overline{\phi}\right)^2\right)^{1/2} \quad (3)$$

Then, given a percentage p that have to be included in the acceptance region, the amplitude and phase values, which defines the frontier of the region symmetrical with respect to the averages (figure 2), are:

$$x = \overline{x}_A \pm s_A z, \quad \phi = \overline{\phi} \pm s_\phi z, \quad p = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right], \quad z = \sqrt{2}\operatorname{erf}^{-1}(2p-1) \quad (4)$$

where the error function erf is:

$$\operatorname{erf}(\xi) = \frac{2}{\sqrt{\pi}}\int_0^{\xi} e^{-t^2} dt \quad (5)$$

Thus, the obtained region is a circular ring sector.

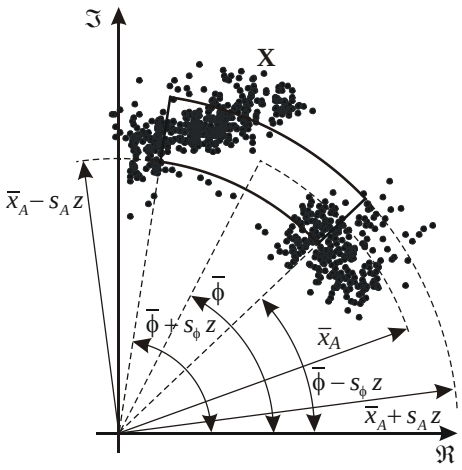

Figure 2. Sector position and shape.

With regards to the descriptive statistics aspect, since the present method uses a parametric estimate, the considered characteristics are related to the moments of the each univariate distribution of vibration vector amplitude and phase. In particular, the location, dispersion, skewness and kurtosis are defined respectively in terms of the first, the second, the third and the fourth moment. Therefore they descend "automatically" from the estimated distribution.

The location of the sample is estimated by means of the averages of vectors $\mathbf{X}_A$ and $\mathbf{\Phi}$, i.e. by vector $\overline{x}_A e^{i\overline{\phi}}$ see eq. (2) and figure 2, but it is well known that average is not a robust with respect to outliers [9,10,11].

The spread is related to the standard deviations of vectors $\mathbf{X}_A$ and $\mathbf{\Phi}$, thus to the second moment: the height of the sector is $2s_A$ and the angular extension $2s_\phi$. The robustness is similar to that of the location, since sample standard deviations are employed.

The orientation of the sector is fixed, due to its geometrical definition. Moreover the sector is always centred in the origin of the reference system.

The skewness of the sample is not evaluated nor reproduced, since the sector is symmetrical with respect to to vector $\overline{x}_A e^{i\overline{\phi}}$. Moreover the estimated univariate distribution of vectors $\mathbf{X}_A$ and $\mathbf{\Phi}$ is symmetric and the third moment is always null for normal distribution.

The kurtosis is related to that of the two univariate normal distributions and therefore equal to 3. So the capability to indicate outliers is rather poor if the sample distribution has not near-normal distribution for the amplitude and phase of the vibration.

These drawbacks are rather evident in figure 3, where some acceptance region, 99%-thick line, 94%-dashed line, 63%-thin line, are drawn on a data cloud coming from the condition monitoring system of the generator of a 50 MW combined cycle power plant. Even the sample is quite pathological and two data cluster are evident, if no preventive clustering is made, $\mathbf{X}_A$ and $\mathbf{\Phi}$ empirical distribution is not near-normal and the actual data estimation is not good. The acceptance region estimation is not robust and this fact can be evaluated also graphically.

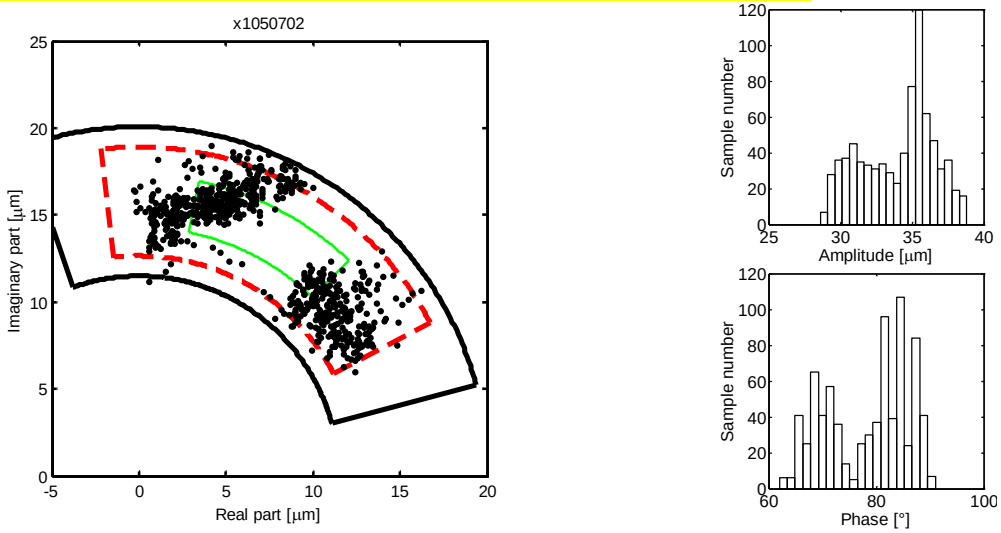

Figure 3. 1X vibration collected by one of the two proximity probes in a bearing of the generator of a 50 MW combined cycle power plant (left). Amplitude and phase distributions (right).

3 BIVARIATE PARAMETRIC METHOD

Whilst the definition of acceptance regions by means of circle sectors uses a univariate normal distribution separately for the amplitudes and the phases of the collected vibration data, the normal estimate uses a normal bivariate distribution for the real and the imaginary part of the vibrations. Also in this case it is questionable the hypothesis that the χ^2 distribution of $\mathbf{X}$ can be split into two χ distributions. Let:

$$\mathbf{X}_1 = \{x_{11}, x_{21}, \ldots, x_{m1}\} \text{ and } \mathbf{X}_2 = \{x_{12}, x_{22}, \ldots, x_{m2}\} \tag{6}$$

respectively the vectors of real and imaginary parts of the vibration vector.
The continuous probability density function of the normal bivariate distribution is given by:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_1\sqrt{1-\rho_{12}^2}} \exp\left(-\frac{1}{2\left(1-\rho_{12}^2\right)} B(x_1, x_2)\right) \tag{7}$$

where

$$B(x_1, x_2) = \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right] \tag{8}$$

and μ_1, μ_2, σ_1 and σ_2 are the mean value and the standard deviation of the marginal distribution of x_1 and x_2, ρ_{12} the correlation coefficient defined as:

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1\sigma_2} \tag{9}$$

and σ_{12} is the covariance of x_1 and x_2.
A relative frequency given by eq. (7) is corresponding to each couple of values (x_1, x_2), while probability is given by the cumulative distribution function that is the volume bounded by eq. (7) and the plane parallel to plane x_1x_2 passing through (x_1, x_2). Since the acceptance regions of a vector quantity are regions of the polar plane, in which a certain amount of the vector tips of the measured value is included, if the data distribution **F** is normal, the region definition is possible by means of the cumulative distribution function and its sections made by planes parallel to the polar plane x_1x_2. The sliced sections are ellipses as it is shown in appendix 1
Then, distribution parameters of normal bivariate are substituted by their estimators:

$$\mu_k \leftarrow \overline{x}_k = \frac{1}{m}\sum_{i=1}^{m} x_{ik} \tag{10}$$

$$\sigma_k \leftarrow s_k = \left(\frac{1}{m-1}\sum_{i=1}^{m}\left(x_{ik}-\overline{x}_k\right)^2\right)^{1/2} \tag{11}$$

$$\sigma_{12} \leftarrow s_{12} = \left(\frac{1}{m-1}\sum_{i=1}^{m}\left(x_{i1}-\overline{x}_1\right)\left(x_{i2}-\overline{x}_2\right)\right)^{1/2} \tag{12}$$

Once parameter C is determined as function of the given percentage included in the region, ellipse semi-axes, see eq. (30) and figure 4, are given by:

$$a = C\,s_1,\ b = C\,s_2 \tag{13}$$

while the angle between semi-axis a and x is given by:

$$\alpha = \frac{1}{2}\tan^{-1}\frac{2\sum_{i=1}^{m}\left(x_{i1}-\overline{x}_1\right)\left(x_{i2}-\overline{x}_2\right)}{\sum_{i=1}^{m}\left(x_{i1}-\overline{x}_1\right)^2 - \sum_{i=1}^{m}\left(x_{i2}-\overline{x}_2\right)^2} \tag{14}$$

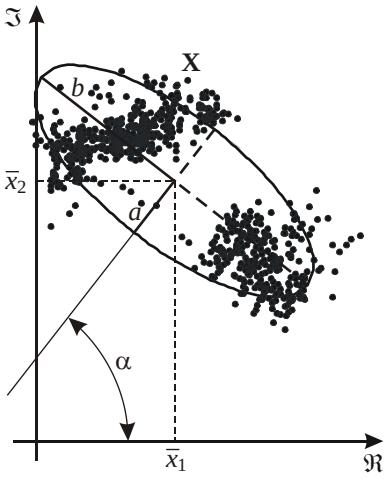

Figure 4. Ellipse position and shape.

Once again it is worth to remember that if the data distribution is not exactly normal bivariate, the actual quantity of the data included inside an ellipse could not be corresponding to the probabilistic estimate. An example is shown using the data of the 1X vibration collected by one of the two proximity probes in a bearing of the exciter of a 660MW power unit measured with the rotor current ranging from 2000 A to 3600 A with a short circuit at initial stage (figure 5). Also the two marginal distributions are shown and are not near-normal. The sample has a "banana shape" and the region containing the 99%-thick line, 94%-dashed line, 63%-thin line are calculated using a normal estimate. If these regions were assumed to be acceptance regions, the actual percentage of the sample included inside them is reported in table 1. From the analysis of figure 5 and table 1, the normal estimate does not give a good estimation of the data in this case and a not robust estimation of the acceptance region.

Table 1. Actual amount of data contained in the acceptance region.

Nominal percentage	63%	94%	99%
Normal estimate	73.00%	98.37%	100%

In regards to the descriptive statistics aspect, since also the present method uses a parametric estimate, the considered characteristics are related to the moments of the bivariate distribution.
The location of the sample is estimated by means of the averages of vectors $\mathbf{X}_1$ and $\mathbf{X}_2$, see eq. (10), since all the ellipses are centred in $(\bar{x}_1, \bar{x}_2)$, see eq. (30), but again these estimators of the means are not robust with respect to outliers [8][9][10]. Moreover, notice that in general:

$$\bar{x}_A \neq \sqrt{\bar{x}_1^2 + \bar{x}_2^2}, \quad \bar{\phi} \neq \arg(\bar{x}_1 + i\,\bar{x}_2) \tag{15}$$

so that the location given by bivariate estimation is different from that of univariate.
The dispersion is related to the standard deviations of vectors $\mathbf{X}_1$ and $\mathbf{X}_2$, thus to the second moment, since ellipse axes are proportional to sample standard deviations, see eq. (13). The robustness is similar to that of the location, since sample standard deviations are employed. Sophisticated parametric methods have been developed, in order to have robust estimation of the dispersion, see [11] and one of them is used in the following in eq. (24), but a detailed discussion about them is far from the scope of this work.
The orientation of the sample is again related to the first and second moments, see eq. (14) that allows the axis inclination to be calculated.
The skewness of the sample is impossible to be evaluated, since the estimated distribution is symmetric and the third moment is always equal to zero.

The kurtosis is related to that of the single marginal normal distribution and therefore equal to 3. So the capability to indicate outliers is rather poor if the sample distribution is not near-normal.

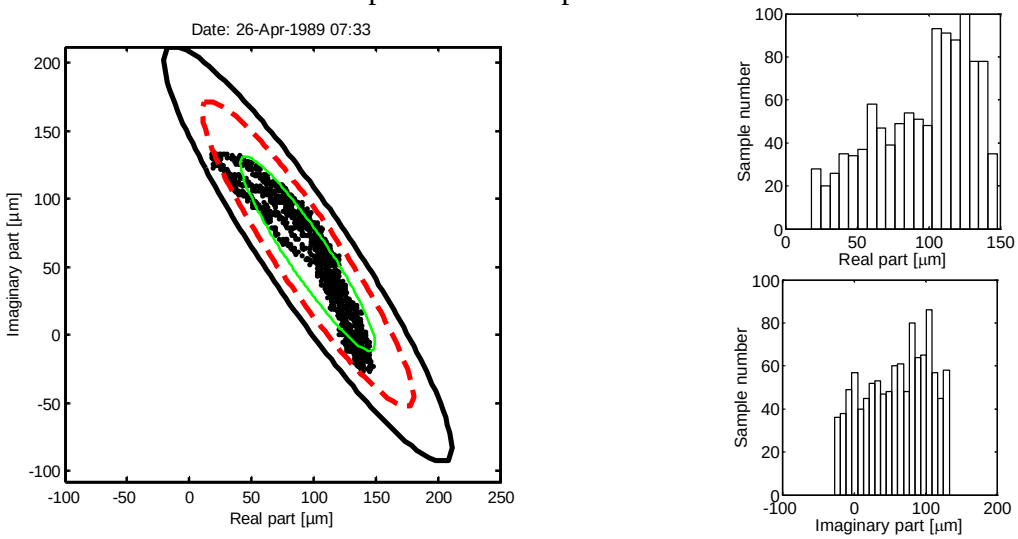

Figure 5. 1X vibration collected by one of the two proximity probes in a bearing of the exciter of a 660MW power unit measured with the rotor current ranging from 2000 A to 3600 A with a short circuit at initial stage (left). Real and imaginary part marginal distributions (right).

4 NONPARAMETRIC APPROACH BASED ON DATA DEPTH

Whilst nonparametric approaches are well-known in Statistics for sample analysis [12], they are not commonly applied in engineering, so it is useful an introduction about the main concepts. In general terms, a *data depth* $DD(\mathbf{F};\theta)$ is a way of measuring how deep (or central) a given point $\theta \in \mathbb{R}^d$, $d \geq 1$ with respect to a continuous probability distribution $\mathbf{F}$ or to a given data cloud $\mathbf{F}_m$. A unique definition of data depth does not exist and some of the proposed are:

- The *Mahalanobis' depth* [13]:

$$M_hD(\mathbf{F};\theta) = \left[1+\left(\theta-\mu_{\mathbf{F}}\right)\Sigma_{\mathbf{F}}^{-1}\left(\theta-\mu_{\mathbf{F}}\right)\right]^{-1} \tag{16}$$

  where $\mu_{\mathbf{F}}$ is the mean vector and Σ the covariance matrix of $\mathbf{F}$. In the sample version $\mu_{\mathbf{F}}$ and Σ are replaced by their estimators.

- The *Oja's depth* [14]:

$$OD(\mathbf{F};\theta) = \left[1+E_{\mathbf{F}}\left\{volume\left(S\left[\theta,\mathbf{X}_1,\ldots,\mathbf{X}_d\right]\right)\right\}\right]^{-1} \tag{17}$$

  where $S\left[\theta,\mathbf{X}_1,\ldots,\mathbf{X}_d\right]$ is the closed simplex with vertices θ and d random observation $\mathbf{X}_1,\ldots,\mathbf{X}_d$ from $\mathbf{F}$, E is the expected value operator. A simplex, sometimes called a hypertetrahedron [15], is the generalization of a tetrahedral region of space to n dimensions. The simplex is so-named because it represents the simplest possible polytope (a finite region of n-dimensional space enclosed by a finite number of hyperplanes) in any given space.,
  For the sample version:

$$OD(\mathbf{F}_m;\theta) = \binom{m}{d}^{-1}\left[1+\sum_{*}\left\{volume\left(S\left[\theta,\mathbf{X}_1,\ldots,\mathbf{X}_d\right]\right)\right\}\right]^{-1} \tag{18}$$

  where * indicates that the sum is extended to all the d-plets $(i_1,\ldots,i_d)$ such that $1 \leq i_1 \leq \ldots \leq i_d \leq m$.

- The *simplicial depth* [16] developed by Liu:

$$SD(\mathbf{F};\theta) = P_{\mathbf{F}}\left\{\theta \in S\left[\mathbf{X}_1,\ldots,\mathbf{X}_{d+1}\right]\right\} \tag{19}$$

where $S\left[\mathbf{X}_1,\ldots,\mathbf{X}_{d+1}\right]$ is a closed simplex formed by $d+1$ random observation from **F**. The sample version replaces **F** with $\mathbf{F}_m$ or alternatively:

$$SD(\mathbf{F}_m;\theta) = \binom{m}{d+1}\sum_{*} I_{(\theta \in S[\mathbf{X}_1,\ldots,\mathbf{X}_{d+1}])} \tag{20}$$

where $I_{(.)}$ is the indicator function, i.e. 1 if the condition (.) is true, 0 otherwise, and * indicates that the sum is extended to all the *d*-plets $(i_1,\ldots,i_{d+1})$ such that $1 \le i_1 \le \ldots \le i_{d+1} \le m$.

- The *majority depth* [6] developed by Singh:

$$MJD(\mathbf{F};\theta) = P_{\mathbf{F}}\left\{\theta \text{ is in a major side determined by } \left(\mathbf{X}_1,\ldots,\mathbf{X}_d\right)\right\} \tag{21}$$

major side is a half-space bounded by the hyperplane containing $\left(\mathbf{X}_1,\ldots,\mathbf{X}_d\right)$ which has probability ≥ 0.5. The sample version replaces **F** with $\mathbf{F}_m$.

- The *likelihood depth LD* [6] is simply the probability density function. The empirical version can be any consistent density estimate at θ.
- The *projection depth* [7][17]:

$$PD(\mathbf{F};\theta) = \left(1+O(\theta)\right)^{-1} \tag{22}$$

where

$$O(\theta) = \sup_{\|\mathbf{u}\|=1} \frac{\left|\mathbf{u}'\theta - \mathrm{Med}_{1\le i\le m}\left\{\mathbf{u}'\mathbf{X}_i\right\}\right|}{\mathrm{MAD}_{1\le i\le m}\left\{\mathbf{u}'\mathbf{X}_i\right\}} \tag{23}$$

is the robust measure of the outlyingness of θ with respect to **F** [18] and MAD denoted the univariate median absolute deviation:

$$\mathrm{MAD}(\mathbf{Y}) = \mathrm{Med}\left(\left|\mathbf{Y} - \mathrm{Med}(\mathbf{Y})\right|\right) \tag{24}$$

The sample version replaces **F** with $\mathbf{F}_m$.

- The *convex hull peeling depth CD* [5] at the sample point $\mathbf{X}_k$ with respect to the data set $\left\{\mathbf{X}_1,\ldots,\mathbf{X}_m\right\}$ is simply the level of the convex layer $\mathbf{X}_k$ belongs to. To build the first convex layer, the smallest convex hull that encloses all sample points $\left\{\mathbf{X}_1,\ldots,\mathbf{X}_m\right\}$ is considered. The sample points on the perimeter are designed as the first convex layer and removed. The next convex hull is considered and the second convex layer defined. The procedure iterates and sequence of nested convex layers is formed. The higher layer a point belongs to, the deeper the point is within the data cloud.
- The *half-space depth* [6][19] developed by Tukey:

$$HD(\mathbf{F};\theta) = \inf_{H}\left[P(H) : H \text{ is a closed half-space in } \mathbb{R}^d \text{ and } \theta \in H\right] \tag{25}$$

The sample version replaces **F** with $\mathbf{F}_m$. If bivariate finite sample is considered, which is more pertinent to this paper, $d = 2$ and an equivalent definition is:

$$HD(\mathbf{F};\theta) = \min_{H} \#\left\{i;\mathbf{x}_i \in H\right\} \tag{26}$$

where # indicates the number of the data samples satisfying the condition inside the brace parentheses and H ranges over all the half-planes of which the boundary line passes through

θ. The definition in eq. (26), given in [18][20], is not normalized like those before, but is geometrically meaningful: practically it represents the smallest number of $\mathbf{x}_i$ contained in any half-plane with boundary line passing through θ. The consideration of figure 6 is useful to understand the definition given in eq. (26): the considered θ on the three centre-left side pictures has depth equal to 1. Practically, a line, passing through each point of the $\mathbb{R}^2$ field to which the collected data belong, is drawn. The $\mathbb{R}^2$ field is divided into two half-planes and the number of the data point in each half-plane is counted.
Obviously, a point θ outside the convex hull of **X** has $HD(\mathbf{F};\theta)$ equal to zero, as shown in rightmost side of figure 6 and this limits the points θ to be considered.

Several researchers have compared the presented depth measures with respect to different statistical criteria [6][7][18][21][22][23] and *HD* appears to have the best overall performances and robustness. Therefore we will consider it only, using definition in eq. (26) that have the advantage of giving an integer number as data depth measure. Anyhow, data ordering criterion (see [24][25][26]) described in the following is applicable whichever *DD* definition is chosen (except for *CD* that directly orders the element of the data) with few adjustments.

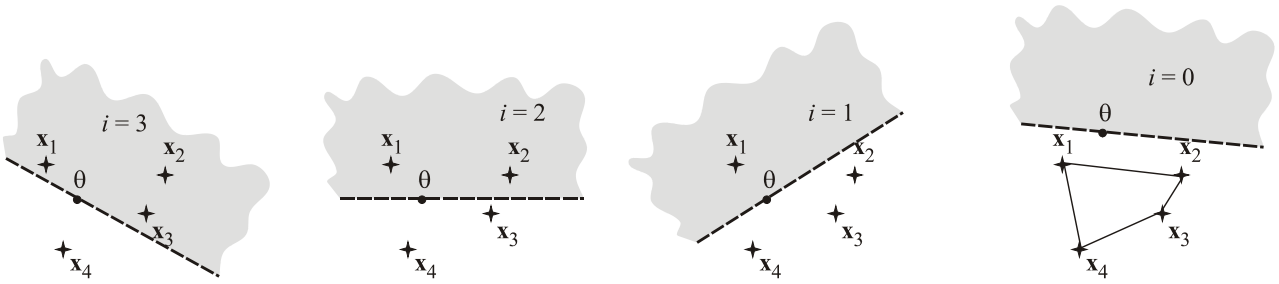

Figure 6. Calculation of the half-space depth of a general point θ.

The *depth region* of depth *k* is defined as the set D_k of points θ having $HD(\mathbf{F};\theta) \geq k$. Likewise, D_k is the intersection of all the closed half-planes that contain at least $m-k+1$ data points, so D_k is a limited and convex polytope. Moreover $D_{k+1} \subset D_k$ as proven in [18]. The boundary of D_k is a convex polygon which is called *contour of depth k* and each vertex is the intersection point of two lines each one is passing through two data points.
Since *HD* is used, if only one point θ exists that has the maximum $HD(\mathbf{F};\theta)$ then it is defined as *Tukey's median* **T*** of **X**, otherwise **T*** is the centre of mass of the maximum depth region. If others *DD* are used, the deepest point concept is retained, but it is generally defined as *depth median*. Now, the *i*-percentile region A_i that includes the percentile Q_i of the data is built as follows: if $\#D_k$ is the number of the experimental data in D_k, first the value *k* is determined such as:

$$\#D_k \leq \lfloor m Q_i \rfloor \leq \#D_{k-1} \tag{27}$$

where $\lfloor \cdot \rfloor$ indicates the floor function, i.e. it rounds $\cdot$ to the nearest integer less than or equal to $\cdot$.
Then, a linear interpolation between D_k and D_{k-1} relative to point **T*** is made in order to obtain the region A_i for which $\#A_i = \lfloor m Q_i \rfloor$. Note that the *i*-percentile region A_i is a convex polygon too, but it has not necessarily vertexes belonging to data cloud **X**.
By considering now the engineering application of these concepts, two criteria can selected to define acceptance region on the basis of data depth analysis.
The first is to define acceptance region simply choosing the confidence level of the percentile region, similarly to the parametric methods presented in paragraphs 2 and 3. In regards to the descriptive statistics aspect, since all the *DD* methods are nonparametric, moments are not considered: the location is estimated by means of the depth median, the spread by the size of the percentile region, the correlation by region orientation, the skewness by of the shape of the region. Data points outside the i-percentile region are considered outliers at confidence level *i*.

The second criterion is to use the bagplot, introduced by Rousseeuw et al. [24], which is a generalization of Tukey's univariate *boxplot* to bivariate data set distributions. As the fundamental concept is the *rank* in the univariate case, as is the *DD*, in the bivariate case. To build a bagplot, the previous exposed ordering method is followed up to eq. (27) in which the *k* value of region of the 50-percentile is instead determined:

$$\#D_k \le \lfloor m/2 \rfloor \le \#D_{k-1} \tag{28}$$

The set A_{50} for which $\#A_{50} = \lfloor m/2 \rfloor$, obtained by a linear interpolation between D_k and D_{k-1} relative to point **T*** is defined as *bag*. Similarly to the boxplot, also in this case a *fence* is defined and it is obtained by inflating bag A_{50} by a certain factor relatively to **T***. Conventionally, after the results of simulations reported in [24], the inflating factor has been chosen equal to 3. The *loop* contains all the points between the bag and the fence, so its vertexes belong to **X**. An example of bagplot of bivariate data is shown in figure 7, compared with the boxplots of the univariate marginal distributions.

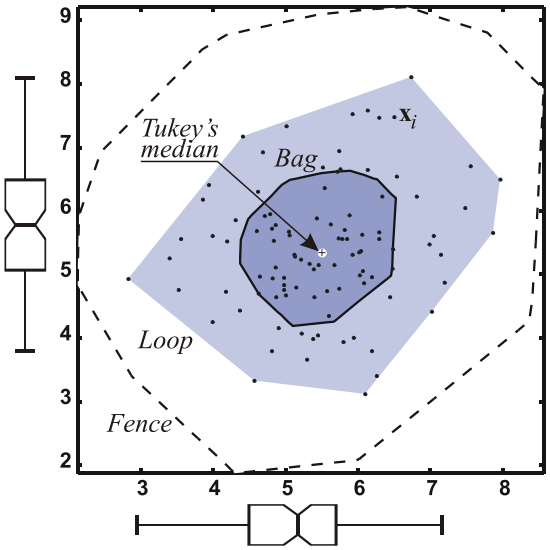

Figure 7. Bagplot of a bivariate data.

In Rousseeuw's opinion, the bagplot is effective in giving an immediate graphical and descriptive representation of some fundamental characteristics of a bivariate data distribution:

- its location or position (by means of the Tukey's median);
- its spread or dispersion (by means of the bag size);
- its correlation (by means of the bag orientation);
- its asymmetry or skewness (by means of the shape of the bag and the loop);
- the presence of tails (by means of the points next to the boundary of the loop and outside of the fence, which are considered as outliers).

The main drawback of *DD* methods with respect to parametric methods of paragraphs 2 and 3 is that specific algorithms for depth calculation have to be used. Anyhow, for the *HD* and the bagplot some public domain algorithms have been developed by Rousseeuw et al. [27][28] and they are used for the data analysis presented in the following. Due to *HD* definition, the algorithm that implements its calculation is rather time consuming since it is $O(m^2 \log m)$, therefore some approximation are made if *m* is greater than 200, thus in this case A_i region could not contain exactly the percentile Q_i of the sample, but this has a marginal relevance in the analyses presented in the following. The authors have developed an alternative algorithm that overrides this limit and takes advantage from the increasing calculation speed of nowadays computers and it has been used for the calculation of the percentile regions.

5 CASE 1

The first experimental case presented is relative to the condition monitoring of a 50 MW combined cycle power plant. Figure 8 shows the time history of the amplitude and phase of the synchronous (1X in the following) vibration over little less than six days. No evident conditions of different load appear from the diagram. Figure 9 shows the polar plot of the same data, from which the data cloud has an "almond" shape, but no further information about data density is manifest. In order to have a better idea the consideration of the 3D histogram of the data in figure 10 suggest that the distribution could be considered with a certain degree of approximation as a near-normal bivariate, even if the tails are not very symmetric. Therefore a good estimation of the actual distribution can be expected by ellipses.

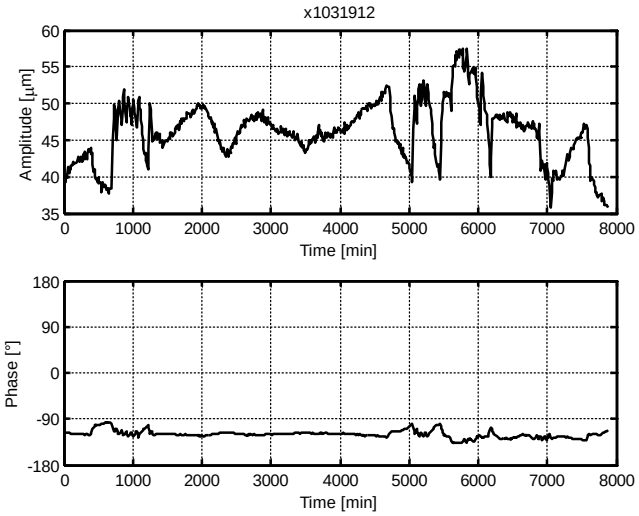

Figure 8. Case 1: time history 1X of vibration collected by one of the two proximity probes in a bearing of the generator of a 50 MW combined cycle power plant.

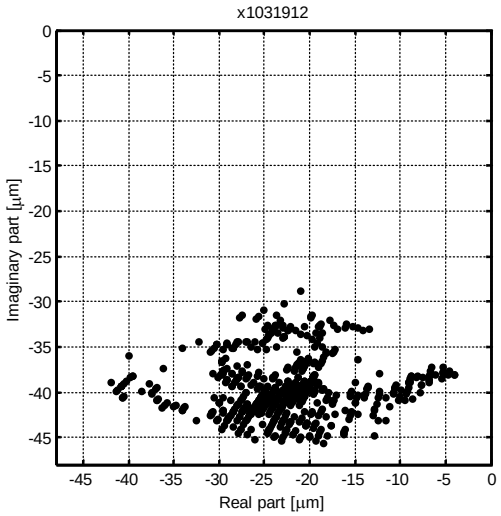

Figure 9. Case 1: polar plot 1X of vibration collected by one of the two proximity probes in a bearing of the generator of a 50 MW combined cycle power plant.

The acceptance regions defined by means of the four different criteria described in paragraphs 2, 3 and 4 are shown in figure 11 for the circle sectors, in figure 12 for the ellipses, in figure 14 for the bagplot and in figure 13 for the percentile regions. For the circle sectors and the ellipses, the sectors are shown for confidence levels from 10% to 90% with a 10% step (thin lines) and for the 99% (bold dashed line). The same confidence levels are shown also for the percentile regions, but note that the 99% value is less significant in this case, since it is very close to the first convex hull of the data. In all the diagrams, the estimated location of the data obtained by means of the respective method is indicated by a target sign (note that for the bagplot and the percentile regions the estimator is the Tukey's median in both cases).

The location of the data obtained by all the four criteria can be deemed as good, even if Tukey's median appears a little bit more "central" then the estimates given by eq. (2) and eq. (10), when figure 10 is considered. The dispersion estimation is good for all criteria except for circle sectors, since they overestimate acceptance regions for higher percentiles. Similar comments can be made in regards to the presence of tails.

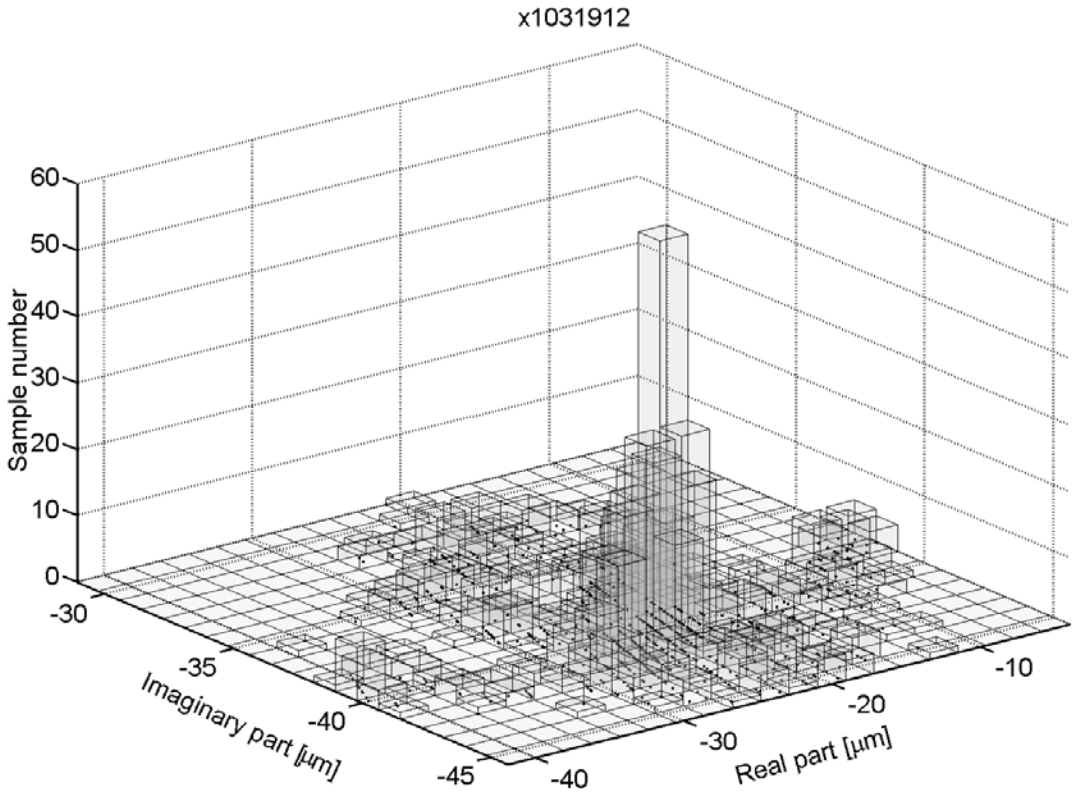

Figure 10. 3D histogram of case 1.

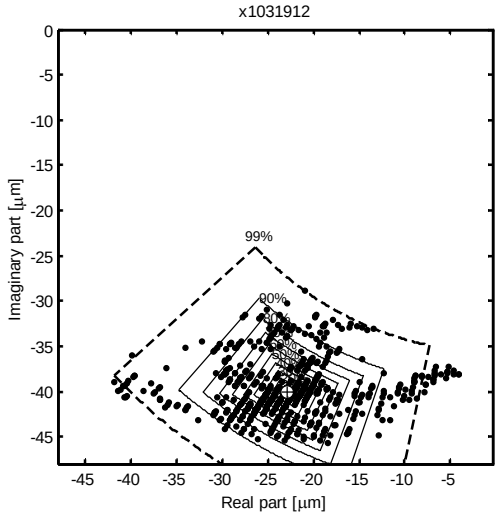

Figure 11. Case 1: acceptance regions defined by means of circle sectors.

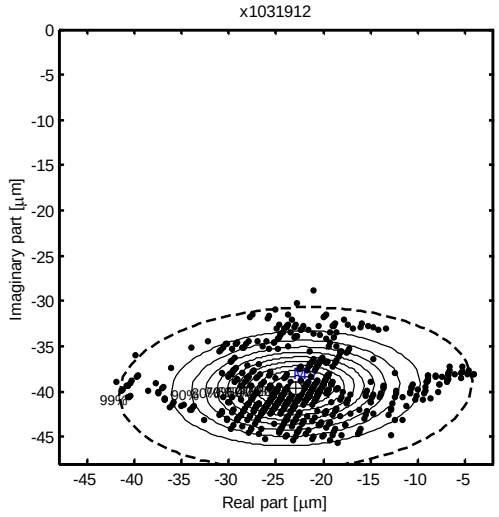

Figure 12. Case 1: acceptance regions defined by means of ellipses.

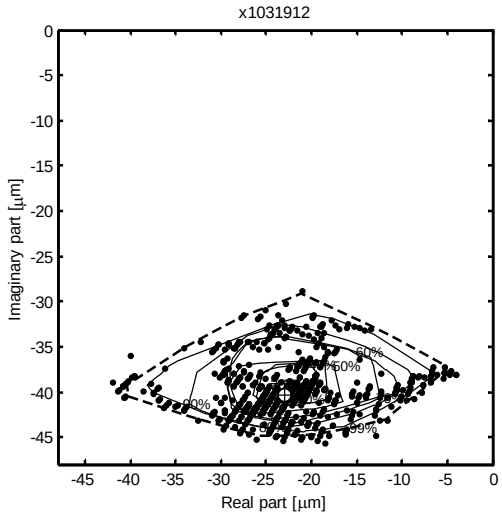

Figure 13. Case 1: acceptance regions defined by means of percentile regions.

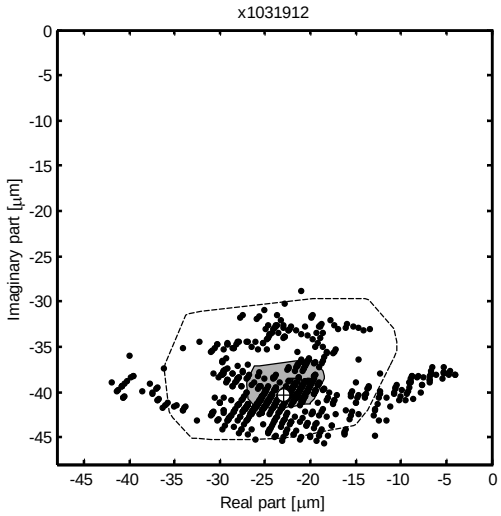

Figure 14. Case 1: acceptance regions defined by means of bagplot.

6 CASE 2

The second experimental case is relative to the same power plant of Case 1, but in a different period. The time history of the amplitude and phase of the 1X vibration over about six days (figure 15) shows almost two different operating conditions, depending on the day-light vs. night-time load. The polar plot in figure 16 has an "S" shape which can hardly be alike to a normal distribution.

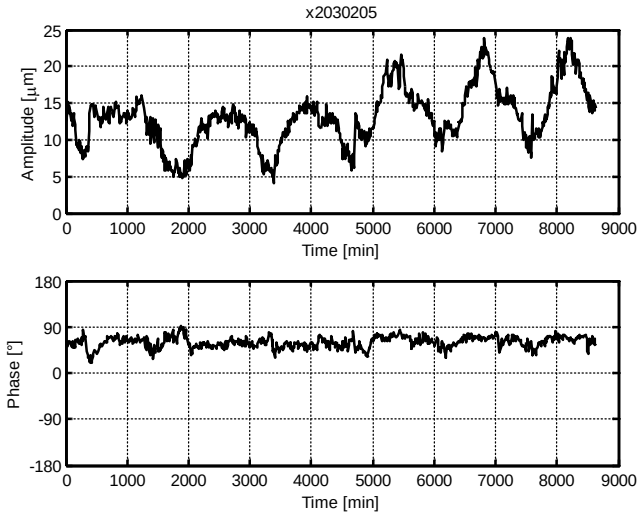

Figure 15. Case 2: time history 1X of vibration collected by one of the two proximity probes in a bearing of the generator of a 50 MW combined cycle power plant.

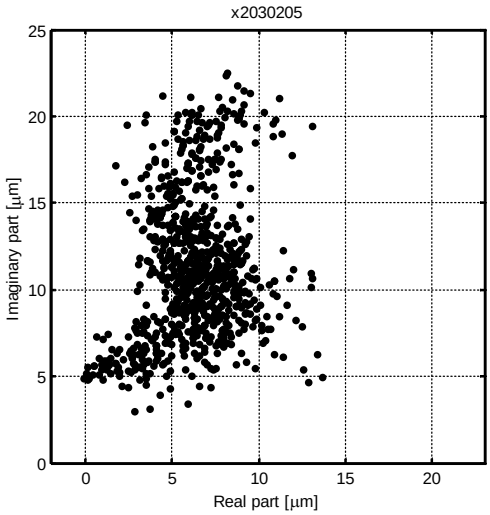

Figure 16. Case 2: polar plot 1X of vibration collected by one of the two proximity probes in a bearing of the generator of a 50 MW combined cycle power plant.

Thus, instead of analyzing the raw data, a preliminary cluster analysis has been performed. This fact implies that the condition monitoring system is able to perform by itself this kind of operation. Some proposal in this sense in rotor dynamics have been presented in literature [29] using artificial neural networks. However, for the purposes of the paper, more traditional clustering techniques have been used, which are briefly described in appendix 2.

6.1 Case 2: Analysis of the Data Cluster

Since data of case 2 are relative to almost two different operating conditions, depending on the day-light vs. night-time load, it is appropriate to analyze data grouped in clusters having the same

operating conditions. Even if this operation might be performed manually, being the data records time and day based, it has been preferred to apply an automatic clustering in order to be independent from the operator action and to simulate the procedure of an automated condition monitoring system. Among the possible criteria that can be applied to cluster data, a hierarchical clustering of the data has been used. A detailed explanation about hierarchical clustering can be found in [30] [32] and is far from the scope of this paper, while some basic concepts are reported in appendix. Figure 17 shows the result of the clustering, in which the two clusters having the highest level of consistency, when the different criteria described in appendix 2 are used for the calculation of the distance between points and the linkage, are indicated by black and grey solid dots respectively. In the hierarchical clustering terminology the clusters with the highest level of consistency are said to have the highest cophenetic coefficient, see eq. (48).

The implicit criterion that clustering seems to have applied is to discriminate the clusters on the basis of an amplitude threshold, but the automatic clustering can be deemed as acceptable, since indicate acceptably the different operating conditions.

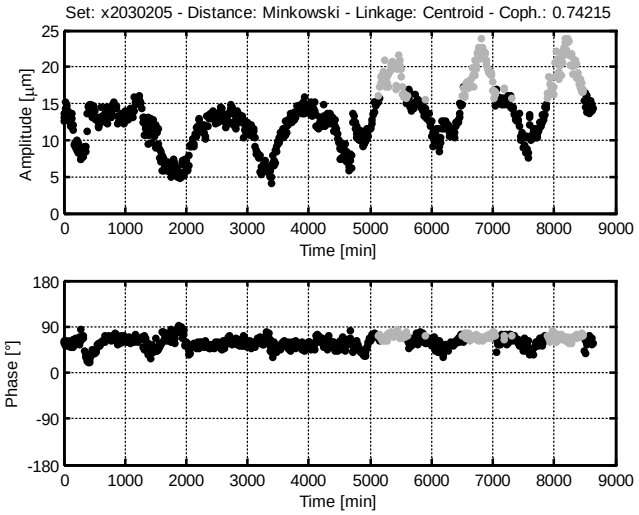

Figure 17. Case 2: time history with the data grouped into two clusters. Cluster 1 is indicated by solid black dots, cluster 2 by solid grey dots.

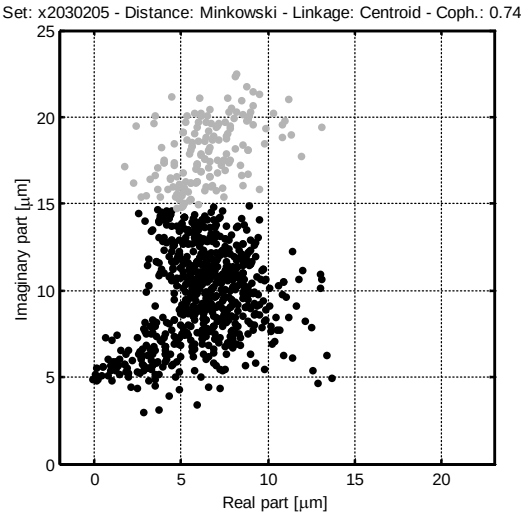

Figure 18. Case 2: polar plot of the two data cluster. Cluster 1 is indicated by solid black dots, cluster 2 by solid grey dots.

From the polar plot of the two clusters shown in figure 18, it is difficult to recognise a near-normal distribution for both the clusters. This opinion is confirmed by the 3D histograms in figure 19 and figure 20.

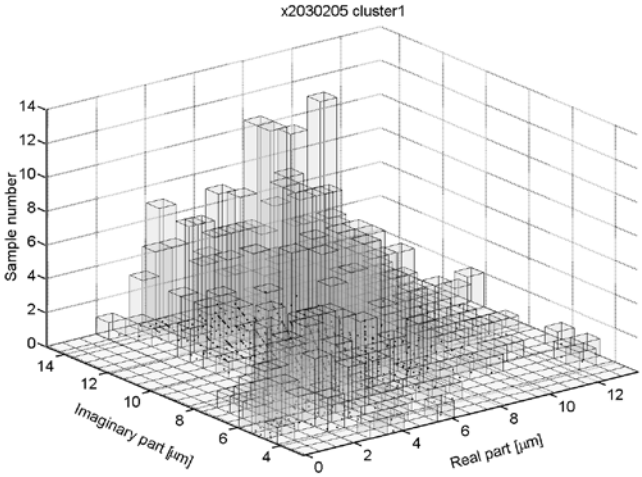

Figure 19. 3D histogram of cluster 1.

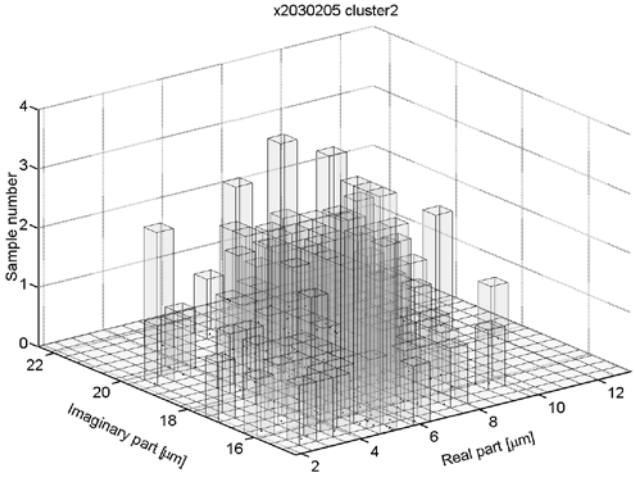

Figure 20. 3D histogram of cluster 2.

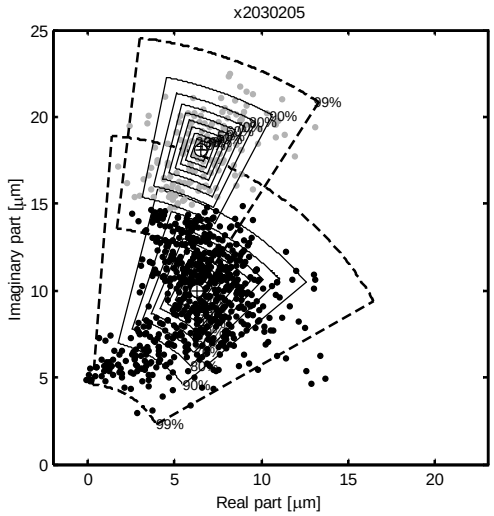

Figure 21. Case 2: acceptance regions defined by means of circle sectors for both clusters.

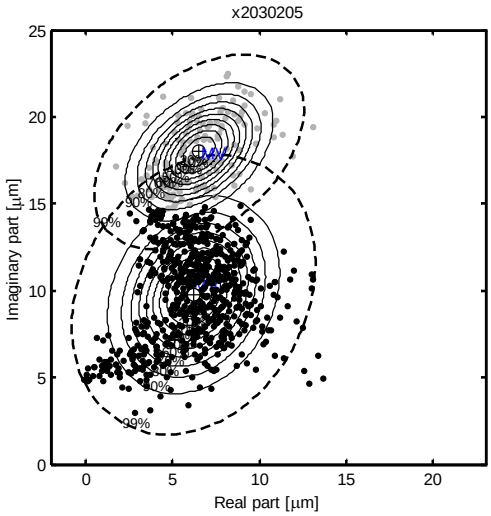

Figure 22. Case 2: acceptance regions defined by means of ellipses for both clusters.

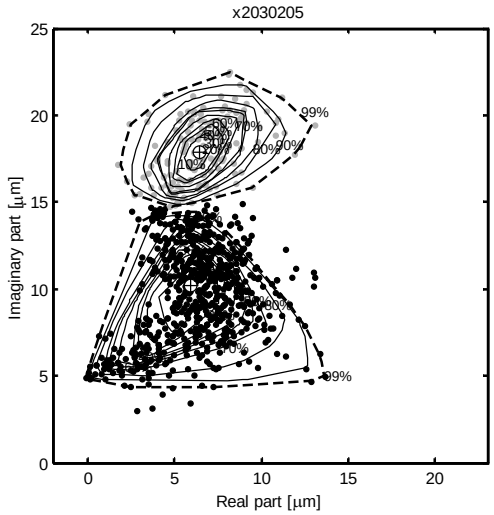

Figure 23. Case 2: acceptance regions defined by means of percentile regions for both clusters.

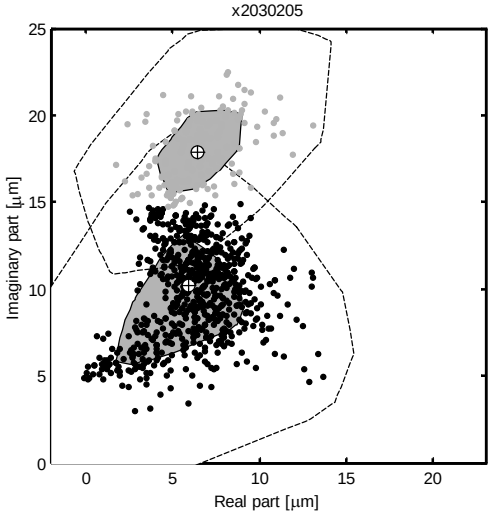

Figure 24. Case 2: acceptance regions defined by means of bagplot for both clusters.

The location of the data obtained by all the four estimation criteria can be deemed as good. These results are not confirmed for the dispersion estimation. Only percentile regions, see figure 23, keeps the two clusters separated. The application of the other estimation methods determines the overlapping of the acceptance regions of the data clusters, see figure 21, figure 22 and figure 24, indicating that these methods are not effective in the estimation of outliers. The last result can be surprising if the bagplot is accounted for, but probably the inflating factor proposed by the authors that introduced it [24] should be less than 3. This way, the shape of the fence would be retained, but it would be closer to the bag and a better estimation of outliers would be obtained.

7. CASE 3

The last experimental case is again relative to the same power plant of Case 1 and 2, but in a different period. The time history of the amplitude and phase of the 1X vibration over about five days (figure 25) shows almost two different operating conditions, depending in this case on the different load in work-days or holydays. These conditions cause a rather strange “atoll” shape for

the data distribution in the polar plot (figure 26). The 3D histogram in figure 27 confirms that the data distribution is far from near-normal.
Even if it is evident that two cluster of data are present, the complete sample was analyzed using the four methods described (figure 28 for circle sectors, figure 29 for ellipses, figure 30 for percentile regions and figure 31 for bagplot) in order to evaluate their capability with respect to data characterized by strong "seasonal" behaviour. Obviously, acceptance regions are largely overestimated by parametric methods (figure 28 and figure 29). The necessity to correct the fence inflating factor is confirmed (figure 31). Also the performance of the percentile regions is not optimal, since they are necessarily convex (see the properties described in paragraph 4) and thus cannot reproduce the "hole" of the "atoll", but in any case there is not the overestimation of the region since, as already reported, they tend to the first convex hull of the data. Therefore, percentile regions are able to acceptably handle also data affected by strong "seasonal" behaviour.

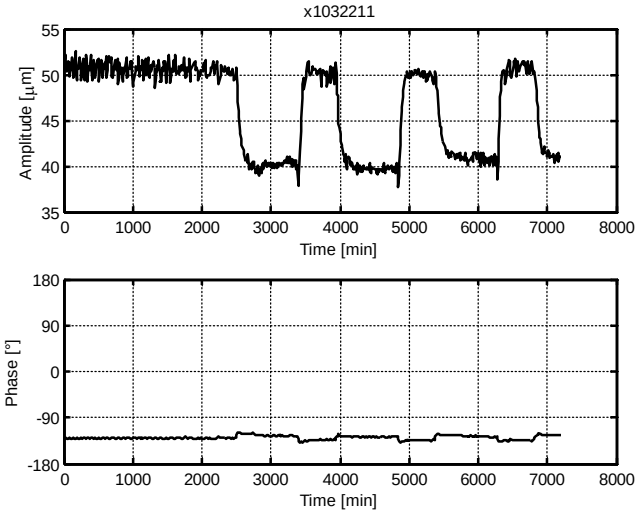

Figure 25. Case 3: time history 1X of vibration collected by one of the two proximity probes in a bearing of the generator of a 50 MW combined cycle power plant.

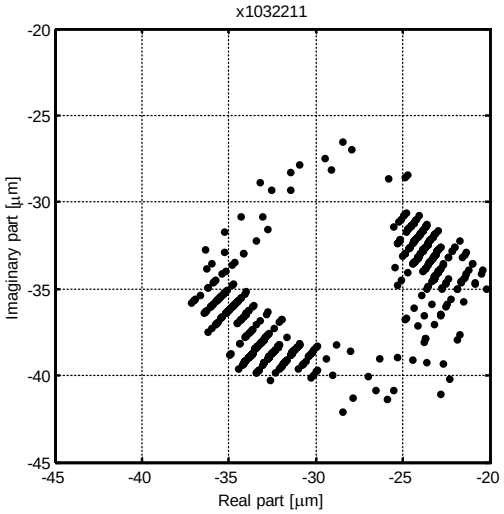

Figure 26. Case 3: polar plot 1X of vibration collected by one of the two proximity probes in a bearing of the generator of a 50 MW combined cycle power plant.

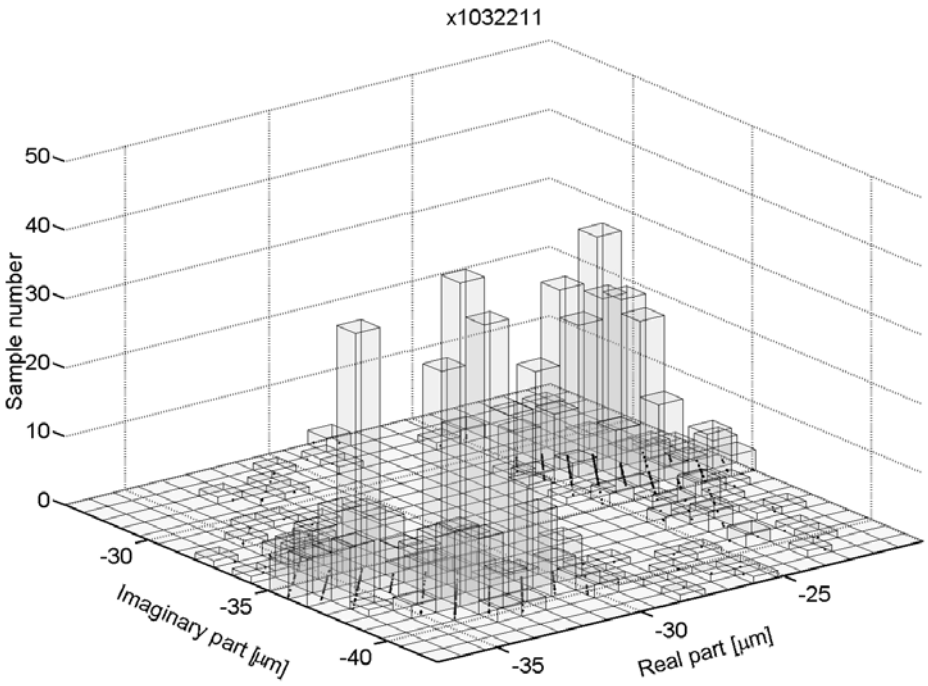

Figure 27. 3D histogram of case 3.

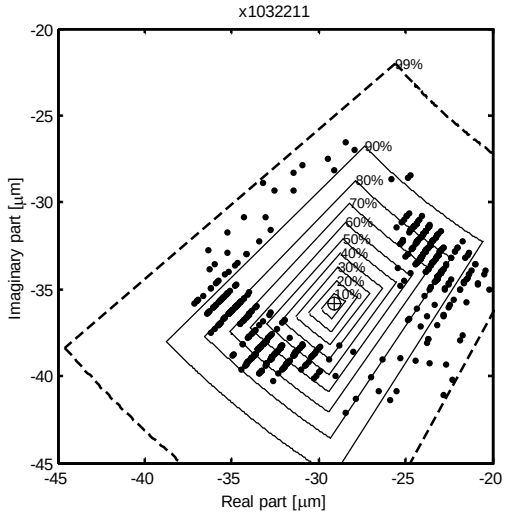

Figure 28. Case 3: acceptance regions defined by means of circle sectors.

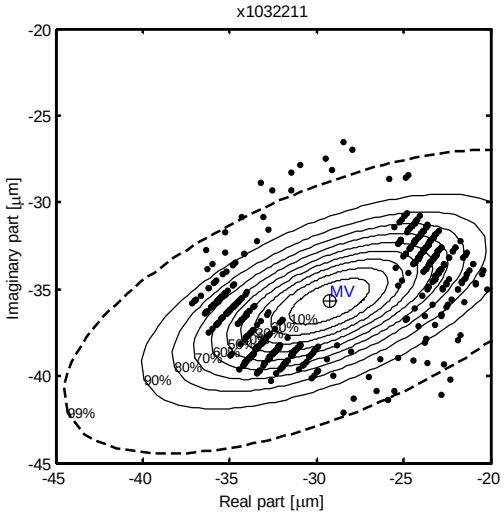

Figure 29. Case 3: acceptance regions defined by means of ellipses.

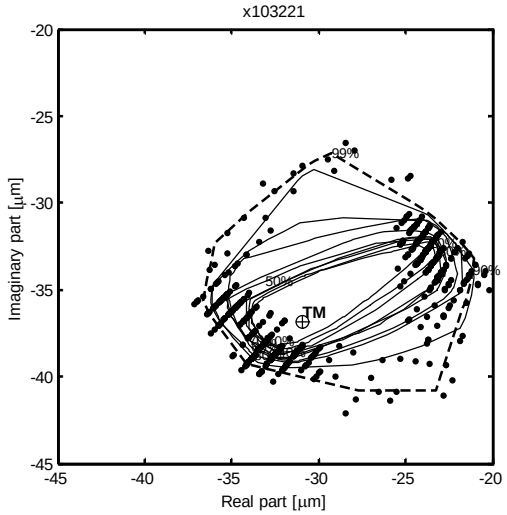

Figure 30. Case 3: acceptance regions defined by means of percentile regions.

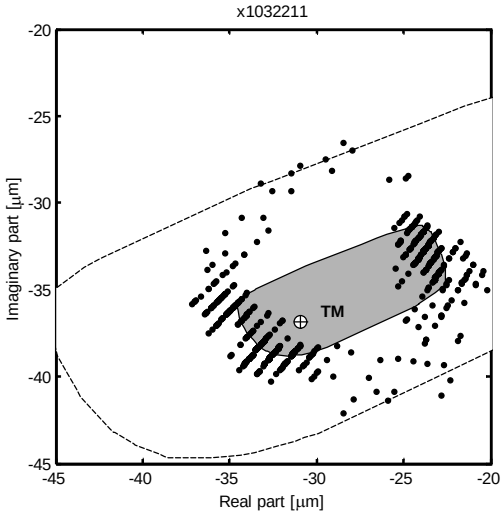

Figure 31. Case 1: acceptance regions defined by means of bagplot.

7.1 Data clustering

Figure 32 shows the result of the clustering which has the higher cophenetic coefficient, see eq. (48) The implicit criterion that clustering seems to have applied is to discriminate the clusters on the basis of both amplitude and phase thresholds, and the automatic clustering can be deemed as acceptable, discriminate the different operating conditions and the "atoll" is split into two parts (figure 33). By comparing the 3D histogram of all the data in figure 27 with figure 33, a preliminary analysis indicates that cluster 1 has a camel-back shape, whilst cluster 2 can be considered near-normal.

The definition of the acceptance regions (figure 34, figure 35, figure 36 and figure 37) can be considered rather satisfactory for all the methods for cluster 2. For cluster 1, the estimation of the dispersion is rather good, but the Tukey's median gives a better estimation of the location as results by considering 3D histogram of figure 27. Moreover, the good performances of circle sectors can be

ascribed to the favourable “geometry” and position with respect to the axis origin of the data cloud clusters, which are suitable to be inscribed in circle sectors.

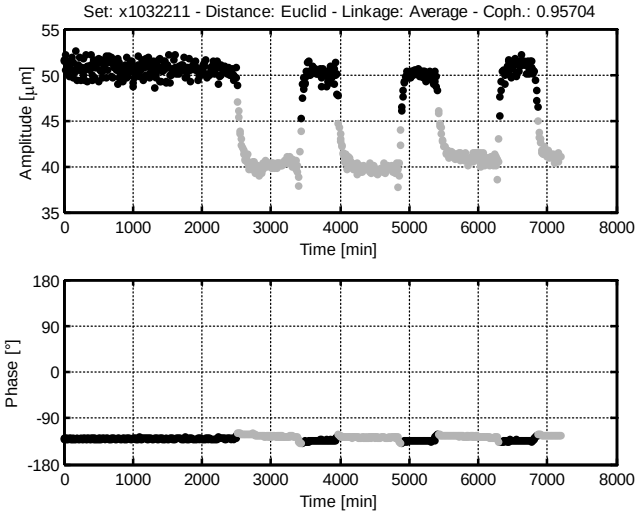

Figure 32. Case 3: time history with the data grouped into two clusters. Cluster 1 is indicated by solid black dots, cluster 2 by solid grey dots.

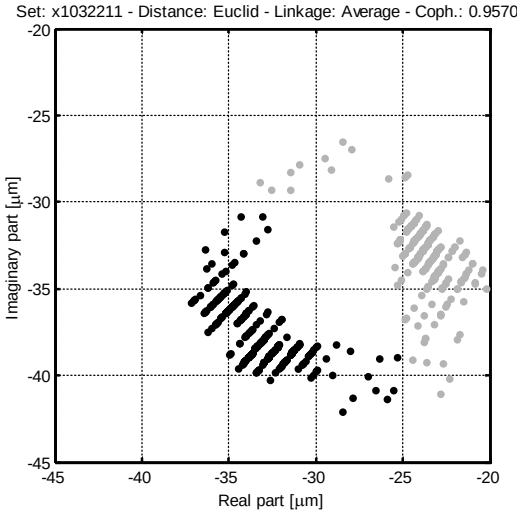

Figure 33. Case 3: polar plot of the two data cluster. Cluster 1 is indicated by solid black dots, cluster 2 by solid grey dots.

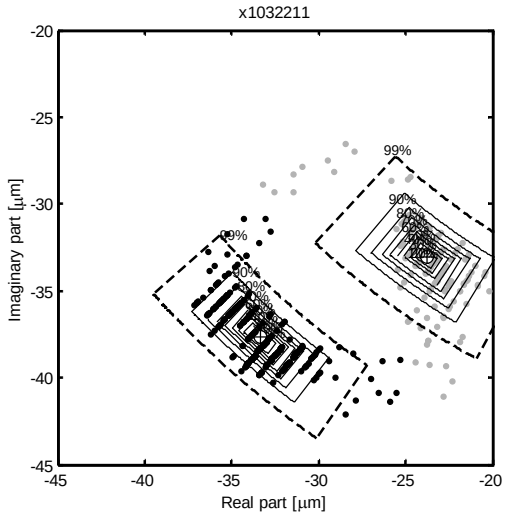

Figure 34. Case 3: acceptance regions defined by means of circle sectors for both clusters.

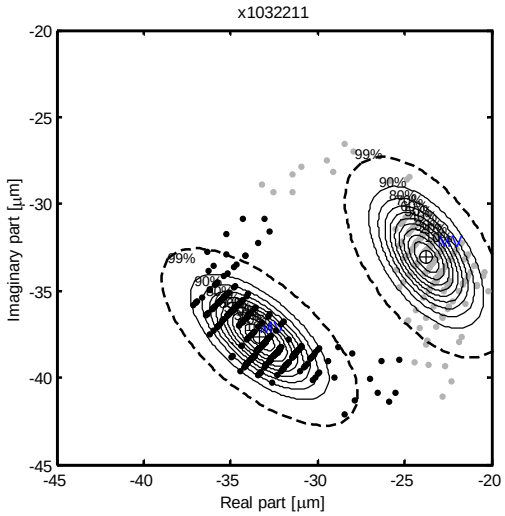

Figure 35. Case 3: acceptance regions defined by means of ellipses for both clusters.

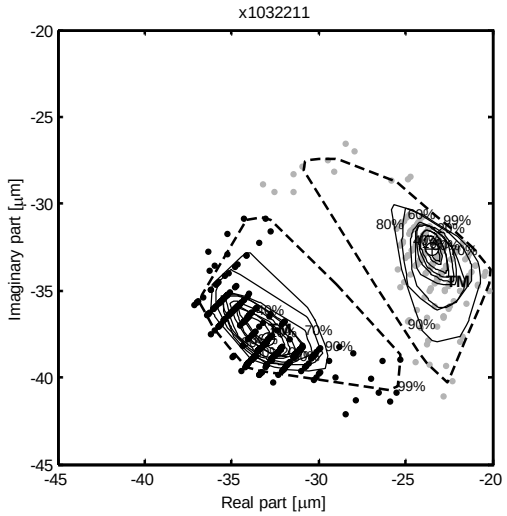

Figure 36. Case 3: acceptance regions defined by means of percentile regions for both clusters.

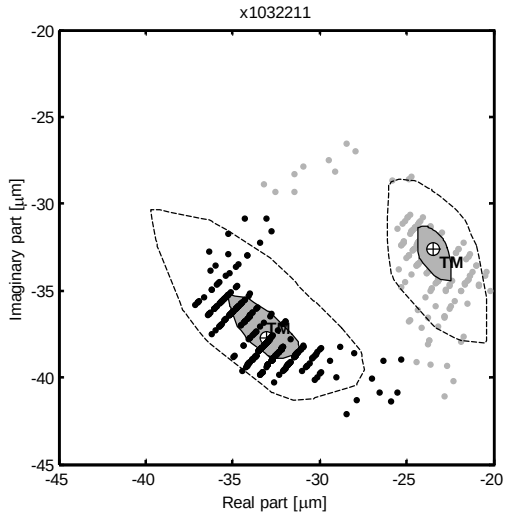

Figure 37. Case 3: acceptance regions defined by means of bagplot for both clusters.

9 CONCLUSIONS

A review on existing parametric techniques applied to bivariate data has been presented in the first part of the paper in order to have analytical tools to define acceptance regions of vibration vectors for condition monitoring systems in rotating machinery. Then, new non-parametric bivariate data description methods have been introduced. They are based on the general concept of data depth and a method to give a ranking to the data is presented. Two possible criteria to define acceptance regions using data depth are discussed: the bagplot and the percentile regions. Obviously, the methods presented here are suitable to all bivariate data, not only to vibration in rotating machinery. Some experimental cases, coming from combined cycle power plant have been presented in order to evaluate the performances and the robustness in the estimation of the acceptance region of the different methods. Two of the experimental cases present also strong seasonal behaviour depending on the operating conditions and loads. The analyses were performed either pre-processing or not the data by means of clustering techniques. The results of the analyses indicates that parametric methods are generally less robust than percentile regions, while the bagplot definition given in literature have to be tuned before being considered acceptable in this field of application. Another interesting result is that percentile regions are rather robust also in case of data presenting strong seasonal behaviour. Since calculation algorithms for non-parametric methods, for percentile regions in particular, are now available and computers are sufficiently powerful, their application into on-field condition monitoring systems will by very useful and limit the errors in alarm signals.

REFERENCES

[1] Bently Nevada Corporation, 3360/61 Dual Vector Monitor, 1999, pp. 1-7.

[2] Bently Nevada Corporation, Shaft Crack Detection using the Acceptance Region, *Orbit*, 1987, August, p. 17.

[3] Kaplan R., The power of Position, *Machine, Plant & Systems Monitor*, 1999, March/April, pp. 14-19.

[4] Puri D., Pyne T., Tipre M.V. and Chandajkar A.S., Refinery Fuel Gas Compressor – A Case Study on Vibration Diagnostics & Solution, *Proc. of HIMER National Conference*, 2001, pp. 1-12.

[5] Barnett V., The Ordering of Multivariate Data, *Journal of Royal Statistical Society - Series A*, (1976), 139, pp. 319-354.

[6] Liu R.Y., Parelius J.M. and Singh K., Multivariate Analysis by Data Depth: Descriptive Statistics, graphics and Inference, *The Annals of Statistics*, (1999), 27(3), pp. 783-858.
[7] Hwang J., Jorn H. and Kim J., On the Performance of Bivariate Robust Location Estimators under Contamination, *Computational Statistics & Data Analysis*, (2004) 44, pp. 587-601.
[8] Huber P.J., *Robust Statistics*, (1981), Wiley, New York.
[9] Hampel F.R, Ronchetti E.M, Rousseeuw P.J. and Stahel W.A., *Robust Statistics: The Approach Based On Influence Functions*, (1986), Wiley, New York.
[10] Collins J.R. Robustness Comparisons of Some Classes of Location Parameter Estimators, *Ann. Inst. Statist. Math.*, (2000) 52(2), pp. 351-366.
[11] Ma Y. and Genton M.G., Highly Robust Estimation of Dispersion Matrices, *Journal of Multivariate Analysis*, (2001) 78(1), pp. 11-36.
[12] Zani, S., Analisi dei dati statistici, Vol. II, osservazioni multidimensionali, ISBN 88-14-07902-1, Giuffrè Editore, Milano, 2000 (in Italian).
[13] Mahalanobis P.C., On the Generalized Distance in Statistics, *Proc. of Nat. Acad. Sci. India*, (1936) 12, pp. 49-55.
[14] Oja H., Descriptive Statistics for Multivariate Distributions, *Statist. Probab. Lett.*, (1983) 1, pp. 327-332.
[15] Buekenhout, F. and Parker, M. The Number of Nets of the Regular Convex Polytopes in Dimension, *Disc. Math*. (1998) 186, 69-94.
[16] Liu R., On a Notion of Data Depth Based on Random Simplices, *Ann. Statist.*, (1990) 18, pp. 405-414.
[17] Zuo Y., Cui H. and Young D., Influence function and maximum bias of projection depth based estimators, *Ann. Statist.* (2004) 32(1), pp. 189–218.
[18] Donoho D.L. and Gasko M., Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness, *Ann. Statist.*, (1992) 20(4), pp. 1803-1827.
[19] Struyf A. and Rosseeuw P.J., Halfspace Depth and Regression Depth Characterize the Empirical Distribution, *Journal of Multivariate Analysis*, (1999) 69, pp. 135-153.
[20] Tukey J.W., Mathematics and Picturing of Data, *Proc. of International Congress of Mathematicians*, Vancouver, (1975), pp. 523-531.
[21] Serfling R., Generalized Quantile Processes Based on Multivariate Depth Functions, with Applications in Nonparametric Multivariate Analysis, *Journal of Multivariate Analysis*, (2002), 83, pp. 232-247.
[22] Zuo Y. and Serfling R., Nonparametric Notions of Multivariate "Scatter Measure" and "More Scattered" Based on Statistical Depth Functions, *Journal of Multivariate Analysis*, (2000), 75, pp. 62-78.
[23] Zuo Y. and Serfling R., On the Performance of Some Robust Nonparametric Location Measures Relative to a General Notion of Multivariate Symmetry, *Journal of Statistical Planning and Inference*, (2000), 84, pp.55-79.
[24] Rousseeuw P.J., Ruts I. and Tukey J.W., The Bagplot: a Bivariate Box Plot, *The American Statistician*, (1999) 53(4), pp. 382-387.
[25] Bachschmid N., Pennacchi P. and Vania A., Introducing More Accurate Criteria to Define Acceptance Regions in Vibrational Behaviour Analysis, in *Condition Monitoring and Diagnostic Engineering Management*, Om P. Shrivastav, Basim Al-Najjar, Raj B. K. N. Rao Editors, Växjö University Press, Växjö, Sweden, 2003, pp. 749-758.
[26] Pennacchi P. and Vania, A., On the Use of the "Bagplot" Representation for a Sensitivity Analysis on the Effects of Bearing Dynamic Stiffness Variations on the Vibrations of a Steam Turbine, *Proc. of ISCORMA-2*, Gdańsk, (2003), on CD-ROM.
[27] Ruts I. and Rosseeuw P.J., Computing Depth Contours of Bivariate Point Clouds, *Computational Statistics & Data Analysis*, (1996) 23, pp. 153-168.
[28] Rosseeuw P.J. and Ruts I., Constructiong the Bivariate Tukey Median, *Statistica Sinica*, (1998) 8, pp. 827-839.

[29] Zhang, S., Ganesan, R. and A Xistris, G. D., Self-organising Neural Networks for Automated Machinery Monitoring Systems, *Mechanical Systems and Signal Processing*, (1996) 10(5), pp. 517-532.
[30] Hubert L.J., Hierarchical Cluster Analysis, in *Encyclopedia of Statistical Sciences*, Eds. S. Kotz, N.L. Johnson, C.B. Read, N. Balakrishnan and B. Vidakovic, Vol. 3 (1983), John Wiley & Sons, Ltd, 623-630.
[31] Anonymous, Cophenetic Matrix, in *Encyclopedia of Statistical Sciences*, Eds. S. Kotz, N.L. Johnson, C.B. Read, N. Balakrishnan and B. Vidakovic, supplement volume (1989), John Wiley & Sons, Ltd, 35-36.
[32] Kaufman L. and Rousseeuw P.J., *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley, 1990.

APPENDIX 1

Starting from the continuous probability density function of the normal bivariate distribution given by eq. (7), the acceptance regions are obtained in the polar plane as:

$$z = f\left(x_1, x_2\right) = \text{constant} \tag{29}$$

and passing to natural logarithms:

$$\frac{x_1^2}{\sigma_1^2} - \frac{2\,\rho_{12}\,x_1\,x_2}{\sigma_1\,\sigma_2} + \frac{x_2^2}{\sigma_2^2} = C^2 \tag{30}$$

Eq. (30) is the analytical expression (parametric in the constant C) of the ellipses that can be projected on the polar plane x_1x_2. The percentage of the data enclosed in each of them is given by the integral of the probability density function over the region itself. The centre co-ordinates of the ellipses is given by the mean values of x_1 and x_2, i.e. (μ_1, μ_2), while it is necessary to define the value of the constant C, in order to have an ellipse containing a certain percentage of the data. This is possible by considering that the surface A of a generic ellipse is:

$$A = \pi\,C^2\sqrt{\sigma_1^2\,\sigma_2^2\left(1-\rho_{12}^2\right)} \tag{31}$$

Differentiating eq. (31), we have:

$$dA = 2\,\pi\,C\sqrt{\sigma_1^2\,\sigma_2^2\left(1-\rho_{12}^2\right)}\,dC \tag{32}$$

The height of the probability density function over surface dA is given by:

$$h = \frac{1}{2\,\pi\sqrt{\sigma_1^2\,\sigma_2^2\left(1-\rho_{12}^2\right)}}\,e^{-C^2/2} \tag{33}$$

so that the volume included over the plane points that are exterior to the ellipse is:

$$V = \int_{C^*}^{\infty} 2\,\pi\,C\sqrt{\sigma_1^2\,\sigma_2^2\left(1-\rho_{12}^2\right)}\,\frac{1}{2\,\pi\sqrt{\sigma_1^2\,\sigma_2^2\left(1-\rho_{12}^2\right)}}\,e^{-C^2/2}\,dC = e^{-C^2/2} \tag{34}$$

Eq. (34) is also the probability that some data are outside of the ellipse:

$$P\left[\frac{x_1^2}{\sigma_1^2} - \frac{2\,\rho_{12}\,x_1\,x_2}{\sigma_1\,\sigma_2} + \frac{x_2^2}{\sigma_1^2} > C^2\right] = e^{-C^2/2} \tag{35}$$

while the probability that the data are inside of the ellipse is:

$$P\left[\frac{x_1^2}{\sigma_1^2}-\frac{2\,\rho_{12}\,x_1\,x_2}{\sigma_1\,\sigma_2}+\frac{x_2^2}{\sigma_2^2}\le C^2\right]=1-e^{-C^2/2} \tag{36}$$

Solving eq. (36) for C, the relation between the parameter C and the percentage of the data included is obtained:

$$C=\sqrt{\ln\left(\frac{1}{1-P}\right)^2} \tag{37}$$

APPENDIX 2

Given the experimental data organized in a $m\times 2$ matrix **X**, composed by m row vectors $\mathbf{x}_j$, $j\in(1,...,m)$, of real x_{j1} (1st column) and imaginary x_{j2} (2nd column) part of a data sample, the various distances between vector $\mathbf{x}_r$ and $\mathbf{x}_s$ are defined as follows:
Euclidean distance:

$$d^2_{Euclidean}(r,s)=(\mathbf{x}_r-\mathbf{x}_s)(\mathbf{x}_r-\mathbf{x}_s)^{\mathrm{T}} \tag{38}$$

Standardized euclidean distance:

$$d^2_{Seuclidean}(r,s)=(\mathbf{x}_r-\mathbf{x}_s)\,\mathbf{D}^{-1}\,(\mathbf{x}_r-\mathbf{x}_s)^{\mathrm{T}} \tag{39}$$

where **D** is the diagonal matrix of the variance σ_j^2 of j-th variable over m.
Mahalanobis distance:

$$d^2_{Mahalanobis}(r,s)=(\mathbf{x}_r-\mathbf{x}_s)^{\mathrm{T}}\,\mathbf{V}^{-1}\,(\mathbf{x}_r-\mathbf{x}_s) \tag{40}$$

where **V** is the sample covariance matrix.
City block metric:

$$d_{Cityblock}(r,s)=\sum_{j=1}^{2}\left|x_{rj}-x_{sj}\right| \tag{41}$$

Minkowski metric:

$$d_{Minkowski}(r,s)=\left(\sum_{j=1}^{2}\left|x_{rj}-x_{sj}\right|^{p}\right)^{1/p} \tag{42}$$

Note that in case of p = 1, Minkowski metric coincides to City block metric, of p = 2 to euclidean distance, so in the paper we consider p = 3.
Distances calculated by means of eqs. (38) to (42) for all the $m(m-1)/2$ pairs of points are then organized in a vector **Y**. The distance information contained in **Y** is used first to link pairs of objects that are close together into binary clusters and then to link these newly formed clusters to other objects to create bigger clusters until all the objects in the original data set are linked together in a hierarchical tree called *dendrogram* that has on the horizontal axis the indices j of the objects in the original data set the incremental distance between the object pairs on the vertical axis. The link pairs are ordered in matrix **Z** of $(m-1)\times 3$ in which the columns 1 and 2 are the indexes of the points and column 3 is the distance.
Given n_R the number of points in cluster R, n_S the number of points in cluster S, x_{Ri} the i-th point in cluster R and x_{Sj} the j-th point in cluster S. Five algorithms have been used to generate the gerarchical cluster tree information.

Single linkage, or nearest neighbour, algorithm uses the shortest distance between objects in the two groups:

$$lk_{single}(R,S) = \min\left(d(x_{Ri}, x_{Sj})\right), \quad i \in (1,...,n_R),\ j \in (1,...,n_S) \tag{43}$$

Complete linkage, or furthest neighbour, algorithm uses the longest distance between objects in the two groups:

$$lk_{complete}(R,S) = \max\left(d(x_{Ri}, x_{Sj})\right), \quad i \in (1,...,n_R),\ j \in (1,...,n_S) \tag{44}$$

Average linkage algorithm uses the average distance between all pairs of objects in the two groups:

$$lk_{average}(R,S) = \frac{1}{n_R\, n_S}\sum_{i=1}^{n_R}\sum_{j=1}^{n_S} d(x_{Ri}, x_{Sj}) \tag{45}$$

Centroid linkage algorithm uses the distance between the centroids of the two groups:

$$lk_{centroid}(R,S) = d(\overline{x}_R, \overline{x}_S), \quad \overline{x}_R = \frac{1}{n_R}\sum_{i=1}^{n_R} x_{Ri}, \overline{x}_S = \frac{1}{n_S}\sum_{j=1}^{n_S} x_{Sj} \tag{46}$$

Ward linkage uses the incremental sum of squares; that is, the increase in the total within-group sum of squares as a result of joining groups *R* and *S*:

$$lk_{Ward}(R,S) = \frac{n_R\, n_S\, d^2_{centroid}(R,S)}{n_R + n_S} \tag{47}$$

The within-group sum of squares of a cluster is defined as the sum of the squares of the distance between all objects in the cluster and the centroid of the cluster.
An example is of dendrogram for data of Case 3 using Euclidean distance eq. (38) and average linkage eq. (45) is shown in figure 38. Due to the high number of data points, unfortunately the object indexes are not visible.
The evaluation of the validity of the cluster information generated by the link algorithms of eqs. (43) to (47) is to compare it with the original proximity data calculated using the distances in eqs. (38) to (42). If the clustering is valid, i.e. clusters are consistent, a strong correlation have to exist. In the clustering literature, this fact is evaluated by means of the construction of two matrices, the so called *dissimilarity matrix* and *cophenetic matrix*, and by the calculation of the correlation coefficient between the entries of the two matrices. The full explanation is far from the scope of the paper, so only the coefficient, which is called *cophenetic coefficient c*, is introduced and it can be calculated as:

$$c = \frac{\sum_{i<j}\left(Y_{ij} - \overline{y}\right)\left(Z_{ij} - \overline{z}\right)}{\sqrt{\sum_{i<j}\left(Y_{ij} - \overline{y}\right)^2 \sum_{i<j}\left(Z_{ij} - \overline{z}\right)^2}} \tag{48}$$

where Y_{ij} is the distance between objects *i* and *j* in **Y**, Z_{ij} is the distance between objects *i* and *j* in the last column of matrix **Z** and $\overline{y}$ and $\overline{z}$ are the respective averages. The closer the *c* value is to 1, the better is the clustering solution. More details about the cophenetic matrix and coefficient can be found in a concise form in [30] [31] and in a detailed form in [32].
For Case 2 and 3 of the paper, all the combinations of distance and linkage algorithms have been used to find out the higher cophenetic coefficient and the corresponding clustering and dendrogram. Since two operating conditions are looked for in the paper, the two required cluster are build up splitting the two higher leaves of the dendrogram (ideally with the dash-dot line of figure 38).

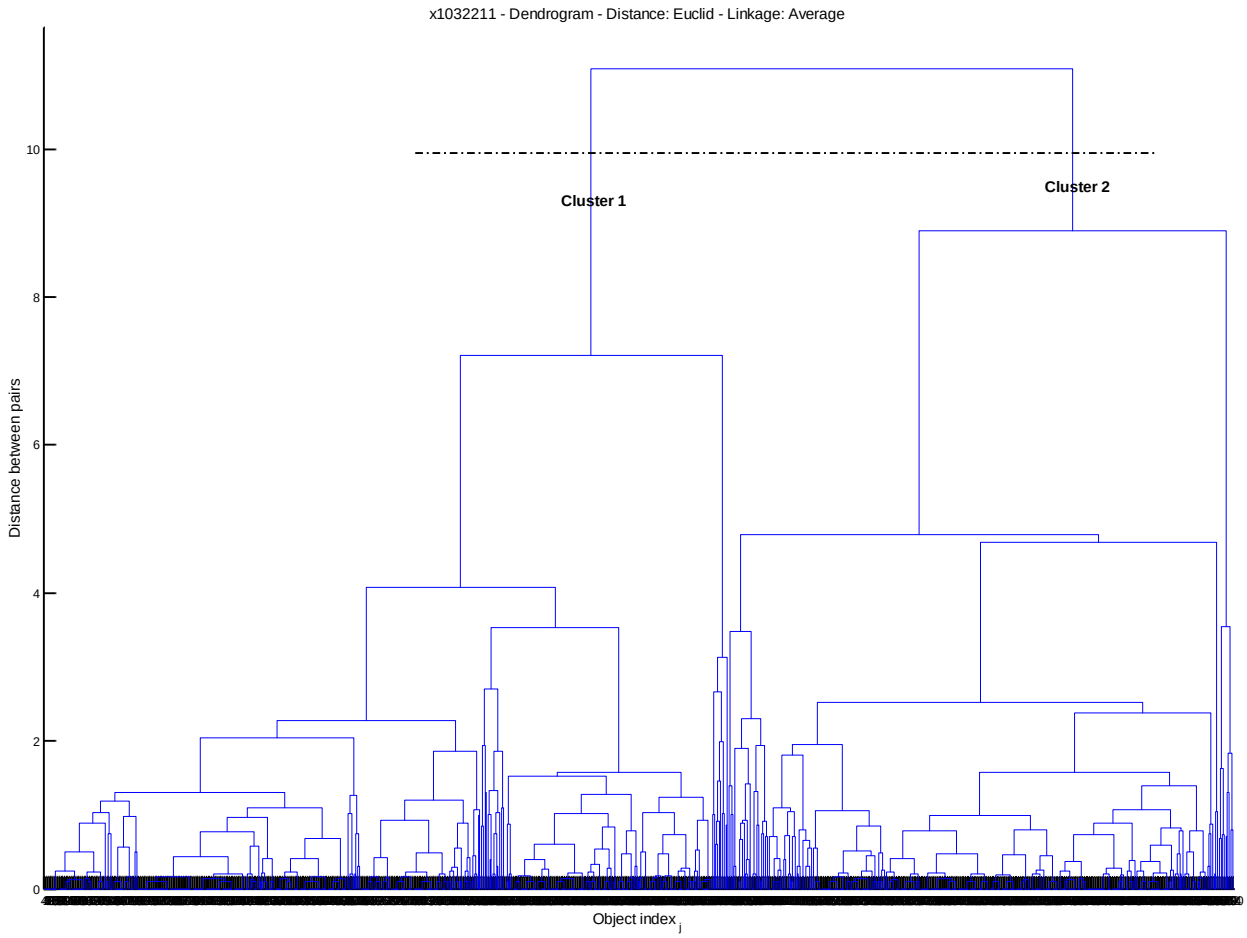

Figure 38. Dendrogram for Case 3 data clustering.