Tape surfaces characterization with persistence images

Tarek Frahi, Clara Argerich, Minyoung Yun, Antonio Falco, Anaïs Barasinski, Francisco Chinesta

Tape surfaces characterization with persistence images

Tarek Frahi[1,*], **Clara Argerich**[2], **Minyoung Yun**[2], **Antonio Falco**[3], **Anais Barasinski**[4] **and Francisco Chinesta**[1]

[1] PIMM & ESI Group International Chair, Arts et Metiers Institute of Technology, 151 boulevard de l'Hôpital, 75013 Paris, France

[2] PIMM, Arts et Metiers Institute of Technology, 151 boulevard de l'Hôpital, 75013 Paris, France

[3] ESI-CEU International Chair, Universidad Cardenal Herrera-CEU, San Bartolome 55, 46115 Alfara del Patriarca, Valencia, Spain

[4] Université de Pau et des Pays de l'Adour, E2S UPPA, CNRS, IPREM, Pau, France

* **Correspondence:** Email: tarek.frahi@ensam.eu; Tel: +33(0)144246299.

Abstract: The aim of this paper is to leverage the main surface topological descriptors to classify tape surface profiles, through the modelling of the evolution of the degree of intimate contact along the consolidation of pre-impregnated preforms associated to a composite forming process. It is well-known at an experimental level that the consolidation degree strongly depends on the surface characteristics (roughness). In particular, same process parameters applied to different surfaces produce very different degrees of intimate contact. It allows us to think that the surface topology plays an important role along this process. However, solving the physics-based models for simulating the roughness squeezing occurring at the tapes interface represents a computational effort incompatible with online process control purposes. An alternative approach consists of taking a population of different tapes, with different surfaces, and simulating the consolidation for evaluating for each one the progression of the degree of intimate contact –DIC– while compressing the heated tapes, until reaching its final value at the end of the compression. The final goal is creating a regression able to assign a final value of the DIC to any surface, enabling online process control. The main issue of such an approach is the rough surface description, that is, the most precise and compact way of describing it from some appropriate parameters easy to extract experimentally, to be included in the just referred regression. In the present paper we consider a novel, powerful and very promising technique based on the topological data analysis –TDA– that considers an adequate metrics to describe, compare and classify rough surfaces.

Keywords: surface characterization; ATP composites manufacturing; tape surfaces; topological data analysis; persistence; homology; *Code2Vect*; classification; regression; random forests; machine learning

1. Introduction

Among composite forming processes for manufacturing structural parts based on the consolidation of pre-impregnated preforms, e.g., sheets, tapes, etc., the automated tape placement (ATP) appears as one of the most interesting techniques due to its versatility and its in-situ consolidation, thus avoiding the use of autoclave. In particular, to obtain the cohesion of two thermoplastic layers two specific physical conditions are needed (a) an almost perfect contact (intimate contact) and (b) a temperature enabling molecular diffusion within the process time window, while avoiding thermal degradation. To reach this goal, a tape is placed and progressively bonded to the substrate consisting of the tapes previously laid-up. Due to the low thermal conductivity of usual resins, an intense local heating is usually considered (laser, gas torches, etc.) in conjunction with a local pressure applied by the consolidation roller moving with the heating head, as sketched in Figure 1. Thus, the two main factors to ensure the intimate contact at the plies surfaces are pressure and heat. Intimate contact is required to promote the molecular diffusion. In this process heat plays a double role, on one hand it enhances molecular mobility and on the other hand, the decrease of the material viscosity with the temperature increase, facilitates the squeeze flow of the heated asperities located on the ply surfaces under the compression applied by the consolidation roller.

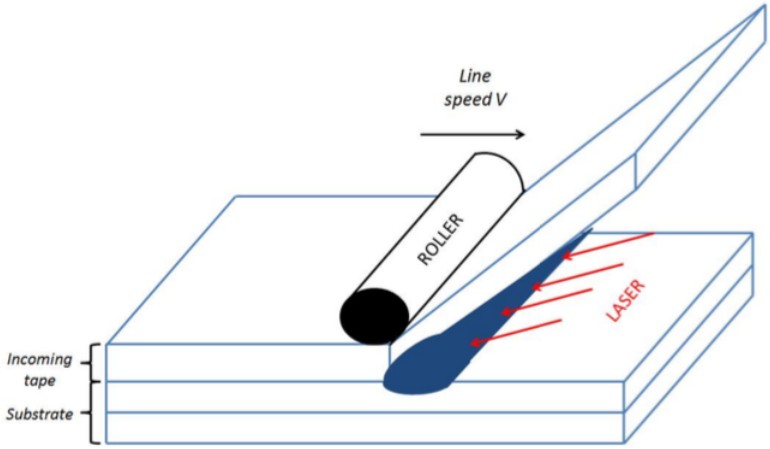


Figure 1. The automated tape placement (ATP).

The numerical model of ATP was introduced in [1] by using the so-called Proper Generalized Decomposition (PGD) [2–6]. The separated representation involved in the PGD enables the 3D high-resolution solution of models defined in degenerated domains where at least one of their characteristic dimensions remains much smaller than the others and also constructing solutions of parametric models where the model parameters are considered as extra-coordinates [7, 8].

Physical modelling and simulation for Automated Tape Placement (ATP) have been proposed in [9] to study the influence of material and process parameters, while consolidation modelling and sPGD-based non-linear regression have been used in [10] to identify the main surface descriptors for a comprehensive characterization of the tape surfaces.

The present paper revisits first the consolidation modeling and its high resolution simulation, enabling the evaluation of the time evolution of the degree of intimate contact –DIC– when two rough

surfaces are put in contact, heated and compressed.

In the present work, as we are addressing tapes involved in the ATP process sketched in Figure 1, the roughness squeezing mainly occurs along the transverse direction (the one related to the tape width) induced by the roller compression. Thus, the flow occurs in the transverse section in which the surface reduces to a one-dimensional curve (the so-called surface profile).

In order to extract a concise and complete description of rough surfaces, topological data analysis –TDA– [11–14] is then introduced, with all the main techniques that it involves, in particular the so-called persistence diagrams and images.

Then, the persistence images are considered for classifying surfaces or for acting as surface descriptors that will be involved in a regression relating them to a quantity of interest –QoI–, in the present study the final DIC reached in the consolidation process, enabling real-time decision making.

2. Consolidation modelling

In our recent works [9, 15, 16] we proposed simulating the consolidation on the real surfaces instead of the, sometimes too crude, approximations of them based on the use of fractal representations or the ones based on the description of asperities from the use of rectangular elements [17, 18].

As sketched in Figure 2, a Haar-based wavelet representation [9] of a rough surface results in a multi-scale sequence of rectangular patterns, from the coarse scale (level 0) to the finest one (level 8) that constitutes a quite precise representation of the considered surface (the one illustrated in Figure 2). The smoother is the surface, the less levels in the description are required.

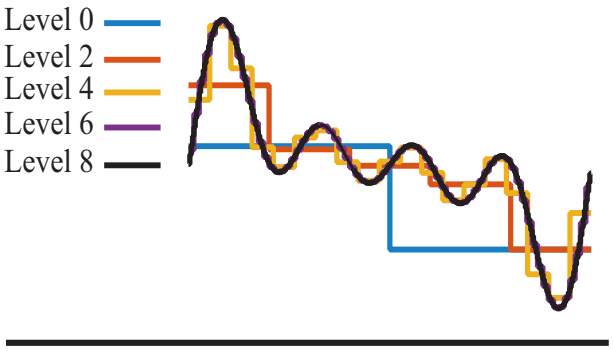


Figure 2. Simple surface representation using Haar wavelets.

The advantage of such a representation consisting of hierarchical rectangles is double: (i) from one side it facilitates the high-resolution of the thermal problem while accounting for all the interface details and their time evolution; and on the other (ii) it allows squeezing the rectangles of a certain level (from the finest level to the coarser one) while retaining the lubrication hypotheses, fact that simplifies significantly the flow modeling and the calculation of the interface evolution when squeezing the asperities. Both aspects are revisited below.

1. As soon as the rough surface profile is represented in a step-wise way consisting of R rectangular elements, with each rectangle r having a length l_r and a height h_r, assumed centered at x_r, each rectangle can be expressed by its characteristic function in a separated form $\chi_r(x,z) = \mathcal{L}_r(x)\mathcal{H}_r(z)$,

with $\mathcal{L}_r(x)$ and $\mathcal{H}_r(z)$ given respectively by Eqs 1 and 2,

$$\mathcal{L}_r(x) = \begin{cases} 1 & \text{if } x \in (x_r - l_r/2, x_r + l_r/2) \\ 0 & \text{elsewhere} \end{cases}, \tag{1}$$

and

$$\mathcal{H}_r(z) = \begin{cases} 1 & \text{if } z \in (0, h_r) \\ 0 & \text{elsewhere} \end{cases}, \tag{2}$$

that allows expressing the conductivity at the interface level according to Eq 3,

$$\mathbf{K}(x,z) = \left(1 - \sum_{r=1}^{\mathrm{R}} \mathcal{L}_r(x)\mathcal{H}_r(z))\right)\mathbf{K}_c + \left(\sum_{r=1}^{\mathrm{R}} \mathcal{L}_r(x)\mathcal{H}_r(z))\right)\mathbf{K}_a, \tag{3}$$

where $\mathbf{K}_a$ and $\mathbf{K}_c$ represent the air and composites conductivities, with the former assumed isotropic and the last concerning the composite conductivity transverse components.
This separated representation of the thermal conductivity allows looking for a separated representation of the temperature field within the proper generalized decomposition –PGD– framework, according to Eq 4.

$$T(x,z) \approx \sum_{i=1}^{M} X_i(x)\, Z_i(z), \tag{4}$$

that by decoupling the 2D heat equation solution into a sequence of 1D problems for computing the functions $X_i(x)$ and $Z_i(z)$ allows an extremely fine resolution as discussed in [9].
As the asperities squeezing progresses, the surface evolves and with it the height of the different rectangular elements. The conductivity separated representation must be updated and the thermal problem solved again to compute the updated temperature field (4).

2. As soon as the temperature field is available, the polymer viscosity can be evaluated and the asperities will flow under the applied pressure. As commented, the description of the surface by using rectangular elements, with their characteristic length $\bar{l}$ much larger that its characteristic height $\bar{h}$, i.e. $\bar{l} \gg \bar{h}$, makes possible the use of the lubrication hypotheses, widely addressed in our former works [16].
   The surface updating procedure is quite simple. We consider all the compressed rectangles, and solve in them the squeeze flow model, while assuming that the pressure in all the other elements vanishes. As soon as the pressures are available in all the rectangles that are being compressed, the velocity field and more precisely the flow rates can be obtained at the lateral boundaries. The fluid leaving each rectangular element that is being compressed is transferred to the neighbor rectangular element that increases its height accordingly in order to ensure the mass conservation.

As it can be noticed, this procedure allows unimaginable level of accuracy, however, despite of the speed-up that separated representation offers, its use online for predicting the thermal and flow coupled problem for any incoming rough tape is not an option.

3. Surface descriptors based on homology persistence

In this section we introduce the data and methods used, in particular the TDA and its related procedures (persistent diagrams and images), even if other approaches exist, e.g. [10, 19].

The proposed methodology proceeds in three main stages:

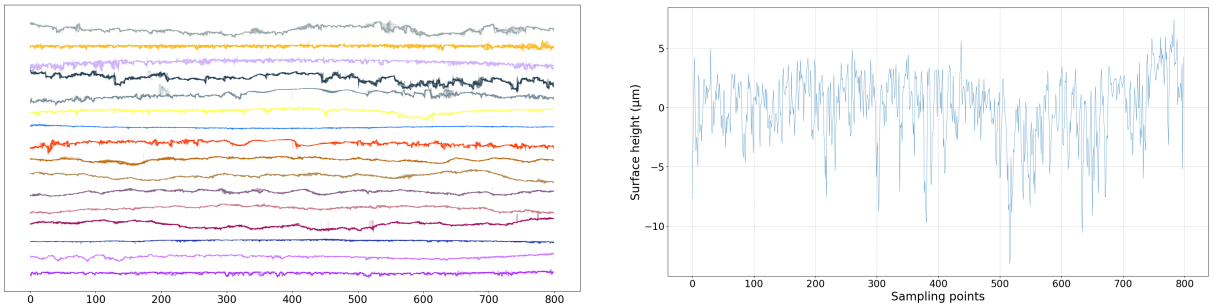
1. Processing the surface profiles data;
2. Compute persistent diagrams and images;
3. Construct the regressions relating the surface topological descriptors and the quantities of interest –QoI–, concretely the DIC.

3.1. Processing the surface profiles data

In order to classify the main surface descriptors of a tape surface, we will consider samples scanned with a 3D non contact profilometer, with a 3.5 um resolution and where each sample has a length of approximately 3 mm (along the tape width). A set of 1359 surface profiles were extracted from 16 different pre-impregnated composite tapes provided by different customers using different impregnation process, each one represented by 800 measured data points $\left\{S_\ell^{(k)} : 1 \le \ell \le 800,\ 1 \le k \le 1359\right\}$.

The main goal is to give a procedure to construct a classification $C(S)$, that is, a map ensuring $C(S^{(k)}) = i$ if and only if $S^{(k)}$ was extracted from the tape i, with $i = \{1, 2, \ldots, 16\}$.

In particular, to facilitate data comparison the profiles are corrected by subtracting the average height. Figure 3 depicts the different surfaces in each of the 16 classes, as well as normalized profile.



(a) **(b)**

Figure 3. Surface profiles data. (a) The 16 surface classes, (b) corrected surface profile when subtracting its averaged height.

3.2. Persistence diagrams and images

The persistence diagram consists of a one-to-one local-minimum-local-maxixum pairing. To illustrate the procedure we consider a simple case of a profile described by 9 heights, $S = \{11, 14, 9, 7, 9, 7, 8, 10, 9\}$, that corresponds to the 9 data points depicted in Figure 4: $\{(0, 11), (1, 14), (2, 9), (3, 7), (4, 9), (5, 7), (6, 8), (7, 10), (8, 9)\}$.

Now, we consider the 4 local minimum: $\{(0, 11), (3, 7), (5, 7), (8, 9)\}$ and the only 3 local maximum: $(1, 14), (4, 9), (7, 10)$. We associate $(0, 11)$ to $(1, 14)$, then $(8, 9)$ to $(7, 10)$ and finally $(5, 7)$ to $(4, 9)$. The remaining local minimum $(3, 7)$ can not be paired to any other local maximum because all of them

have already been paired. The local-minimum-local-maximum paired heights constitutes the so-called persistence diagram $\mathcal{PD}(S)$, in our example $\mathcal{PD}(S) = \{(7,9),(9,10),(11,14)\}$, with the minimum of the pair representing the topological occurrence birth, whereas the associated maximum its death.

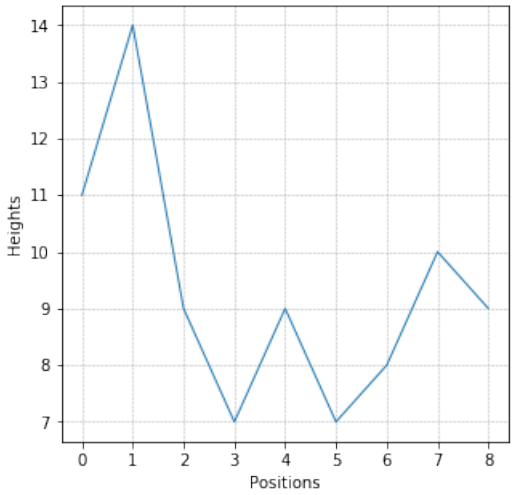


Figure 4. Profile consisting of 9 measured height at 9 positions.

In our example, consisting of the 9 data and the three topological occurrences composing, the associated persistence diagram is shown in Figure 5.

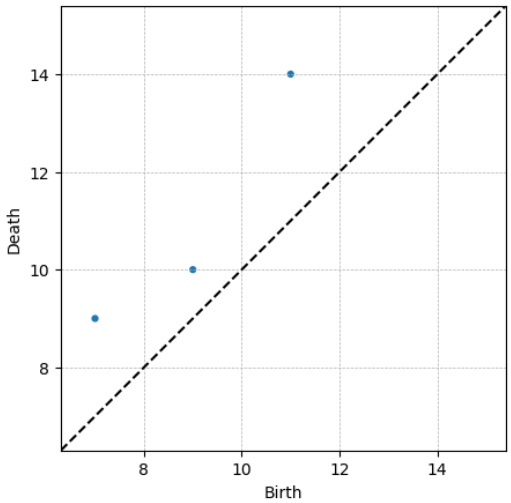


Figure 5. Persistence diagram $\mathcal{PD}(S)$.

In the two dimensional representation associated to the persistence diagram, each data point (x, y) verifies the relationship $y \geq x$ (the topological occurrence birth precedes its death) and then points locate above the bisector $x = y$. The topological representation provided by the persistence diagram offers a very concise description of any curve (e.g. surface profiles, time-series, etc.) and the change in their topological features as considered in [13, 14, 20, 21].

In order to use the persistence diagram and perform vectorial operations such as the ones required

in classification, we must transform the persistence diagram into a vectorial representation of it, the so-called *persistence image* [21, 22].

For that purpose, we first introduce the so-called *lifetime diagram* $\mathcal{T}(S)$ associated to the function $\mathcal{PD}(S)$, defined in Eq 5.

$$\mathcal{T}(S) = \{(x, y) \in \mathcal{PD}(S) \rightarrow (x, y - x) \in \mathbb{R}^2\}, \tag{5}$$

where $y - x$ represent the lifetime of the topological occurrence. In our example we have $\mathcal{T}(S) = \{(7, 2), (9, 1), (11, 3)\}$, that is illustrated in Figure 6.

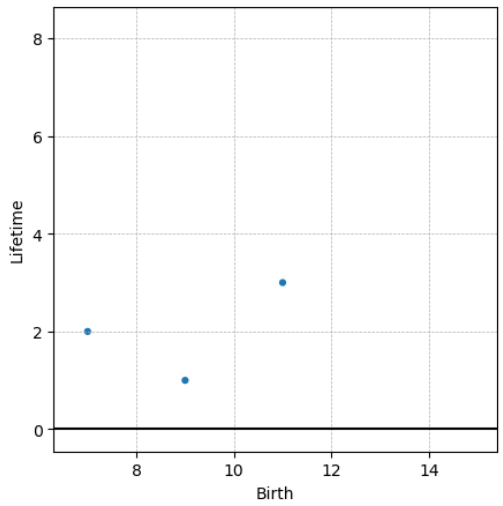


Figure 6. Lifetime diagram $\mathcal{T}(S)$ associated to $\mathcal{PD}(S)$.

Next, we will construct a persistent image as follows. We consider a continuous piecewise derivable non-negative weighting function (with $(x, y) \in \mathcal{T}(S)$, $w(x, 0) = 0$ and $w(x, y_{max}) = 1$, with $y_{max} = \max(y)$, that can be approximated by a linear function of the lifetime y, e.g. $w(x, y) = y/y_{max}$) and a bivariate normal distribution $g_{x,y}(u, v)$ centered at each point $(x, y) \in \mathcal{T}(S)$ and with its variance σ, $\sigma > 0$, scaling with the maximum of the lifetime diagram [21, 22], then we define the variable $\rho_S(u, v)$ expressed in Eq 6:

$$\rho_S(u, v) = \sum_{(x,y)\in\mathcal{T}(S)} w(x, y)\, g_{(x,y)}(u, v), \tag{6}$$

with $(u, v) \in \mathcal{D}$, with $\mathcal{D}$ a compact domain (for example the domain in which $\mathcal{T}(S)$ is defined).

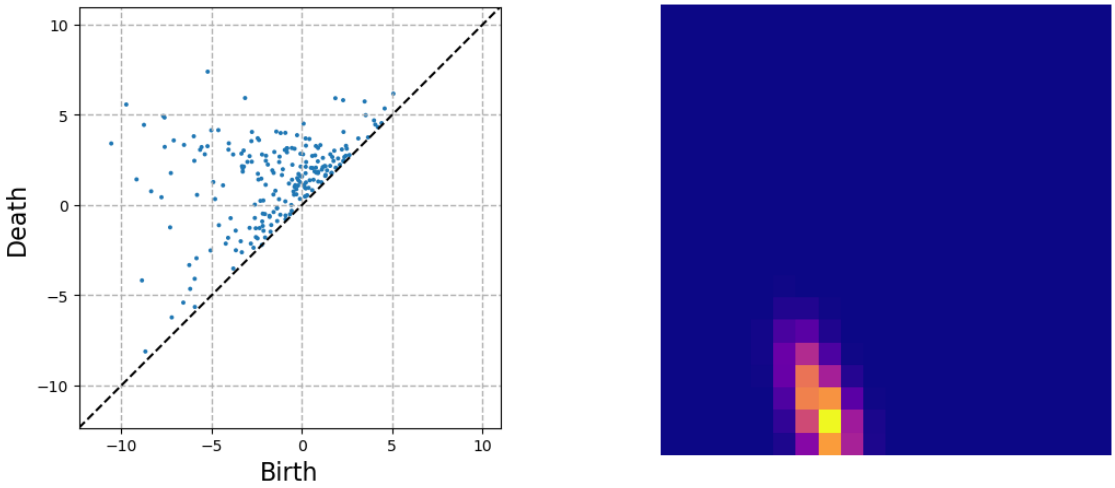
Now, the domain $\mathcal{D}$ is partitioned in a series of non-overlapping subdomains covering it, the so-called pixels P_i, with $\mathcal{D} = \cup_{i=1}^{\mathtt{P}} P_i$, and function $\rho_S(u, v)$ averaged in each of those pixels, that will define the *persistence image* $\mathcal{PI}(S)$. Thus each of the $\mathtt{P}$ pixels of the persistence image $\mathcal{PI}(S)$ takes the value given by Eq 7:

$$\mathcal{PI}_{P_i}(S) = \iint_{P_i} \rho_S(u, v)\, du\, dv. \tag{7}$$

As the profile that served to illustrate the different concepts contains too few topological occurrences, to illustrate what a persistence image resembles to, we consider a profile related to one of the measured rough surfaces S, compute the persistence diagram $\mathcal{PD}(S)$, then its associated

lifetime diagram $\mathcal{T}(S)$, and finally its persistence image $\mathcal{PI}(S)$. Figure 7 shows $\mathcal{PD}(S)$ and $\mathcal{PI}(S)$, the last employing 20×20 pixels, i.e. $\mathtt{P} = 400$ with a variance σ in the normal distribution $g_{x,y}(u, v)$ given by Eq 8:

$$\sigma = \frac{max_{(x,y)\in\mathcal{T}(S)}\{y\}}{20}. \tag{8}$$



(a) **(b)**

Figure 7. TDA analysis of a real rough surface. (a) Persistence diagram $\mathcal{PD}(S)$, (b) persistence image $\mathcal{PI}(S)$.

3.3. Images classification versus clustering

When applying the rationale just described to the 1359 rough surfaces $S^{(k)}$, $k = 1, \ldots, 1359$, we will obtain the associated 1359 persistence diagrams $\mathcal{PD}(S^{(k)})$, lifetime diagrams $\mathcal{T}(S^{(k)})$ and persistence images $\mathcal{PI}(S^{(k)})$, $k = 1, \ldots, 1359$.

Thus each surface produced a persistence image composed of $\mathtt{P} = 400$ pixels. These images are expected belonging to 16 different classes, the 16 families of composite tapes. Obviously, trying to proceed to that classification directly from the surface raw data $S^{(k)}$ seems a tricky issue because the proximity is not well defined when using a standard Euclidean metric. The same surface taken with a small shift will induce a significant difference. Metrics based on the topology seem more robust because the appealing associated invariance properties. Thus, more than trying to classify from the raw data, persistence images seem to be the right starting point.

3.3.1. Images classification

Image classification is a procedure to automatically categorize images into classes, by assigning to each image a label representative of its class. A supervised classification algorithm requires a training sample for each class, that is, a collection of data points whose class of interest is known. Labels are assigned to each class of interest. The classification is thus based on how close a new point is to each training sample. The Euclidean distance is the most common distance metric used in low dimensional data sets. The training samples are representative of the known classes of interest to the analyst.

In order to classify the persistence images we can use any state of the art technique. In our case we considered the Random Forest classification [23]. We train the random forest (consisting of 400 trees) by using 65% of the the persistence images (the remaining 35% serving to evaluate the classification performances), where a label was attached to each one, a label precisely specifying the family, among the 16 composites considered, to which it belongs.

With the trained random forest one expects, from a given persistence image, obtaining in almost real-time the family to which it belongs, of major interest in process control.

3.3.2. Images clustering

Unsupervised learning algorithms aim at finding unknown patterns in data sets without pre-existing labels. Clustering is used in unsupervised learning to group, or segment, data that has not been labelled, classified or categorized. It is based on the presence or absence of commonalities in each new piece of data. This approach also helps detect anomalous data points that do not fit into either group.

One of the most popular clustering techniques, the k-means, aims at partitioning the observations into k clusters in which each observation belongs to the cluster with the nearest mean or center [23]. The cluster center serves as a prototype of the cluster population. The observations are then allocated according to the criterion of minimizing the within-cluster variances, which is a squared Euclidean distance. The data can be then labelled according to their respective clusters (arbitrarily numbered).

To determine the optimal number of clusters we proceed as follows. For different values of k, k-means is trained with the whole data-set, and the data-labelled depending on the cluster to which each data belongs. Then, k-means is applied again but now with only 65% of the data. Then, for each data the cluster to which it belongs is compared to the label (cluster to which it belonged when all the data was employed in the k-means). A parametric variance analysis allows determining the optimal value of k, that in our case resulted as expected $k = 16$.

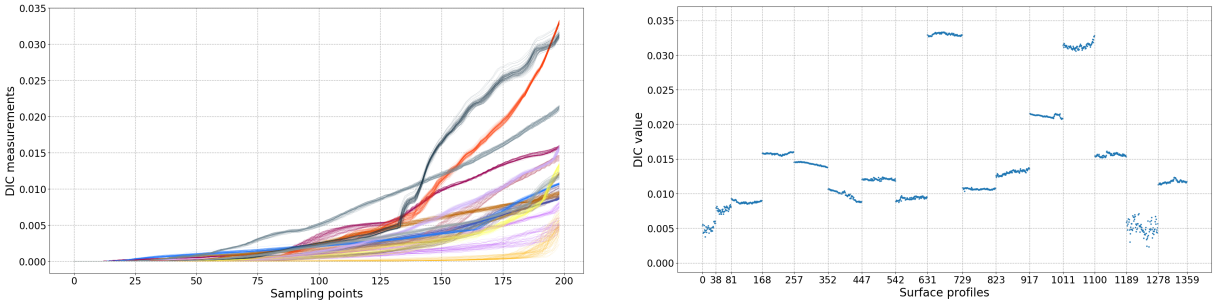
As soon as the best number of cluster is determined, $k = 16$ in our case, k-means proceeds with the whole data to generate the reference labels (cluster to which each data belongs). Then, the process repeats but now employing only 65% of the data. Finally we estimate for the remaining 35% of the data to which cluster it is associated and compare with its label to have an estimation of the clustering performances.

3.4. Predicting the degree of intimate contact

The consolidation process of all the available surfaces (1359) was simulated by using the PGD-based high-resolution solver described in Section 2. The evolution of the DIC, that is, the fraction of the surface in perfect contact, was evaluated at the different time steps. Figure 8a depicts the DIC evolution for the 1359 surfaces during the first 200 time steps of the consolidation process. As it can be noticed the dispersion of the DIC is quite small within each one of the 16 composite tapes (classes), however it exhibits large differences from one composite to another.

In what follows we are interested in the DIC prediction at the last time step (the number 200), that will consist of our quantify of interest –QoI– $\mathcal{O}$, that for each surface results in the values $\mathcal{O}^{(k)}$, $k = 1, \ldots, 1359$ depicted in Figure 8b.

Now, we are interested in constructing a regression for expressing the QoI, $\mathcal{O}$, as a function of the considered surface, the geometry of the last expressed through its persistence image.



(a) **(b)**

Figure 8. Simulated degree of intimate contact. (a) Simulated time evolution of the DIC, (b) DIC reached at time step 200 for each surface.

For that purpose we are considering two regression techniques: (i) the so-called *Code2Vect* [24] summarized in the Appendix, and (ii) the random forest.

- *Code2Vect* maps the surfaces described by the 400 values related to the pixels of their associated persistence images into another low dimensional vector space where the distance between any two points (representing two surfaces) scales with the QoI difference, that is, with respect to the difference of their DIC. However, as for usual nonlinear regression techniques, the complexity scales with the number of parameters involved in the regression, and here 400 seems a bit excessive with respect to the available data.
  For this reason, and prior to use de *Code2Vect* regression, the 1359 persistence images, each represented by 20×20 pixels, are first analyzed by using the principal component analysis – PCA– to remove linear correlations [23] where the two most significant modes were retained, and each persistence image described by its projection on both models. Thus, the reduction is impressive, each persistence image, and in consequence each surface, is now described from only two parameters. Then, the *Code2Vect* was employed to establish the regression between these two parameters and the quantity of Interest $\mathcal{O}$, the final DIC [24].
  Again, to evaluate the regression performances, *Code2Vect* was trained by using 80% of the data, and the remaining 20% served for evaluating the prediction performances.

- As previously indicated a regression based on the use of Random Forest [23] (using 400 trees) was considered, with 65% of the data used in the training and 35% for evaluating the prediction performances.

3.5. Models evaluation

For evaluating the model performances we consider different procedures:

- Confusion matrix
  The component (i, j) of the confusion matrix contains de number of surfaces that belonging to a class i are predicted belonging to class j. Obviously the classification is perfect when this matrix becomes diagonal.
- Classification scoring. Evaluating a classification model is determining how often labels are correctly or wrongly predicted for the testing samples. In other words, it is counting how many

times a sample is correctly or wrongly labelled into a particular class. We distinguish four qualities:

- TP (True Positive): the correct prediction of a sample into a class;
- TN (True Negative): the correct prediction of a sample out of a class;
- FP (False Positive): the incorrect prediction of a sample into a class;
- FN (False Negative): the incorrect prediction of a sample out of class.

These quantities are involved in the definition of different estimators of the model performances:

- The Precision (P) is the number of correct positive results divided by the number of all positive results, expressed by Eq 9:

$$P = \frac{TP}{TP + FP} \tag{9}$$

- The Recall (R) is the number of correct positive results divided by the number of all relevant samples, expressed by Eq 10:

$$R = \frac{TP}{TP + FN} \tag{10}$$

- The F1 score is the harmonic mean of precision and recall, expressed by Eq 11:

$$F1 = 2\frac{P \cdot R}{P + R} \tag{11}$$

- The Accuracy (A) is the number of correct predictions over the number of all samples, expressed by Eq 12:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

- Regression scoring
  We evaluate our regression prediction using the $\mathcal{R}^2$ coefficient, defined in Eq 13:

$$\mathcal{R}^2 = 1 - \frac{\sum_i^n \left(O_i^{true} - O_i^{pred}\right)^2}{\sum_i^n \left(O_i^{true} - \overline{O}^{true})\right)^2}. \tag{13}$$

  We also use the mean absolute percentage error $MAPE$, defined in Eq 14:

$$MAPE = \frac{100\%}{n} \sum_i^n \left| \frac{O_i^{true} - O_i^{pred}}{O_i^{true}} \right|, \tag{14}$$

  with best model having the closest MAPE to 0%.
- Features importance. In decision trees, every node is a condition on how to split values for a single feature, so that similar values of the dependent variable end up in the same set after the split. The condition is based on impurity, which in the case of classification problems is the Gini impurity or the information gain (entropy), while for regression trees it is the variance. So when training a tree, we can compute how much each feature contributes to decreasing the weighted impurity, and in the case of Random Forest, we are talking about averaging the decrease in impurity over all the trees [23]. Although this method is known to be statistically biased for categorical variables, it should not be affected in our case, as we only have homogeneous and continuous variables, 20×20 pixels images.

4. Results

In this section we provide the numerical results and evaluations associated to each of the previously introduced models: Random Forest classification, *k*-means clustering, *Code2Vect* and Random Forest regression.

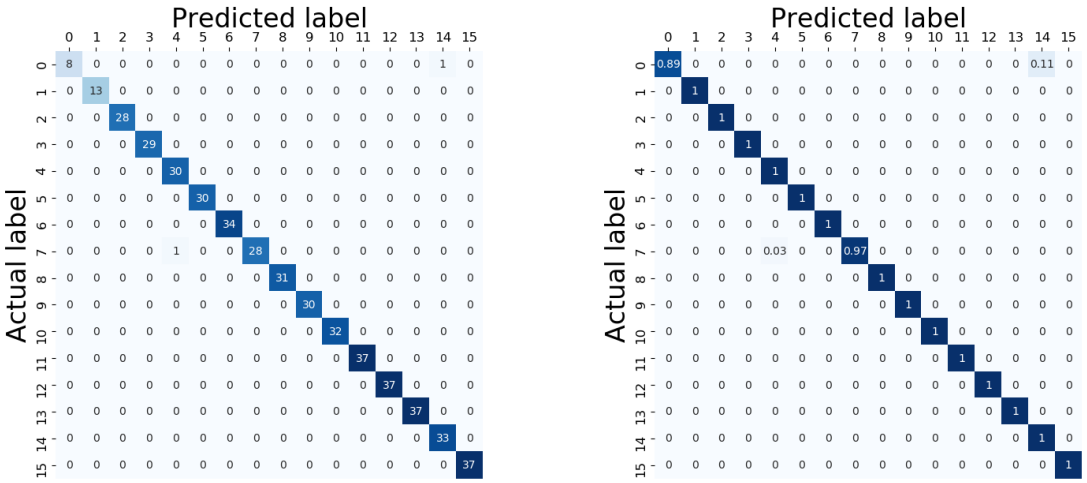
4.1. Classification results

The trained random forest classifier for the persistence images shows high accuracy scores (over 99%), suggesting a strong differentiation of the images with respect to their generating surface profiles. The classification performance report shown in Figure 9 summarizes the precision, recall, f1-score estimators over each of the 16 classes (surface labels) from the test dataset. The number of samples for each class is also provided. The accuracy score estimator is computed over the complete test dataset, along with the macro and weighted averages of the previously cited estimators.

	precision	recall	f1-score	support
Surface 01	1.00	0.89	0.94	9
Surface 02	1.00	1.00	1.00	13
Surface 03	1.00	1.00	1.00	28
Surface 04	1.00	1.00	1.00	29
Surface 05	1.00	1.00	1.00	30
Surface 06	1.00	1.00	1.00	30
Surface 07	1.00	1.00	1.00	34
Surface 08	1.00	1.00	1.00	29
Surface 09	1.00	1.00	1.00	31
Surface 10	1.00	1.00	1.00	30
Surface 11	1.00	1.00	1.00	32
Surface 12	1.00	1.00	1.00	37
Surface 13	1.00	1.00	1.00	37
Surface 14	1.00	1.00	1.00	37
Surface 15	0.97	1.00	0.99	33
Surface 16	1.00	1.00	1.00	37
accuracy			1.00	476
macro avg	1.00	0.99	1.00	476
weighted avg	1.00	1.00	1.00	476

Figure 9. Classification performance report.

The confusion matrix given in Figure 10 shows that images are accurately labelled across all classes, reporting also the normalized scores. It was proved that these results are quite insensible to randomizing and changing the ratio between the training and testing samples.

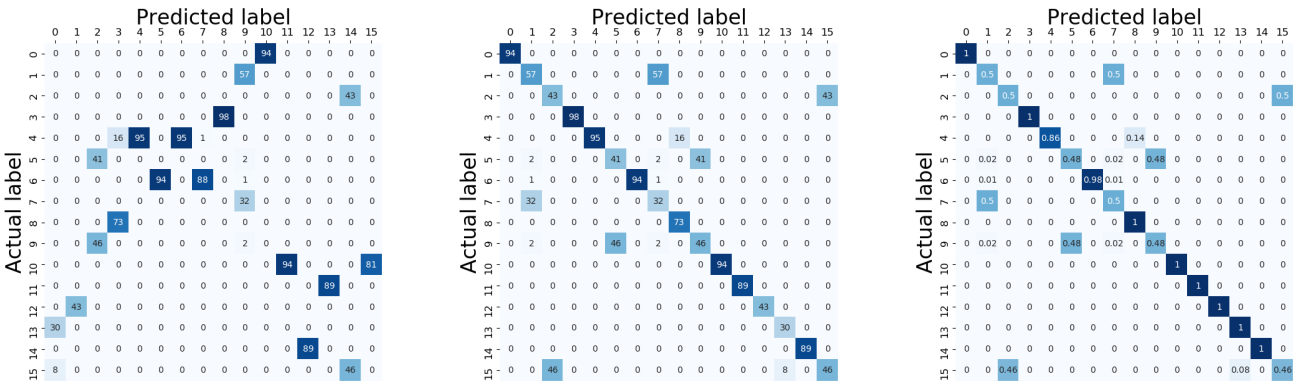


(a) **(b)**

Figure 10. Confusion matrix for the random forest classifier. (a) Original, (b) normalized.

4.2. Clustering results

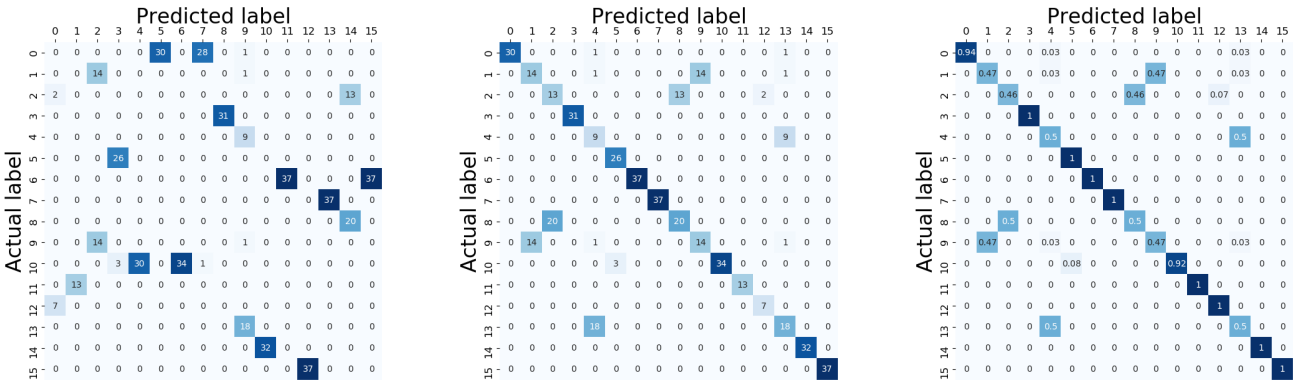
Given the disparity between clusters labels and original labels (*k*-means algorithm assigns clusters labels arbitrarily), the confusion matrix is the best way to evaluate the model performance. It shows a majority of one-to-one classes correspondence, meaning that given a certain permutation of the columns (clusters labels), we can obtain a rearranged matrix. The permuted confusion matrix given in Figure 11b shows a good accuracy (80%) for the clustering compared to the original profiles labels.



(a) **(b)** **(c)** d

Figure 11. Confusion matrix for *k*-means clustering of the complete dataset. (a) Original, (b) permuted, and (c) normalized.

In order to evaluate the predictive performances of the trained model, we compare the predicted labels (clusters) of the test data against their actual labels. The labelling disparity still remains, with a majority of one-to-one classes correspondence. After reordering the confusion matrix, depicted in Figure 12b, we can observe a good enough accuracy of the clustering (77%) for predicting labels. Thus, the model allows to identify the surface of new incoming profiles, when proceeding in an unsupervised way.

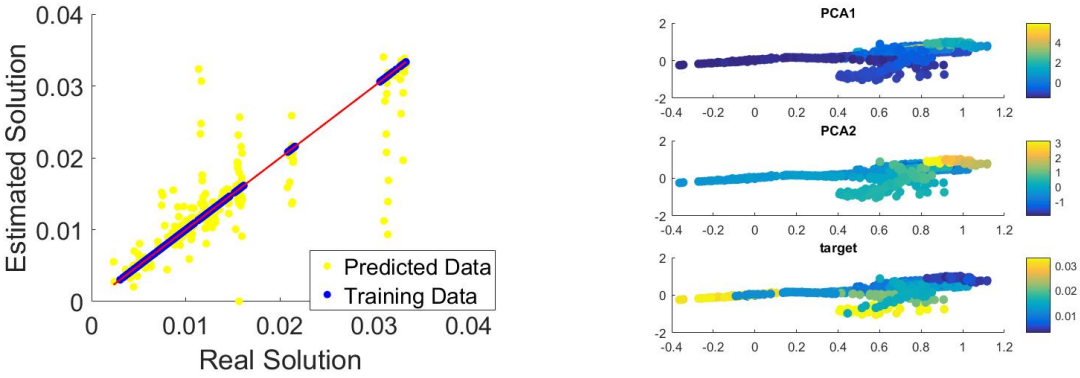


(a) **(b)** **(c)**

Figure 12. Confusion matrix for *k*-means predictions over the test dataset. (a) Original, (b) permuted, and (c) normalized.

4.3. DIC prediction by regression

Code2Vect performs an accurate regression of the DIC, with a MAPE of 2.3% when considering all the data and a MAPE of 12.86% when applied on the points that were not used in training, as shown in Figure 13. Thus, it can be concluded that the reduction of the persistence images to only two quantities (the weights of the two most relevant modes extracted from the PCA applied on the persistence images) has not a significant impact in the regression performances, proving that the combination of *Code2Vect* and PCA constitutes an excellent nonlinear dimensionality reduction technique. The correlation between these two parameteres (PCA weights) and the QoI (DIC) is also shown in Figure 13b.



(a) **(b)**

Figure 13. *Code2Vect* regression performance. (a) Prediction error, (b) projected space.

Similarly, the random forest regression shows a high reliability to accurately predict our quantity of interest, with an R^2 score over 96%.

5. Conclusion

Composite tapes have been successfully classified using the persistence images related to their rough surfaces. Topological Data Analysis seems a very valuable way of describing accurately and concisely those surfaces, in particular its roughness that constitutes the main factor when evaluating the consolidation performances, from the time evolution of the DIC (degree of intimate contact).

Different classification (supervised) and clustering (unsupervised) were successfully applied for associating the different surfaces to the composites from which they were extracted. On the other hand, by using advanced regression techniques, the degree of intimate contact was associated to the surface topological content, with excellent and fast predictions of the expected DIC for a given surface.

These procedures open unimaginable possibilities in process control and the online adaptation of processing parameters for ensuring the adequate DIC at the end of the process.

5.1. Code2Vect

Code2Vect maps data, eventually heterogenous, discrete, categorial, ... into a vector space equipped of an euclidean metric allowing computing distances, and in which points with similar outputs O remain close one to other as sketched in Figure 14.

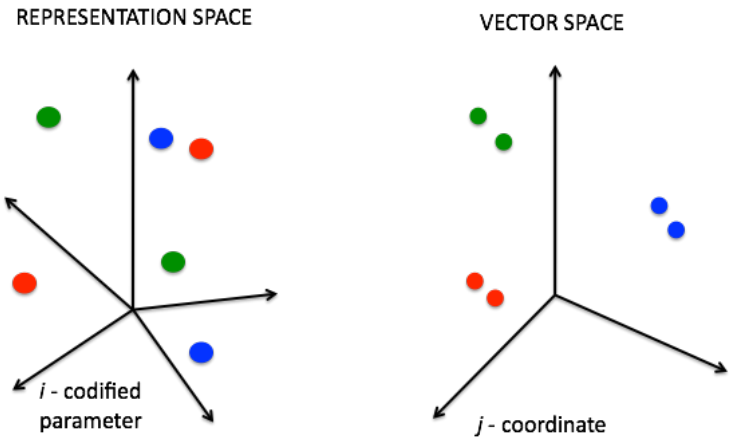


Figure 14. Input space (a) and target vector space (b).

We assume that points in the origin space (space of representation) consists of P arrays composed on D entries, noted by $\mathbf{y}_i$. Theirs images in the vector space are noted by $\mathbf{x}_i \in \mathbb{R}^d$, with $d \ll D$. The mapping is described by the $d \times D$ matrix $\mathbf{W}$, according to Eq 15:

$$\mathbf{x} = \mathbf{W}\mathbf{y}, \tag{15}$$

where both, the components of $\mathbf{W}$ and the images $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, \cdots, \mathrm{P}$, must be calculated. Each point $\mathbf{x}_i$ keep the label (value of the output of interest) associated with is origin point $\mathbf{y}_i$, denoted by O_i.

We would like placing points $\mathbf{x}_i$, such that the Euclidian distance with each other point $\mathbf{x}_j$ scales with their outputs difference, as expressed in Eq 16:

$$(\mathbf{W}(\mathbf{y}_i - \mathbf{y}_j)) \cdot (\mathbf{W}(\mathbf{y}_i - \mathbf{y}_j)) = \|\mathbf{x}_i - \mathbf{x}_j\|^2 = |O_i - O_j|, \tag{16}$$

where the coordinates of one of the points can be arbitrarily chosen. Thus, there are $\frac{\mathtt{P}^2}{2} - \mathtt{P}$ relations to determine the $d \times D + \mathtt{P} \times d$ unknowns.

Linear mappings are limited and do not allow proceeding in nonlinear settings. Thus, a better choice consists of the nonlinear mapping $\mathbf{W}(\mathbf{y})$ [24].

Acknowledgments

Authors acknowledge the ESI Chair @ Arts et Métiers Institute of Technology, the ESI Chair @ Universidad Cardenal Herrera, the AWESOME E2S/Arkema/Canoe Chair and the French ANR through the DataBEST project.

Conflict of interests

There is no conflict of interests between authors.

References

1. Chinesta F, Leygue A, Bognet B, et al. (2014) First steps towards an advanced simulation of composites manufacturing by automated tape placement. *Int J Mater Form* 7: 81–92.
2. Chinesta F, Ammar A, Cueto E (2010) Recent advances and new challenges in the use of the Proper Generalized Decomposition for solving multidimensional models. *Arch Comput Method Eng* 17: 327–350.
3. Chinesta F, Ladeveze P, Cueto E (2011) A short review in model order reduction based on Proper Generalized Decomposition. *Arch Comput Method Eng* 18: 395–404.
4. Chinesta F, Keunings R, Leygue A (2014) *The Proper Generalized Decomposition for Advanced Numerical Simulations: A primer*, Springer-Cham.
5. Chinesta F, Ladeveze P (2014) *Separated Representations and PGD Based Model Reduction: Fundamentals and Applications*, Springer-Verlag.
6. Falcó A, Nouy A (2012) Proper generalized decomposition for nonlinear convex problems in tensor banach spaces. *Numer Math* 121: 503–530.
7. Chinesta F, Leygue A, Bordeu F, et al. (2013) Parametric PGD based computational vademecum for efficient design, optimization and control. *Arch Comput Method Eng* 20: 31–59.
8. Falcó A, Montés N, Chinesta F, et al. (2018) On the existence of a progressive variational vademecum based in the proper generalized decomposition for a class of elliptic parametrised problems. *J Comput Appl Math* 330: 1093–1107.
9. Leon A, Argerich C, Barasinski A, et al. (2018) Effects of material and process parameters on in-situ consolidation. *Int J Mater Form* 12: 491–503.
10. Argerich C, Ruben I, Leon A, et al. (2018) Tape surface characterization and classification in automated tape placement processability: Modeling and numerical analysis. *AIMS Mater Sci* 5: 870–888.

11. Rabadan R, Blumberg AJ (2020) *Topological Data Analysis For Genomics And Evolution*, Cambridge: Cambridge University Press.
12. Oudot SY (2010) Persistence theory: From quiver representation to data analysis, *Mathematical Surveys and Monographs*, American Mathematical Society, 209.
13. Chazal F, Michel B (2017) An introduction to topological data analysis: fundamental and practical aspects for data scientists. *arXiv* 1710.04019.
14. Carlsson G (2009) Topology and data. *Bull Amer Math Soc* 46: 255–308.
15. Leon A, Barasinski A, Nadal E, et al. (2015) High-resolution thermal analysis at thermoplastic pre-impregnated composite interfaces. *Compos Interface* 22: 767–777.
16. Leon A, Barasinski A, Chinesta F (2017) Microstructural analysis of pre-impreganted tapes consolidation. *Int J Mater Form* 10: 369–378.
17. Yang F, Pitchumani R (2001) A fractal cantor set based description of interlaminar contact evolution during thermoplastic composites processing. *J Mater Sci* 36: 4661–4671.
18. Levy A, Heider D, Tierney J, et al. (2014) Inter-layer thermal contact resistance evolution with the degree of intimate contact in the processing of thermoplastic composite laminates. *J Compos Mater* 48: 491–503.
19. Krishnapriyan AS, Haranczyk M, Morozov D (2020) Robust topological descriptors for machine learning prediction of guest adsorption in nanoporous materials. *arXiv* 2001.05972.
20. Carlsson G, Zomorodian A, Colling A, et al. (2004) Persistence barcodes for shapes. Avaliable from: http://dx.doi.org/10.2312/SGP/SGP04/127-138.
21. Saul N, Tralie C (2019) Scikit-TDA: topological data analysis for python.
22. Adams H, Chepushtanova S, Kirby M, et al. (2017) Persistence images: A stable vector representation of persistent homology. *J Mach Learn Res* 18: 218–252.
23. Pedregosa F, Varoquaux G, Gramfort A, et al. (2011) Scikit-learn: Machine learning in python. *arXiv*1201.0490v4.
24. Argerich C, Ruben I, Leon A, et al. (2019) *Code2Vect*: An efficient heterogenous data classifier and nonlinear regression technique. *CR Mecanique* 347: 754–761.