DIVIS: A Semantic Distance to Improve the Visualization of Incomplete Heterogeneous Phenotypic Datasets

Rayan Eid
IRHS: Institut de recherche en Horticulture et Semences
Claudine Landès
IRHS: Institut de recherche en Horticulture et Semences
Alix Pernet
IRHS: Institut de recherche en Horticulture et Semences
Emmanuel Benoît
Institut superieur des sciences agronomiques agroalimentaires horticoles et du paysage: AGROCAMPUS OUEST
Pierre Santagostini
IRHS: Institut de recherche en Horticulture et Semences
Angélina El Ghaziri
IRHS: Institut de recherche en Horticulture et Semences
Julie Bourbeillon (julie.bourbeillon@agrocampus-ouest.fr)
IRHS: Institut de recherche en Horticulture et Semences https://orcid.org/0000-0002-3365-1286

Research

Keywords: mixed datasets, heterogeneous datasets, phenotypic traits, multivariate analysis, ontologies, semantic distance, clustering, visualization

Posted Date: August 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-742853/v1

RESEARCH

DIVIS: a semantic DIstance to improve the VISualization of incomplete heterogeneous phenotypic datasets

Rayan Eid[1], Claudine Landès[1], Alix Pernet[1], Emmanuel Benoît[2], Pierre Santagostini[1], Angelina El Ghaziri[1] and Julie Bourbeillon[1*]

*Correspondence: julie.bourbeillon@agrocampus-ouest.fr
[2]AGROCAMPUS OUEST, Institut Agro, 2 rue André le Nôtre, F-49045 Angers cedex 1, France
Full list of author information is available at the end of the article

Abstract

Background: Thanks to the wider spread of high-throughput experimental techniques, biologists are accumulating large amounts of datasets which often mix quantitative and qualitative variables and are not always complete, in particular when they regard phenotypic traits. In order to get a first insight into these datasets and reduce the data matrices size scientists often rely on multivariate analyses. However such approaches are not always easily practicable in particular when faced with mixed datasets with missing values. Moreover displaying large numbers of individuals leads to cluttered visualizations which are difficult to interpret.

Results: We introduce a new methodology to overcome these limits. The underlying principle consists in (i) grouping similar individuals, (ii) representing each group by emblematic individuals we call archetypes and (iii) build sparse visualizations based on these archetypes. As a preliminary step to the clustering we design a new semantic distance tailored for both quantitative and qualitative variables which allows a realistic representation of the relationships between individuals. This semantic distance is based on ontologies which are engineered to represent real life knowledge regarding the underlying variables. Our approach is implemented as a Python pipeline and illustrated by a rosebush dataset including passport and phenotypic data.

Conclusions: The introduction of our new semantic distance and of the archetype concept allows us to build a comprehensive representation of an incomplete dataset characterized by large proportion of qualitative data. The methodology described here could have wider use beyond information characterizing organisms or species and beyond plant science. Indeed we could apply the same approach to any incomplete mixed dataset.

Keywords: mixed datasets; heterogeneous datasets; phenotypic traits; multivariate analysis; ontologies; semantic distance; clustering; visualization

Background

The 2000s and the sequencing of complete genomes sparked a scientific revolution in the study of living beings. The now accessible no *a priori* approach results in the wider spread of high-throughput experimental techniques such as transcriptomics, proteomics, metabolomics or phenomics and the increase in the volume of publicly available data. As a consequence, biologists are accumulating large amounts of datasets which are characterized by an increasing heterogeneity:

- Information sources heterogeneity: multiple databanks, which can be local or distant, with various formats and interfaces, files with various formats,
- Data heterogeneity: various scales (from the molecule to the population), various natures (quantitative and qualitative), various modes (text or images), various structuring levels (database fields, structured text, free text).

Therefore the demand by biologists to integrate heterogeneous and large datasets from "omics" and phenotyping activities is rapidly expanding [1].

In this context where large complex datasets are becoming more and more widespread, biologists often rely on multivariate analyses to project individuals in a new coordinates space to get a first insight into the data and have smaller matrices to process. However such approaches are not always easily practicable in particular when faced with mixed (qualitative and quantitative) incomplete (that is to say including missing values) datasets. Moreover displaying large numbers of individuals leads to cluttered visualizations which are difficult to interpret.

In this paper we introduce a new methodology designed to overcome these limits. The approach relies on a new semantic distance which is designed for both quantitative and qualitative variables and allows for a realistic representation of the relationships between individuals. This semantic distance is based on ontologies which are engineered to represent real life knowledge regarding the underlying variables. We associate this new distance definition with an archetype concept to overcome the cluttered displays issue. Indeed we define archetypes as individuals representing groups of similar individuals from the dataset. Limiting the visualizations to these archetypes leads to a sparser representation which still provides valuable insight into the data.

More precisely the structuring of the population in groups is conducted through clustering, for which numerous approaches exist [2, 3]. A common characteristic of clustering techniques is that they group individuals based on their similarity. This similarity is estimated based on distances between the features of the individuals.

However most clustering methods rely on numeric arithmetic. Therefore the features have to be represented by numeric values. This causes problems with qualitative variables and even more in the case of mixed datasets. Classical approaches consist in the discretization or dummy-coding of qualitative variables to transform them into numeric variables. But if the number of modalities is very different between variables the weight of each variable in the resulting similarity between individuals might be unbalanced [4]. Some distances are designed to cope with qualitative data. For instance we can cite Jaccard's coefficient [5], Dice's coefficient [6], Gower's distance [7], or the Chi-Square [8]. These metrics are widely used in biology, and in particular in ecology, to characterize species populations. For example *Pandey et al* rely on Jaccard's coefficient to cluster sesame (*Sesamum indicum* L.) populations [9], Pavoine and colleagues extends Gower's distance to characterize periurban woodland plant species populations [10] and *de Bello et al* propose a solution to overcome the issue of the disproportionate contribution of certain traits which exists with Gower's distance [11]. For a review on current clustering approaches for heterogeneous data, see [12].

Even if methods exist to account for qualitative variables, the fact that they do not always consist in a flat list of categories is often overlooked. Indeed in many cases these categories are structured. For instance a variable corresponding to the months of the year can be considered as a circular variable and proposals have been made to take this fact into account in distance calculations, through an extension of Gower's distance [10]. But a lot of qualitative variables modalities are structured according to more elaborate schemes and qualitative variables can be described as ontologies [13]. Ontologies structure knowledge as graphs where nodes represent concepts or terms and edges the relationships between them. Ontologies are heavily developed and used in life sciences to annotate data, in particular in almost every major biological database, and reason over domain knowledge [14].

In an ontology representing the modalities of a variable, modalities/values could be viewed as concepts, while the complex links between them would be materialized by the graph of relationships between concepts. We therefore propose to use the distance between concepts in corresponding ontologies to measure the distance between modalities of qualitative variables.

The measurement of distances in ontologies is a fundamental Semantic Web notion which is exploited for clustering, data mining or information retrieval [15]. Numerous formulas or algorithms [16, 17] exist to define the distances between concepts in ontologies but most are based on two main approaches or a mix of the two:

- Edge-based approaches imply counting the number of edges between two concepts in the ontology graph,
- Node-based approaches compare the properties of the concepts involved, be it the concepts themselves, their parents, or their children. They generally rely on the Information Content (IC) notion which evaluates how specific and informative a concept is.

However these approaches rely on the topology of the graph, with no regard for the reality of what the concepts represent. This can lead to inaccurate measurements. For instance, a geographical ontology graph usually positions France, Italy and Denmark as three concepts that are part of Europe. A classical ontological distance calculation would lead to identical pairwise distances between these countries. This is obviously false from a geographical point of view: Italy is closer to France than it is to Denmark.

We therefore intend to augment the ontology graphs with *a priori* knowledge represented as distance values associated with the relations between concepts.

Moreover clustering and distance calculations usually cannot be performed as is on datasets including missing data. But data matrices in biology are often incomplete, for example because of the cost of some experimental techniques or because an individual hasn't been available for the full duration of the study. The traditional approach to coping with missing data is to exclude the affected individuals or to ignore the variables with too many missing values. But this reduces the amount of data available for analysis. In the case of matrices with a large quantity of holes it could make the study completely useless. Estimating missing values using imputation is often presented as a better approach. A review of available methods is

performed in [18] in the epidemiology field. However Johnson and colleagues show that estimating missing data is not always appropriate and none of the methods they tested could deal effectively with severe biases [19] which can be common in trait datasets.

We therefore decide to define a distance which can be calculated even when the features of an individual are not all described, i.e. in the presence of missing data.

To reduce the number of individuals we also propose to represent each cluster by a limited number of individuals we call archetypes. In order to define these archetypes, different strategies can be considered depending on the clustering results:

- In the case of a large number of small clusters, representing each group by a single individual is probably more relevant. In such a case we can imagine basing the archetypes definition on the cluster centroids.
- In the case of a small number of large clusters a single individual might not be sufficient to represent the intra-cluster diversity. In these conditions, better to select several individuals with one of the existing sampling techniques [20, 21].

A visualization of the archetypes allows to declutter the initial display of the population and can be associated with means to access the whole groups.

In this paper we develop the new approach we hinted at to overcome the listed limitations of classical methods to manipulate large datasets. We apply it to a rosebush dataset which includes passport data and a collection of qualitative and quantitative phenotypic traits.

Methods

Use case: rosebushes phenotypic traits

To illustrate our study we rely on information associated with the rosebush collection of the RosePom Biological Resource Center (BRC) in Angers, France. The dataset consists in passport data and the phenotypic traits evaluated during the study of French roses (*Rosa* sp.) performed by *Liorzou et al.* [22]. It includes 1434 rosebushes from European garden roses from the 18^{th} and 19^{th} centuries. Each rosebush is described by the variables listed in Table 1. With the exception of the number of flowers all these variables are qualitative. Their respective modalities have been defined by domain experts.

Table 1 Variables in the rosebush dataset.

Variable	Type	Nature	Number of modalities
Horticultural group	Passport	Qualitative	17
Geographic origin	Passport	Qualitative	9
Breeding period	Passport	Qualitative	16
Ploidy level	Passport	Qualitative	5
Petal colour	Phenotypic	Qualitative	12
Bush height	Phenotypic	Qualitative	6
Quantity of prickles	Phenotypic	Qualitative	4
Perfume intensity	Phenotypic	Qualitative	2
Repeat flowering level	Phenotypic	Qualitative	6
Number of flowers by volume	Phenotypic	Quantitative	-
Duplicature type	Phenotypic	Qualitative	4

Passport variables come from [22] or are inferred from them. The horticulture group is defined according to the American Rose Society (ARS) classification. Breeding dates are grouped into time periods. Phenotypic variables have been evaluated by the RosePom BRC and its partners: breeders and rose gardens.

The dataset is far from complete: the "Quantity of prickles", "Perfume intensity", "Repeat flowering level" and "Number of flowers by volume" variables are only filled in for a small number of rosebushes. Some individuals include information for only one or two variables.

Such a dataset could be difficult to analyze with classical approaches and we choose it to test our method and see if we can provide better insight into the data. The dataset is therefore subjected to the pipeline presented in Figure 1. This pipeline is developed in *Python 3.7*. It relies on the *NumPy* [23] and *pandas* [24] libraries to manipulate the data, *scikit-learn* [25] to perform machine learning and *matplotlib* [26] to draw the figures. The next subsections detail it more precisely.

Build ontologies and capture the distance between concepts

General principle

The first stages of the process consist in associating each qualitative variable in the dataset with an ontology, which corresponds to steps (1) and (2) from the pipeline in Figure 1. The various modalities of a variable will then become concepts in an ontology. Two cases can be considered:

- A publicly available ontology corresponding to the variable exists: we can use and eventually adapt it to fit our variable modalities,
- No public reference ontology exists: we rely on expert knowledge to transform the list of modalities of the variable into the concept graph of an ontology.

For each ontology and each pair of concepts in the ontology we have to define a pairwise distance, as indicated in step (3) of the pipeline in Figure 1. Here again two cases exist:

- A distance can be defined based on what the variable represents in real life: we use this distance and calculate it as needed,
- No simple distance can be calculated: we rely on expert distance estimations. These estimations are stored along the ontology graph.

These information are then processed to build a distance matrix for each variable as for step (4) of Figure 1.

Therefore the qualitative variables in our dataset are handled as follows.

Variables associated with public ontologies

Public ontologies exist in relation to colours and geographic information and they can be used for our "Petal colour" and "Geographic origin" variables.

Regarding colours we extract their descriptions from DBpedia [27], using its SPARQL endpoint. These descriptions include coordinates of reference in different colour spaces. We choose to use the *L*a*b** colour space because it is designed to approach the perception of colours by human vision. In this space, $L*$ represents perceptual lightness, $a*$ the green–red opponent colors and $b*$ the blue to yellow tones. We then use as distance the ΔE (CIE 2000) which quantifies the visual difference between two *L*a*b** colours and is presented in Equation (1).

$$\Delta E = \sqrt{({L_2}^* - {L_1}^*)^2 + ({a_2}^* - {a_1}^*)^2 + ({b_2}^* - {b_1}^*)^2} \quad (1)$$

In order to evaluate this distance we use the implementation from the Python *colormath* library [28].

Regarding regions of origin we take advantage of the *GeoPy* library [29] and the Nominatim geocoder to access OpenStreetMap data [30] and associate the locations with coordinates. Some region names in our dataset do not exist in OpenStreetMap. It is for example the case for the subdivision of France into four main quadrants. In such cases we consider the list of named areas composing the region and associate it with the mean latitude and the mean longitude of the areas in the list as a proxy for its location.

Variables with no associated public ontology

For the other variables no existing ontology can be located. The structuring in the form of a graph of possible values is carried out for each variable in collaboration with rosebush experts and stored in an ontology file in OWL format using the Protégé editor [31]. We then have to define a distance between pair of concepts in each graph.

For the time periods we consider we can calculate such a distance. If $S1$ and $S2$ are the start years of two periods and $E1$ and $E2$ the end years, we define the distance between the two periods Δt as the number of years between the middle of each period as presented in Equation (2).

$$\Delta t = \left\lfloor \left| \left(S1 + \frac{E1 - S1}{2} - S2 + \frac{E2 - S2}{2} \right) \right| \right\rfloor \tag{2}$$

Among the time periods modalities two have just one date: "< 1700" and "> 1920". For the calculations we consider 1600 as the start date for the first one and 2020 as the end year for the second.

Distances between the other phenotypic variables are defined with the help of rosebush experts. The corresponding ontologies are usually organized as trees. Moreover the modalities of the variables in the original sets are often ordered. For instance the set of modalities for the "Quantity of prickles" variable ("low", "average", "high" and "very high") can be ordered from the lowest quantity to the highest. In the ontology these modalities are organized in two subgroups as presented Figure 2. Distances between pairs of leaf concepts are defined with arbitrary but not random values: we choose values so that inter-subgroups distances are higher than distances within a subgroup and so that the original order is conserved when it is relevant. The resulting distance matrix for this example is presented in Table 2.

Table 2 Distance matrix for the quantity of prickles ontology.

	Low	Medium	High	Very high
Low	0	1	5	6
Medium	1	0	4	5
High	5	4	0	1
Very high	6	5	1	0

Having defined values for pairwise distances between concepts we then need to store them into the OWL ontology. In order to do this we introduce a *has_distance* relationship, that is to say an Object Property. We associate it with a *distance*

Data Property of type $owl : real$. This *distance* is the Range of *has_distance*. This principle is illustrated in Figure 3.

If we consider the previous example, the distances between the "Low" concept and the others in the "Quantity of prickles" ontology is represented as in Figure 4.

Build distance matrices for the ontologies

In order to build the distance matrix for the colour and region ontologies, each pair of concepts in the ontology file is processed and the distance is calculated according to the previously defined methods. The OWL file containing the other ontologies that we engineered is read using the Python *Owlready2* library [32]. It allows us to retrieve the list of concepts for each ontology along with the pairwise distances. These are formatted as distance matrices stored in a global Microsoft Excel file.

The range of distance values for each variable are very different. In order to prevent some variables from out-weighting the others in future calculations, each distance matrix is normalized on a $[0 - 100]$ scale.

Build individuals distance matrix

The following step in Figure 1, that is to say step (5), consists in building the individuals distance matrix. We first of all have to define how to calculate the pairwise distance. Each individual can be represented as a vector of variable values. If we consider two individuals represented by the A and B vectors, the values of the i^{th} variable can be represented as A_i and B_i respectively. The distance $d_{A,B}(i)$ between A and B for the i^{th} variable can be found in the corresponding distance matrix as the distance between A_i and B_i. The distance $D(A, B)$ between A and B can therefore be expressed as Equation (3).

$$D(A, B) = \frac{1}{M} \sum_{i=1}^{N} d_{A,B}(i) \tag{3}$$

where N is the total number of variables in the vectors and M the number of variables for which a distance can be defined. Indeed we have missing data in our dataset. If either A_i or B_i or both are missing then $d_{A,B}(i)$ is missing too. An example for a subset of variables is presented Figure 5.

We can then repeat the operation for all pairs of individuals to build the final distance matrix. In the matrix we store for each pair of individuals both the $D(A, B)$ distance and M, the number of variables used in the calculation.

In order to have a baseline for comparison, we also calculate a distance matrix based on Gower's distance [7]. There is no official implementation of this distance in Python libraries. Moreover we have to adapt a version which would handle missing values the same way as our semantic distance. We therefore develop our own based on the Dice and Manhattan distances as found in the *scikit-learn* library. The algorithm goes through all variables in the dataset and builds a distance matrix for each variable. If the variable is quantitative, it uses the Manhattan distance. If the variable is qualitative it converts it into binary indicator variables including one

for missing values. It calculates the Dice distance on the new dummy variables and marks as missing the pairwise distances which implied the missing value indicator. The final distance matrix is calculated as the by element average of all the individual variables distance matrices.

Projection in coordinates space and clustering

The next stage of the process would be to group similar individuals based on the distance matrix. Since different clustering algorithms can produce different results depending on the structure of the population to cluster we choose to test several algorithms. However not all clustering algorithms can use a distance matrix as input.

Therefore we perform a Multi-Dimensional Scaling (MDS) [33] to project the distance matrix in a coordinates space and use the projection as input for all clustering algorithms, as indicated in step (6) of the Figure 1 pipeline. The distance matrix is subjected to a metric MDS as implemented in the *scikit-learn* Python library [25]. The MDS function provides a value of $STRESS$ which quantifies the quality of the representation. This indicator is normalized to obtain the "Kruskal stress" (*stress1*), defined in Equation (4):

$$stress1 = \sqrt{\frac{\sum_{j>i}(\hat{\delta}_{ij} - \delta_{ij})^2}{\sum_{j>i}(\delta_{ij})^2}} \tag{4}$$

where δ_{ij} corresponds to the observed distance between pairs of individuals (i, j) supplied as input to the multidimensional positioning algorithm and $\hat{\delta}_{ij}$ is the reconstructed distance in the Euclidean space representing the data.

$stress1$ is a widely used indicator in the literature [33] and thresholds exist to guide the selection of the number of dimensions to keep in the new space to have a sufficiently good representation.

Regarding the clustering *per se*, that is to say step (7) from the pipeline in Figure 1, we rely on the $scikit-learn$ implementations of the following algorithms:

- Birch [34],
- Gaussian Mixture [35],
- Hierarchical Clustering with Ward linkage [36],
- KMeans [37],
- KMedoids [38],
- Spectral Clustering [39].

The objective is to compare the results between the various algorithms.

Most of these algorithms require a number of clusters as parameter. In order to assist in this choice we perform a Silhouette analysis [40] using the *scikit-learn* implementation of the Silhouette coefficient calculation.

To compare the results of the various clustering algorithms we divert classification evaluation approaches for our purpose. Indeed these evaluation techniques usually compare a prediction with a ground truth. We don't have a ground truth but the results of several algorithms. We consider the KMeans clustering results as "ground truth" and the result of each of the other algorithms as predictions. We

then compute concordance matrices using *pandas* and confusion matrices using *scikit-learn*.

We use the same approach to compare the clusters between the two distances.

Archetypes definition and visualization

The next stage (8) of the process described in Figure 1 consists in representing each group by a small number of individuals. As previously stated two main strategies can be considered:

- in the case of a large number of small clusters, we represent each group by a single individual.
- in the case of a small number of large clusters we represent each group by several individuals through sampling.

To define the single archetype, we identify the cluster centroids as implemented by the *scikit-learn* library. We then calculate the Euclidean distance between each individual and its cluster centroid. The single archetype is the closest individual to the centroid. To choose several archetypes per cluster, we perform a random sampling of 5 % of the population in each cluster, using the *pandas* library.

The results visualizations are constructed using the *seaborn* library [41]. Given the number of dimensions may be superior to three we choose to draw pairplots. Given pairplots are symmetrical along the diagonal we use each half to present different elements: the original scatter plots in the bottom left part and archetypes associated with kernel density estimations to give a rough estimate of the groups envelops in the top right section.

Results

Classical analysis: MCA

Before subjecting the dataset to the pipeline presented here we explore it with a classical analysis. Given the types of the variables in the dataset, a Mutiple Correspondence Analysis (MCA) [42] is seemingly appropriate. However we have to deal with the missing data. Three approaches are usually used to do so:

- Delete the rows with missing values. In our case, four variables (out of 11) have a very high percentage of missing values (98.9%, 98.0%, 94.6% and 84.8%). We remove these variables before proceeding to the missing values by row. 405 individuals (out of 1434) don't contain any missing values. Thus it almost reduce by a third the number of variables and two-thirds the individuals in the population.
- Consider the missing data as a particular category. Here it is not relevant since the missing data for sure are not in the same category.
- Use multiple imputation methods [43]. It is not appropriate to rely on such an approach in our context since the rosebush varieties are by definition different. The similarity in some features doesn't reflect a similarity in the varieties.

We apply MCA with the *prince* Python package [44], beginning with the first method (removing all the missing values). We present in Table 3 the percentage of inertia explained by the first six components. With the first two components, the total inertia is 7.5% and reaches 19.2% with six components. Therefore the projection in the MCA space is very poor.

Table 3 MCA explained inertia by component for the rosebush dataset without missing data.

Component	1	2	3	4	5	6
Explained inertia	0.041	0.034	0.032	0.029	0.028	0.028

Another point that one needs to be careful about in MCA is the presence of rare categories (categories of small size). These categories can affect the results since the associated inertia will be high. Several solutions can be considered to remedy this. In particular, we can group the categories if there are natural groupings. By grouping categories and removing the missing values, we increase the inertia of the axes of the MCA. This percentage reaches 12.2% for two components and 30.8% with six components. However, for some variables combining into bigger categories is not relevant. For instance grouping underrepresented flower colours leads to treating very different colours together and to distinguishing similar colors.

In the end, this attempt to use an MCA approach is not conclusive for our objective and for our type of data.

Distance matrices

Following the described pipeline for our dataset we first of all produce distances matrices between individuals using both distances: Gower's and our own semantic one. These matrices are displayed as heatmaps in Figures 6 and 7.

The matrix is sparser in the Gower case and a larger proportion of values are closer to the maximum. This can be explained by the way the two distances are constructed. The distance between individuals is based on a majority of qualitative variables and a single quantitative variable. The qualitative variable is associated with a very limited amount of data. In the Gower case we mainly represent the proportion of variables whose values are different between individuals. Indeed qualitative variables are somewhat "interchangeable" given the distance between two individuals is binary. The values in the distance matrix are often superior to 0.5 because our rosebushes don't share a large proportion of values and we necessarily have a distance of 1 for each variable where the values are different. In the semantic case, the possible distance values between modalities are different from 1 and depend on the variable. The range of possible distances between individuals is therefore quite large but with a smaller maximum. The frequency of values in the two cases presented in Figure 8 illustrates this.

Looking at the heatmaps from Figures 6 and 7 both distances seem to structure the population in 3 or 4 groups but the interpretation is less clear between the two larger groups in the semantic distance case.

Multi Dimensional Scaling

In order to choose the appropriate number of dimensions for the MDS we plot the value of $stress1$ from Equation 4 for an increasing number of axes, as presented in Figure 9. The $stress1$ values are similar for both distances. Its is generally admitted that a $stress1$ value below 0.2 corresponds to a good representation of the distance matrix in a coordinates space [33]. Thus we choose 4 dimensions for the MDS.

Looking at individuals coordinates in the new space it appears that data points are more spread for the semantic distance than for Gower's distance. It is related to the fact that we have a wider range of distances between individuals.

Number of clusters choice

For both distances and as a preliminary step for the clustering process, we perform a Silhouette analysis using two strategies:

- Plot the mean Silhouette coefficient as a function of the number of groups for three clustering algorithms: KMeans, KMedoids and Hierarchical Clustering,
- Perform a Silhouette analysis at the individuals level for several number of clusters for the KMeans algorithm.

The mean Silhouette graphs for the KMeans algorithm are presented Figure 10.

For the semantic distance the number of clusters which maximizes the Silhouette coefficient is 6. Profiles are similar for the KMedoids and Hierarchical Clustering algorithms and suggest 5 or 6 clusters. A more precise rendering of Silhouette values at the individual level is presented for the KMeans algorithm and 5, 6 and 7 clusters in Figure 11 [see Additional file 1 for renderings for 2 to 19 clusters]. This figure confirms that the clustering quality from a Silhouette point of view is similar and we chose to perform the next steps for 5, 6 and 7 clusters.

For Gower's distance the situation is very different since it seems that the more clusters the better the representation [see Additional file 2 for Silhouette coefficient renderings for 2 to 19 clusters]. We anyway choose to perform the next steps with 5 to 7 clusters so that the results can be compared with the semantic distance.

Cluster analysis

As part of the analysis of the clustering results we want to:

- Compare the results between the different clustering algorithms for each distance (Gower's and semantic). In order to do so we consider the KMeans results as ground truth and compare each of the other algorithms with it.
- Compare the results of our new semantic distance with Gower's for all algorithms. Here we consider the results for the semantic distance as ground truth.

In order to do so we calculate for 5, 6 and 7 clusters and for the two sets of comparisons:

- Confusion matrices. These allow us to determine the number of rosebushes which are classified the same way or differently between two methods,
- Concordance tables. These provide us a mapping between the clusters built according to the two methods.

As illustration of the comparison of the results between algorithms Figure 12 presents the concordance matrices for the semantic distance [for the Gower distance see Additional file 3]. In each heatmap rows correspond to the KMeans clusters and columns to the clusters for the other algorithm. The other algorithms are Birch, Hierarchical Clustering and Gaussian Mixture for the top three heatmaps and KMedoids and Spectral Clustering for the two bottom ones.

Looking at the confusion and concordance matrices for both distances and all three numbers of clusters it appears that:

- Hierarchical Clustering and Birch clusters are very close to the KMeans clusters,
- KMeans and KMedoids results are very close except for KMeans cluster 1 which is split in two by the KMedoids algorithm,

- Spectral Clustering and Gaussian Mixture results are more different. Moreover the results of these two algorithms are not similar.

We have a good concordance between three of the algorithms, the fourth one has only one major difference and the last two have some clusters which are mixed up. Moreover the concordance is better for the Gower distance for all algorithms except Gaussian Mixture, in particular for 6 clusters. Therefore we decide to focus on the results of the KMeans algorithm with 6 clusters for the following analyses.

Regarding the comparison between the two distances the concordance matrix is presented as Figure 13 for the KMeans algorithm and 6 clusters. Group 3 is the only one which is almost identical between the two distances. Group 0 and 2 from the Gower distance are spread among the various semantic groups, with the larger subgroups in semantic groups 5 and 4 respectively. Almost all individuals from the Gower group 1 are in the semantic group 1 but the semantic distance associates them with individuals from the Gower groups 4 and 5. Group sizes are more balanced with the Gower distance. We can suppose it is once again linked to the fact that the data points are more spread in the semantic case.

Archetypes and visualisations

We calculate the archetypes positions for both distances and both approaches: one or several archetypes per group. The resulting visualizations are presented for the KMeans algorithm, 6 clusters and the semantic distance in Figure 14 for the single archetype approach and Figure 15 for the multiple archetype approach. Equivalent figures for Gower's distance are provided as Additional file 4 and Additional file 5.

The projection according to the 4 axes of the MDS (1/2 matrix at the bottom left) shows a structuring of the 6 groups along plan 2,3 which separates the pink / light blue groups from the four other groups and axis 4 which separates the three light blue / yellow / red groups from the others. The plans which provide the better representation of the six groups are 2,3 and 3,4. The complex structure of the point cloud is captured by the K-Means clustering in 6 groups. We can also see that Gower's distance does not provide such a clear structure between the groups.

Comparing our representations (top right corner of the pairplot) with the whole dataset scatter plot (bottom left corner of the pairplot) it appears our representation provides a good overview of the dispersion of the population and of its structuring as groups. Both approaches (single or multiple archetypes per group) seem relevant and the choice of one or the other is a matter of user preference and of the number of clusters: the more clusters there are, the less archetypes per cluster are required to provide a good overview of the dataset.

Discussion

The introduction of our new semantic distance and of the archetype concept allowed us to build a comprehensive representation of an incomplete dataset characterized by large proportion of qualitative data. This can be useful from several perspectives.

Incomplete datasets including mixed (quantitative and qualitative) data are becoming more and more common in life sciences. Classical statistics approaches such are MCA present limits when it comes to providing a first insight into such data: it is

often necessary to drop part of the original information to build complete matrices. Data imputation is a solution to the problem but it also has drawbacks and imputed values remain probabilistic and might not represent reality. The approach we developed allows to overcome these problems. Indeed, as long as a pairwise distance can be calculated for all pairs of individuals, they can all be taken into into account in the subsequent MDS projection, clustering, archetype definition and visualization.

Regarding distances, we introduced a semantic distance as an alternative to distances tailored for mixed data such as Gower's. This semantic distance (Equation (3)) allows to account for the underlying meaning of qualitative variables. It can be attached to real life measures such as geographical distances or associated with specific calculations such as the distances between time periods we defined in Equation (2). It can also be based on expert knowledge regarding both the structuring of the modalities of the variable as the concepts graph of an ontology and the distance values between two concepts. This semantic distance brings more precision regarding how two individuals relate to each other compared to Gower's, which is more binary. This allows a wider range of possible distance values in the dataset, and as a consequence a more realistic spread of the data points in the MDS coordinates space. Moreover this semantic distance is defined as a weighted sum. Therefore it allows to give more or less importance to some variables compared to others, thus granting the ability to fine tune the way each facet of the dataset is managed.

Relying on *ad hoc* distances in concept graphs for some variables, we had to capture this information in an ontology format. We did just that in OWL format: we defined a *has_distance* relationship. This relationship allows to link two concepts and store the distance value within a *distance* property attached to it. Giving a numeric value to the distance between two concepts is difficult for domain experts but such an approach also present advantages. These distances are data and they can be easily changed, which once again brings flexibility to the way we process the datasets.

Moreover we tackled the problem of cluttered scatter plots by reducing the number of displayed individuals.

On the application point of view, we illustrated our approach with passport and phenotypic traits associated with a collection of rosebushes held by a BRC. But it could be used for any dataset describing a large set of organisms, for instance in ecology, and including other types of data, for instance genomic. More widely it could be used for any incomplete dataset mixing qualitative and quantitative variables. A problem which present similar premises (reduce the number of individuals representing a population) is the constitution of core-collections by BRCs. Indeed BRCs store large collections of biological material and associated information and they often need to constitute sub-samples of a more manageable size *e.g.* for experimental purposes. These core-collections include, with a minimum of repeatability, the maximum diversity of the species in question [45] and are designed by exploiting the maximum amount of data available: origin of the samples, genetic and phenotypic characteristics, etc. The existing strategies for the selection of inputs are diverse: random strategy, partitioning (also called "stratification"), maximization, and some other so-called "hybrid" strategies [46]. The methodology presented here could add a new tool to the arsenal of BRCs.

However, even if it is functional, the methodology presents some limits.

First of all the method relies heavily on ontologies which have to be defined. Efforts towards reference characterization of individuals in the plant sciences domain exist, for instance the MIAPPE (Minimum Information About a Plant Phenotyping Experiment) [47] minimum requirements or the ontologies of the Planteome (https://planteome.org) databank [48], in particular the Plant Trait Ontology. Sharing more reference ontologies would reduce the knowledge engineering burden. Moreover we specifically defined the distances between concepts in our ontologies. This again may not be practical, depending on the number of concepts. Indeed these concepts correspond to modalities of qualitative variables. Some classes may be associated with measures, such as our time periods, colours and geographic locations, but it is not always the case. In this situation distances have to be defined artificially and not only is the process time consuming but it might bias the results. Methods to better anchor the distances with quantifiable information have therefore to be designed.

Secondly the management of missing data could be further refined. The distance between individuals defined in Equation (3) allows to calculate a distance with each individual having missing data. However, depending on the number of variables where two individuals share values, the pairwise distance can be calculated based on different numbers of variables. For instance in our rosebush example we have distances calculated from 1 to 9 variables out of 11 potential variables for individuals with a complete record. In this context we might want to consider that a distance calculated based on more variables is more accurate than one calculated with less. One approach to represent this accuracy might be to represent the distance not as a number but as an interval or a fuzzy number. Another approach would be to associate an error to the distance. We then would have to perform the next stages of the process (MDS, clustering, archetype definition and visualization) based either on fuzzy data or error prone data. Methods are described in the literature, for instance for fuzzy MDS [49] or fuzzy clustering [50]. We however have to study the topic more thoroughly and find implementations of the described approaches or develop our own.

Thirdly the approaches we used to build the archetypes representing the clusters may not be the most relevant. We might want to better link the construction of these individuals with the values of the variables in the original dataset. A better archetype might indeed be an "artificial" one whose variable values are the most represented in its cluster. Defining an archetype this way however introduces new problems. It would have to be projected in the new coordinates space created by the MDS so that it could be represented in the visualizations. It isn't a trivial task given we can only calculate distances between individuals. The topic would have to be explored further.

Fourthly the vizualisations we produced are static pairplots. A big improvement would be to render them dynamically and make the visualisation interactive. A graphic interface to choose which display to render (which distance, which clustering algorithm, how many clusters, etc.) would be most welcome. We could imagine allowing to rotate and zoom in and out of the display. Tooltips associated with the archetypes could provide information regarding the cluster they represent such

as number of individuals, main characteristics regarding the original variables, etc. Clicking on an archetype could change the display and lead to a visualization of the individuals composing the corresponding cluster (or a larger subset of it). The pipeline presented here was developed as proof of concept regarding the interest of the semantic distance and archetype notions. Dynamic visualizations would regard future work.

Conclusion

In this paper, we present a new method to integrate heterogeneous datasets including missing data. The approach relies on a new semantic distance which is designed for both quantitative and qualitative variables and can be considered as an alternative to for instance Gower's. This distance allows for a more realistic representation of the relationships between individuals and a wider spread of the data points. This semantic distance can be linked to real-life knowledge regarding the modalities of the underlying variable or to distance measures captured in ontologies. In this respect, we defined how to describe the distance value between two concepts in OWL format. We associated this new distance definition with an archetype concept to overcome the cluttered displays issue. Indeed we defined archetypes as individuals representing groups of similar individuals from the dataset. Limiting the visualizations to these archetypes leads to a sparser representation which still provides valuable insight into the data.

The methodology described here was applied to a dataset describing rosebush passport and phenotypic traits but it could have wider use beyond information characterizing organisms or species and beyond plant science. Indeed we could apply the same approach to any incomplete mixed dataset. Moreover, the selection of a representative subset of a population is a widespread problem. It is for instance faced by BRCs willing to build core collections for the species they are conserving. Our technique could provide a complementary methodology to existing ones.

The method is fully functional and has been implemented in Python 3.7. However, some aspects imply future work. The method relies heavily on ontologies. Sharing more reference ontologies would reduce the knowledge engineering burden. The design and choice of the pairwise distances in ontologies also have to be studied further so that it remains anchored in real-life information while still scaling up. Taking into account missing data and some kind of confidence in the pairwise distance between individuals based on the number of variables used to calculate this distance has to be studied further. An interactive visualization could improve the overall usability.

Acknowledgements

We thank all rose gardens owners for providing access to their precious resources: Loubert rose garden (Rosiers-sur-Loire, France), Val-de-Marne rose garden (Haÿ-les-roses, France), La Cour de Commer rose garden (Cour-de-Commer, France), Jumaju rose garden (Montchamp, France), Jardin botanique de la Tête d'Or (Lyon, France), Désert rose garden (Bouzon Gellenave, France), Grande roseraie de Lyon (Lyon, France), La Beaujoire rose garden (Nantes, France), SCRADH (Hyères, France), Arboretum des Barres (Nogent-sur-Vernisson, France), and BRC RosePom (Beaucouzé, France). We also thank the Gentyane platform (Clermont-Ferrand, France) for the genomic data and Jordan Marie-Magdelaine for the preparation of the initial dataset.

Funding

This research was conducted in the framework of the regional program "Objectif Végétal, Research, Education and Innovation in Pays de la Loire", supported by the French Region Pays de la Loire, Angers Loire Métropole and the European Regional Development Fund, as part of the DIVIS project.

Abbreviations

- BRC: Biological Ressource Center
- IC: Information Content
- PCA: Principal Component Analysis
- MCA: Multiple Correspondance Analysis
- MDS: Multi-Dimensional Scaling
- OWL: Web Ontology Language
- SPARQL: SPARQL Protocol and RDF Query Language

Availability of data and materials

The software developed to implement the pipeline presented in this paper is available as follows:

- Project name: DIVIS
- Project home page: https://forgemia.inra.fr/irhs-bioinfo/Divis
- Archived version: v1.0
- Operating system(s): Platform independent
- Programming language: Python 3.7
- Other requirements: Described as requirement.txt file for pip in the code repository
- License: CeCILL. See LICENCE file in the code repository
- Any restrictions to use by non-academics: None

The OWL ontology (in French) is bundled with the code.
The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Authors' contributions

JB developed the methodology with the help of CL and EB. RE developed the initial Python code and JB refined it with the help of EB. AP provided the rosebush dataset, rosebush expertise and biological insight into the results. PS and AEG provided statistical expertise for both the development of the methodology and the analysis of the results. JB was the main contributor in writing the manuscript with significant inputs by CL, AEG and EB. All authors read and approved the final manuscript.

Author details

[1]Univ Angers, Institut Agro, INRAE, IRHS, SFR QUASAV, F-49000 Angers, France. [2]AGROCAMPUS OUEST, Institut Agro, 2 rue André le Nôtre, F-49045 Angers cedex 1, France.

References

1. Hendler, J.: Data integration for heterogenous datasets. Big Data **2**(4), 205–215 (2014). doi:10.1089/big.2014.0068
2. Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O.P., Tiwari, A., Er, M.J., Ding, W., Lin, C.-T.: A review of clustering techniques and developments. Neurocomputing **267**, 664–681 (2017). doi:10.1016/j.neucom.2017.06.053
3. Mehta, V., Bawa, S., Singh, J.: Analytical review of clustering techniques and proximity measures. Artificial Intelligence Review, 1–29 (2020). doi:10.1007/s10462-020-09840-7
4. Foss, A.H., Markatou, M.: kamila: Clustering Mixed-Type Data in R and Hadoop. Journal of Statistical Software, Articles **83**(13), 1–44 (2018). doi:10.18637/jss.v083.i13
5. Jaccard, P.: THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. New Phytologist **11**(2), 37–50 (1912). doi:10.1111/j.1469-8137.1912.tb05611.x
6. Dice, L.R.: Measures of the Amount of Ecologic Association Between Species. Ecology **26**(3), 297–302 (1945). doi:10.2307/1932409
7. Gower, J.C.: A General Coefficient of Similarity and Some of Its Properties. Biometrics **27**(4), 857–871 (1971). doi:10.2307/2528823
8. Ariosto Serna, L., Alejandro Hernández, K., Navarro González, P.: A K-Means Clustering Algorithm: Using the Chi-Square as a Distance. In: Tang, Y., Zu, Q., Rodríguez García, J.G. (eds.) Human Centered Computing, Mérida, Mexico, pp. 464–470 (2019). doi:10.1007/978-3-030-15127-0
9. Pandey, S.K., Das, A., Rai, P., Dasgupta, T.: Morphological and genetic diversity assessment of sesame (*Sesamum indicum L.*) accessions differing in origin. Physiology and Molecular Biology of Plants **21**(4), 519–529 (2015). doi:10.1007/s12298-015-0322-2
10. Pavoine, S., Vallet, J., Dufour, A.-B., Gachet, S., Daniel, H.: On the challenge of treating various types of variables: application for improving the measurement of functional diversity. Oikos **118**(3), 391–402 (2009). doi:10.1111/j.1600-0706.2008.16668.x
11. de Bello, F., Botta-Dukát, Z., Lepš, J., Fibich, P.: Towards a more balanced combination of multiple traits when computing functional differences between species. Methods in Ecology and Evolution **12**(3), 443–448 (2021). doi:10.1111/2041-210X.13537
12. Preud'homme, G., Duarte, K., Dalleau, K., Lacomblez, C., Bresso, E., Smaïl-Tabbone, M., Couceiro, M., Devignes, M.-D., Kobayashi, M., Huttin, O., Ferreira, J.P., Zannad, F., Rossignol, P., Girerd, N.: Head-to-head

comparison of clustering methods for heterogeneous data: a simulation-driven benchmark. Scientific Reports **11**(1), 4202 (2021). doi:10.1038/s41598-021-83340-8
13. Gruber, T.: A translation approach to portable ontology specifications. Knowledge Acquisition **5**, 199–220 (1993). doi:10.1006/knac.1993.1008
14. Walls, R.L., Athreya, B., Cooper, L., Elser, J., Gandolfo, M.A., Jaiswal, P., Mungall, C.J., Preece, J., Rensing, S., Smith, B., Stevenson, D.W.: Ontologies as integrative tools for plant science. American Journal of Botany **99**(8), 1263–1275 (2012). doi:10.3732/ajb.1200222
15. Kulmanov, M., Smaili, F.Z., Gao, X., Hoehndorf, R.: Semantic similarity and machine learning with ontologies. Briefings in Bioinformatics (2020). doi:10.1093/bib/bbaa199
16. Pesquita, C., Faria, D., Falcão, A.O., Lord, P., Couto, F.M.: Semantic Similarity in Biomedical Ontologies. PLOS Computational Biology **5**(7), 1–12 (2009). doi:10.1371/journal.pcbi.1000443
17. Guzzi, P.H., Mina, M., Guerra, C., Cannataro, M.: Semantic similarity analysis of protein data: assessment with biological features and issues. Briefings in Bioinformatics **13**(5), 569–585 (2011). doi:10.1093/bib/bbr066
18. Carpenter, J.R., Smuk, M.: Missing data: A statistical framework for practice. Biometrical Journal (2021). doi:10.1002/bimj.202000196
19. Johnson, T.F., Isaac, N.J.B., Paviolo, A., González-Suárez, M.: Handling missing values in trait data. Global Ecology and Biogeography **30**(1), 51–62 (2021). doi:10.1111/geb.13185
20. Cochran, W.G.: Sampling Techniques, 3rd Edition. John Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore (1977)
21. Tille, Y.: Sampling and Estimation from Finite Populations. John Wiley & Sons, Hoboken, Chichester (2020)
22. Liorzou, M., Pernet, A., Li, S., Chastellier, A., Thouroude, T., Michel, G., Malécot, V., Gaillard, S., Briée, C., Foucher, F., Oghina-Pavie, C., Clotault, J., Grapin, A.: Nineteenth century French rose (*Rosa sp.*) germplasm shows a shift over time from a European to an Asian genetic background. Journal of Experimental Botany **67**(15), 4711–4725 (2016). doi:10.1093/jxb/erw269
23. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. Nature **585**, 357–362 (2020). doi:10.1038/s41586-020-2649-2
24. Wes McKinney: Data Structures for Statistical Computing in Python. In: Stéfan van der Walt, Jarrod Millman (eds.) Proceedings of the 9th Python in Science Conference, pp. 56–61 (2010). doi:10.25080/Majora-92bf1922-00a
25. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011). doi:10.5555/1953048.2078195
26. Hunter, J.D.: Matplotlib: A 2D graphics environment. Computing in Science & Engineering **9**(3), 90–95 (2007). doi:10.1109/MCSE.2007.55
27. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal **6**(2), 167–195 (2015). doi:10.3233/SW-140134
28. Taylor, G.: colormath [Color math and conversion library for Python] (2009). https://pypi.org/project/colormath Accessed Accessed 03 June 2021
29. GeoPy Contributors: GeoPy [Geocoding library for Python] (2006). https://pypi.org/project/geopy/ Accessed Accessed 03 June 2021
30. Haklay, M., Weber, P.: OpenStreetMap: User-Generated Street Maps. IEEE Pervasive Computing **7**(4), 12–18 (2008). doi:10.1109/MPRV.2008.80
31. Musen, M.A.: The protégé project: A look back and a look forward. AI Matters **1**(4), 4–12 (2015). doi:10.1145/2757001.2757003
32. Lamy, J.-B.: Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. Artificial Intelligence in Medicine **80**, 11–28 (2017). doi:10.1016/j.artmed.2017.07.002
33. Härdle, W.K., Simar, L.: Applied Multivariate Statistical Analysis, 4th edn. Springer, Berlin, Heidelberg (2015). doi:10.1007/978-3-662-45171-7
34. Zhang, T., Ramakrishnan, R., Livny, M.: Birch: An efficient data clustering method for very large databases. In: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data. SIGMOD '96, pp. 103–114. Association for Computing Machinery, New York, NY, USA (1996). doi:10.1145/233269.233324
35. Reynolds, D.: Gaussian mixture models. In: Li, S.Z., Jain, A.K. (eds.) Encyclopedia of Biometrics, pp. 827–832. Springer, Boston, MA (2015). doi:10.1007/978-1-4899-7488-4_196
36. Ward, J.H.J.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association **58**(301), 236–244 (1963). doi:10.1080/01621459.1963.10500845
37. Lloyd, S.: Least squares quantization in pcm. IEEE Transactions on Information Theory **28**(2), 129–137 (1982). doi:10.1109/TIT.1982.1056489
38. Park, H.-S., Jun, C.-H.: A simple and fast algorithm for k-medoids clustering. Expert Systems with Applications **36**(2, Part 2), 3336–3341 (2009). doi:10.1016/j.eswa.2008.01.039
39. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. NIPS'01, pp. 849–856. MIT Press, Cambridge, MA, USA (2001)
40. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53–65 (1987). doi:10.1016/0377-0427(87)90125-7
41. Waskom, M.L.: seaborn: statistical data visualization. Journal of Open Source Software **6**(60), 3021 (2021). doi:10.21105/joss.03021

42. Greenacre, M., Blasius, J. (eds.): Multiple Correspondence Analysis and Related Methods, 1st edn. Chapman and Hall/CRC, New York (2006). doi:10.1201/9781420011319
43. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, 3rd edn. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, Chichester (2019)
44. Halford, M.: prince [Python factor analysis library (PCA, CA, MCA, MFA, FAMD)] (2020). https://github.com/MaxHalford/prince Accessed Accessed 25 June 2021
45. Brown, A.H.D.: Core collections: a practical approach to genetic resources management. Genome **31**(2), 818–824 (1989). doi:10.1139/g89-144
46. Corrado, G., Caramante, M., Piffanelli, P., Rao, R.: Genetic diversity in Italian tomato landraces: Implications for the development of a core collection. Scientia Horticulturae **168**, 138–144 (2014). doi:10.1016/j.scienta.2014.01.027
47. Papoutsoglou, E.A., Faria, D., Arend, D., Arnaud, E., Athanasiadis, I.N., Chaves, I., Coppens, F., Cornut, G., Costa, B.V., Cwiek-Kupczynska, H., Droesbeke, B., Finkers, R., Gruden, K., Junker, A., King, G.J., Krajewski, P., Lange, M., Laporte, M.-A., Michotey, C., Oppermann, M., Ostler, R., Poorter, H., Ramirez-Gonzalez, R., Rasak, Z., Reif, J.C., Rocca-Serra, P., Sansone, S.-A., Scholz, U., Tardieu, F., Uauy, C., Usadel, B., Visser, R.G.F., Weise, S., Kersey, P.J., Miguel, C.M., Adam-Blondon, A.-F., Pommier, C.: Enabling reusability of plant phenomic datasets with MIAPPE 1.1. New Phytologist **227**(1), 260–273 (2020). doi:10.1111/nph.16544
48. Cooper, L., Meier, A., Laporte, M.-A., Elser, J.L., Mungall, C., Sinn, B.T., Cavaliere, D., Carbon, S., Dunn, N.A., Smith, B., Qu, B., Preece, J., Zhang, E., Todorovic, S., Gkoutos, G., Doonan, J.H., Stevenson, D.W., Arnaud, E., Jaiswal, P.: The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic Acids Research **46**(D1), 1168–1180 (2017). doi:10.1093/nar/gkx1152
49. Masson, M., Denœux, T.: Multidimensional scaling of fuzzy dissimilarity data. Fuzzy Sets and Systems **128**(3), 339–352 (2002). doi:0.1016/S0165-0114(01)00162-2
50. Ramos-Guajardo, A.B., Ferraro, M.B.: A fuzzy clustering approach for fuzzy data based on a generalized distance. Fuzzy Sets and Systems **389**, 29–50 (2020). doi:10.1016/j.fss.2019.09.010

Figures

Additional Files

Additional file 1 — Silhouette analysis for the semantic distance
Silhouette values at the individual level for the KMeans algorithm, 2 to 19 clusters and the semantic distance.

Additional file 2 — Silhouette analysis for Gower's distance
Silhouette values at the individual level for the KMeans algorithm, 2 to 19 clusters and Gower's distance.

Additional file 3 — Concordance between algorithms, Gower's distance
Heatmaps of the concordance tables between KMeans clusters for 6 clusters (columns) and the other tested clustering algorithms (rows), Gower's distance. In each heatmap rows correspond to the KMeans clusters and columns to the clusters for the other algorithm. This other algorithm correspond to Birch, CAH (Hierarchical clustering) and Gaussian Mixture for the top three heatmaps and to KMedoids and Spectral Clustering for the two bottom ones.

Additional file 4 — Visualization with single archetype and Gower's distance
Clusters and archetype visualizations, single archetype, for Gower's distance, KMeans algorithm, 6 clusters.

Additional file 5 — Visualization with single archetype and Gower's distance
Clusters and archetype visualizations, multiple archetypes, for Gower's distance, KMeans algorithm, 6 clusters.

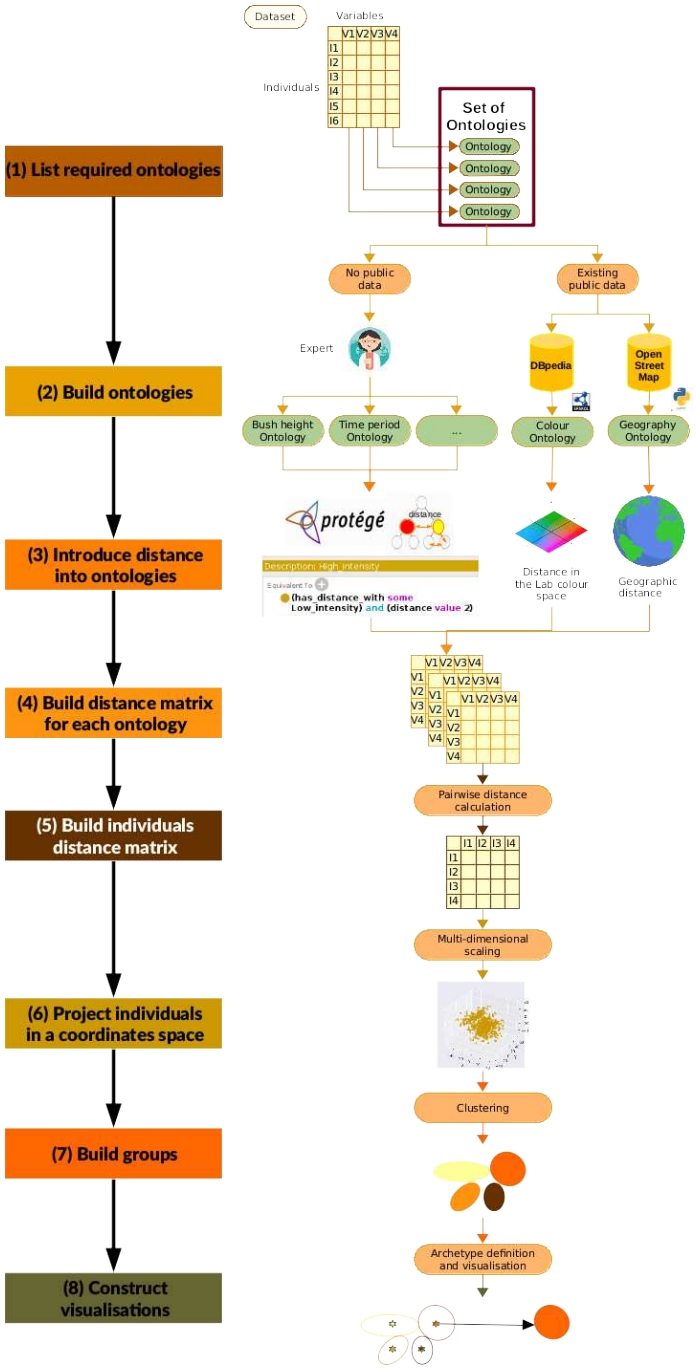

Figure 1 The dataset processing pipeline. Based on the list of qualitative variables we define the list of required ontologies (one for each variable). For each ontology, if relevant data are publicly available we retrieve it. Otherwise we rely on expert knowledge to build the ontology graph. We introduce distances between concepts in the ontologies based either on real life distances or expert knowledge. These ontologies including distance between concepts information are used to build a distance matrix between variable modalities for each qualitative variable. Based on the vector of variable values which represent each individual in the dataset we calculate pairwise distances to build a distance matrix between individuals. Individuals are then projected in a coordinates space using multi-dimensional scaling. Individuals coordinates are used during the clustering process to build groups. Representative individuals for each group are estimated to define the groups archetypes which are used as part of the visualizations.

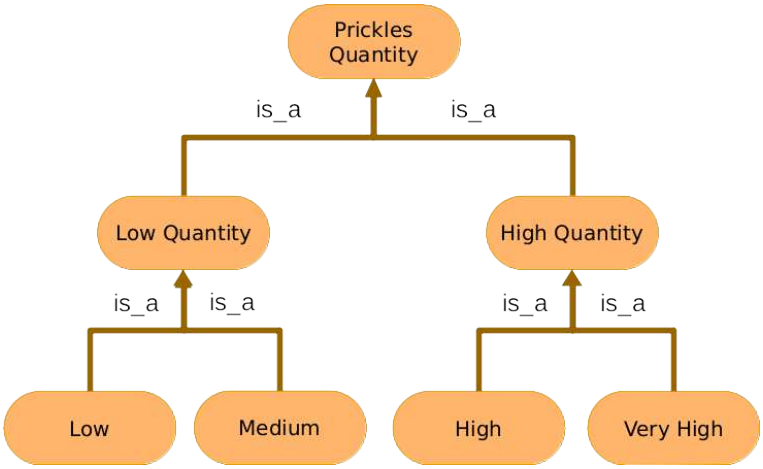

Figure 2 Ontology graph for the quantity of prickles variable.

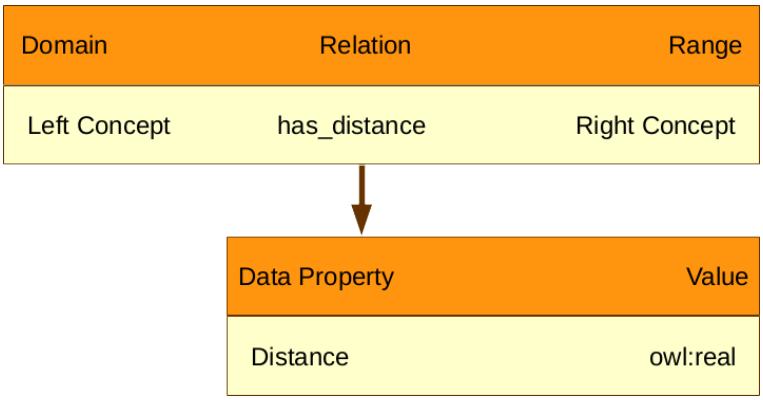

Figure 3 Principe used to store a distance between concepts in the OWL ontology.

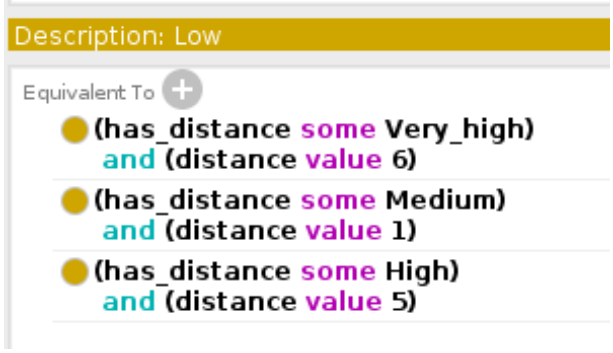

Figure 4 Quantity of prickles ontology - representation of the distances between the "Low" concept and the others.

Variables	Petal colour	Bush height	Quantity of prickles	Geographic origin	Breeding period
Indiv A	rose	low	medium	France	[1820-1829]
Indiv B	yellow	High	NaN	China	[1810-1819]
Distance D(A,B)	**d1**	**d2**	**NaN**	**d4**	**d5**

D(A,B) = (d1 + d2 + d4 + d5) / 4

Figure 5 Example of a calculation of distance between individual rosebushes.

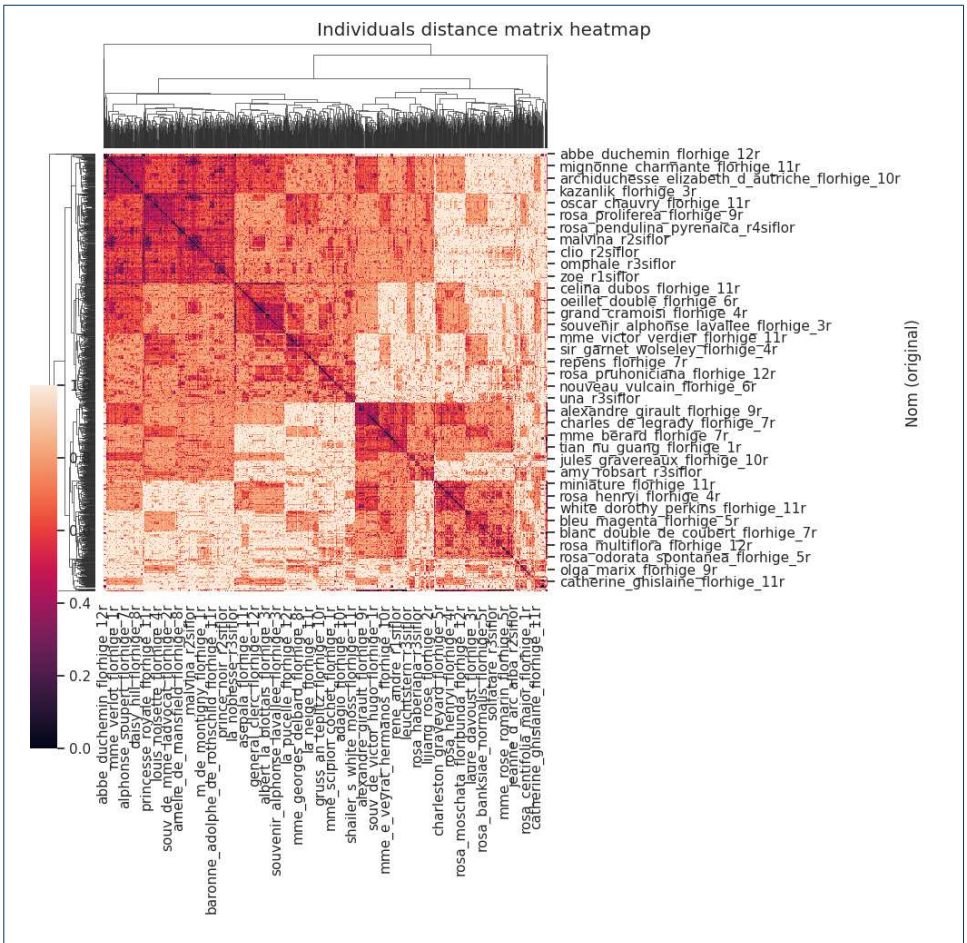

Figure 6 Distance matrix between rosebushes - Gower distance.

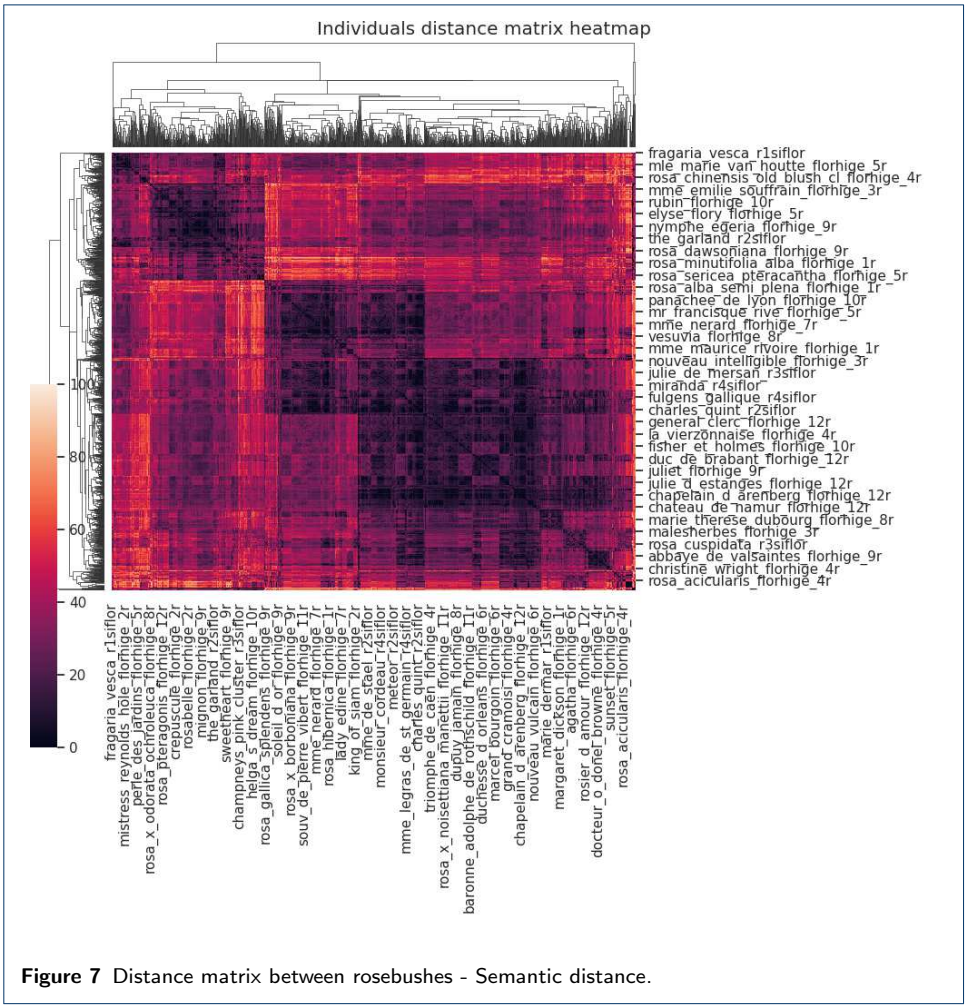

Figure 7 Distance matrix between rosebushes - Semantic distance.

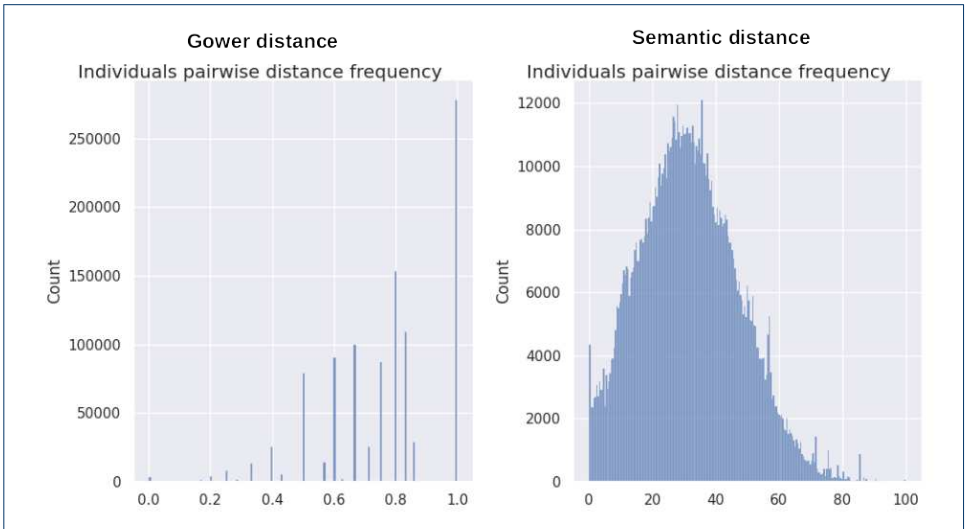

Figure 8 Frequency of distance values in the individuals distance matrices for Gower's distance and our semantic distance.

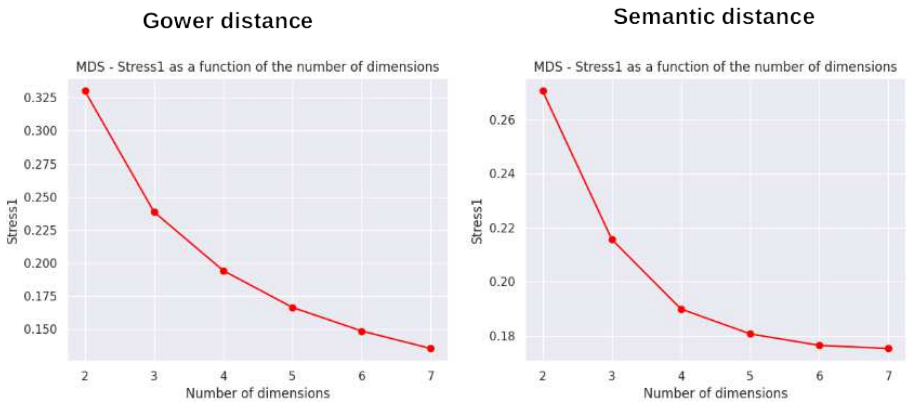

Figure 9 Value of $stress1$ according to the number of dimensions for the MDS. Distance matrix based on Gower's distance (left) and semantic distance (right).

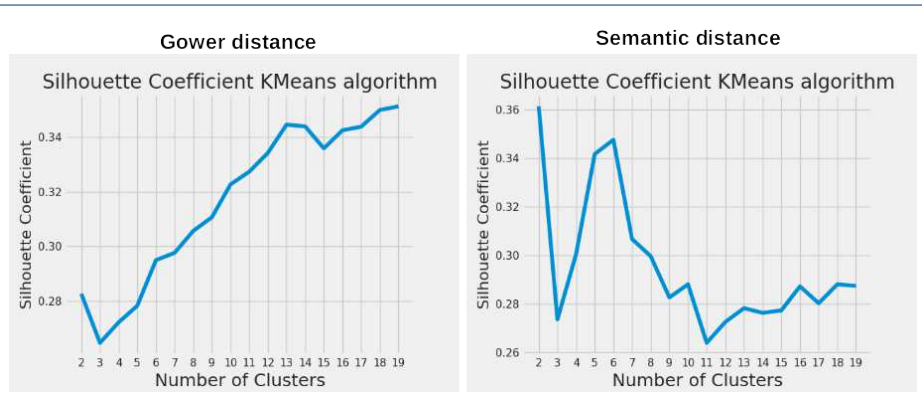

Figure 10 Mean Silhouette coefficient as a function of the number of clusters, KMeans algorithm. Gower's distance (left) and semantic distance (right).

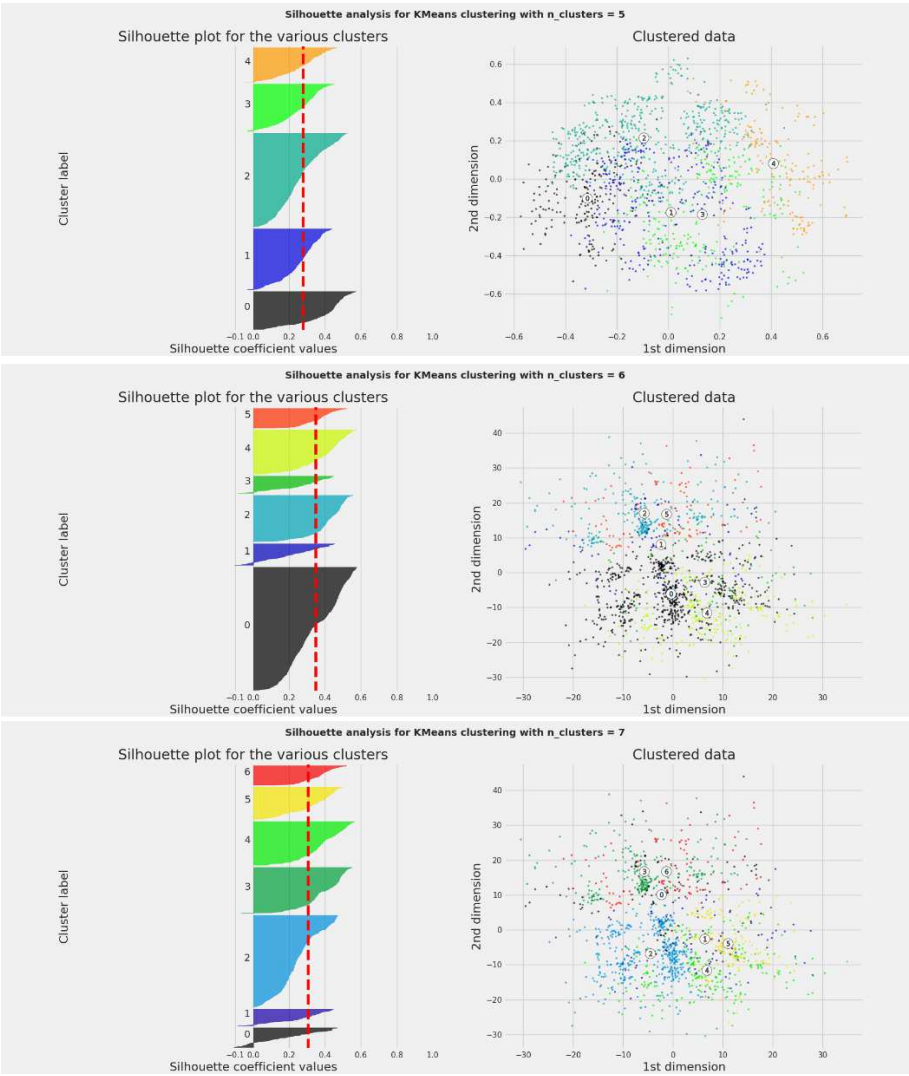

Figure 11 Silhouette analysis for 5, 6 and 7 clusters, KMeans algorithm, semantic distance. The left hand part of each figure presents the Silhouette value for each individual (colour coded per cluster) and the mean Silhouette value as the red dash vertical line. The right hand part presents the individuals projected in the first two dimensions.

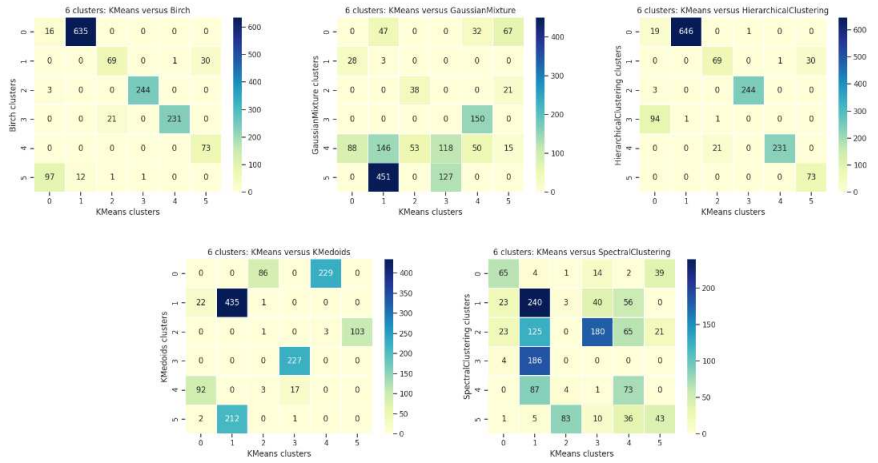

Figure 12 Heatmaps of the concordance tables between KMeans clusters for 6 clusters (columns) and the other tested clustering algorithms (rows), semantic distance. In each heatmap rows correspond to the KMeans clusters and columns to the clusters for the other algorithm. This other algorithm correspond to Birch, Hierarchical Clustering and Gaussian Mixture for the top three heatmaps and to KMedoids and Spectral Clustering for the bottom tow.

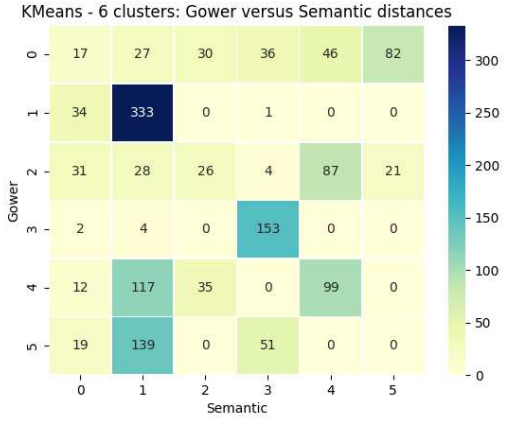

Figure 13 Heatmap of the concordance table between KMeans clusters for 6 clusters built with Gower's distance (rows) and the semantic distance (columns).

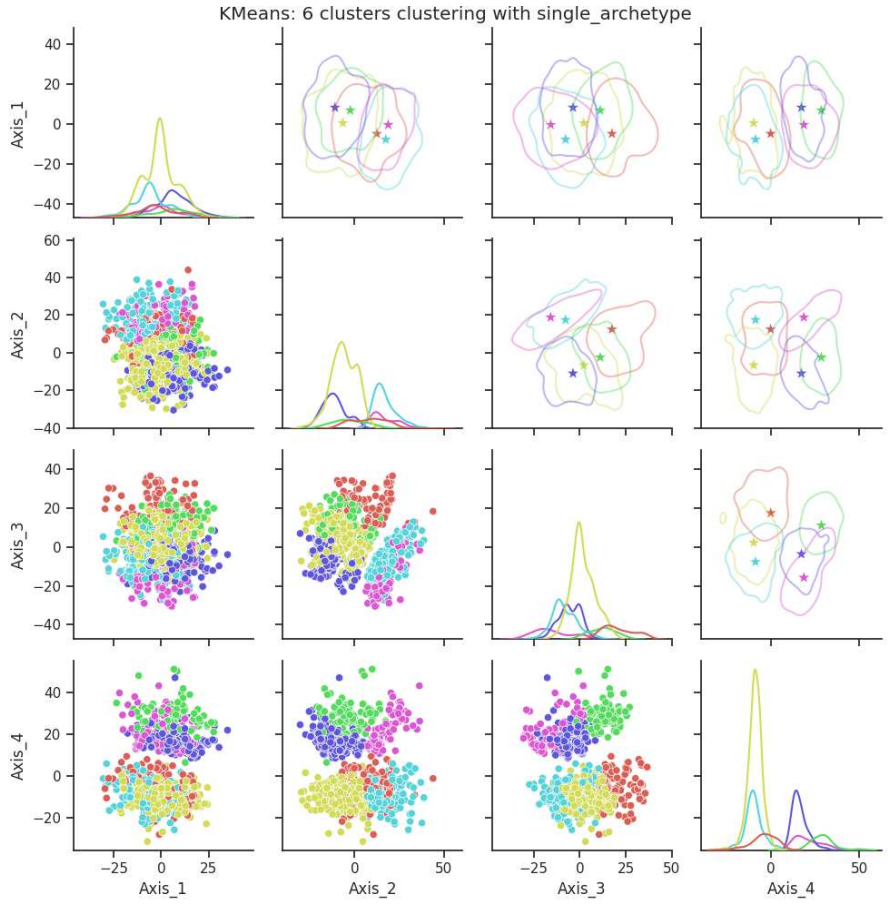

Figure 14 Clusters and archetype visualizations, single archetype, for the semantic distance, KMeans algorithm, 6 clusters.

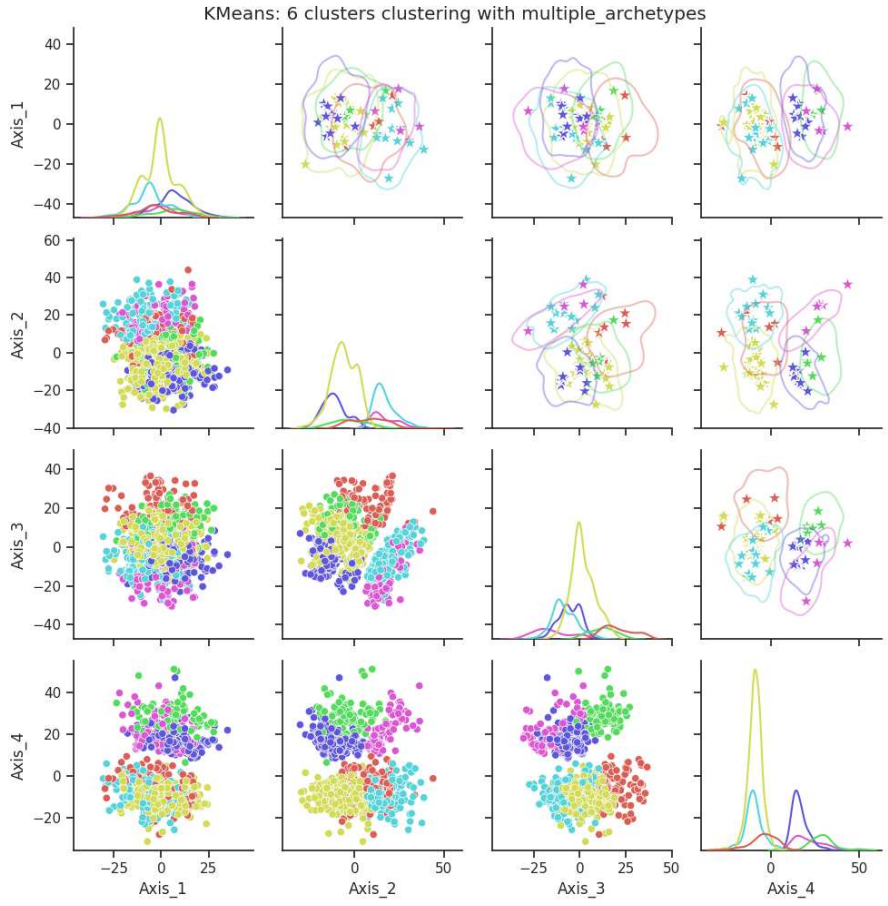

Figure 15 Clusters and archetype visualizations, multiple archetypes, for the semantic distance, KMeans algorithm, 6 clusters.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- additionalfile1.pdf
- additionalfile2.pdf
- additionalfile3.png
- additionalfile4.png
- additionalfile5.png