Archetypal analysis with missing data: see all samples by looking at a few based on extreme profiles

Irene Epifanio [*], M. Victoria Ibáñez and Amelia Simó
Departament de Matemàtiques-IMAC, Universitat Jaume I, Castelló 12071, Spain

October 12, 2018

Abstract

In this paper we propose several methodologies for handling missing or incomplete data in Archetype analysis (AA) and Archetypoid analysis (ADA). AA seeks to find archetypes, which are convex combinations of data points, and to approximate the samples as mixtures of those archetypes. In ADA, the representative archetypal data belong to the sample, i.e. they are actual data points. With the proposed procedures, missing data are not discarded or previously filled by imputation and the theoretical properties regarding location of archetypes are guaranteed, unlike the previous approaches. The new procedures adapt the AA algorithm either by considering the missing values in the computation of the solution or by skipping them. In the first case, the solutions of previous approaches are modified in order to fulfill the theory and a new procedure is proposed, where the missing values are updated by the fitted values. In this second case, the procedure is based on the estimation of dissimilarities between samples and the projection of these dissimilarities in a new space, where AA or ADA is applied, and those results are used to provide a solution in the original space. A comparative analysis is carried out in a simulation study, with favorable results. The methodology is also applied to two real data sets: a well-known climate data set and a global development data set. We illustrate how these unsupervised methodologies allow complex data to be understood, even by non-experts.

Keywords: Incomplete data set, Archetype Analysis, Multidimensional Scaling, Partial Distance Strategy

[*] This work is supported by the following grants: DPI2017-87333-R from the Spanish Ministry of Ciencia, Innovacin y Universidades (AEI/FEDER, EU) and UJI-B2017-13 from Universitat Jaume I.

1 Introduction

In everyday life, archetypes are original examples or perfect models that embody the fundamental characteristics of a thing. For example, in Greek mythology there are many hybrid beings composed of archetypes, such as a minotaur that is a mixture of two archetypes, a man and a bull. Archetypes are common in behavior, modern psychological theory and literary analysis, and they form the basis of eclecticism art. The concept of archetypes or pure types in Statistics follows the same ideas and was formulated by Cutler and Breiman (1994). It is a matrix factorization technique where the data are explained as compositions of a few pure patterns.

Data mining seeks to discover unknown and unanticipated structure in the data (Hand, 1998; Costantini et al., 2010), together with the visualization of that structure (Daszykowski et al., 2003). Apart from understanding and describing the entire data set, we would like to be able to extract information that is easily interpretable, even by non-experts. Dimension reduction techniques are very useful for exploring multivariate data (Larose, 2006; Giudici and Figini, 2009; Fogel et al., 2013). As only input and no output features are present, this is an unsupervised statistical learning problem (see Hastie et al. (2009, Chapter 14) for a complete review of unsupervised learning techniques). Data decomposition techniques are frequently used to find the latent components. A data set is viewed as a linear combination of several factors. Different unsupervised techniques with specific objectives are generated depending on the constraints on the factors and how they are combined (Mørup and Hansen, 2012; Thurau et al., 2012; Vinué et al., 2015). At one extreme, we find Principal Component Analysis (PCA), which explains data variability satisfactorily at expense of the interpretability of the factors, since this is compromised due to the construction of the factors as linear combinations of variables. At the other extreme, we find clustering techniques such as k-means or k-medoids, whose factors are readily interpreted. These factors are the centroids of the clusters, which are averages of groups of data in the case of k-means and medoids (concrete instances) in the case of k-medoids. Nonetheless, in contrast with PCA, their modeling flexibility is undermined by the binary assignment of data to the clusters.

1.1 Archetypal analysis

Archetype analysis (AA) lies somewhere in between these two techniques, since its factors can be interpreted as easily as those of clustering methodologies, but its modeling flexibility is higher than for clustering techniques. A table summarizing the relationship between several unsupervised multivariate techniques is provided by Mørup and Hansen (2012) and Vinué et al. (2015).

Cutler and Breiman (1994) formulated AA in such a way that each instance of the data set is approximated by a mixture (convex combination) of pure or extremal types called archetypes. Furthermore, archetypes are built as mixtures of the cases in the data set. Although archetypes are easily interpretable, as they are artificial constructions, there may not be instances in the data set with characteristics similar to those of the archetypes, which may make it unsuitable in some fields (Seiler and Wohlrabe, 2013). To solve this question the new concept of archetypoids was introduced by Vinué et al. (2015). In Archetypoid Analysis (ADA) each instance in the data set is approximated by a mixture of a set of actual extreme observations called archetypoids.

This procedure not only allows us to relate the cases in the data set to extreme patterns but also facilitates comprehension of the data. Humans understand the data better when the instances are shown through their extreme constituents (Davis and Love, 2010) or when characteristics of one instance are shown as opposed to those of another (Thurau et al., 2012). In fact, Jones and Rice (1992) used functions with extreme principal component scores to describe and display the important characteristics of a set of functions. This could be considered as seeking the archetypoid functions. Nonetheless, unlike PCA, the objective of AA is to recover extreme instances, and functions with extreme PCA scores do not necessarily correspond to archetypal cases. This is explained in Cutler and Breiman (1994) and shown in Epifanio et al. (2013) through a problem where archetypes could not be restored with PCA even if all the components had been taken into account. Not only that, Stone and Cutler (1996) also showed that AA may be more appropriate than PCA when the data do not have elliptical distributions. In summary, as regards human interpretability, the central points returned by clustering techniques do not seem as favorable as extreme types, which are also more readily comprehensible than a linear combination of data, such as that returned by PCA.

This has meant that the applications of AA and ADA have spread to many different fields, such as astrophysics (Chan et al., 2003), biology (D'Esposito et al., 2012), climate (Steinschneider and Lall, 2015; Su et al., 2017), developmental psychology (Ragozini et al., 2017), e-learning (Theodosiou et al., 2013), genetics (Thøgersen et al., 2013), human development (Epifanio, 2016), industrial engineering (Epifanio et al., 2013, 2018; Millán-Roures et al., 2018), machine learning (Mørup and Hansen, 2012; Seth and Eugster, 2016a,b; Ragozini and D'Esposito, 2015), market research (Li et al., 2003; Porzio et al., 2008; Midgley and Venaik, 2013), multi-document summarization (Canhasi and Kononenko, 2013, 2014), nanotechnology (Fernandez and Barnard, 2015), neuroscience (Tsanousa et al., 2015; Hinrich et al., 2016) and sports (Eugster, 2012; Vinué and Epifanio, 2017).

1.2 Missing data

AA and ADA, like the majority of statistical techniques, assume the completeness of the data. Nevertheless, incomplete data, i.e. data with missing values, are common in real applications (Lott and Reiter, 2018; Wang and Johnson, 2018; Xia and Yang, 2016; Akande et al., 2017; Maity et al., 2018). According to Little and Rubin (2002), three types of missing data can be established: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In simple terms, MCAR means that the probability that an observation is missing is not related to its value or to any other values in the data set, unlike MAR, where that probability is related to the values for some other observed variables. However, when that probability is related to its value, we say that the data are not missing at random (MNAR).

One way to solve this problem is to ignore the incomplete instances and to work only with the complete cases. However, this means discarding information that may be valuable, especially if the percentage of missing cases is high. Another way to solve this problem is to impute data, and once the missing values are filled, then proceed with the statistical procedure as normal. This approach could work well if the percentage of missing values is not too high, otherwise errors derived by imputation become increasingly important (Eirola et al., 2013). Furthermore, probabilistic imputation methods are usually computationally expensive. Another approach in factorial methods and cluster analysis is to adapt the corresponding algorithm by either skipping or considering the missing values in the computation of outputs (Dray and Josse, 2015). Some PCA algorithms that consider missing values are the Non-linear Iterative Partial Least Squares (NIPALS) algorithm (Dray and Dufour, 2007) or the iterative PCA method (Kiers, 1997), also known as the EM-PCA algorithm (Josse and Husson, 2012), while in the clustering case some examples are the k-POD algorithm (Chi et al., 2016) and mixture model clustering to handle missing data (Hunt and Jorgensen, 2003). With regard to PCA algorithms that skip missing values, i.e. the PCA algorithm is adapted so that missing values are not considered in the computation, some examples are the pairwise correlation approach and the approach based on Principal Coordinates Analysis of the Euclidean distance matrix computed between the individuals as described by Dray and Josse (2015), while in the clustering case we can cite, for instance, the Partitioning Around Medoids (PAM) algorithm (Kaufman and Rousseeuw, 1990, Ch. 2).

Two approaches for computing AA with incomplete data have been considered to date, the first by Mørup and Hansen (2012) and the second by Epifanio et al. (2018). Both approaches modified the original AA algorithm to consider the missing values. Nonetheless, in both approaches the results did not fit with the expected theoretical results, as will be shown in the supplemental material. On the one hand, we propose a modification of the solutions of both approaches in order to fulfill the theory. Furthermore, we propose a new procedure, which also considers the missing values, based on updating the missing values with the fitted values.

On the other hand, we consider a different point of view that could, in fact, be used not only for AA but also for ADA and that skips missing values. We propose to estimate the Euclidean distances between cases, then a multidimensional scaling technique is applied to these dissimilarities so that distances between the points returned in the Euclidean space are approximately equal to the original dissimilarities. Once complete vectors are available, AA or ADA can be applied and then those results are used to build the AA and ADA solution in the original space. In other words, as we cannot work in the original space due to the incompleteness of the data, a kind of projection is performed on the data (in fact, on the dissimilarity matrix), and the concrete statistical technique is applied in this new space. Then, we take advantage of the learning obtained in this new space to provide a solution in the initial space. Note that if the dissimilarities in n cases were Euclidean distances, they could be represented exactly in at most n - 1 dimensions (Mardia et al., 1979, Theorem 14.4.1) by means of classical multidimensional scaling (cMDS), so we could recover the original vectors and the solution in the cMDS space would coincide with that of the initial space.

In summary, the aim of this paper is to propose a new methodology with missing data for AA and ADA that guarantees the fulfillment of the theoretical properties regarding the location of archetypes and archetypoids. Furthermore, on the one hand, the proposed procedures do not substantially increase the computational burden of the original algorithms, and on the other hand, they do not require any assumptions on the missingness patterns due to the absence of the completely observed data formulation. Moreover, these methods can provide an imputation for the missing values despite this not being their objective. With the proposed procedures, missing values can appear in the archetypal profiles, which conforms to the fact that instances with missing values can be archetypal cases.

The outline of the paper is as follows: In Section 2 we review archetype and archetypoid analysis and the current AA algorithms that consider missing values. In Section 3 we introduce our proposal for handling missing data with archetypal analysis. A toy example illustrates our proposed methodology and the flaws of the previous approaches in the supplementary material. In Section 4, a comparison is made in a simulation study. In Section 5, our proposal is applied to two real data sets. Section 6 contains conclusions and some ideas for future work. The data sets and code in R (R Development Core Team, 2018) for reproducing the results are available as supplementary material.

2 Background

2.1 AA and ADA with complete data

Let $\mathbf{X}$ be an $n \times m$ matrix with n instances and m variables. Three matrices are searched in AA: 1) the $k \times m$ matrix $\mathbf{Z}$, whose rows contain the k archetypes $\mathbf{z}_j$; 2) an $n \times k$ matrix $\alpha = (\alpha_{ij})$ with the mixture coefficients that approximate each instance $\mathbf{x}_i$ by a mixture of the archetypes ($\hat{\mathbf{x}}_i = \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j$); and 3) a $k \times n$ matrix $\beta = (\beta_{jl})$ with the mixture coefficients that set each archetype ($\mathbf{z}_j = \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l$). To establish these matrices, the minimization of the following residual sum of squares (RSS) with the respective restrictions is carried out ($\|\cdot\|$ denotes the Euclidean norm):

$$RSS = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l\|^2, \tag{1}$$

under the constraints

a) $\sum_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, \ldots, n$ and

b) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \geq 0$ for $j = 1, \ldots, k$.

It is imperative to highlight that archetypes do not necessarily match actual instances. Specifically, this will only occur when one and only one β_{jl} is equal to one for each archetype, i.e. when each archetype is made up of only one instance. Accordingly, in ADA the former restriction b) is modified by the following one, and as a result the former continuous optimization problem of AA is converted into a mixed-integer optimization problem:

b) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \in \{0, 1\}$ and $j = 1, \ldots, k$.

To solve the AA problem, Cutler and Breiman (1994) developed an alternating minimizing algorithm that alternates between computing the optimum α for given archetypes $\mathbf{Z}$ and the best archetypes $\mathbf{Z}$ for a given α. This involves solving convex least squares problems at each stage, which is done through a penalized version of the non-negative least squares algorithm (Lawson and Hanson, 1974). That algorithm was implemented in the R package **archetypes** by Eugster and Leisch (2009), although with various changes (previous data standardization and use of the spectral norm in equation 1 instead of the Frobenius norm for matrices). However, in our R implementation these changes do not occur and the data are not standardized by default, and equation (1) is indeed the objective function to minimize. A total of 20 random starts is considered in the experiments and the best repetition is chosen as the solution.

To solve the ADA problem, Vinué et al. (2015) proposed an algorithm inspired by the scheme of the PAM clustering algorithm (Kaufman and Rousseeuw, 1990). Two phases make up the algorithm: the BUILD stage and the SWAP stage. A starting set of archetypoids is determined in the BUILD stage, which is upgraded during the SWAP stage by swapping the chosen instances for unselected observations and inspecting whether these interchanges decrease the RSS. That algorithm was implemented by Vinué (2017) in the R package **Anthropometry** with three possible starting sets in the BUILD phase: $cand_{ns}$, $cand_\alpha$ and $cand_\beta$. The nearest neighbors in Euclidean distance to the k archetypes form the $cand_{ns}$ set. The $cand_\alpha$ set is constituted by the instances with the maximum α value for each archetype j, while the $cand_\beta$ set is composed of the instances with the maximum β value for each archetype j. Each of these three sets goes through the SWAP stage and three sets are retrieved. From these three sets, the one with the minimum RSS (often the same set is returned from the three initializations) is chosen as the ADA solution.

Let us see the locations of the archetypal representatives. On the one hand, if $k = 1$, the archetype corresponds to the mean and the archetypoid to the medoid (Kaufman and Rousseeuw, 1990). Nevertheless, if $k > 1$, the archetypes are located on the boundary of the convex hull of the data (see Cutler and Breiman (1994)), although this is not necessarily the case for archetypoids (see Vinué et al. (2015)). On the other hand, archetypes are not necessarily nested, and neither are archetypoids. As a consequence, different ks may reflect different structures of the data. As with any unsupervised technique, the selection of the number of components k is an open question. If the user has prior knowledge of the arrangement of the data, the value of k can be chosen based on that information. Otherwise, the elbow criterion, which has been widely used (Cutler and Breiman, 1994; Eugster and Leisch, 2009; Vinué et al., 2015) could be considered. The elbow criterion means representing the RSS for different k values and selecting the value k as the point where the elbow is found.

2.2 Previous techniques for AA with missing data

Let $\mathbf{X}$ be an $n \times m$ matrix as before, but now missing values (NAs) can be present in the data. Let us suppose that there is no row or column with all its values missing; otherwise, that row or column would have to be erased. The simplest approach is to remove the incomplete instances and analyze only the complete ones, with which usable information is wasted. This approach is referred to as COM. Another approach, which is referred to as IMP, is to estimate the missing values and analyze all the instances. Missing value estimators range from the simplest one, such as using the mean values of the non-missing values in the respective feature, to more complex ones that exploit the information from other features. In the experiments carried out in Sections 4 and 5, multiple imputation using additive regression, bootstrapping and predictive mean matching is used, which is implemented with the *aregImpute* function in the R package **Hmisc** (Harrell Jr et al., 2016) (the default type 'pmn' is only changed to 'regression' if imputations are not completed due to a high percentage of missingness). The number of multiple imputations is 5 and the 5 imputed data sets are appended to be combined.

To the best of our knowledge, two specific techniques have been developed for AA with missing data. The first was introduced by Mørup and Hansen (2012), which is referred to as AAMOHAN. For handling incomplete cases, the objective function for minimizing RSS was modified and includes a parameter (ϵ) as a regularization for avoiding division by zero. In the experiments, two values for ϵ are used: 1e-3, the default, and 1e-9. Furthermore, in the implementation, the missing values are substituted by zeros, which could mean that archetypes were located outside the convex hull. As a consequence, the results did not fit with the expected theoretical results. The second was introduced by Epifanio et al. (2018) and is referred to as AAEIS. Different weights for non-missing and missing values (zero in that case) are considered in order to solve the problem. However, those weights could also mean that archetypes were located outside the convex hull of the data for $k > 1$.

3 Proposed methodology

3.1 Procedures using dissimilarities

3.1.1 Projecting dissimilarities for computing AA and ADA

The outline of the procedure is as follows:

Step 1 Estimate an $n \times n$ dissimilarity matrix from all the pairwise dissimilarities between data points.

Step 2 Project the dissimilarities in such a way that the cases are embedded in a Euclidean space.

Step 3 Apply AA or ADA to the projected space according to the analysis pursued. Keep β.

Step 4 Compute the archetypes or archetypoids in the original space with β from Step 3. Compute α and RSS in the original space.

For Step 1, the dissimilarities are estimated directly; no previous missing data imputation is carried out. This leads to more reliable estimates according to Eirola et al. (2013).

In the experiments carried out in Sections 4 and 5, the well-known Partial Distance Strategy (PDS) is used (Dixon, 1979). PDS estimates the squared Euclidean distance by calculating the sum of squared differences of the mutually known variables, and scaling the value proportionally to account for the missing values. The Euclidean distances (ED) are estimated in this way by *daisy* function from the R package **cluster** (Maechler et al., 2017) and recorded in the matrix $\mathbf{D}$ with elements d_{ij}.

Many mappings can be chosen for Step 2. In the experiments in Sections 4 and 5, only two are considered. The first is the well-known classical multidimensional scaling (cMDS), which takes a set of dissimilarities and returns a set of points such that the Euclidean distances between these points are approximately equal to those dissimilarities. If possible, we consider n - 1 as the maximum dimension of the space in which the data are to be represented. If the dissimilarities are Euclidean distances, they can be represented exactly in at most n - 1 dimensions (Mardia et al., 1979, Theorem 14.4.1). $\boldsymbol{D}$ is only Euclidean if $\boldsymbol{B}$ is positive semidefinite (Mardia et al., 1979, Theorem 14.2.1), where $\boldsymbol{B} = (\boldsymbol{I} - n^{-1}\boldsymbol{e}\boldsymbol{e}')\boldsymbol{M}(\boldsymbol{I} - n^{-1}\boldsymbol{e}\boldsymbol{e}')$, $\boldsymbol{I}$ is the $n \times n$ identity matrix, $\boldsymbol{e}$ is the $n \times 1$ vector with all its elements equal to unity and $\boldsymbol{M}$ is a matrix with elements m_{ij} = -0.5* d^2_{ij}. Due to the missing values, in the experiments $\mathbf{D}$ is not Euclidean. Therefore, in second place we also consider a method that also works when the dissimilarity is not a distance, the h-plot (HP) (Epifanio, 2013). Note that cMDS tries to preserve the original interpoint dissimilarities in the new space, unlike h-plot, which tries to preserve relationships between dissimilarity variables. This perspective is particularly suitable when working with non-metric dissimilarities, because the dissimilarities cannot be represented exactly in a Euclidean space, since the matrix is not Euclidean, as in this case. For the HP method, both the estimated Euclidean distances and their squares, as originally defined by Dixon (1979), have been used in the experiments, since raising dissimilarities to a power could be useful in this method (Epifanio, 2013).

For Step 4, the archetypoids can be directly determined in the original space, as they correspond to concrete cases. But to determine the archetypes it is necessary to define $\mathbf{Z} = \beta \times \mathbf{X}$ when $\mathbf{X}$ has missing values. Note that archetypes are mixtures of the observations, i.e. they are defined as weighted averages of the observations. Therefore, if we have missing values, those weights have to be scaled to account for the missing values in a similar way to PDS. In particular,

$$z_{jh} = \sum_{l=1}^{n} \beta_{jl} \frac{x_{lh} w_{lh}}{\sum_{l=1}^{n} \beta_{jl} w_{lh}} \;\; \forall j = 1, \ldots, k, \text{ and } \;\forall h = 1, \ldots, m, \tag{2}$$

and where $\mathbf{W}$ is an $n \times m$ matrix with 0 whenever the element is missing in $\mathbf{X}$ and 1 otherwise.

Then, we find the best α for a given $\mathbf{Z}$ by solving n convex least squares problems ($i = 1, \ldots, n$). It should be noted that each problem is independent of the rest, and its result can be calculated by excluding the coordinates with missing values

$$min_{\alpha_i} \|x_i - \alpha_i Z\|_2 \tag{3}$$

subject to $\alpha_i \geq 0$ and $\sum_{j=1}^{k} \alpha_{ij} = 1$.

Finally, RSS can be computed as the sum of the n distances between $\mathbf{x}_i$ and $\sum_{j=1}^{k} \alpha_{ij} \mathbf{z}_j$ using PDS, which is the same approach followed by Epifanio et al. (2018). Note that we can provide an imputation for the missing values, despite this not being the objective, by using $\hat{\mathbf{x}}_i$.

In summary, for AA or ADA, according to the analysis to be carried out, three possible alternatives are considered in the experiments, since we contemplate two mappings in Step

2 and two different dissimilarities in Step 1 for the second mapping, which handles non-metric dissimilarity matrices. These three combinations will be labeled as: AAEDcMDS, archetypal analysis using Euclidean distance in step 1 and cMDS in Step 2; AAEDHP, archetypal analysis using Euclidean distance in Step 1 and HP in Step 2, and AAPDSHP, archetypal analysis using partial distance in Step 1 and HP in Step 2.

3.1.2 Kernel method for AA

Instead of projecting the dissimilarities, the kernel-AA algorithm by Mørup and Hansen (2012), which generalizes the AA algorithm to kernel representations, can be used to compute AA, as it is based on pairwise relations between the data points. So, we need to define the kernel matrix. In particular, the Gaussian radial basis function $K(x, y) = exp(-\gamma\|x - y\|^2)$, which is one of the most popular choices for a kernel function, can be formulated only in terms of the dissimilarities between samples. We use this kernel in the experiments, with the parameter that sets the spread of the kernel, γ, equal to 0.1. No significant improvements are observed if we use other γ values. This approach is referred to as AAK.

3.2 Modified AAMOHAN and AAEIS for AA

To ensure that the archetypes are not outside the convex hull, instead of working with the archetypes returned by AAMOHAN and AAEIS, we consider the β values returned by these methods, and build the archetypes as in equation 2, and estimate the α values by means of equation 3. This approach is referred to as MAAMOHAN and MAAEIS. Furthermore, estimating the archetypes in this way allows there to be archetypes with missing values.

3.3 Intrinsic imputation for AA

The original AA algorithm can be adapted to handle missing data by making internal imputations during the parameter updates. In each iteration of the standard iterative alternating procedure of estimating α and β, an imputation step is introduced, in which missing data entries are imputed according to the current AA model. For the initialization, we need to start with archetypes without missing values and that fulfill the constraints. So, the starting archetypes are built considering only complete cases. This does not prevent us from obtaining archetypes formed by mixtures of cases with missing values, during the iterative phase of improvement. To fulfill the theory, i.e. that archetypes are a mixture of the data, the final archetypes are built from β values returned by this algorithm, as in equation 2, and α values are estimated by means of equation 3. This approach is referred to as AAII.

3.4 ADA with missing values

The methods by Mørup and Hansen (2012) are not defined for ADA, while extending AAEIS and AAII to ADA would imply that only the complete cases could act as archetypoids, as explained by Epifanio et al. (2018). This would considerably restrict the feasible set of solutions. Therefore, we consider that the best options for computing ADA with missing data are the methodology introduced in Section 3.1.1 and applying ADA after imputing values.

4 Simulation study

Three simulation studies are carried out to compare AA with missing data. Two of them are similar to those followed by Epifanio et al. (2018), and the other is inspired by Chi et al. (2016). A toy example, where we show that AAMOHAN and AAEIS do not fulfill the theory, is analyzed in the supplemenatry material. Here, two well-known benchmark data sets are analyzed: an artificial one (waveform data) and a real one (wine data set). Both of them have a factor with labels that is discarded in the computation but is used to assess the solution. Although the objective of AA is not data clustering, we can assign each case to the group in which its corresponding alpha is the maximum. We will use this to compare the clustering AA solutions with those of k-POD in two sets in which differentiated (separate) clusters do not exist. Like Chi et al. (2016), we use the Rand score (Rand, 1971) to compare each clustering result to the true class label variable, with the *adjustedRand* function in the R package **clues** (Chang et al., 2010). This index ranges between 0 and 1, where higher scores indicate greater agreement and, as a consequence, more accurate clustering performance. The algorithm for k-POD is available in the R package **kpodclustr** (Chi and Chi, 2014).

4.1 Simulation study with wine and waveform data

We follow the same experimental set-up performed by Chi et al. (2016) and Epifanio et al. (2018). The first set considered is the wine data set from the UCI Machine Learning repository (Dheeru and Karra Taniskidou, 2017). A total of 13 chemical analyses are recorded for each of the 178 samples, together with the wine type (there are three classes). Secondly, a benchmark data set, the waveform data set defined by Breiman et al. (1984, pp. 49-55), is generated in each trial using the function *mlbench.waveform* from the R package **mlbench** (Leisch and Dimitriadou, 2010). Each data set is constituted by $n = 150$ samples and $m = 21$ continuous variables, together with a factor recording the 3 classes (33% for each of the 3 classes). Each class is generated from a convex combination of 2 of 3 "base" waves.

Each of the following steps is repeated 100 times. To simulate the MCAR mechanism, we randomly remove entries to obtain approximately 10 and 30% overall missingness in the wine data set and 50% overall missingness in the waveform data sets. To simulate the MAR mechanism, we randomly remove 50% of the values in the 1st, 4th, and 7th variables in the wine data set, and in the 5th, 10th and from the 15th to 21st variables in the waveform data sets, which means approximately 12% and 21% overall missingness, respectively. To simulate the MNAR mechanism, we randomly remove 95% of the entries in the bottom 25th quantile in each of the variables in the wine data set, and 75% of the entries in the bottom 50th quantile in each of the variables of the waveform data set, which means approximately 24% and 38% overall missingness, respectively. We consider high percentages of missingness since, although they were uncommon in the past, nowadays they are frequent in data from online social networks and recommender systems, as explained in Chi et al. (2016). The different approaches are then applied to each data set with $k = 3$. The data sets generated from wine are standardized. As in the previous Section, $\mathbf{D}$ is not Euclidean.

A summary (mean and standard deviation) of the RSS/n from each approach is displayed in Table 1. As we also know the original data sets, we compute the matrix α that approximates the original data (without missing values) using the archetypes kept for each strategy and compute the RSS/n. The idea is to judge the capacity of each method to recover the original data. The results obtained using the original data without missing values are referred to as ORG. A summary (mean and standard deviation) of these quantities is displayed in Table 2. In Table 3 we show the summary of the Frobenius norm of the difference between the archetypes obtained with each approach and the original ones (we consider the permutation that gives the least error to match the archetypes). Finally, a summary of the Rand scores is shown in Table 4. The best result for each scenario is highlighted in bold. An empty space in the tables indicates failure to complete the experiment for a given scenario. This happens with waveform data due to the high percentage of missingness. For example, there are no complete cases. This means that AAII could not be initialized. For that reason, if there are no complete cases, we have imputed the minimum value of each variable to the missing entries to build the starting archetypes; for the rest of the algorithm this imputation is discarded. For IMP, in some cases the imputation mechanism fails with the first seed used, then another seed is used to run the imputation function. For the MCAR mechanism with the waveform data, there are a few missing entries in the dissimilarities, since a few rows do not have any variables without missing values in common. Therefore, we have worked with the complete cases of the dissimilarity matrix.

Table 1: Mean RSS/n (st. deviation in brackets) for each approach for the wine and waveform data.

	Wine				Waveform		
Method	MCAR(10%)	MCAR(30%)	MAR	MNAR	MCAR	MAR	MNAR
COM	7.255 (0.313)	15.64 (4.289)	8.304 (0.659)	9.785 (0.092)	-	-	-
IMP	**6.583** (0.111)	8.172 (0.260)	7.315 (0.137)	8.223 (0.066)	39.32 (2.580)	24.29 (0.843)	26.56 (1.021)
MAAMOHAN 1e-3	6.601 (0.109)	**8.130** (0.525)	6.800 (0.259)	**8.178** (0.060)	**35.97** (1.790)	24.18 (0.818)	26.25 (0.950)
MAAMOHAN 1e-9	6.601 (0.109)	8.130 (0.524)	**6.800** (0.259)	8.178 (0.060)	35.98 (1.792)	**24.18** (0.818)	**26.24** (0.946)
MAAEIS	6.710 (0.112)	8.561 (0.304)	6.871 (0.083)	8.538 (0.089)	41.02 (2.426)	24.77 (0.857)	42.57 (7.227)
AAII	6.626 (0.111)	8.236 (0.254)	6.830 (0.065)	8.253 (0.060)	36.90 (19.17)	24.35 (0.819)	26.94 (1.025)
AAK	7.219 (0.126)	9.767 (0.536)	7.415 (0.083)	9.237 (0.104)	69.85 (17.24)	31.21 (2.756)	50.90 (5.563)
AAEDcMDS	6.699 (0.113)	8.434 (0.259)	6.849 (0.064)	8.600 (0.096)	39.38 (2.285)	24.58 (0.822)	28.19 (1.086)
AAEDHP	6.770 (0.114)	8.516 (0.266)	6.917 (0.068)	8.675 (0.084)	39.95 (2.325)	25.06 (0.856)	29.09 (1.108)
AAPDSHP	6.838 (0.122)	8.549 (0.260)	7.072 (0.126)	8.928 (0.128)	40.16 (23.94)	25.73 (1.002)	29.13 (1.055)

The worst results are achieved in all scenarios and for all measures with COM, as it discards information. In the cases where the percentage of missingness is high, as no complete instances are available, no results are returned. Let us analyze results for each

Table 2: Mean RSS/n (st. deviation in brackets) for each approach for the wine and waveform data with the original data.

	Wine				Waveform		
Method	MCAR(10%)	MCAR(30%)	MAR	MNAR	MCAR	MAR	MNAR
ORG	6.015 (0)	6.015 (0)	6.015 (0)	6.015 (0)	19.40 (0.553)	19.40 (0.553)	19.40 (0.553)
COM	6.536 (0.260)	11.129 (2.976)	7.227 (0.571)	7.671 (0.056)	-	-	-
IMP	**5.985** (0.016)	**6.011** (0.032)	6.894 (0.208)	**6.150** (0.032)	21.65 (1.299)	19.76 (0.590)	23.93 (1.126)
MAAMOHAN 1e-3	6.006 (0.015)	6.022 (0.028)	6.028 (0.021)	6.183 (0.031)	21.85 (0.706)	19.56 (0.553)	23.35 (0.958)
MAAMOHAN 1e-9	6.006 (0.015)	6.015 (0.028)	**6.028** (0.021)	6.178 (0.027)	**19.78** (0.618)	**19.56** (0.553)	23.28 (0.904)
MAAEIS	6.101 (0.039)	6.288 (0.099)	6.131 (0.108)	6.534 (0.083)	22.14 (0.975)	20.21 (0.636)	36.39 (6.475)
AAII	6.029 (0.017)	6.099 (0.062)	6.067 (0.040)	6.209 (0.045)	20.29 (0.637)	19.80 (0.586)	23.36 (1.026)
AAK	6.526 (0.052)	7.045 (0.419)	6.518 (0.050)	7.071 (0.075)	36.07 (8.734)	24.74 (2.093)	52.38 (5.957)
AAEDcMDS	6.094 (0.021)	6.206 (0.053)	6.103 (0.050)	6.854 (0.131)	21.01 (0.811)	19.99 (0.575)	**23.17** (1.138)
AAEDHP	6.157 (0.037)	6.261 (0.067)	6.157 (0.053)	6.764 (0.065)	21.32 (0.851)	20.48 (0.669)	23.66 (0.938)
AAPDSHP	6.235 (0.073)	6.350 (0.102)	6.400 (0.226)	7.402 (0.247)	21.57 (0.916)	21.35 (0.901)	23.51 (1.015)

Table 3: Mean Frobenius norm (st. deviation in brackets) of the difference between the archetypes for each approach and the original ones for the wine and waveform data.

	Wine				Waveform		
Method	MCAR(10%)	MCAR(30%)	MAR	MNAR	MCAR	MAR	MNAR
COM	4.540 (6.573)	39.00 (27.07)	11.81 (11.89)	34.55 (1.538)	-	-	-
IMP	0.330 (0.115)	1.049 (0.283)	3.827 (0.761)	**2.259** (0.360)	26.87 (17.69)	6.718 (1.832)	21.77 (6.329)
MAAMOHAN 1e-3	0.320 (0.124)	1.010 (0.364)	0.520 (0.200)	2.842 (0.322)	7.119 (2.171)	2.488 (0.780)	17.68 (12.508)
MAAMOHAN 1e-9	0.320 (0.129)	**1.005** (0.362)	**0.513** (0.197)	2.785 (0.271)	**6.965** (2.038)	**2.466** (0.760)	17.57 (13.589)
MAAEIS	0.556 (0.215)	1.689 (0.573)	0.772 (0.347)	4.230 (1.107)	15.43 (4.48)	4.888 (1.559)	53.27 (23.954)
AAII	**0.293** (0.118)	1.224 (0.603)	0.636 (0.369)	2.577 (0.423)	9.482 (2.780)	3.456 (1.245)	18.33 (4.642)
AAK	4.084 (0.411)	6.350 (2.034)	4.083 (0.404)	7.190 (0.650)	91.89 (38.45)	37.18 (13.76)	169.87 (30.23)
AAEDcMDS	0.469 (0.146)	1.027 (0.266)	0.731 (0.258)	5.975 (1.208)	9.427 (3.113)	3.511 (1.003)	**17.44** (6.327)
AAEDHP	0.793 (0.180)	1.295 (0.274)	0.933 (0.273)	4.327 (0.434)	11.04 (3.023)	6.471 (2.089)	19.40 (3.441)
AAPDSHP	1.168 (0.428)	1.886 (0.560)	2.480 (1.619)	11.915 (3.371)	12.52 (38.76)	10.765 (3.129)	18.81 (4.398)

Table 4: Mean Rand score (st. deviation in brackets) for each approach for the wine and waveform data.

	Wine				Waveform		
Method	MCAR(10%)	MCAR(30%)	MAR	MNAR	MCAR	MAR	MNAR
ORG	0.9392 (0)	0.9392 (0)	0.9392 (0)	0.9392 (0)	0.6674 (0.006)	0.6674 (0.006)	0.6674 (0.006)
k-POD	0.9332 (0.015)	0.893 (0.022)	0.911 (0.011)	0.821 (0.017)	0.6543 (0.023)	0.6676 (0.011)	**0.6670** (0.007)
COM	0.9006 (0.040)	0.710 (0.085)	0.846 (0.074)	0.711 (0.006)	-	-	-
IMP	0.9304 (0.014)	0.894 (0.021)	0.892 (0.016)	**0.858** (0.013)	0.6645 (0.010)	0.6673 (0.006)	0.6639 (0.011)
MAAMOHAN 1e-3	0.9281 (0.012)	0.889 (0.020)	0.908 (0.013)	0.856 (0.013)	0.6650 (0.008)	0.6671 (0.005)	0.6654 (0.009)
MAAMOHAN 1e-9	0.9282 (0.012)	0.889 (0.020)	0.909 (0.013)	0.856 (0.012)	0.6650 (0.008)	0.6671 (0.005)	0.6656 (0.010)
MAAEIS	0.9269 (0.017)	0.886 (0.021)	0.912 (0.015)	0.832 (0.014)	0.6653 (0.011)	0.6675 (0.006)	0.6530 (0.024)
AAII	0.9287 (0.012)	0.888 (0.020)	0.910 (0.015)	0.852 (0.014)	0.6645 (0.007)	0.6670 (0.006)	0.6653 (0.009)
AAK	**0.9333** (0.013)	0.866 (0.031)	0.913 (0.012)	0.841 (0.015)	0.6424 (0.035)	**0.6725** (0.011)	0.6162 (0.035)
AAEDcMDS	0.9325 (0.012)	0.894 (0.019)	**0.915** (0.014)	0.829 (0.010)	0.6650 (0.008)	0.6673 (0.006)	0.6666 (0.009)
AAEDHP	0.9309 (0.014)	**0.897** (0.019)	0.908 (0.012)	0.843 (0.012)	**0.6661** (0.009)	0.6685 (0.007)	0.6655 (0.008)
AAPDSHP	0.9180 (0.018)	0.888 (0.020)	0.885 (0.016)	0.812 (0.013)	0.6651 (0.009)	0.6673 (0.005)	0.6652 (0.008)

measure. For the measure in Table 1, the best strategy is MAAMOHAN, except in one scenario where IMP is the best, but MAAMOHAN is quite near. AAII, despite not being the most accurate in any of the scenarios, it is always near the best for all the scenarios.

However, IMP is not a good approach for any of the scenarios: IMP is not among the most accurate for MAR with the wine data set and MCAR with the waveform data set.

For the measure in Table 2, again the most accurate approaches are IMP and MAAMOHAN, with the exception of MNAR with the waveform data set, for which the best result is obtained by AAEDcMDS. As before, IMP is not among the most accurate for all the scenarios. However, MAAMOHAN, AAII and also AAEDcMDS consistently return good results in all the scenarios. Note that the gold standard reference is ORG, as it uses all the information, without missing values. The proposed procedures provide results that are close to ORG, despite working with missing values, with the exception of MNAR with the waveform data set, where the distance is higher. In two scenarios, IMP reports better results than ORG, since the percentage of missingness is not very high and more accurate results can be achieved, as we work with more data due to the multiple imputations.

The best results are obtained for different strategies in each scenario for the measure in Table 3. The methods that give the best results in any of the scenarios are (the number of scenarios where each method is the best appears in brackets): AAII (1), AAEDcMDS (1), MAAMOHAN (4) and IMP (1). With the waveform data set scenarios, where a high percentage of missingness is used, IMP does not provide good results. However, AAII and MAAMOHAN and AAEDcMDS (with the exception of MNAR with the wine data set) consistently give good results in all the scenarios.

Finally, as regards the results in Table 4, in the majority of the scenarios a method based on dissimilarities is best. In particular, the methods that provide the most accurate results are (the number of scenarios where each method is the best appears in brackets): AAK (2), AAEDHP (2), AAEDcMDS (1), IMP (1) and k-POD (1). With the exception of MNAR for the waveform data set, the best results are obtained using a method for AA with missing data instead of one intended for clustering, such as k-POD, since those data sets do not have differentiated clusters. With the exception of MNAR for the wine data set, the results with missing data are not so far from those obtained without missing data (ORG). In fact, for AAK with the MAR scenario and the waveform data set, the mean is higher.

MAAEIS and AAPDSHP are not the best method in any of the scenarios or measures.

But AAPDSHP is the best in the toy example of supplementary material and MAAEIS is the best method, with lowest RSS, for the real data set in Section 5.1.

5 Real data sets

5.1 Air quality

It is normal for temperature data (and climate data in general) to include missing observations, unreasonable readings, spurious zeroes, and so on (Kotsiantis et al., 2006). In this section, we are going to analyze a well-known data set that contains daily readings of ozone, solar radiation, average wind speed and maximum daily temperatures in New York from May 1, 1973 to September 30, 1973. Ozone is measured as mean ozone in parts per billion from 13:00 to 15:00 hours on Roosevelt Island. Solar radiation is measured in Langleys in the frequency band 4000 - 7700 Angstroms from 08:00 to 12:00 hours in Central Park. Average wind speed is measured in miles per hour at 07:00 and 10:00 hours at La Guardia Airport and the maximum daily temperature is measured in degrees Fahrenheit at La Guardia Airport. The data set is available from the R package **datasets** (R Development Core Team, 2018; Chambers et al., 1983).

However, this data set has 37 missing ozone observations (24.03%) and 7 missing solar radiation measurements (4.55%). Our aim is to find a set of archetypes (mixtures of real days) to reflect extreme patterns and to facilitate comprehension of the data set.

First of all, the variables are standardized, since their range and meaning are very different. We apply the procedures for different numbers of archetypes and represent the screeplots. As stated in Section 2.1, the elbow criterion suggests that $k = 3$ archetypes should be chosen in all cases. We omit the figures in the interests of brevity. Table 5 shows the RSS/n for each approach with $k = 3$. The lowest value is achieved by MAAEIS. We analyze those results in detail, although the three archetypes found with the different strategies have similar characteristics and similar comments could be done for the rest of strategies. The screeplot for MAAEIS is displayed in Fig 1.

Table 6 shows the percentiles of the three archetypes regarding the variables analyzed. The first archetype presents very high values for solar radiation and wind and medium-

Table 5: RSS/n for each approach for the air quality data. The best result is highlighted in bold.

Method	RSS/n	Method	RSS/n	Method	RSS/n	Method	RSS/n	Method	RSS/n
COM	1.0153	MAAMOHAN 1e-3	0.9852	MAAEIS	**0.9799**	AAK	1.8994	AAEDHP	1.1584
IMP	1.0112	MAAMOHAN 1e-9	0.9852	AAII	0.9955	AAEDcMDS	1.030	AAPDSHP	1.0082

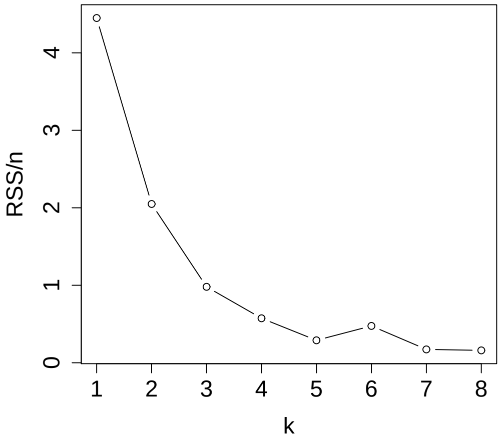


Figure 1: Screeplot of the RSS in descending order against the number of archetypes of MAAEIS for airquality data.

low values for ozone and temperature. The second archetype shows very low values for ozone, solar radiation and temperature and high values for wind, while the third archetype presents very high values for ozone and temperature and very low values for wind, with medium values for solar radiation. In other words, archetype 3 is the archetypal really hot day. However, archetypes 1 and 2 are very windy days, archetype 2 being colder than archetype 1, although what really contrasts between these two archetypes is the solar radiation, which is extremely high for archetype 1, but extremely low for archetype 2. For MAAEIS, archetype 1 is basically formed by the 22nd May. Archetype 2 is mainly a mixture of 9th and 21st May, while archetype 3 is basically a mixture of 29th, 30th August and 3rd September.

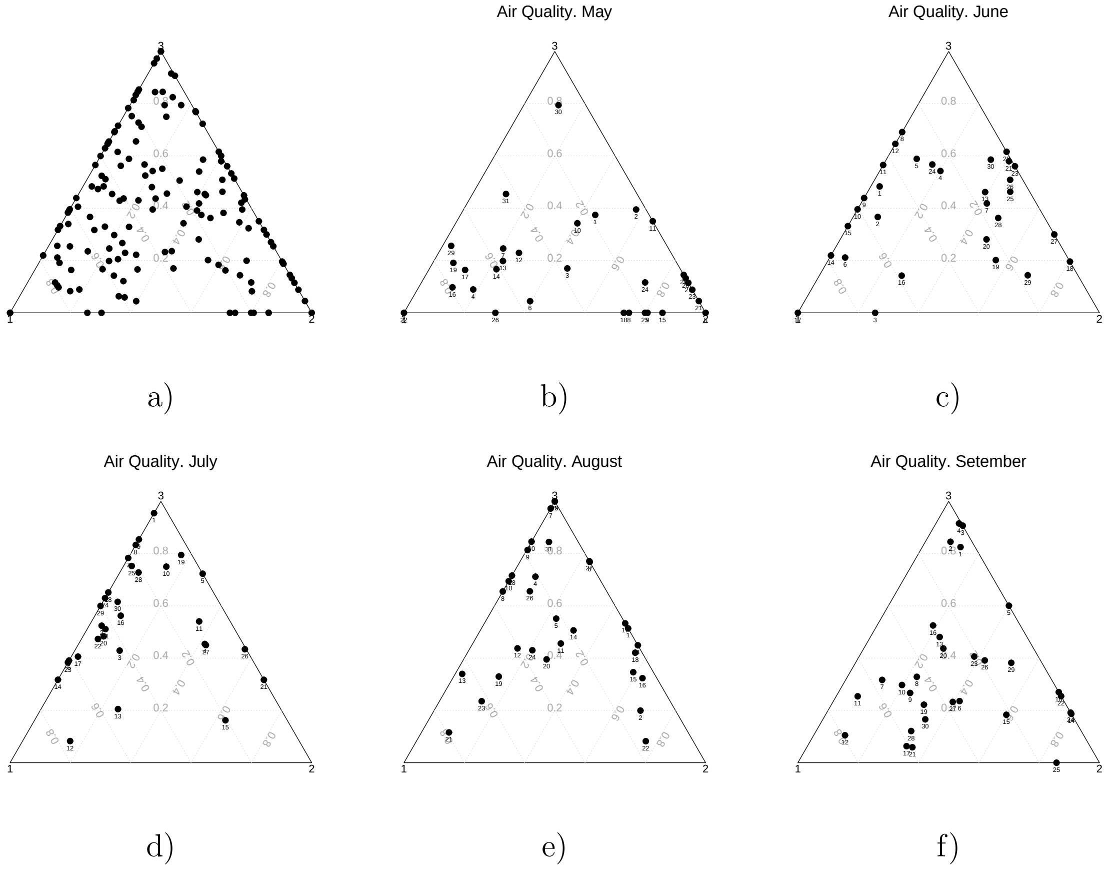
As an illustration, let us consider the three archetypes obtained with the MAAEIS

Table 6: Percentile profiles of the archetypes of MAAEIS for the air quality data.

Variables	A1	A2	A3
Ozone	17	1	94
Solar radiation	96	1	52
Wind	96	78	2
Temperature	25	5	99

strategy. Having obtained the 3 archetypes with this strategy, it is possible to obtain the $n \times 3$ matrix α with the coefficients that approximates each observation as mixture of the three archetypes. Fig. 2 shows ternary plots for these coefficients. In the first plot all the observations are jointly represented, and later, a ternary plot is shown with the observations for each month, so differences between months are easily observed. The points in each plot are labeled with the number of the day that it is being represented in each case. In these figures, the three archetypes represent the three vertices of the triangles, and each point represents a different observation (day) as a mixture of the three archetypes. The scales are included in the triangles in gray in segments parallel to each side. May days cluster close to archetypes 1 and 2, in particular, with low values for archetype 3 except 30th May. However, June days show medium values for archetype 3, i.e. they are hotter than May, but not as hot as in summer. July and August days display similar profiles, except some specific days; the majority of the days have high alpha values for archetype 3, i.e. they are hot days. In September, however, the alpha values for archetype 3 diminish; they are medium-low values, except for the first days of September. Many days in September are explained by approximately a mixture of 50% archetype 1, 25% archetype 2 and 25% archetype 3.

As carried out by Cutler and Breiman (1994), the alpha coefficients can also be used to see how the individual variables vary as functions of archetypes. For example, we regressed on terms of up to 3rd degree in α_{i1} and α_{i2} for each variable. The values of α_{i3} are not considered as $\alpha_{i1} + \alpha_{i2} + \alpha_{i3} = 1$. The R^2 are 0.89, 0.91, 0.82 and 0.79 for Ozone, Solar Radiation, Wind and Temperature, respectively. Therefore, the data can be surprisingly



Air Quality. May

Air Quality. June

a) b) c)

Air Quality. July

Air Quality. August

Air Quality. Setember

d) e) f)

Figure 2: Ternary plots with the mixture coefficients of MAAEIS for the air quality data.

well represented as a mixture of three archetypal days.

5.2 A snapshot of the world's countries

A solid understanding of the world is the first step for improving living conditions for all people worldwide. However, global development databases, with indicators for each country or region over the years, are not complete; in fact, missing values are quite common, especially in certain indicators.

In this section, we analyze six socio-demographic aspects of society, specifically the following indicators: Total fertility rate (TFR), Life expectancy at birth (LEB), Maternal mortality ratio (MMR), Infant mortality rate (IMR), Adult obesity rate (AOR) and Children under the age of 5 years underweight (CUW), for each country in the world in 2013. TFR and LEB from 1960 to 2013 of countries with nearly complete data over the years were analyzed in Epifanio (2016), who used them to find functional archetypoids to represent the big picture of global development. They were also previously selected by Rosling in the TED talk 'The best stats you've ever seen' (Rosling, 2006). Here, to aid an understanding the world today, we focus on the 2013 data as a snapshot of where we are now. The data are freely available at Shackman (2013) and come from the CIA World Factbook 2013 (Central Intelligence Agency, 2013).

TFR represents the number of children who would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with current age-specific fertility rates. LEB means the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life. MMR is the annual number of female deaths per 100000 live births from any cause related to or aggravated by pregnancy or its management (excluding accidental or incidental causes). This ratio includes deaths during pregnancy, childbirth, or within 42 days of termination of pregnancy, irrespective of the duration and site of the pregnancy, for a specified year. IMR measures the number of deaths of infants under one year old in a given year per 1,000 live births in the same year. This rate is often used as an indicator of the level of health in a country. AOR gives the percentage of a country's population considered to be obese. Obesity is defined as an adult having a Body Mass Index (BMI) greater than or equal to 30.0, and finally CUW gives the percentage of children under five years old considered to be underweight. This statistic is an indicator of the nutritional status of a community.

All the variables present missing observations, but only two countries, Montenegro and Tokelau, have missing data in the variable TFR. These two countries have been removed from the analysis, otherwise 'Nas' appear in the dissimilarity matrix. So finally, a total of 224 countries are considered and the percentages of missing data are of 18.75% for MMR; 0.45% for IMR; 0.89% for LEB; 0% for TFR, 15.63% for AOR and 41.52% for CUW.

The variables are standardized because they are measured in non-compatible units. Interpretation is easier if the representative are 'extreme countries' rather than 'combinations of countries'; therefore, ADA is considered in this case. We only present results for the EDcMDS strategy, which gives the lowest RSS for all ks and strategies based on projecting dissimilarities. In the next Section we will analyze the results when imputation and clustering methods are used.

Although archetypoids are not necessarily nested, in this case they are somehow nested, and when the number of archetypoids k, increases, new finer patterns appear. Therefore, let us see what happens when different numbers of archetypoids are considered. Table 7 shows the archetypoids obtained, together with their respective percentiles. In all cases, the archetypoids, are representative of extreme patterns.

For $k = 2$, the archetypoids are Malta and Nigeria. As can be seen in Table 7, Nigeria shows high percentiles for MMR, IMR, TFR and CUW, while Malta shows low percentiles for these indicators and high percentiles for LEB and AOR. In order to see how the patterns of other countries are expressed as mixtures of these archetypoids, coefficients α are estimated (see Table 8 for some examples) and represented in Fig. 3. The first map shows α_{i1}, the coefficients with respect to the first archetypoid (Malta), and the second map shows α_{i2}, the coefficients with respect to the second archetypoid (Nigeria). The darker the color on the map, the greater the value α_{ij} for country i and archetypoid j, according to the heat color palette (light yellow signals low values, whereas red indicates high values). So, darker colors on the map indicate that the indicators of each country are better explained by the corresponding archetypoid. Territories with no information are displayed in green.

Table 7: Percentile profiles of the archetypoids for the development indicator data.

k	Archetypoids	MMR	IMR	LEB	TFR	AOR	CUW
2	Malta	15	8	85	17	84	NA
	Nigeria	95	94	6	95	24	83
3	Greece	2	20	86	9	51	NA
	Nigeria	95	94	6	95	24	83
	American Samoa	NA	33	53	75	100	NA
4	Greece	2	20	86	9	51	NA
	Zambia	87	92	4	97	10	61
	Chad	100	98	0	90	8	94
	American Samoa	NA	33	53	75	100	NA
5	Japan	6	1	99	8	19	NA
	Bosnia and Herzegovina	15	23	62	3	77	8
	Burkina Faso	80	96	8	98	6	82
	Chad	100	98	0	90	8	94
	American Samoa	NA	33	53	75	100	NA
6	Japan	6	1	99	8	19	NA
	Czech Republic	6	9	72	5	90	12
	Zambia	87	92	4	97	10	61
	Chad	100	98	0	90	8	94
	American Samoa	NA	33	53	75	100	NA
	India	72	78	27	65	4	99

Table 8: Mixture coefficients α_{ij} with respect to the j-th archetypoid for some countries, depending on the number k of archetypoids considered.

k	Arch.	USA	Bolivia	Brazil	Angola	Centr.Afr.Rep.	Armenia	India	Afghan.
2	Malta	0.958	0.597	0.815	0	0	0.882	0.531	0
	Nigeria	0.042	0.403	0.185	1	1	0.118	0.469	1
4	Greece	0.724	0.532	0.798	0.028	0	0.705	0.554	0
	Zambia	0.026	0	0	0.542	0.028	0.088	0	0
	Chad	0	0.385	0.177	0.413	0.972	0	0.445	1
	Am Sam	0.249	0.082	0.025	0.045	0	0.207	0	0
6	Japan	0.445	0.122	0.025	0	0	0	0	0
	Czech Rep.	0.198	0	0.553	0	0	0.709	0	0
	Zambia	0.031	0	0.001	0.539	0	0	0	0
	Chad	0	0.159	0	0.413	0.949	0	0	1
	Am Sam	0.325	0.213	0	0.042	0.002	0	0	0
	India	0	0.505	0.421	0	0.048	0.290	1	0

As can be seen (Fig. 3), most ‘developed’ countries have high coefficients with respect to the archetypoid Malta, while the poorest African countries together with Afghanistan have high coefficients with respect with the archetypoid Nigeria, and other countries such as Bolivia or India have similar coefficients in both archetypoids (Table 8); they are explained fifty-fifty by each archetypoid. This kind of maps is usually referred to as ‘abundance maps’ in the hyperspectral imaging field.

For $k = 3$, the archetypoids are Greece, Nigeria and American Samoa. Table 7 shows that in terms of the variables analyzed, the profile of Greece is very similar to the profile of Malta, and the third archetypoid, American Samoa introduces a profile characterized by a very high percentage for obese population, a high percentile for TFR and a medium percentile for LEB. If we estimate the mixture coefficients α to formulate the other countries as mixtures of these archetypoids, the countries with highest coefficients for the third

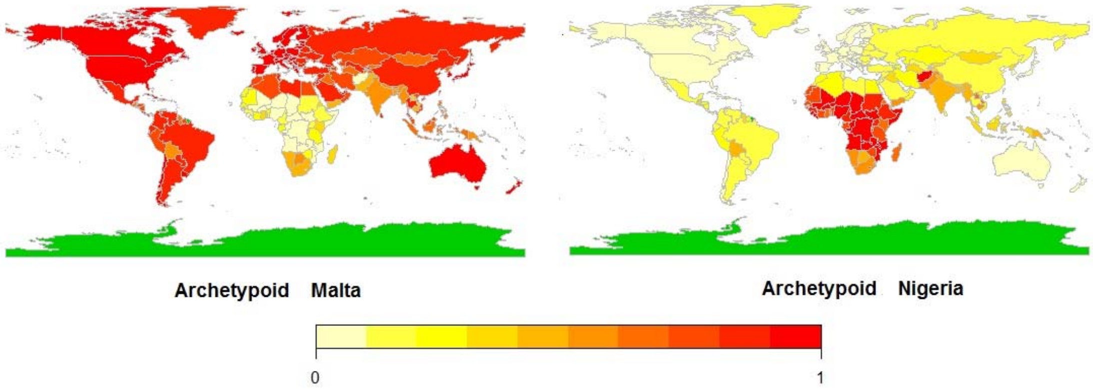


Figure 3: Abundance maps for the mixture coefficients for $k = 2$ archetypoids.

archetypoid are islands such as Nauru (0.9154), Tonga (0.7294), Cook Islands (0.7841), the Gaza Strip (0.8337) and the West Bank (0.7469). As they are all very small countries, they cannot be seen in Fig. 4.

For $k = 4$, the profile of the poorest countries is subdivided into two profiles, represented by Zambia and Chad, respectively. As can be seen in Table 8 and Fig. 4, countries like Angola or the Central African Republic are now better represented by one archetypoid or the other. Note that Chad's percentiles are very extreme, with very high or very low values. However, Zambia's percentiles are not as extreme as Chad's, except for TFR, the percentile for which is higher for Zambia.

For $k = 5$, the profile of the 'developed countries', represented by Greece, is now subdivided into two new profiles characterized by Japan and Bosnia and Herzegovina, respectively. The great difference between these two countries lies in LEB (whose percentile is 62 for Bosnia Herzegovina and 99 for Japan), but especially in AOR (whose percentile is 77 for Bosnia Herzegovina and 19 for Japan). Finally, for $k = 6$ a new profile appears, represented by India (see Table 7 and Fig. 5). It can be seen in Fig. 5 that Japan-like countries are developed countries in Asia, such as Singapore or South Korea, Western European countries and Australia and New Zealand, while Czech Republic-like countries are mainly Eastern European countries. As mentioned before, Zambia-like countries are mainly in Africa, as are Chad-like countries, although Afghanistan is also well explained by Chad. American Samoa-like countries are mainly small islands that cannot be seen

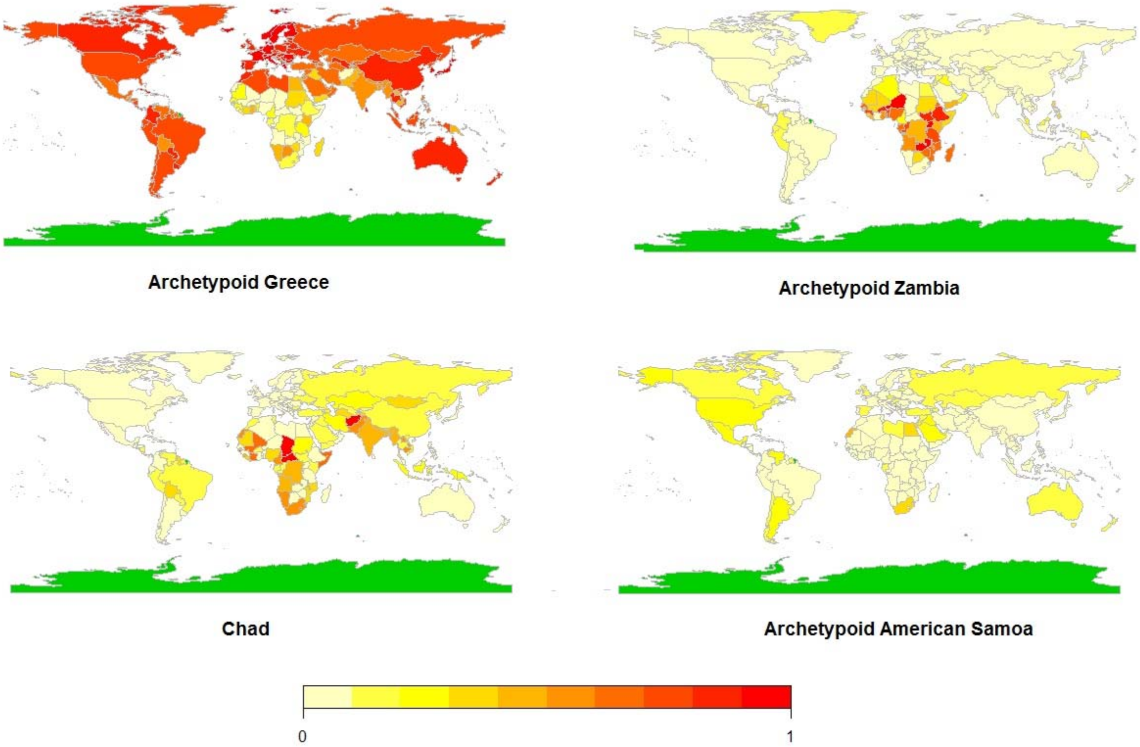


Figure 4: Abundance maps for the mixture coefficients for $k = 4$ archetypoids.

on the maps. Finally, India-like countries correspond to many Southeast Asian countries. Moreover, the maps reveal the composition of other countries. For example, the USA is approximately formed by a mixture between Japan (45%), American Samoa (33%) and the Czech Rep. (20%), while Brazil is a mixture between the Czech Rep. (55%) and India (42%) (see also Table 8).

Our intuition tells us that the variables vary continuously across countries, i.e. we do not expect there to be clearly differentiated (separate) groups of countries. This is corroborated by Figure 6, which displays dissimilarities between countries. Although the objective of AA is not data clustering, we can assign each country to the group in which its corresponding alpha is the maximum. With $k = 6$, the number of countries for each archetypoid are: 62 (Japan), 72 (Czech Republic), 30 (Zambia), 18 (Chad), 17 (American Samoa) and 25 (India).

5.2.1 Results for other alternatives

First, we can remove the cases with missing entries, which reduces the sample size from 224 to 128, nearly half of the countries. The archetypoids with $k = 5$ are: China, Uganda, Kuwait, India and Chad. Note that none of those countries have very low percentiles in MMR, IMR, TFR and a very high percentile in LEB, i.e. the profiles corresponding to Japan and Czech Rep., which were the most numerous groups. So, when working with complete data we cannot recover the most common extreme profiles. Second, we can impute missing values and apply ADA. The archetypoids obtained with $k = 5$ are: British Virgin Islands, Gabon, Macau, American Samoa and Somalia, whose percentiles are not as extreme as those of the archetypoids obtained using cMDS. Furthermore, the RSS with imputation is 0.415, which is higher than the RSS by cMDS (0.338).

Third, we apply k-POD with $k = 5$. To find a country that is representative of each cluster, we consider the nearest country to the center of each cluster. The representative cluster centers are therefore: Curaçao, Nigeria, Samoa, Puerto Rico and Haiti. Note that many of those countries belong to the same zone, the Caribbean sea, which makes it difficult for humans to understand the differences in profiles between clusters. Their profiles are not as differentiated as those of archetypoids. This also happen if we apply PAM with

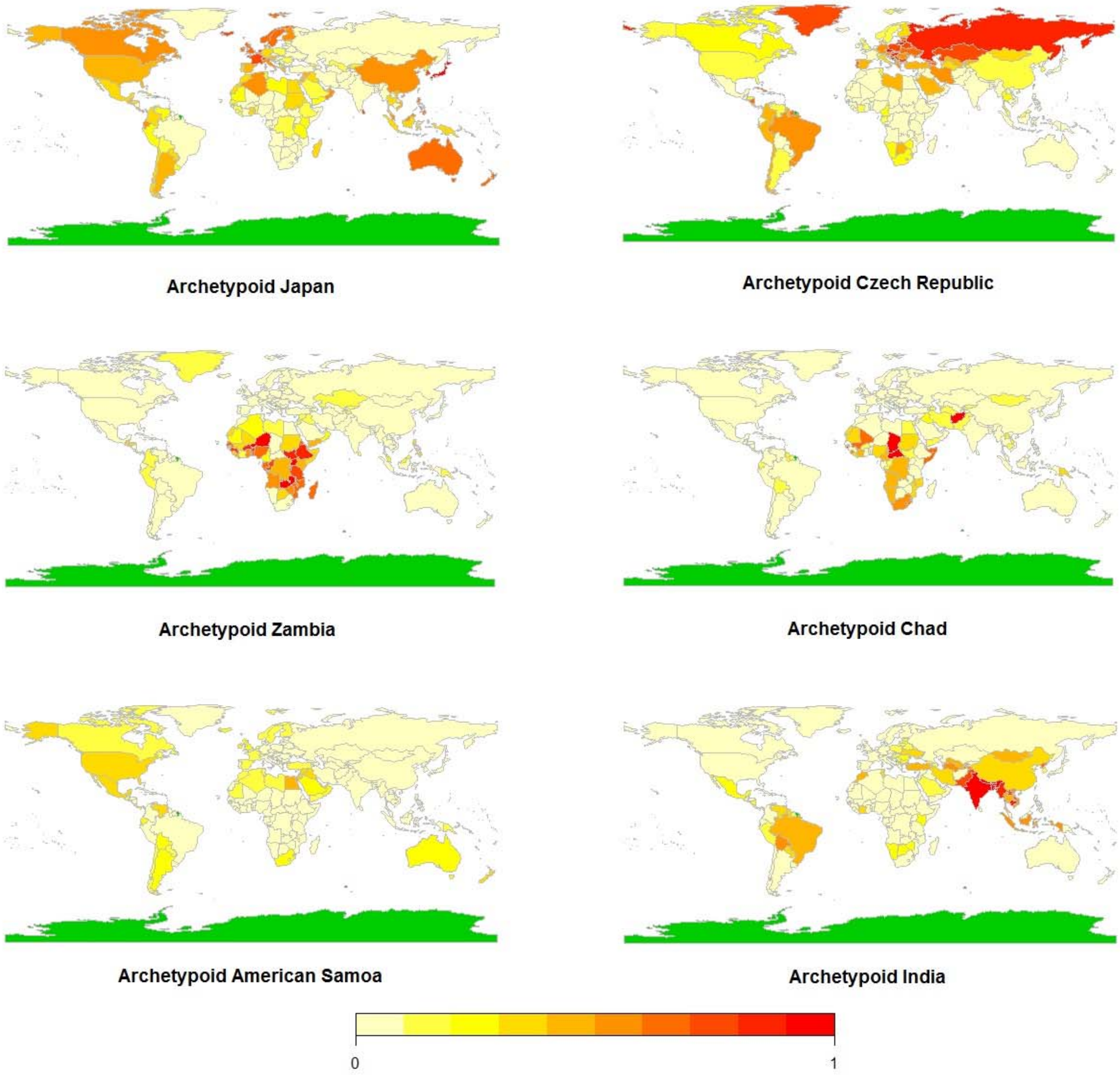


Figure 5: Abundance maps for the mixture coefficients for $k = 6$ archetypoids.

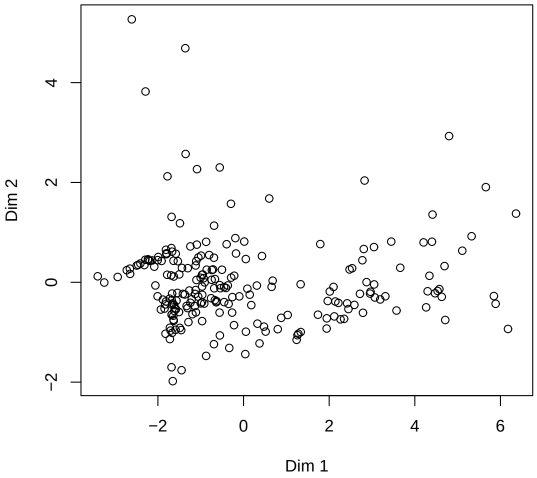


Figure 6: Scatter plot for countries: cMDS.

k =5 with the estimated dissimilarities; the medoids are: South Sudan, Puerto Rico, West Bank, Curaçao and Western Sahara, and the silhouette coefficient is 0.16, which indicates that there is no clear cluster structure (Kaufman and Rousseeuw, 1990), as seen in Figure 6. Furthermore, note that clustering results return assignations to each cluster without any explanation about the way they belong to each cluster, unlike the information returned by alpha values of ADA. For example, k-POD includes countries such as USA, Brazil (and many other South and Central America countries), Czech Republic (and other Eastern European countries), Cape Verde, Gaza Strip, Iran, Iraq, etc. in the same cluster, when their variable values are not so similar, as can be seen in the alpha coefficients shown in Table 8. The information returned by ADA is richer, since we can know the composition of data. That information is not returned either if we apply fuzzy clustering to the dissimilarities, with the *fanny* function from the R package **cluster** (Maechler et al., 2017). The memberships, which range from 0 to 1, are spread out between the groups, they are not sparse like the alpha values. In fact, none of the fuzzy memberships is zero and the maximum value is 0.49. This happens because, as mentioned above, there is no no clear cluster structure. The groups returned by ADA are therefore more reasonable and their interpretation easier since ADA identifies a sparser representation of each country in terms of the archetypoids.

6 Conclusion

New procedures for handling missing data in AA or ADA have been proposed and compared in a simulation study with previous approaches, offering favorable results. With the proposed procedures, the theoretical properties regarding location of archetypes are guaranteed, unlike the previous approaches. Furthermore, our procedures are the only ones to date that can return archetypal representatives with NAs, which are a natural part of a data set with missing values. The information gathered in cases with missing values is not discarded, and possible errors due to imputation do not occur either. Moreover, no assumption on the missingness mechanism is needed. The procedures do not substantially increase the computational cost of the original algorithms either, since the proposed modifications are not computationally expensive.

Based on the simulation results, it seems that depending on the data set, one method could work better than another, but there are no significant differences between the best methods. Depending also on whether the interest is in fitting (low RSS) or obtaining accurate archetypes or clustering the data, one method could be more appropriate than another: MAAMOHAN seems a good alternative for the first cases, whereas a method based on dissimilarities seems the best option for the last case. IMP has return good results in many settings, but it failed with wine data set for MAR. Furthermore, IMP was not good for recovering the archetypes in waveform data set, which had a high percentage of missigness. Our contribution is to provide a range of methods to choose from depending on the problem.

For the methodology based on dissimilarities, the cornerstone of the procedure is the estimation of the dissimilarities between samples with missing values and their subsequent projection in a new space. Simple methods have been used for the estimation of the dissimilarities, but even so, promising results have been obtained. As future work, more sophisticated estimators, such as those proposed in Eirola et al. (2014) or Mesquita et al. (2017), could be used. Another open question is the computation of standard errors of the proposed estimates, which could be based on a resampling technique (bootstrapping).

In terms of fields of the application, we have focused on socio-demographic aspects of society, but many other aspects could be analyzed, such as politics, economics, technology, etc. In fact, practically any application is possible, since missing data occur in almost all areas of research.

Another possibility for future work would be to extend AA and ADA to data sets with categorical or mixed data with missing values. Note that the "don't know" and "no opinion" answers are not uncommon in surveys. Furthermore, the perspective of using dissimilarities and their projection can be analyzed with other multivariate techniques.

Acknowledgments

The authors are grateful to the Editor, Associate Editor and two reviewers for their very constructive suggestions, which have led to improvements in the manuscript.

SUPPLEMENTARY MATERIAL

Toy example: A toy example illustrates our proposed methodology and the flaws of the previous approaches. (.pdf file)

Data and code: Data sets and code in R for reproducing the results. (.rar file)

References

Akande, O., F. Li, and J. Reiter (2017). An empirical comparison of multiple imputation methods for categorical data. *The American Statistician 71*(2), 162–170.

Breiman, L., J. Friedman, C. Stone, and R. Olshen (1984). *Classification and Regression Trees.* The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, Monterey.

Canhasi, E. and I. Kononenko (2013). Multi-document summarization via archetypal analysis of the content-graph joint model. *Knowledge and Information Systems*, 1–22.

Canhasi, E. and I. Kononenko (2014). Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. *Expert Systems with Applications 41*(2), 535 – 543.

Central Intelligence Agency (2013). The World Factbook CIA. `http://www.cia.gov/library/publications/the-world-factbook/`.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey (1983). *Graphical Methods for Data Analysis*. Wadsworth.

Chan, B., D. Mitchell, and L. Cram (2003). Archetypal analysis of galaxy spectra. *Monthly Notices of the Royal Astronomical Society 338*.

Chang, F., W. Qiu, R. H. Zamar, R. Lazarus, and X. Wang (2010). clues: An R package for nonparametric clustering based on local shrinking. *Journal of Statistical Software 33*(4), 1–16.

Chi, J. T. and E. C. Chi (2014). kpodclustr: An R package for clustering partially observed data. version 1.0.

Chi, J. T., E. C. Chi, and R. G. Baraniuk (2016). k-POD: A method for k-means clustering of missing data. *The American Statistician 70*(1), 91–99.

Costantini, P., M. Linting, and G. C. Porzio (2010). Mining performance data through nonlinear PCA with optimal scaling. *Applied Stochastic Models in Business and Industry 26*(1), 85–101.

Cutler, A. and L. Breiman (1994). Archetypal Analysis. *Technometrics 36*(4), 338–347.

Daszykowski, M., B. Walczak, and D. Massart (2003). Projection methods in chemistry. *Chemometrics and Intelligent Laboratory Systems 65*(1), 97 – 112.

Davis, T. and B. Love (2010). Memory for category information is idealized through contrast with competing options. *Psychological Science 21*(2), 234–242.

D'Esposito, M. R., F. Palumbo, and G. Ragozini (2012). Interval Archetypes: A New Tool for Interval Data Analysis. *Statistical Analysis and Data Mining 5*(4), 322–335.

Dheeru, D. and E. Karra Taniskidou (2017). UCI machine learning repository.

Dixon, J. K. (1979). Pattern recognition with partly missing data. *IEEE Transactions on Systems, Man, and Cybernetics 9*(10), 617–621.

Dray, S. and A. Dufour (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software 22*(4), 1–20.

Dray, S. and J. Josse (2015). Principal component analysis with missing values: a comparative survey of methods. *Plant Ecology 216*(5), 657–667.

Eirola, E., G. Doquire, M. Verleysen, and A. Lendasse (2013). Distance estimation in numerical data sets with missing values. *Information Sciences 240*, 115 – 128.

Eirola, E., A. Lendasse, V. Vandewalle, and C. Biernacki (2014). Mixture of Gaussians for distance estimation with missing data. *Neurocomputing 131*, 32 – 42.

Epifanio, I. (2013). h-plots for displaying nonmetric dissimilarity matrices. *Statistical Analysis and Data Mining 6*(2), 136–143.

Epifanio, I. (2016). Functional archetype and archetypoid analysis. *Computational Statistics & Data Analysis 104*, 24 – 34.

Epifanio, I., M. V. Ibáñez, and A. Simó (2018). Archetypal shapes based on landmarks and extension to handle missing data. *Advances in Data Analysis and Classification 12*(3), 705–735.

Epifanio, I., G. Vinué, and S. Alemany (2013). Archetypal analysis: contributions for estimating boundary cases in multivariate accommodation problem. *Computers & Industrial Engineering 64*(3), 757–765.

Eugster, M. J. and F. Leisch (2009). From Spider-Man to Hero - Archetypal Analysis in R. *Journal of Statistical Software 30*(8), 1–23.

Eugster, M. J. A. (2012). Performance profiles based on archetypal athletes. *International Journal of Performance Analysis in Sport 12*(1), 166–187.

Fernandez, M. and A. S. Barnard (2015). Identification of nanoparticle prototypes and archetypes. *ACS Nano 9*(12), 11980–11992.

Fogel, P., D. M. Hawkins, C. Beecher, G. Luta, and S. S. Young (2013). A tale of two matrix factorizations. *The American Statistician 67*(4), 207–218.

Giudici, P. and S. Figini (2009). *Applied Data Mining for Business and Industry (2nd Ed.).* John Wiley & Sons, Chichester.

Hand, D. J. (1998). Data mining: Statistics and more? *The American Statistician 52*(2), 112–118.

Harrell Jr, F. E., with contributions from Charles Dupont, and many others. (2016). *Hmisc: Harrell Miscellaneous.* R package version 4.0-2.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning. Data mining, inference and prediction.* 2nd ed., Springer-Verlag, New York.

Hinrich, J. L., S. E. Bardenfleth, R. E. Roge, N. W. Churchill, K. H. Madsen, and M. Mørup (2016). Archetypal analysis for modeling multisubject fMRI data. *IEEE Journal on Selected Topics in Signal Processing 10*(7), 1160–1171.

Hunt, L. and M. Jorgensen (2003). Mixture model clustering for mixed data with missing information. *Computational Statistics & Data Analysis 41*(3), 429 – 440.

Jones, M. C. and J. A. Rice (1992). Displaying the important features of large collections of similar curves. *The American Statistician 46*(2), 140–145.

Josse, J. and F. Husson (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Societe Française de Statistique 153*(2), 79–99.

Kaufman, L. and P. J. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley, New York.

Kiers, H. A. L. (1997). Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika 62*(2), 251–266.

Kotsiantis, S., A. Kostoulas, S. Lykoudis, A. Argiriou, and K. Menagias (2006). Filling missing temperature values in weather data banks. In *The 2nd IET International Conference on Intelligent Environments - IE 06*, Volume 1, pp. 327–334.

Larose, D. (2006). *Data Mining Methods and Models.* John Wiley & Sons, Hoboken.

Lawson, C. L. and R. J. Hanson (1974). *Solving Least Squares Problems.* Prentice Hall, Englewood Cliffs.

Leisch, F. and E. Dimitriadou (2010). *mlbench: Machine Learning Benchmark Problems.* R package version 2.1-1.

Li, S., P. Wang, J. Louviere, and R. Carson (2003). Archetypal Analysis: A New Way To Segment Markets Based On Extreme Individuals. In *ANZMAC 2003 Conference Proceedings*, pp. 1674–1679.

Little, R. and D. Rubin (2002). *Statistical analysis with missing data.* Wiley.

Lott, A. and J. P. Reiter (2018). Wilson confidence intervals for binomial proportions with multiple imputation for missing data. *The American Statistician 0*(0), 1–7.

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik (2017). *cluster: Cluster Analysis Basics and Extensions.* R package version 2.0.6 — For new features, see the 'Changelog' file (in the package source).

Maity, A. K., V. Pradhan, and U. Das (2018). Bias reduction in logistic regression with missing responses when the missing data mechanism is nonignorable. *The American Statistician 0*(0), 1–10.

Mardia, K., J. Kent, and J. Bibby (1979). *Multivariate Analysis.* Academic Press.

Mesquita, D. P., J. P. Gomes, A. H. S. Junior, and J. S. Nobre (2017). Euclidean distance estimation in incomplete datasets. *Neurocomputing 248*, 11 – 18.

Midgley, D. and S. Venaik (2013). Marketing strategy in MNC subsidiaries: pure versus hybrid archetypes. In *P. McDougall-Covin and T. Kiyak, Proceedings of the 55th Annual Meeting of the Academy of International Business*, pp. 215–216.

Millán-Roures, L., I. Epifanio, and V. Martínez (2018). Detection of anomalies in water networks by functional data analysis. *Mathematical Problems in Engineering 2018*(Article ID 5129735), 13.

Mørup, M. and L. K. Hansen (2012). Archetypal analysis for machine learning and data mining. *Neurocomputing 80*, 54–63.

Porzio, G. C., G. Ragozini, and D. Vistocco (2008). On the use of archetypes as benchmarks. *Applied Stochastic Models in Business and Industry 24*, 419–437.

R Development Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ragozini, G. and M. R. D'Esposito (2015). Archetypal networks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, New York, NY, USA, pp. 807–814. ACM.

Ragozini, G., F. Palumbo, and M. R. D'Esposito (2017). Archetypal analysis for data-driven prototype identification. *Statistical Analysis and Data Mining: The ASA Data Science Journal 10*(1), 6–20.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*(336), 846–850.

Rosling, H. (2006). The best stats you've ever seen. `https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen`. TED conference.

Seiler, C. and K. Wohlrabe (2013). Archetypal scientists. *Journal of Informetrics 7*(2), 345–356.

Seth, S. and M. J. A. Eugster (2016a). Archetypal analysis for nominal observations. *IEEE Trans. Pattern Anal. Mach. Intell. 38*(5), 849–861.

Seth, S. and M. J. A. Eugster (2016b). Probabilistic archetypal analysis. *Machine Learning 102*(1), 85–113.

Shackman, G. (2013). International data sets. `http://gsociology.icaap.org/dataupload.html`. Accessed: 2018-03-15.

Steinschneider, S. and U. Lall (2015). Daily precipitation and tropical moisture exports across the Eastern United States: An application of archetypal analysis to identify spatiotemporal structure. *Journal of Climate 28*(21), 8585–8602.

Stone, E. and A. Cutler (1996). Introduction to archetypal analysis of spatio-temporal dynamics. *Physica D: Nonlinear Phenomena 96*(1), 110 – 131.

Su, Z., Z. Hao, F. Yuan, X. Chen, and Q. Cao (2017). Spatiotemporal variability of extreme summer precipitation over the Yangtze river basin and the associations with climate patterns. *Water 9*(11).

Theodosiou, T., I. Kazanidis, S. Valsamidis, and S. Kontogiannis (2013). Courseware usage archetyping. In *Proceedings of the 17th Panhellenic Conference on Informatics*, PCI '13, New York, NY, USA, pp. 243–249. ACM.

Thøgersen, J. C., M. Mørup, S. Damkiær, S. Molin, and L. Jelsbak (2013). Archetypal analysis of diverse pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. *BMC Bioinformatics 14*, 279.

Thurau, C., K. Kersting, M. Wahabzada, and C. Bauckhage (2012). Descriptive matrix factorization for sustainability: Adopting the principle of opposites. *Data Mining and Knowledge Discovery 24*(2), 325–354.

Tsanousa, A., N. Laskaris, and L. Angelis (2015). A novel single-trial methodology for studying brain response variability based on archetypal analysis. *Expert Systems with Applications 42*(22), 8454 – 8462.

Vinué, G. (2017). Anthropometry: An R package for analysis of anthropometric data. *Journal of Statistical Software 77*(6), 1–39.

Vinué, G. and I. Epifanio (2017). Archetypoid analysis for sports analytics. *Data Mining and Knowledge Discovery 31*(6), 1643–1677.

Vinué, G., I. Epifanio, and S. Alemany (2015). Archetypoids: A new approach to define representative archetypal data. *Computational Statistics & Data Analysis 87*, 102 – 115.

Wang, J. and D. E. Johnson (2018). An examination of discrepancies in multiple imputation procedures between sas and spss. *The American Statistician 0*(0), 1–9.

Xia, Y. and Y. Yang (2016). Bias introduced by rounding in multiple imputation for ordered categorical variables. *The American Statistician 70*(4), 358–364.