HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

M. Halynchyk, F. Carvalho, A. Parinov, B. Mirkin

# Versions of least-squares k-means algorithm for interval data

Working Paper WP7/2024/01
Series WP7

Mathematical methods
for decision making in economics,
business and politics

Moscow
2024

УДК 519.2
ББК 22.172
F97



**Versions of least-squares k-means algorithm for interval data** [Electronic resource] : Working Paper WP7/2024/01 / M. Halynchyk, F. Carvalho, A. Parinov, B. Mirkin; National Research University Higher School of Economics. – Electronic text data (530 Kb). – Moscow: HSE Publishing House, 2024. – (Series WP7 "Mathematical methods for decision making in economics, business and politics"). – 34 p.

Recently, k-means clustering has been extended to the so-called interval data. In contrast to conventional data case, the interval data feature values are intervals rather than single reals. This paper further explores the least-squares criterion for k-means clustering to tackle the issue of initialization, that is, finding a proper set of initial cluster centers at interval data clustering. Specifically, we extend, for the interval data, a Pythagorean decomposition of the data scatter in the sum of two items, one being a genuine k-means least-squares criterion, the other, a complementary criterion, requiring the clusters to be numerous and anomalous. Therefore we propose a method for one-by-one obtaining anomalous clusters. After a run of the method, we start k-means iterations from the centers of the most numerous of the found anomalous clusters. We test and validate our proposed BIKM algorithm at versions of two newly introduced interval datasets.

Key words: Interval data, K-means, Least-Squares, Anomalous cluster, Feature weights

УДК 519.2
ББК 22.172

*Maksim Halynchyk*, Department of Data Analysis and Artificial Intelligence, HSE University, Moscow, Russian Federation, maksgalinchik@gmail.com

*Francisco de Carvalho*, Federal University of Pernambuco, UFPE Center of Informatics, Brazil, fatc@cin.ufpe.br

*Andrey Parinov*, Department of Data Analysis and Artificial Intelligence, HSE University, Moscow, Russian Federation, aparinov@hse.ru

*Boris Mirkin*, Department of Data Analysis and Artificial Intelligence, International Laboratory of Decision Choice and Analysis, HSE University, Moscow, RF; School of Computing and Mathematical Sciences, Birkbeck University of London, UK, bmirkin@hse.ru

# Contents

# 1 Introduction: Background and motivation

Data clustering is a major approach in data analysis and machine learning. It is oriented toward finding homogeneous groups of entities, aka clusters, conventionally represented as numerical multidimensional points. Clusters can be further used straightforwardly as they are in such applications as image or market segmentation, or they can be used as building blocks for further generalizations in such applications as annotation of text collections.

The most popular clustering method is k-means (see Table 1 in [18]), a procedure for alternating minimization of a least squares criterion. This criterion can be expressed by the following formula

$$L(S,c) = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v=1}^{V} (x_{iv} - c_{kv})^2. \quad (1)$$

Here $S = \{S_1, S_2, ..., S_K\}$ is a partition of the $N$-element entity set $I$, whose parts $S_k$ are clusters to be found. Each entity $i \in I$ is represented by a $V$-dimensional vector $x_i = (x_{i1}, x_{i2}, ..., x_{iV})$, whereas each cluster $S_k$ is represented by its $V$-dimensional center $c_k = (c_{k1}, c_{k2}, ..., c_{kV})$, also to be found. It is assumed that the data are represented by an entity-to-feature $N \times V$ data matrix $X$ whose rows correspond to entities $i = 1, 2, , , , N$ and columns to features $v = 1, 2, ..., V$. Entry $(i, v)$ in such a table is a real $x_{iv}$ representing the value of feature $v$ at the entity $i$.

The criterion in (1) is to be minimized so that the centers $c_k$ should be positioned in such a way that the clusters $S_k$ are formed by points $x_i$ "around" them.

Globally minimizing the criterion (1) is prohibitively difficult. The k-means method minimizes criterion (1) with respect to either $S$ or $c$, alternatingly. It starts with a randomly chosen set of $K$ entities considered as the initial centers, $c = (c_1, c_2, ..., c_K)$. Then it works in iterations consisting of two steps each. Step 1 (cluster update): each entity $i$ is assigned to its nearest center $c_k$, these form the current $S_k$. Step 2 (center update): new centers $cc_k$ are computed as the average vectors for $S_k$, so that $cc_k = \sum_{i \in S_k} x_i / N_k$, where $N_k$ is the number of elements in $S_k$. It is easy to prove that Step 1 finds an optimal $S_k$ at a given $c_k$. Similarly, Step 2 finds optimal $cc_k$ at the given $S_k$. After the iteration is finished, the new centers are compared with the previous ones. If some $cc_k$ and $c_k$ are not equal, up to the computation error, another iteration is run starting from $cc_k$ as $c_k$, $k = 1, 2, ..., K$. Otherwise, the process is finished and found $S_k$, $c_k$ and $L(S, c)$ are treated as the output of the algorithm.

This process is simple and intuitive. However, there are two issues: in a frequent situation where there is little knowledge of the phenomenon under study, how can one define: (a) the number of clusters $K$ and (b) the initial centers $c_k$? No good answer has come out so far, in spite of multiple attempts; those interested may wish to consult recent reviews in [7] and [14].

Meanwhile, there is an approach to both issues that is based on the properties of the least-squares criterion (1), unlike many others mentioned in [7, 14]. This approach involves the following "Pythagorean" decomposition of the data scatter $T = \sum_{i=1}^{N} \sum_{v=1}^{V} x_{iv}^2$ on the sum of criterion $L(S, c)$ in (1) and the complementary criterion $F(S, c)$,

$$T = L(S,c) + F(S,c), \quad (2)$$

where

$$F(S,c) = \sum_{k=1}^{K} N_k \langle c_k, c_k \rangle . \quad (3)$$

where $N_k$ is the number of entities in $S_k$, $k = 1, 2, ..., K$. This is proven in, say, [18, 26].

Since $T$ does not depend on $S$, the complementary criterion $F(S, c)$ in (3) is to be maximized to minimize $L(S, c)$, according to (2). Although the criterion $F(S, c)$ does depend on the location of the origin, unlike the original criterion $L(S, c)$, the change does not depend on $S$. The complementary criterion is the sum of individual cluster contributions, $N_k \langle c_k, c_k \rangle$. To maximize them, clusters should be as populous as possible and as far away from the origin as possible.

A greedy approach to this is to find clusters one-by-one, in sequence, by maximizing the cluster contribution. Originally coined in [17] as a "separate-and-conquer" procedure, a current version of this approach is described in [26] as "anomalous clustering".

This is the core of the clustering algorithm BANCO [26]: the algorithm selects the largest anomalous clusters and uses their centers to initialize k-means clustering iterations.

Although not necessarily successful at synthetic datasets with generated Gaussian clusters [4], this strategy leads to decent results in applications such as reconstruction of the history of genomes [19], analysis of socially responsible strategies of large companies [26], or modeling of the dynamics of off-coastal phenomenon upwelling [20].

The goal of this paper is to extend this approach to interval data clustering. Interval data are data with structural values, sometimes referred to as "symbolic data" [1]. Interval data emerge when units of observation, the entities, are not individual objects but rather categories of them. Such are data of fungi (mushroom) species (842 of them are listed in [9]) or research production categorized over dominating themes such as mathematics, mechanical engineering, or law [21]. An interval data value is an interval $(x_1, x_2)$, which is a set of all reals $x$ such that $x_1 \leq x \leq x_2$.

Clustering methods for interval data are being developed as extensions of those for conventional numeric data. Probably, the very first clustering method for interval data was described in [3]. Several others followed, such as [2] for probabilistic clustering, [22, 27] for fuzzy clustering, [24] for similarity clustering, [28] with a genetic algorithm, and [16] with overlap distance. Extensions of k-means clustering to the interval data were considered in [3, 23, 25]. We specifically rely on the method developed in [25] because it involves a cluster-specific feature-weighting scheme. Feature weighting in k-means clustering started by paper [12] and was further extended in [5, 8]. In these methods, feature weights $w_v$ were to satisfy the normalizing condition,

$$\sum_{v=1}^{V} w_v = 1. \tag{4}$$

According to [25], in interval clustering, any feature is to have two weights, $w_{v1}$ for the bottom ends of feature intervals and $w_{v2}$ for the tops. The normalizing condition is transformed here into equation

$$\prod_{v=1}^{V} w_{v1} w_{v2} = 1. \tag{5}$$

We are going to use both normalization conditions.

# 2 Least squares clustering for interval data

## 2.1 Least squares criteria

Consider an $N \times V$ data table $X = (x_{iv})$ where $N$ is the number of objects, or entities; $V$, the number of features, and $x_{iv}$ is the value of feature $v$ at entity $i$ represented by an interval $x_{iv} = (x_{iv1}, x_{iv2})$ where $x_{iv1}$ is the left, and $x_{iv2}$) the right end of the interval. Our goal is to partition the entity set in $K$ non-overlapping subsets $S_1, S_2, ..., S_K$, each represented by a center $c_k = (c_{kv})$ whose components are intervals as well, $c_{kv} = (c_{kv1}, c_{kv2})$, $k = 1, 2, ..., K$, $v = 1, 2, ..., V$. The goodness-of-fit criterion for clustering is

$$L(S, c, w) = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v=1}^{V} (w_{v1}^{\beta}(x_{iv1} - c_{kv1})^2 + w_{v2}^{\beta}(x_{iv2} - c_{kv2})^2) \quad (6)$$

where $S$s and $x$s and $c$s, with indices, are defined above, whereas $w_{kv1}$ and $w_{kv2}$ are feature weights, possibly cluster-specific. The exponent $\beta$ is a user-defined parameter to re-scale the effect of the weights on the distances which are parts of criterion (6):

$$D(x_{i1}, c_{v1}) = \sum_{v=1}^{V} w_{v1}^{\beta}(x_{iv1} - c_{kv1})^2, D(x_{i2}, c_{v2}) = \sum_{v=1}^{V} w_{v2}^{\beta}(x_{iv2} - c_{kv2})^2 \quad (7)$$

The weights are supposed to be positive and satisfy the normalization condition in (5).

This criterion, to be minimized with respect to $S_k$ and interval-valued $c_k, w_v$, $k = 1, 2, ..., K$. is, basically, the clustering criterion defined by equation (7) in [25].

It should be mentioned that criterion (6) extends the corresponding least-squares criterion for conventional data tables. A conventional data table, denoted here by $Y = (y_{iv})$, has real values $y_{iv}$ of features $v = 1, 2, ..., V$ at objects $i = 1M, 2, ..., N$ as its entries. A convenient form of the least squares criterion is

$$Ll(S, c, w) = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v=1}^{V} w_v^{\beta}(y_{iv} - c_{kv})^2 \quad (8)$$

introduced and explored by Huang et al. [12].

## 2.2 Optimal centers

Let us derive some properties of the optimal solutions for criterion (6) by using the first order optimality conditions. First of all, one can state that the optimal centers here are within-cluster gravity centers whose components are the within-cluster averages of the corresponding components in the data table. Indeed, the partial derivative of $L(S, c)$ with respect to $c_{kv1}$ is

$$\frac{\partial L}{\partial c_{kv1}} = \sum_{i \in S_k} 2w_{v1}^{\beta}(x_{iv1} - c_{kv1})(-1).$$

Making this equal to zero, one arrives at equation $\sum_{i \in S_k}(x_{iv1} - c_{kv1}) = 0$ by dropping off all the constant factors. This implies $\sum_{i \in S_k} x_{iv1} = N_k c_{kv1}$ where $N_k$ is the number of elements in cluster $S_k$. Therefore, $c_{kv1} = \sum_{i \in S_k} x_{iv1}/N_k$ which is the average value of all the lower boundaries of the intervals within cluster $S_k$. Similar equation for the upper boundaries of the centers, $c_{kv2}) = \sum_{i \in S_k} x_{iv2}/N_k$, can be derived analogously.

## 2.3 Optimal weights

To find optimal feature weights in criterion (6) under condition (5), let us consider the corresponding Lagrange function:

$$M = L(S, c, w) - \delta(\prod_{v=1}^{V} w_{v1} * w_{v2} - 1).$$

The partial derivative of $M$ with regard to $w_{v1}$ is equal to

$$\partial M w v1 = \sum_{k=1}^{K} \sum_{i \in S_k} \beta w_{v1}^{\beta-1} (x_{iv1} - c_{kv1})^2 - \frac{\delta}{w_{v1}} = 0.$$

The last equation holds according to the first-order necessary condition for optimality. This equation leads to the following solution:

$$w_{v1} = \left( \frac{\delta}{\beta \sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv1} - c_{kv1})^2} \right)^{\frac{1}{\beta}}.$$

A similar solution can be found for $w_{v2}$ analogously. To determine $\delta$, one may use equation (7). Indeed,

$$\prod_{j=1}^{V} \left( \frac{\delta}{\beta \sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv1} - c_{kv1})^2} \right)^{\frac{1}{\beta}} * \left( \frac{\delta}{\beta \sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv2} - c_{kv2})^2} \right)^{\frac{1}{\beta}} = 1.$$

This implies:

$$\delta = \beta \left\{ \prod_{v=1}^{V} \left[ \sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv1} - c_{kv1})^2 \right] \left[ \sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv2} - c_{kv2})^2 \right] \right\}^{\frac{1}{2V}}$$

By substituting this into formulas for the weights, we finally obtain:

$$w_{v1} = \left( \frac{\left\{ \prod_{v=1}^{V} \left[ \sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv1} - c_{kv1})^2 \right] \left[ \sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv2} - c_{kv2})^2 \right] \right\}^{\frac{1}{2V}}}{\sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv1} - c_{kv1})^2} \right)^{\frac{1}{\beta}} \tag{9}$$

$$w_{v2} = \left( \frac{\left\{ \prod_{v=1}^{V} \left[ \sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv1} - c_{kv1})^2 \right] \left[ \sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv2} - c_{kv2})^2 \right] \right\}^{\frac{1}{2V}}}{\sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv2} - c_{kv2})^2} \right)^{\frac{1}{\beta}} \tag{10}$$

These formulas imply that the weights $w^\beta$ in the criterion (6) do not depend on $\beta$ at all. This means that the criterion may be equivalently formulated by using just $w$ weights by themselves. The normalizing condition (5) is the cause.

### 2.4 Convenient weights

To add flexibility to our analysis, we propose using one more definition of feature weights inherited from the paper by Huang et al. [12] in which the weights have been subject to the conventional normalization condition (4), summing to unity, so that the coefficients in the criterion do depend on $\beta$. Let us consider those values, derived for the conventional, not interval-based, criterion in (8):

$$h_v = \left( \frac{p(S,c)}{\sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv1} - c_{kv1})^2} \right)^{\frac{1}{\beta-1}} \tag{11}$$

where

$$p(S,c) = \left\{ \prod_{v=1}^{V} \left[ \sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv} - c_{kv})^2 \right] \right\}^{\frac{1}{2V}}.$$

It is not difficult to check that these weights do satisfy the normalizing condition (4).

Applied to interval data. these formulas can be rewritten as follows:

$$h_{v1} = \left( \frac{P(S,c)}{\sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv1} - c_{kv1})^2} \right)^{\frac{1}{\beta-1}}, h_{v2} = \left( \frac{P(S,c)}{\sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv2} - c_{kv2})^2} \right)^{\frac{1}{\beta-1}}. \tag{12}$$

Here

$$P(S,c) = \left\{ \prod_{v=1}^{V} \left[ \sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv1} - c_{kv1})^2 \right] \left[ \sum_{k=1}^{K} \sum_{i \in S_k} (x_{iv2} - c_{kv2})^2 \right] \right\}^{\frac{1}{2V}}. \tag{13}$$

# 3 Algorithms

## 3.1 Algorithm IKM

Here is the algorithm IKM from [25] adapted to the case at which initial cluster centers are found with a different algorithm.

Input: Data matrix $X$, the number of clusters K, weight exponent $\beta$, initial cluster centers in interval format.

Output: Clusters $S_k$ and their interval centers $c_k$, $k = 1, 2, ..., K$.

1. Data pre-processing:
   (a) Compute the grand mean vectors for the left and right interval boundaries, respectively, $g_1$ and $g_2$.
   (b) Centering: Subtract interval vector $[g1, g2]$ from all the rows of the data matrix.
   (c) Normalization: Divide the feature left and right bound values over their standard deviations.
2. Initialization:
   (a) Weights: Put unity for all the weight values.

(b) Centers: At each $k$, $k = 1, 2, ..., K$ take a random row of $X$ as the initial center $c_k$.

(c) Clusters: Initialize $K$ empty clusters.

3. Loop until convergence:

(a) Cluster update:

i. For each object $i = 1, ..., N$ compute its distances to the cluster centers;

ii. Each object is assigned to its nearest center.

(b) Center update: current $c_k$ boundaries are computed as the within-cluster averages of the corresponding boundaries.

(c) Weight update:

- Global feature weights are computed according to formulas (9), (10) (optimal) or (12) (convenient).
- Cluster-specific weights are computed according to cluster-specific versions of formulas (9), (10):

$$v_{k,1,j} = \frac{\left\{\prod_{v=1}^{V}\left[\sum_{i\in S_k}(x_{i,1,v} - c_{k,1,v})^2\right]\left[\sum_{i\in S_k}(x_{i,2,v} - c_{k,2,v})^2\right]\right\}^{\frac{1}{2V}}}{\sum_{i\in S_k}(x_{i,1,j} - c_{k,1,j})^2}$$

$$v_{k,2,j} = \frac{\left\{\prod_{v=1}^{V}\left[\sum_{i\in S_k}(x_{i,1,v} - c_{k,1,v})^2\right]\left[\sum_{i\in S_k}(x_{i,2,v} - c_{k,2,v})^2\right]\right\}^{\frac{1}{2V}}}{\sum_{i\in S_k}(x_{i,2,j} - x_{k,2,j})^2},$$

as the optimal ones, or (12) as those convenient.

## 3.2 KM Algorithm

Here we consider a conventional way to the analysis of interval data. According to this approach, every feature $v$ is substituted by its double versions $v1$, corresponding to the left boundary of the interval value, and $v2$, corresponding to its right boundary. We utilize a version of k-means, k-means++, implemented in the library Scikit-Learn [15]. This version differs from the conventional random start k-means by its initialization. According to this approach, the first center is a randomly chosen entity. The general step: having a subset of centers $c$ already selected, define the distance to $c$, for every entity outside of $c$, as the minimum distance to the entities in $c$. Assign to each of the entities a probability proportional to its distance to $c$. Choose the next center randomly according to the specified probabilities. This version usually finds deeper minima of the least-squares criterion than the random start initialized versions. We also apply a feature-weighted version of k-means with feature weights updated according to formulas in (11). Both versions have the number of clusters as the input and cluster partition $S$ and cluster centers as output.

## 3.3 Anomalous clustering algorithm

### 3.3.1 One cluster modeling

We follow here the version described in [26] and extend it to the interval data case.

Consider a reduced clustering problem at which only one cluster $S \subset I$ is sought, along with its center. Then the criteria in (8) and (6) should be reformulated, in respect, as follows:

$$ll(S, c, w) = \sum_{i \in S} \sum_{v=1}^{V} w_v^\beta (y_{iv} - c_v)^2 \tag{14}$$

where $c = (c_v)$ is the cluster's center and $w_v$ are feature weights, for the ordinary data case, and

$$l(S, c, w) = \sum_{i \in S} \sum_{v=1}^{V} (w_{v1}^\beta (x_{iv1} - c_{v1})^2 + w_{v2}^\beta (x_{iv2} - c_{v2})^2), \tag{15}$$

for the interval data case. Here indices 1 and 2 correspond to the left and right interval boundaries, respectively. The weights are supposed to be positive and satisfy the normalization condition (5).

As mentioned in the introduction, both models yield a complementary criterion via the corresponding Pythagorean decomposition (2).

### 3.3.2 Pythagorean decomposition and complementary criterion for interval data

Indeed, let us make elementary transformations of the least-squares criterion:

$$L(S, c, w) = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v \in V} \left[ w_{kv1}^\beta (x_{iv1} - c_{kv1})^2 + w_{kv2}^\beta (x_{iv2} - c_{kv2})^2 \right] = \sum_{i=1}^{N} \sum_{v \in V} \left[ w_{kv1}^\beta x_{iv1}^2 + w_{kv2}^\beta x_{iv2}^2 \right] +$$

$$\sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v \in V} \left[ w_{kv1}^\beta c_{kv1}^2 + w_{kv2}^\beta c_{kv2}^2 \right] - 2 \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v \in V} \left[ w_{kv1}^\beta x_{iv1} c_{kv1} + w_{kv2}^\beta x_{iv2} c_{kv2} \right]$$

Equations

$$\sum_{i \in S_k} x_{iv1} = N_k \overline{x}_{kv1}, \sum_{i \in S_k} x_{iv2} = N_k \overline{x}_{kv2}$$

hold because of the first-order necessary conditions:

$$\frac{\partial L}{\partial c_{kv1}} = \sum_{i \in S_k} 2 w_{kv1}^\beta (x_{iv1} - c_{kv1})(-1) = 0 => c_{kv1} = \overline{x}_{kv1}$$

$$\frac{\partial L}{\partial c_{kv2}} = \sum_{i \in S_k} 2 w_{kv2}^\beta (x_{iv2} - c_{kv2})(-1) = 0 => c_{kv2} = \overline{x}_{kv2}$$

Therefore,

$$L = T(X) - \sum_k N_k \sum_v w_{kv1}^\beta \overline{x}_{kv1} - \sum_k N_k \sum_v w_{kv2}^\beta \overline{x}_{kv2},$$

where

$$T(X) = \sum_{i=1}^{N} \sum_{v \in V} \left[ w_{kv1}^\beta x_{iv1}^2 + w_{kv2}^\beta x_{iv2}^2 . \right]$$

This is the interval data scatter.

Since $T(X)$ does not depend on partition $S$, then one may find clusters by maximizing

$$F(S,c) = \sum_k N_k \sum_v \left( w_{kv1}^{\beta} \overline{x}_{kv1} + w_{kv2}^{\beta} \overline{x}_{kv2} \right)$$

rather than by minimizing $L$.

By applying a part of this criterion to a single cluster $S$ rather than a partition $S$, we obtain a single cluster criterion: maximize

$$F(S,c) = |S| \sum_v \left( w_{v1}^{\beta} \overline{x}_{v1} + w_{v2}^{\beta} \overline{x}_{v2} \right) \tag{16}$$

labelnf1

Here $S$ is a single cluster, $S \subseteq I$; $|S|$, the number of elements in $S$; and $c = \sum_{i \in S} x_i / |S|$, its gravity center.

Preliminarily, matrix $X$ is standardized by subtraction of the grand mean $g$ from each its row, so that the origin moves into $g$. Conventionally, each feature is re-scaled by dividing over its standard deviation, the square root of the variance. The cluster $S$ is initialized by putting there a single object, that one furthest away from 0. This very object serves as the cluster center as well.

Given cluster $S$ and its center $c$, the following rules apply to update the cluster and weights.

### 3.3.3 Cluster update rule (CUR)

This follows the alternating optimization path: given a cluster center $c = (c_v)$ where $c_v$ is the interval $c_v = (c_{v1}, c_{v2})$, the optimal rule CUR requires:

- Remove $i \in S$ from $S$ if:

$$f(S,c) > 2|S| \langle x_{i1}, c_1 \rangle_{w_1^{\beta}} + 2|S| \langle x_{i2}, c_2 \rangle_{w_2^{\beta}} - \langle x_{i1}, x_{i1} \rangle_{w_1^{\beta}} - \langle x_{i2}, x_{i2} \rangle_{w_2^{\beta}}$$

- Add $i \notin S$ to $S$ if:

$$f(S,c) < 2|S| \langle x_{i1}, c_1 \rangle_{w_1^{\beta}} + 2|S| \langle x_{i2}, c_2 \rangle_{w_2^{\beta}} + \langle x_{i1}, x_{i1} \rangle_{w_1^{\beta}} + \langle x_{i2}, x_{i2} \rangle_{w_2^{\beta}}$$

Here:

$$f(S,c) = |S| \langle c_1, c_1 \rangle_{w_1^{\beta}} + |S| \langle c_2, c_2 \rangle_{w_2^{\beta}}$$

and the inner product is weighted so that, for example,

$$\langle x_{i1}, c_1 \rangle_{w_1^{\beta}} = \sum_v w_{v1} x_{iv1} c_{v1}$$

### 3.3.4 Proof of the Interval Cluster Update Rule CUR

$$f(S-i,c') = (|S|-1)\left\langle \frac{\sum_{j\in S} y_{j1}-y_{i1}}{|S|-1}, \frac{\sum_{j\in S} y_{j1}-y_{i1}}{|S|-1}\right\rangle_{w_1^\beta} +$$
$$(|S|-1)\left\langle \frac{\sum_{j\in S} y_{j2}-y_{i2}}{|S|-1}, \frac{\sum_{j\in S} y_{j2}-y_{i2}}{|S|-1}\right\rangle_{w_2^\beta} =$$
$$\frac{1}{|S|-1}\left\langle \sum_{j\in S} y_{j1}-y_{i1}, \sum_{j\in S} y_{j1}-y_{i1}\right\rangle_{w_1^\beta} + \frac{1}{|S|-1}\left\langle \sum_{j\in S} y_{j2}-y_{i2}, \sum_{j\in S} y_{j2}-y_{i2}\right\rangle_{w_2^\beta} =$$
$$\frac{1}{|S|-1}\left(\left\langle \sum_{j\in S} y_{j1}, \sum_{j\in S} y_{j1}\right\rangle_{w_1^\beta} - 2\left\langle y_{i1}, \sum_{j\in S} y_{j1}\right\rangle_{w_1^\beta} + \langle y_{i1}, y_{i1}\rangle_{w_1^\beta}\right) +$$
$$\frac{1}{|S|-1}\left(\left\langle \sum_{j\in S} y_{j2}, \sum_{j\in S} y_{j2}\right\rangle_{w_2^\beta} - 2\left\langle y_{i2}, \sum_{j\in S} y_{j2}\right\rangle_{w_2^\beta} + \langle y_{i2}, y_{i2}\rangle_{w_2^\beta}\right) \quad (17)$$

Similarly,

$$f(S,c) = \frac{1}{|S|}\left(\left\langle \sum_{j\in S} y_{j1}, \sum_{j\in S} y_{j1}\right\rangle_{w_1^\beta} + \left\langle \sum_{j\in S} y_{j2}, \sum_{j\in S} y_{j2}\right\rangle_{w_2^\beta}\right) =$$
$$\frac{1}{|S|-1}\left(1-\frac{1}{|S|}\right)\left(\left\langle \sum_{j\in S} y_{j1}, \sum_{j\in S} y_{j1}\right\rangle_{w_1^\beta} + \left\langle \sum_{j\in S} y_{j2}, \sum_{j\in S} y_{j2}\right\rangle_{w_2^\beta}\right) =$$
$$\frac{1}{|S|-1}\left(\left\langle \sum_{j\in S} y_{j1}, \sum_{j\in S} y_{j1}\right\rangle_{w_1^\beta} + \left\langle \sum_{j\in S} y_{j2}, \sum_{j\in S} y_{j2}\right\rangle_{w_2^\beta} - f(S,c)\right) \quad (18)$$

By subtracting the last expression from the first formula, one obtains:

$$f(S-i,c') - f(S,c) = \frac{1}{|S|-1}$$
$$\left(f(S,c) - 2\left\langle y_{i1}, \sum_{j\in S} y_{j1}\right\rangle_{w_1^\beta} + \langle y_{i1}, y_{i1}\rangle_{w_1^\beta} - 2\left\langle y_{i2}, \sum_{j\in S} y_{j2}\right\rangle_{w_2^\beta} + \langle y_{i2}, y_{i2}\rangle_{w_2^\beta}\right) \quad (19)$$

This is positive if and only if:

$$f(S,c) - 2\left\langle y_{i1}, \sum_{j\in S} y_{j1}\right\rangle_{w_1^\beta} + \langle y_{i1}, y_{i1}\rangle_{w_1^\beta} - 2\left\langle y_{i2}, \sum_{j\in S} y_{j2}\right\rangle_{w_2^\beta} + \langle y_{i2}, y_{i2}\rangle_{w_2^\beta} > 0$$

that is, if

$$f(S,c) > 2|S|\,\langle y_{i1}, c_1\rangle_{w_1^\beta} - \langle y_{i1}, y_{i1}\rangle_{w_1^\beta} + 2|S|\,\langle y_{i2}, c_2\rangle_{w_2^\beta} - \langle y_{i2}, y_{i2}\rangle_{w_2^\beta}$$

This proves one part of CUR rule. The other part, for $i \notin S$, is proved analogously by considering the difference $f(S+i,c') - f(S,c)$ where c' is the center of S + i

### 3.3.5 Weighting update rule (WUR)

Feature weights initialise as equal to each other. Then, given a cluster $S$ with its (interval) center $c$. new left boundary weights are computed as:

$$w_{v1} = \frac{1}{\sum_{u \in V} [D_{v1}/D_{u1}]^{\frac{1}{\beta-1}}}$$

where

$$D_{v1} = \sum_{i \in S} (x_{iv1} - c_{v1})^2,$$

the within-cluster dispersion of the interval left boundaries. A similar formula holds for the right boundaries:

$$w_{v2} = \frac{1}{\sum_{u \in V} [D_{v2}/D_{u2}]^{\frac{1}{\beta-1}}}$$

with

$$D_{v2} = \sum_{i \in S} (x_{iv2} - c_{v2})^2.$$

To avoid division by zero, in computations each $D_u$ is added by the average variance of feature $u$.

These weights satisfy the normalizing conditions:

$$\sum_{v \in V} w_{k,1,v} = 1,$$

$$\sum_{v \in V} w_{k,2,v} = 1.$$

Our algorithm finds a cluster and its interval center by iterating the following three steps:

1. Apply WUR to obtain feature weights;
2. Apply CUR to obtain cluster $S$;
3. For each feature $v$, compute the average feature interval within $S$, $c_v$; define cluster $S$ center as $c = (c_v)$.

We refer to this algorithm as EXTAN following [26].

### 3.3.6 Algorithm BANCO for interval data

input: Data matrix $X$, number of clusters $K$, the weight exponent $\beta$.

Output: Cluster $K$-part partition $S$ and cluster centers.

1. Data preprocessing:
   (a) Compute grand means, $g_1$ and $g_2$, for both lower and upper bounds of the interval values.

(b) Subtract $[g_1, g_2]$ out of all the data rows.

(c) Normalize by the standard deviations.

(d) Set k=1 and $I_k = I$. Define $X(I_k)$ as the part of matrix $X$ obtained by removing all the rows $i \in I$ which do not belong to $I_k$.

2. Iterated EXTAN:

(a) Apply EXTAN to $X(I_k)$;

(b) Denote the resulting cluster by $S_k$ and its center by $c_k$;

(c) Define $I_{k+1} = I_k - S_k$ and $k = k + 1$.

(d) If $I_k \neq \emptyset$ go to **item (1) at step 2**. Otherwise, halt.

3. Return $K$ clusters $S_k$ of maximal cardinality together with their centers $c_k$.

# 4 Computation: Validation and Comparison

## 4.1 Two interval datasets

This paper introduces two novel interval datasets extracted from existing databases. One is $180 \times 5$ California Fungi dataset; the other, $76 \times 6$ Brazilian Science Production dataset.

### 4.1.1 Fungi dataset

Here we extend the 55-strong fungi dataset from [25] to include all the data available from [9]. This data relate to 588 taxa of fungi found in California [9]. Each taxon is characterized by 5 interval features:

1. pileus width,
2. stipes width,
3. stipes thickness,
4. spores height, and
5. spores width,

and a target categorical feature 'species'. Unfortunately, there are only 180 taxa with this target feature available. They form our dataset.

Here is a list of all its 26 species categories (see Table 1).

| Species | Frequency | Species | Frequency | Species | Frequency |
|---|---|---|---|---|---|
| Agaricus | 26 | Inocybe | 10 | Stropharia | 4 |
| Boletus | 19 | Suillus | 9 | Laccaria | 3 |
| Amanita | 16 | Lactarius | 8 | Coprinus | 3 |

Continued on next page

Table 1: List of species categories in Fungi data.

| | | | | | |
|---|---|---|---|---|---|
| Mycena | 14 | Hygrocybe | 8 | Strobilurus | 3 |
| Tricholoma | 13 | Marasmius | 7 | Leccinum | 2 |
| Russula | 12 | Pholiota | 5 | Hypholoma | 2 |
| Clitocybe | 10 | Psilocybe | 4 | Tylopilus | 2 |

Table 1: List of species categories in Fungi data.

The data for the most numerous 10 species are publicly available in the Github repository [10].

In our experiments, we used three datasets consisting, in respect, of the three, four, or five most numerous species in the list. They are respectively denoted as Fungi_3, Fungi_4, Fungi_5.

#### 4.1.2 Brazilian Scientific Production dataset

The original data table on research output by Brazilian academics is publicly available at [21]. This data have been assembled from databases in the National Brazilian Council on Science and Technology (CNPq) and CAPES Foundation (Coordination for the Improvement of Higher Education Personnel). According to the site, research activities of each researcher are characterized by 33 continuous variables and by three categorical features. These three features are: the institution, the field of science (grand-area-predominante), and the scientific sub-field (area predominante). The continuous features are average research output for the years 2006, 2007 and 2008 in the following 33 items:

1. National journal
2. International Journal
3. Paper
4. Monograph
5. Book chapter
6. Other publication
7. Journal abstract
8. Conference abstract
9. Publication
10. PjD completed
11. Master Program completed
12. Special training completed
13. Bachelor degree obtained
14. UR(Utilization Review) completed
15. PhD bot completed

16. Master Program not completed
17. Special training not completed
18. Bachelor Program not completed
19. UR unfinished
20. Educational training completed
21. Educational training unfinished
22. Other intelligent products
23. Other production
24. Program codes registered
25. Program codes unregistered
26. Product unregistered
27. Technology registered
28. Technology unregistered
29. Technology work
30. Technology presentation
31. Other product-related technology
32. Technology
33. Art work

These data have been summed within institutions and scientific sub-fields to obtain a $5620 \times 33$ data table used in [21].

We additionally grouped together all rows within the same scientific sub-field in a field of science, using the within-group median as the science field feature value. After this, we removed all the features that have their bottom boundary equal to zero for all the rows. The resulting interval data table has its size $76 \times 6$ and it is divided in 8 science fields (see table below). The data table is publicly available from GitHub in [10].

In our experiments, we used three datasets consisting, in respect, of the three, four, or five most numerous categories in the list.

| Species | Frequency | Species | Frequency |
|---|---|---|---|
| Biology (Ciências Biológicas) | 13 | Health Sciences (Ciências da Saúde) | 9 |
| Social Sciences (Ciências Sociais Aplicadas) | 13 | Earth Sciences (Ciências Exatas e da Terra) | 8 |
| Engineering (Engenharias) | 13 | Agricultural Sciences (Ciências Agrárias) | 7 |

Continued on next page

Table 2: List of categories in BSP data.

| Species | Frequency | Species | Frequency |
|---|---|---|---|
| Humanitarian Sciences (Ciências Humanas) | 10 | Linguistics, Literature and Arts (Lingüística, Letras e Artes) | 3 |

Table 2: List of categories in BSP data.

## 4.2 Algorithms

Here are the algorithms under investigation:

1. IKM with five versions of feature weighting:

   n no feature weighting;

   o optimal feature weighting;

   os optimal feature weighting cluster specific;

   c convenient feature weighting;

   cs convenient feature weighting cluster specific.

2. IKM preceded by Banco (BIKM), with various versions of feature weighting each.
3. KM with three versions of feature weighting

   n no feature weighting;

   c convenient feature weighting;

   cs convenient feature weighting cluster specific.

4. KM preceded by Banco (BKM) with three versions of feature weighting each

## 4.3 Assessment of results

To evaluate and compare obtained clustering results, we use two popular metrics of similarity between partitions: 1) Adjusted Rand Index (ARI) [13], and 2) Normalised Mutual Information (NMI) [6].

To define the Adjusted Rand Index, one uses what is referred to as contingency table in statistics. Given two partitions, $S = \{S_1, S_2, ..., S_K\}$ and $T = \{T_1, T_2, ..., T_L\}$, a contingency table is a two-way table whose rows correspond to parts $S_k$ $(k = 1, 2, ..., K)$ of $S$, and its columns, to parts $T_l$ $(l = 1, 2, ..., L)$ of $T$. The $(k, l)$-th entry is $n_{kl} = |S_k \cap T_l|$, the frequency of $(k, l)$ co-occurrence. The so-called marginal row $a$ and marginal column $b$ are defined by $a_k = \sum_{l=1}^{L} n_{kl} = |S_k|$ and $b_l = \sum_{k=1}^{K} n_{kl} = |T_l|$.

The Adjusted Rand Index is defined as:

$$ARI(S,T) = \frac{\sum_{k,l}\binom{n_{kl}}{2} - [\sum_k \binom{a_k}{2} \sum_l \binom{b_l}{2}]/\binom{N}{2}}{\frac{1}{2}[\sum_k \binom{a_k}{2} + \sum_l \binom{b_l}{2}] - [\sum_k \binom{a_k}{2} \sum_l \binom{b_l}{2}]/\binom{N}{2}] } \quad (20)$$

Considering partition $S$ as the ground truth, whereas partition $T$ - the found clusters, ARI value gives an estimation of the similarity between the two. The closer the value of ARI to unity, the better the match between the partitions; ARI=1 if and only if $S = T$. If one of the partitions consists of just one part, the set $I$ itself, then ARI=0.

The NMI index is defined by using the concept of entropy. The entropy of partition $\mathbf{S}$ is defined as $H(\mathbf{S}) = -\sum_{k=1}^{K} p(k)\log(p(k))$ where $p(k) = |S_k|/N = a(k)/N$ is the probability that an object picked at random falls into $S_k$. Given a partition $\mathbf{T}$, $H(T)$ is defined similarly. The mutual information (MI) between $\mathbf{S}$ and $\mathbf{T}$ is defined as:

$$MI(\mathbf{S},\mathbf{T}) = \sum_{k}^{K}\sum_{l=1}^{L} p_{kl}\log(\frac{p_{kl}}{a(k)b(l)}), \tag{21}$$

where $p_{kl} = n_{kl}/N$ is the probability that a random object falls into both $S_k$ and $T_l$ ($k = 1, 2, ..., K;\ l = 1, 2, ..., L$). The normalised mutual information is defined as

$$NMI = \frac{MI(\mathbf{S},\mathbf{T})}{max(H(\mathbf{S}), H(\mathbf{T}))}. \tag{22}$$

NMI ranges between 0 and 1. Its values close to zero indicate random clustering results, whereas the closer NMI to unity the better is the match between $S$ and $T$.

## 4.4 Computational results

### 4.4.1 K-means clustering preceded by Banco

For the sake of convenience, we present computational results of our k-means algorithms preceded by Banco in two tables. Table 3 presents results for computations at which both Banco and k-means, in both versions, IKM and KM, used the same feature weighting scheme. Table 4 presents results for computations at which Banco and k-means used different feature weighting schemes. The entries in both tables are ARI index values separated by slash from NMI index values. Both indexes show the degree of cluster recovery by the corresponding clustering algorithm.

| Name | Fungi_3 | Fungi_4 | Fungi_5 | BSP_3 | BSP_4 | BSP_5 |
|---|---|---|---|---|---|---|
| BIKMn | .62 / .56 | .59 / .64 | .47 / .58 | .78 / .78 | .69 / .79 | .50 / .57 |
| BIKMc | .81 / .76 | .75 / .74 | .57 / .67 | .85 / .87 | .72 / .75 | .57 / .68 |
| BIKMo | .81 / .76 | .58 / .60 | .40 / .51 | .85 / .87 | .76 / .81 | .57 / .67 |
| BIKMcs | .81 / .76 | .42 / .49 | .44 / .57 | .92 / .92 | .65 / .70 | .55 / .66 |
| BIKMos | .81 / .76 | .47 / .53 | .37 / .50 | .92 / .92 | .69 / .79 | .49 / .59 |
| BKMn | .77 / .70 | .64 / .69 | .50 / .61 | .73 / .75 | .65 / .78 | .50 / .57 |
| BKMc | .85 / .82 | .78 / .78 | .58 / .68 | .85 / .87 | .68 / .77 | .61 / .71 |
| BKMcs | .73 / .69 | .45 / .52 | .49 / .61 | .92 / .92 | .62 / .73 | .61 / .71 |

Table 3: Values of ARI/NMI indexes at the datasets under investigation obtained by k-means clustering algorithms preceded by Banco: the same feature weighting schemes).

| Name | Fungi_3 | Fungi_4 | Fungi_5 | BSP_3 | BSP_4 | BSP_5 |
|---|---|---|---|---|---|---|
| BnIKMc | .81 / .76 | .75 / .74 | .54 / .63 | .85 / .87 | .68 / .74 | .50 / .57 |
| BnIKMo | .81 / .76 | .75 / .74 | .56 / .65 | .85 / .87 | .71 / .79 | .50 / .57 |
| BnIKMcs | .81 / .76 | .67 / .70 | .49 / .58 | .92 / .92 | .71 / .79 | .51 / .58 |
| BnIKMos | .81 / .76 | .64 / .69 | .49 / .59 | .92 / .92 | .69 / .79 | .50 / .57 |
| BcIKMn | .47 / .53 | .44 / .52 | .43 / .54 | .85 / .87 | .64 / .69 | .54 / .65 |
| BcIKMo | .81 / .76 | .78 / .77 | .57 / .66 | .85 / .87 | .68 / .74 | .68 / .77 |
| BcIKMos | .81 / .76 | .47 / .53 | .50 / .62 | .92 / .92 | .64 / .69 | .61 / .70 |
| BoIKMn | .51 / .55 | .42 / .51 | .43 / .54 | .78 / .78 | .69 / .79 | .49 / .62 |
| BoIKMc | .81 / .76 | .58 / .63 | .43 / .55 | .85 / .87 | .68 / .75 | .53 / .62 |
| BoIKMcs | .64 / .62 | .42 / .49 | .34 / .49 | .92 / .92 | .72 / .79 | .47 / .58 |
| BnKMc | .81 / .76 | .75 / .74 | .50 / .60 | .85 / .87 | .68 / .74 | .50 / .57 |
| BnKMcs | .81 / .76 | .67 / .70 | .49 / .59 | .92 / .92 | .61 / .68 | .50 / .57 |
| BcKMn | .52 / .54 | .64 / .69 | .49 / .60 | .73 / .75 | .67 / .78 | .57 / .70 |
| BoKMn | .56 / .52 | .44 / .52 | .38 / .50 | .73 / .75 | .65 / .78 | .51 / .65 |
| BoKMc | .81 / .76 | .59 / .60 | .38 / .50 | .85 / .87 | .72 / .79 | .53 / .62 |
| BnKMcs | .55 / .58 | .45 / .52 | .38 / .50 | .92 / .92 | .65 / .70 | .51 / .64 |

Table 4: Values of ARI/NMI indexes at the datasets under investigation obtained by k-means clustering algorithms preceded by Banco: different feature weighting options).

Let us point out general features of the tables.

- Values of the indexes, ARI and NMI, generally, agree: the larger values of ARI correspond to larger values of NMI. Therefore, we are going to consider only ARI values for further analyses.

- Index values are greater, at both data tables, at the 3-part partition, and are much smaller at the 5-part partition, so that they are at medium levels at 4-part partitions. This goes in line with our expectations: the greater the granularity of a partition, the more difficult to reproduce that with clustering.

Now we can turn to specifics of our algorithms.

- In contrast to our expectations, the variability in weighting options between Banko and K-means, in general, yields no better results. The maxima of ARI in Table 3 overall are greater than those in Table 4 for both Fungi and BSP. Specifically, for Fungi_3,_4,_5, the maxima in Table 3 are 0.85, 0.78, 0. 58, respectively, whereas those in Table 4 are somewhat smaller: 0.81, 0.75, 0.56, respectively. Similarly, the maxima for BSP_3, _4, in table 3, 0.92, 0.76, are respectively better than those in Table 4, 0.92, 0.72. The only exception from this rule occurs at BSP_5: The value in Table 4, 0.68 is greater than that in Table 3, 0.61.

- Using our feature weighting schemas does bring forth the best results. For example, the maximum ARI value of 0.92 has been reached at clustering set BSP_3 with algorithm BoIKMcs at which both Banko and K-means used feature weighting schemes. It should be noted that the same ARI value can be reached with the generic Banko, with no weight adjustments at all, by using BnIKMcs, BnIKMos and BmKMcs. Each of these, however, involves a cluster-specific feature-weighting scheme.

- Our last observation concerns relation between two approaches to clustering interval data: a genuine one and a naive one. Our genuine approach leads to a series of IKM algorithms, whereas our naive approach – dismissing intervals altogether and just doubling the number of features – leads to a series of KM algorithms. They both show more or less similar results. Sometimes it is the KM series which

wins, as, for example, at recovery of Fungi_3, Fungi_4, and Fungi_5 partitions. Here the best recovery results are shown by the algorithm BcKMc out of KM series: its ARI values 0.85, 0.78, and 0.58 are the maxima over the respective partitions. In contrast, the winner at BSP_4 is BoIKMo (ARI=0.76), and at BSP_5, the BcIKMo (ARI=0.68). These observations show that interval modeling needs a deeper insight to make it more effective.

It remains to take a look at performances of Banco-preceded k-means methods in comparison with the conventional multi-start k-means clustering at which the final cluster solution is selected from results of multiple runs of k-means starting from random initial centers each.

| alg \ dataset | Fungi_3 | Fungi_4 | Fungi_5 | BSP_3 | BSP_4 | BSP_5 |
|---|---|---|---|---|---|---|
| IKMn | .77 / .40 / .17 | .64 / .43 / .10 | .54 / .38 / .07 | 1.00 / .56 / .23 | .73 / .55 / .13 | .58 / .43 / .09 |
| IKMc | .85 / .39 / .26 | .78 / .49 / .16 | .59 / .40 / .10 | 1.00 / .44 / .20 | .72 / .43 / .17 | .63 / .35 / .13 |
| IKMo | .85 / .42 / .25 | .78 / .46 / .16 | .59 / .40 / .08 | 1.00 / .49 / .25 | .76 / .45 / .18 | .68 / .32 / .14 |
| IKMcs | .90 / .40 / .24 | .76 / .33 / .15 | .58 / .37 / .09 | 1.00 / .48 / .20 | .74 / .44 / .15 | .67 / .33 / .12 |
| IKMos | .86 / .40 / .21 | .82 / .42 / .14 | .56 / .38 / .09 | 1.00 / .46 / .22 | .73 / .45 / .15 | .67 / .34 / .11 |
| KMn | .81 / .54 / .13 | .64 / .51 / .10 | .55 / .41 / .07 | 1.00 / .57 / .22 | .69 / .57 / .12 | .61 / .44 / .08 |
| KMc | .85 / .40 / .29 | .78 / .45 / .15 | .62 / .40 / .09 | 1.00 / .45 / .21 | .72 / .48 / .17 | .68 / .36 / .13 |
| KMcs | .86 / .15 / .14 | .78 / .43 / .14 | .59 / .36 / .09 | 1.00 / .32 / .29 | .72 / .33 / .24 | .71 / .05 / .14 |

Table 5: ARI index at the datasets under investigation after a hundred random-start runs of the corresponding k-means algorithms. Every entry consists of three values: the maximum, the mean, and the standard deviation of the ARI values.

Table 5 represents results found at 100 random-start runs of our IKM and KM algorithms. The value $\beta$ has been adjusted each time in such a way that the best recovery results are achieved (see Appendix C.) Each entry contains three ARI index values, the maximum, the average and the standard deviation. For example, the very first entry on top of the table, .77/.40/.17, reports that at a hundred runs of the generic IKM algorithm, with no feature weighting involved, the maximum ARI index value was 0.77, the average 0.40, and the standard deviation from the average, 0.17.

Table 6 shows similar results for the NMI index.

The ARI values, presented in Table 5, show that the best partition recovery results almost always are greater than the best results achieved with Banco-preceded algorithms. But the difference is not that large: it is a fraction of the standard deviation value, usually of the order of 0.2-0.3 of that.

This allows us to conclude that using Banco for initialization of k-means algorithms modified for interval data is highly beneficial. It involves just a single run of the algorithm instead of a multitude of random-start runs leading to many solutions. Usually, the ground truth is unknown, so the user, in the latter case, faces a cumbersome task of selection of a most appropriate solution.

| alg \ dataset | Fungi_3 | Fungi_4 | Fungi_5 | BSP_3 | BSP_4 | BSP_5 |
|---|---|---|---|---|---|---|
| IKMn | .70 / .39 / .14 | .69 / .52 / .08 | .65 / .50 / .06 | 1.00 / .63 / .19 | .85 / .66 / .12 | .74 / .56 / .07 |
| IKMc | .82 / .42 / .23 | .78 / .57 / .12 | .68 / .52 / .08 | 1.00 / .53 / .17 | .81 / .54 / .15 | .71 / .50 / .10 |
| IKMo | .82 / .44 / .21 | .78 / .55 / .12 | .68 / .52 / .07 | 1.00 / .59 / .21 | .81 / .57 / .16 | .77 / .47 / .12 |
| IKMcs | .88 / .43 / .20 | .78 / .45 / .13 | .68 / .49 / .08 | 1.00 / .56 / .18 | .81 / .56 / .13 | .74 / .49 / .10 |
| IKMos | .82 / .42 / .17 | .81 / .52 / .11 | .66 / .50 / .07 | 1.00 / .54 / .18 | .85 / .57 / .13 | .74 / .51 / .10 |

Continued on next page

Table 6: NMI index at the datasets under investigation after a hundred random-start runs of the corresponding k-means algorithms. Every entry consists of three values: the maximum, the mean, and the standard deviation of the NMI values.

| | | | | | | |
|---|---|---|---|---|---|---|
| KMn | .76 / .54 / .09 | .69 / .59 / .08 | .65 / .53 / .06 | 1.00 / .63 / .18 | .78 / .69 / .11 | .72 / .60 / .07 |
| KMc | .82 / .42 / .25 | .78 / .54 / .12 | .69 / .53 / .08 | 1.00 / .55 / .19 | .84 / .60 / .14 | .77 / .50 / .10 |
| KMcs | .82 / .22 / .15 | .78 / .53 / .11 | .67 / .49 / .08 | 1.00 / .43 / .24 | .84 / .48 / .21 | .75 / .22 / .14 |

Table 6: NMI index at the datasets under investigation after a hundred random-start runs of the corresponding k-means algorithms. Every entry consists of three values: the maximum, the mean, and the standard deviation of the NMI values.

# 5 Conclusion

Interval data is an important class of complex structure data. Clustering is an important data science approach recently extended to interval data with a most popular tool, k-means clustering. On par with many advantages, k-means suffers from some shortcomings. One of them is lack of instruments for choosing initial cluster centers. This paper proposes using anomalous clusters as adequate center bearers. This approach is consistent with the meaning of the least squares criterion. As follows from equation (2), to minimize it, one needs to find most numerous anomalous clusters. We propose a method, Banco, for one-by-one finding most anomalous clusters, so that k-means computations start with the centers of $K$ most numerous of them.

Also, we propose several competing feature weighting schemes to use within the k-means clustering framework.

We introduce two novel interval datasets with innate cluster structure. One of them. Fungi, further extends the dataset used in previously [25] from 55 specimen to 180. The other, is a categorisation of the data related to research output of Brazilian scientists into a $76 \times 6$ tata table. Both datasets have external categories assigned to them: taxa, for fungies, and research domains, for research outputs.

We take sets of three or four or five the most numerous categories out of the two data tables – six sets altogether, and compare various versions of k-means approach with respect to their ability to recover the category structures from the data. The level of recovery is assessed by conventional indexes of similarity between the innate partition and that found by an algorithm, the ARI and NMI coefficients.

The variety of clustering algorithms under investigation stems from three divisions. One of the divisions comes from the view of interval data. One, genuine, view takes the intervals as feature values. The other, a naive view, removes the intervals altogether, by considering interval data as a double data table at which every interval feature is represented by two conventional features, one for the left, the other for the right boundary of the interval. Another division relates to the fact whether our Banco algorithm is involved or not. The third division concerns the way we assign weights to features. There are three ways for feature weighting: constant weights, optimally adjusted weights, and conventionally adjusted weights. Further differences emerge depending on whether Banco algorithm and follow-up k-means algorithm involve the same or different feature weighting schemas.

Our experimental computations show that using Banco algorithm for initialization is beneficial for clustering. Other findings concern more specific properties, as for example, our observation that using the same weighting scheme at both Banco and k-means overall leads to better results than using different weighting schemes.

Further work should obtain better insights into the nature of interval data, perhaps by using within-interval distributions, to obtain superior cluster recovery results.

All the data tables and the code of the algorithms from this article, as well as testing results are publicly available in the GitHub repository [11].

# Appendices

# A Fungi dataset 5 clusters

| index | genera | name | spores 1d | spores 2d | pileus width | stipes long | stipes thick |
|---|---|---|---|---|---|---|---|
| 0 | Agaricus | moronii | [.06, .075] | [.04, .05] | [600, 1200] | [200, 700] | [150, 300] |
| 1 | Agaricus | subrutilescens | [.04, .06] | [.035, .045] | [600, 1400] | [600, 1600] | [100, 200] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2 | Agaricus | smithianus | [.07, .09] | [.05, .055] | [700, 1200] | [500, 1200] | [200, 300] |
| 3 | Agaricus | sylvicola | [.055, .065] | [.035, .04] | [600, 1200] | [600, 1200] | [150, 200] |
| 4 | Agaricus | semotus | [.045, .055] | [.03, .035] | [200, 600] | [300, 700] | [40, 80] |
| 5 | Agaricus | perobscurus | [.065, .08] | [.045, .05] | [800, 1200] | [600, 1200] | [150, 200] |
| 6 | Agaricus | pattersonae | [.07, .09] | [.06, .065] | [500, 1500] | [600, 1500] | [250, 350] |
| 7 | Agaricus | micromegathus | [.045, .05] | [.03, .035] | [250, 400] | [250, 450] | [40, 70] |
| 8 | Agaricus | lilaceps | [.05, .065] | [.04, .05] | [800, 2000] | [900, 1900] | [300, 500] |
| 9 | Agaricus | hondensis | [.04, .06] | [.03, .045] | [700, 1400] | [800, 1400] | [150, 250] |
| 10 | Agaricus | fuscovelatus | [.07, .08] | [.05, .06] | [350, 800] | [400, 1000] | [100, 200] |
| 11 | Agaricus | diminutivus | [.04, .05] | [.03, .04] | [150, 250] | [300, 600] | [25, 35] |
| 12 | Agaricus | deardorffensis | [.04, .06] | [.035, .045] | [700, 1900] | [800, 1500] | [200, 350] |
| 13 | Agaricus | xanthodermus | [.05, .06] | [.04, .055] | [500, 1700] | [400, 1400] | [100, 350] |
| 14 | Agaricus | comtulus | [.04, .05] | [.03, .035] | [250, 400] | [300, 500] | [40, 70] |
| 15 | Agaricus | arorae | [.045, .05] | [.03, .035] | [300, 800] | [400, 900] | [50, 250] |
| 16 | Agaricus | incultorum | [.07, .08] | [.05, .06] | [250, 600] | [150, 350] | [100, 150] |
| 17 | Agaricus | augustus | [.075, .105] | [.05, .065] | [600, 3200] | [1000, 3700] | [600, 600] |
| 18 | Agaricus | benesii | [.05, .06] | [.03, .04] | [400, 800] | [500, 1100] | [100, 200] |
| 19 | Agaricus | bernardii | [.055, .07] | [.055, .065] | [700, 1600] | [400, 700] | [300, 450] |
| 20 | Agaricus | fissuratus | [.065, .09] | [.045, .06] | [600, 2100] | [400, 1400] | [100, 350] |
| 21 | Agaricus | subrufescens | [.055, .065] | [.04, .045] | [600, 1300] | [600, 1200] | [150, 250] |
| 22 | Agaricus | brunneofibrillosus | [.05, .065] | [.035, .04] | [400, 1500] | [400, 1500] | [150, 250] |
| 23 | Agaricus | californicus | [.05, .075] | [.04, .055] | [400, 1100] | [300, 700] | [40, 100] |
| 24 | Agaricus | campestris | [.055, .08] | [.035, .05] | [500, 1000] | [300, 600] | [100, 200] |
| 25 | Agaricus | bitorquis | [.05, .065] | [.04, .055] | [500, 1500] | [400, 1000] | [200, 400] |
| 26 | Amanita | vernicoccora | [.09, .12] | [.06, .08] | [800, 2000] | [700, 2500] | [400, 400] |
| 27 | Amanita | velosa | [.085, .12] | [.07, .11] | [500, 1100] | [400, 1100] | [100, 250] |
| 28 | Amanita | vaginata | [.08, .115] | [.075, .1] | [550, 1000] | [600, 1300] | [120, 200] |
| 29 | Amanita | smithiana | [.105, .12] | [.065, .09] | [700, 1400] | [700, 1700] | [200, 400] |
| 30 | Amanita | phalloides | [.07, .12] | [.06, .1] | [350, 1500] | [400, 1800] | [100, 300] |
| 31 | Amanita | pantherina | [.095, .13] | [.07, .095] | [400, 1500] | [700, 1100] | [100, 250] |
| 32 | Amanita | pachycolea | [.115, .14] | [.1, .12] | [800, 1800] | [1000, 2500] | [100, 300] |
| 33 | Amanita | ocreata | [.09, .125] | [.07, .09] | [500, 1300] | [1000, 2200] | [150, 300] |
| 34 | Amanita | muscaria | [.09, .13] | [.065, .095] | [600, 3900] | [700, 1600] | [200, 300] |
| 35 | Amanita | gemmata | [.08, .13] | [.065, .09] | [300, 1100] | [400, 1400] | [100, 200] |
| 36 | Amanita | constricta | [.095, .095] | [.115, .115] | [600, 1200] | [900, 1700] | [100, 200] |
| 37 | Amanita | calyptratoides | [.098, .14] | [.065, .089] | [300, 1000] | [500, 1450] | [60, 200] |
| 38 | Amanita | calyptroderma | [.08, .11] | [.05, .06] | [800, 2500] | [1000, 2000] | [150, 400] |
| 39 | Amanita | augusta | [.08, .12] | [.06, .08] | [400, 1200] | [500, 1500] | [100, 200] |
| 40 | Amanita | aprica | [.08, .13] | [.06, .085] | [500, 1500] | [350, 900] | [350, 350] |
| 41 | Amanita | novinupta | [.07, .085] | [.055, .06] | [500, 1400] | [600, 1200] | [150, 350] |
| 42 | Boletus | rex veris | [.125, .18] | [.04, .05] | [900, 1800] | [500, 1000] | [200, 600] |
| 43 | Boletus | Rubropulcherrimus | [.13, .155] | [.05, .06] | [900, 1700] | [700, 1400] | [800, 800] |
| 44 | Boletus | Butyriautumniregius | [.13, .155] | [.04, .05] | [800, 1500] | [500, 900] | [300, 400] |
| 45 | Boletus | Calorubripes | [.12, .165] | [.045, .055] | [600, 1600] | [600, 1500] | [300, 500] |
| 46 | Boletus | X. mendocinensis | [.12, .15] | [.045, .06] | [500, 1000] | [500, 1000] | [150, 250] |
| 47 | Boletus | X. subtomentosus | [.1, .15] | [.04, .05] | [400, 1200] | [400, 800] | [100, 200] |
| 48 | Boletus | X. atropurpureus | [.11, .15] | [.04, .06] | [400, 1100] | [500, 1000] | [100, 300] |
| 49 | Boletus | orovillus | [.055, .065] | [.035, .04] | [800, 1500] | [500, 900] | [250, 450] |
| 50 | Boletus | smithii | [.135, .135] | [.16, .16] | [700, 1500] | [700, 1500] | [350, 700] |
| 51 | Boletus | Calofrustosus | [.11, .14] | [.04, .05] | [700, 1500] | [500, 1000] | [250, 350] |
| 52 | Boletus | Aureocitriniporus | [.12, .135] | [.0375, .045] | [400, 800] | [400, 700] | [100, 300] |
| 53 | Boletus | Rubroeastwoodiae | [.11, .15] | [.035, .06] | [1000, 2200] | [700, 1400] | [1300, 1300] |
| 54 | Boletus | X. dryophilus | [.115, .16] | [.05, .065] | [400, 1200] | [400, 800] | [100, 250] |
| 55 | Boletus | X. diffractus | [.115, .14] | [.04, .06] | [400, 900] | [500, 1000] | [100, 150] |
| 56 | Boletus | Butyripersolidus | [.115, .135] | [.035, .045] | [700, 1400] | [500, 900] | [300, 600] |
| 57 | Boletus | S. amygdalinus | [.11, .14] | [.05, .065] | [400, 1000] | [400, 700] | [150, 300] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 58 | Boletus | regineus | [.115, .135] | [.035, .045] | [700, 1400] | [700, 1300] | [300, 400] |
| 59 | Boletus | Aureoflaviporus | [.12, .15] | [.05, .06] | [600, 1100] | [600, 1200] | [100, 200] |
| 60 | Boletus | edulis | [.12, .17] | [.04, .06] | [700, 2500] | [700, 2000] | [300, 800] |
| 61 | Mycena | purpureofusca | [.07, .1] | [.05, .06] | [70, 300] | [300, 700] | [10, 40] |
| 62 | Mycena | pura | [.06, .085] | [.03, .04] | [150, 450] | [200, 600] | [20, 70] |
| 63 | Mycena | overholtsii | [.055, .07] | [.03, .035] | [200, 600] | [1500, 1500] | [150, 150] |
| 64 | Mycena | oregonensis | [.065, .085] | [.03, .035] | [20, 80] | [100, 250] | [100, 100] |
| 65 | Mycena | maculata | [.075, .095] | [.05, .055] | [150, 400] | [200, 900] | [15, 40] |
| 66 | Mycena | haematopus | [.075, .09] | [.045, .055] | [100, 300] | [250, 700] | [20, 30] |
| 67 | Mycena | galericulata | [.085, .105] | [.06, .075] | [200, 500] | [300, 1400] | [20, 50] |
| 68 | Mycena | nivicola | [.085, .115] | [.05, .06] | [150, 300] | [250, 900] | [20, 30] |
| 69 | Mycena | californiensis | [.075, .09] | [.04, .045] | [70, 200] | [200, 700] | [10, 20] |
| 70 | Mycena | aurantiomarginata | [.075, .09] | [.04, .055] | [100, 200] | [250, 700] | [10, 20] |
| 71 | Mycena | amicta | [.08, .095] | [.04, .05] | [50, 150] | [300, 700] | [10, 30] |
| 72 | Mycena | tenerrima | [.08, .105] | [.04, .06] | [20, 40] | [40, 100] | [10, 10] |
| 73 | Mycena | acicula | [.085, .115] | [.03, .04] | [20, 80] | [100, 500] | [5, 5] |
| 74 | Mycena | capillaripes | [.08, .11] | [.04, .065] | [100, 200] | [400, 600] | [10, 20] |
| 75 | Tricholoma | sejunctum | [.05, .08] | [.035, .055] | [400, 900] | [300, 1000] | [100, 150] |
| 76 | Tricholoma | saponaceum | [.05, .065] | [.035, .045] | [400, 900] | [450, 800] | [150, 200] |
| 77 | Tricholoma | muricatum | [.045, .06] | [.03, .035] | [500, 1200] | [300, 600] | [100, 350] |
| 78 | Tricholoma | moseri | [.065, .1] | [.035, .05] | [200, 450] | [200, 500] | [50, 100] |
| 79 | Tricholoma | imbricatum | [.055, .07] | [.04, .05] | [600, 1500] | [500, 1000] | [200, 350] |
| 80 | Tricholoma | dryophilum | [.05, .06] | [.04, .0425] | [500, 1500] | [600, 1300] | [100, 450] |
| 81 | Tricholoma | fracticum | [.055, .075] | [.04, .055] | [500, 1000] | [300, 800] | [150, 250] |
| 82 | Tricholoma | atroviolaceum | [.075, .09] | [.05, .06] | [350, 900] | [400, 800] | [150, 300] |
| 83 | Tricholoma | vernaticum | [.085, .11] | [.04, .06] | [400, 1400] | [500, 1300] | [200, 350] |
| 84 | Tricholoma | murrillianum | [.05, .07] | [.045, .055] | [500, 2500] | [400, 1500] | [100, 600] |
| 85 | Tricholoma | equestre | [.06, .075] | [.035, .05] | [500, 1300] | [400, 800] | [150, 300] |
| 86 | Tricholoma | griseoviolaceum | [.05, .07] | [.035, .05] | [400, 1100] | [600, 1400] | [100, 200] |
| 87 | Tricholoma | myomyces | [.05, .075] | [.035, .045] | [150, 500] | [250, 500] | [50, 100] |

# B Fungi dataset 180 species

| index | genera | name | spores 1d | spores 2d | pileus width | stipes long | stipes thick |
|---|---|---|---|---|---|---|---|
| 0 | Agaricus | moronii | [.060, .075] | [.040, .050] | [ 600, 1200] | [ 200, 700] | [ 150, 300] |
| 1 | Agaricus | xanthodermus | [.050, .060] | [.040, .055] | [ 500, 1700] | [ 400, 1400] | [ 100, 350] |
| 2 | Agaricus | subrutilescens | [.040, .060] | [.035, .045] | [ 600, 1400] | [ 600, 1600] | [ 100, 200] |
| 3 | Agaricus | smithianus | [.070, .090] | [.050, .055] | [ 700, 1200] | [ 500, 1200] | [ 200, 300] |
| 4 | Agaricus | sylvicola | [.055, .065] | [.035, .040] | [ 600, 1200] | [ 600, 1200] | [ 150, 200] |
| 5 | Agaricus | semotus | [.045, .055] | [.030, .035] | [ 200, 600] | [ 300, 700] | [ 40, 80] |
| 6 | Agaricus | perobscurus | [.065, .080] | [.045, .050] | [ 800, 1200] | [ 600, 1200] | [ 150, 200] |
| 7 | Agaricus | pattersonae | [.070, .090] | [.060, .065] | [ 500, 1500] | [ 600, 1500] | [ 250, 350] |
| 8 | Agaricus | micromegathus | [.045, .050] | [.030, .035] | [ 250, 400] | [ 250, 450] | [ 40, 70] |
| 9 | Agaricus | lilaceps | [.050, .065] | [.040, .050] | [ 800, 2000] | [ 900, 1900] | [ 300, 500] |
| 10 | Agaricus | fuscovelatus | [.070, .080] | [.050, .060] | [ 350, 800] | [ 400, 1000] | [ 100, 200] |
| 11 | Agaricus | diminutivus | [.040, .050] | [.030, .040] | [ 150, 250] | [ 300, 600] | [ 25, 35] |
| 12 | Agaricus | deardorffensis | [.040, .060] | [.035, .045] | [ 700, 1900] | [ 800, 1500] | [ 200, 350] |
| 13 | Agaricus | hondensis | [.040, .060] | [.030, .045] | [ 700, 1400] | [ 800, 1400] | [ 150, 250] |
| 14 | Agaricus | comtulus | [.040, .050] | [.030, .035] | [ 250, 400] | [ 300, 500] | [ 40, 70] |
| 15 | Agaricus | arorae | [.045, .050] | [.030, .035] | [ 300, 800] | [ 400, 900] | [ 50, 250] |
| 16 | Agaricus | fissuratus | [.065, .090] | [.045, .060] | [ 600, 2100] | [ 400, 1400] | [ 100, 350] |
| 17 | Agaricus | incultorum | [.070, .080] | [.050, .060] | [ 250, 600] | [ 150, 350] | [ 100, 150] |
| 18 | Agaricus | benesii | [.050, .060] | [.030, .040] | [ 400, 800] | [ 500, 1100] | [ 100, 200] |
| 19 | Agaricus | bernardii | [.055, .070] | [.055, .065] | [ 700, 1600] | [ 400, 700] | [ 300, 450] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 20 | Agaricus | augustus | [.075, .105] | [.050, .065] | [ 600, 3200] | [ 1000, 3700] | [ 600, 600] |
| 21 | Agaricus | subrufescens | [.055, .065] | [.040, .045] | [ 600, 1300] | [ 600, 1200] | [ 150, 250] |
| 22 | Agaricus | brunneofibrillosus | [.050, .065] | [.035, .040] | [ 400, 1500] | [ 400, 1500] | [ 150, 250] |
| 23 | Agaricus | californicus | [.050, .075] | [.040, .055] | [ 400, 1100] | [ 300, 700] | [ 40, 100] |
| 24 | Agaricus | campestris | [.055, .080] | [.035, .050] | [ 500, 1000] | [ 300, 600] | [ 100, 200] |
| 25 | Agaricus | bitorquis | [.050, .065] | [.040, .055] | [ 500, 1500] | [ 400, 1000] | [ 200, 400] |
| 26 | Amanita | pachycolea | [.115, .140] | [.100, .120] | [ 800, 1800] | [ 1000, 2500] | [ 100, 300] |
| 27 | Amanita | vernicoccora | [.090, .120] | [.060, .080] | [ 800, 2000] | [ 700, 2500] | [ 400, 400] |
| 28 | Amanita | vaginata | [.080, .115] | [.075, .100] | [ 550, 1000] | [ 600, 1300] | [ 120, 200] |
| 29 | Amanita | smithiana | [.105, .120] | [.065, .090] | [ 700, 1400] | [ 700, 1700] | [ 200, 400] |
| 30 | Amanita | phalloides | [.070, .120] | [.060, .100] | [ 350, 1500] | [ 400, 1800] | [ 100, 300] |
| 31 | Amanita | pantherina | [.095, .130] | [.070, .095] | [ 400, 1500] | [ 700, 1100] | [ 100, 250] |
| 32 | Amanita | ocreata | [.090, .125] | [.070, .090] | [ 500, 1300] | [ 1000, 2200] | [ 150, 300] |
| 33 | Amanita | velosa | [.085, .120] | [.070, .110] | [ 500, 1100] | [ 400, 1100] | [ 100, 250] |
| 34 | Amanita | muscaria | [.090, .130] | [.065, .095] | [ 600, 3900] | [ 700, 1600] | [ 200, 300] |
| 35 | Amanita | gemmata | [.080, .130] | [.065, .090] | [ 300, 1100] | [ 400, 1400] | [ 100, 200] |
| 36 | Amanita | constricta | [.095, .095] | [.115, .115] | [ 600, 1200] | [ 900, 1700] | [ 100, 200] |
| 37 | Amanita | calyptratoides | [.098, .140] | [.065, .089] | [ 300, 1000] | [ 500, 1450] | [ 60, 200] |
| 38 | Amanita | calyptroderma | [.080, .110] | [.050, .060] | [ 800, 2500] | [ 1000, 2000] | [ 150, 400] |
| 39 | Amanita | augusta | [.080, .120] | [.060, .080] | [ 400, 1200] | [ 500, 1500] | [ 100, 200] |
| 40 | Amanita | aprica | [.080, .130] | [.060, .085] | [ 500, 1500] | [ 350, 900] | [ 350, 350] |
| 41 | Amanita | novinupta | [.070, .085] | [.055, .060] | [ 500, 1400] | [ 600, 1200] | [ 150, 350] |
| 42 | Boletus | C. frustosus | [.110, .140] | [.040, .050] | [ 700, 1500] | [ 500, 1000] | [ 250, 350] |
| 43 | Boletus | orovillus | [.055, .065] | [.035, .040] | [ 800, 1500] | [ 500, 900] | [ 250, 450] |
| 44 | Boletus | Rubropulcherrimus | [.130, .155] | [.050, .060] | [ 900, 1700] | [ 700, 1400] | [ 800, 800] |
| 45 | Boletus | Butyriautumniregius | [.130, .155] | [.040, .050] | [ 800, 1500] | [ 500, 900] | [ 300, 400] |
| 46 | Boletus | X. subtomentosus | [.100, .150] | [.040, .050] | [ 400, 1200] | [ 400, 800] | [ 100, 200] |
| 47 | Boletus | smithii | [.135, .135] | [.160, .160] | [ 700, 1500] | [ 700, 1500] | [ 350, 700] |
| 48 | Boletus | edulis | [.120, .170] | [.040, .060] | [ 700, 2500] | [ 700, 2000] | [ 300, 800] |
| 49 | Boletus | X. mendocinensis | [.120, .150] | [.045, .060] | [ 500, 1000] | [ 500, 1000] | [ 150, 250] |
| 50 | Boletus | X. atropurpureus | [.110, .150] | [.040, .060] | [ 400, 1100] | [ 500, 1000] | [ 100, 300] |
| 51 | Boletus | Calorubripes | [.120, .165] | [.045, .055] | [ 600, 1600] | [ 600, 1500] | [ 300, 500] |
| 52 | Boletus | Rubroeastwoodiae | [.110, .150] | [.035, .060] | [ 1000, 2200] | [ 700, 1400] | [ 1300, 1300] |
| 53 | Boletus | rex-veris | [.125, .180] | [.040, .050] | [ 900, 1800] | [ 500, 1000] | [ 200, 600] |
| 54 | Boletus | X. diffractus | [.115, .140] | [.040, .060] | [ 400, 900] | [ 500, 1000] | [ 100, 150] |
| 55 | Boletus | Butyripersolidus | [.115, .135] | [.035, .045] | [ 700, 1400] | [ 500, 900] | [ 300, 600] |
| 56 | Boletus | S. amygdalinus | [.110, .140] | [.050, .065] | [ 400, 1000] | [ 400, 700] | [ 150, 300] |
| 57 | Boletus | regineus | [.115, .135] | [.035, .045] | [ 700, 1400] | [ 700, 1300] | [ 300, 400] |
| 58 | Boletus | Aureoflaviporus | [.120, .150] | [.050, .060] | [ 600, 1100] | [ 600, 1200] | [ 100, 200] |
| 59 | Boletus | Aureocitriniporus | [.120, .135] | [.037, .045] | [ 400, 800] | [ 400, 700] | [ 100, 300] |
| 60 | Boletus | X. dryophilus | [.115, .160] | [.050, .065] | [ 400, 1200] | [ 400, 800] | [ 100, 250] |
| 61 | Clitocybe | tarda | [.055, .080] | [.035, .040] | [ 200, 600] | [ 150, 500] | [ 30, 70] |
| 62 | Clitocybe | sclerotoidea | [.075, .075] | [.100, .100] | [ 100, 300] | [ 100, 400] | [ 40, 80] |
| 63 | Clitocybe | odora | [.050, .070] | [.030, .050] | [ 250, 700] | [ 300, 700] | [ 50, 120] |
| 64 | Clitocybe | nebularis | [.055, .085] | [.035, .045] | [ 500, 2500] | [ 500, 1500] | [ 150, 400] |
| 65 | Clitocybe | glacialis | [.055, .070] | [.035, .045] | [ 200, 600] | [ 200, 600] | [ 100, 150] |
| 66 | Clitocybe | P. flaccida | [.040, .045] | [.034, .034] | [ 200, 900] | [ 300, 700] | [ 40, 60] |
| 67 | Clitocybe | rivulosa | [.040, .050] | [.020, .030] | [ 200, 400] | [ 200, 400] | [ 40, 80] |
| 68 | Clitocybe | deceptiva | [.060, .075] | [.035, .040] | [ 120, 500] | [ 150, 400] | [ 30, 50] |
| 69 | Clitocybe | nuda | [.060, .080] | [.040, .050] | [ 400, 1400] | [ 300, 650] | [ 100, 250] |
| 70 | Clitocybe | albirhiza | [.050, .060] | [.025, .035] | [ 200, 900] | [ 200, 600] | [ 50, 120] |
| 71 | Coprinus | calyptratus | [.000, .190] | [.095, .110] | [ 400, 700] | [ 600, 1000] | [ 50, 70] |
| 72 | Coprinus | comatus | [.120, .160] | [.070, .080] | [ 500, 1400] | [ 800, 2000] | [ 100, 150] |
| 73 | Coprinus | sterquilinus | [.175, .225] | [.110, .135] | [ 300, 600] | [ 400, 900] | [ 60, 100] |
| 74 | Hygrocybe | flavifolia | [.070, .090] | [.040, .055] | [ 150, 350] | [ 200, 400] | [ 40, 60] |
| 75 | Hygrocybe | coccinea | [.070, .095] | [.040, .050] | [ 250, 500] | [ 250, 550] | [ 50, 100] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 76 | Hygrocybe | singeri | [.095, .115] | [.050, .065] | [ 200, 500] | [ 400, 1400] | [ 50, 100] |
| 77 | Hygrocybe | miniata | [.060, .090] | [.040, .060] | [ 150, 350] | [ 200, 400] | [ 30, 50] |
| 78 | Hygrocybe | flavescens | [.075, .090] | [.040, .050] | [ 200, 600] | [ 350, 700] | [ 70, 120] |
| 79 | Hygrocybe | conica | [.090, .130] | [.050, .065] | [ 200, 900] | [ 500, 1000] | [ 50, 100] |
| 80 | Hygrocybe | G. psittacinus | [.080, .100] | [.050, .060] | [ 150, 400] | [ 400, 900] | [ 30, 50] |
| 81 | Hygrocybe | punicea | [.080, .110] | [.050, .060] | [ 400, 1200] | [ 300, 1400] | [ 50, 200] |
| 82 | Hypholoma | capnoides | [.060, .075] | [.035, .050] | [ 250, 600] | [ 500, 700] | [ 40, 100] |
| 83 | Hypholoma | fasciculare | [.065, .080] | [.035, .045] | [ 200, 700] | [ 200, 900] | [ 40, 150] |
| 84 | Inocybe | P. sororium | [.100, .140] | [.060, .080] | [ 250, 650] | [ 400, 1000] | [ 30, 80] |
| 85 | Inocybe | pudica | [.075, .100] | [.045, .050] | [ 200, 400] | [ 200, 400] | [ 50, 80] |
| 86 | Inocybe | griseolilacina | [.080, .105] | [.045, .060] | [ 150, 300] | [ 200, 400] | [ 40, 70] |
| 87 | Inocybe | pallidicremea | [.075, .105] | [.045, .050] | [ 120, 300] | [ 250, 400] | [ 30, 40] |
| 88 | Inocybe | insinuata | [.075, .090] | [.045, .050] | [ 200, 400] | [ 250, 500] | [ 30, 60] |
| 89 | Inocybe | fraudans | [.090, .115] | [.055, .070] | [ 250, 650] | [ 400, 800] | [ 50, 170] |
| 90 | Inocybe | citrifolia | [.114, .114] | [.053, .057] | [ 200, 400] | [ 30, 70] | [ 30, 70] |
| 91 | Inocybe | brunnescens | [.080, .105] | [.050, .060] | [ 300, 700] | [ 400, 900] | [ 100, 150] |
| 92 | Inocybe | I. adaequatum | [.090, .120] | [.060, .075] | [ 100, 100] | [ 400, 800] | [ 100, 200] |
| 93 | Inocybe | corydalina | [.075, .110] | [.050, .060] | [ 400, 600] | [ 400, 900] | [ 100, 200] |
| 94 | Laccaria | laccata | [.070, .090] | [.070, .085] | [ 150, 500] | [ 300, 600] | [ 20, 60] |
| 95 | Laccaria | fraterna | [.080, .105] | [.075, .090] | [ 150, 400] | [ 150, 500] | [ 20, 50] |
| 96 | Laccaria | amethysteo-occidentalis | [.075, .105] | [.070, .160] | [ 100, 650] | [ 200, 1200] | [ 30, 120] |
| 97 | Lactarius | rubrilacteus | [.070, .090] | [.060, .075] | [ 500, 1200] | [ 200, 500] | [ 100, 250] |
| 98 | Lactarius | deliciosus | [.075, .110] | [.060, .075] | [ 500, 1300] | [ 300, 600] | [ 150, 250] |
| 99 | Lactarius | argillaceifolius | [.070, .090] | [.070, .090] | [ 900, 2100] | [ 700, 1400] | [ 200, 500] |
| 100 | Lactarius | alnicola | [.070, .100] | [.060, .080] | [ 600, 1300] | [ 200, 500] | [ 150, 250] |
| 101 | Lactarius | xanthogalactus | [.070, .080] | [.060, .065] | [ 400, 1100] | [ 300, 600] | [ 100, 200] |
| 102 | Lactarius | pubescens | [.065, .075] | [.045, .045] | [ 300, 700] | [ 250, 400] | [ 150, 200] |
| 103 | Lactarius | rubidus | [.065, .075] | [.065, .075] | [ 150, 450] | [ 200, 500] | [ 40, 100] |
| 104 | Lactarius | pallescens | [.090, .100] | [.070, .080] | [ 500, 1100] | [ 400, 800] | [ 120, 200] |
| 105 | Leccinum | scabrum | [.140, .180] | [.050, .060] | [ 500, 1400] | [ 800, 1400] | [ 200, 400] |
| 106 | Leccinum | manzanitae | [.130, .175] | [.040, .050] | [ 500, 1800] | [ 900, 1700] | [ 200, 400] |
| 107 | Marasmius | curreyi | [.090, .120] | [.040, .050] | [ 40, 80] | [ 150, 300] | [ 10, 10] |
| 108 | Marasmius | calhouniae | [.090, .100] | [.035, .045] | [ 100, 300] | [ 150, 400] | [ 20, 50] |
| 109 | Marasmius | armeniacus | [.085, .105] | [.030, .040] | [ 40, 120] | [ 100, 300] | [ 5, 5] |
| 110 | Marasmius | plicatulus | [.110, .145] | [.050, .065] | [ 100, 400] | [ 500, 1100] | [ 15, 35] |
| 111 | Marasmius | M. copelandii | [.130, .180] | [.025, .035] | [ 50, 200] | [ 300, 800] | [ 10, 30] |
| 112 | Marasmius | oreades | [.070, .085] | [.040, .055] | [ 150, 400] | [ 200, 600] | [ 20, 50] |
| 113 | Marasmius | C. quercophila | [.075, .090] | [.030, .045] | [ 20, 50] | [ 100, 250] | [ 10, 10] |
| 114 | Mycena | maculata | [.075, .095] | [.050, .055] | [ 150, 400] | [ 200, 900] | [ 15, 40] |
| 115 | Mycena | haematopus | [.075, .090] | [.045, .055] | [ 100, 300] | [ 250, 700] | [ 20, 30] |
| 116 | Mycena | nivicola | [.085, .115] | [.050, .060] | [ 150, 300] | [ 250, 900] | [ 20, 30] |
| 117 | Mycena | overholtsii | [.055, .070] | [.030, .035] | [ 200, 600] | [ 1500, 1500] | [ 150, 150] |
| 118 | Mycena | pura | [.060, .085] | [.030, .040] | [ 150, 450] | [ 200, 600] | [ 20, 70] |
| 119 | Mycena | purpureofusca | [.070, .100] | [.050, .060] | [ 70, 300] | [ 300, 700] | [ 10, 40] |
| 120 | Mycena | galericulata | [.085, .105] | [.060, .075] | [ 200, 500] | [ 300, 1400] | [ 20, 50] |
| 121 | Mycena | capillaripes | [.080, .110] | [.040, .065] | [ 100, 200] | [ 400, 600] | [ 10, 20] |
| 122 | Mycena | californiensis | [.075, .090] | [.040, .045] | [ 70, 200] | [ 200, 700] | [ 10, 20] |
| 123 | Mycena | oregonensis | [.065, .085] | [.030, .035] | [ 20, 80] | [ 100, 250] | [ 100, 100] |
| 124 | Mycena | amicta | [.080, .095] | [.040, .050] | [ 50, 150] | [ 300, 700] | [ 10, 30] |
| 125 | Mycena | tenerrima | [.080, .105] | [.040, .060] | [ 20, 40] | [ 40, 100] | [ 10, 10] |
| 126 | Mycena | acicula | [.085, .115] | [.030, .040] | [ 20, 80] | [ 100, 500] | [ 5, 5] |
| 127 | Mycena | aurantiomarginata | [.075, .090] | [.040, .055] | [ 100, 200] | [ 250, 700] | [ 10, 20] |
| 128 | Pholiota | velaglutinosa | [.065, .075] | [.035, .045] | [ 300, 500] | [ 350, 600] | [ 40, 80] |
| 129 | Pholiota | terrestris | [.040, .065] | [.035, .045] | [ 200, 800] | [ 350, 900] | [ 50, 100] |
| 130 | Pholiota | squarrosa | [.060, .080] | [.040, .050] | [ 300, 1200] | [ 400, 1200] | [ 150, 150] |
| 131 | Pholiota | spumosa | [.060, .095] | [.040, .055] | [ 200, 600] | [ 200, 600] | [ 30, 80] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 132 | Pholiota | flammans | [.040, .050] | [.025, .030] | [ 400, 800] | [ 500, 1000] | [ 100, 100] |
| 133 | Psilocybe | D. subviscida | [.065, .075] | [.040, .045] | [ 100, 200] | [ 150, 400] | [ 10, 30] |
| 134 | Psilocybe | D. coprophila | [.110, .140] | [.070, .090] | [ 100, 250] | [ 150, 500] | [ 10, 30] |
| 135 | Psilocybe | cyanescens | [.090, .120] | [.060, .080] | [ 200, 450] | [ 300, 600] | [ 30, 60] |
| 136 | Psilocybe | D. montana | [.070, .095] | [.045, .060] | [ 70, 150] | [ 100, 300] | [ 10, 20] |
| 137 | Russula | brevipes | [.080, .105] | [.065, .090] | [ 600, 1200] | [ 400, 600] | [ 200, 300] |
| 138 | Russula | cantharellicola | [.075, .100] | [.065, .070] | [ 700, 1200] | [ 200, 750] | [ 250, 350] |
| 139 | Russula | olivacea | [.085, .105] | [.075, .090] | [ 800, 1600] | [ 800, 1300] | [ 200, 350] |
| 140 | Russula | silvicola | [.067, .105] | [.057, .086] | [ 400, 900] | [ 400, 1000] | [ 100, 300] |
| 141 | Russula | dissimulans | [.060, .110] | [.060, .090] | [ 500, 2000] | [ 300, 800] | [ 100, 400] |
| 142 | Russula | densifolia | [.070, .095] | [.055, .070] | [ 700, 1300] | [ 300, 750] | [ 200, 400] |
| 143 | Russula | cyanoxantha | [.065, .095] | [.055, .070] | [ 400, 1500] | [ 500, 1300] | [ 100, 300] |
| 144 | Russula | basifurcata | [.070, .095] | [.065, .080] | [ 400, 700] | [ 300, 700] | [ 100, 300] |
| 145 | Russula | fragrantissima | [.060, .090] | [.060, .080] | [ 750, 2000] | [ 700, 1500] | [ 150, 600] |
| 146 | Russula | aeruginea | [.060, .085] | [.050, .070] | [ 500, 900] | [ 400, 600] | [ 100, 200] |
| 147 | Russula | sanguinea | [.078, .095] | [.065, .085] | [ 400, 1000] | [ 500, 1000] | [ 100, 250] |
| 148 | Russula | cerolens | [.070, .080] | [.050, .060] | [ 400, 1100] | [ 300, 700] | [ 100, 250] |
| 149 | Strobilurus | albipilatus | [.040, .065] | [.030, .035] | [ 150, 300] | [ 150, 600] | [ 10, 20] |
| 150 | Strobilurus | diminutivus | [.045, .050] | [.025, .030] | [ 12, 40] | [ 10, 30] | [ 5, 5] |
| 151 | Strobilurus | trullisatus | [.035, .060] | [.020, .030] | [ 40, 170] | [ 150, 450] | [ 10, 20] |
| 152 | Stropharia | P. semiglobata | [.150, .200] | [.075, .100] | [ 200, 400] | [ 300, 800] | [ 20, 50] |
| 153 | Stropharia | L. riparius | [.120, .150] | [.060, .075] | [ 200, 600] | [ 500, 1300] | [ 500, 1300] |
| 154 | Stropharia | ambigua | [.100, .150] | [.060, .090] | [ 400, 1400] | [ 700, 1700] | [ 100, 200] |
| 155 | Stropharia | coronilla | [.070, .085] | [.045, .055] | [ 200, 500] | [ 150, 450] | [ 40, 70] |
| 156 | Suillus | volcanalis | [.070, .100] | [.030, .035] | [ 800, 1500] | [ 400, 600] | [ 200, 450] |
| 157 | Suillus | umbonatus | [.080, .100] | [.035, .040] | [ 200, 800] | [ 200, 500] | [ 50, 100] |
| 158 | Suillus | tomentosus | [.080, .110] | [.030, .040] | [ 500, 1100] | [ 500, 900] | [ 150, 300] |
| 159 | Suillus | megaporinus | [.070, .100] | [.035, .040] | [ 200, 700] | [ 100, 200] | [ 50, 100] |
| 160 | Suillus | lakei | [.075, .100] | [.030, .040] | [ 400, 1200] | [ 300, 700] | [ 150, 250] |
| 161 | Suillus | fuscotomentosus | [.095, .115] | [.035, .045] | [ 400, 1500] | [ 400, 700] | [ 200, 350] |
| 162 | Suillus | brevipes | [.075, .100] | [.030, .045] | [ 350, 1000] | [ 150, 600] | [ 150, 350] |
| 163 | Suillus | pungens | [.090, .100] | [.030, .035] | [ 500, 1300] | [ 300, 800] | [ 150, 200] |
| 164 | Suillus | caerulescens | [.065, .095] | [.030, .040] | [ 600, 1300] | [ 200, 700] | [ 100, 350] |
| 165 | Tricholoma | vernaticum | [.085, .110] | [.040, .060] | [ 400, 1400] | [ 500, 1300] | [ 200, 350] |
| 166 | Tricholoma | sejunctum | [.050, .080] | [.035, .055] | [ 400, 900] | [ 300, 1000] | [ 100, 150] |
| 167 | Tricholoma | saponaceum | [.050, .065] | [.035, .045] | [ 400, 900] | [ 450, 800] | [ 150, 200] |
| 168 | Tricholoma | muricatum | [.045, .060] | [.030, .035] | [ 500, 1200] | [ 300, 600] | [ 100, 350] |
| 169 | Tricholoma | moseri | [.065, .100] | [.035, .050] | [ 200, 450] | [ 200, 500] | [ 50, 100] |
| 170 | Tricholoma | imbricatum | [.055, .070] | [.040, .050] | [ 600, 1500] | [ 500, 1000] | [ 200, 350] |
| 171 | Tricholoma | fracticum | [.055, .075] | [.040, .055] | [ 500, 1000] | [ 300, 800] | [ 150, 250] |
| 172 | Tricholoma | dryophilum | [.050, .060] | [.040, .043] | [ 500, 1500] | [ 600, 1300] | [ 100, 450] |
| 173 | Tricholoma | atroviolaceum | [.075, .090] | [.050, .060] | [ 350, 900] | [ 400, 800] | [ 150, 300] |
| 174 | Tricholoma | murrillianum | [.050, .070] | [.045, .055] | [ 500, 2500] | [ 400, 1500] | [ 100, 600] |
| 175 | Tricholoma | equestre | [.060, .075] | [.035, .050] | [ 500, 1300] | [ 400, 800] | [ 150, 300] |
| 176 | Tricholoma | griseoviolaceum | [.050, .070] | [.035, .050] | [ 400, 1100] | [ 600, 1400] | [ 100, 200] |
| 177 | Tricholoma | myomyces | [.050, .075] | [.035, .045] | [ 150, 500] | [ 250, 500] | [ 50, 100] |
| 178 | Tylopilus | P. porphyrosporus | [.145, .170] | [.060, .075] | [ 700, 1200] | [ 700, 1500] | [ 150, 300] |
| 179 | Tylopilus | indecisus | [.090, .120] | [.030, .040] | [ 600, 1300] | [ 600, 1200] | [ 300, 450] |

# C Brazilian Science Production dataset

| Index | GRANDE AREA PREDOM. | AREA PREDOM. | BIBL TRABALHO | BIBL | ORIE CONC | DEMAIS | OUTRAS | TECN |
|---|---|---|---|---|---|---|---|---|
| 0 | Ciências Agrárias | Agronomia | [0. 5.] | [0.25 27.] | [0. 7.75] | [0. 20.25] | [0. 20.75] | [0. 10.5] |
| 1 | Ciências Agrárias | Ciência e Tecnologia de Alimentos | [0. 2.75] | [0.5 19.] | [0. 7.75] | [0.5 16.] | [0.5 16.] | [0. 8.25] |
| 2 | Ciências Agrárias | Engenharia Agrícola | [0. 5.25] | [0.75 19.5] | [0. 5.] | [0.5 15.375] | [0.5 15.375] | [0. 8.5] |
| 3 | Ciências Agrárias | Medicina Veterinária | [0. 2.25] | [0.5 23.25] | [0. 6.75] | [0.25 18.75] | [0.25 18.75] | [0. 9.5] |
| 4 | Ciências Agrárias | Recursos Florestais e Engenharia Florestal | [0. 4.75] | [0.5 18.5] | [0. 5.25] | [0. 15.25] | [0. 15.25] | [0. 7.] |
| 5 | Ciências Agrárias | Recursos Pesqueiros e Engenharia de Pesca | [0. 1.375] | [0.5 14.875] | [0. 5.625] | [0.875 15.875] | [1. 15.875] | [0. 5.75] |
| 6 | Ciências Agrárias | Zootecnia | [0. 4.25] | [0.5 28.5] | [0. 6.75] | [0.5 20.] | [0.5 19.25] | [0. 8.75] |
| 7 | Ciências Biológicas | Biofísica | [0. 3.] | [1.25 15.25] | [0. 4.] | [0.667 11.] | [0.667 11.125] | [0. 6.375] |
| 8 | Ciências Biológicas | Biologia Geral | [0. 1.5] | [1. 18.] | [0. 4.5] | [0.75 13.25] | [0.75 13.25] | [0. 8.5] |
| 9 | Ciências Biológicas | Bioquímica | [0. 1.75] | [0.75 18.5] | [0. 5.25] | [0.333 14.75] | [0.333 14.75] | [0. 6.] |
| 10 | Ciências Biológicas | Botânica | [0. 1.25] | [0.75 17.75] | [0. 5.] | [0.75 14.75] | [0.75 14.75] | [0. 7.] |
| 11 | Ciências Biológicas | Ecologia | [0. 2.25] | [0.333 16.] | [0. 6.25] | [0.25 15.] | [0.25 15.] | [0. 8.25] |
| 12 | Ciências Biológicas | Farmacologia | [0. 1.] | [0.75 18.75] | [0. 5.75] | [0.417 15.875] | [0.417 15.875] | [0. 7.375] |
| 13 | Ciências Biológicas | Fisiologia | [0. 1.25] | [1. 19.5] | [0. 4.25] | [0.75 13.25] | [0.75 13.5] | [0. 5.5] |
| 14 | Ciências Biológicas | Genética | [0. 2.] | [0.875 23.75] | [0. 5.75] | [0.5 17.625] | [0.5 17.625] | [0. 9.5] |
| 15 | Ciências Biológicas | Imunologia | [0. 1.] | [1. 17.5] | [0. 4.5] | [0.5 17.25] | [0.5 17.25] | [0. 8.5] |
| 16 | Ciências Biológicas | Microbiologia | [0. 1.75] | [1. 17.25] | [0. 5.75] | [0.5 15.25] | [0.5 15.25] | [0. 6.] |
| 17 | Ciências Biológicas | Morfologia | [0. 1.25] | [1. 16.25] | [0. 4.75] | [1. 16.25] | [1. 17.25] | [0. 7.25] |
| 18 | Ciências Biológicas | Parasitologia | [0. 1.] | [1. 16.5] | [0. 4.75] | [0.333 13.75] | [0.333 13.75] | [0. 6.25] |
| 19 | Ciências Biológicas | Zoologia | [0. 1.] | [1. 17.875] | [0. 5.125] | [0.5 13.625] | [0.5 13.625] | [0. 8.] |
| 20 | Ciências Exatas e da Terra | Astronomia | [0. 1.625] | [1. 12.375] | [0. 2.25] | [0. 6.625] | [0. 6.625] | [0. 3.75] |
| 21 | Ciências Exatas e da Terra | Ciência da Computação | [0. 5.] | [0. 10.] | [0. 6.125] | [0. 13.5] | [0. 13.5] | [0. 6.25] |
| 22 | Ciências Exatas e da Terra | Física | [0. 3.25] | [0.417 14.875] | [0. 3.75] | [0. 9.25] | [0. 9.375] | [0. 4.375] |
| 23 | Ciências Exatas e da Terra | Geociências | [0. 4.875] | [0. 15.25] | [0. 6.5] | [0. 15.5] | [0. 15.75] | [0. 7.75] |

| Index | GRANDE AREA PREDOM. | AREA PREDOM. | BIBL TRABALHO | BIBL | ORIE CONC | DEMAIS | OUTRAS | TECN |
|---|---|---|---|---|---|---|---|---|
| 24 | Ciências Exatas e da Terra | Matemática | [0. 2.625] | [0. 7.875] | [0. 4.] | [0.25 10.5] | [0.25 10.5] | [0. 5.25] |
| 25 | Ciências Exatas e da Terra | Oceanografia | [0. 3.] | [0.5 17.875] | [0. 5.875] | [0. 18.] | [0. 18.] | [0. 7.] |
| 26 | Ciências Exatas e da Terra | Probabilidade e Estatística | [0. 2.75] | [0.5 14.25] | [0. 4.] | [0.25 11.] | [0.25 11.] | [0. 5.25] |
| 27 | Ciências Exatas e da Terra | Química | [0. 3.25] | [0.417 19.875] | [0. 6.] | [0.125 15.5] | [0.125 15.5] | [0. 6.875] |
| 28 | Ciências Humanas | Antropologia | [0. 2.25] | [0. 9.75] | [0. 6.] | [0.5 16.] | [0.75 17.] | [0. 10.] |
| 29 | Ciências Humanas | Arqueologia | [0. 2.25] | [0.25 9.25] | [0. 4.25] | [0.75 15.5] | [0.75 15.5] | [0. 8.] |
| 30 | Ciências Humanas | Ciência Política | [0. 2.75] | [0.583 9.5] | [0. 7.] | [1.125 17.875] | [1.125 17.75] | [0. 9.125] |
| 31 | Ciências Humanas | Educação | [0. 5.5] | [0. 16.75] | [0. 10.] | [0. 23.75] | [0. 23.75] | [0. 16.25] |
| 32 | Ciências Humanas | Filosofia | [0. 1.5] | [0.25 11.25] | [0. 6.] | [0.333 16.75] | [0.5 17.5] | [0. 10.75] |
| 33 | Ciências Humanas | Geografia | [0. 4.5] | [0.333 13.5] | [0. 7.125] | [1. 17.125] | [1.125 17.75] | [0. 8.75] |
| 34 | Ciências Humanas | História | [0. 2.75] | [0.25 12.25] | [0. 7.25] | [1. 19.5] | [1. 19.5] | [0. 11.25] |
| 35 | Ciências Humanas | Psicologia | [0. 2.625] | [0.25 17.625] | [0. 8.125] | [0.667 21.625] | [0.667 21.75] | [0. 13.5] |
| 36 | Ciências Humanas | Sociologia | [0. 2.75] | [0.25 12.75] | [0. 6.25] | [0.5 19.75] | [0.5 20.] | [0. 11.5] |
| 37 | Ciências Humanas | Teologia | [0. 1.875] | [0.125 13.75] | [0. 7.75] | [0.375 19.] | [0.5 19.] | [0. 12.125] |
| 38 | Ciências Sociais Aplicadas | Administração | [0. 6.25] | [0.25 13.25] | [0. 10.75] | [0.5 19.75] | [0.5 20.] | [0. 10.] |
| 39 | Ciências Sociais Aplicadas | Arquitetura e Urbanismo | [0. 4.] | [0. 10.5] | [0. 6.5] | [0.333 18.] | [0.333 18.5] | [0. 8.75] |
| 40 | Ciências Sociais Aplicadas | Ciência da Informação | [0. 3.75] | [0. 11.125] | [0. 6.375] | [0.25 19.25] | [0.292 19.25] | [0. 8.125] |
| 41 | Ciências Sociais Aplicadas | Comunicação | [0. 3.25] | [0. 11.75] | [0. 8.25] | [0.5 20.5] | [0.708 20.75] | [0. 10.375] |
| 42 | Ciências Sociais Aplicadas | Demografia | [0. 3.125] | [1. 9.875] | [0.125 4.375] | [0.75 16.625] | [0.75 16.75] | [0.375 8.375] |
| 43 | Ciências Sociais Aplicadas | Desenho Industrial | [0. 4.75] | [0.25 9.] | [0. 8.875] | [0.5 19.75] | [0.583 20.625] | [0. 10.875] |
| 44 | Ciências Sociais Aplicadas | Direito | [0. 1.5] | [0. 11.75] | [0. 12.125] | [0.5 27.25] | [0.583 27.25] | [0. 9.125] |
| 45 | Ciências Sociais Aplicadas | Economia | [0. 4.5] | [0. 12.5] | [0. 6.75] | [0. 18.5] | [0. 18.5] | [0. 7.5] |
| 46 | Ciências Sociais Aplicadas | Economia Doméstica | [0. 4.25] | [1. 14.] | [0. 7.5] | [0. 23.5] | [0. 23.5] | [0. 18.5] |
| 47 | Ciências Sociais Aplicadas | Museologia | [0. 3.625] | [0.625 11.] | [0. 4.875] | [1.5 20.75] | [1.5 18.125] | [0.375 13.875] |

| Index | GRANDE AREA PREDOM. | AREA PREDOM. | BIBL TRABALHO | BIBL | ORIE CONC | DEMAIS | OUTRAS | TECN |
|---|---|---|---|---|---|---|---|---|
| 48 | Ciências Sociais Aplicadas | Planejamento Urbano e Regional | [0. 3.5] | [0.25 10.5] | [0. 6.5] | [0.5 17.75] | [0.5 18.] | [0. 7.5] |
| 49 | Ciências Sociais Aplicadas | Serviço Social | [0. 3.5] | [0. 11.375] | [0. 8.] | [0.583 22.] | [0.667 22.125] | [0. 13.] |
| 50 | Ciências Sociais Aplicadas | Turismo | [0. 2.875] | [0.417 9.625] | [0. 6.] | [1.75 16.] | [1.75 16.] | [0.292 6.375] |
| 51 | Ciências da Saúde | Educação Física | [0. 2.75] | [0.25 15.75] | [0. 7.75] | [0.75 17.75] | [0.75 17.75] | [0. 9.25] |
| 52 | Ciências da Saúde | Enfermagem | [0. 1.5] | [0. 16.] | [0. 7.75] | [1. 22.75] | [1. 22.75] | [0. 12.] |
| 53 | Ciências da Saúde | Farmácia | [0. 1.25] | [0.25 17.875] | [0. 6.5] | [0.417 16.375] | [0.417 16.5] | [0. 8.75] |
| 54 | Ciências da Saúde | Fisioterapia e Terapia Ocupacional | [0. 1.5] | [0.25 11.5] | [0. 6.625] | [1.125 14.25] | [1.125 14.25] | [0. 9.125] |
| 55 | Ciências da Saúde | Fonoaudiologia | [0. 1.375] | [0.625 15.25] | [0. 5.625] | [1.875 20.25] | [2.25 20.625] | [0. 12.5] |
| 56 | Ciências da Saúde | Medicina | [0. 2.] | [0.25 24.25] | [0. 6.5] | [0. 22.75] | [0. 22.75] | [0. 12.75] |
| 57 | Ciências da Saúde | Nutrição | [0. 1.75] | [0.5 17.] | [0. 7.25] | [0.75 17.] | [0.75 17.] | [0. 8.] |
| 58 | Ciências da Saúde | Odontologia | [0. 1.875] | [0.5 22.25] | [0. 6.375] | [0.5 21.125] | [0.5 21.125] | [0. 13.125] |
| 59 | Ciências da Saúde | Saúde Coletiva | [0. 1.625] | [0. 16.875] | [0. 7.] | [0.25 19.625] | [0.292 19.75] | [0. 11.125] |
| 60 | Engenharias | Engenharia Aeroespacial | [0.292 7.25] | [0.458 11.625] | [0. 3.875] | [0.417 10.625] | [0.417 10.625] | [0. 4.5] |
| 61 | Engenharias | Engenharia Biomédica | [0. 5.5] | [1. 13.] | [0. 5.] | [0. 12.5] | [0. 12.5] | [0. 4.5] |
| 62 | Engenharias | Engenharia Civil | [0. 8.25] | [0. 15.] | [0. 7.] | [0. 15.5] | [0. 15.5] | [0. 7.75] |
| 63 | Engenharias | Engenharia Elétrica | [0. 6.] | [0. 9.25] | [0. 5.] | [0. 11.5] | [0. 11.5] | [0. 4.25] |
| 64 | Engenharias | Engenharia Mecânica | [0. 7.75] | [0. 13.25] | [0. 5.875] | [0. 13.625] | [0. 13.625] | [0. 6.5] |
| 65 | Engenharias | Engenharia Naval e Oceânica | [0.333 6.625] | [0.333 9.625] | [0. 2.25] | [0.375 9.5] | [0.375 9.5] | [0. 8.375] |
| 66 | Engenharias | Engenharia Nuclear | [0. 8.375] | [0.375 31.75] | [0. 4.75] | [0. 15.5] | [0. 15.5] | [0. 7.] |
| 67 | Engenharias | Engenharia Química | [0. 6.75] | [0.667 17.5] | [0. 6.25] | [0. 14.25] | [0. 14.25] | [0. 4.75] |
| 68 | Engenharias | Engenharia Sanitária | [0. 7.25] | [0.667 14.25] | [0. 6.5] | [0.25 16.5] | [0.25 16.5] | [0. 7.75] |
| 69 | Engenharias | Engenharia de Materiais e Metalúrgica | [0. 6.25] | [0.5 15.75] | [0. 4.75] | [0. 10.5] | [0. 10.5] | [0. 5.5] |
| 70 | Engenharias | Engenharia de Minas | [0. 5.25] | [0.667 10.5] | [0. 3.75] | [0. 7.] | [0. 7.] | [0. 5.75] |
| 71 | Engenharias | Engenharia de Produção | [0. 6.75] | [0.25 14.25] | [0. 7.75] | [0. 16.75] | [0. 16.75] | [0. 7.25] |
| 72 | Engenharias | Engenharia de Transportes | [0. 5.625] | [0.333 8.75] | [0. 3.875] | [0.292 10.625] | [0.292 10.625] | [0. 7.25] |

| Index | GRANDE AREA PREDOM. | AREA PREDOM. | BIBL TRABALHO | BIBL | ORIE CONC | DEMAIS | OUTRAS | TECN |
|---|---|---|---|---|---|---|---|---|
| 73 | Lingüística, Letras e Artes | Artes | [0. 2.5] | [0. 8.75] | [0. 6.25] | [0.5 16.75] | [1. 20.75] | [0. 9.5] |
| 74 | Lingüística, Letras e Artes | Letras | [0. 2.75] | [0.25 13.25] | [0. 7.] | [0.75 17.25] | [0.75 18.] | [0. 11.5] |
| 75 | Lingüística, Letras e Artes | Lingüística | [0. 2.5] | [0.25 12.875] | [0. 7.375] | [1. 20.625] | [1. 21.] | [0. 14.] |

# D Choice of *beta*

## D.1 Table of best beta values:

| Name | Fungi_3 | Fungi_4 | Fungi_5 | BSP_3 | BSP_4 | BSP_5 |
|---|---|---|---|---|---|---|
| BIKMn | 1.0 \| 1.0 | 1.0 \| 1.0 | 1.0 \| 1.0 | 1.0 \| 1.0 | 1.0 \| 1.0 | 1.0 \| 1.0 |
| BIKMc | 1.1 \| 1.1 | 1.1 \| 2.5 | 1.2 \| 2.2 | 1.7 \| 1.8 | 2.2 \| 1.4 | 1.1 \| 1.8 |
| BIKMo | 1.0 \| 1.0 | 1.0 \| 2.6 | 2.5 \| 2.8 | 1.4 \| 1.0 | 2.5 \| 2.1 | 1.0 \| 1.1 |
| BIKMcs | 1.5 \| 2.2 | 2.6 \| 2.1 | 1.7 \| 2.1 | 1.5 \| 1.9 | 1.9 \| 1.3 | 1.2 \| 2.2 |
| BIKMos | 2.1 \| 1.0 | 1.0 \| 1.1 | 1.9 \| 1.0 | 1.8 \| 2.4 | 1.0 \| 2.4 | 1.7 \| 1.0 |
| BKMn | 1.0 \| 1.0 | 1.0 \| 1.0 | 1.0 \| 1.0 | 1.0 \| 1.0 | 1.0 \| 1.0 | 1.0 \| 1.0 |
| BKMc | 1.2 \| 1.9 | 1.1 \| 2.4 | 1.3 \| 2.7 | 1.1 \| 1.1 | 1.9 \| 2.3 | 1.3 \| 2.7 |
| BKMcs | 1.5 \| 2.4 | 2.6 \| 2.8 | 1.7 \| 2.7 | 1.5 \| 1.7 | 1.1 \| 2.4 | 1.1 \| 2.4 |

| Name | Fungi_3 | Fungi_4 | Fungi_5 | BSP_3 | BSP_4 | BSP_5 |
|---|---|---|---|---|---|---|
| BnIKMc | 1.0 \| 1.5 | 1.0 \| 2.2 | 1.0 \| 2.3 | 1.0 \| 1.8 | 1.0 \| 1.7 | 1.0 \| 1.4 |
| BnIKMo | 1.0 \| 1.0 | 1.0 \| 1.0 | 1.0 \| 1.6 | 1.0 \| 1.0 | 1.0 \| 1.7 | 1.0 \| 1.0 |
| BnIKMcs | 1.0 \| 1.1 | 1.0 \| 2.8 | 1.0 \| 1.2 | 1.0 \| 2.3 | 1.0 \| 1.5 | 1.0 \| 2.3 |
| BnIKMos | 1.0 \| 1.0 | 1.0 \| 1.0 | 1.0 \| 1.0 | 1.0 \| 2.4 | 1.0 \| 2.5 | 1.0 \| 1.0 |
| BcIKMn | 1.2 \| 1.0 | 1.2 \| 1.0 | 1.6 \| 1.0 | 1.5 \| 1.0 | 1.9 \| 1.0 | 1.1 \| 1.0 |
| BcIKMo | 1.2 \| 1.0 | 1.1 \| 1.0 | 1.3 \| 1.8 | 1.5 \| 1.0 | 1.9 \| 1.0 | 1.2 \| 1.2 |
| BcIKMos | 1.5 \| 1.0 | 2.6 \| 1.1 | 1.3 \| 2.7 | 1.5 \| 2.3 | 1.9 \| 1.0 | 1.1 \| 1.0 |
| BoIKMn | 2.6 \| 1.0 | 1.0 \| 1.0 | 2.6 \| 1.0 | 1.8 \| 1.0 | 1.0 \| 1.0 | 1.4 \| 1.0 |
| BoIKMc | 1.0 \| 1.1 | 1.0 \| 1.1 | 2.5 \| 1.2 | 1.4 \| 1.8 | 1.0 \| 1.7 | 2.0 \| 2.1 |
| BoIKMcs | 2.1 \| 2.4 | 1.0 \| 2.1 | 1.0 \| 1.4 | 1.4 \| 2.0 | 1.0 \| 2.7 | 1.0 \| 2.2 |

| | | | | | | |
|---|---|---|---|---|---|---|
| BnKMc | 1.0 \| 1.2 | 1.0 \| 1.8 | 1.0 \| 2.0 | 1.0 \| 1.2 | 1.0 \| 2.3 | 1.0 \| 2.5 |
| BnKMcs | 1.0 \| 1.9 | 1.0 \| 1.7 | 1.0 \| 2.4 | 1.0 \| 1.7 | 1.0 \| 2.0 | 1.0 \| 2.7 |
| BcKMn | 1.2 \| 1.0 | 1.4 \| 1.0 | 1.8 \| 1.0 | 1.5 \| 1.0 | 1.1 \| 1.0 | 1.3 \| 1.0 |
| BoKMn | 1.7 \| 1.0 | 1.0 \| 1.0 | 1.0 \| 1.0 | 1.0 \| 1.0 | 1.0 \| 1.0 | 1.0 \| 1.0 |
| BoKMc | 2.1 \| 2.3 | 1.0 \| 1.2 | 1.0 \| 2.0 | 1.0 \| 1.2 | 2.5 \| 2.3 | 1.0 \| 2.3 |
| BnKMcs | 2.1 \| 2.4 | 1.0 \| 2.8 | 1.0 \| 2.2 | 1.0 \| 1.7 | 1.0 \| 2.5 | 2.6 \| 2.2 |

# References


[1] Bock, H. H. and Diday, E. (Eds.). (2012). Analysis of symbolic data: exploratory methods for extracting statistical information from complex data. Springer Science & Business Media.

[2] Brito, P., Silva, A., and Dias, J. G. (2015). Probabilistic clustering of interval data. Intelligent Data Analysis, 19(2), 293-313.

[3] De Carvalho, F. D. A., De Souza, R. M., Chavent, M., and Lechevallier, Y. (2006). Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. Pattern Recognition Letters, 27(3), 167-179.

[4] Chiang, M. M. T., and Mirkin, B. (2010). Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads. Journal of Classification, 27, 3-40.

[5] Cordeiro de Amorim R., and Mirkin, B. Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering, Pattern Recognition, vol. 45, no. 3, pp. 1061–1075, Mar. 2012, doi: https://doi.org/10.1016/j.patcog.2011.08.012.

[6] Cover, T., and Thomas, J. (1991) Elements of Information Theory, J. Wiley and Sons.

[7] Fahim, A. (2021). K and starting means for k-means algorithm. Journal of Computational Science, 55, 101445.

[8] Ferreira, M.R.P and de Carvalho, F.A.T. (2014) Kernel-based hard clustering methods in the feature space with automatic variable weighting. Pattern Recognition, 47, 3082-3095.

[9] The Fungi of California: Species Index. URL: https://www.mykoweb.com/CAF/species_index.html#1_2, accessed March 29, 2024.

[10] M. Halynchik, "mhalynchik/interval_clustering/data" GitHub, Sep. 5, 2024. link (accessed Sep. 05, 2024).

[11] M. Halynchik, "mhalynchik/interval_clustering," GitHub, Sep. 5, 2024. link (accessed Sep. 05, 2024).

[12] Z. Huang, M.K. Ng, H. Rong, and Z. Li. (2005) Automated variable weighting in k-means type clustering, IEEE Transactions on Pattern Analysis and Machine Learning, 27(5), 657–668.

[13] Hubert, L. and Arabie, P. (1985) Comparing partitions, Journal of Classification, 2 (1), 193-218.

[14] Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B. and Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences, 622, 178-210.

[15] KMeans scikit-learn: "sklearn.cluster.KMeans — scikit-learn 1.30.0 documentation," Scikit-learn.org, 2023. link

[16] Lethikim, N., Lehoang, T. and Vovan, T. (2023). Automatic clustering algorithm for interval data based on overlap distance. Communications in Statistics-Simulation and Computation, 52(5), 2194-2209.

[17] Mirkin, B. (1999). Concept learning and feature selection based on square-error clustering. Machine Learning, 35, 25-39.

[18] Mirkin, B. (2018). Braverman's Spectrum and matrix diagonalization versus iK-means: a unified framework for clustering. In "Braverman Readings in Machine Learning. Key Ideas from Inception to Current State" (pp. 32-51). Springer International Publishing.

[19] Mirkin, B., Camargo, R., Fenner, T., Loizou, G., and Kellam, P. (2010). Similarity clustering of proteins using substantive knowledge and reconstruction of evolutionary gene histories in herpesvirus. Theoretical Chemistry Accounts, 125, 569-581.

[20] Nascimento, S., Martins, A., Relvas, P., Luís, J. F., and Mirkin, B. (2023). Core–shell clustering approach for detection and analysis of coastal upwelling. Computers and Geosciences, 179, 105421.

[21] "Institutes's Scientific Production," Ufpe.br, 2024. URL: https://www.cin.ufpe.br/ bap/ScientificProduction (accessed March 07, 2024).

[22] B. A. Pimentel and R. M. C. R. de Souza, "A weighted multivariate Fuzzy C-Means method in interval-valued scientific production data," Expert Systems with Applications, vol. 41, no. 7, pp. 3223–3236, Jun. 2014, doi: https://doi.org/10.1016/j.eswa.2013.11.013.

[23] Rao, W., Xia, J., Lyu, W., and Lu, Z. (2019). Interval data-based k-means clustering method for traffic state identification at urban intersections. IET Intelligent Transport Systems, 13(7), 1106-1115.

[24] Rico, N., Huidobro, P., Bouchet, A., and Díaz, I. (2022). Similarity measures for interval-valued fuzzy sets based on average embeddings and its application to hierarchical clustering. Information Sciences, 615, 794-812.

[25] S. I. Rizo Rodríguez and F. de A. Tenório de Carvalho, "Clustering interval-valued data with adaptive Euclidean and City-Block distances" Expert Systems with Applications, vol. 198, p. 116774, Jul. 2022, doi: https://doi.org/10.1016/j.eswa.2022.116774.

[26] Taran, Z., and Mirkin, B. (2020). Exploring patterns of corporate social responsibility using a complementary K-means clustering criterion. Business Research, 13(2), 513-540.

[27] D'Urso, P., De Giovanni, L., Alaimo, L. S., Mattera, R., and Vitale, V. (2023). Fuzzy clustering with entropy regularization for interval-valued data with an application to scientific journal citations. Annals of Operations Research, 1-24.

[28] Vovan, T., Phamtoan, D., Tuan, L. H., and Nguyentrang, T. (2021). An automatic clustering for interval data using the genetic algorithm. Annals of Operations Research, 303, 359-380.

**Разработка модификаций метода к-средних для кластерного анализа интервальных данных с использованием аномальных кластеров** [Электронный ресурс]: препринт WP7/2024/01 / М. Галынчик[1], Ф. Карвальо[2], А. Паринов[3], Б. Миркин[4] ; Нац. исслед. ун-т «Высшая школа экономики». — Электрон. текст. дан. (530 Кб) – М.: Изд. дом ВШЭ, 2024. — (Серия WP7 «Математические методы для принятия решений в экономике, бизнесе и политике»). – 34 с.

В последнее время наблюдается интерес к возможности распространения популярного метода кластерного анализа, к-средних, на так называемые интервальные данные. В отличие от случая обычных данных, значениями признаков здесь являются не отдельные числа, а интервалы вещественной оси. Как известно, одна из проблем метода к-средних — это инициализации метода, то есть определение местоположения гипотетических центров кластеров для начала итераций метода. Хотя результаты работы метода сильно зависят от инициализации, никакого универсального подхода к настоящему времени не существует.

В данной работе исследуется возможность использования аномальных кластеров для инициализации для метода при интервальных данных. А именно, мы используем «интервальную» версию пифагоровского разложения разброса данных на два слагаемых, одно из которых – минимизируемый критерий наименьших квадратов для метода к-средних, а второе – дополнительный критерий, требующий, чтобы кластеры были большие и аномальные. Мы получаем такие аномальные кластеры один за другим и используем центры самых больших из них для инициализации метода к-средних. При этом возникают различные версии за счет использования адаптивно настраиваемых весовых коэффициентов признаков. Мы показываем, что предложенный метод вполне конкурентоспособен на примере двух таблиц интервальных данных, впервые вводимых в научный оборот в данном тексте.

Ключевые слова: интервальные данные, К-средних, метод наименьших квадратов, аномальный кластер, весовые коэффициенты признаков

[1]Департамент анализа данных и искусственного интеллекта, Национальный исследовательский университет «Высшая школа экономики», Москва, Российская Федерация, maksgalinchik@gmail.com

[2]Федеральный университет Пернамбуку, Центр информатики UFPE, Бразилия, fatc@cin.ufpe.br

[3]Департамент анализа данных и искусственного интеллекта, Национальный исследовательский университет «Высшая школа экономики», Москва, Российская Федерация, aparinov@hse.ru

[4]Департамент анализа данных и искусственного интеллекта, Международная лаборатория анализа и выбора решений, Национальный исследовательский университет «Высшая школа экономики», Москва, Российская Федерация; Школа вычислительной техники и математических наук, Университет Биркбек, Лондон, Великобритания, bmirkin@hse.ru

**Препринты Национального исследовательского университета «Высшая школа экономики» размещаются по адресу: http://www.hse.ru/org/hse/wp**

*Препринт WP7/2024/01*
*Серия WP7*

Математические методы анализа решений
в экономике, бизнесе и политике

М. Галынчик, Ф. Карвальо, А. Паринов, Б. Миркин

# Разработка модификаций метода к-средних для кластерного анализа интервальных данных с использованием аномальных кластеров

(на английском языке)

Публикуется в авторской редакции