A hierarchical semantic-based distance for nominal histogram comparison

Camille Kurtz, Pierre Gançarski, Nicolas Passat, Anne Puissant

▶ **To cite this version:**

Camille Kurtz, Pierre Gançarski, Nicolas Passat, Anne Puissant. A hierarchical semantic-based distance for nominal histogram comparison. Data and Knowledge Engineering, 2013, 87, pp.206-225. 10.1016/j.datak.2013.06.002 . hal-01719116

HAL Id: hal-01719116
https://hal.univ-reims.fr/hal-01719116v1

Submitted on 28 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL
archives-ouvertes.fr

A hierarchical semantic-based distance for nominal histogram comparison

Camille Kurtz, Pierre Gançarski, Nicolas Passat, Anne Puissant

▶ **To cite this version:**

Camille Kurtz, Pierre Gançarski, Nicolas Passat, Anne Puissant. A hierarchical semantic-based distance for nominal histogram comparison. 2011. <hal-00634304v2>

HAL Id: hal-00634304
https://hal.archives-ouvertes.fr/hal-00634304v2

Submitted on 20 Dec 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A hierarchical semantic-based distance for nominal histogram comparison

Camille Kurtz, *Student Member, IEEE,* Pierre Gançarski, Nicolas Passat and Anne Puissant

Abstract—We propose a new distance, devoted to the comparison of nominal histograms equipped with a dissimilarity matrix providing the correlations between the bins. The computation of this distance is based on a hierarchical strategy, progressively merging the considered instances (and their bins) according to their semantic proximity. For each level of this hierarchy, a standard bin-to-bin distance is computed between the corresponding pair of histograms. In order to obtain the proposed distance, these bin-to-bin distances are then fused by taking into account the semantic coherency of their associated level. From this modus operandi, the proposed distance can handle histograms which are generally compared thanks to cross-bin distances. It preserves the advantages of such cross-bin distances (namely robustness to histogram translation and histogram bin size), while inheriting from the low computational cost of bin-to-bin distances. Validations in the context of geographical data classification emphasize the relevance and usefulness of the proposed distance.

Index Terms—histogram distance, nominal histogram, semantic-based distance, clustering.

1 Introduction

1.1 Context

A HISTOGRAM represents the distribution of quantified values of a measurement among the samples of a studied set. Such a set can gather, for example, the results of an experiment, or a population of individuals. In various domains, including the data mining field, it is necessary to classify large datasets, in which each data is characterized by one or more histograms. For instance, it is often necessary to classify populations in terms of the distribution of a particular measurement/feature (*e.g.*, the distribution of the size of the individuals contained in these populations). Histograms are then useful structures to model numerous kinds of data and enable to take into consideration their statistical properties.

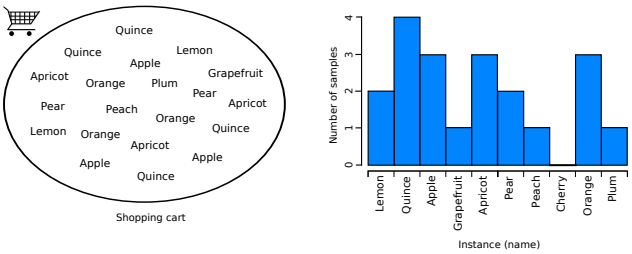
There exist different kinds of histograms related to specific types of measurements: nominal, ordinal (plus modulo, which are a special case of ordinal measurements) [1]. In a nominal measurement, each value is named and/or can represent an instance of a particular semantic concept (*e.g.*, the concept FRUIT can take values/instances such as Lemon, Quince, Apple, Grapefruit, Apricot, *etc.*). Then, a nominal type histogram can model the composition of a shopping cart according to the number and the kinds of fruits it contains (see Figure 1(a)). In such histogram, the measurement levels can be permuted since there is no (total) ordering among them (shuffling invariance property). On the contrary, in an ordinal measurement, the values are totally ordered (*e.g.*, the price of fruits can be quantized into 10 discrete values between 1 and 10 pounds). Thus, an ordinal type histogram can model the composition of a shopping cart according to the prices of the articles (see Figure 1(b)).

Measuring the similarity between histograms is a crucial operation in various domains such as clustering [2], [3], pattern classification and recognition [4], [5], text categorization [6], [7], time series analysis [8], or image retrieval [9], [10]. Indeed, the distance between pairs of histograms enables to assess the similarity of their corresponding statistical properties. For the last decades, several measures of similarity between histograms have been proposed. Histogram distances can be divided into two categories: *bin-to-bin* (or *vector*) and *cross-bin* (or *probabilistic*) distances. The bin-to-bin distances consider a histogram as a fixed-dimensional vector and only compare the content of corresponding histogram bins, while the cross-bin distances consider a histogram as an estimation of a *probability density function* and compare corresponding bins as well as non-corresponding ones.

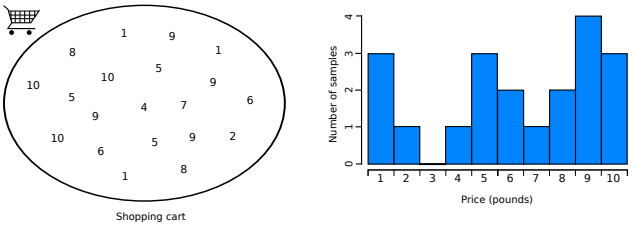
1.2 Motivation

In this work, we consider the comparison of possibly large datasets where data are characterized by nominal histograms for which (semantic) proximity information between the bins can be provided. In order to illustrate such histograms, let us go back to the example depicted in Figure 1(a). Each bin of the histogram represents the proportion of a kind of fruit which is an instance of the semantic concept FRUIT. Since Lemon and Orange are both citrus fruits, the

- *C. Kurtz, P. Gançarski and N. Passat are with Image Sciences, Computer Sciences and Remote Sensing Laboratory UMR CNRS 7005, University of Strasbourg, France.*
  E-mail: ckurtz@unistra.fr, gancarski@unistra.fr, passat@unistra.fr
- *A. Puissant is with the Image, City and Environment Laboratory ERL CNRS 7230, University of Strasbourg, France.*
  E-mail: anne.puissant@live-cnrs.unistra.fr



(a) Nominal histogram.



(b) Ordinal histogram.

Fig. 1. Two kinds of histograms modelling the composition of a shopping cart. (a) Nominal histogram modelling the distribution of the instances of the semantic concept FRUIT. (b) Ordinal histogram modelling the distribution of the prices of articles.

instance Lemon could be considered as closer to the instance Orange than the instance Plum. It is then possible to evaluate semantic similarities between the bins composing these nominal histograms. In this context, it can be advisable to take into account such kinds of semantic similarities to improve the comparison of these histograms.

On one hand, the bin-to-bin distances are well fitted to process large datasets, in particular thanks to their low computational cost. Nevertheless, by definition, such distances cannot consider the semantic proximity between the different bins. (In particular, they suffer from both problems of histogram translations and bin size changes.) For instance, let us consider the three shopping carts C_1, C_2, C_3, provided in Table 1. Their fruits can take values in {Lemon, Quince, ..., Plum}. Then, the composition of a shopping cart C_i can be modelled by a histogram $H_i(\#\text{Lemon}, \#\text{Quince}, \ldots, \#\text{Plum})$ where $\#x$ denotes the number of occurrences of the instance x in the shopping cart C_i. With a standard bin-to-bin distance, for instance the Manhattan one d_{L_1}, there is the same distance between H_1 and H_2 than between H_1 and H_3 ($d_{L_1}(H_1, H_2) = d_{L_1}(H_1, H_3) = 16$). However, C_1 is semantically closer to C_3 than to C_2 because C_1 and C_3 are both citrus fruits shopping carts.

On the other hand, cross-bin distances, which compare more exhaustively both corresponding and non-corresponding bins, enable to consider the semantic proximity between the different bins. Practically, this can be done by assigning, to each pair of instances, a weight (*i.e.*, a numerical value) modelling the degree of semantic proximity between the compared instances. For instance, in the shopping cart example, the weight associated to the couple of instances (Lemon, Orange) should be lower than the one associated to the couples (Lemon, Plum) or (Orange, Plum). The counterpart of such strategies is the quadratic cost induced by these multiple bin comparisons.

TABLE 1
Histograms modelling the composition of three shopping carts C_1, C_2, C_3, composed each of 10 fruits.

	Lemon	*Quince*	*Apple*	*Grapefruit*	*Apricot*	*Pear*	*Peach*	*Cherry*	*Orange*	*Plum*
H_1	9	0	0	0	0	0	0	0	1	0
H_2	1	0	0	0	0	0	0	0	1	8
H_3	1	0	0	0	0	0	0	0	8	1

Based on these considerations, it appears that, when comparing nominal histograms, the handling of semantic proximity between their instances seems incompatible with a low computational cost. In this work, we propose a new distance addressing this issue. Its computation is based on a hierarchical strategy, progressively merging the considered instances (and their bins) according to their semantic proximity. For each level of this hierarchy, a standard bin-to-bin distance is computed between the corresponding pair of histograms. In order to obtain the proposed distance, these bin-to-bin distances are then fused by taking into account the semantic coherency of their associated level. From this modus operandi, the proposed distance preserves the advantages of cross-bin distances (namely robustness to histogram translation and histogram bin size), while inheriting from the low computational cost of bin-to-bin distances.

1.3 Outline

This article is organized as follows. Section 2 introduces useful definitions and notations. Section 3 recalls different histogram similarity measures proposed in the literature. Section 4 describes the proposed hierarchical distance, dedicated to compare nominal histograms. Section 5 gathers experiments enabling to assess the relevance of this distance. Conclusions and perspectives will be found in Section 6.

2 DEFINITIONS

An interval on $\mathbb{R}$, bounded by $a, b \in \mathbb{R}$, will be noted $[a, b]$ while an interval on $\mathbb{Z}$, bounded by $a, b \in \mathbb{Z}$, will be noted $[\![a, b]\!]$. A list $\mathcal{L}_v$ of v elements e_i with $i \in [\![0, v-1]\!]$ is denoted by $\langle e_i \rangle_0^{v-1} = \langle e_0, e_1, \ldots, e_{v-1} \rangle$.

2.1 Histogram

For the sake of readability, we follow the same notations as in [1] and several subsequent articles. Let x

be a measurement, or an attribute, which can take v values in the set $X = \{x_0, x_1, \ldots, x_{v-1}\}$. Let A be a set of n elements/objects. Each element of A is associated to a value a by the measurement x. The "observation" set resulting from this measurement is denoted by $A_x = \{a_1, a_2, \ldots, a_n\}$ where $a_i \in X$. The histogram of the set A_x according to the measurement x of A, noted $H(x, A)$ is a list of v elements counting the number of occurrences of the values of x among the a_i. For the sake of concision, we will use $H(A)$ instead of $H(x, A)$. The histogram $H(A)$ can be defined as $H(A) = \langle H_0(A), H_1(A), \ldots, H_{v-1}(A)\rangle$ where $H_i(A)$, $i \in [\![0, v-1]\!]$, denotes the number of elements of A_x that have value x_i. Each $H_i(A)$ can be computed as

$$H_i(A) = \sum_{j=1}^{n} c_{ij} \quad \text{with} \quad c_{ij} = \begin{cases} 1 & \text{if } a_j = x_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In the literature, the different $H_i(A)$ are generally called the bins of the histogram $H(A)$. If $P_i(A)$ denotes the probability of samples in the i-th bin, then $P_i(A) = H_i(A)/n$. Then, $P(A)$ can be considered as a normalized histogram.

The v values of the measurement x are generally called *measurement levels* when they are used in $H(A)$ to index the distributions of the sample values. Without loss of generality, the v values of the measurement x are also called *instances* when they are used in $H(A)$ to index the distributions of the possible instances of a semantic concept.

2.2 Distance between measurement levels

A histogram $H(A)$ represents the distribution and the frequency of quantified values of a measurement x among the samples of a set A. Corresponding to the two types of measurements (*i.e.*, ordinal and nominal), we define two functions d_{ord} and d_{nom} which measure the difference between two measurement levels $x_i, x_j \in X$. In the literature, the difference between two measurement levels is called the *ground distance*.

2.2.1 Ordinal measurement

In an ordinal measurement, the values x_i are totally ordered and it is possible to determine a basic distance $\Delta_{(x_i,\ x_{i+1})} \in \mathbb{R}_+$ between each successive levels x_i and x_{i+1} of the measurement. Thus, we define the ground distance between two ordinal measurement values x_i and x_j as the sum of the basic distances between each successive levels from i to j:

$$d_{ord}(x_i, x_j) = \sum_{k=i}^{j-1} \Delta_{(x_k,\ x_{k+1})} \tag{2}$$

When the ordinal measurement values are numerical ones (*i.e.*, each $x_i \in \mathbb{R}$), the ground distance between two ordinal measurement values is the absolute difference between them:

$$d_{ord}(x_i, x_j) = \sum_{k=i}^{j-1} |x_k - x_{k+1}| = |x_i - x_j| \tag{3}$$

TABLE 2
Dissimilarity matrix $\mathcal{M}^{dis}$ associated to the instances of the concept FRUIT. As d_{nom} is symmetric, $\mathcal{M}^{dis}$ is a symmetric matrix (we only depict its upper right part).

x_i	Lemon	Quince	Apple	Grapefruit	Apricot	Pear	Peach	Cherry	Orange	Plum
Lemon	0.00	**0.80**	**0.90**	**0.20**	**0.70**	**0.80**	**0.90**	**0.80**	**0.10**	**0.85**
Quince	–	0.00	**0.20**	**0.75**	**0.40**	**0.20**	**0.45**	**0.50**	**0.78**	**0.48**
Apple	–	–	0.00	**0.90**	**0.40**	**0.02**	**0.50**	**0.45**	**0.95**	**0.45**
Grapefruit	–	–	–	0.00	**0.92**	**0.85**	**0.75**	**0.90**	**0.15**	**0.95**
Apricot	–	–	–	–	0.00	**0.40**	**0.15**	**0.07**	**0.90**	**0.10**
Pear	–	–	–	–	–	0.00	**0.40**	**0.40**	**0.90**	**0.40**
Peach	–	–	–	–	–	–	0.00	**0.10**	**0.90**	**0.05**
Cherry	–	–	–	–	–	–	–	0.00	**0.90**	**0.10**
Orange	–	–	–	–	–	–	–	–	0.00	**0.90**
Plum	–	–	–	–	–	–	–	–	–	0.00

2.2.2 Nominal measurement

In a nominal measurement, two cases can occur:
1) It is not possible to determine proximity relations between the values x_i. Thus, we define the ground distance between them as either match or mismatch:

$$d_{nom}(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

2) It is possible to determine semantic proximity relations between the values x_i. As the previous definition of d_{nom} does not enable to consider the semantic proximity between the bins of a nominal histogram, we extend the ground distance between two semantic nominal measurement values as:

$$d_{nom}(x_i, x_j) = d_{nom}(x_j, x_i) = \begin{cases} 0 & \text{if } x_i = x_j \\ \alpha_{(x_i,x_j)} & \text{otherwise} \end{cases} \tag{5}$$

where $\alpha_{(x_i,x_j)} \in\ ]0, 1]$ reflects the semantic dissimilarity between x_i and x_j that has been provided by the background knowledge of the expert (*e.g.*, the human perception). Then, it is possible to define a $v \times v$ dissimilarity matrix $\mathcal{M}^{dis}$ that models the relations between each instance $x \in X = \{x_0, x_1, \ldots, x_{v-1}\}$ of the concept linked to the histogram:

$$\mathcal{M}^{dis} = \begin{bmatrix} \alpha_{(x_0,x_0)} & \cdots & \alpha_{(x_0,x_{v-1})} \\ \vdots & \ddots & \vdots \\ \alpha_{(x_{v-1},x_0)} & \cdots & \alpha_{(x_{v-1},x_{v-1})} \end{bmatrix} \tag{6}$$

Table 2 presents an example of a dissimilarity matrix for the concept FRUIT[1] introduced in Section 1.

1. The dissimilarity values contained in this matrix have been defined for example purpose only and do not reflect a true semantic reality.

Metric property

The measures d_{ord} and d_{nom} presented in Equations (3–4) satisfy the following metric properties [1] and are then distances:

1) Non-negativity: $d(x_i, x_j) \geq 0$
2) Symmetry: $d(x_i, x_j) = d(x_j, x_i)$
3) Identity: $d(x_i, x_j) = 0 \Leftrightarrow x_i = x_j$
4) Triangle inequality: $d(x_i, x_k) \leq d(x_i, x_j) + d(x_j, x_k)$

By construction, the measure in Equation (5) satisfies, at least, the following properties:

1) Non-negativity: $d_{nom}(x_i, x_j) \geq 0$
2) Symmetry: $d_{nom}(x_i, x_j) = d_{nom}(x_j, x_i)$
3) Identity: $d_{nom}(x_i, x_j) = 0 \Leftrightarrow x_i = x_j$

However, as this measure is built using the human perception and semantics, the triangle inequality property is generally not satisfied, and thus, the extended version of d_{nom} is not a distance.

3 HISTOGRAM DISTANCES

For the last decades, several measures of similarity between histograms have been proposed. This section starts by introducing and discussing the advantages and limitations of the most used ones. The purpose and contribution of this article are then presented.

3.1 Related works

3.1.1 Bin-to-bin distances

As stated above, bin-to-bin distances consider a histogram as a fixed-dimensional vector and only compare the contents of the corresponding bins of the histograms. To compare these bins, it is possible to use various metrics/distance functions. The most commonly used are (non-exhaustively): Manhattan (L_1) distance, Euclidean (L_2) distance, Intersection distances, χ^2 distances, *etc.* In the sequel, they are denoted as D_{L_1}, D_{L_2}, $D_\cap$, and D_{χ^2}. For two histograms $H(A)$ and $H(B)$ with v bins, these distances can be formulated as:

$$D_{L_1}(H(A), H(B)) = \sum_{i=0}^{v-1} |H_i(A) - H_i(B)| \quad (7)$$

$$D_{L_2}(H(A), H(B)) = \sqrt{\sum_{i=0}^{v-1} |H_i(A) - H_i(B)|^2} \quad (8)$$

$$D_\cap(H(A), H(B)) = \sum_{i=0}^{v-1} \min(H_i(A), H_i(B)) \quad (9)$$

$$D_{\chi^2}(H(A), H(B)) = \sum_{i=0}^{v-1} \frac{|H_i(A) - H_i(B)|^2}{2(H_i(A) + H_i(B))} \quad (10)$$

These distances present specific properties, the respective advantages and drawbacks of which are discussed in [11].

As they only compare corresponding histogram bins, and then ignore the correlations between neighboring bins, bin-to-bin distances are fast to compute and can be used to measure similarities for large datasets. Moreover, they do not require any ordering or similarity relations among the bins, and can then be used to compare both nominal or ordinal histograms. However, they suffer from the translation problem: a small translation of the histogram values may significantly affect the histogram distance. Furthermore, bin-to-bin distances are deeply linked to the bin size of histograms: a too coarse binning will not have a sufficient discriminative capacity while a too fine one will separate similar/correlated features in different bins which will not be matched. The cross-bin distances enable to overcome these limitations.

3.1.2 Cross-bin distances

Different cross-bin distances have been proposed to compare pairs of histograms in a more accurate fashion than bin-to-bin distances. They can be divided in two families: those requiring a (dis)similarity matrix to model the proximity relations among the bins, and those requiring a total ordering on the bins.

Matrix-based distances Among the matrix-based distances, the quadratic form ones [12], [13] use a similarity matrix $\mathcal{M}^{sim}$ (which is the "opposite" of $\mathcal{M}^{dis}$, see Equation (6)) to model the similarity relationships between the bins and compute the histograms distance as a matrix product:

$$D_{quad}(H(A), H(B)) = \left[(H(A) - H(B))^T \mathcal{M}^{sim} (H(A) - H(B))\right]^{\frac{1}{2}} \quad (11)$$

These approaches provide a first solution to take into account the similarity among the bins. However, they suffer from an important computational cost. For instance, it is shown in [14] that using the quadratic form distance in image retrieval tasks leads to weak results, where the mutual similarity of color distributions is overestimated.

The match distances are another form of cross-bin distances. The principle of these approaches is to estimate the cost of mapping two histograms. Among them, the Earth Mover's Distance (EMD)[2] is recognized as the most efficient distance measure. This distance has been deeply studied during the last decades [17]. In the EMD, two features are considered as the "earth" and the "holes", respectively. Then, the distance measure problem is transformed into the earth moving problem, where the minimum cost of moving all the "earth" into the "holes" is calculated. As for quadratic distances, the EMD makes use of a similarity matrix $\mathcal{M}^{sim}$ to model the similarity relationships between the bins. The EMD method enables

2. The Earth Mover's Distance is also called the Mallows distance [15] in the field of statistics [16].

to consider the correlations between the bins and to reduce the sensitivity of the distance to the bin size. Furthermore, it was shown in [18] that the EMD outperforms most of the existing cross-bin distances for image retrieval tasks. The main drawback of the EMD is its high computational cost. Indeed, solving the associated linear programming problem requires huge computation time, *a fortiori* when the dimensionality of the feature space is high. Although several methods have been proposed to reduce this complexity (*e.g.*, see [19]), it remains difficult to use the EMD for large-database applications.

Order-based distances When a total ordering is available on the instances, the EMD can be simplified [1], [20] as:

$$D_{EMD}(H(A),H(B)) = \sum_{i=0}^{v-1} \left| \sum_{j=0}^{i} (H_j(A) - H_j(B)) \right| \quad (12)$$

This definition is then equivalent to the one proposed in [21] where the match distance between two one-dimensional histograms is defined as the L_1 distance between their cumulative corresponding histograms. This distance is efficient for one-dimensional histograms but it is difficult to use it to compare high dimensional ones, as the mapping in high dimensional spaces requires to solve complex graph matching problems [22].

Temporal measures of similarity can also be viewed as a special case of order-based cross-bin distances. For instance, in [8], the similarity is related to the closeness of positions and shapes of peaks in the compared histograms. The main drawback of such measures is to focus on the peaks of the histograms that is generally not sufficient to compare data structures where empty bins also carry information. Another solution consists of using the Dynamic Time Warping (DTW) similarity measure [23], [24]. This temporal similarity measure enables small distortions when matching pairs of histograms. However, it requires to set a distortion parameter. If this parameter is set to small values, this measure is similar to the Manhattan bin-to-bin distance. On the contrary, if it is set to a high value, this measure can handle histogram distortions (but with a higher computational cost).

On one hand, bin-to-bin distances enable to compare histograms with a low computational cost but without considering the possible semantic correlations between the bins. On the other hand, cross-bin distances, and in particular the most general ones, namely the matrix-based distances can deal with such semantic correlations but present a much higher computational cost (despite efforts conducted to reduce this cost, in particular in the case of ordinal histograms).

TABLE 3
Histograms modelling the compositions of three shopping carts C_1, C_2, C_3 (see Table 1) after the creation of the Citrus instance.

	Citrus	*Quince*	*Apple*	*Apricot*	*Pear*	*Peach*	*Cherry*	*Plum*
H'_1	**10**	0	0	0	0	0	0	0
H'_2	**2**	0	0	0	0	0	0	8
H'_3	**9**	0	0	0	0	0	0	1

3.2 Histograms and hierarchies

For the last decades, it has been experimentally proved that organizing the semantic information, carried out by the data, into hierarchies can facilitate knowledge extraction tasks, as illustrated, *e.g.*, in image analysis [25]. Indeed, there exist a wide range of histogram-based approaches relying on this paradigm. In such approaches, the considered hierarchy of histograms is generally composed by the histograms of a hierarchy of data which are compared with standard distances.

However, hierarchical strategies directly linked to histograms seem well adapted to deal with the issue evoked in Section 1.2. In order to illustrate this assertion, let us go back to the fruit shopping cart example introduced above.

Let us consider the three shopping carts C_1, C_2, C_3 (composed each of 10 fruits, see Table 1). We recall that the composition of a shopping cart C_i is modelled by a histogram $H_i(\#\text{Lemon}, \#\text{Quince}, \ldots, \#\text{Plum})$. Suppose now that we fuse the instances Lemon, Orange and Grapefruit to create a new instance called Citrus. The composition of a shopping cart C_i is now modelled by a histogram $H'_i(\#\textbf{Citrus}, \#\text{Quince}, \ldots, \#\text{Plum})$ where $\#\textbf{Citrus} = \#\text{Lemon} + \#\text{Grapefruit} + \#\text{Orange}$. The resulting composition of the three shopping carts C_1, C_2, C_3 is presented in Table 3. The Manhattan bin-to-bin distance d_{L_1} becomes now higher between H'_1 and H'_2 ($d_{L_1}(H'_1, H'_2) = 16$) than between H'_1 and H'_3 ($d_{L_1}(H'_1, H'_3) = 2$). This measure value reflects better the semantic similarities between C_1 and C_3 which are both citrus fruits shopping carts.

As illustrated by this example, hierarchical distances naturally enable to consider the multilevel semantic correlations between the distributions modelled by the histograms. To the best of our knowledge, the only histogram distance based on such hierarchical strategy has been proposed in [26]. Its computation relies on the iterative merging of the closest bins of the histograms to create coarser histograms. As the distance measure value is obtained by computing iteratively a chosen bin-to-bin distance, its computational cost is lower than those required for cross-bin distances. This distance, which has been involved in image retrieval applications, has provided

encouraging results. Nevertheless, as its merging process requires a total ordering among the bins, this distance only deals with ordinal histograms. In the case considered in this work, namely the comparison of nominal histograms equipped with a dissimilarity matrix, this distance is therefore irrelevant.

In the next section, we propose to address this issue by defining a hierarchical distance dealing with such nominal histograms. In particular, this new distance, called *Hierarchical Semantic-Based Distance* (HSBD), combines the efficiency of bin-to-bin distances (*e.g.*, low computational cost) and the advantages offered by cross-bin distances (*e.g.*, robustness to both histogram translation and bin size issues).

4 THE HSBD DISTANCE

4.1 Workflow

The computation of the HSBD distance between two histograms $H(A)$ and $H(B)$ of v bins, requires two parameters:

1) a dissimilarity matrix $\mathcal{M}^{dis}$ modelling the semantic proximity values between the v instances of the concept represented by these histograms;
2) a bin-to-bin histogram distance D_{bin}.

Before effectively computing the distance between $H(A)$ and $H(B)$, the adopted strategy requires to define a way to hierarchically merge the different instances of the histograms into clusters (*i.e.*, instances of higher semantic levels). This "pre-processing" step, described in Section 4.2, mainly consists of building a dendrogram $\mathfrak{D}$ induced by $\mathcal{M}^{dis}$ modelling this instance merging hierarchy. (Note that this step has to be performed only once for a given matrix $\mathcal{M}^{dis}$.)

Once the dendrogram $\mathfrak{D}$ has been built, the HSBD distance can be computed. This computation is organized in two main steps:

- Step 1. **Hierarchical bin-to-bin sub-distances computation** During an iterative merging process (scanning each stage of the dendrogram from the leaves to the root), the histograms $H^k(A)$ and $H^k(B)$ associated to $H(A)$ and $H(B)$, which are induced by the merging of the instances composing each cluster of the stage S_k, are built. After each iteration, a bin-to-bin sub-distance D_{bin} is then computed between the couple of coarser histograms $H^k(A)$ and $H^k(B)$ created previously.
- Step 2. **Bin-to-bin sub-distances fusion** The bin-to-bin sub-distances computed for all the stages of the dendrogram, and the "semantic energy" required to go from one stage to the next, are then fused into a function which is finally integrated to provide the HSBD distance.

These two steps are fully described in Section 4.3. The reader may also refer to Figure 3 for a visual outline of the computation of HSBD. Finally, Section 4.4 provides a computational complexity study of HSBD.

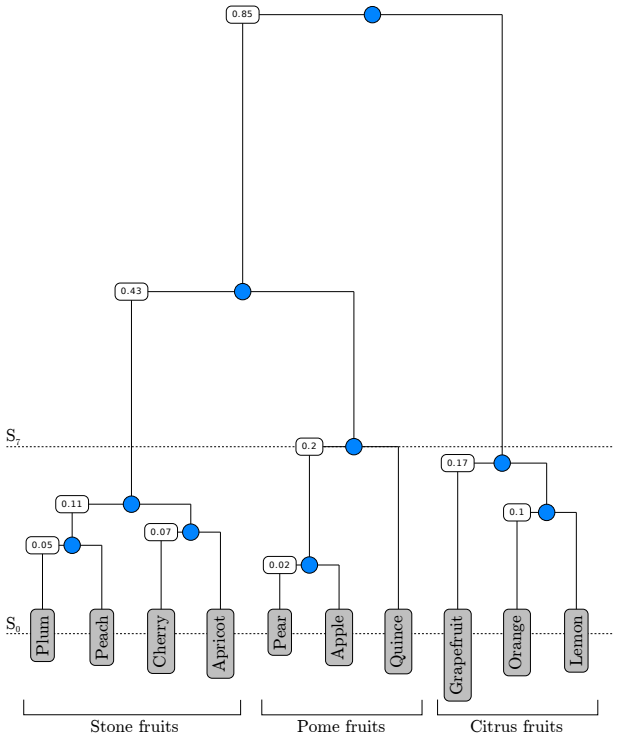


Fig. 2. Dendrogram associated to the dissimilarity matrix presented in Table 2. The basic instances are represented by gray rectangles while the instances of higher semantic level are represented by blue disks.

4.2 Building the merging hierarchy

The principle of the proposed approach is to compute several times a bin-to-bin distance between pairs of histograms by progressively merging the semantically closest bins/instances to create coarser histograms of higher semantic levels. To this end, it is necessary to determine the order of the fusions between the instances of the semantic concept. Such order can be naturally determined by defining an instance merging hierarchy.

Starting from the values contained in $\mathcal{M}^{dis}$, it is possible to compute the instance merging hierarchy by using the Ascendant Hierarchical Clustering (AHC) algorithm [27]. It performs in four steps:

- Step 1: Begin with groups containing only one basic instance (*i.e.*, v groups where v is the number of instances).
- Step 2: Compute the dissimilarity values between every group couples, and update the dissimilarity matrix $\mathcal{M}^{dis}$.
- Step 3: Merge the two closest groups (*i.e.*, the groups which have the lowest dissimilarity value in $\mathcal{M}^{dis}$), and modify $\mathcal{M}^{dis}$ accordingly (by merging the two lines/columns associated to these two groups).
- Step 4: If there are more groups than desired (generally, one group), go to step 2.

This algorithm hierarchically builds clusters of in-

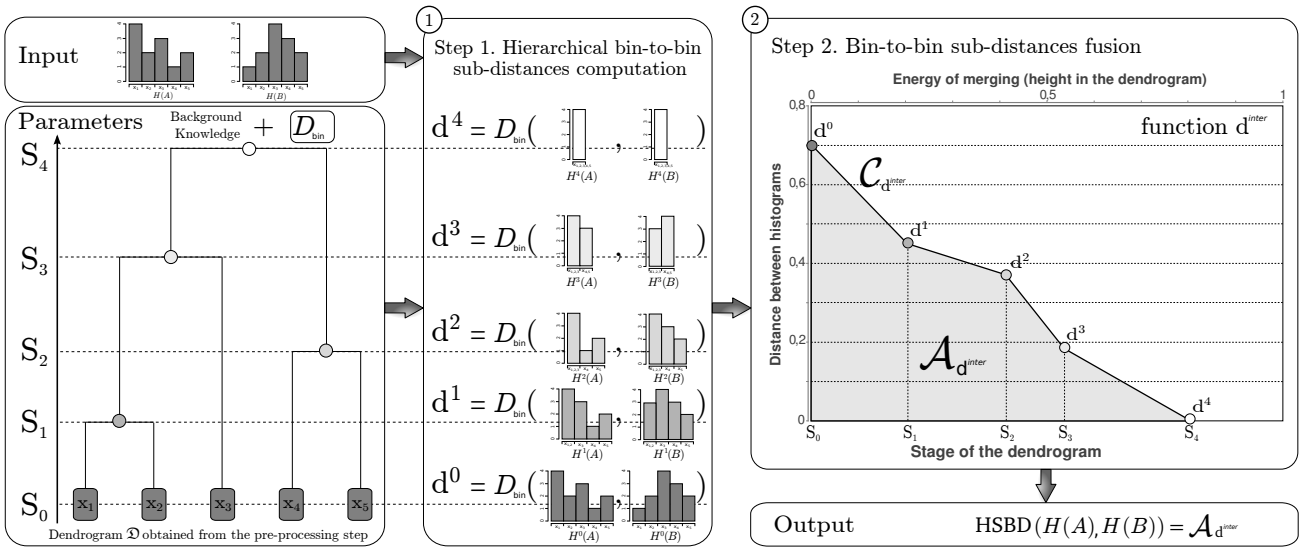


Fig. 3. Computation workflow of the HSBD distance.

stances while minimizing their intra-group inertia. To compute the dissimilarity values between every group couple, it is necessary to choose a linkage criteria. In this work, we have chosen to use the well-known Average Linkage criterion which generally provides satisfactory results.

A merging hierarchy is usually modelled by a dendrogram $\mathfrak{D}$ of s stages[3], whose root is the cluster that contains all the instances. Each stage of $\mathfrak{D}$ corresponds to a particular semantic level. The minimal value of s ($s_{min} = 2$) is reached when $\mathcal{M}^{dis}$ is a matrix where $\alpha_{(x_i,x_j)} = 1$ if $x_i \neq x_j$ and $\alpha_{(x_i,x_j)} = 0$ otherwise (*i.e.*, no background knowledge). In this case, the dendrogram presents one stage for the leaves which are the basic instances and one stage for the root. On the contrary, the maximal value of s ($s_{max} = v$) is reached when $\mathfrak{D}$ is a totally unbalanced dendrogram.

From the considered dendrogram, we define:

- a function $f_{\mathfrak{D}}$ which takes as input the index k of the stage S_k ($k \in [\![0, s-1]\!]$) and provides as output the list $\mathcal{L}_m^k = \langle \mathcal{L}_{v_0}, \ldots, \mathcal{L}_{v_{m-1}} \rangle$ composed of the m instance merging lists $\mathcal{L}_{v_i}$ ($i \in [\![0, m-1]\!]$) induced by $\mathfrak{D}$ at this stage (*i.e.*, m clusters).
- a function $h_{\mathfrak{D}}$ which takes as input the index k of the stage S_k and provides as output its height $h_{\mathfrak{D}}(k)$ in the dendrogram $\mathfrak{D}$.

This height $h_{\mathfrak{D}}(k)$ corresponds to the "semantic energy" required to build the considered clusters of instances (*i.e.*, the inter-group inertia computed when running the AHC algorithm).

For instance, Figure 2 illustrates the dendrogram associated to the dissimilarity matrix defined in Table 2. In this example, the height of the stage S_7 is given by the function $h_{\mathfrak{D}}(7) = 0.2$. The list $\mathcal{L}_3^7$ of the 3 instance merging lists at stage S_7 is given by $f_{\mathfrak{D}}(7) = \langle \langle Plume, Peach, Cherry, Apricot \rangle, \langle Pear, Apple, Quince \rangle, \langle Grapefruit, Orange, Lemon \rangle \rangle$. (Note that these 3 instance merging lists correspond to the three classes of higher semantic level: Stone fruits, Pome fruits and Citrus fruits.)

3. Such stages are generally indexed by their depth in the dendrogram (*i.e.*, from the root to the leaves). In our case, and for readability purpose, we index these stages by their level in the dendrogram (*i.e.*, from the leaves to the root).

4.3 Computation of HSBD

Based on the dendrogram $\mathfrak{D}$ provided from the pre-processing step, it becomes possible to compute HSBD between the two histograms $H(A)$ and $H(B)$ of v bins.

4.3.1 Step 1. ***Hierarchical bin-to-bin sub-distances computation***

To compute the hierarchical bin-to-bin distance during the iterative merging process, we define the function $\mathbf{d}^k$ for any $k \in [\![0, s-1]\!]$. This function, which provides the distance D_{bin} between the coarser versions $H^k(A)$ and $H^k(B)$ of the histograms $H(A)$ and $H(B)$ at stage S_k of $\mathfrak{D}$, is defined as

$$\mathbf{d}^k(H(A), H(B)) = D_{bin}(H^k(A), H^k(B)) \tag{13}$$

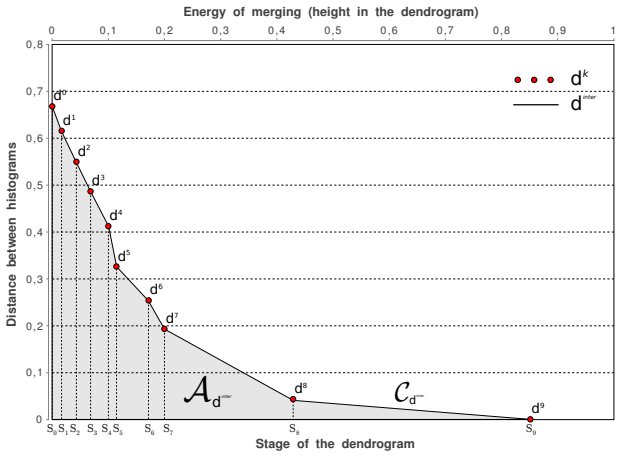
Such coarser histograms (linked to a higher level of semantics) can be built using the function $f_{\mathfrak{D}}(k)$ which provides a list $\mathcal{L}_m^k = \langle \mathcal{L}_{v_0}, \ldots, \mathcal{L}_{v_{m-1}} \rangle$ composed of the m instance merging lists induced by the stage S_k. More formally, the histogram $H^k(Y)$ is defined as

$$H^k(Y) = \langle H_0^k(Y), H_1^k(Y), \ldots, H_{m-1}^k(Y) \rangle \tag{14}$$

where each bin $H_i^k(Y)$ is computed as

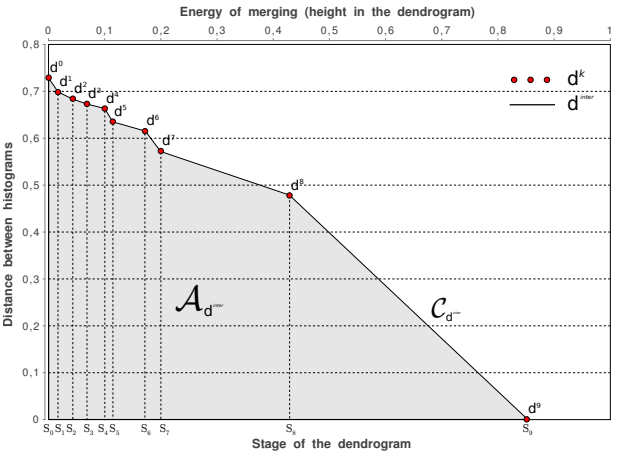
$$H_i^k(Y) = \sum_{j \in \mathcal{L}_{v_i}} H_j(Y) \tag{15}$$

For the sake of concision, $\mathbf{d}^k(H(A), H(B))$ will be simply denoted as $\mathbf{d}^k$. Furthermore, the values produced



	Lemon	Quince	Apple	Grapefr.	Apricot	Pear	Peach	Cherry	Orange	Plum
$H(A)$	25	15	0	20	0	0	25	0	10	5
$H(B)$	5	0	5	40	0	10	0	0	30	10

(a) Comparison of *semantically* similar histograms.



	Lemon	Quince	Apple	Grapefr.	Apricot	Pear	Peach	Cherry	Orange	Plum
$H(A)$	0	25	25	0	5	15	0	0	0	30
$H(B)$	20	0	10	0	30	0	0	5	25	0

(b) Comparison of *semantically* different histograms.

Fig. 4. Graphical representation of the $\mathbf{d}^k$ values and the function $\mathbf{d}^{inter}$, computed between pairs of example histograms. Each histogram models the composition of a shopping cart composed of 100 fruits. The $\mathbf{d}^k$ values are represented by red disks while the function $\mathbf{d}^{inter}$ is represented by the curve $\mathcal{C}_{\mathbf{d}^{inter}}$ with black lines. Depending on the contents of the histograms $H(A)$ and $H(B)$, the behavior of the function $\mathbf{d}^{inter}$ widely differs.

by this function will also be called the *sub-distance* values in the remainder of this article.

The iterative hierarchical merging process performs as follows (Figure 3-①)): Firstly, the bin-to-bin distance $\mathbf{d}^0$ is computed for $H(A)$ and $H(B)$ by considering all the bins of the histograms (*i.e.*, $f_{\mathfrak{D}}(0) = \langle\langle x_0\rangle, \langle x_1\rangle, \ldots, \langle x_{v-1}\rangle\rangle$ and $h_{\mathfrak{D}}(0) = 0$). Then, by climbing to the next stage S_i of the dendrogram, the closest bins of the histograms (given by $f_{\mathfrak{D}}(i)$) are merged, and a new sub-distance $\mathbf{d}^i$ is computed between the two resulting histograms $H^i(A)$ and $H^i(B)$. This sub-distance enables to assess the similarity at a specific level of binning and semantics. This step is repeated for each stage $S_k, k \in [\![1, s-1]\!]$ until the number of bins is equal to 1 (*i.e.*, the process stops when the root of the dendrogram is reached), and a series of fine-to-coarse sub-distances is stored as $\mathbf{d}^0, \ldots, \mathbf{d}^{s-1}$. Note that $\mathbf{d}^{s-1}$ is always equal to 0 since $H^{s-1}(A)$ and $H^{s-1}(B)$ are always composed of only one bin representing the instance of highest semantic level of the dendrogram (*i.e.*, the root).

4.3.2 Step 2. ***Bin-to-bin sub-distances fusion***

Once all the fine-to-coarse sub-distances $\mathbf{d}^0, \ldots, \mathbf{d}^{s-1}$ have been computed, it is possible to fuse them to get the HSBD distance between the two considered histograms $H(A)$ and $H(B)$.

In order to introduce and motivate the proposed approach, let us consider the examples provided in Figure 4. This figure illustrates the graphical representation of the $\mathbf{d}^k$ values, computed between pairs of example histograms. One can note that for semantically/thematically similar histograms (see Figure 4(a)), the $\mathbf{d}^k$ values tend to decrease more rapidly than for dissimilar ones (see Figure 4(b)).

This behavior is linked to analytic properties of the function induced by the $\mathbf{d}^k$ values. More precisely, let us consider the piecewise affine function $\mathbf{d}^{inter} : [h_{\mathfrak{D}}(0), h_{\mathfrak{D}}(s-1)] \to \mathbb{R}_+$ defined by $\mathbf{d}^{inter}(h_{\mathfrak{D}}(k)) = \mathbf{d}^k$ for any $k \in [\![0, s-1]\!]$ (see Figure 4). The decreasing rate of $\mathbf{d}^{inter}$, which characterizes the similarity between histograms, is directly linked to the integral value $\mathcal{A}_{\mathbf{d}^{inter}}$ of this function (Figure 3-②).

This assertion justifies the definition of the Hierarchical Semantic-Based Distance (HSBD) as

$$\mathrm{HSBD}(H(A), H(B)) = \int_{h_{\mathfrak{D}}(0)}^{h_{\mathfrak{D}}(s-1)} \mathbf{d}^{inter}(t).dt \quad (16)$$

Practically, this distance can be computed by using the Trapezoidal Rule, then leading to the following discrete formulation

$$\mathrm{HSBD}(H(A), H(B)) = \frac{1}{2}\sum_{k=0}^{s-2}\left[\left(\mathbf{d}^{k+1} + \mathbf{d}^k\right)\left(h_{\mathfrak{D}}(k+1) - h_{\mathfrak{D}}(k)\right)\right] \quad (17)$$

Remark 1. *It is possible to use the HSBD similarity measure with a "partial" dissimilarity matrix $\mathcal{M}^{dis}$. Indeed, in*

the case where a proximity cannot be established between some of the considered instances, the associated dissimilarity values are set to 1 *in* $\mathcal{M}^{dis}$*. For such parts of the matrix, the proposed similarity measure will act as the underlying bin-to-bin distance* D_{bin} *for the associated instances. (Note that the matrix considered in the experiments of Section 5 (see Table 4) is an example of such partial matrix.)*

In particular, if no background knowledge is available, $\mathcal{M}^{dis}$ *is a matrix where* $\alpha_{(x_i,x_j)} = 1$ *if* $x_i \neq x_j$ *and* $\alpha_{(x_i,x_j)} = 0$ *otherwise. In such conditions, the HSBD similarity measure is coherently equivalent to the underlying bin-to-bin distance:*

$$HSBD(H(A), H(B)) = \frac{1}{2} D_{bin}(H(A), H(B)) \quad (18)$$

Metric property

To be a distance, a measure has to satisfy the following properties: non-negativity, symmetry, identity and triangle inequality. The HSBD measure can be defined as a weighted sum of sub-distances $\mathbf{d}^k$ (Equation (17)) which is equivalent to a sum of bin-to-bin distances D_{bin}. We demonstrate hereafter that the HSBD measure inherits from the metric properties of D_{bin} and is then a distance.

Fact 1 (Non-negativity property)**.** *HSBD has non-negativity property:* $HSBD(H(A), H(B)) \geq 0$*.*

Proof: The result derives from the non-negativity property of the function $\mathbf{d}^k$ and the decreasingness of the function $h_{\mathfrak{D}}$. □

Fact 2 (Symmetry property)**.** *HSBD has symmetry property:* $HSBD(H(A), H(B)) = HSBD(H(B), H(A))$*.*

Proof: The result derives from the symmetry property of D_{bin}. □

Fact 3 (Identity property)**.** *HSBD has identity property* $HSBD(H(A), H(B)) = 0 \Leftrightarrow H(A) = H(B)$*.*

Proof: The "$\Leftarrow$" part of the result straightforwardly derives from the identity property of D_{bin}.

The "$\Rightarrow$" part of the result is detailed hereafter:

$$\begin{aligned}
&\text{HSBD}(H(A), H(B)) = 0 \\
&\Rightarrow \tfrac{1}{2} \textstyle\sum_{k=0}^{s-2} \big[\underbrace{(\mathbf{d}^{k+1} + \mathbf{d}^k)}_{\geq 0} \underbrace{(h_{\mathfrak{D}}(k+1) - h_{\mathfrak{D}}(k))}_{>0} \big] = 0 \\
&\Rightarrow \textstyle\sum_{k=0}^{s-2} \big[D_{bin}(H^{k+1}(A), H^{k+1}(B)) \\
&\qquad\qquad + D_{bin}(H^k(A), H^k(B)) \big] = 0 \\
&\Rightarrow \forall k \in [0, s-2], D_{bin}(H^k(A), H^k(B)) = 0 \\
&\Rightarrow H(A) = H(B)
\end{aligned}$$

We obtain $\text{HSBD}(H(A), H(B)) = 0 \Rightarrow H(A) = H(B)$. □

Fact 4 (Triangle inequality property)**.** *HSBD has triangle inequality property* $HSBD(H(A), H(B)) + HSBD(H(B), H(C)) \geq HSBD(H(A), H(C))$*.*

Proof: The result derives from the triangle inequality property of D_{bin}. □

4.4 Computational complexity

We detail hereafter the computational complexity of the proposed distance. To compute HSBD, it is first necessary to run a pre-processing step (see Section 4.2), which consists of building the merging order (*i.e.*, to build the dendrogram $\mathfrak{D}$). To this end, we use the Ascendant Hierarchical Clustering (AHC) algorithm. The complexity of the naive AHC algorithm is $\Theta(v^3)$ where v is the number of basic instances, since it is necessary to exhaustively scan the $v \times v$ matrix $\mathcal{M}^{dis}$ for the smallest dissimilarity in each of the $v-1$ iterations (see Section 4.2, Step 4). However, this complexity can be reduced to $\Theta(v^2 \log v)$ by using a priority-queue algorithm.

Once the dendrogram has been built, it becomes possible to compute the HSBD distance. This requires the pre-computation of the series of sub-distances $\mathbf{d}^0, \ldots, \mathbf{d}^{s-1}$ (see Equation (17)) where s corresponds to the number of stages in $\mathfrak{D}$. The complexity of the computation of each $\mathbf{d}^k$ is directly linked to the one of D_{bin} which is, in general, $\Theta(v)$ where v denotes the number of bins in the histograms (see Equations (7-10)). Thus, the complexity of HSBD depends on both the number of bins v in the histograms and the number of stages s in the dendrogram $\mathfrak{D}$.

Depending on the value of s (which is correlated to the shape of $\mathfrak{D}$), different cases can appear:

- if $\mathfrak{D}$ is a "flat" dendrogram (*i.e.*, a 2-stage one), then $s = 2$ and the complexity of HSBD becomes

$$\Theta(v + 1) = \Theta(v) \quad (19)$$

- if $\mathfrak{D}$ is a totally balanced dendrogram, then $s = \log_2(v)$ and the complexity of HSBD becomes

$$\Theta(\underbrace{v + \frac{v}{2} + \frac{v}{4} + \ldots + 1}_{\log_2(v) \text{ terms}}) = \Theta(2v) = \Theta(v) \quad (20)$$

- if $\mathfrak{D}$ is a totally unbalanced dendrogram, then $s = v$ and the complexity of HSBD becomes

$$\Theta(\underbrace{v + (v-1) + \ldots + 1}_{v \text{ terms}}) = \Theta(v^2) \quad (21)$$

Thus, the computational cost of HSBD is bounded by $\Theta(v)$ and $\Theta(v^2)$. To this complexity, it is necessary to add the one required to build the dendrogram $\mathfrak{D}$ which takes $\Theta(v^2 \log v)$ operations. However, as this operation is only performed once, its complexity can be considered as insignificant when comparing large histogram datasets.

To conclude this complexity study, the computation of HSBD requires more time/operations than the computation of classical bin-to-bin distances but much less than the cross-bin distances that can require, in the worst case, supercubic time. Thus, HSBD can be relevantly used to compare large histogram datasets.

5 Experimental studies

To assess the relevance of the proposed similarity measure, we have applied it to the clustering of geographical data. We start by introducing the context of this experimental study in Section 5.1. The datasets that were used to test the method are then described in Section 5.2. Finally, the experiments performed using these data are presented in Section 5.3.

5.1 Context: Classification of geographical data

Urban planning and development organizations, disaster or environment agencies need to follow the increase of urban settlements. To this end, it is necessary to map urban areas from satellite images. Since the last decades, numerous efforts have been conducted to automatically extract features from satellite images, in order to involve them into learning systems. The classical methodology consists of classifying the data into land cover classes by using supervised or unsupervised classification [28].

The mapping of urban areas can be realized at different scales corresponding each to a particular level of analysis. To validate the propose distance, we have applied it to the classification of urban blocks, which can be defined by the minimal cycles closed by communication ways. The main originality of this task is to classify sets of urban blocks that are characterized by their "ground" compositions in terms of basic urban objects (*e.g.*, individual houses, gardens, roads, *etc.*) [29]. For instance, a urban block U_i can be characterized by a histogram $H_i(\#\text{Red tile roof}, \#\text{Slate roof}, \ldots, \#\text{Herbaceous vegetation})$ where Red tile roof, Slate roof, ..., Herbaceous vegetation are the instances of the concept Urban Object (Table 4).

The main issue is to succeed in classifying into a same cluster several objects that are not characterized by similar histograms. For instance, let us consider a block U_i characterized by a histogram $H_i(21, 4, \ldots, 10)$ (*i.e.*, U_i is composed of 21 red tile roofs, 4 slate roofs, ..., 10 vegetation parcels) and a block U_j characterized by a histogram $H_j(3, 22, \ldots, 10)$ (*i.e.*, U_j is composed of 3 red tile roofs, 22 slate roofs, ..., 10 vegetation parcels). From the expert point of view, these two blocks have to be grouped into the same class "Urban fabric with individual houses" because they are both composed of individual houses (with red tile roofs or slate ones) and vegetation parcels.

To deal with this issue, a solution consists of using a classification process associated to a distance that takes into consideration the semantic correlations of the data. We propose to validate the usefulness of the HSBD distance by integrating it into a classification algorithm to classify such data.

5.2 Datasets

We consider three datasets (denoted **Dataset-1**, **-2** and **-3**) composed each of (i) a set of urban blocks manually extracted by an expert and (ii) a high resolution map providing the composition of each block in terms of basic urban objects.

Each urban block U_i (of each dataset) has then been characterized by a "composition" histogram $H(U_i)$ which models the composition of the urban block U_i in terms of the distribution of the eleven instances

TABLE 4
Dissimilarity matrix $\mathcal{M}^{dis}$ associated to the instances of the concept Urban Object.

x_i	Shadow	Water	Road	Railway	Bare soil	Herbac. veget.	Forest	Slate roof	Red tile roof	Gray tile roof	Metallic roof
Shadow	0.00	**0.10**	**1.00**	**1.00**	**1.00**	**1.00**	**0.80**	**0.80**	**0.80**	**0.80**	**0.80**
Water	–	0.00	**0.50**	**0.50**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Road	–	–	0.00	**0.10**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Railway	–	–	–	0.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Bare soil	–	–	–	–	0.00	**0.10**	**0.80**	**1.00**	**1.00**	**1.00**	**1.00**
Herbac. veget.	–	–	–	–	–	0.00	**0.80**	**1.00**	**1.00**	**1.00**	**1.00**
Forest	–	–	–	–	–	–	0.00	**1.00**	**1.00**	**1.00**	**1.00**
Slate roof	–	–	–	–	–	–	–	0.00	**0.10**	**0.40**	**0.90**
Red tile roof	–	–	–	–	–	–	–	–	0.00	**0.40**	**0.90**
Gray tile roof	–	–	–	–	–	–	–	–	–	0.00	**0.10**
Metallic roof	–	–	–	–	–	–	–	–	–	–	0.00

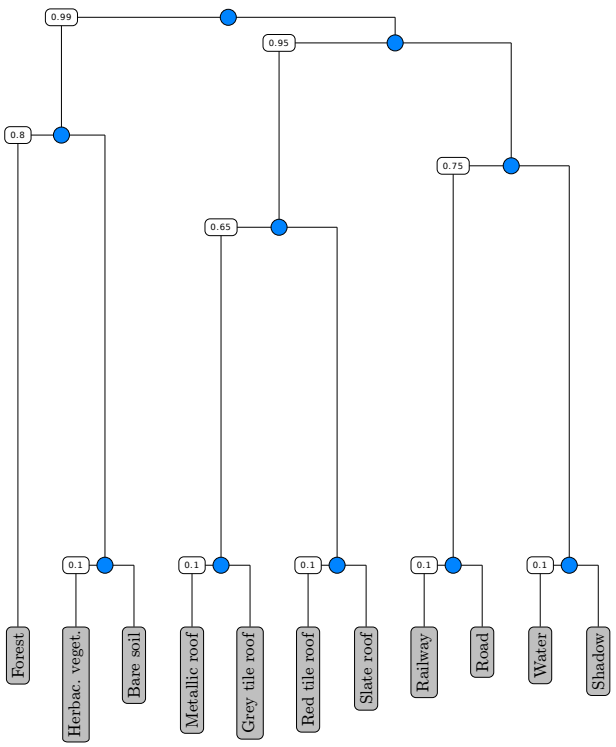


Fig. 5. Dendrogram associated to the dissimilarity matrix presented in Table 4. The basic instances are represented by gray rectangles while the instances of higher semantic levels are represented by blue disks.

of the semantic concept URBAN OBJECT in the high resolution map. These eleven instances considered in this experiment are listed in Table 4. Note that as each urban block U_i is not composed of the same number of urban objects in the high resolution map, we use the normalized histogram $P(U_i)$ instead of $H(U_i)$.

5.3 Experiments

5.3.1 Experimental settings

To model the semantic relations between each level of the histograms, a 11×11 dissimilarity matrix $\mathcal{M}^{dis}$ has been provided by the expert (Table 4). The dissimilarity values, which are modelled by this matrix, enable to take into consideration the semantic "proximity" between the considered levels and the land use function attached to each one of these levels. From this matrix, a 7-stage dendrogram has been built (using the pre-process, see Section 4.2) to model the merging order of these different levels (Figure 5).

Supervised classification algorithms require training examples to learn the classification model. In our case, the definition of such training examples remains a complex task for the expert. Indeed, the high number of considered classes induced a high number of examples to define. Moreover, such training examples are strongly data-dependent and can not be reused directly to classify other datasets. For these reasons, we have chosen to use an unsupervised classification algorithm that does not require the definition of such examples. We have applied the K-MEANS clustering algorithm [30], which does not require *a priori* parameters, to classify the urban blocks created previously.

To process, the distance HSBD has been directly integrated into the K-MEANS clustering algorithm to compare the classified histograms. We have respectively run the K-MEANS clustering algorithm with the HSBD distance associated to different bin-to-bin subdistances (HSBD_{L_1}, HSBD_{L_2}, and HSBD_{χ^2}). To compare HSBD to other existing distances, we have also run the K-MEANS algorithm using classical bin-to-bin distances D_{L_1}, D_{L_2}, and D_{χ^2}. These comparisons enable to assess the advantages of using HSBD instead of a classical bin-to-bin distance (*e.g.*, HSBD_{L_1} *vs.* D_{L_1}, HSBD_{L_2} *vs.* D_{L_2} and HSBD_{χ^2} *vs.* D_{χ^2}). As the results provided by the K-MEANS algorithm are sensitive to the initialization step of the algorithm, each run has been repeated 10 times by varying the "seeds" of the algorithm. We have then computed the variance value σ obtained for each considered evaluation index (described hereafter) and for each series of run.

From these datasets, we have chosen, in agreement with the expert, to extract nine classes of urban blocks: c1 - Dense urban fabric; c2 - Urban fabric with housing blocks; c3 - Urban fabric with individual houses; c4 - Industrial urban fabric; c5 - Water surfaces; c6 - Roads; c7 - Agricultural zones; c8 - Urban vegetation; c9 - Forest areas; except for the **DATASET-2** and **DATASET-3** where the "Dense urban fabric" class can not be extracted. Thus, the K-MEANS algorithm has been run respectively with ten clusters for the **DATASET-1** and nine clusters for the **DATASET-2** and **DATASET-3**.

TABLE 5
Evaluation measures.

Symbol	Evaluation measure	Type
$\mathcal{P}$	Precision index	Per-class accuracy
$\mathcal{R}$	Recall index	Per-class accuracy
$\mathcal{F}$	F-measure	Per-class accuracy
$\mathcal{K}$	Kappa index	Global accuracy
$\overline{\mathcal{F}}$	Weighted harmonic mean of $\mathcal{F}$	Global accuracy

5.3.2 Results evaluation

Evaluating clustering results is a complex task since it is difficult to find an objective measure of quality of clusters. A common strategy consists of assessing and comparing the intra/inter cluster inertia of the different clustering results (*i.e.*, unsupervised evaluation). Nevertheless, in our case the clustering algorithm is run by varying the distance used to compare the data (and then the inertia definitions vary). Thus, such measure of goodness seems not relevant to assess the quality of the clustering results obtained.

We decided to consider supervised evaluation techniques which consists of comparing a clustering result to a set of data manually labelled by the expert. Thus, we have compared the obtained clustering results to different land cover reference maps (extracted from geographical databases or provided by the expert). To this end, we have computed both standard local and global evaluation indexes (see Table 5).

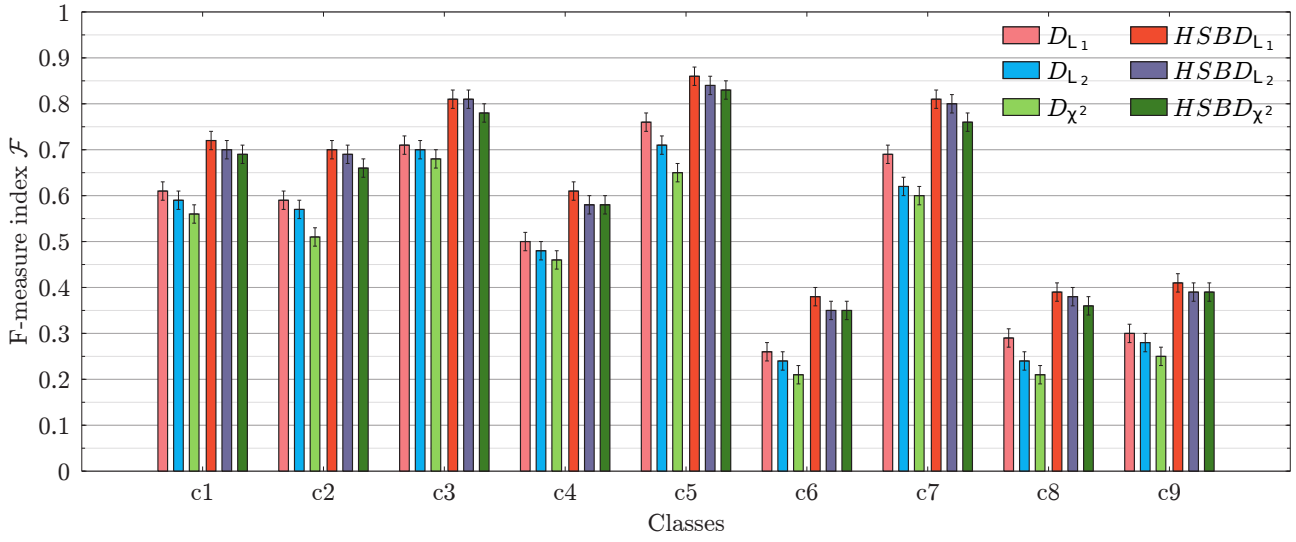
Local evaluation Local evaluation indexes enable to independently assess the extraction of each thematic class. To process, for each thematic class, the best corresponding clusters were extracted. Then, we have computed: the percentage of false positives, denoted by $f^{(p)}$, the percentage of false negatives, denoted by $f^{(n)}$, and the percentage of true positives, denoted by $t^{(p)}$. These measures are used to estimate the precision $\mathcal{P}$ and the recall $\mathcal{R}$ of the results obtained by using the proposed method:

$$\mathcal{P} = \frac{t^{(p)}}{t^{(p)} + f^{(p)}} \quad \text{and} \quad \mathcal{R} = \frac{t^{(p)}}{t^{(p)} + f^{(n)}} \tag{22}$$

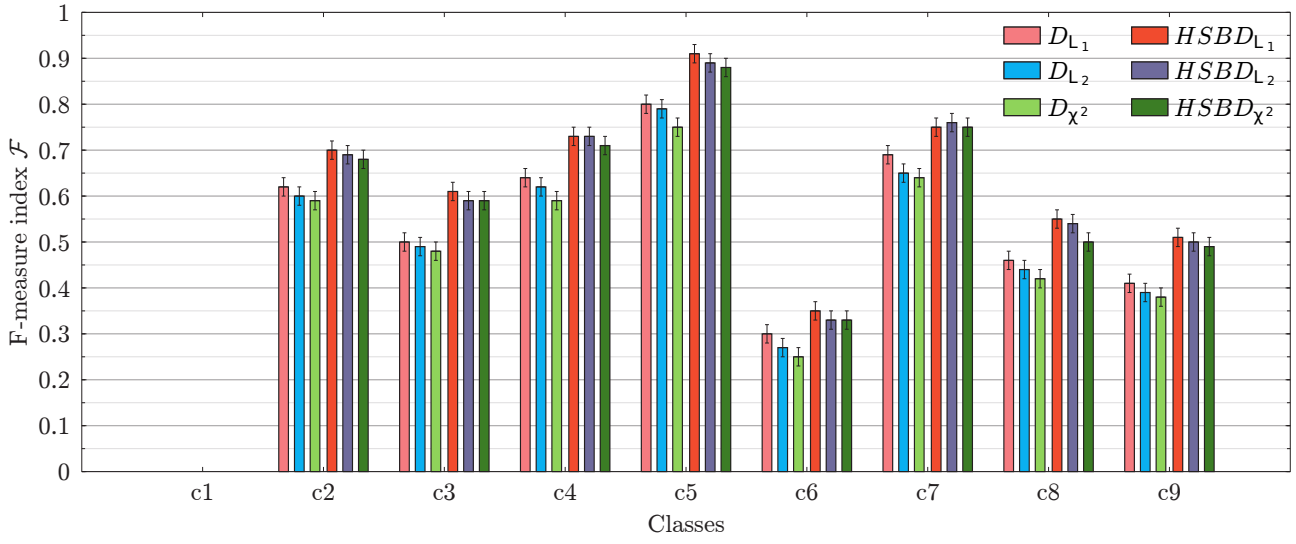
To fuse these measures, we have computed the standard F-measure $\mathcal{F}$ which is the harmonic mean of precision and recall:

$$\mathcal{F} = 2 \cdot \frac{\mathcal{P} \cdot \mathcal{R}}{\mathcal{P} + \mathcal{R}} \tag{23}$$

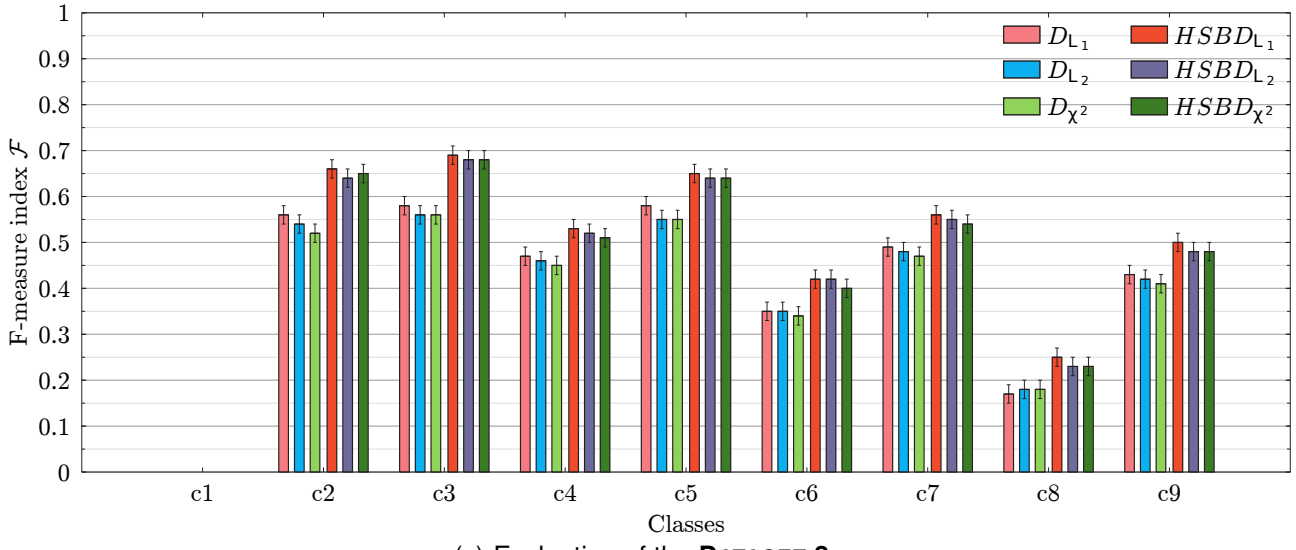
Global evaluation To assess the relevance of the results, we also provide global classification accuracy indexes. For each experiment, we have computed



(a) Evaluation of the **DATASET-1**.



(b) Evaluation of the **DATASET-2**.



(c) Evaluation of the **DATASET-3**.

Fig. 6. Local evaluation results on **DATASET-1**, **DATASET-2**, **DATASET-3**.

TABLE 6
Global evaluation results on **DATASET-1**, **DATASET-2**, **DATASET-3**.

		$\overline{\mathcal{F}} \pm \sigma$		$\mathcal{K} \pm \sigma$	
Dataset	Bin-to-bin distance	D_{bin}	HSBD$_{bin}$	D_{bin}	HSBD$_{bin}$
DATASET-1	L_1	0.65 ± 0.02	$\mathbf{0.71 \pm 0.01}$	0.76 ± 0.02	$\mathbf{0.79 \pm 0.01}$
	L_2	0.63 ± 0.03	$\mathbf{0.69 \pm 0.02}$	0.75 ± 0.03	$\mathbf{0.78 \pm 0.02}$
	χ^2	0.59 ± 0.02	$\mathbf{0.65 \pm 0.02}$	0.72 ± 0.03	$\mathbf{0.75 \pm 0.02}$
DATASET-2	L_1	0.67 ± 0.02	$\mathbf{0.72 \pm 0.01}$	0.77 ± 0.02	$\mathbf{0.86 \pm 0.02}$
	L_2	0.64 ± 0.01	$\mathbf{0.70 \pm 0.02}$	0.76 ± 0.02	$\mathbf{0.83 \pm 0.01}$
	χ^2	0.61 ± 0.02	$\mathbf{0.68 \pm 0.03}$	0.73 ± 0.01	$\mathbf{0.76 \pm 0.02}$
DATASET-3	L_1	0.63 ± 0.01	$\mathbf{0.66 \pm 0.01}$	0.73 ± 0.01	$\mathbf{0.76 \pm 0.01}$
	L_2	0.60 ± 0.03	$\mathbf{0.63 \pm 0.01}$	0.73 ± 0.02	$\mathbf{0.75 \pm 0.01}$
	χ^2	0.58 ± 0.02	$\mathbf{0.62 \pm 0.02}$	0.71 ± 0.01	$\mathbf{0.73 \pm 0.02}$

the weighted harmonic mean $\overline{\mathcal{F}}$ of the F-measures (weighted by the cardinals of the thematic classes), and the Kappa index [31] defined as:

$$\mathcal{K} = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \tag{24}$$

where $\Pr(a)$ is the relative agreement among the observers, and $\Pr(e)$ is the hypothetical probability of chance agreement. The Kappa index takes value in $[0, 1]$ and decreases as the classification is in disagreement with the ground-truth map. We have computed this index using the strategy proposed in [5].

5.3.3 Results

These different indexes have been used to quantitatively evaluate the clustering results obtained with HSBD and to compare them to the results obtained with classical bin-to-bin distances. We present hereafter both local and global evaluation results.

The local evaluation results obtained on the three datasets are presented in Figure 6. From these graphs, one can see that the F-measure scores obtained for each extracted class are always higher when the K-MEANS algorithm is run with the HSBD distance instead of its corresponding bin-to-bin distance D_{bin}. In particular, the best scores have been obtained when the HSBD distance is run with the Manhattan D_{L_1} distance as sub-distance, while the worst scores have been obtained when HSBD is run with the χ^2 as sub-distance. However, the scores obtained with the χ^2 as sub-distance remain always higher than those obtained by using only standard bin-to-bin distances. Such local evaluation results mean that the HSBD distance enables to enhance the precision and the recall of the results.

The global evaluation results obtained on the three datasets are presented in Table 6. They lead to the same observations as previously. Best global evaluation scores obtained are always higher when the K-MEANS algorithm is run with the HSBD distance instead of its corresponding bin-to-bin distance D_{bin}. Furthermore, the best scores have been obtained when the HSBD distance is run with the D_{L_1} sub-distance.

From these different results, one can see that the proposed distance outperforms the classical bin-to-bin ones when comparing semantic nominal histograms. Finally, such validations, in the context of geographical data classification, emphasize the relevance and usefulness of HSBD for data mining tasks.

6 CONCLUSION AND PERSPECTIVES

This article has presented a new distance dedicated to compare nominal histograms equipped with a dissimilarity matrix modelling the semantic proximity relations between the bins. Thanks to a hierarchical strategy, this distance enables to consider the multilevel semantic correlations between the bins. Moreover, by opposition to cross-bin distances (which can handle such histograms), it inherits from the low computational cost of bin-to-bin distances, while keeping the advantages of cross-bin ones, namely robustness to histogram translation and histogram bin size issues.

To validate this distance, we have applied it to the clustering of geographical data. The results appear to be sufficient to further accurately perform both supervised classification or object recognition tasks. This seems to validate the relevance of the proposed distance and the soundness of the approach.

This work opens up several perspectives and different research directions. From a methodological point of view, we plan to study more formally the possible behaviors of the sub-distance function $\mathbf{d}^k$ in order to enhance the distance computational cost. Furthermore, it could be relevant to integrate an approach enabling to help the user for building the dissimilarity matrix. Indeed, by asking him for constraint examples between the data (*e.g.*, must-link or cannot-link constraints), semi-supervised clustering approaches could be used to learn the α values. From an applicative point of view, this distance could be used for several applications, including the classification of large text datasets or the clustering of symbolic patterns.

REFERENCES

[1] S. H. Cha and S. N. Srihari, "On measuring the distance between histograms," *Pattern Recognition*, vol. 35, no. 6, pp. 1355–1370, 2002.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[3] E. K. K. Ng, A. W. C. Fu, and R. C. W. Wong, "Projective clustering by histograms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 369–383, 2005.

[4] M. Bressan, D. Guillamet, and J. Vitrià, "Using an ICA representation of local color histograms for object recognition," *Pattern Recognition*, vol. 36, no. 3, pp. 691–701, 2003.

[5] C. Kurtz, N. Passat, P. Gançarski, and A. Puissant, "Extraction of complex patterns from multiresolution remote sensing images: A hierarchical top-down methodology," *Pattern Recognition*, vol. 45, no. 2, pp. 685–706, 2012.

[6] M. Capdevila and O. W. M. Florez, "A communication perspective on automatic text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 7, pp. 1027–1041, 2009.

[7] S. Fabrizio, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, pp. 1–47, 2002.

[8] V. V. Strelkov, "A new similarity measure for histogram comparison and its application in time series analysis," *Pattern Recognition Letters*, vol. 29, no. 13, pp. 1768–1774, 2008.

[9] R. Brunelli and O. Mich, "Histograms analysis for image retrieval," *Pattern Recognition*, vol. 34, no. 8, pp. 1625–1637, 2001.

[10] Y. Liu, D. Zhang, G. Lu, and W. Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, 2007.

[11] S. H. Cha, "Taxonomy of nominal type histogram distance measures," in *Proceedings of the American Conference on Applied Mathematics*. World Scientific and Engineering Academy and Society (WSEAS), 2008, pp. 325–330.

[12] C. W. Niblack, R. Barber, W. Equitz, M. D. Flickner, E. H. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin, "QBIC project: Querying Images By content, Using Color, Texture and Shape," in *Proceedings of the SPIE Conference on Storage and Retrieval for Image and Video Databases*, vol. 1908, no. 1, 1993, pp. 173–187.

[13] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 7, pp. 729–736, 1995.

[14] M. A. Stricker and M. Orengo, "Similarity of color images," in *Proceedings of the SPIE Conference on Storage and Retrieval for Image and Video Databases*, vol. 2420, no. 1, 1995, pp. 381–392.

[15] C. L. Mallows, "A note on asymptotic joint normality," *Annals of Mathematical Statistics*, vol. 43, no. 2, pp. 508–515, 1972.

[16] E. Levina and P. Bickel, "The Earth Mover's Distance is the Mallows distance: some insights from statistics," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, 2001, pp. 251–256.

[17] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, pp. 99–121, 2000.

[18] Y. Rubner, J. Puzicha, C. Tomasi, and J. M. Buhmann, "Empirical evaluation of dissimilarity measures for color and texture," *Computer Vision and Image Understanding*, vol. 84, no. 1, pp. 25–43, 2001.

[19] H. Ling and K. Okada, "An Efficient Earth Mover's Distance algorithm for robust histogram comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 840–853, 2007.

[20] F. Serratosa and A. Sanfeliu, "Signatures versus histograms: Definitions, distances and algorithms," *Pattern Recognition*, vol. 39, no. 5, pp. 921–934, 2006.

[21] H. C. Shen and A. K. C. Wong, "Generalized texture representation and metric," *Computer Vision, Graphics, and Image Processing*, vol. 23, no. 2, pp. 187–206, 1983.

[22] M. Werman, S. Peleg, and A. Rosenfeld, "A distance metric for multidimensional histograms," *Computer Vision, Graphics, and Image Processing*, vol. 32, no. 3, pp. 328–336, 1985.

[23] T. M. Rath and R. Manmatha, "Word image matching using Dynamic Time Warping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2003, pp. 521–527.

[24] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognition*, vol. 44, no. 3, pp. 678–693, 2011.

[25] A. M. Tousch, S. Herbin, and J. Y. Audibert, "Semantic hierarchies for image annotation: A survey," *Pattern Recognition*, vol. 45, no. 1, pp. 333–345, 2012.

[26] Y. Ma, X. Gu, and Y. Wang, "Histogram similarity measure using variable bin size distance," *Computer Vision and Image Understanding*, vol. 114, no. 8, pp. 981–989, 2010.

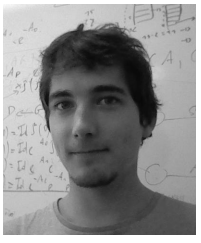
[27] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[28] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 1, pp. 2–16, 2010.

[29] C. Kurtz, N. Passat, P. Gançarski, and A. Puissant, "Multiresolution region-based clustering for urban analysis," *International Journal of Remote Sensing*, vol. 31, no. 22, pp. 5941–5973, 2010.

[30] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281–297.

[31] R. Congalton, "A review of assessing the accuracy of classifications of remotely sensed data," *Remote Sensing of Environment*, vol. 37, no. 1, pp. 35–46, 1991.



Camille Kurtz was born in 1986. He received his M.Sc. degree in Computer Science from the University of Strasbourg (France) in 2009. He is currently working toward the Ph.D. degree at the Image Sciences, Computer Sciences and Remote Sensing Laboratory, Strasbourg. His research interests include hierarchical segmentation, multiresolution clustering, image analysis and remote sensing.

Pierre Gançarski was born in 1959. He received his Ph.D. degree in 1989 and his Habilitation degree in 2007 in Computer Science from the University of Strasbourg. He is currently a full Professor at the Department of Computer Science from the University of Strasbourg. His current research interests include collaborative multistrategical clustering with applications to complex data mining and remote sensing image analysis.

Nicolas Passat was born in 1978. He received his Ph.D. degree in 2005 and his Habilitation degree in 2011 in Computer Science from the University of Strasbourg. Since 2006, he has been Assistant Professor of Computer Science at the University of Strasbourg. His scientific interests include medical and remote sensing image processing and analysis, mathematical morphology, and discrete topology.

Anne Puissant was born in 1975. She received her PhD degree in 2003 in Geography and Remote Sensing from the University of Strasbourg. She is currently Assistant Professor in the Geography Department at the University of Strasbourg. Her research topics are focused on the utility of Earth observation data to improve the knowledge of landscapes and to manage their state and dynamics and the spatial analysis of natural processes.