Intelligent Data Analysis 15 (2011) 109–130
DOI 10.3233/IDA-2010-0459
IOS Press

Measure based metrics for aggregated data

V.J. Rayward-Smith
School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK
E-mail: vjrs@uea.ac.uk

Abstract. Aggregated data arises commonly from surveys and censuses where groups of individuals are studied as coherent entities. The aggregated data can take many forms including sets, intervals, distributions and histograms. The data analyst needs to measure the similarity between such aggregated data items and a range of metrics are reported in the literature to achieve this (e.g. the Jaccard metric for sets and the Wasserstein metric for histograms). In this paper, a unifying theory based on measure theory is developed that establishes not only that known metrics are essentially similar but also suggests new metrics.

Keywords: Measure theory, metric space, similarity, clustering, symbolic data, aggregated data

1. Introduction

Large data sets concerning individual entities found in medical, census or financial databases, for example, are often too large and/or too sensitive to be released to a wider community. To facilitate analysis of such data, it is common practice to aggregate the data based on individuals into data based upon groups of individuals. For example, with census data, data might be generated and analysed that describes geographically based communities. This can not only protect the individual but also be used as a means of comparing communities and thereby targeting strategic government funding. For medical data, aggregated data might be based on hospitals or health authorities and comparisons between them can then be made. Once the data has been aggregated, not only is it more manageable but it can often be safely released to a wide community, perhaps even to the general public.

Aggregated data, often referred to as symbolic data [2,3], usually have a markedly different structure from that of an individual. An entry for an individual might have a field describing the individual's age. The corresponding data for a group of individuals might be a set of ages, an interval of ages or a histogram describing the distribution of ages within the group.

The analyst needs to compare one group with another and thus needs techniques to measure the similarity between aggregated data. To measure similarities between items, it is common to seek a metric (or perhaps a pseudometric). In this paper, metrics and pseudometrics are defined over various types of aggregated data. These measures can then be combined to get an overall measure of similarity between the aggregated groups. Defining such measures correctly is important because it can affect policy and investment, internationally, nationally and locally, by companies, organisations and by governments.

In this paper, metrics and pseudometrics for aggregated data are studied. By introducing some simple measure theory, such metrics and pseudometrics are seen to have much in common; they are all special cases of one of two (pseudo)metrics defined in terms of a measure over an algebra. Section 2 introduces the measure theory required and Section 3 defines a metric space and explains how metrics and pseudometrics can be derived for an algebra over which a measure is defined. Then, in Section 4,

the theory that has been developed is applied to generate metrics and pseudometrics over sets, intervals and histograms. Throughout, care is taken to distinguish between categorical data that is nominal and that which is ordinal, as well as distinguishing between categorical and numeric data. The last section of the paper presents conclusions and some suggestions for further research.

Metrics are important in the analysis of unaggregated data, especially in clustering applications (see, e.g. [17]). Their use is dicussed further in [19] together with various metric based maesures for cluster qulaity. A scalable, metric based algorithm is described in [9]. Studying metrics for aggregated data is also not new. There is a large and growing corpus of work in this area both of a theoretical and of an applied nature, see [2–4,6–8,10–12,15], several of which include case studies relating to the analysis of census data. This paper provides a unifying theory for many existing metrics used in these articles and also produces some new metrics.

2. Finitely additive measures

Let S be a set and let Σ be a non-empty set of subsets of S that is closed under complement and union. Thus, if A is in Σ then so is the complement of A, $A' = S \setminus A$. Similarly if A, B are in Σ then $A \cup B$ is also in Σ. Providing these properties are satisfied, (S, Σ) is called an *algebra*. By applying de Morgan's law, any algebra, (S, Σ), will also be closed under intersection.

A *finitely additive measure*, μ, on an algebra, (S, Σ), is a function

$$\mu : \Sigma \to R \cup \{\infty\}$$

such that

1. $\mu(A) \geqslant 0$ for all $A \in \Sigma$,
2. $\mu(A) = 0$ if $A = \emptyset$,
3. If A, B are disjoint sets in Σ then
$$\mu(A \cup B) = \mu(A) + \mu(B).$$

A finitely additive measure is a relaxed form of a measure. A measure is defined on a σ-algebra, which is an algebra that is also closed under the union of a countable number of sets, rather than just a finite number of sets, see for example, [1,13]. A measure then has all the properties of a finitely additive measure but also satisfies the additional propert that if $A_1, A_2, \ldots$ is a countably infinite sequence of disjoint sets in the σ-algebra then

$$\mu(\bigcup A_i) = \sum \mu(A_i).$$

If $\mu(A)$ is finite for all $A \in \Sigma$, a finitely additive measure is called *finite* and all of the example measures used in this paper are indeed finite.

A finitely additive measure, μ, will be called *strong* if $\mu(A) = 0 \Rightarrow A = \emptyset$. Not all finitely additive measures discussed here are strong but, when they are, a metric can be constructed rather than just a pseudometric.

For any sets $A, B \in \Sigma$, the sets $A \setminus B, B \setminus A$ and $A \cap B$ are mutually disjoint. Thus,the following can be deduced.

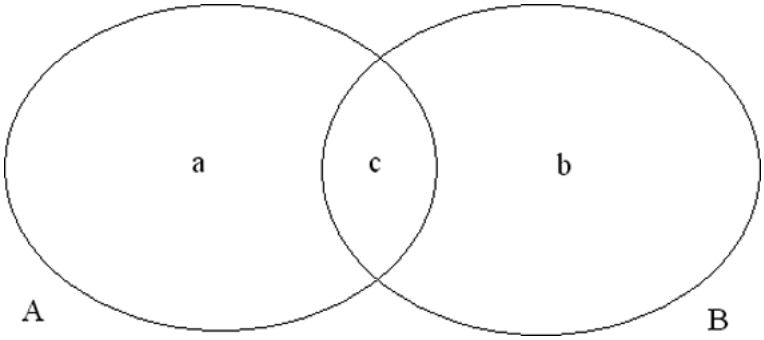

Fig. 1. Two intersecting sets.

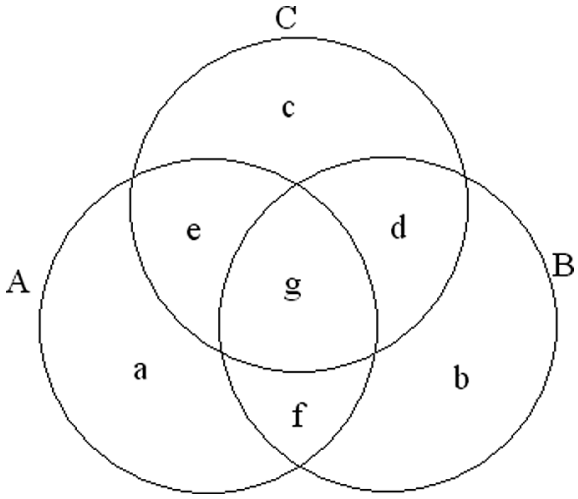

Fig. 2. Three intersecting sets.

Proposition 1 *For any finitely additive measure, μ, on the algebra (S, Σ) and for any sets, $A, B \in \Sigma$, if $a = \mu(A \setminus B), b = \mu(B \setminus A)$ and $c = \mu(A \cap B)$, as in Fig. 1, then*

$$\mu(A) = \mu(A \setminus B) + \mu(A \cap B) = a + c,$$
$$\mu(B) = \mu(B \setminus A) + \mu(A \cap B) = b + c \text{ and}$$
$$\mu(A \cup B) = \mu(A \setminus B) + \mu(B \setminus A) + \mu(A \cap B) = a + b + c.$$

Similarly,

Proposition 2 *For any three sets, A, B, C in Σ, if $a = \mu(A \setminus B \setminus C), b = \mu(B \setminus A \setminus C), c = \mu(C \setminus A \setminus B), d = \mu((B \cap C) \setminus A), e = \mu((C \cap A) \setminus B), f = \mu((A \cap B) \setminus C)$ and $g = \mu(A \cap B \cap C)$, as in Fig. 2, then*

$$\mu(A) = a + e + f + g,$$
$$\mu(B) = b + d + f + g,$$
$$\mu(C) = c + d + e + g,$$

$$\mu(A \cup B) = a + b + d + e + f + g,$$
$$\mu(B \cup C) = b + c + d + e + f + g \text{and}$$
$$\mu(A \cup C) = a + c + d + e + f + g.$$

These two propositions are key to proving the following results on metrics and to understanding this paper.

3. Metrics

To be a *metric* on Σ, a distance function $\delta : \Sigma \times \Sigma \to R_0^+$ must satisfy:

1. $\delta(A, B) \geqslant 0$,
2. $\delta(A, B) = 0$ if and only if $A = B$,
3. δ is symmetric, i.e. $\delta(A, B) = \delta(B, A)$ for all $A, B \in \Sigma$, and
4. δ satisfies the *triangle inequality*, i.e. $\delta(A, B) + \delta(B, C) \geqslant \delta(A, C)$ for all $A, B, C \in \Sigma$.

(Σ, δ) is then called a *metric space*.

Metrics are used to define the difference between objects in the set Σ and are widely used both to compare objects and within clustering algorithms, see e.g. [17].

If δ satisfies all the conditions of being a metric, except that $\delta(A, B) = 0$ can occur when $A \neq B$, then δ is called a *pseudometric*. Clearly, any pseudometric will infer a metric on the equivalence classes of Σ defined by the equivalence relation $A \sim B$ iff $\delta(A, B) = 0$.

Given a finitely additive measure, μ, on an algebra, (S, Σ), the distance function, $\delta_1 : \Sigma \times \Sigma \to R_0^+$ is defined by

$$\delta_1(A, B) = \mu(A \cup B) - \mu(A \cap B) = \mu(A \setminus B) + \mu(B \setminus A).$$

Then,

1. $\delta_1(A, B) \geqslant 0$ since, by Proposition 1, $\mu(A \cup B) - \mu(A \cap B) = \mu(A \setminus B) + \mu(B \setminus A)$, which must be $\geqslant 0$,
2. $\delta_1(A, B) = 0$ if $A = B$ since then $\mu(A \cup B) - \mu(A \cap B) = \mu(A) - \mu(A) = 0$,
3. if $\delta_1(A, B) = 0$ then $\mu(A \setminus B) + \mu(B \setminus A) = 0$ and thus both $\mu(A \setminus B) = 0$ and $\mu(B \setminus A) = 0$. If μ is a strong measure then it follows that $A \setminus B = \emptyset$ and $B \setminus A = \emptyset$ and hence, $A = B$,
4. $\delta_1(A, B) = \delta_1(B, A)$ by the symmetry of the definition, and
5. using the notation of Fig. 2, for any sets, $A, B, C \in \Sigma$,

$$\begin{aligned}\delta_1(A, B) + \delta_1(B, C) &= a + e + b + d + b + f + c + e \\ &\geqslant a + f + c + d \\ &= \delta_1(A, C)\end{aligned}$$

By the results listed above for δ_1, the following result is established.

Theorem 1. *Given a finitely additive measure, μ, on an algebra, (S, Σ), a pseudometric, $\delta_1 : \Sigma \times \Sigma \to R_0^+$ can be defined by*

$$\delta_1(A, B) = \mu(A \cup B) - \mu(A \cap B).$$

Moreover, if μ is a strong measure then δ_1 is a metric.

An alternative distance measure, $\delta_2 : \Sigma \times \Sigma \to R_0^+$, is defined by

$$\delta_2(A,B) = \begin{cases} 0 & \text{if } A = B = \emptyset, \\ 1 - \frac{\mu(A \cap B)}{\mu(A \cup B)} & \text{otherwise.} \end{cases}$$

Then $0 \leqslant \delta_2(A,B) \leqslant 1$ since $0 \leqslant \mu(A \cap B) \leqslant \mu(A \cup B)$. Clearly

1. $\delta_2(A,A) = 0$,
2. $\delta_2(A,B) = 0$ implies $\mu(A \cap B) = \mu(A \cup B)$ and hence $\mu(A \setminus B) = \mu(B \setminus A) = 0$. Thus, if μ is strong, this implies $A \setminus B = B \setminus A = \emptyset$ and hence $A = B$,
3. $\delta_2(A,B) = \delta_2(B,A)$.

The triangle inequality is also satisfied as shown in the lemma below.

Lemma 1. *For any subsets, A, B, C of Σ,*

$$\delta_2(A,B) + \delta_2(B,C) \geqslant \delta_2(A,C).$$

Proof: Referring to Fig. 2 and assuming $S = a + b + c + d + e + f + g$ and $S - c \neq 0$ then

$$\delta_2(A,B) = 1 - \frac{f+g}{S-c}.$$

Similarly,

$$\delta_2(B,C) = 1 - \frac{d+g}{S-a}$$

provided $S - a \neq 0$ and

$$\delta_2(A,C) = 1 - \frac{e+g}{S-b}$$

provided $S - b \neq 0$.

Then, provided $(S-a)(S-b)(S-c) \neq 0$,

$$\begin{aligned} &\delta_2(A,B) + \delta_2(B,C) - \delta_2(A,C) \\ &= 1 - \frac{f+g}{S-c} + 1 - \frac{d+g}{S-a} - 1 + \frac{e+g}{S-b} \\ &= 1 - \frac{(f+g)(S-a)(S-b) + (d+g)(S-b)(S-c) - (e+g)(S-a)(S-c)}{(S-a)(S-b)(S-c)}. \end{aligned}$$

After some tedious algebra, the numerator of this expression evaluates to a sequence of terms that are all non-negative.

On the assumption that $(S-a)(S-b)(S-c) \neq 0$, the denominator is also positive, so we can deduce that $\delta_2(A,B) + \delta_2(B,C) - \delta_2(A,C) \geqslant 0$ and thus the triangle inequality holds.

The above argument relies on the assumption that $(S-a)(S-b)(S-c) \neq 0$. Now $(S-a)(S-b)(S-c) = 0$ iff one or more of $(S-a)$, $(S-b)$ or $(S-c)$ is zero iff at least two of the sets are empty. If $A = B = C = \emptyset$ then $\delta_2(A,B) = \delta_2(B,C) = \delta_2(A,C) = 0$ and the triangle inequality holds. If $A = B = \emptyset$ and $C \neq \emptyset$ then $\delta_2(A,C) = \delta_2(B,C) = 1$ and $\delta_2(A,B) = 0$ so the triangle inequality holds. The argument for the remaining cases are similar. The proof is thereby completed.

Hence, the following can be deduced.

Theorem 2. *The distance function* $\delta_2 : \Sigma \times \Sigma \to R_0^+$ *defined by*

$$\delta_2 = \begin{cases} 0 & \text{if } A = B = \emptyset, \\ 1 - \frac{\mu(A \cap B)}{\mu(A \cup B)} & \text{otherwise} \end{cases}$$

is a pseudometric on Σ *and, moreover, if* μ *is a strong measure then* δ_2 *is a metric.*

Proof: The pseudometric result follows from the lemma above and the preceeding observations. If μ is strong then

$$\begin{aligned} \delta_2(A, B) = 0 &\Rightarrow \mu(A \cup B) = \mu(A \cap B) \\ &\Rightarrow \mu(A \setminus B) = \mu(B \setminus A) = 0 \\ &\Rightarrow A \setminus B = B \setminus A = \emptyset \\ &\Rightarrow A = B \end{aligned}$$

and hence δ_2 is a metric.

4. Applications to aggregated data

In this section, we consider examples of aggregated data and show how measures can be defined and metrics deduced.

4.1. Finite sets

One of the most common examples of aggregated data is a set. Say a database has a field, F, with values that are categorical. Now, consider aggregating data from field F from n records, $r_1, r_2, \ldots, r_n$. The result may be a set of values taken by field F for these n records.

Given a finite set, S, the *cardinality* function, $\mu_c : 2^S \to Z \subset R$ is defined by

$$\mu_c(A) = |A|.$$

Clearly this is a finitely additive measure on the algebra 2^S and, moreover, it is a strong measure. Hence the following.

Corollary 1 *If* S *is a finite set then the following are both metrics on* 2^S*:*

1. $\delta_1^c(A, B) = |A \cup B| - |A \cap B| = |A \setminus B| + |B \setminus A|,$
2. $\delta_2^c = \begin{cases} 0 & \text{if } A = B = \emptyset, \\ 1 - \frac{|A \cap B|}{|A \cup B|} = \frac{|A \setminus B| + |B \setminus A|}{|A \cup B|}, & \text{otherwise.} \end{cases}$

The first metric is the usual metric for sets, the second is known as the Jaccard metric [16]. Both have been used to cluster sets; for example, in [20], sets of support for partial classification rules were clustered using δ_2^c in order to identify rules that were similar semantically and thereby to gain a better understanding of the data.

4.2. Finite sets of ordinals

Field values may be ordinal; if the values lie in a finite ordinal set, S, then there will be a function $\rho : S \to R^+$. This may be a simple ranking function whereby the i^{th} element of the set is assigned i or it may be a more sophisticated assignment. For example, DEGREECLASS may contain values from 1st, 2(i), 2(ii), 3rd, Pass, Fail and a simple ranking would assign these values to integers 1, 2, 3, 4, 5 and 6, respectively. An alternative assignment that perhaps better reflects their relative merit would be to assign each classification to the average of the marks in the span. Using a UK marking scheme, this might result in an assignment of 85, 65, 55, 45, 37, 17.5, respectively.

Let S be a finite set of ordinal data and $\rho : S \to R^+$ be an injection. Then $(S, 2^S)$ is an algebra and the *rank measure induced by* ρ is

$$\mu_\rho(A) = \sum_{x \in A} \rho(x).$$

Then μ_ρ is a finite measure and, since $\rho > 0$, $\mu_\rho(A) = 0$ only when $A = \emptyset$. Thus μ_ρ is also a strong measure and hence the following result.

Corollary 2 *If S is a finite ordinal set and $\rho : S \to R^+$ is an injection then the following are both metrics on 2^S:*

1. $$\delta_1^\rho(A, B) = \sum_{x \in A \cup B} \rho(x) - \sum_{x \in A \cap B} \rho(x) = \sum_{x \in A \setminus B} \rho(x) + \sum_{x \in B \setminus A} \rho(x),$$
2. $$\delta_2^\rho(A, B) = \begin{cases} 0 & \text{if } A = B = \emptyset, \\ 1 - \dfrac{\sum_{x \in A \cap B} \rho(x)}{\sum_{x \in A \cup B} \rho(x)} & \text{otherwise.} \end{cases}$$

   which is equivalent to

$$\delta_2^\rho(A, B) = \begin{cases} 0 & \text{if } A = B = \emptyset, \\ \dfrac{\sum_{x \in A \setminus B} \rho(x) + \sum_{x \in B \setminus A} \rho(x)}{\sum_{x \in A \cup B} \rho(x)} & \text{otherwise.} \end{cases}$$

4.3. Intervals

Let S be the interval of the real line, $[a, b]$ say, and let Σ denote all the finite sets of subintervals of $[a, b]$. A subinterval is either the empty set or may be open, closed or half open, i.e. of the form

1. $(c, d) = \{x \mid a \leqslant c < x < d \leqslant b\}$,
2. $[c, d) = \{x \mid a \leqslant c \leqslant x < d \leqslant b\}$,
3. $(c, d] = \{x \mid a \leqslant c < x \leqslant d \leqslant b\}$ or
4. $[c, d] = \{x \mid a \leqslant c \leqslant x \leqslant d \leqslant b\}$.

Then (S, Σ) is an algebra.

The width measure is defined on any interval, $I = (c, d), [c, d), (c, d]$, or $[c, d]$ by

$$\mu_w(I) = d - c.$$

Two intervals are said to be *disjoint* if their union is not itself an interval. A union of a finite number of intervals can clearly be expressed uniquely as a union of pairwise disjoint intervals. If A is an element

of Σ, i.e. a set of intervals in $[a, b]$, then $\hat{A}$ will represent the corresponding set of pairwise disjoint intervals.

The function μ_w can then be extended to elements of Σ in the obvious way by defining

1. $\mu_w(\Phi) = 0$, and
2. for any nonempty set of intervals $A \in \Sigma$, $\mu_w(A)$ is the sum of the widths of the pairwise disjoint intervals in $\hat{A}$, i.e.

$$\mu_w(A) = \sum_{I \in \hat{A}} \mu_w(I).$$

This measure is finite since for any set of intervals, A in $[a, b]$, $\mu_w(A) \leqslant b - a$. However, since $\mu_w\{[x, x]\} = 0$ for any $x \in [a, b]$, it is not a strong measure.

Corollary 3 *If $A, B \in \Sigma$ denote finite sets of intervals in $[a, b]$ then the following are both pseudometrics on Σ:*

1. $\delta_1^w(A, B) = \mu_w(A \cup B) - \mu_w(A \cap B)$,
2. $\delta_2^w(A, B) = \begin{cases} 0 & \text{if } A = B = \emptyset, \\ 1 - \frac{\mu_w(A \cap B)}{\mu_w(A \cup B)} & \text{otherwise.} \end{cases}$

4.4. Regions of the Euclidean Plain

Let S be a finitely bounded, closed region of R^2. S has a finite perimeter and contains all the points on the perimeter and in the region bounded by that perimeter. Let B denote all finitely bounded regions within S. An element of B is a subspace of S and will be contained by a perimeter but may or may not contain points on that perimeter, i.e. it may be be closed or open. Now, let Σ denote the closure of B under union and complement so that Σ is an algebra.

Every $A \in \Sigma$ has a finite area less than or equal to the finite area of S. The area of A, denoted by $\mu_a(A)$ provides a finitely additive measure on (S, Σ). It is not a strong measure since if A is an open region in S, i.e. does not contain its boundary, whilst $\bar{A}$ is the corresponding closed region, i.e. A together with its boundary, then $\mu_a(A) = \mu_a(\bar{A})$ although $A \neq \bar{A}$.

Area measures are of particular interest to analysts of aggregated data when applied to distributions and to histograms.

Let $I = [a, b]$ be an interval and let c be a positive real. Then $\mathcal{F}_{a,b,c}$ denotes the set of continuous functions on I bounded so that

$$\text{if } f \in \mathcal{F}_{a,b,c} \text{ then } 0 \leqslant f(x) \leqslant c \text{ for all } x \in I.$$

Any $f \in \mathcal{F}_{a,b,c}$ defines a region, X_f, bounded by the perimeter comprising four lines

1. $\{(x, y) \mid x = a,\ 0 \leqslant y \leqslant f(a)\}$,
2. $\{(x, y) \mid x = b,\ 0 \leqslant y \leqslant f(b)\}$,
3. $\{(x, y) \mid a \leqslant x \leqslant b,\ y = 0\}$,
4. $\{(x, y) \mid a \leqslant x \leqslant b,\ y = f(x)\}$

as in Fig. 3.

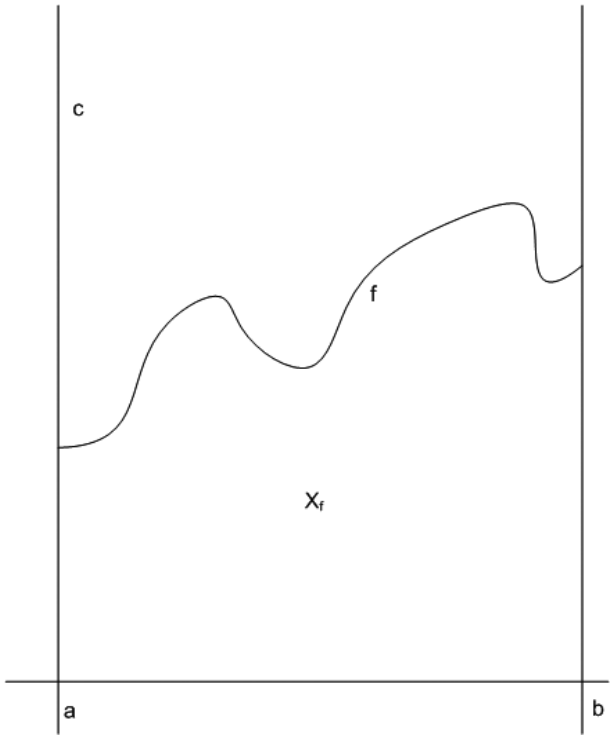

Fig. 3. The region X_f.

The area of the region X_f is then

$$\int_{x=a}^{b} f(x)dx.$$

By applying the area pseudometric, the following can be deduced.

Corollary 4 *The following are metrics on elements of* $\mathcal{F}_{a,b,c}$.

1. $\displaystyle\int_{x=a}^{b} |f(x) - g(x)| = \int_{x=a}^{b} (\max(f(x), g(x)) - \min(f(x), g(x))),$
2. $\displaystyle 1 - \frac{\int_{x=a}^{b} \min(f(x), g(x))}{\int_{x=a}^{b} \max(f(x), g(x))}$ providing f, g are not both everywhere 0.

Proof:

1. The integral simply gives the value of $\mu_a(X_f \setminus X_g) + \mu_a(X_g \setminus X_f)$ and hence, by Theorem 1, is a pseudometric. However,
$$\int_{x=a}^{b} |f(x) - g(x)| = 0 \Rightarrow f(x) = g(x)\ \forall x \in [a, b], \text{i.e.} f = g \text{ on } I$$
and hence the integral is also a metric.
2. This follows from Theorem 2 using a similar argument.

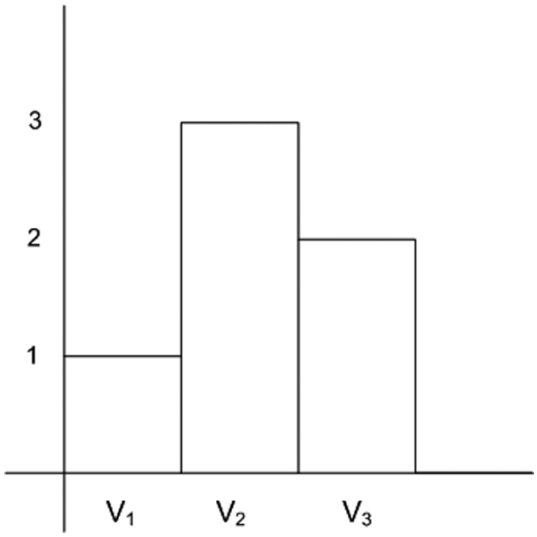

Fig. 4. Representing a simple histogram on $\{V_1, V_2, V_3\}$.

4.5. Histograms

Let F be a field of a database of n records that has been aggregated to produce a histogram. The field, F, may be nominal, ordinal or real-valued. Each case will be considered separately.

4.5.1. Histograms over nominal sets

In the nominal case, the possible values in F will be finite in number. Let V denote the set of values that are enumerated as $\{v_1, v_2, \ldots, v_m\}$. A histogram for F, H, over V is determined by

1. a partition of V into disjoint, nonempty, subsets, $V_1, V_2, \ldots, V_k$, $k \leqslant m$ and,
2. for each $1 \leqslant i \leqslant k$, a count, $c^H(V_i) \in Z_0^+$, of the number of occurrences of elements in V_i that occur in field F of the database.

Commonly, but not necessarily, each V_i is a singleton set. If H is such a histogram and $V_i = \{v_i\}$ then $c^H(\{v_i\})$ may be expressed as $c^H(v_i)$.

Note that in all cases

$$n = \sum_{k=1}^{m} c^H(V_i)$$

is the number of elements in the underlying database and this will be called the *base number* of the histogram.

There are two ways of representing a histogram, H over V, in R^2. The first provides equal width partitions of the x-axis for each of $V_1, V_2, \ldots, V_k$ and comprises a series of rectangles R_i, $1 \leqslant i \leqslant k$, where $R_i = [i-1, i) \times (0, c^H(V_i))$.

Thus, if F comprises the values 1,2,2,3,4,4 and $V_1 = \{1\}, V_2 = \{2, 3\}, V_3 = \{4\}$, the histogram can be represented as in Fig. 4.

However, it is also common to label the x-axis with the elements of V_1, followed by the elements of V_2, etc. Then a rectangle, R_i, is drawn for each V_i of width $|V_i|$ and height $\frac{c^H(V_i)}{|V_i|}$. For the above example and using the listing of V to be v_1, v_2, v_3, v_4, this results in the histogram of Fig. 5.

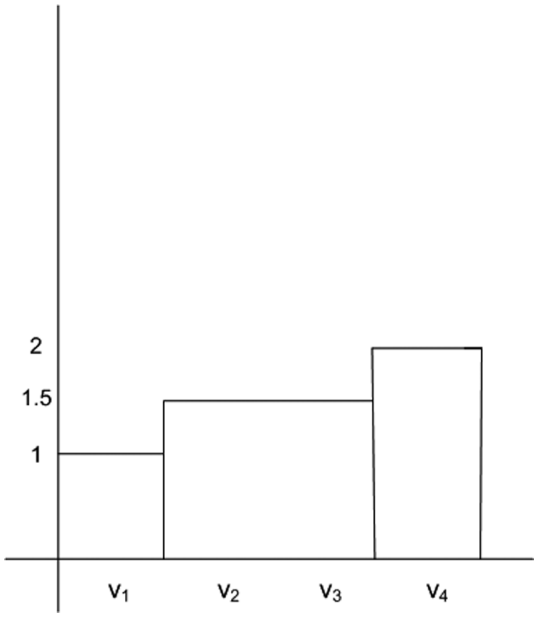

Fig. 5. Representing a simple histogram on $\{v_1, v_2, v_3, v_4\}$.

Note that, with either representation, the total area of the representation of the histogram is its base number.

Assume now that H_1 and H_2 are two distinct histograms over V_1 and V_2, respectively, where $|V_1| = m_1$ and $|V_2| = m_2$. These two histograms are to be compared.

If $V_1 \neq V_2$, then set $V = V_1 \cup V_2$ and regard each of H_1 and H_2 to be a histogram over V setting $c^{H_1}(V \setminus V_1) = 0$ if $V \setminus V_1 \neq \emptyset$ and, likewise, $c^{H_2}(V \setminus V_2) = 0$ if $V \setminus V_2 \neq \emptyset$. Let n_1 denote the base number of H_1 and n_2 denote the base number of H_2. The histograms are then scaled by setting $n = lcm(n_1, n_2)$, and multiplying $c^{H_1}(V_i)$ by n/n_1 and $c^{H_2}(V_i)$ by n/n_2. The two histograms are then over the same set (although not necessarily using the same partition of this set) and have the same base number, n.

For example, consider two histograms, H_1 and H_2, where

1. H_1 is defined over $\{1, 2, 3, 4\}$ and partitions this set into the three subsets $V_{11} = \{1\}, V_{12} = \{2, 3\}$ and $V_{13} = \{4\}$ with $c^{H_1}(V_{11}) = 1, c^{H_1}(V_{12}) = 3$ and $c^{H_1}(V_{13}) = 2$.
2. H_2 is defined over $\{1, 2, 3, 4, 5\}$ and partitions this set into the three subsets $V_{21} = \{1, 4\}, V_{22} = \{3, 5\}$ and $V_{23} = \{2\}$ with $c^{H_2}(V_{21}) = 2, c^{H_2}(V_{22}) = 4$ and $c^{H_2}(V_{23}) = 2$.

In this case, the base numbers of H_1 and H_2 are 6 and 8, respectively. Both can be regarded as histogramms over $\{1, 2, 3, 4, 5\}$ and their revised, scaled values are

1. $c_s^{H_1}(V_{11}) = 4, c_s^{H_1}(V_{12}) = 12, c_s^{H_1}(V_{13}) = 8, c_s^{H_1}(\{5\}) = 0$.
2. $c_s^{H_2}(V_{21}) = 6, c_s^{H_2}(V_{22}) = 12$ and $c_s^{H_2}(V_{23}) = 6$.

Any two histograms that are to be compared will thus be assumed to be over the same set, V, and both to have base number, n. Let H_1, H_2 denote two such histograms.

If two histograms, H_1 and H_2 are defined over the same partition of V into singleton sets, $V = \{v_1\} \cup \{v_2\} \ldots \cup \{v_m\}$ then the obvious metric to use to compare H_1 and H_2 is

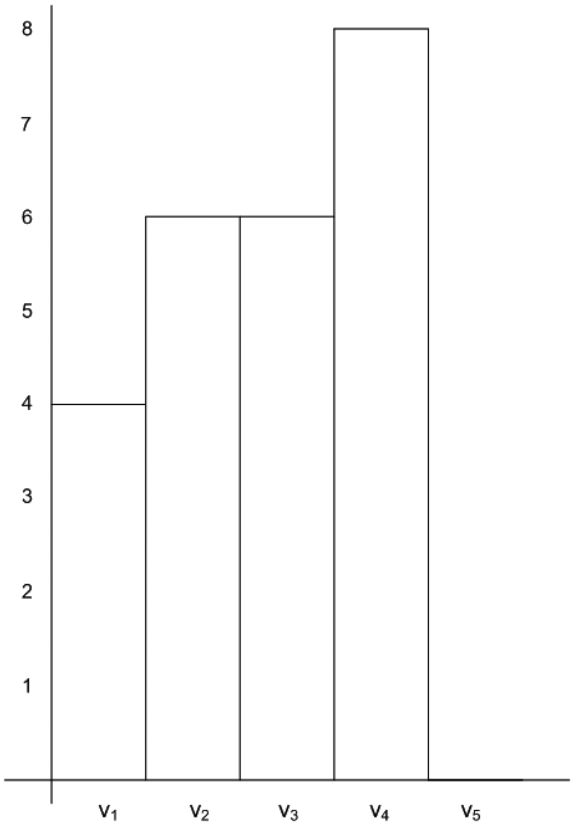

Fig. 6. Derived Representation of (scaled) H_1.

$$\delta(H_1, H_2) = \sum_{i=1}^{m} |c^{H_1}(v_i) - c^{H_2}(v_i)|.$$

However, when the histograms use different partitions, the metric is not so immediate but is an obvious generalisation. With respect to a histogram, H, over V, each $v \in V$ can be assigned a derived count value

$$d^H(v) = \frac{c^H(V_v^H)}{|V_v^H|},$$

where V_v^H is the set in the partition of H containing v. Then the following is clear.

Corollary 5 *One metric to compare H_1 and H_2 is simply*

$$\delta_1(H_1, H_2) = \sum_{i=1}^{m} |d^{H_1}(v_i) - d^{H_2}(v_i)|.$$

The two histograms, H_1 and H_2, can both be presented diagrammatically in R^2 where they both have an identically labelled x-axis, which will be some enumeration of V, $v_1, v_2, \ldots, v_m$. The *derived representation* of histogram $H_i (i = 1, 2)$ is constructed as follows. For each label, v_j, the rectangle

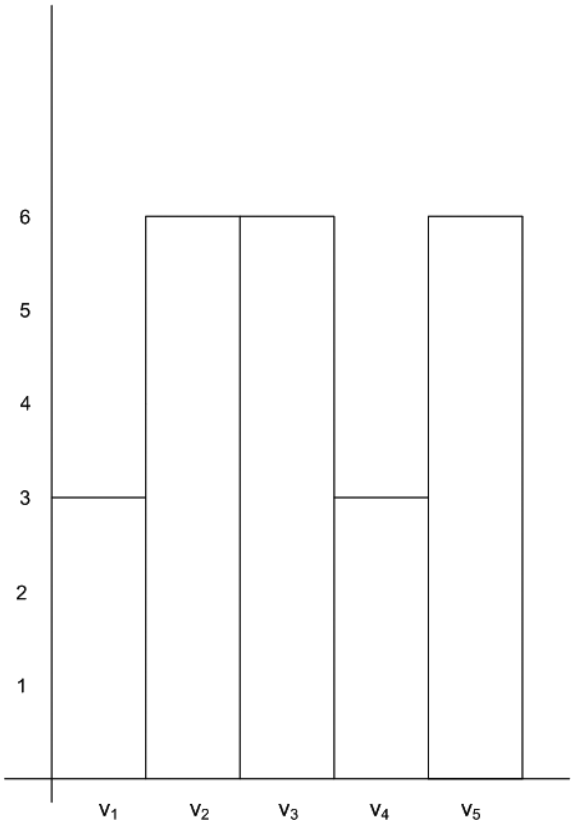

Fig. 7. Derived Representation of (scaled) H_2.

$[j-1, j] \times [0, d_i^H(v_j)]$ is drawn. Thus if H_1 and H_2 are the scaled histograms above, their derived representations are as in Figs 6 and 7. Note that the representation necessarily has area n.

The above metric then corresponds to the first metric that can be deduced using proposition 1 from the area measure applied to the two histograms viewed as regions of $[0, m] \times [0, n]$. A second metric then follows from Proposition 2.

Corollary 6

$$\begin{aligned}\delta_2(H_1, H_2) &= \frac{\sum_{i=1}^m |d^{H_1}(v_i) - d^{H_2}(v_i)|}{\sum_{i=1}^m \max(d^{H_1}(v_i), d^{H_2}(v_i))} \\ &= 1 - \frac{\sum_{i=1}^m \min(d^{H_1}(v_i), d^{H_2}(v_i))}{\sum_{i=1}^m \max(d^{H_1}(v_i), d^{H_2}(v_i))}\end{aligned}$$

is a metric. Proof:

$\sum_{i=1}^m |d^{H_1}(v_i) - d^{H_2}(v_i)|$ is the area of the symmetric difference of the two regions defined by H_1 and H_2 and is equal to $\sum_{i=1}^m \max(d^{H_1}(v_i), d^{H_2}(v_i)) - \sum_{i=1}^m \min(d^{H_1}(v_i), d^{H_2}(v_i))$. $\sum_{i=1}^m \max(d^{H_1}(v_i), d^{H_2}(v_i))$ is the area of the union of these two regions.

For our example histograms, the first metric provides a distance value of

$$|4-3|+|6-6|+|6-6|+|8-3|+|6-0|=12$$

and the second normalises this as

$$\frac{12}{4+6+6+8+6}=\frac{4}{10}.$$

Of course, it may not have been wise to share out the count of the number of occurrences of elements of a partition equally between the elements in that partition as was done with d^H. Since $d^H(v)$ may not be integer valued, there may be no possible database of n elements that could give rise to such a distribution. Also, given two histograms with different partitions but constructed from the same database, the above metrics constructed from d_1^H and d_2^H are quite unlikely to measure H_1 and H_2 as being distance zero apart.

One might argue that a more reasonable distance measure is to use alternative derived functions e^{H_1}, e^{H_2}, which are both integer valued and are such that

1. If the partition of H_1 is $V_{11}, V_{12}, \ldots, V_{1k_1}$ then
$$\sum_{v\in V_{1j}} e^{H_1}(v) = c^{H_1}(V_{1j}) \text{ for all } 1 \leqslant j \leqslant k_1.$$
2. If the partition of H_2 is $V_{21}, V_{22}, \ldots, V_{2k_2}$ then
$$\sum_{v\in V_{2j}} e^{H_2}(v) = c^{H_2}(V_{2j}) \text{ for all } 1 \leqslant j \leqslant k_2.$$
3. Subject to the above,
$$\delta_{\min}(H_1,H_2)=\sum_{i=1}^{m}|e^{H_1}(v_i)-e^{H_2}(v_i)|$$
is minimised.

Note, such a distance measure may not itself be a metric since it may not satisfy the triangle inequality. However, if two histograms are constructed from the same database, they will necessarily be distance zero apart as measured by $\delta_{\min}$. This distance measure can be computed using a maximum flow algorithm. A network is constructed as follows.

1. There is a source node labelled, S, and from this node, directed edges go to nodes labelled, $V_{11}, V_{12}, \ldots, V_{1k_1}$, where the arc from S to V_{1j} has capacity $c^{H_1}(V_{1j})$ for all $1 \leqslant j \leqslant k_1$.
2. There is a node labelled with each $v \in V$ and, for each node labelled V_{1j}, $1 \leqslant j \leqslant k_1$, there are directed edges to each $v \in V_{1j}$; these edges all have capacity $c^{H_1}(V_{1j})$.
3. There are nodes labelled, $V_{21}, V_{22}, \ldots, V_{2k_2}$ and, for each V_{2j}, $1 \leqslant j \leqslant k_2$, there are directed edges from each $v \in V_{2j}$; these edges all have capacity $c^{H_2}(V_{2j})$.
4. There is a sink node labelled T, and directed edges go to the node T from nodes labelled, $V_{21}, V_{22}, \ldots, V_{2k_2}$, where the arc from V_{2j} to T has capacity $c^{H_2}(V_{2j})$ for all $1 \leqslant j \leqslant k_2$.

As a simple example, consider two histograms H_1 and H_2, where

1. $V_{11}=\{a,b\}, V_{12}=\{c,e\}, V_{13}=\{d\}$,

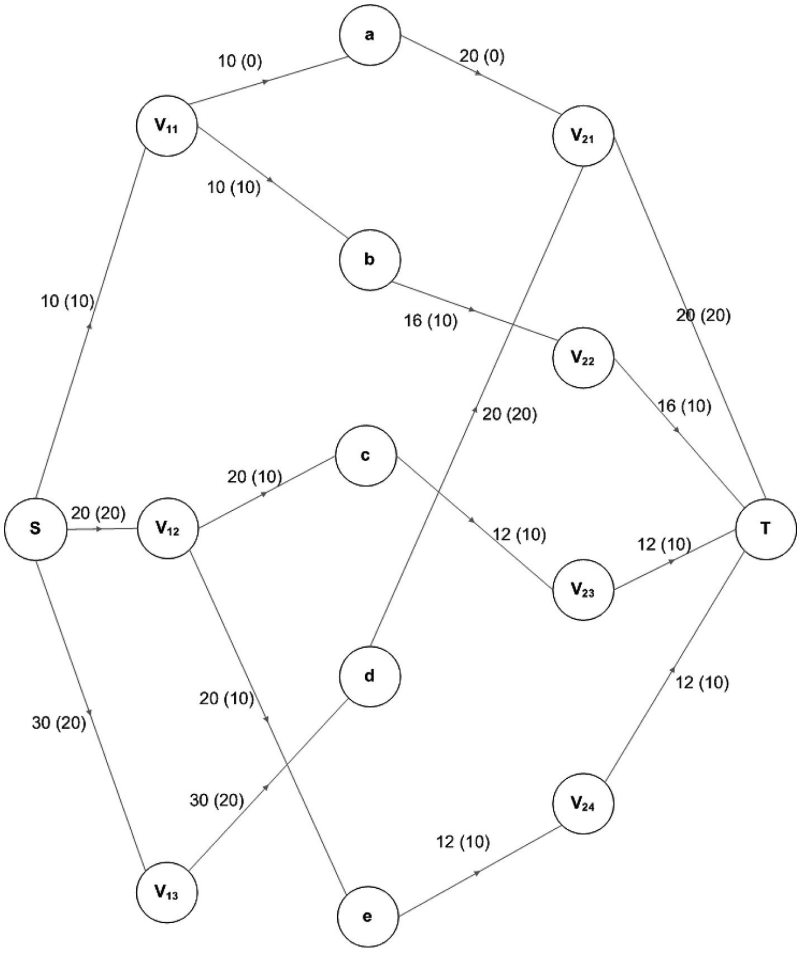

Fig. 8. Constructed network.

2. $V_{21} = \{a, d\}, V_{22} = \{b\}, V_{23} = \{c\}, V_{24} = \{e\}$,
3. $c^{H_1}(V_{11}) = 10, c^{H_1}(V_{12}) = 20, c^{H_1}(V_{13}) = 30$, and
4. $c^{H_2}(V_{21}) = 20, c^{H_2}(V_{22}) = 16, c^{H_2}(V_{23}) = c^{H_2}(V_{24}) = 12$.

The network constructed is then as in Fig. 8.

Let F denote the maximum flow that can be put through such a network for arbitrary H_1, H_2 from the source node to the sink node. This can be computed in $O(m^2)$ time using the well known Ford-Fulkerson maximum flow algorithm [18]. For the example of Fig. 8, F is 50. One possible maximum flow is given in parentheses alongside the arcs in Fig. 8 and it can be seen that this is a maximum flow since $\{S, V_{13}, d\}$ are separated from the remaining nodes by edges that are saturated.

Theorem 3. $\delta_{\min}(H_1, H_2) = 2(n - F)$.

Proof:

Consider the maximum flow, F and let the flow through the node labelled v be $f(v)$. For each V_{1j}, $1 \leqslant j \leqslant k_1$, if $\sum_{v \in V_{1j}} f(v) = c^{H_1}(V_{1j})$ then set $g^{H_1}(v) = f(v)$ for all $v \in V_{1j}$. Otherwise

$\sum_{v\in V_{1j}} f(v) < c^{H_1}(V_{1j}$ and then select an arbitrary element, $a_{1j} \in V_{1j}$, and assign $g^{H_1}(a_{1j}) = f(a_{1j}) + c^{H_1}(V_{1j}) - \sum_{v\in V_{1j}} f(v) > f(a_{1j}$, whilst setting $g^{H_1}(v) = f(v)$ for all $v \in V_{1j} \setminus \{a_{1j}\}$.

Similarly, For each V_{2j}, $1 \leqslant j \leqslant k_2$, if $\sum_{v\in V_{2j}} f(v) = c^{H_2}(V_{2j})$ then set $g^{H_2}(v) = f(v)$ for all $v \in V_{2j}$; otherwise select an arbitrary element, $a_{2j} \in V_{2j}$ and assign $g^{H_2}(a_{2j}) = f(a_{2j}) + c^{H_2}(V_{2j}) - \sum_{v\in V_{2j}} f(v)$ whilst setting $g^{H_2}(v) = f(v)$ for all $v \in V_{2j} \setminus \{a_{2j}\}$. Then,

$$\sum_{v\in V_{1j}} g^{H_1}(v) = c^{H_1}(V_{1j}) \text{ for all } 1 \leqslant j \leqslant k_1$$

and

$$\sum_{v\in V_{2j}} g^{H_2}(v) = c^{H_2}(V_{2j}) \text{ for all } 1 \leqslant j \leqslant k_2.$$

Note also that for all nodes labelled $v \in V$,

$$\min(g^{H_1}(v), g^{H_2}(v)) = f(v)$$

since if there is any node where that does not occur, the flow through that node can be increased. For each V_{1j}, if $\sum_{v\in V_{1j}} f(v) < c^{H_1}(V_{1j}$ then there is some arbitrary element of V_{1j} whose g^{H_1} value has been increased to take up the slack, viz. $C^{H_1}(V_{1j}) - \sum_{v\in V_{1j}} f(v)$. The total slack across all sets $V_{11}, V_{12}, \ldots, V_{1k_1}$ is $n - F$. This also applies to sets $V_{21}, V_{22}, \ldots, V_{2k_2}$ and hence $\sum_{i=1}^{m} |g^{H_1}(v_i) - g^{H_2}(v_i)| = 2(n - F)$.

All that is now needed to be established is that $\sum_{i=1}^{m} |e^{H_1}(v_i) - e^{H_2}(v_i)|$ is minimised by g. Say it was not and that there is some other choice of functions, $h^{H_1}(v_i), h^{H_2}(v_i)$ that satify $\sum_{v\in V_{1j}} h^{H_1}(v) = c^{H_1}(V_{1j})$ for all $1 \leqslant j \leqslant k_1$ and $\sum_{v\in V_{2j}} h^{H_2}(v) = c^{H_2}(V_{2j})$ for all $1 \leqslant j \leqslant k_2$ but where $\sum_{i=1}^{m} |h^{H_1}(v_i) - h^{H_2}(v_i)| < \sum_{i=1}^{m} |g^{H_1}(v_i) - g^{H_2}(v_i)|$.

Now, set $f'(v_i) = \min(h^{H_1}(v_i), h^{H_2}(v_i))$ and consider a flow, F', through the network where $f'(v_i)$ passes through node v_i. This will be a valid flow through the other nodes of the network as well. $\sum_{i=1}^{m} |h^{H_1}(v_i) - h^{H_2}(v_i)| \leqslant 2(n - F')$ so $2F' \geqslant 2n - \sum_{i=1}^{m} |g^{H_1}(v_i) - g^{H_2}(v_i)| > 2n - \sum_{i=1}^{m} |e^{H_1}(v_i) - e^{H_2}(v_i)| = 2F$ and this is a contradiction since F is a maximum flow. The theorem is thus established.

Returning now to the above simple example. If the derived values, d are used then $\delta(H_1, H_2) = |5 - 10| + |5 - 16| + |10 - 12| + |30 - 10| + |10 - 12| = 40$. However, constructing the flow network of Fig. 8, the maximum flow is found to be 50, comprising (say) a flow through a of 0, through b of 10, through c of 10, through d of 20 and through e of 10. Hence, $\delta_{\min}(H_1, H_2) = 2(60–50) = 20$. This could arise if H_1 was a histogram for a database with 0 occurrences of a, 10 occurrences of b, 10 occurrences of c, 30 occurrences of d and 10 occurrences of e and H_2 was a histogram for a database with 0 occurrences of a, 16 occurrences of b, 12 occurrences of c, 20 occurrences of d and 12 occurrences of e.

4.5.2. Histograms on ordinal sets

If a histogram, H, is over an ordinal field, F, with values in a finite ordered set, V, then there is an injective ranking function $\rho : V \to R^+$. A histogram is then based on a partitioning of V into subsets $V_1, V_2, \ldots, V_k$ for some $k > 1$. In the case where V is ordinal, $v \in V_i$ and $w \in V_j$ must satisfy $i < j \Leftrightarrow \rho(v) < \rho(w)$. The elements of V are assumed to be ordered by their ρ-value, i.e. $v_i < v_j \Leftrightarrow \rho(v_i) < \rho(v_j)$, and then, for each $1 \leqslant i \leqslant k$, $V_i = \{v_{l_i}, v_{l_i+1}, \ldots, v_{r_i}\}$, where

1. $v_{l_1} = v_1$ and $v_{r_k} = v_m$,
2. $v_{r_i+1} = v_{l_{i+1}}$ for $1 \leqslant i < k$.

A histogram over $V = V_1 \cup V_2 \ldots \cup V_k$ will then assign a count, $c^H(V_i) \in Z_0^+$ for each $1 \leqslant i \leqslant k$, of the number of occurrences of elements in V_i that occur in field F of the database.

As in Subsection 4.5.1, the assumption is made when comparing two histograms, H_1, H_2, on ordinal sets that both histograms have been scaled if necessary so that they both have the same base number, n, and are both defined over the same set, V.

The fact that V is ordered can be ignored and, if wished, V can be treated as nominal data. Hence the two metrics of Corollaries 5 and 6 can be used on histograms over ordinal sets. However, such metrics do not exploit the ordering; to do so, a cumulative histogram should be constructed. If H is a histogram over an ordinal set, the *cumulative histogram*, $\bar{H}$, corresponding to H has the same partition $V_1, V_2, \ldots, V_k$ as H but has count

$$c^{\bar{H}}(V_i) = \sum_{j=1}^{i} c^H(V_j).$$

For example, in Fig. 9, H_1 and H_2 are two histograms on an ordered set $V = \{v_1, v_2, \ldots, v_6\}$. H_1 partitions V into $\{v_1, v_2\}$, $\{v_3, v_4, v_5\}$ and $\{v_6\}$. H_2 partitions V into $\{v_1, v_2, v_3\}$, $\{v_4\}$ and $\{v_5, v_6\}$. $\bar{H}_1$ and $\bar{H}_2$ are their corresponding cumulative histograms.

If H is a histogram over an ordinal set $V = \{v_1, v_2, \ldots, v_m\}$ where V is partitioned into $V_1 \cup V_2 \ldots \cup V_k$ then, as for nominal data, a derived count can be computed for each $v_i \in V$,

$$d^H(v_i) = \frac{c^H(V_{f(i)})}{|V_{f(i)}|},$$

where $V_{f(i)}$ is the set in the partition that contains v_i. The *derived cumulative count* for $v_i, 1 \leqslant i \leqslant m$, is then

$$d_c^H(v_i) = \begin{cases} \frac{p_i}{|V_1|} c^{\bar{H}}(V_1) & \text{if } f(i) = 1, \\ c^{\bar{H}}(V_{f(i)-1}) + \frac{p_i}{|V_{f(i)}|} c^H(V_{f(i)}) & \text{otherwise,} \end{cases}$$

where v_i is the p_ith element of $V_{f(i)}$. A simple induction argument can be used to show that the following result holds.

Proposition 3 $d_c^H(v_i) = \sum_{j=1}^{i} d^H(v_j)$.

If H_1 and H_2 are histograms over an ordinal set, V, two new metrics can be deduced by applying Corollaries 5 and 6 to the histograms H_1' and H_2', both with partition $\{\{v_1\}, \{v_2\}, \ldots, \{v_m\}\}$ and with counts $d_c^{H_1}(v_i), d_c^{H_2}(v_i)$, respectively. This gives the following result.

Corollary 7 *If H_1 and H_2 are histograms over an ordinal set, $V = \{v_1, v_2, \ldots, v_m\}$, then the following are metrics:*

1.

$$\delta_3(H_1, H_2) = \sum_{i=1}^{m} |d_c^{H_1}(v_i) - d_c^{H_2}(v_i)|$$

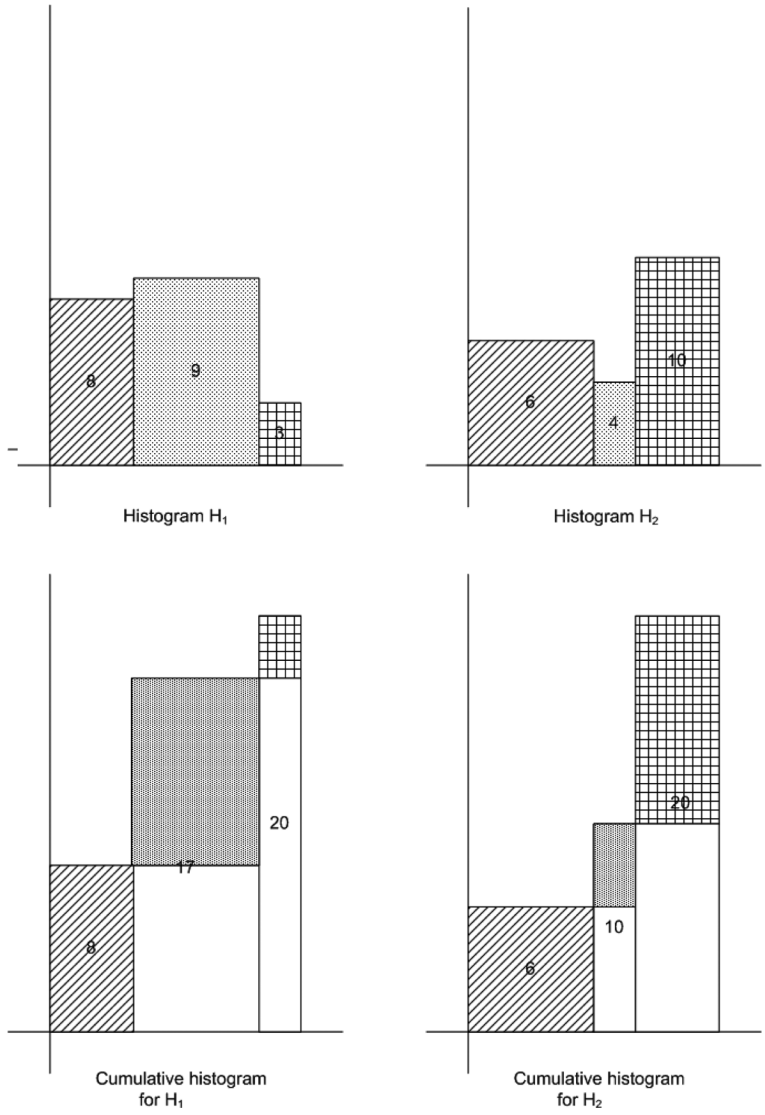

Fig. 9. Two histograms and their correponding cumulative histograms.

$$= \sum_{i=1}^{m} |\sum_{j=1}^{i} d^{H_1}(v_j) - \sum_{j=1}^{i} d^{H_2}(v_j)|.$$

2.

$$\delta_4(H_1, H_2) = 1 - \frac{\sum_{i=1}^{m} \min(d_c^{H_1}(v_i), d_c^{H_2}(v_i))}{\sum_{i=1}^{m} \max(d_c^{H_1}(v_i), d_c^{H_2}(v_i))}$$
$$= 1 - \frac{\sum_{i=1}^{m} \min(\sum_{j=1}^{i} d^{H_1}(v_j), \sum_{j=1}^{i} d^{H_2}(v_j))}{\sum_{i=1}^{m} \max(\sum_{j=1}^{i} d^{H_1}(v_j), \sum_{j=1}^{i} d^{H_2}(v_j))}.$$

As an example, consider the two histograms, H_1 and H_2 of Fig. 9. The derived cumulative counts for H_1 for the 6 elements v_1, v_2, v_3, v_4, v_5 and v_6 are 4, 8, 11, 14, 17, 20, respectively and for H_2 they are 2, 4, 6, 10, 15, 20. Hence

$$\delta_3(H_1, H_2) = 2 + 4 + 5 + 4 + 2 + 0 = 17$$

and

$$\delta_4(H_1, H_2) = 1 - \frac{2 + 4 + 6 + 10 + 15 + 20}{4 + 8 + 11 + 14 + 17 + 20} = \frac{17}{74}.$$

4.5.3. Histograms on intervals of the real line

If a field, F, is real-valued with values in the range $I = (a, b]$, a histogram, H, over I then comprises

1. a partition of the interval I into subintervals $I_1 = (l_1, r_1]$, $I_2 = (l_2, r_2], \ldots, I_k = (l_k, r_k]$ where $l_1 = a, r_k = b$ and $r_i = l_{i+1}$ for $1 \leqslant i \leqslant m - 1$,
2. for each interval, I_i, $1 \leqslant i \leqslant k$, a count $c^H(I_i) \in Z_0^+$ of the number of occurrences of elements of F that lie in I_i.

Consider two interval histograms, H_1 over $(a, b]$, and H_2 over $(c, d]$. Both can be regarded as acting over the same interval, $(\min(a, c), \max(b, d)]$, by setting

1. $c^{H_1}(c, a] = 0$ if $c < a$ and $c^{H_2}(a, c] = 0$ if $a < c$,
2. $c^{H_1}(b, d] = 0$ if $b < d$ and $c^{H_2}(d, b] = 0$ if $d < b$.

By applying a scaling function

$$x \mapsto \frac{x - \min(a, c)}{\max(b, d) - \min(a, c)},$$

it can then be assumed that both histograms are over $(0, 1]$. This will be assumed to have been done for any histograms that are to be compared. Moreover, it will be assumed that the base number, n, is also the same.

For any histogram, H over $(0, 1]$, the *cumulative distribution function* associated with H is defined as a continuous line from $(0, 0)$ to $(1, n)$ such that,

1. over the segment, $I_1 = (0, r_1]$, it corresponds to the straight line joining $(0, 0)$ to $(r_1, c^H(I_1))$, and
2. for $1 < i \leqslant k$ over the segment, $I_i = (l_i, r_i]$, it corresponds to the straight line joining $(r_{i-1}, \sum_{j=1}^{i-1} c^H(I_j))$ to $(r_i, \sum_{j=1}^{i} c^H(I_j))$.

The cumulative distribution function associated with H will be denoted by f_H.

Figure 10 gives two histograms over (0,1] together with their cumulative distibution functions. H_1 partitions $(0, 1]$ into $(0, \frac{1}{2}]$ and $(\frac{1}{2}, 1]$; H_2 partitions $(0, 1]$ into $(0, \frac{1}{3}]$, $(\frac{1}{3}, \frac{2}{3}]$ and $(\frac{2}{3}, 1]$; $c_{H_1}(0, \frac{1}{2}] = 10$, $c_{H_1}(\frac{1}{2}, 1] = 6$, $c_{H_2}(0, \frac{1}{3}] = 8$, $c_{H_2}(\frac{1}{3}, \frac{2}{3}] = 2$, $c_{H_2}(\frac{2}{3}, 1] = 6$.

Corollary 8 *If H_1, H_2 are histograms over the interval (0,1] then the following are metrics*

1.

$$\begin{aligned}\delta_1^I(H_1, H_2) &= \int_0^1 |f_{H_1}(x) - f_{H_2}(x)| \\ &= \int_0^1 (\max(f_{H_1}(x), f_{H_2}(x)) - \min(f_{H_1}(x), f_{H_2}(x))),\end{aligned}$$

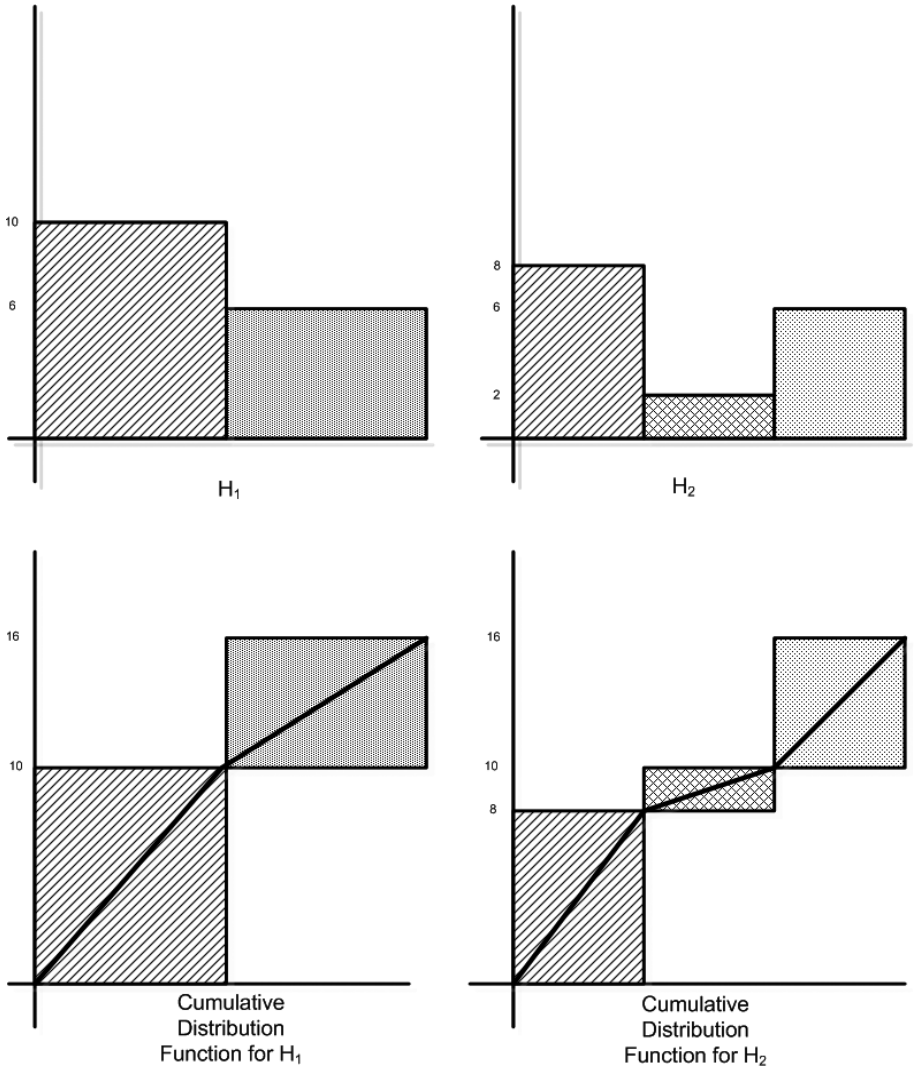

Fig. 10. Two histograms over $(0, 1]$.

2.

$$\delta_2^I(H_1, H_2) = 1 - \frac{\int_0^1 \min(f_{H_1}(x), f_{H_2}(x))}{\int_0^1 \max(f_{H_1}(x), f_{H_2}(x))}.$$

Proof: Both results follow immediately from Corollary 4. The first of these metrics is known as the *Wasserstein metric* and is the usual metric for comparing histograms on intervals, see, e.g. [5,14].

To compute $\delta_1^I(H_1, H_2)$, the points of intersections of $f_{H_1}(x)$ and $f_{H_2}(x)$ need to be found and then the integral is simply the sum of the differences of areas of trapezia. For example, referring to Fig. 10, the functions $f_{H_1}(x)$ and $f_{H_2}(x)$ are superimposed in Fig. 11. These two lines only intersect at a single point in $(0, 1]$ other than $(1, 1)$, i.e. the point of intersection of the line joining $(0, 0)$ with $(\frac{1}{2}, 10)$ with the line joining $(\frac{1}{3}, 8)$ and $(\frac{2}{3}, 10)$, viz. $(\frac{2}{7}, \frac{4}{7})$. The Wasserstein metric in this case can be computed by computing the difference between the areas of two trapezia in each of the five regions shown. In general,

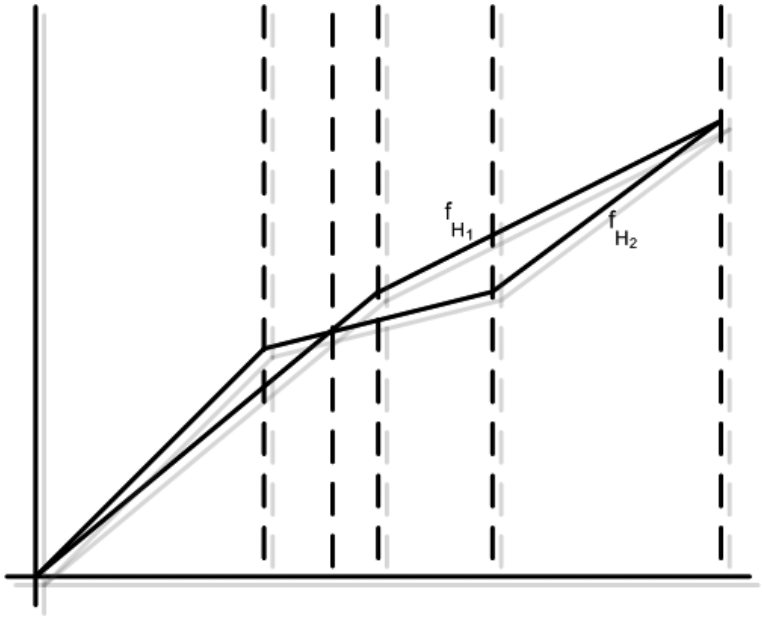

Fig. 11. The intersection of f_{H_1} and f_{H_2}.

if H_1 has k_1 intervals and H_2 has k_2 intervals, the number of intersection points is at most $\min(k_1, k_2)$. The Wasserstein metric can thus be computed in $O(k^3)$ time where $k = \max(k_1, k_2)$.

5. Conclusions and topics for further research

Two pseudometrics, one of which is normalised, have been shown to exist on an algebra, (S, Σ), over which a finitely additive measure, μ, is defined. Provided the measure is strong, both of these pseudometrics have been shown to be full metrics. The first of these metrics is known in the measure theory literature. The normalised version appears to be new.

From these results, metrics or pseudometrics have been deduced for aggregated data in the form of sets, intervals and histograms. Neither of the metrics deduced for nominal sets is new but the second metrics for ordinal sets does appear to be new. The first width metric for intervals is known but again the normalised version has not been found in the literature.

With histograms, it is important to distinguish between histograms over nominal sets, over ordinal sets and over intervals of the reals. The metrics discussed here act on histograms that do not necessarily assume the base set has been partitioned in the same way in both of the two histograms being compared. For histograms over nominal sets, Corollary 6 gives a novel normalised metric. Theorem 3 provides a lower bound on the similarity of two histograms and is new. Finding a similar bound in the ordinal case is an open problem. Of the two metrics for histograms over ordinal sets given in Corollary 7, the first is the obvious one and, once again, it is the second, normalised metric that appears to be new. This is also the case for histograms over intervals of the real line where the first metric is the Wasserstein metric and the normalised metric appears to be new.

Aggregated data arises following a summarisation process of large databases and may be used as a way of hiding sensitive information on individuals or may be purely part of an analysis process. Important,

strategic planning decisions can result from the comparison of groups described by aggregated data and a key step in this is to define metrics to measure the difference between aggregated data items relating to two different groups. Normalised metrics have an obvious appeal and, in this paper, a unified theory and notation has been developed from which they can be deduced. Given two groups of individuals, each may have a number of fields, each describing aggregated data. The difference between any two fields can be now be measured but how these differences are best combined to produce a fair and honest, single measure of the difference betwreen the two groups remains a topic for research.

Acknowledgment

Daniel Sandoval Izarraras provided much of the motivation for this research and discussed some of the underlying ideas with the author.

References

[1] K.B. Athreya and S.N. Lahiri, *Measure Theory and Probability Theory*, Springer-Verlag, New York, USA, 2006.
[2] Lynne Billard and Edwin Diday, *Symbolic Data Analysis: Conceptual Statistics and Data Mining* (*Wiley Series in Computational Statistics*), John Wiley & Sons, 2007.
[3] Hans Hermann Bock and E. Diday, *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2000.
[4] Francisco De Carvalho, Paula Brito and Hans-Hermann Bock, Dynamic clustering for interval data based on l2 distance, *Comput Stat* **21**(2) (2006), 231–250.
[5] T.F. Chan, S. Esedoglu and K. Ni, Histogram based segmentation using Wasserstein distances. In *Scale Space and Variational Methods in Computer Vision, Springer Lecture Notes in Computer Science 4485*, 2007, pages 697–708.
[6] Marie Chavent, Francisco Carvalho, Yves Lechevallier, and Rosanna Verde. New clustering methods for interval data, *Computational Statistics* **21**(2) (June 2006), 211–229.
[7] Francisco de A.T de Carvalho, Renata M.C.R. de Souza, Marie Chavent and Yves Lechevallier, Adaptive hausdorff distances and dynamic clustering of symbolic interval data, *Pattern Recogn Lett* **27**(3) (2006), 167–179.
[8] Renata M.C.R. de Souza and Francisco de A.T. de Carvalho, Clustering of interval data based on city-block distances, *Pattern Recognition Letters* **25**(3) (2004), 353–365.
[9] V. Ganti, R. Ramakrishnan, J. Gehrke, A. Powell and J. French, Clustering large datasets in arbitrary metric spaces, *International Conference on Data Engineering*, 1999.
[10] Antonio Giusti and Laura Grassini, Cluster analysis of census data using the symbolic data approach, *Advances in Data Analysis and Classification* **2**(2) (2008), 163–176.
[11] K. Chidananda Gowda and T.V. Ravi, Agglomerative clustering of symbolic objects using the concepts of both similarity and dissimilarity, *Pattern Recognition Letters* **16**(6) (1995), 647–652.
[12] D.S. Guru, Bapu B. Kiranagi and P. Nagabhushan, Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns, *Pattern Recognition Letters* **25**(10) (2004), 1203–1213.
[13] P.R. Halmos, *Measure Theory*, van Nostrand, New York, USA, 1968.
[14] A. Irpino, R. Verde and Y. Lechevallier, Dynamic clustering of histograms using Wasserstein metric. In *17th COMPSTAT Symposium of the IASC*, 2006.
[15] Antonio Irpino and Rosanna Verde, Dynamic clustering of interval data using a wasserstein-based distance, *Pattern Recogn Lett* **29**(11) (2008), 1648–1658.
[16] P. Jaccard, Étude comparative de la distribution florale dans une portion des Alpes et des Jura, *Bulletin de la Société Vaudoise de la Sciences Naturelles* **37** (1901), 547–579.
[17] A.K. Jain, M.N. Murty and P.J. Flynn, Data clustering: a review, *ACM Comput Surv* **31**(3) (1999), 264–323.
[18] L.R. Ford, Jr. and D.R. Fulkerson, Maximal flow through a network, *Canadian Journal of Mathematics* **8** (1956), 399–404.
[19] Q.H. Nguyen and V.J. Rayward-Smith, Internal quality measures for clustering in metric spaces, *International Journal of Business Intelligence and Data Mining* **3**(1) (2008).
[20] A. Reynolds, G. Richards, B. de laIglesia and V.J. Rayward-Smith, Clustering rules: A comparison of partitioning and hierarchical clustering algorithms, *Journal of Mathematical Modelling and Algorithms* **5**(4) (2006), 475–504.