Information Sciences 178 (2008) 4644–4655

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A measure of variance for hierarchical nominal attributes

Josep Domingo-Ferrer *, Agusti Solanas

Universitat Rovira i Virgili, UNESCO Chair in Data Privacy, Department of Computer Engineering and Maths, Av. Päisos Catalans 26, E-43007 Tarragona, Catalonia, Spain

ARTICLE INFO

Article history:
Received 11 January 2008
Received in revised form 3 July 2008
Accepted 7 August 2008

Keywords:
Variance
Spread
Diversity
Nominal attributes
Categorical hierarchy
Clustering
Data privacy
Data anonymization

ABSTRACT

The need for measuring the dispersion of nominal categorical attributes appears in several applications, like clustering or data anonymization. For a nominal attribute whose categories can be hierarchically classified, a measure of the variance of a sample drawn from that attribute is proposed which takes the attribute's hierarchy into account. The new measure is the reciprocal of "consanguinity": the less related the nominal categories in the sample, the higher the measured variance. For non-hierarchical nominal attributes, the proposed measure yields results consistent with previous diversity indicators. Applications of the new nominal variance measure to economic diversity measurement and data anonymization are also discussed.

1. Introduction

An attribute in a data set can be classified as being either numerical or categorical, depending on whether its values can be operated using standard arithmetic or not. In turn, a categorical attribute can be either ordinal or nominal, depending on whether an order relationship is defined within the set of its possible values or not. As an example, consider a data set related to individuals containing a numerical attribute "Weight", an ordinal attribute "Instruction level" and a nominal attribute "Favorite sport".

The statistical treatment of categorical attributes without underlying numerical attributes in an under-researched topic. In [13], a regression procedure was proposed for ordinal attributes which had no obvious numerical mapping. In [4], weighted arithmetic means for ordinal attributes are introduced. In [6], Mahalanobis-type distances are defined for ordinal and numerical attributes. For other recent investigations on the treatment of ordinal attributes see [2,5]. In this paper, we focus on the problem of measuring the dispersion or spread of a sample of a hierarchical nominal attribute, that is, a nominal attribute whose categories are drawn from a hierarchical classification. Even if this problem is seldom addressed in common statistics or data analysis, it is of substantial concern in several important applications, such as the following:

- *Measurement of biodiversity*: Given a sample of plants or animals observed in an ecosystem, measuring their diversity over time is of paramount importance to decide whether the ecosystem is degrading and protective action should be taken. Hierarchical classifications exist for plants and animals, so that biodiversity can be regarded as the variance of a hierar-

[*] Corresponding author. Tel.: +34 977558109; fax: +34 977559710.
E-mail addresses: josep.domingo@urv.cat (J. Domingo-Ferrer), agusti.solanas@urv.cat (A. Solanas).

0020-0255/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2008.08.003

chical nominal attribute. Surprisingly enough, though, biodiversity indicators in use tend to ignore the available hierarchical information (e.g. [1,16]).

- *Measurement of economic diversity*: Diversity in the economic activities in a country is often regarded as an asset by economists. Several hierarchical classifications for economic activities exist (e.g. NACE, [7]), so that economic diversity can also be modeled as a hierarchical nominal variance. This application is illustrated in Section 5.1.
- *Database privacy*: A common approach to privacy-preserving release of individual data is to suppress identifier attributes and perturb key attributes (those which could be used to link released records with specific individuals) so that any intruder wishing to determine the released record corresponding to a specific target individual (re-identification) is confronted to a group of equally likely records. This approach offers protection as long as the values of all confidential attributes in the records of that group are diverse enough: indeed, if the intruder finds that the record corresponding to patient John Smith belongs to a group where all patients have the confidential attribute ‘‘Disease" equal to ‘‘AIDS", then the intruder knows that John Smith suffers from AIDS even if she cannot determine John Smith's precise record. Therefore, under this privacy-preserving approach, the diversity of confidential attributes is a privacy metric. In the previous example, measuring the diversity of the ‘‘Disease" attribute amounts to computing a hierarchical nominal variance (diseases can be hierarchically classified). This application is discussed in greater detail in Section 5.2.

1.1. Contribution and plan of this article

The contribution of this article is a variance measure for the spread of a sample drawn from a hierarchical nominal attribute which effectively captures the semantics of the hierarchy of the attribute (unlike previous nominal spread measures). Section 2 contains background on nominal spread measures. Section 3 describes the proposed variance measure. Section 4 shows that the new measure is well-defined and reflects the spread of nominal values. Section 5 deals with two applications of the proposed measure: a case study on economic diversity measurement and a discussion of the role of the nominal variance in database anonymization. Section 6 is the conclusion.

2. Background on spread measures for nominal attributes

Measures for the spread or diversity of nominal attributes were developed in the context of natural sciences from the 1940s onwards. Among the earliest proposals, Simpson's measure [11] deserves mention. Given a sample with N nominal values, Simpson's diversity measure l is defined as

$$l = \frac{\sum_{i=0}^{Z} n_i(n_i - 1)}{N(N-1)},$$

where Z is the number of different categories to which the N nominal values belong, and n_i is the number of values belonging to the ith category. Simpson's measure can be simply understood as the probability that two elements chosen at random and independently from the sample will be found to belong to the same category. This measure is a ‘‘flat" one, in the sense that it does not take into account any hierarchical relationships that may exist between categories.

In [16], as many as 18 diversity indices for biological application (some of which based on information-theoretic measures) are compared and the author concludes that Simpson's measure is one of the three best, even if it was already 35 years old when [16] was written. A common feature of the 18 considered measures is that all of them are flat ones, that is, they do not make use of any available hierarchical information. More recent sources, like [1], still only consider flat measures.

If spread measures arisen in the context of natural sciences do ignore hierarchy, measures devised by computer scientists working in data privacy are not much better. The latter typically tend to ignore the semantics of the nominal values of the attribute. For example, a measure as simple as the number of different nominal values that appear in the sample is often employed [15,8]. Such a rough spread estimation has at least two shortcomings:

- It fails to detect low dispersion due to some of the nominal values appearing with much higher frequency than others. Clearly, the spread of a sample of size 1000 with 100 different nominal values appearing 10 times each is larger than the spread of a sample of size 1000 with 100 different nominal values one of which appears 901 times and the rest of which appear only once each.
- It fails to detect low dispersion due to possible similarity between the various nominal values in a sample. Arguably, there is more disease dispersion in a sample of 1000 patients whose diseases are uniformly distributed in the set

$$\{\text{“Cold”}, \text{“Stomach ache”}, \text{“Cancer”}, \text{“Broken leg”}\}$$

than in a sample of 1000 patients whose diseases are uniformly distributed in a range of 15 possible cancer types.

The use of spread measures based on entropy (e.g. Shannon's entropy, [10]) has been proposed to overcome the first shortcoming listed above (the one related to frequencies, see for example, [8]). However, entropy still does not detect low dispersion due to similarity. Attempts at taking semantics into account when measuring dispersion are scarce and largely dependent on subjective criteria which cannot be easily automated (e.g. [14]).

3. A measure of nominal variance

The minimum requirement to define a semantic variance for a sample drawn from a nominal attribute is that the attribute categories can be hierarchically classified. Let this hierarchy be expressed as a tree, with the root being the top-level category (e.g. "Disease" for an attribute "Disease"), the children being general categories (e.g. "Cold", "Cancer", etc., for an attribute "Disease"), the grand-children being more specific categories (e.g. "Nose cold", "Chest cold", "Lung cancer", "Liver cancer", etc.) and so on.

Intuitively, a set of nominal values belonging to categories which are all children of the same parent category (flat hierarchy or no hierarchy) has smaller variance that a set with children from different parents. The less "consanguinity" between a set of nominal values, i.e. the less related they are, the more variance. To capture this notion, we use the following procedure to compute the variance of a sample drawn from a nominal attribute.

Algorithm 1. Nominal variance computation

(1) Let the hierarchy of categories of a nominal attribute X be such that b is the maximum number of children that a parent category can have in the hierarchy.

(2) Given a sample T_X of nominal categories drawn from X, place them in the tree representing the hierarchy of X. If there are repeated values, there will be several nominal values associated to one or more leaves (lowest-level categories) in the tree. Consider the edges of the tree leading to leaves with no value in the sample associated to them, and prune those edges and leaves.

(3) Label as follows the edges remaining in the tree from the root node to each of its children:
- If b is odd, consider the following succession of labels $l_0 = (b-1)/2, l_1 = (b-1)/2 - 1, l_2 = (b-1)/2 + 1, l_3 = (b-1)/2 - 2, l_4 = (b-1)/2 + 2, \ldots, l_{b-2} = 0, l_{b-1} = b-1$.
- If b is even, consider the following succession of labels $l_0 = (b-2)/2, l_1 = (b-2)/2 + 1, l_2 = (b-2)/2 - 1, l_3 = (b-2)/2 + 2, l_4 = (b-2)/2 - 2, \ldots, l_{b-2} = 0, l_{b-1} = b-1$.
- Label the edge leading to the child with most categorical values associated to its descendant *leaves* (lowest-level categories) as l_0, the edge leading to the child with the second highest number of categories associated to its descendant leaves as l_1, the one leading to the child with the third highest number of categories associated to its descendant leaves as l_2 and, in general, the edge leading to the child with the ith highest number of categories associated to its descendant leaves as l_{i-1}. Since there are at most b children, the set of labels $\{l_0, \ldots, l_{b-1}\}$ should suffice. Thus an *edge label* can be viewed as a b-ary digit (to the base b).

(4) Recursively repeat Step 3 taking instead of the root node each of the root's child nodes.

(5) Assign to values associated to each leaf in the hierarchy a *leaf label* consisting of a b-ary number constructed from the edge labels, more specifically as the concatenation of the b-ary digits labeling the edges along the path from the root to the leaf: the label of the edge starting from the root is the most significant one and the edge label closest to the leaf is the least significant one.

(6) Let L be the maximal length of the leaf b-ary labels. Append as many ~~zeroes~~ as needed in the least significant positions to the shorter labels so that all of them eventually consist of L digits.

(7) Let $T_X(0)$ be the set of b-ary digits in the least significant positions of the leaf labels (the "units" positions); let $T_X(1)$ be the set of b-ary digits in the second least significant positions of the leaf labels (the "tens" positions), and so on, until $T_X(L-1)$ which is the set of digits in the most significant positions of the leaf labels.

(8) Compute the variance of the sample as

Erratum: "zeroes" should read "l_0 digits"

$$\mathrm{Var}(T_X) = \mathrm{Var}(T_X(0)) + b^2 \cdot \mathrm{Var}(T_X(1)) + \cdots + b^{2(L-1)} \cdot \mathrm{Var}(T_X(L-1)). \tag{1}$$

The rationale of expression (1) is to compute the variance of the hierarchical nominal attribute sample as a combination of the variances of the numerical labels assigned for each level of the hierarchy, in such a way that the variance for the higher levels is given more weight in the combination. The idea is that diversity at the higher levels of the hierarchy should weigh more than diversity at the lower levels. A more accurate analysis of the validity and consistency of the above variance expression is given in Section 4.

We next give an example to illustrate the application of the above-defined nominal variance.

3.1. Example of use

This subsection 3.1 should be corrected according to the erratum marked above for Algorithm 1, Step 6

Assume a nominal attribute "*Disease*", for which a sample is available whose nominal values can be hierarchically classified as shown in Fig. 1. The classification is not accurate in medical terms and is only meant for illustration. The hierarchy has been pruned, so that only leaves with some value in the sample are depicted. The sample consists of two patients with nose cold and one patient for each of the remaining leaves. We will now show step by step how Algorithm 1 is applied to compute the variance of this sample.

For convenience of representation, we map disease names to numerical identifiers as described in Table 1. Identifiers are assigned by traversing the hierarchy from top to bottom and from left to right.

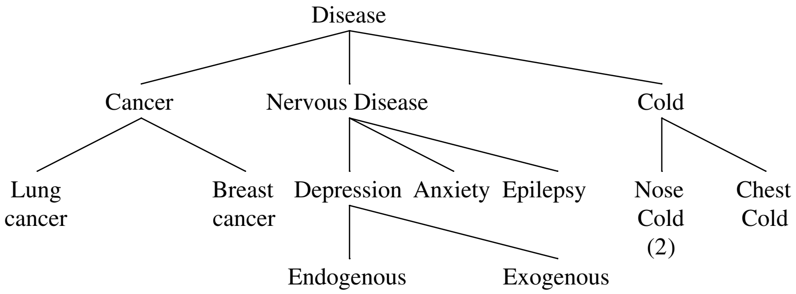


Fig. 1. Example pruned hierarchy of a first sample of the "*Disease*" attribute.

Table 1
Mapping of disease names to numerical identifiers

Disease name	ID
Disease	0
Cancer	1
Nervous disease	2
Cold	3
Lung cancer	4
Breast cancer	5
Depression	6
Anxiety	7
Epilepsy	8
Nose cold	9
Chest cold	10
Endogenous depression	11
Exogenous depression	12

In Step 1 of Algorithm 1 the maximum number b of children that a parent node can have is determined. In the example hierarchy of diseases, we assume that before pruning (that is, whatever the sampled diseases) a parent disease can at most have three child diseases, i.e. $b = 3$.

Step 2 (hierarchy tree pruning) can be skipped because we started with a pruned hierarchy.

Steps 3 and 4 label with a b-ary digit each edge in the tree; since we have $b = 3$ in our example, edge labels will be ternary digits. The left-hand side of Fig. 2 shows the labeling of the first level of the hierarchy. The edge leading from node 0 to its child node with most patients associated to its descendant leaves (that is, node 2, to which four patients are associated) is labeled with $l_0 = (b-1)/2 = 1$. The edge leading to the child node with the second highest number of patients associated to its descendant leaves (that is, node 3, to which three patients are associated) is labeled with $l_1 = (b-1)/2 - 1 = 0$. Finally, the remaining edge (leading to node 1, to which two patients are associated) is labeled with $l_2 = (b-1)/2 + 1 = 2$. Labels in the remaining levels of the hierarchy are assigned in a similar fashion. The complete edge labeling is shown in the right-hand side of Fig. 2.

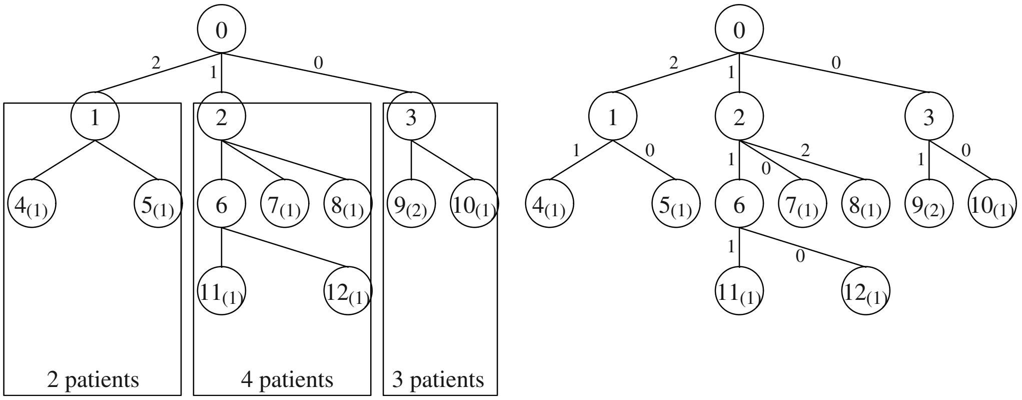


Fig. 2. Left: Edge labeling for the first level of the hierarchy. Right: Edge labeling for the full hierarchy.

In Step 5 of Algorithm 1, a *b*-ary number is associated to each leaf in the hierarchy which is constructed as the concatenation of the digits labeling the edges along the path from the root to the leaf: the label in the edge starting from the root is the most significant one and the label closest to the leaf is the least significant one. Fig. 3 shows how labels for leaves 4 and 5 are constructed as ternary numbers from the ternary digits labeling the edges connecting the root to those two leaves.

Since leaves can be at different levels of the hierarchy tree, some leaf labels may be shorter than others. In Step 6, the length of all leaf labels is made equal to the depth L of the hierarchy tree (in our example we have $L = 3$) by appending as many zeroes as needed in the least significant positions to the shorter labels. Table 2 shows leaf labels before and after uniformizing their length.

In Step 7 the following level sets are formed:

$$T_{\text{Disease}}(0) = \{0, 0, 0, 0, \mathbf{0}, \mathbf{0}, 0, 1, 0\},$$
$$T_{\text{Disease}}(1) = \{1, 0, 0, 2, \mathbf{1}, \mathbf{1}, 0, 1, 1\},$$
$$T_{\text{Disease}}(2) = \{2, 2, 1, 1, \mathbf{0}, \mathbf{0}, 0, 1, 1\},$$

where digits emphasized in bold correspond to leaf 9. Those digits appear twice because there are two patients associated to leaf 9.

Finally, in Step 8 the variance at each hierarchy level is computed

$$\text{Var}(T_{\text{Disease}}(0)) = 0.098765,$$
$$\text{Var}(T_{\text{Disease}}(1)) = 0.395062,$$
$$\text{Var}(T_{\text{Disease}}(2)) = 0.543210$$

and the level variances are combined to obtain the variance of the sample represented in the hierarchy tree

$$\text{Var}(T_{\text{Disease}}) = 0.098765 + 3^2 \times 0.395062 + 3^4 \times 0.543210 = 47.654321.$$

If we now remove the patient with breast cancer and add a second patient with epilepsy, the intuitive idea is that the variance should decrease. Let us check this. The new sample T'_{Disease} is pictorially represented in Fig. 4 and its leaf labels are

$$\{210, 111, 110, 120, 100, 100, 010, 010, 000\}.$$

The new level sets are

$$T'_{\text{Disease}}(0) = \{0, 1, 0, 0, 0, 0, 0, 0, 0\},$$
$$T'_{\text{Disease}}(1) = \{1, 1, 1, 2, 0, 0, 1, 1, 0\},$$
$$T'_{\text{Disease}}(2) = \{2, 1, 1, 1, 1, 1, 0, 0, 0\}$$

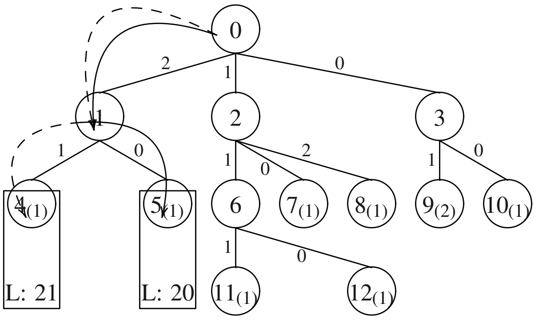


Fig. 3. Labeling of leaves 4 and 5. The dashed line shows the path followed to label leaf 4 while the solid line shows the path for leaf 5.

Table 2
Left: leaf labels before length uniformization; Right: leaf labels after length uniformization

Node ID	Most significant		Least significant	Node ID	Most significant		Least significant
4		2	1	4	2	1	0
5		2	0	5	2	0	0
7		1	0	7	1	0	0
8		1	2	8	1	2	0
9		0	1	9	0	1	0
10		0	0	10	0	0	0
11	1	1	1	11	1	1	1
12	1	1	0	12	1	1	0

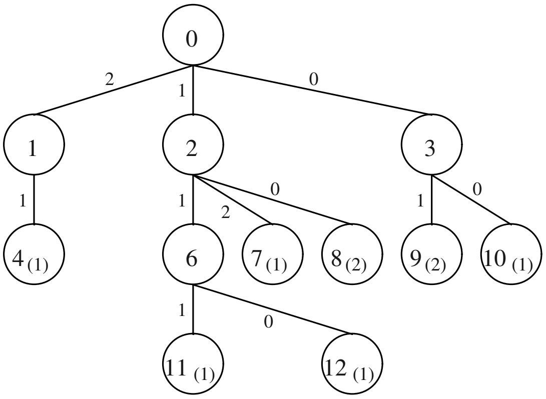


Fig. 4. Example pruned hierarchy of a second sample of the "*Disease*" attribute.

and the new level variances are

$$\mathrm{Var}(T'_{\mathrm{Disease}}(0)) = 0.098765,$$
$$\mathrm{Var}(T'_{\mathrm{Disease}}(1)) = 0.395062,$$
$$\mathrm{Var}(T'_{\mathrm{Disease}}(2)) = 0.395062.$$

Therefore, the variance of the new sample is

$$\mathrm{Var}(T'_{\mathrm{Disease}}) = 0.098766 + 3^2 \times 0.395062 + 3^4 \times 0.395062 = 35.654321.$$

So clearly the variance of this second sample is less than the variance of the first sample.

4. Validity and consistency

The proposed variance measure is intended for hierarchical nominal attributes, so that it is not directly comparable to the classical definition of variance for numerical attributes used in statistics. In order to support the appropriateness of our definition, we will first show that its definition is valid (Section 4.1) and then we will show that it is consistent with previous spread measures for non-hierarchical nominal attributes (Section 4.2).

4.1. Validity

The purpose of this section is to show that the proposed measure is well-defined and really measures how close the values in a sample of a hierarchical nominal attribute are within the attribute's hierarchy.

First it is shown that the proposed categorical value labeling does not depend on how the tree representing the hierarchy of categories is depicted.

Lemma 1. *For a given set of categories, the category labeling described in Algorithm 1 is invariant to isomorphisms of the tree representing the hierarchy of categories.*

Proof. For any edge in the tree hierarchy, the digit assignment described in Algorithm 1 depends only on the number of categorical values associated to the edge's descendant leaves. Thus, the digit assignment is independent of the relative position of the edge among those in its same depth level. The lemma follows: if two subtrees are swapped, each subtree is coded as it would have been before swapping. □

The specific numerical outcome yielded by the proposed variance measure depends on the coding of nominal categories. We show next that the coding upon which the measure is based yields the minimum possible variance for a given set of categorical values. Thus, the measure is well-defined and can be used to compare the variance of two hierarchical nominal samples.

Lemma 2. *For a given set of categorical values, the category labeling described in Algorithm 1 minimizes the variance yielded by expression (1).*

Proof. Given a set of n categorical values, if we prove that each of $\mathrm{Var}(T_X(0)), \mathrm{Var}(T_X(1)), \ldots, \mathrm{Var}(T_X(L-1))$ are minimum, it will follow that expression (1) is minimum. For any i, consider

$$\mathrm{Var}(T_X(i)) = \frac{\sum_{j=1}^{n}(d_j^i - \bar{d}^i)^2}{n}, \tag{2}$$

where

$$d_j^i \in \{l_0, l_1, \ldots, l_{b-1}\} =\{\, 0, \ldots, b-1\}$$

is the digit in the ith position (level) for the leaf label corresponding to the jth categorical value. By construction of the coding, the most frequent value l_0 in

$$T_X(i) = \{d_0^i, \ldots, d_n^i\}$$

is the most central value in the range $[0, b-1]$; the second most frequent value is the second most central value l_1 in the range, the third most frequent value is the third most central value l_2, etc. Therefore l_0 is as close as an integer digit can be to the average $\bar{d}^i$ of $T_X(i)$. On the other hand, the more different are values in $T_X(i)$ from l_0, the less frequent they are. Therefore, coding the set $T_X(i)$ as described in Algorithm 1 minimizes expression (2). □

Next, we give a lemma and a corollary that bound the proposed nominal variance.

Lemma 3. *For any sample T_X of a nominal attribute X, it holds that $0 \leqslant \mathrm{Var}(T_X(j)) \leqslant ((b-1)/2)^2$, for $j = 0, \ldots, L-1$, where L is the depth of the hierarchy of X and b is the maximum number of children that a node can have in the hierarchy of X.*

Proof. $\mathrm{Var}(T_X(j))$ is obviously nonnegative for any j. To check the upper bound, let us first assume that the size n of the sample T_X is even. Since values in T_X must be in the range $\{0, \ldots, b-1\}$, $\mathrm{Var}(T_X(j))$ is maximized when $T_X(j)$ consists of $n/2$ digits equal to 0 and $n/2$ digits equal to $b-1$. Therefore, for n even, we have

$$\mathrm{Var}(T_X(j)) \leqslant \frac{(n/2)(0-(b-1)/2)^2 + (n/2)((b-1)-(b-1)/2)^2}{n} = ((b-1)/2)^2. \tag{3}$$

Now, let us assume that n is odd. Since the values in T_X must be in $\{0, \ldots, b-1\}$, $\mathrm{Var}(T_X(j))$ is maximized when $T_X(j)$ consists of $(n+1)/2$ digits equal to 0 and $(n-1)/2$ digits equal to $b-1$. Therefore, for n odd, we have

$$\mathrm{Var}(T_X(j)) \leqslant \frac{((n+1)/2)(0-(b-1)/2)^2 + ((n-1)/2)((b-1)-(b-1)/2)^2}{n} = ((b-1)/2)^2. \quad \square \tag{4}$$

The corollary below follows from Lemma 3 and expression (1).

Corollary 1. *For any sample T_X of a nominal attribute X, it holds that*

$$0 \leqslant \mathrm{Var}(T_X) \leqslant ((b-1)/2)^2 \cdot b^{2(L-1)},$$

where L is the depth of the hierarchy of X and b is the maximum number of children that a node can have in the hierarchy of X.

Finally, we state a theorem to show that the proposed nominal variance really measures the closeness of the nominal values within the hierarchy of the nominal attribute.

Theorem 1. *Given two samples T_X and T'_X drawn from a nominal categorical attribute X, it holds that $\mathrm{Var}(T_X) < \mathrm{Var}(T'_X)$ if and only if nominal values in T_X are closer in the hierarchy of X than values in T'_X, where higher closeness means values descending from a smaller variety of higher-level edges in the tree hierarchy of X.*

Proof. First, assume that $\mathrm{Var}(T_X) < \mathrm{Var}(T'_X)$. By Lemma 3, $\mathrm{Var}(T_X(j)) \leqslant b^2 - 1$ and $\mathrm{Var}(T'_X(j)) \leqslant b^2 - 1$ for $j = 0, \ldots, L-1$. Therefore, expression (1) for $\mathrm{Var}(T_X)$ can be viewed as a b^2-ary expansion of $\mathrm{Var}(T_X)$ where the jth order term is $\mathrm{Var}(T_X(j))$, for $j = 0, \ldots, L-1$. The same holds for expression (1) applied to $\mathrm{Var}(T'_X)$. (If all terms in those expansions were integer, terms could be regarded as b^2-ary digits; in general, terms are non-integer variances, but this does not affect what follows.) Since $\mathrm{Var}(T_X) < \mathrm{Var}(T'_X)$, the b^2-ary expansions of both variances must satisfy that

$$\exists k \in \{0, \ldots, L-1\} \text{ s.t. } \mathrm{Var}(T_X(k)) < \mathrm{Var}(T'_X(k)) \quad \text{for } j = k+1 \text{ to } L-1 \ \mathrm{Var}(T_X(j)) \leqslant \mathrm{Var}(T'_X(j)). \tag{5}$$

Inequalities (5) mean that, from the kth level of the tree hierarchy of X upwards, values in T_X descend from a set of higher-level edges that is less diverse than the set of higher-level edges from which values in T'_X descend. Therefore, values in T_X are closer within the hierarchy of X than those in T'_X.

Conversely, assume that values in T_X are closer within the hierarchy of X than values in T'_X. This means that, at the higher levels of the tree hierarchy of X, values in T_X descend from a set of higher-level edges that is less diverse than the set of higher-level edges from which values in T'_X descend. So there must be $k \in \{0, \ldots, L-1\}$ such that inequalities (5) hold. Using that $\mathrm{Var}(T_X)$ and $\mathrm{Var}(T'_X)$ can be expressed as b^2-ary expansions with $\mathrm{Var}(T_X(j))$ and $\mathrm{Var}(T'_X(j))$, respectively, being the jth order terms for $j = 0$ to $L-1$, it follows that $\mathrm{Var}(T_X) < \mathrm{Var}(T'_X)$. □

4.2. Consistency with previous diversity measures

The new variance measure for hierarchical nominal attributes cannot be compared with the classical statistical variance for numerical data, precisely because the latter is only defined for numerical data. It can, however, be shown to yield results that are consistent with those obtained with the diversity measures for nominal attributes mentioned in Section 2. Recall that all those measures differ from ours in that they are flat, that is, they do not make use of any hierarchical information available for the nominal attribute. Based on the discussion in Section 2, we have chosen Simpson's measure as the most representative one for comparison.

We first show that for a nominal variable with a flat hierarchy (that is, with a single level), both Simpson's measure and ours behave similarly. Fig. 5 compares the results obtained with Simpson's measure and ours for three samples of a flat hierarchy consisting of one root category and three child nominal categories. In the first sample, there are two values from each category; in the second sample, there are two values from each of the first two categories and three values from the third category; in the third sample, there are two values from the each of the first two categories and four from the third category. If we take into account that Simpson's measure is *inversely* proportional to sample diversity, we can conclude that its diversity criterion coincides with ours for the three samples: the first sample is the most diverse one and the third sample is the least diverse one. This consistency has a more general explanation and should not be surprising: in case of a flat hierarchy, what Algorithm 1 does is to replace each sample value by a leaf label consisting of a single *b*-ary digit (the label of the edge connecting the root to the corresponding leaf) and our measure simply becomes the classical statistical variance of those numerical leaf labels.

However, all measures discussed in Section 2 and Simpson's in particular fail to capture the diversity differences between the three samples in Fig. 6. The three samples consist of six values, but each sample comes from a different hierarchy. The first sample comes from the flat hierarchy in Fig. 5; in the second sample, there are two values from each of sibling categories 1 and 2, and two values from category 4 which is a "niece" category of categories 1 and 2; in the third sample, there are two values from each of sibling categories 1 and 2 and two values from category 5, which is a "grandniece" of categories 1 and 2.

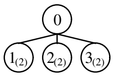
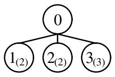
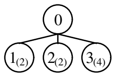
Sample			
Simpson's measure	$\frac{6}{30} = 0.2$	$\frac{10}{42} = 0.23$	$\frac{16}{56} = 0.285$
Ours	0.66667	0.571429	0.500000

Fig. 5. Simpson's measure vs. ours for three samples of a flat hierarchy. Note: Simpson's measure is inversely proportional to the sample diversity.

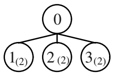
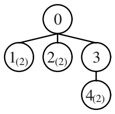
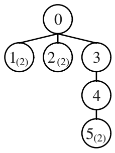
Sample			
Simpson's measure	$\frac{6}{30} = 0.2$	$\frac{6}{30} = 0.2$	$\frac{6}{30} = 0.2$
Ours	0.66667	6.222222	56.222222

Fig. 6. Simpson's measure vs. ours for three samples of non-flat hierarchies.

Simpson's measure detects no diversity differences between the three samples, whereas our measure says that the third sample is the most diverse one and the first sample is the least diverse one.

5. Applications

Beyond the obvious applications of the variance of a hierarchical nominal attribute as a descriptive statistic, we will focus here on two applications, economic diversity measurement and database privacy. For the case of economic diversity, we contribute empirical work on a real data set.

5.1. Measuring economic diversity: a case study

In this section we apply the proposed measure to a real-life problem: comparing the diversity of the economy of a set of European countries. To this end, we use data from the Structure of Earnings Survey 2002 (SES 2002, [12]) collected by Eurostat (the statistical office of the European Union). This survey contains a sample of 222,364 companies from several European countries. The list of countries can be seen in Fig. 8. The number of sampled companies differs for each country. Several attributes corresponding to year 2002 are available for each company, one of which is the company's activity sector, coded according to the NACE rev. 1.1 classification [7].

The NACE classification is a hierarchy with up to four levels. The anonymized SES 2002 data set we used contains NACE codes with only the two upper levels, represented in Fig. 7. The top-level categories are called sections and the second-level categories are called subsections. To illustrate what NACE categories look like, we list NACE sections in Table 3.

Now, if we compute our nominal variance measure for the NACE attribute over the SES 2002 companies sampled for each country, we will get an indication on how diversified is that country's economy. Or perhaps it would be more accurate to say on how *NACE-diversified* is that country's economy. Indeed, NACE categories are far from homogeneous: for example, Section D is huge because it includes all manufacturing, whereas Sections F and P are more specific. This means that a heavily industrial country like Germany is likely to be less NACE-diversified than a less developed country whose companies are more evenly distributed across NACE sections (e.g. a country with a more balanced proportion of manufacturing companies, construction companies and hotels–restaurants).

With the above cautions, we present in Fig. 8 a comparison of the nominal variance for the NACE attribute of every European country included in the SES 2002 survey. The height of a country's bar in the figure is the NACE variance for that country.

5.2. Measuring privacy in database anonymization

In database privacy and, specifically, in the extensions to the *k*-anonymity privacy model, the need to guarantee a lower bound for the dispersion of confidential nominal attributes appears. As explained below, the goal of such lower-bounding is

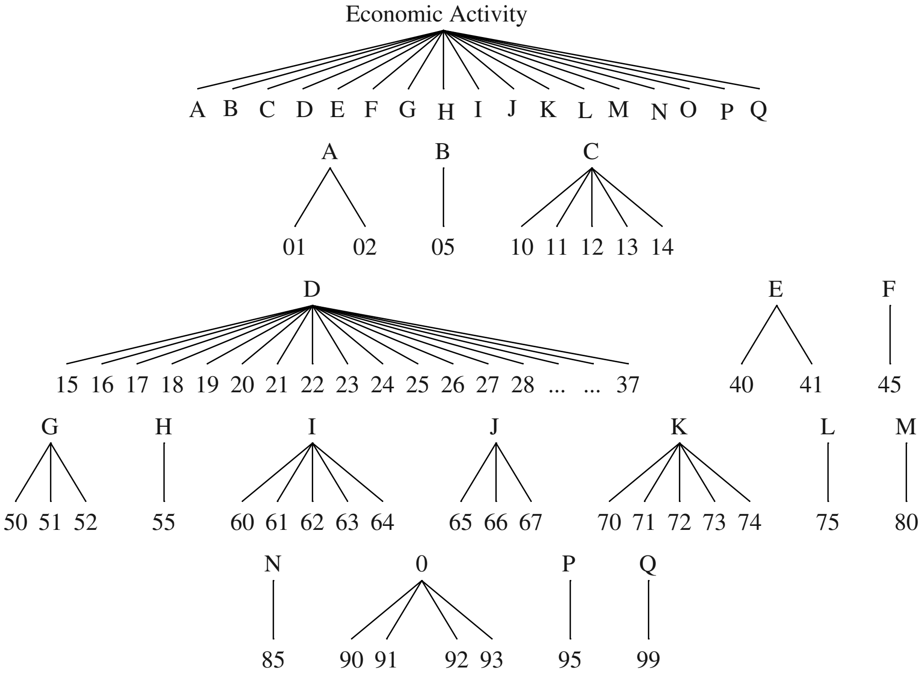


Fig. 7. The two upper levels of the NACE classification.

Table 3
Sections of the NACE classification

Section	Meaning
A	Agriculture, hunting and forestry
B	Fishing
C	Mining and quarrying
D	Manufacturing
E	Electricity, gas and water supply
F	Construction
G	Wholesale and retail trade
H	Hotels and restaurants
I	Transport, storage and communications
J	Financial intermediation
K	Real state, renting and business activities
L	Public administration and defense
M	Education
N	Health and social work
O	Other community and social activities
P	Activities of households
Q	Extra-territorial organizations and bodies

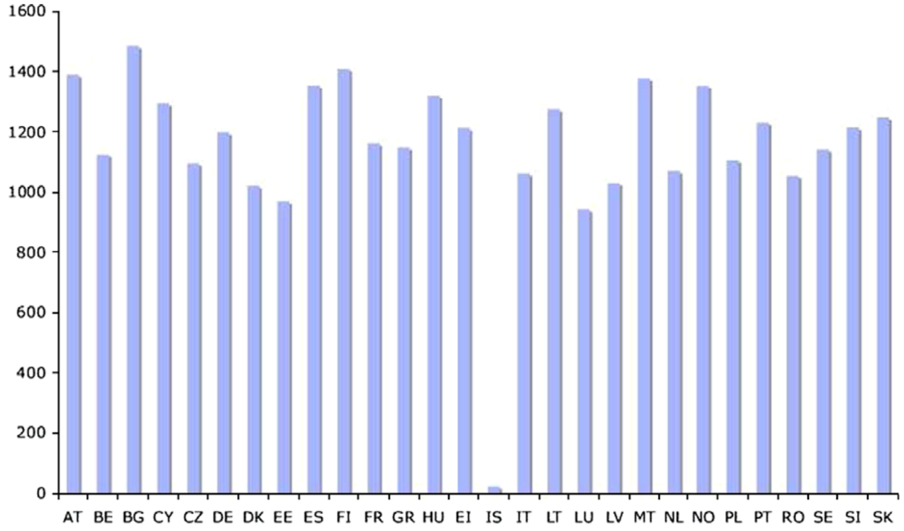


Fig. 8. NACE-diversification of the economy in several European countries (a taller bar indicates more NACE-diversification).

to thwart attribute disclosure. However, we will show that the approaches in the literature poorly capture the dispersion of the values of nominal attributes. We briefly recall the *k*-anonymity model and then two of its extensions. Then the improvement that can be provided by nominal variance is explained.

5.3. k-Anonymity

Depending on their disclosiveness, the attributes in a data set **D** can be classified as *identifiers* (to be suppressed), *key attributes* and *confidential outcome attributes*. Key attributes are those in **D** that, in combination, can be linked with external information to re-identify (some of) the respondents to whom (some of) the records in **D** refer. Example key attributes in a data set about citizens are age, job, zipcode, etc. Confidential outcome attributes are those conveying sensitive information on the respondent (e.g. disease, political opinion, etc.).

Definition 1. (*k*-Anonymity, [9]). A protected data set is said to satisfy *k*-anonymity for $k > 1$ if, for each combination of key attributes, at least *k* records exist in the data set sharing that combination.

A group of records sharing a combination of key attributes will be hereafter called a block. *k*-Anonymity prevents identity disclosure, but it fails to protect against attribute disclosure. Imagine that an individual's health record is *k*-anonymized into a block with *k*-anonymized key attribute values Age = "30", Height = "180 cm" and Weight = "80 kg". Now, if all *k* patients in the block share the confidential attribute value Disease = "AIDS", *k*-anonymization is useless, because an intruder who uses the key attributes (Age, Height, Weight) can link an external identified record

$$(\text{Name} = \text{“John Smith”},\ \text{Age} = \text{“31”},\ \text{Height} = \text{“179”},\ \text{Weight} = \text{“81”})$$

with the above block and infer that John Smith suffers from AIDS (attribute disclosure). There are several extensions to *k*-anonymity which try to fix this problem [3]; we will mention here *p*-sensitive *k*-anonymity and *l*-diversity.

5.4. *p-Sensitive k-anonymity*

The *p*-sensitive *k*-anonymity model is defined as follows.

Definition 2. (*p*-Sensitive *k*-anonymity, [15]). A data set is said to satisfy *p*-sensitive *k*-anonymity for $k > 1$ and $p \leqslant k$ if it satisfies *k*-anonymity and, for each block of tuples with the same combination of key attribute values that exists in the data set, the number of distinct values for each confidential attribute is at least *p* within the block.

A problem which may still arise with *p*-sensitive *k*-anonymity or *l*-diversity is that the *p* values of a confidential attribute that occur within a block *are too similar*. Imagine that there is a nominal confidential attribute “Disease”; even if three-sensitive *k*-anonymity is enforced, if a block of *k* patients sharing a combination of key attributes have diseases in the set

$$\{\text{“Liver cancer”},\ \text{“Lung cancer”},\ \text{“Stomach cancer”}\}$$

then an intruder re-identifying a patient in that block knows that the patient has some kind of cancer. To fight this *similarity* problem, extended *p*-sensitive *k*-anonymity was proposed in [14]. To define this new property, the concepts of strong value and protected subtree are needed.

Definition 3. (Strong value). A category in the generalization hierarchy of a confidential attribute is called strong if none of its ascendants (including the root) leaks sensitive information.

For example, “Cancer” (which generalizes “Liver cancer”, “Lung cancer” and “Stomach cancer”) would be a strong value if its only two ascendants were noncommittal labels such as “Non-infectious disease” and “Disease”, respectively.

Definition 4. (Protected subtree). A protected subtree in the hierarchy of a confidential attribute is a subtree which has a strong value as root.

E.g., if “Cancer” is a strong value, all types of cancer constitute a protected subtree.

Definition 5. (Extended *p*-sensitive *k*-anonymity). A data set is said to satisfy extended *p*-sensitive *k*-anonymity if it satisfies *k*-anonymity and for each block that exists in the data set, the values of each confidential attribute *X* within that block belong to at least *p* different protected subtrees in the generalization hierarchy of *X*.

Thus, enforcing extended *p*-sensitive *k*-anonymity would prevent releasing data on a block of *k* patients sharing a combination of key attribute values and suffering all of them from cancer diseases.

5.5. *l-Diversity*

The basic principle of *l*-diversity [8] is very similar to *p*-sensitive *k*-anonymity: *k*-anonymity is not required, but at least *l* “well-represented” values for the confidential attribute must be present in a block sharing a combination of key attributes. The simplest meaning of “well-represented” is just “different”: however, this has the same shortcomings as plain *p*-sensitive *k*-anonymity (no protection against similarity nor against dominant values with high frequency). A sophistication called entropy *l*-diversity is to require that “well-represented” mean that the sample entropy of the confidential attribute values within each block be at least $\log(l)$. A further sophistication called (c, l)-recursive diversity is to require that, if *m* different values of the confidential attribute appear in a block, the most frequent value does not appear more than *c* times the pooled frequency of the $m - l$ least frequent values.

5.6. *The contribution of nominal variance*

As discussed in the previous sections, the simplest versions of *p*-sensitive *k*-anonymity and *l*-diversity primarily focus on the *number* of different values that a confidential categorical attribute takes in a block of records sharing a combination of key attributes.

In the more advanced versions of both models, it is realized that guaranteeing a minimum number of different values is not enough:

- Entropy *l*-diversity and (c, l)-recursive diversity focus on the *frequencies* of the various values adopted by the confidential attribute within the block. The aim is to ensure that the frequency distribution is not dominated by one or very few values, which would facilitate attribute disclosure. However, note that the *semantic diversity* of the values is not captured; hence, the similarity problem is not addressed, i.e. for a “Disease” attribute, a block may still contain very similar diseases, like cancer types.

- Extended p-sensitive k-anonymity tries to capture the *semantics* of the attribute values, by requiring that the values in a block belong to at least p different protected subtrees, each rooted by a different strong value. However, the notion of strong value may be a bit awkward and largely subjective in many hierarchies.

Nominal variance, on its side, captures the dispersion of nominal values in the hierarchy, without requiring any subjective decision to be made. Therefore, *nominal variance is useful and easily applicable as a criterion to decide whether the values of a confidential nominal attribute are scattered enough to rule out problems of similarity or value dominance*. More specifically, the data protector could require that, for a data set to be considered safe for release, the variance of its confidential nominal attributes within each block ought to be at least a certain pre-selected percentage of the upper bound given by Corollary 1.

6. Conclusion

A measure of nominal variance has been presented which measures the dispersion of a sample drawn from a hierarchical nominal attribute by taking the attribute's hierarchy into account. Using the available hierarchical information to measure spread seems a novelty, even if it is a natural thing to do: spread depends not only on whether values belonging to different categories appear in a sample, but also on how far are the categories of the sampled values within the hierarchy of the attribute. Detailed examples on how the proposed measure is to be applied have been given. Validity and consistency have been discussed in the following sense:

- It has been shown that the proposed measure is well-defined and really measures how close the values in a sample of a hierarchical nominal attribute are within the attribute's hierarchy.
- For non-hierarchical nominal attributes (i.e. those whose categories can only be listed as a flat hierarchy), the proposed measure yields results consistent with previous diversity indicators, mainly devised in the area of natural sciences.

Beyond its use as a descriptive statistic, specific applications of the new nominal variance measure to economic diversity measurement and data anonymization have been illustrated and discussed.

Acknowledgements

The authors are with the UNESCO Chair in Data Privacy, but they are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization.

This work was partly supported by the Spanish Government through Projects TSI2007-65406-C03-01 "E-AEGIS" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the Government of Catalonia under Grant 2005 SGR 00446.

References

[1] M. Beals, L. Gross, S. Harrell, Diversity indices: Shannon's H and E. University of Tennessee at Knoxville, 2000. <http://www.tiem.utk.edu/mbeals/shannonDI.html>.

[2] S. Díaz, J.L. García-Lapresta, S. Montes, Consistent models of transitivity for reciprocal preferences on a finite ordinal scale, Information Sciences 178 (2008) 2832–2848.

[3] J. Domingo-Ferrer, V. Torra, A critique of k-anonymity and some of its enhancements, in: Proceedings of the Third International Conference on Availability, Reliability and Security, ARES 2008, IEEE Computer Society, 2008, pp. 990–993.

[4] A. Kolesárová, G. Mayor, R. Mesiar, Weighted ordinal means, Information Sciences 177 (2007) 3822–3830.

[5] W. Kotlowski, K. Dembczyński, S. Grecop, R. Slowiński, Stochastic dominance-based rough set model for ordinal classification, Information Sciences 178 (2008) 4019–4037.

[6] B. McCane, M. Albert, Distance functions for categorical and mixed variables, Pattern Recognition Letters 29 (2008) 986–993.

[7] NACE-Nomenclature générale des Activités économiques dans les Communautés Européennes (Statistical Classification of Economic Activities in the European Community), Revision 1.1. Eurostat, European Commission, 2002. <http://ec.europa.eu/comm/eurostat/ramon/>.

[8] A. Machanavajjhala, J. Gehrke, D. Kiefer, M. Venkitasubramanian, L-diversity: privacy beyond k-anonymity, in: Proceedings of the 22nd International Confernce on Data Engineering, ICDE 2006, IEEE, 2006, p. 24.

[9] P. Samarati, L. Sweeney, Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, Technical report, SRI International, 1998.

[10] C.E. Shannon, A mathematical theory of communication, Bell Systems Technical Journal 14 (1948) 7–13.

[11] E.H. Simpson, Measurement of diversity, Nature 163 (1949) 688.

[12] Structure of Earnings Survey 2002 – Eurostat's arrangements for implementing the Council Regulation 530/1999 and the Commission Regulation 1916/2000, Eurostat, European Commission, 6 April 2004.

[13] V. Torra, J. Domingo-Ferrer, J.M. Mateo-Sanz, M. Ng, Regression for ordinal variables without underlying continuous variables, Information Sciences 176 (2006) 465–474.

[14] T.M. Truta, A. Campan, P. Meyer, Generating microdata with p-sensitive k-anonymity, in: Secure Data Management – 4th VLDB Workshop, SDM'2007, Lecture Notes in Computer Science, vol. 4721, Berlin-Heidelberg, 2007, pp. 124–141.

[15] T.M. Truta, B. Vinay, Privacy protection: p-sensitive k-anonymity property, in: Proceedings of the Second International Workshop on Privacy Data Management, PDM 2006, IEEE Computer Society, Atlanta GA, 2006, p. 94.

[16] H.G. Washington, Diversity, biotic and similarity indices, Water Research 18 (1984) 653–694.