Food Quality and Preference 32 (2014) 107–112

Contents lists available at SciVerse ScienceDirect

Food Quality and Preference

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

Determination of the consensus partition and cluster analysis of subjects in a free sorting task experiment

Ph. Courcoux *, P. Faye, E.M. Qannari

LUNAM University, ONIRIS, USC Sensometrics and Chemometrics Laboratory, Nantes F-44322, France
INRA, Nantes F-44316, France

ARTICLE INFO

Article history:
Received 27 September 2012
Received in revised form 5 March 2013
Accepted 8 May 2013
Available online 29 May 2013

Keywords:
Rand index
Adjusted Rand index
Free sorting
Partitions
Consensus partition
Clustering

ABSTRACT

The purpose of this study is to investigate the problem of clustering subjects in a free sorting task. We compare different measures of agreement between partitions. From a simulation study, we advocate using the Adjusted Rand index. On the basis of this index, we propose a technique for determining a consensus partition as a summary of the initial partitions given by the subjects after a categorization task. Thereafter, the problem of clustering the subjects is explored. For this purpose, a method combining hierarchical clustering and a partitioning algorithm is described. These techniques are applied to a case study of the perception of wine aromas by a panel of subjects.

1. Introduction

Free sorting task is a widely used technique for investigating perception of a set of stimuli by a panel of subjects. In the course of a free sorting experiment, each subject is asked to give a partition of the whole set of objects. This task is very natural and simple and may be used with untrained assessors. In most cases, the practitioner is interested in a product-oriented analysis where the main purpose is to describe the proximities between stimuli and/or relate them to instrumental variables or hedonic responses. For this purpose, the responses from the different subjects are aggregated and the analysis generally focuses on the differences between products. Recently, numerous studies compared free sorting task to conventional profiling technique by comparing factorial representations of the products resulting from the two techniques (Faye et al., 2004; Saint-Eve, Kora, & Martin, 2004; Cartier et al., 2006). The categorization task is considered as a potential alternative to quantitative descriptive analysis.

In situations where the study involves untrained assessors, the analysis may be subject-oriented and the aim is to interpret the differences of perception between subjects. Generally, the groups of subjects are known a priori. For instance, some studies involved the comparison of groups of subjects with different cultural backgrounds (Chrea et al., 2004; Blancher et al., 2007) or with different levels of expertise (Lelièvre, Chollet, Abdi, & Valentin, 2008; Ballester, Symoneaux, & Valentin, 2008). However, in some cases, the groups of subjects are not known a priori and the purpose of the study may be to reveal segments of subjects with different perception of the stimuli. In such studies, the focus is more on individual perception of the different subjects than on a global description of the stimuli. Naturally, the clustering of the subjects appears to be the main clue of the analysis and this question is analogous to the problem of comparing several partitions of the same set of objects. In a first stage, we will compare some measures of agreement between partitions. In a second stage, the definition of a consensus between partitions will be studied and an algorithm will be proposed for determining a central partition that summarizes the different partitions given by the subjects after a sorting task. A technique for clustering the subjects around their central partition will be proposed and allows us to study the differences in perception of a panel of subjects.

It is worth noting that the investigation of the differences among subjects participating to a free sorting task has always been a constant preoccupation of scientists. Very often, this is done by highlighting how the subjects agree on a group average configuration of the stimuli. This strategy of analysis is typified by DISTATIS (Abdi, Valentin, Chollet, & Chrea, 2007) where each subject is assigned a weight that reflects his agreement with an average configuration. Alternative methods highlight how the subjects agree upon the various dimensions underlying the average configuration

* Corresponding author at: LUNAM University, ONIRIS, USC Sensometrics and Chemometrics Laboratory, Nantes F-44322, France. Tel.: +33 0251785436.

E-mail address: philippe.courcoux@oniris-nantes.fr (Ph. Courcoux).

0950-3293/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.foodqual.2013.05.004

of the stimuli (Cadoret, Lê, & Pagès, 2009; Qannari, Cariou, Teillet, & Schlich, 2009). Obviously, our strategy of analysis is complimentary to these approaches since we ultimately seek segments of homogenous subjects. Once these segments are formed, they can be separately subjected to one or the other of the (multivariate) methods mentioned above.

2. Agreement between partitions

The problem of comparing partitions has been widely studied and numerous criteria have been introduced for evaluating the proximity (or the dissimilarity) between partitions of the same set of objects. A comprehensive review of different measures of agreement between partitions is given by Youness and Saporta (2004a,b). Among these indices, one of the most popular is Rand index. Developed by Rand (1971), it is generally used for comparing partitions of a set of objects obtained by means of different clustering algorithms.

2.1. Rand index

Let us consider a set S of N objects and let us assume that we have two partitions $U=\{u_i,\dots u_Q\}$ and $V=\{v_i,\dots v_Q\}$ of the objects in S. As indicated, U and V have respectively Q and R classes. The concordance between the partitions U and V may be described by the cross tabulation of the pairs of elements of S (Table 1).

The quantities a and d may be interpreted as agreements between the two partitions whereas b and c represent disagreements between them.

The Rand index (Rand, 1971) is a very natural measure of agreement between partitions: the proportion of agreements among the pairs of objects involved in the two partitions to be compared.

$$\mathrm{RI}(U,V)=\frac{a+d}{a+b+c+d}=\frac{a+d}{\binom{N}{2}}=\frac{a+d}{N(N-1)/2}$$

Rand index (RI) lies between 0 and 1. The maximum value of 1 is obtained when the two partitions perfectly agree (the two partitions are similar).

Alternatively, calculation of Rand index can be derived from the contingency table which cross-tabulates the two partitions. In Table 2, n_{ij} represents the number of stimuli that are classified in the ith cluster of partition U and the jth cluster of partition V.

RI can be computed as

$$\mathrm{RI}(U,V)=\frac{\binom{N}{2}+\sum_{i=1}^{Q}\sum_{j=1}^{R}n_{ij}^2-\frac{1}{2}\left(\sum_{i=1}^{Q}n_{i.}^2+\sum_{j=1}^{R}n_{.j}^2\right)}{\binom{N}{2}}$$

Notwithstanding its intuitive appeal, Rand index suffers from several deficiencies as it is likely to take large values for two partitions which are generated at random. The reason is that the number d may be relatively high even in the case of random partitions. In addition, expected value of RI of random partitions does not take a constant value and highly depends on the number of products and the number of classes of the two partitions.

Table 1
Cross tabulation of pairs of objects.

		V	
		Same class	Different classes
U	Same class	*a*	*b*
	Different classes	*c*	*d*

a: number of pairs of objects placed in the same class in *U* and in the same class in *V*.
b: number of pairs of objects placed in the same class in *U* and in different classes in *V*.
c: number of pairs of objects placed in different classes in *U* and in the same class in *V*.
d: number of pairs of objects placed in different classes in *U* and in different classes in *V*.

Table 2
Contingency table which cross-tabulates partitions *U* and *V*.

	Class	Partition V				Sums
		v_1	v_2	–	v_R	
Partition U	u_1	n_{11}	n_{12}		n_{1R}	$n_{1.}$
	u_2	n_{21}	n_{22}		n_{2R}	$n_{2.}$
	–					
	u_Q	n_{Q1}	n_{Q2}		n_{QR}	$n_{Q.}$
	Sums	$n_{.1}$	$n_{.2}$		$n_{.R}$	N

As an illustration of these properties, we conducted a simulation study to highlight the behavior of Rand index. Partitions of 20 objects were randomly generated with a number of classes varying between 2 and 19. Rand index for each pair of partitions was computed and the distribution of these values (for all the pairs of 1000 random partitions) is shown in Fig. 1.

The mean value of the Rand index is 0.7236 and the 95th percentile of this distribution is equal to 0.952. It can be seen that this distribution is skewed with bulk of the values including the mean and the median lying to the right (close to 1). This means that, for a setting (i.e. number of stimuli and groups) which is common in free sorting task, very high values of Rand index are likely to be observed due only to agreement by chance.

In order to illustrate the effect of the number N of stimuli and the number of classes of the two partitions being compared, we simulated 1000 random partitions. The number N varied from 10 to 28 and, for each value of N, we randomly generated partitions with number of classes from 3 to $N-1$. The Fig. 2 shows the results of this simulation.

In Fig. 2, each curve is associated to a given number of stimuli (i.e. $N = 10$, 16, 22 and 28) and depicts the evolution of the average Rand index as a function of the number of groups in the partitions being compared.

For a fixed number of stimuli, the mean Rand index clearly increases with the number of classes of the partitions. The reason is that the number of pairs separated in the two partitions being compared becomes larger when the number of classes increases

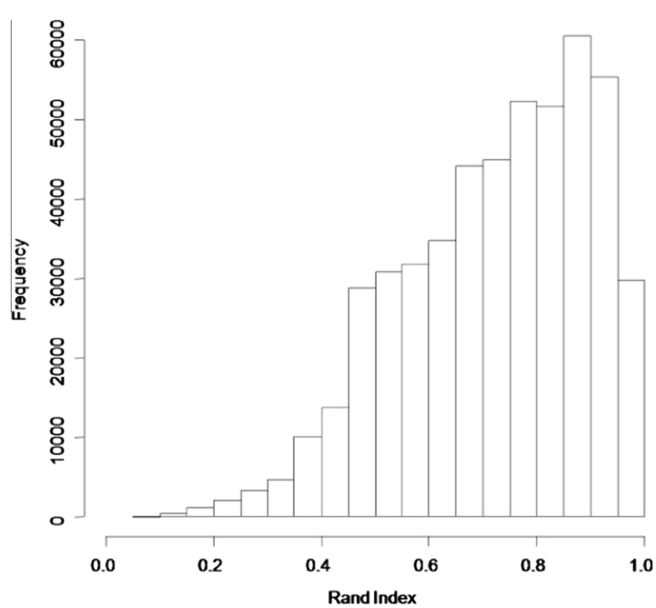

Fig. 1. Distribution of Rand index between pairs of partitions from 1000 randomly simulated partitions of 20 stimuli.