Food Quality and Preference 32 (2014) 35–40

Contents lists available at SciVerse ScienceDirect

Food Quality and Preference

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

An original methodology for the analysis and interpretation of word-count based methods: Multiple factor analysis for contingency tables complemented by consensual words

Belchin Kostov [a,b,*], Mónica Bécue-Bertaut [b], François Husson [c]

[a] Transverse Group for Research in Primary Care, IDIBAPS, Mejia Lequerica, s/n, 08028 Barcelona, Spain
[b] Department of Statistics and Operational Research, Universitat Politècnica de Catalunya, Campus North C5, Jordi Girona 1-3, 08034 Barcelona, Spain
[c] Applied Mathematics Department, Agrocampus Rennes, 65 rue de Saint-Brieuc, 35042 Rennes Cedex, France

ARTICLE INFO

Article history:
Received 12 October 2012
Received in revised form 18 March 2013
Accepted 22 June 2013
Available online 2 July 2013

Keywords:
Word-count based method
Free-text descriptions
Multiple factor analysis for contingency tables
Product descriptions
Consensual word
Correspondence analysis

ABSTRACT

In sensory analysis, results from word-count based methods are customary analyzed through correspondence analysis applied to the global table products × words summing the citations of a same word given by all the panelists. This approach assumes that a same word mentioned by different panelists corresponds to a similar perception, which is not always verified. To solve this problem, we propose a new methodology based on multiple factor analysis for contingency tables. This methodology offers a mean configuration of the products taking into account all the individual words but spots these that are consensual to ease the interpretation. The consensual words have the same meaning for most of the consumers as far as they describe the same products. A test, based on resampling techniques, allows for assessing the significance of the consensus. A real example shows how this methodology eases the interpretation of the word-count based methods by solving problems arising from the large diversity of vocabulary and the different meanings possibly associated to a same word.

1. Introduction

In order to design, promote and market food products that meet the consumers' sensory expectations, food companies need information about how consumers perceive the sensory characteristics of the products. In this sense, interest for collecting descriptions of products, less costly and time-consuming than conventional profiling methods led to design new sensory methods (Abdi & Valentin, 2007; Chollet, Lelièvre, Abdi, & Valentin, 2011; Delarue & Sieffermann, 2004; Faye et al., 2004; Perrin et al., 2008). These methods include, among others, free-text descriptions and frequency-of-citations based techniques (Campo, Ballester, Langlois, Dacremont, & Valentin, 2010; Varela & Ares, 2012) that can be gathered under the name of word-count based methods. The most widely used are:

- *Open-ended questions* (ten Kleij & Musters, 2003): panelists are asked to provide a free-description of the sensory characteristics of a set of products.
- *Check-all-that-apply* (CATA; Lancaster & Foley, 2007): panelists have to select, for each product, the items that they consider appropriate in a list of pre-established phrases or words. Words related to hedonic perception, or even emotions, can be included.
- *Ultra-flash profiling* (UFP; Perrin & Pagès, 2009), a complement to Napping® (Pagès, 2003, 2005): the panelists are asked to write words describing each product after performing a napping task.
- *Labeled sorting task* (Abdi & Valentin, 2007; Cadoret, Lê, & Pagès, 2009), an extension of sorting task (Abdi, Valentin, Chollet, & Chrea, 2007; Lawless, Sheng, & Knoops, 1995): the panelists have to sort products into groups and then label each group with descriptive words.

These methods aim at understanding the consumers' sensory perceptions through collecting the product descriptions from the consumer's own vocabulary (Antmann et al., 2011; Carr, Craig-Petsinger, & Hadlich, 2001). They are usually encoded into a products × words frequency table, called lexical table. The cell (i, j) contains the number of times word j has been used to describe product i.

Correspondence analysis (CA; Benzécri, 1973, 1981; Lebart, Salem, & Berry, 1998; Murtagh, 2005) is a reference method to

* Corresponding author at: Transverse Group for Research in Primary Care, IDIBAPS, Mejia Lequerica, s/n, 08028 Barcelona, Spain. Tel.: +34 93 227 56 83; fax: +34 93 227 55 97.
E-mail address: badriyan@clinic.ub.es (B. Kostov).

0950-3293/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.foodqual.2013.06.009

analyze lexical tables. It is frequently used to tackle word-count based methods either associated to labeled sorting task (Bécue-Bertaut & Lê, 2011; Chollet et al., 2011), ultra flash profiling (UFP, Perrin & Pagès, 2009), under the form of check-all-that-apply (CATA, Ares, Barreiro, Deliza, Giménez, & Gámbaro, 2010; Ng, Chaya, & Hort, 2012) or open-ended questions (Bécue-Bertaut, Àlvarez-Esteban, & Pagès, 2008; ten Kleij & Musters, 2003). CA offers a visualization of: (1) the similarities between products: two products are all the closer as they are described by the same words; (2) the similarities between words: two words are all the closer as they are frequently associated to the same products and (3) the associations between products and words: a word is at the centroid of the products that it describes and a product is at the centroid of the words that describe it.

However, the large diversity of the vocabulary used by the panelists makes the products map difficult to interpret (Chollet et al., 2011). Moreover, the panelists understand and perceive some words in a different way and thus associate them to different products (Cadoret et al., 2009; Veinand, Godefroy, Adam, & Delarue, 2011). As discussed in Delarue and Sieffermann (2004), the ability of the panelists to communicate their sensory perceptions using a common base of vocabulary is sometimes doubtful. If a same word has different meanings for the panelists, it is not possible to use this word to interpret the products map since the reading can lead to different interpretations.

In fact, CA applied to a global products × words table relies on the assumption that a same word mentioned by different panelists reports a similar perception. If this assumption is not verified, summing up the occurrences of a same word used by different panelists may not be meaningful because different perceptions are merged into a same word.

We propose to start from a different encoding of the results issued from word-count based methods; this encoding both preserves all the individual words provided by all the panelists and uncovers the consensual words to guide the interpretation. From each panelist $t(t=1,\ldots,T)$ a lexical table is built with J_t columns corresponding to the individual-words $(j,t),(j=1,\ldots,J_t)$ used by panelist t to describe the products. The columns of the T tables can or cannot correspond to the same words. At that first step, no relationships are established between the homologous individual-words. Then, the T tables are juxtaposed row-wise in a multiple frequency table where cell (i,j,t) is equal to "1" if the individual-word (j,t) was used to describe the product i by the panelist t and "0" if not used. All the individual-words pronounced by the panelists are included in the multiple frequency table. This codification preserves all the individual structures on the products as induced by the different sets of column-words.

The aim of this paper is to present a new methodology to deal with this data encoding. Our proposal combines multiple factor analysis for contingency tables (MFACT; Bécue-Bertaut & Pagès, 2004, 2008), designed to tackle multiple frequency tables, and an original technique to assess which words are consensual and have to guide the interpretation.

The following outline is adopted. In Section 2, the example used to illustrate the methodology is given, data coding and notation are specified and methodology is detailed. Section 3 presents the results. Concluding remarks are offered in Section 4.c

2. Materials and methods

2.1. Dataset

Ninety eight consumers carried out a labeled sorting task on twelve luxury perfumes: Angel, Aromatics Elixir, Chanel n°5, Cinéma, Coco Mademoiselle, L'instant, Lolita Lempicka, Pleasures, Pure Poison, Shalimar, J'adore (eau de parfum), J'adore (eau de toilette). The consumers were mostly women (69.4%) and quite young (mean age: 25 years; range: 18–58). The full dataset can be downloaded from http://factominer.free.fr/book/perfume.csv.

Consumers were placed in individual booths, each perfume was sprayed on a small piece of cotton wool placed into a pill box, and all twelve pill boxes were presented at each consumer. The pill boxes were ordered according to William's Latin squares. Consumers had to evaluate the products in the presentation order but were allowed to go back to any sample; they were asked to make at least two and at most eleven groups of perfumes. After, they had to describe each group with a few words.

All the words are kept without applying any kind of spelling correction, lemmatization, stop-list or frequency threshold. One hundred and ninety-eight distinct-words and six hundred and eighty-four individual-words were used to characterize the perfumes.

2.2. Data coding and notation

As detailed above, the description of the perfumes is coded through perfumes × individual-words tables noted $\boldsymbol{F}_1,\ldots,\boldsymbol{F}_t,\ldots,\boldsymbol{F}_T$, collecting and keeping all the data obtained from the consumers (Fig. 1). Individual table $\boldsymbol{F}_t$, with dimensions $I\times J_t$, has as many columns as different individual-words as the consumer $t(t=1,\ldots,T)$ used. Different individual-words, belonging to different consumers, can correspond to the same word. For example, the word sweet is used by 32 consumers and thus corresponds to 32 different individual-words/columns.

All the individual tables are juxtaposed row-wise into the multiple frequency table $\boldsymbol{F}$, with dimensions $I\times J$ (12 perfumes × 684 individual-words). Each cell f_{ijt} of $\boldsymbol{F}$ is equal to "1" if perfume $i(i=1,\ldots,I)$ is described with individual-word $(j,t),(j=1,\ldots,J_t)$ used by consumer $t(t=1,\ldots,T)$ and "0" if not used. The row margin of table $\boldsymbol{F}$, $f_{i..}=\sum_{jt}f_{ijt}$ $(f_{i..},i=1,\ldots,I)$ corresponds to the counts of individual-words used to describe the perfume i by all the consumers. The column margin, $f_{.jt}=\sum_i f_{ijt}$ $(f_{.jt},(j=1,\ldots,J_t),t=1,\ldots,T)$ corresponds to the counts of perfumes described by the individual-word (j,t) used by consumer t. $f_{...}$ is the grand total of table $\boldsymbol{F}$. The multiple frequency table $\boldsymbol{F}$ is transformed into proportion table $\boldsymbol{P}$ with general term $(p_{ijt}=\frac{f_{ijt}}{f_{...}})$; $\sum_{ijt}p_{ijt}=1$. From $\boldsymbol{P}$, row and column weights are computed: $p_{i..}=\sum_{jt}p_{ijt}$ $(p_{i..},i=1,\ldots,I)$ and $p_{.jt}=\sum_i p_{ijt}$ $(p_{.jt},(j=1,\ldots,J_t),t=1,\ldots,T)$. The internal row weights $p_{i.t}=\sum_j p_{ijt}$ $(i=1,\ldots,I)$ and grand total $p_{..t}=\sum_{ij}p_{ijt}$ of each subtable $\boldsymbol{P}_t$ are computed.

2.3. Multiple factor analysis for contingency tables (MFACT)

MFACT is an extension of multiple factor analysis (MFA; Escofier & Pagès, 1998) dedicated to analyze multiple frequency/contingency tables in a CA-like way but in reference to the intra-tables independence model (Bécue-Bertaut & Pagès, 2004). It adopts the MFA point of view to balance the influence of the different tables in the global analysis. MFACT consists of a classical MFA applied to a multiple table $\boldsymbol{Z}$, built from the multiple frequency table, whose general term is given by (1), endowing the rows i with weights $p_{i..}$ $(p_{i..},i=1,\ldots,I)$ and the columns (j,t) with weights $p_{.jt}$ $(p_{.jt},(j=1,\ldots,J_t),t=1,\ldots,T)$.

$$Z=\frac{p_{ijt}-\left(\frac{p_{i.t}}{p_{..t}}\right)p_{.jt}}{p_{i..}\times p_{.jt}}=\frac{1}{p_{i..}}\left(\frac{p_{ijt}}{p_{.jt}}-\frac{p_{i.t}}{p_{..t}}\right) \qquad (1)$$

This multiple table $\boldsymbol{Z}$ juxtaposes the tables of the weighted residuals with respect to the intra-tables independence model. The influence of each subtable $\boldsymbol{Z}_t$ (t = 1,...,T) in the global analysis