Tourism Management 69 (2018) 524–536

Contents lists available at ScienceDirect

Tourism Management

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

# Analysing cluster evolution using repeated cross-sectional ordinal data

Marta Disegna[a,*], Pierpaolo D'Urso[b], Riccardo Massari[b]

[a] Faculty of Management, Department of Accounting, Finance & Economics, Bournemouth University, 89 Holdenhurst Road, Bournemouth, BH8 8EB, United Kingdom
[b] Dipartimento di Scienze Sociali ed Economiche, Sapienza University of Roma, P.le Aldo Moro 5, 00185, Roma, Italy

ARTICLE INFO

*Keywords:*
Repeated cross-sectional data
Matching
Evolution
Ordinal data
Fuzzy data
Fuzzy clustering

ABSTRACT

This study contributes to the existing literature on tourism market segmentation by providing a new matching-clustering procedure that allows patterns of behaviours to be identified using repeated cross-sectional surveys. By extracting equivalent samples over time, the matching method allows inter-temporal cluster analyses to be performed so that a deeper insight into a phenomenon can be obtained beyond the traditional aggregate level of understanding. The paper provides a step-by-step description of the matching-clustering procedure that can be easily replicated, both within and outside the tourism field, when repeated cross-sectional data are available. From a practical and managerial perspective, the proposed procedure helps destination managers and municipalities to describe and verify the efficacy of policy and strategies adopted over years without the necessity to rely on longitudinal surveys, which are often difficult to conduct.

## 1. Introduction

In the recent tourism literature, the necessity to perform accurate analyses to obtain information on temporal, spatial or cultural changes regarding a particular phenomenon has been highlighted (Cang, Sun, & Li, 2017; Amaro, Duarte, & Henriques, 2016; Song, van der Veen, Li, & Chen, 2012, 2011), especially in consideration of the increased multiculturalism that characterises the worldwide tourism markets (Jin, Moscardo, & Murphy, 2017). In this paper, a new matching-clustering procedure is suggested to obtain information on the evolution of tourists' behaviour when repeated cross-sectional data are available. The increasing availability of large repeated cross-sectional and longitudinal surveys, at both national and international levels, has given researchers the possibility to study how different tourism-related phenomena evolve either within or between countries. While in longitudinal surveys the same sample of units (individuals, households, firms, cities, etc.) is tracked over time, in the repeated cross-sectional surveys different units are involved each time. Therefore, longitudinal surveys are designed to track individual behaviour gathering information on the evolution of a phenomenon while repeated cross-sectional surveys allows comparing years by using aggregate data (Wooldridge, 2012). When repeated cross-sectional surveys are used to obtain temporal comparisons at an individual level the results can be biased producing wrong policy and managerial conclusions. On the other hand, repeated cross-sectional surveys are frequently designed to capture feelings, emotions, motivation, and tourism behaviour (see for instance the Eurobarometer survey on "Preferences of Europeans towards tourism" or the World Tourism Barometer collected by the UNWTO), making it a necessity to develop an adequate procedure to compare this kind of individual information over time.

The idea behind the matching-clustering procedure suggested in this paper is that more detailed information on the evolution of a particular phenomenon can be gathered at cluster-level when units, observed by repeated cross-sectional surveys, are comparable across time. Fig. 1 schematically describes the suggested 7-steps procedure and the "What-How-Why" questions (i.e. the aims, the methods, and the reasons) behind the adoption of each particular step and technique. The first and the last steps, i.e. the "Matching" and the "Comparing partitions", are the main innovations of this study. In fact, these two steps are not usually included in any traditional cluster analysis since they serve the particular aim of performing and describing comparisons when the same set of segmentation variables is collected on different samples of units.

The core five steps, going from the "Data recoding" to the "Labelling & profiling" phase, are the ones that generally characterise any clustering analysis (Everitt, Landau, Leese, & Stahl, 2011). Consequently, the aims ("What") of these five steps are commonly known in the literature of market segmentation. However, the techniques ("How") that can be adopted to accomplish each of these steps are not unique since the choice depends on both data and research aim ("Why").

The matching-clustering procedure illustrated in this paper combines (1) the matching method with (2) fuzzy numbers and (3) fuzzy clustering algorithm to overcome three problems, respectively: (1) the comparison between cross-sectional samples, (2) the imprecision

* Corresponding author. Faculty of Management, Bournemouth University, Executive Business Centre, 89 Holdenhurst Road, Bournemouth, BH8 8EB, United Kingdom.
*E-mail addresses:* disegnam@bournemouth.ac.uk (M. Disegna), pierpaolo.durso@uniroma1.it (P. D'Urso), riccardo.massari@uniroma1.it (R. Massari).

https://doi.org/10.1016/j.tourman.2018.06.028
Received 1 January 2018; Received in revised form 21 June 2018; Accepted 26 June 2018

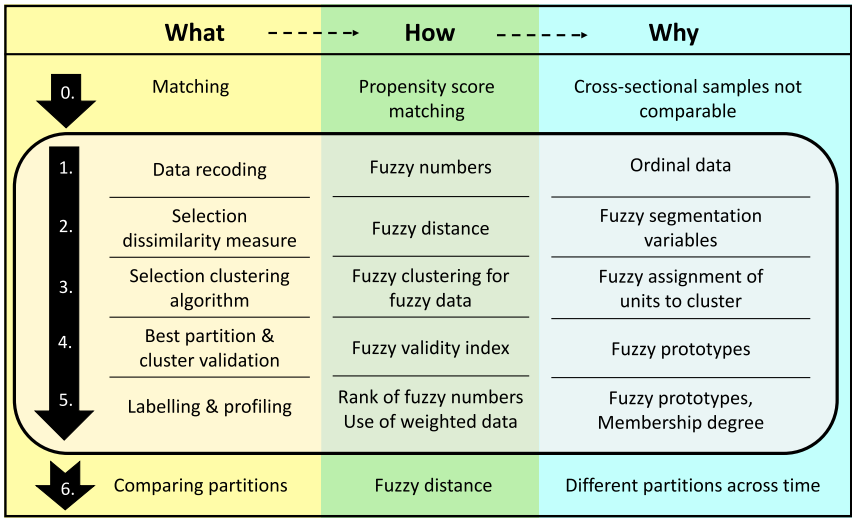


**Fig. 1.** Steps of the suggested matching-clustering procedure.

embedded in ordinal data, and (3) the uncertainty in the assignment of each unit to clusters. Therefore, this paper specifically focuses on the situation in which (1) repeated cross-sectional surveys are available, (2) segmentation variables are measured by means of ordinal scale, such as Likert-type scale, and (3) the hidden market structure is unclear, making it unreasonable to assign each unit solely to one cluster. Readers should note that the focus of this paper is on the procedure itself, not on the techniques ("How") suggested. In fact, each technique can be replaced by any other suitable technique better able to address the "What" questions according to either the research aims or the nature of the segmentation variables; however, the matching-clustering procedure should remain the same if repeated cross-sectional surveys are used.

The next section provides a review of the relevant literature describing these three problems while the third section provides a more technical discussion of each step of the suggested procedure. The fourth section describes the empirical data used to illustrate the matching-clustering procedure, while the findings are discussed in the fifth section. The final section provides conclusions, highlighting both limitations and further directions.

## 2. Literature review

### 2.1. Statistical matching

In order to carry out meaningful temporal comparisons, it is necessary to perform quantitative analyses on harmonised samples, i.e. samples made up from the same group of units (individuals, households, companies, cities, etc.) observed over time, or paired samples, i.e. samples of units that are similar according to a set of controlling factors defined *a priori*. Harmonised samples result from either longitudinal or panel surveys. Even if the same group of units is studied across time, researchers have to interpret carefully changes observed at the micro-level (Brandt, 2018). In fact, results can be compromised, for instance, by a socio-demographic or economic change: an individual could change civil status, becoming separated or getting married, or employment status, becoming employee or retired; household composition can change on the arrival of new babies or the death of a family member; a company can become bigger changing its legal status. The main clustering approaches suggested in the literature to aggregate units characterised by similar behaviour across time are the (1) model- (2) feature- and (3) observation-based approaches (for more details, see Disegna, D'Urso, & Durante, 2017; Caiado, Maharaj, & D'Urso, 2015).

When repeated cross-sectional surveys are used to see how a key relationship has changed over time, a common procedure is to create a pooled dataset aggregating cross-sectional data from different years. Although extensive econometric techniques have been developed for the analysis of pooled cross sectional data (Wooldridge, 2012), in tourism market segmentation the cluster analysis is usually performed on the pooled dataset losing the information on the evolution of the phenomenon (see for instance Ferrer-Rosell & Coenders, 2018). Another common approach adopted in tourism literature, is to compare directly clusters over time, even if samples made by different units are used (see for instance Cang et al., 2017). However, as stated at the beginning of this section, comparison should be made on either harmonised or paired samples. In fact, comparing clusters that have been obtained on different samples can lead to misleading and imprecise conclusions, since it is impossible to know whether the changes finally observed in the phenomenon are due to actual changes in the phenomenon or if they are due to structural differences in the samples.

In this study the adoption of a statistical matching procedure is suggested to overcome this limitation and identify paired samples in repeated cross-sectional surveys. As discussed in D'Orazio et al. (2006), statistical matching procedures aim to match two or more datasets in order to create a more informative integrated dataset. Statistical matching procedures can be adopted when (1) a set of variables is commonly observed, (2) different units, belonging to the same population, are observed in different datasets and (3) any couple of datasets are conditionally independent given a set of common variables used for the matching. Since datasets obtained through repeated cross-sectional surveys satisfy all these conditions, it is reasonable to adopt a statistical matching procedure. In this case, the final aim is not, as usual, the creation of an integrated dataset but the identification of a perfectly matched pair of units belonging to different datasets collected over years. When similar sub-samples are identified, a cluster analysis can be performed on each sub-sample and the results can be compared based on the variables used in the matching procedure and any other variables against which the sub-samples are statistically independent. In this way, the comparison will highlight changes in the phenomenon under observation and not changes in the samples used to obtain the clusters.

### 2.2. Data recoding

Feelings, emotions, motivation, and consumer behaviour are complex psychological processes vaguely captured through quantitative imprecise measurements (Hung & Yang, 2005; D'Urso, 2007). In both academia and industry, ordinal scales, such as Likert-type scales, are