Tourism Management 61 (2017) 173–179

Contents lists available at ScienceDirect

Tourism Management

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

Making sense of tourists' photographs using canonical variate analysis

Nika Balomenou [a], Brian Garrod [b,*], Andri Georgiadou [a]

[a] University of Hertfordshire, UK
[b] Aberystwyth University, UK

HIGHLIGHTS

- Tourists' photographs can be a rich source of behavioural, perceptual and attitudinal data.
- Analysis of such data tends to be resource-intensive if coder subjectivity is to be regulated.
- A pragmatic response may be to identify and select the most meaningful variables for coding.
- Canonical variate analysis (CVA) has great potential to accomplish this without loss of data richness or explanatory power.
- CVA has distinct advantages over alternative multivariate techniques.

ARTICLE INFO

Article history:
Received 29 October 2015
Received in revised form
1 February 2017
Accepted 9 February 2017

Keywords:
Tourists
Photographs
Canonical variate analysis
Dimensionality reduction

ABSTRACT

Tourists' photographs can serve as a rich database for researchers wishing to study tourists' perceptions and attitudes towards destinations. Such data can also be useful in examining how tourists behave, where, when, with whom and why. Many researchers favour the qualitative analysis of such data, which requires the use either of relatively small numbers of photographs or a considerable expense of researcher time and effort to undertake. Much of this process is speculative, in that it involves working with variables which may or may not prove to be significant in addressing the hypotheses chosen for the research. This paper recommends the use of a preliminary phase of research in which a quantitative approach is used to reduce the number of variables needing to be coded. Canonical variate analysis is suggested as an appropriate tool for achieving this. Case study results are presented to demonstrate the utility of this approach.

1. The problem

There is a considerable untapped potential for applying visual research methods in tourism (Garrod, 2008). This is despite the significant progress that has been made in recent years in terms of theorising visual tourism research (Scarles, 2011), addressing critics' concerns about the ‘subjective’ nature of visual research (Balomenou & Garrod, 2014; Crang, 2003), and technological advances in personal photography (Straumann, Coltekin, & Andrienko, 2014). More specifically, tourists' photographs can serve as rich datasets to help answer pressing questions about tourists' preferences and behaviours. Such images are increasingly available in large volumes, whether they are collected using participant-generated image (PGI) techniques (Sun, Ryan, &Pan, 2014; Pan, Lee, & Tsai, 2014; Fung & Jim, 2015; Cutler, Doherty, & Carmichael, 2016) or employ images found in the media, notably the burgeoning number of social media sites such as Flickr and Instagram (Kim & Stepchenkova, 2015; Konijn, Sluimer, & Mitas, 2016; Michaelidou, Siamagka, Moraes, & Micevski, 2013). As such, they can be thought of as ‘big data’ and have enormous potential for the application of data-mining techniques, for example to identify the elements of the destination that appeal the most to tourists and can be emphasised in marketing activities.

Big photographic datasets can, however, be exceedingly resource-hungry to prepare, analyse and interpret (Balomenou & Garrod, 2014; Pearce, Wu, & Chen, 2015). Merely the coding-up can take months of researcher time. Pearce et al. (2015), for example, used a team of two researchers who worked full time for four months coding 10,000 photos into 42 variables. One of the authors of this paper, meanwhile, spent two months of full-time work coding 500 photos into 33 variables, and a further four

* Corresponding author. Aberystwyth Business School, Aberystwyth University, Rheidol Building, Llanbadarn Campus, Aberystwyth, Ceredigion, Wales SY23 3AL, UK.

E-mail address: bgg@aber.ac.uk (B. Garrod).

http://dx.doi.org/10.1016/j.tourman.2017.02.010
0261-5177/

months coding 996 photos into 12 variables. These significant resource demands serve to limit the practicality of using visual methods with large numbers of images.

This paper sets out a possible response, which is to identify a reduced set of variables that are of greatest relevance to the research questions involved (Darlington, Weinberg, & Walberg, 1973), thus making the coding-up and subsequent analytical processes more manageable. Researchers have long proposed that a preliminary interpretation phase could be applied to reduce the number of variables to be coded up (Albrecht, 1980).

Principal component analysis (PCA) has, to date, been the most widely used technique (Johnson, Lloyd, Mur, Smith, & Causton, 2007; Taylor, King, Altmann, & Fiehn, 2002; Schulz, Baranska, Belz, Rösch, Strehle, & Popp, 2004) for dimensionality reduction. A proposed advantage of PCA is that it does this by introducing new variables that are composites of the original variables. It is important to note, however, that PCA is fundamentally an unsupervised technique (Martens & Neaes, 1989), so it does not allow *a priori* hypotheses to be tested. Even where correlations are observed, PCA can provide no measure of the significance of these (Johnson et al., 2007). Moreover, PCA cannot provide clear graphical representations of the interrelationships between the variables, which would be particularly useful in the interpretation of large datasets. Assuming unknown weights for the variables in PCA also risks losing valuable information. This is mainly because of correlation between the number of units analysed and the number of variables (Pérez, Guerrero, González, Pérez, & Caballero, 2013). Moreover, PCA cannot be used in cases where the data come from multiple samples, nor for a repeated-measures design. This limits the utility of PCA as a means of dimensionality reduction.

An alternative technique that is sometimes used for dimensionality reduction is factor analysis. Dwyer, Mellor, Livaic, Edwards, and Kim (2004), for example, use it to suggest various indicators that can be used to estimate the competitiveness of tourism destinations. However, as with PCA, there are no established criteria against which to assess the findings.

This paper proposes that canonical variate analysis (CVA) has strong potential as a dimensionality-reduction technique. It can be said to be superior to similar techniques in several important respects. CVA can measure the comparative contribution of each variable in the canonical (composite) relationships that are calculated, hence allowing the relationships between various sets of the independent and dependent variables to be assessed. As Larimore (1997) explains, CVA is a maximum likelihood statistical technique that can be used to classify the relationships between variables. As such, CVA allows for the testing of hypothesis using a measure of prediction accuracy. The following section presents a brief outline of CVA.

2. A proposed solution: canonical variate analysis

Canonical variate analysis (also known as canonical discriminant analysis) can be thought of as a variant of canonical correlation analysis (CCA), where group indicators form one variable set (Gittins, 1985). CCA was developed by Hotelling (1935) as a means of identifying the linear combination of one set of variables, X, that is most correlated with another linear combination of a second set of variables, Y. Beaghen (1997, p. 6) emphasises that canonical correlation has the property of biorthogonality, which is 'the property that each canonical variate in the X-domain is uncorrelated with the canonical variates in the Y-domain except the corresponding Y-variate'. CCA has been used in tourism research in the context of travel motivations and push and pull factors (Gonzalez & Bello, 2002; Oh, Uysal, & Weaver, 1995; Baloglu & Uysal, 1996; Uysal & Jurowski, 1994), tourism behaviour (Wong & Lau, 2001), destination marketing and branding (Ahmed, 1986; Hosany, Ekinci, & Uysal, 2006), e-relationship marketing and hotel financial performance (Jang, Hu, & Bai, 2006), hosts perceptions of impacts (Allen, Long, Perdue, & Kieselbach, 1988) and demand (Uysal & O'Leary, 1986). However, CCA has not been used extensively, nor specifically to analyse tourist photographs.

Muller (1982) proposed a general linear model for canonical correlation techniques. Developed in 1948 by Rao (1948, 2005), CVA can be thought of as being part of this family. As with CCA, the technique works by constructing canonical variables, each of which can include one or more of the original variables. Darlington et al. (1973) explain the mechanics as a two-stage process, with two statistics. Starting with the original variables, the first canonical correlation is the highest correlation possible between a weighted combination of X variables and a weighted combination of Y variables. These are the first canonical variates (CVs). The second canonical correlation is then calculated as the highest correlation that can be found between the X and Y weighted composites that are uncorrelated with the first canonical variates (Fig. 1). These are known as second CVs.

CVA thus works by detecting the optimum dimensionality of each variable that strengthens the relationship between dependent and independent variable sets. It is based on the premise of defining how much of the variance in one set of variables can be explained by the second set. The most common practice to achieve this is by identifying functions where the canonical correlation coefficients are statistically significant beyond some predetermined level, typically 0.05. In so doing, using CVA helps to ensure that proper regard is given to variations within each variable set (Bussell, Gidman, Causton, Gwynn-Jones, Malham, & Jones, 2008; Chatfield & Collins, 1980; Darlington et al., 1973; Russell, Chiang, & Braatz, 2000). Hair, Anderson, Tatham, and Black (1998) recommend three criteria to use in combination to decide which of the canonical functions should be interpreted: (i) the level of statistical significance of the function, (ii) the magnitude of the canonical correlation, and (iii) the redundancy measure for the percentage of variance accounted for from the two data sets.

CVA has thus far been used predominantly in the biological sciences (Albrecht, 1980; Causton, 2008). Few studies have used CVA in a tourism and hospitality context (rare exceptions being Tran, Dauchez, & Szemik, 2013; Tran & Ralston, 2006) and none as a tool to analyse photographs in the tourism field, despite the surge in readily available photographic data that often result in very large photographic datasets (Lee, 2016).

2.1. Justification for the use of CVA

The studies by Brown, Goldman, Inn, and Anderson (1980) and Tran and Ralston (2006) both used CVA to test hypotheses they had already developed based on interviews with informants. This reflects a key advantage of CVA that is reported by non-social science researchers, who suggest that CVA is best used when the researchers have *a priori* knowledge of the data (Alsberg, Wade, & Goodacre, 1998; Johnson et al., 2007). PCA, in contrast, is fundamentally an unsupervised technique. CVA also allows any variable (be it an original variable or a canonical one) to be continuous, categorical or even mixed (Darlington et al., 1973). This can be vital in the social sciences, allowing 'soft' data to be brought in to help the analysis.

It is also argued that CVA is useful for data visualisation, particularly to evaluate inter-relationships (Johnson et al., 2007)