Making sense of tourists' photographs using Canonical Variate Analysis

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 (Crang, 2003; Balomenou & Garrod, 2014), and technological advances in personal photography (Straumann et al., 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 et al.; 2014; Pan et al., 2014; Fung and Jim, 2015; Cutler et al., 2016) or employ images found in the media, notably the burgeoning number of social media sites such as Flickr and Instagram (Michaelidou et al., 2013; Kim & Stepchenkova, 2015; Konijn et al., 2016). 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 (Pearce et al., 2015; Balomenou & Garrod, 2014). 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 note, meanwhile, spent two months of full-time work coding 500 photos into 33 variables, and a further four 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 research note 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 et al., 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 (Taylor et al., 2002; Schultz et al., 2004; Johnson et al., 2007) 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 et al., 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 et al. (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 (Uysal & Jurowski, 1994; Oh et al., 1995; Balogu & Uysal, 1996; Gonzalez & Bello, 2002), tourism behaviour (Wong & Lau, 2001), destination marketing and branding (Ahmed, 1986; Hosany et al., 2006), e-relationship marketing and hotel financial performance (Jang et al, 2006), hosts perceptions of impacts (Allen et al, 1988) and demand (Uysal & O’Leary, 1986). However, CCA has not been used extensively, nor specifically to analyse tourism photographs.

Muller (1982) proposed a General Linear Model (GLM) 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 (Figure 1). These are known as second CVs.

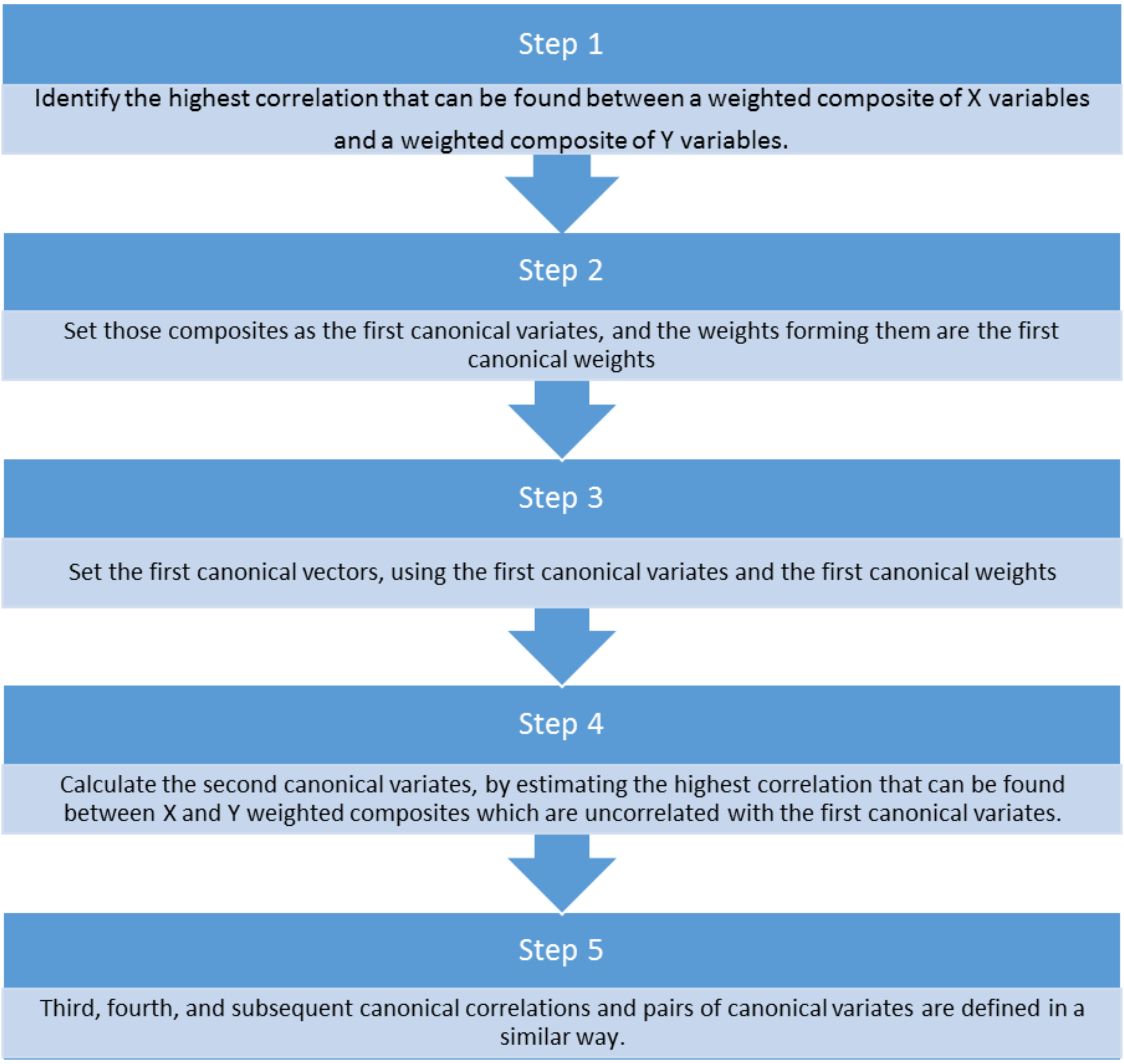

Figure 1: CVA process

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 .05. In so doing, using CVA helps to ensure that proper regard is given to variations within each variable set (Darlington et al., 1973; Chatfield & Collins, 1980; Russell et al., 2000; Bussell et al., 2008). Hair et al. (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 et al., 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 et al. (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 et al., 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) and to reveal the basic structure of complex datasets (Albrecht, 1980). CVA allows the mapping of clusters in two or three dimensions (Hammer and Harper, 2006). Albrecht (1980) explains how CVA helps visualise the dataset on a plot. He regards CVA as a succession of rotational and rescaling transformations of the original variables which protect the integrity of the data while allowing the researcher to interpret them (Albrecht, 1980). He further suggests that using CVA is as if the:

> 'coordinate system defined by the original descriptor variables is suspended in air such that the investigator can walk around it until the most favorable vantage point is located for viewing the differences among the populations. Canonical Variate Analysis simply

defines the most favourable vantage point as being related to the greatest statistical separation among the populations' (Albrecht, 1980, p. 687)

The results of CVA can be conveyed as bivariate plots of one CV versus another, or as three-dimensional plots (Albrecht, 1980). This allows associations between sample groups to become visible (Johnson et al., 2007). The mean of each sample class is plotted against each CV, usually surrounded by a confidence area. The confidence area is circular, and Quinn and Keough (2002) describe it as an 'interim' calculation of the population mean which, according to Johnson et al. (2007), is equivalent to confidence intervals in the univariate situation. If 95% confidence circles are plotted around each mean, significantly different sample groups can be identified visually on the plot (by their lack of overlap).

2.2. An example: The use of CVA in a volunteer-employed photography (VEP) study

This section presents an example of the use of CVA. The dataset used in the study was collected for a tourism planning study in the St David's area of Pembrokeshire Coast National Park, Wales (see Balomenou & Garrod, 2014; this paper presents a different analysis of the data collected in that study). Tourists and residents were given cameras, diaries and a demographic survey, and were asked to photograph positive and negative aspects of holidaying and living in the area. A brief description of the dataset is presented in Table 1:

Table 1: Study dataset in numbers

Total number of participants	278
Overall return rate	64.7% (51.2% locals, 76.5% tourists)
Average survey time per participant	21 minutes
Total data collection time for main study	98 hours
Number of photographs analysed	1496

CVA analysis of this data used only the variables that were already expressed in quantitative form or could sensibly be converted into such. These are shown in Table 2.

Table 2: Survey questions data drawn for the quantitative analysis

Tourists	**Locals**
Question 3: What is your main activity during your visit?	Question 2: How long have you lived in St David's peninsula?
Question 4: Why have you chosen to visit Pembrokeshire Coast National Park?	Question 4: Is your job related to the tourism industry in any way?
Question 5: What is it that you value most about this area?	Question 5: What do you think is special about Pembrokeshire Coast National Park?
Question 6: Have you visited Pembrokeshire Coast National Park before?	Question 6: What is it that you value most about this area?
Question 7: Is this the start, middle or end of your holiday?	Question 9: How might the area be improved?
Question 8: Are you going to spend all your holiday in the St David's area?	Question 10: Given the chance would you ever think of moving elsewhere in this country?
Question 10: How might the area be improved?	Question 11: Our National Parks are under a lot of pressure. Are there any aspects of the area that, if changed, would mean you wouldn't enjoy living in Pembrokeshire Coast National Park anymore?
Question 12: Our National Parks are under a lot of pressure. Are there any aspects of the area that, if changed, would mean that you would not choose to come back to Pembrokeshire Coast National Park for your holidays?	

The software used to run the CVAs for this study was devised by Dr David Causton, from the Institute of Biological, Earth and Rural Sciences at the University of Wales in Aberystwyth and has been used in multiple occasions in biology (Bussell et al., 2008; Johnson et al., 2007). Other software available in the market include CVAGen6 AND PCAGen6.

The data was extracted by coding the answers to these questions. There are three reasons these questions were used. First, the answers to them could be grouped effectively and researcher interpretation was minimal. Second, one of the objectives of the analysis was to compare photos captured by different user groups, so the questions and the answers needed to be comparable. Third, the decision to run a satisfactory number of tests and get the maximum amount of information from the data collected: the data collected from the rest of the questions asked in the survey would be used in the analysis of the survey and the in-depth analysis of all the elements of the technique together.

Maintaining data integrity and avoiding researcher bias was imperative. Thus, instead of the researchers constructing the variables according to their own interpretation of the face value of the photographs, the coding system was based on interviews with the general public and their assessment of the photograph content. Thirty photos were selected randomly from the dataset and copies were placed on a board that could be easily transported. The board was approximately 1m x 80cm and could hold a maximum of 30 photographs. Interviews took place in three different locations in Aberystwyth, another seaside town in the same part of Wales, among people from a similar range of age groups and user groups to those in Pembrokeshire. Stratified sampling was used, based on data drawn from the UK census regarding age and gender. Participants were simply asked to describe what they could see in five photographs of their choice.

Seven sets of variables were produced in the process of identifying the variables for the coding process. After each set was produced, its selection was challenged by the research team and an improved version was produced, which was again challenged and so on. The final set of 30 variables that would be used as a basis for the coding were identified in the seventh attempt and can be seen in Table 3.

Table 3: Thirty variables identified after the interviews

A. Overall percentage	
Water	Natural and man-made features: sea, river, marina, jetty, harbor
Sky, blue	
Sky, clouds	
People	
Trees	
Vegetation	Grass, fern, bracken
Flowers	
Beach	Shingle, sand, when tide is out
Rocks/ hills	In the distance and when this is what was captured, natural features
Signs	Road sign, walking path signs, cycling signs, advertisements, speed signs, etc

Animals	
"Coastal Path"	
Heritage buildings	St David's Cathedral, Treffin's Mill, Solva Mill, etc
Other buildings	
Means of transport	
Other man-made features	Roads, fences, tomb stones, car parks, rubbish bins, benches, chairs, tables
Rubbish	
Tourism paraphernalia	Wind breaks, beach mats, tents, umbrellas
B. Specific, units	
People	Standing, sitting, engaged in activities
Trees	
Signs	
Dogs	
Horses	
Other animals	Mammals, insects, birds, excluding people
Heritage buildings	St David's Cathedral, Treffin's Mill, Solva Mill, etc
Other buildings	
Rubbish bins	
Cars	
Boats	
Flowers	

Initially, 500 randomly selected photographs were coded using these 30 variables. A double-blind coding process was used to enable inter-coding reliability statistics to be calculated. To maintain the integrity of the dataset, CVA was applied to this dataset and it became apparent that 12 variables were responsible for 95% of total variation. These 12 variables (Table 4) were then used to code the rest of the dataset. This greatly reduced the amount of time and resources required to code up the remaining two-thirds of the photographs.

Table 4: The final 12 variables used in the CVA coding

Variable no. 1	Blue sky (proportion of photograph area)
Variable no. 2	Cloudy sky (proportion)
Variable no. 3	People (proportion)
Variable no. 4	Animals (proportion)

Variable no. 5	Car interior (proportion)
Variable no. 6	Other man-made features (proportion)
Variable no. 7	Tourism paraphernalia (proportion)
Variable no. 8	People (number visible in photograph)
Variable no. 9	Signs (number)
Variable no.10	Horses (number)
Variable no.11	Heritage buildings (number)
Variable no.12	Flowers (number)

Twelve hypotheses were then constructed, based on information from the literature and the data from the demographic questionnaires. These hypotheses were then tested using CVA. A high proportion of the original variation (99% to 100%) could be explained in relation to hypotheses with relatively few variables. One such hypotheses will be presented here to illustrate the success of using CVA for this dataset.

> Hypothesis 11: There are significant differences between the photographs taken by members of the local community compared to visitors according to what they value the most about the area.

Although it does not explain the highest proportion of the variation, it is used to indicate how CVA successfully analyses this complicated and rich dataset. The CVA thus compared photographs taken by locals and tourists according to people's perception about what they most value about the area. Eight groups were thus formed, as shown in Table 5:

Table 5: CVA 11 populations

Locals	Tourists
No overdevelopment	No overdevelopment
Quality of life	Quality of life
Location	Location
Community	Other

Table 6: Eigenvalues and canonical correlations

Root No.	Eigenvalue	Pct.	Cum. Pct.
1	0.1299	57.5923	57.5923
2	0.0449	19.9005	77.4927
3	0.0230	10.1863	87.6791
4	0.0132	5.8513	93.5304
5	0.0101	4.4707	98.0011
6	0.0033	1.4810	99.4822
7	0.0012	0.5178	99.9999

The CVA plot (Figure 2) explains 87.7% of the total original variation. There are significant differences between the photographs taken by residents compared to visitors, according to what they value the most about the area. The only two groups whose photographs were not significantly different were locals who appreciate the limited scale of development in the area, and tourists who appreciate the quality of life in the area.

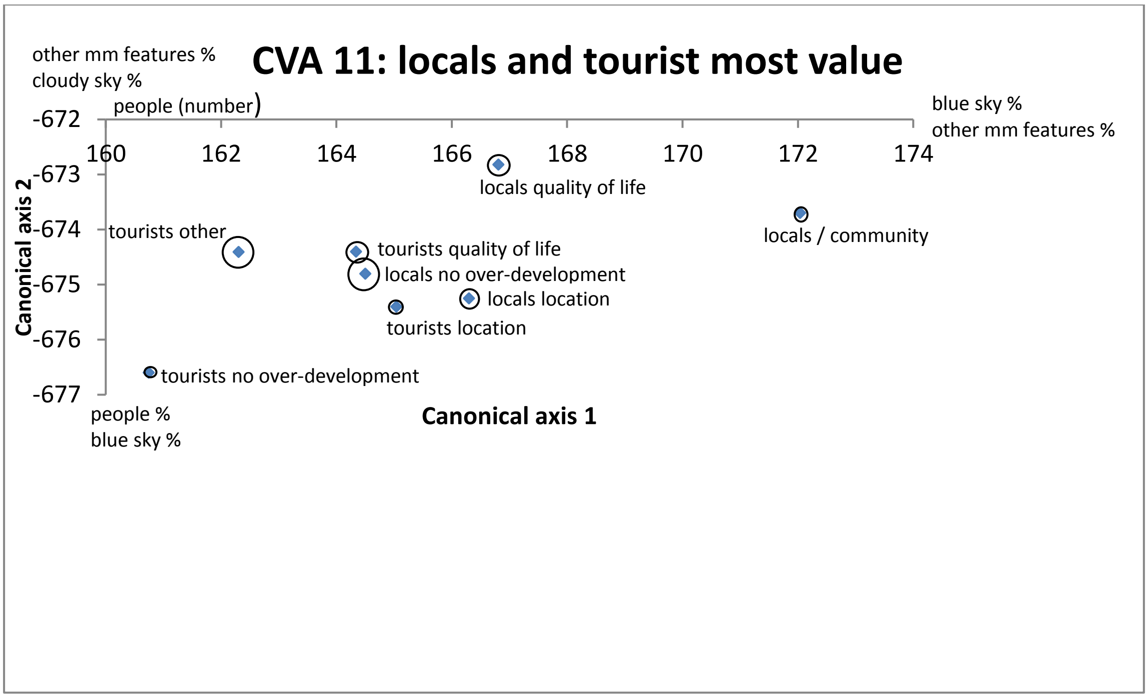

Figure 2: CVA 11 - what locals and tourists most value

The two groups that were placed opposite on both axes were 'tourists who most valued that the area is not overdeveloped' and 'locals who most valued the sense of community in the area'. Participants who fell into the first of these groups tended to include more people in their photographs, and participants in the second group tend to include more blue sky and man-made features.

To verify the validity of the coding of the photographs, a third of the photographs were blind-double coded by an independent researcher. The researcher coded the 500 randomly chosen photographs the principal researcher had used to narrow down the number of the original 33 variables to 12. Both sets of coding were plotted and similarity was observed.

3. Insights and future research

The case study identifies three benefits of using CVA in analysing 'big' visual data. Firstly, it shows how CVA can be used to justify a reduction in the dimensionality of multivariate data. In this case, the identified variables were reduced from 30 to 12. This made a considerable reduction in coding time. It took the researchers almost two months to code 30 variables for 500 photos, implying the need for approximately another four months to complete the remaining 996. Using CVA allowed the elimination of variables that were common to all participants and could not be used to differentiate between photographs, thus reducing the number of variables that needed to be coded up.

Secondly, the richness of the dataset was not compromised in the process. Despite the reduction in the number of the variables, the reduced variable set was responsible for 95% of the total original variation. Future researchers can be confident that by using a robust coding technique followed up by CVA, they can reduce the dimensionality of their dataset without compromising its depth and richness.

Thirdly, an advantage of CVA is that the photos can be traced back to those who took them. This enabled the discrimination and identification of structures and inter-relationships within the multivariate statistical population (Bussell et al., 2008). These were associated with particular sorts of people and differences of opinion among different user groups of the same area. The analysis of the photographs indicated, *inter alia*, that there are significant differences between people who were born in the area compared with those who moved in the area, locals and tourists who were happy to see the character of the area change and those who were not, tourists depending on the stage of their holiday, and so on.

CVA is subject to some limitations, including that it is the CVs that are the ones to be interpreted, rather than the original variables, and that interpretation takes place in pairs. Considering that solutions depend on the level of correlation between and within sets, it is likely that a modification in a variable of the one set will have implications to the structure of the other set.

The findings presented here are invaluable, given the purpose of the study, which was to attempt to identify differences in the destination image construed by visitors, that perceived by residents, and that proposed by marketers (an aim also adopted by Michaelidou et al., (2013). Indeed, Markwell (1997), Urry (2002), and Urry and Larsen (2011) have all observed that the tourism industry can shape a destination image in ways that may be dissonant with that of residents or, indeed, be consistent with the actual experience of tourists. MacKay and

Couldwell (2004), meanwhile, suggest that keeping a visual inventory of the visitors' images of a site can be especially useful for informing marketing efforts. The management implications of this kind of analysis of 'big' visual data are thus substantial. Future research can attempt to identify an 'optimal' image for use in marketing a destination to a target market: one that has all the components that will appeal particularly to their aesthetics. It would therefore be useful to examine whether marketing campaigns can be more effective if they employ such techniques. The analysis of differences in resident and tourist perceptions of the impacts of tourism can also be useful to complement tourism planning decision-making.

4. Conclusions

This research note has demonstrated the utility of CVA as a dimensionality-reduction technique for use with 'big' visual data. Such data is increasingly becoming available, both through the use of PGI techniques and 'found' data available in various media, notably the huge amount of user-generated content on photograph-sharing websites. Using CVA in this way can make the meaningful analysis of such data considerably less resource-hungry, rendering it more tenable for use by destination marketing organisations, tourism planning departments, tour operators and other stakeholders. In an era of ever-shrinking research budgets this represents too an important an option to be overlooked, as it has tended to be to date. CVA also has distinct advantages over PCA and Factor Analysis in achieving this task, including the calculation of a meaningful correlation statistic, preservation of data integrity and the availability of graphical display of data patterns and inter-relationships, making the findings intelligible to a wide audience. As such, this research note argues that CVA opens up the potential for visual tourism research methods as never seen before.

References

Ahmed, S.A. (1986). Understanding residents' reaction to tourism marketing strategies. *Journal of Travel Research, 25*(2), 13-18.

Albrecht, G.H. 1980. Multivariate analysis and the study of form, with special reference to canonical variate analysis. *Integrative and Comparative Biology*, 20, 679-693.

Allen, L.R., Long, P.T., Perdue, R.R., and Kieselbach, S. (1988). The impact of tourism development on residents' perceptions of community life. *Journal of Travel Research, 27*(1), 16-21.

Alsberg, B. K., Wade, W. G. and Goodacre, R. (1998). Chemometric analysis of diffuse reflectance-absorbance fourier transform infrared spectra using rule induction methods: application to the classification of eubacterium species. *Applied Spectroscopy*, 52.

Baloglu, S., and Uysal, M. (1996). Market segments of push and pull motivations: A canonical correlation approach. *International Journal of Contemporary Hospitality Management, 8(*3), 32-38.

Balomenou, N. and Garrod, B. (2014). Using volunteer-employed photography to inform tourism planning decisions: A study of St David's Peninsula, Wales. *Tourism Management*, 44, 126-139.

Beaghen, M. (1997). Canonical variate analysis and related methods with longitudinal data. PhD Dissertation. Virginia Polytechnic Institute and State University.

Brown, S. A., Goldman, M. S., Inn, A. and Anderson, L. R. (1980). Expectations of reinforcement from alcohol: Their domain and relation to drinking patterns*. Journal of consulting and Clinical Psychology*, 48(4), 419-426.

Bussell, J. A., Gidman, E. A., Causton, D. R., Gwynn-Jones, D., Malham, S. K., Jones, M. L. M., Reynolds, B. and Seed, R. (2008). Changes in the immune response and metabolic fingerprint of the mussel, Mytilus Edulis (Linnaeus) in response to lowered salinity and physical stress. *Journal of Experimental Marine Biology and Ecology,* 358**,** 78-85.

Causton, D. (2008). Statistics for research biologists, Canonical Variate Analysis (Discriminant Function Analysis), lecture notes. Aberystwyth University.

Chatfield, C. and Collins, A. J. (1980). *Introduction to Multivariate Analysis*. US: Springer.

Crang, M. (2003). Qualitative methods: Touchy, feeling, look-see? *Progress in Human Geography*, 27, 494-504.

Cutler, S. Q., Doherty, S. and Carmichael, B. (2016). Immediacy, photography and memory: The tourist experience of Machu Picchu. *In:* Bourdeau, L., Gravari-Barbas. M. and Robinson, P. (Eds.) *World heritage, tourism and identity: Inscription and co-production*, (pp. 131-146). London: Routledge.

Darlington, R. B., Weinberg, S. L. and Walberg, H. J. (1973). Canonical variate analysis and related techniques. *Review of Educational Research*, *43*(4), 433-454.

Dwyer, L., Mellor, R., Livaic, Z., Edwards, D. and Kim, C. (2004). Attributes of destination competitiveness: A factor analysis. *Tourism Analysis, 9*, 91-101.

Fung, C. K. and Jim, C. (2015). Unraveling Hong Kong Geopark experience with visitor-employed photography method. *Applied Geography, 62*, 301-313.

Garrod, B. (2008). Exploring place perception: A photo-based analysis. *Annals of Tourism Research,* 35(2), 381-401.

Gittins, R. (1985). *Canonical analysis: a review with applications in ecology*. Berlin: Springer Verlag.

Gonzalez, A.M., and Bello, L. (2002). The construct 'lifestyle' in market segmentation: The behaviour of tourist consumers. *European Journal of Marketing, 36*(1/2), 51-85.

Hair, J.F. Jr., Anderson, R.E., Tatham, R.L., and Black, W.C. (1998). *Multivariate data analysis* (5th eds.). Upper Saddle River, NJ: Prentice-Hall.

Hosany, S., Ekinci, Y., and Uysal, M. (2006). Destination image and destination personality: An application of branding theories to tourism places. *Journal of Business Research, 59*(5), 638-642.

Hotelling, H. (1935). The most predictable criterion. *Journal of Educational Psychology*, 26, 139-142.

Jang, S. S., Hu, C., and Bai, B. (2006). A canonical correlation analysis of e-relationship marketing and hotel financial performance. *Tourism and Hospitality Research, 6*(4), 241-250.

Johnson, H. E., Lloyd, A. J., Mur, L. A. J., Smith, A. R. and Causton, D. (2007). The application of MANOVA to analyse Arabidopsis thaliana metabolomic data from factorially designed experiments. *Metabolomics*, 3, 517-530.

Kim, H. and Stepchenkova, S. (2015). Effect of tourist photographs on attitudes towards destination: Manifest and latent content. *Tourism Management*, 49, 29-41.

Konijn, E., Sluimer, N. and Mitas, O. (2016). Click to share: Patterns in tourist photography and sharing. *International Journal of Tourism Research, 18*(6), 525–632

Larimore, W.E. (1997). Optimal reduced rank modelling, prediction, monitoring, and control using canonical variate analysis. IFAC 1997 International Symposium on Advanced Control of Chemical Processes, 61-66.

Lee, W.-Y. (2016) Multi-modal learning over user-contributed content from cross-domain social media. Thirtieth AAAI Conference on Artificial Intelligence, 2016.

MacKay, K. J. and Couldwell, C. M. (2004). Using visitor-employed photography to investigate destination image. *Journal of Travel Research*, 42(4), 390-396.

Madden, T.J., Hewett, K. and Roth, M.S. (2000). Managing images in different cultures: A cross-national study of color meanings and preferences. *Journal of International Marketing*, 8, 90-107.

Markwell, K.W. (1997). Dimensions of photography in a nature-based tour. *Annals of Tourism Research*, 24, 131-155.

Martens, H and Neaes, T. (1989). *Multivariate calibration*, New York, J.Wiley & Sons Ltd.

Michaelidou, N., Siamagka, N. T., Moraes, C. and Micevski, M. (2013). Do marketers use visual representations of destinations that tourists value? Comparing visitors' image of a destination with marketer-controlled images online. *Journal of Travel Research*, *55,* 588-602.

Muller, K. 1982. Understanding canonical correlation through the general linear model and principal components. *The American Statistician, 36*(4), 342-354.

Oh, H. C., Uysal, M., and Weaver, P.A. (1995). Product bundles and market segments based on travel motivations: A canonical correlation approach. *International Journal of Hospitality Management, 14*(2), 123-137.

Pan, S., Lee, J. and Tsai, H. (2014). Travel photos: Motivations, image dimensions, and affective qualities of places. *Tourism Management*, 40, 59-69.

Pearce, P.L., Wu, M.-Y. and Chen, T. (2015). The spectacular and the mundane: Chinese tourists' online representations of an iconic landscape journey. *Journal of Destination Marketing & Management*, 4, 24-35.

Pérez, V., Guerrero, F., González, M., Pérez, F. and Caballero, R. (2013). Composite indicator for the assessment of sustainability: The case of Cuban nature-based tourism destinations. *Ecological Indicators*, 29, 316-324.

Quinn, G. P. and Keough, M. J. (2002). *Experimental design and data analysis for biologists*. UK: Cambridge University Press.

Rao, C.R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological),* 10**,** 159-203.

Rao, C.R. (2005). *Handbook of statistics. Data mining and data visualization,* San Diego, CA, Elsevier.

Russell, E. L., Chiang, L. H. and Braatz, R. D. (2000). Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 51(1), 81-93.

Scarles, C.E. (2011). Eliciting embodied knowledge and response: Respondent-led photography and visual autoethnography. *In:* Rakic, T. and Chambers, D. (Eds.) *An Introduction to Visual Research Methods in Tourism.* London; Routledge.

Schulz, H., Baranska, M., H.-H.Belz, Rösch, P., Strehle, M. A. and Popp, J. (2004). Chemotaxonomic characterisation of essential oil plants by vibrational spectroscopy measurements. *Vibrational Spectroscopy*, 35, 81-86.

Straumann, R. K., Coltekin, A. and Andrienko, G. (2014). Towards (re)constructing narratives from georeferenced photographs through visual analytics. *The Cartographic Journal*, 51, 152-165.

Sun, M., Ryan, C. and Pan, S. (2014). Assessing tourists' perceptions and behaviour through photographic and blog analysis: The case of Chinese bloggers and New Zealand holidays. *Tourism Management Perspectives*, 12, 125-133.

Taylor, J., King, R. D., Altmann, T. and Fiehn, O. (2002). Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Bioinformatics*, 18, S241-248.

Tran, X., Dauchez, C. and Szemik, A. M. (2013). Hotel brand personality and brand quality. *Journal of Vacation Marketing*, 19(4), 329-341.

Tran, X. and Ralston, L. (2006). Tourist preferences influence of unconscious needs. *Annals of Tourism Research*, 33, 424-441.

Urry, J. (2002). *The tourist gaze*, London, Sage.

Urry, J., and Larsen, J. (2011). *The tourist gaze 3.0*, London, Sage.

Uysal, M., and Jurowski, C. (1994). Testing the push and pull factors. *Annals of Tourism Research, 21*(4), 844-846.

Uysal, M., and O'Leary, J.T. (1986). A canonical analysis of international tourism demand. *Annals of Tourism Research, 13*(4), 651-655.

Wong, S., and Lau, E. (2001). Understanding the behavior of Hong Kong Chinese tourists on group tour packages. *Journal of Travel Research, 40*(1), 57-67.