2011

Three good reasons NOT to use factor-cluster segmentation

Sara Dolnicar
University of Wollongong, s.dolnicar@uq.edu.au

Bettina Grun
Institute for Statistics and Mathematics, WU Wirtschaftsuniversität Wien, bettina@uow.edu.au

Publication Details

This conference paper was originally published as Dolnicar, S and Grun, B, Three good reasons NOT to use factor-cluster segmentation, CAUTHE 2011 : 21st CAUTHE National Conference, Adelaide, Australia, 8-11 February 2011.

Three good reasons NOT to use factor-cluster segmentation

Abstract
Market segmentation is very popular both in tourism industry and among tourism researchers. Tourism industry uses it to identify homogenous subsets of tourists and to select the most suitable of them to target over the medium and long term. Tourism researchers use it to gain a deeper understanding of the heterogeneity of consumer behaviour among tourists. There are two basic forms of market segmentation: a priori (Mazanec, 2000) or commonsense segmentation (Dolnicar, 2004) and post-hoc (Myers and Tauber, 1977), a posteriori (Mazanec, 2000), or data-driven segmentation (Dolnicar, 2004). In commonsense segmentation the users determine in advance which tourist characteristic should be used to group tourists. Typically one single characteristic is used (e.g. age, country of origin, gender), tourists are split according to this criterion and then the resulting groups are described. This makes commonsense segmentation a very simple procedure with no major methodological traps that could lead to solutions of questionable validity. The same does not hold for data-driven segmentation. In data-driven segmentation a set of variables is used as the so-called segmentation base. A mathematical algorithm is then required to determine groups of respondents who have responded similarly to the variables included in the segmentation base. This process is not particularly complex, but it does require solid understanding of the foundations of clustering because a number of decisions need to be made by the data analyst throughout the clustering process which – if made wrongly – can lead to segmentation solutions of questionable validity. One of the problems that data-analysts frequently face is that the number of variables in the data set (or the number of questionnaire questions selected to be included in the segmentation base) is too high for the sample size. The recommended ratio is $5*2^k$ or at least 2^k (Formann, 1984) which means that a sample size including 1000 respondents does not permit clustering with more than 9 variables in the segmentation base. If the data analyst was not involved in the questionnaire design they are frequently asked to use a set of 20 or 30 variables (e.g. benefits sought, travel motivations, emotions, pre-trip information sources, vacation activities etc.), which typically cannot be accommodated with the available data sets. The typical way of dealing with this problem of having too many variables for a given sample size is to conduct something referred to as "factor-cluster segmentation". This term appears to have been introduced by Smith (1989) as it is not used outside of the tourism discipline. It involves first factor analysing the full set of variables included in the segmentation base and then using the resulting factor scores in the cluster analysis. There are (at least) three good reasons why this approach should not be used:

1. Firstly, the segmentation analysis is only based on part of the information collected from respondents. A high percentage of variance explained by the factor analysis in survey data sets is 60%. This still means that 40% of the information contained in the data is thrown away before the segmentation analysis is even conducted. The segmentation solution is therefore based only on slightly more than half of the information that was originally deemed to be important when the data was collected and when the segmentation base was selected.
2. Secondly, the segmentation solution is identified in a transformed space and that means the very nature of the data is altered before the segmentation is undertaken (Arabie and Hubert, 1994; Ketchen and Shook, 1996). It is therefore not legitimate to interpret the solution using the original variables. Instead factors have to be used to interpret the segmentation solution. But factors are an abstraction of items. As a consequence, it is not easy to derive direct marketing action implications from factors which are composites of a number of items, often including some which are not logically related.
3. Finally, and most importantly, factor-cluster analysis has been shown to perform worse in identifying the correct data structure in experiments with artificial data (Sheppard, 1996; Dolnicar and Grün,

2008) than running cluster analysis directly on the raw, untransformed data. Even if the artificial data sets were constructed using a factor-analytic model, which should give the factor-cluster segmentation approach a competitive advantage, the factor-cluster analysis did not perform substantially better.

In contradiction to current practice in tourism research but in line with the recommendations from leading clustering experts (Arabie and Hubert, 1994) as well as researchers who have conducted comparative studies using factor-cluster analysis and clustering without pre processing (Sheppard, 1996; Dolnicar and Grün, 2008), it has to be concluded that factor-cluster analysis is indeed an "outmoded and statistically insupportable practice" (Arabie and Hubert, 1994) and should not be used in data-driven tourism segmentation studies. A number of simple alternatives are available to data analyst to deal with too many variables. Optimally, the data analyst is involved in preparing data the collection and can either ensure that no redundant variables are included or that the sample size chosen is sufficient to allow clustering with the number of variables included. This is the optimal solution as it solves the problem at its origin. If the data analyst is not consulted before data collection, another alternative approach is to eliminate redundant variables from the segmentation base before segmenting. If users are still interested in segment differences with respect to variables that were eliminated, these can be computed after the segmentation task is completed. Whichever option is chosen, using raw data is preferable to transformed data when looking for groups of individuals in a space defined by carefully selected pieces of information (survey questions).

Disciplines
Business | Social and Behavioral Sciences

Publication Details
This conference paper was originally published as Dolnicar, S and Grun, B, Three good reasons NOT to use factor-cluster segmentation, CAUTHE 2011 : 21st CAUTHE National Conference, Adelaide, Australia, 8-11 February 2011.

Three good reasons NOT to use Factor-Cluster Segmentation

Sara Dolnicar*
Institute for Innovation in Business and Social Research
Faculty of Commerce, University of Wollongong,

Wollongong, NSW 2522, Australia
Telephone: (61 2) 4221 3862, Fax: (61 2) 4221 4154
sara_dolnicar@uow.edu.au

Bettina Grün*
Institute for Statistics and Mathematics, WU Wirtschaftsuniversität Wien
and Institute for Innovation in Business and Social Research

Augasse 2-6, A-1090 Vienna, Austria
Telephone: (43 1) 31336 5032, Fax: (43 1) 31336 774
bettina.gruen@wu.ac.at

* Authors listed in alphabetical order.

Working Paper

Market segmentation is very popular both in tourism industry and among tourism researchers. Tourism industry uses it to identify homogenous subsets of tourists and to select the most suitable of them to target over the medium and long term. Tourism researchers use it to gain a deeper understanding of the heterogeneity of consumer behaviour among tourists.

There are two basic forms of market segmentation: a priori (Mazanec, 2000) or commonsense segmentation (Dolnicar, 2004) and post-hoc (Myers and Tauber, 1977), a posteriori (Mazanec, 2000), or data-driven segmentation (Dolnicar, 2004). In commonsense segmentation the users determine in advance which tourist characteristic should be used to group tourists. Typically one single characteristic is used (e.g. age, country of origin, gender), tourists are split according to this criterion and then the resulting groups are described. This makes commonsense segmentation a very simple procedure with no major methodological traps that could lead to solutions of questionable validity.

The same does not hold for data-driven segmentation. In data-driven segmentation a set of variables is used as the so-called segmentation base. A mathematical algorithm is then required to determine groups of respondents who have responded similarly to the variables included in the segmentation base. This process is not particularly complex, but it does require solid understanding of the foundations of clustering because a number of decisions need to be made by the data analyst throughout the clustering process which – if made wrongly – can lead to segmentation solutions of questionable validity.

One of the problems that data-analysts frequently face is that the number of variables in the data set (or the number of questionnaire questions selected to be included in the segmentation base) is too high for the sample size. The recommended ratio is $5*2^k$ or at least 2^k (Formann, 1984) which means that a sample size including 1000 respondents does not permit clustering

with more than 9 variables in the segmentation base. If the data analyst was not involved in the questionnaire design they are frequently asked to use a set of 20 or 30 variables (e.g. benefits sought, travel motivations, emotions, pre-trip information sources, vacation activities etc.), which typically cannot be accommodated with the available data sets.

The typical way of dealing with this problem of having too many variables for a given sample size is to conduct something referred to as "factor-cluster segmentation". This term appears to have been introduced by Smith (1989) as it is not used outside of the tourism discipline. It involves first factor analysing the full set of variables included in the segmentation base and then using the resulting factor scores in the cluster analysis.

There are (at least) three good reasons why this approach should not be used:

1. Firstly, the segmentation analysis is only based on part of the information collected from respondents. A high percentage of variance explained by the factor analysis in survey data sets is 60%. This still means that 40% of the information contained in the data is thrown away before the segmentation analysis is even conducted. The segmentation solution is therefore based only on slightly more than half of the information that was originally deemed to be important when the data was collected and when the segmentation base was selected.
2. Secondly, the segmentation solution is identified in a transformed space and that means the very nature of the data is altered before the segmentation is undertaken (Arabie and Hubert, 1994; Ketchen and Shook, 1996). It is therefore not legitimate to interpret the solution using the original variables. Instead factors have to be used to interpret the segmentation solution. But factors are an abstraction of items. As a consequence, it is not easy to derive direct marketing action implications from factors which are composites of a number of items, often including some which are not logically related.
3. Finally, and most importantly, factor-cluster analysis has been shown to perform worse in identifying the correct data structure in experiments with artificial data (Sheppard, 1996; Dolnicar and Grün, 2008) than running cluster analysis directly on the raw, untransformed data. Even if the artificial data sets were constructed using a factor-analytic model, which should give the factor-cluster segmentation approach a competitive advantage, the factor-cluster analysis did not perform substantially better.

In contradiction to current practice in tourism research but in line with the recommendations from leading clustering experts (Arabie and Hubert, 1994) as well as researchers who have conducted comparative studies using factor-cluster analysis and clustering without pre processing (Sheppard, 1996; Dolnicar and Grün, 2008), it has to be concluded that factor-cluster analysis is indeed an "outmoded and statistically insupportable practice" (Arabie and Hubert, 1994) and should not be used in data-driven tourism segmentation studies.

A number of simple alternatives are available to data analyst to deal with too many variables. Optimally, the data analyst is involved in preparing data the collection and can either ensure that no redundant variables are included or that the sample size chosen is sufficient to allow clustering with the number of variables included. This is the optimal solution as it solves the problem at its origin. If the data analyst is not consulted before data collection, another alternative approach is to eliminate redundant variables from the segmentation base before segmenting. If users are still interested in segment differences with respect to variables that were eliminated, these can be computed after the segmentation task is completed. Whichever

option is chosen, using raw data is preferable to transformed data when looking for groups of individuals in a space defined by carefully selected pieces of information (survey questions).

Acknowledgements

This research was supported by the Australian Research Council (through grants DP0557257, LX0559628 and LX0881890) and the Austrian Science Foundation (through grants P17382 and T351).

References

Aldenderfer, M.S., & Blashfield, R.K. (1984). *Cluster Analysis*. Beverly Hills: Sage Publications.

Arabie, P., & Hubert, L. (1994). Cluster Analysis in Marketing Research. In R. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 160-189). Cambridge: Blackwell.

Dolnicar, S. (2002). Review of Data-Driven Market Segmentation in Tourism. *Journal of Travel and Tourism Marketing*, 12(1), 1-22.

Dolnicar, S. (2004). Beyond "Commonsense Segmentation" – a Systematics of Segmentation Approaches in Tourism. *Journal of Travel Research*, 42(3), 244-250.

Dolnicar, S., & Grün, B. (2008). Challenging "Factor-Cluster Segmentation". *Journal of Travel Research*, 47(1), 63-71.

Formann, A.K. (1984). *Die Latent-Class-Analyse: Einführung in die Theorie und Anwendung.* Weinheim: Beltz.

Ketchen D.J. jr., & Shook, C.L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*, 17, 441-458.

Mazanec, J. (2000). Market Segmentation. In Jafari, J. (Ed.), *Encyclopedia of Tourism*. London: Routledge.

Myers, J.H., & Tauber, E. (1977). *Market structure analysis*. Chicago: American Marketing Association.

Sheppard, A.G. (1996). The sequence of factor analysis and cluster analysis: Differences in segmentation and dimensionality through the use of raw and factor scores. *Tourism Analysis*, 1, 49-57.

Smith, S.L.J. (1989). *Tourism Analysis: a Handbook*. Harlow: Longman.