Research Article

Required Sample Sizes for Data-Driven Market Segmentation Analyses in Tourism

Journal of Travel Research
2014, Vol. 53(3) 296–306

DOI: 10.1177/0047287513496475
jtr.sagepub.com
SAGE

Sara Dolnicar[1], Bettina Grün[2], Friedrich Leisch[3], and Kathrin Schmidt[4]

Abstract

Data analysts in industry and academia make heavy use of market segmentation analysis to develop tourism knowledge and select commercially attractive target segments. Within academic research alone, approximately 5% of published articles use market segmentation. However, the validity of data-driven market segmentation analyses depends on having available a sample of adequate size. Moreover, no guidance exists for determining what an adequate sample size is. In the present simulation study using artificial data of known structure, the impact of the difficulty of the segmentation task on the required sample size is analyzed in dependence of the number of variables in the segmentation base. Under all simulated data circumstances, a sample size of 70 times the number of variables proves to be adequate. This finding is of substantial practical importance because it will provide guidance to data analysts in academia and industry who wish to conduct reliable and valid segmentation studies.

Keywords

market segmentation, cluster analysis, sample size, k-means, simulation study

Introduction

Market segmentation analysis represents one of the key techniques in tourism research used to develop knowledge about consumer behavior of and gain market intelligence about tourists. In the academic tourism literature, approximately 5% of articles published between 1986 and 2005 were related to market segmentation (Zins 2008), which shows that the topic represents a key methodology used by academic tourism researchers to develop knowledge. Between 2011 and 2012 the three main publication outlets for data-driven segmentation studies in tourism (*Journal of Travel Research, Tourism Management*, and the *Journal of Travel and Tourism Marketing*) have published 29 market segmentation studies in total (representing 6% of the articles published during those two years in the selected three journals). Of these 29 articles, 17 used an a posteriori (Mazanec 2000) or data-driven (Dolnicar 2004) segmentation approach. In the tourism industry, many key strategic decisions are made based on market segmentation studies, including the choice of a target segment and the development of an entire marketing mix to suit this segment. It is therefore important that segmentation studies are conducted in a valid way.

Despite the popularity of data-driven market segmentation analysis in tourism (recent examples published in this journal alone include Nicolau 2012; Masiereo and Nicolau 2012; Weaver and Lawton 2011; and Needham et al. 2011), areas of segmentation analysis remain where no recommendations are given to data analysts (Dolnicar and Lazarevski 2009). One such area relates to the sample size required for data-driven market segmentation analysis: no guidelines exist that would allow the data analyst to ensure that the sample size available is sufficient for the analysis, because recommendations or techniques used to determine sample sizes for other statistical techniques cannot be used (such as for example sample sizes derived from power analysis for statistical hypothesis testing and from optimal design for regression analysis). Nevertheless, insufficient sample sizes can have serious negative consequences for the validity of the market segmentation solution, where validity implies that the true structure of the data has been identified.

To understand the potential negative consequences of insufficient sample sizes in data-driven market segmentation studies, we must bear in mind that segmentation analysis is

[1]University of Queensland, Brisbane, Australia
[2]Johannes Kepler University Linz, Linz, Austria
[3]University of Natural Resources and Life Sciences, Vienna, Austria
[4]Telefónica Germany, Munich, Germany

Corresponding Author:
Sara Dolnicar, University of Queensland, St Lucia, Brisbane, Australia.
Email: s.dolnicar@uq.edu.au

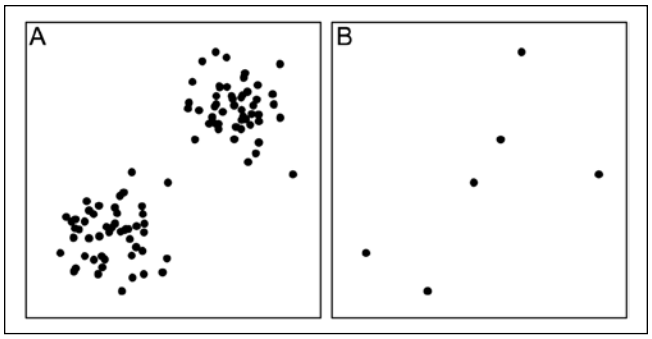

Figure 1. (A) A sample with 100 observations from a bivariate normal mixture with two equally sized components. (B) A subset with 6 observations from the 100.

exploratory by its very nature, and consequently, any segmentation algorithm will always arrive at a grouping of individuals, whether or not the grouping is worthy. Furthermore, in segmentation analysis, each variable represents one dimension in space. For example, a data-driven segmentation study with 20 variables (which is not uncommon in tourism) means that a mathematical problem is solved in 20-dimensional space. To find groups in 20-dimensional space, many data points are required; otherwise no patterns can be detected and any resulting grouping is entirely random. This can be illustrated by imagining the simplest possible situation, that of two-dimensional space, as shown in Figure 1. Figure 1(A) shows 100 data points in this two-dimensional space. It clearly shows that two clusters exist in these data. Figure 1(B) shows the same data situation but with only six data points. Here, it is impossible to determine, based on those six data points only, what the true structure of the data is. Based on the illustration in Figure 1(B), the correct solution may include anything between one and six clusters.

The problem illustrated in Figure 1 becomes exponentially worse as the number of dimensions increases. It is exasperated by the fact that no indicator exists to warn the data analyst if the sample size–to–variable number ratio is critical and thus may lead the data analyst to draw incorrect conclusions.

The issue of selecting an adequate sample size is largely ignored when a posteriori or data-driven segmentation studies are conducted. In a review of 47 data-driven tourism segmentation studies, Dolnicar (2002) highlighted the problem of potentially insufficient numbers of respondents, given large numbers of variables in the segmentation base. Specifically, Dolnicar (2002) reports that the sample sizes of the 47 reviewed studies ranged from a mere 46 to nearly 8000 respondents, with a median of 461. More than one-third of the data sets had fewer than 400 respondents. Simultaneously, the number of variables ranged from 3 to 56, with approximately two-thirds of studies using between 10 and 22 variables in their segmentation bases. The median ratio of the number of respondents divided by the number of variables is 22.4. The correlation between sample size and number of variables is insignificant, indicating that data analysts do not collect larger samples in cases where the data situation is more complex because a high number of variables is included in the segmentation base.

For the present study, we conducted a review similar to that conducted by Dolnicar in 2002 with more recent articles, specifically data-driven market segmentation studies published in the last decade in the *Journal of Travel Research, Tourism Management*, and the *Journal of Travel and Tourism Marketing*. For the purpose of assessing sample sizes and variable numbers, we included only studies that did not reduce the number of variables before segmenting. Typical approaches used in tourism for data preprocessing include factor cluster analysis (which has been shown to lead to suboptimal results; cf. Dolnicar and Grün 2008) or methods that simultaneously select variables and group individuals, such as decision trees (e.g., Van Middelkoop, Borgers, and Timmermans 2003; Dolnicar et al. 2011; Legoherel and Wong 2012) or biclustering (Dolnicar et al. 2012). The results led to similar conclusions as the previous review: although the median has increased significantly to 1,000 (presumably because the collection of survey data has become easier and cheaper) and the average number of variables decreased slightly to 19, no significant correlation exists between sample size and the number of variables used for the segmentation task (Pearson correlation coefficient −0.09) and a substantial number of individual studies can be identified that have critical ratios of sample to number of variables, for example, more than 40 segmentation variables and fewer than 500 respondents.

These results also indicate that in tourism research the sample sizes are at best modest and there is no need to employ subsampling strategies to reduce the computational burden in the segmentation analysis due to large data sets, as suggested for other areas of research where millions of observations are available (cf. Bejarano et al. 2011).

Dolnicar's (2002) results indicate that despite market segmentation being used extensively in the field of tourism research, the fundamental question of how many variables should be used for a certain number of respondents has not yet been explicitly considered, and practically no guidance is available to data analysts with respect to the sample size required.

The contribution of the present study is to derive sample size requirements for data-driven market segmentation analyses. This will allow data analysts to check whether the sample available for their segmentation analysis is sufficient, given the number of variables in the segmentation base, or whether it may be necessary to either collect more data or reduce the number of variables used in the analysis.

In the present study, sample size requirements are derived by conducting an extensive simulation study using artificial data sets whose correct cluster structure is known. Artificial data are required because for empirical survey data the true

segmentation solution is unknown. Consequently, the effects of insufficient sample sizes cannot be studied, because of the lack of a dependent variable (correctness of the segmentation solution). We conduct simulations for a range of scenarios, which have been modeled to be similar in nature to typical empirical tourism data sets to ensure that the final recommendation is adequate—even under the most difficult of data circumstances. Characteristics of typical data sets used in data-driven segmentation studies conducted in the tourism literature have been taken from Dolnicar's (2002) review of data-driven segmentation studies published in tourism.

Prior Work

Only two recommendations about the appropriate ratio of respondents to numbers of variables have been published to date. Both recommendations are not easily accessible to the English-speaking and scientific community, with one being a research monograph in German by a Viennese psychologist (Formann 1984) and the other one a recommendation available from a help page of the add-on package called clusterGeneration (Qiu and Joe 2009) to the statistical software environment R (R Development Core Team 2013).

Formann (1984) proposes including at least 2^d respondents (preferably $5 \cdot 2^d$), where d is the number of variables in the segmentation base. This recommendation is provided in a very specific context, namely, in the context of goodness-of-fit testing with the χ^2 test for latent class analysis grouping binary data. Therefore, strictly speaking, the recommended rule of thumb for sample size requirements applies only to model selection, not to model estimation.

Qiu and Joe (2009) suggest that the sample size should amount to a minimum of 10 times the number of variables in the segmentation base times the number of clusters ($10 \cdot d \cdot k$, with d representing the number of segmentation variables and k representing the number of clusters or segments) in cases where the clusters are of equal size. If clusters are of unequal size, the smallest cluster is required to contain at least $10 \cdot d$ respondents. This rule of thumb, which leads to substantially lower sample size requirements than is produced by the rule recommended by Formann (1984), was postulated in the context of artificial data generation for clustering simulation studies. Qiu and Joe (2009) do not provide an explanation or justification for their recommendation. However, we may assume that the rule aims to ensure that the method used for generating covariance matrices does not produce singular matrices. In terms of empirical data analysis, rather than generation of artificial data sets, the problem with the rule propounded by Qiu and Joe (2009) is that the number of clusters is not known in advance.

According to Qiu and Joe (2009), we may assume that more respondents are needed when data contain more segments or clusters. More generally, we might expect that the sample size requirements will increase with the difficulty of the segmentation task.

Levels of difficulty of segmentation tasks have been discussed by Dolnicar and Leisch (2010). They argue that typical survey data situations range from natural clusters (where density clusters are present in the data) to reproducible clusters (where data contains only a weak structure) and constructive clustering (where virtually no structure is in the data that would allow repeated segmentation analysis to lead to the same results). The further apart market segments are, the more likely it is that the underlying structure is one of natural clusters. The closer they are, the more likely that cluster structure cannot even be identified, making constructive clustering necessary, leading to the assumption that the separation between market segments is a key criterion for determining the required sample size.

Finally, it is not uncommon for survey data to contain variables that do not necessarily contribute to understanding the cluster structure of the data. Data analysts may include such variables in the analysis because they do not know in advance whether or not they contribute to the segmentation solution. If they do not contribute, they may instead mask the cluster structure and, as a consequence, lead to less homogeneous clusters. Such variables are referred to as *noisy* variables and are generally assumed to be identically distributed for each segment. Because of their masking effect, the presence of noisy variables renders the clustering task more challenging. According to Qiu and Joe (2006), such variables are normally distributed and independent of non-noisy variables but are not necessarily independent of one another. Noisy data that "contain little clustering information can cause misleading results" (Carmone, Kara, and Maxwell 1999, p. 501). For example, in the case of a benefit segmentation of tourists traveling to Australia, the item "change of environment" could be such a noisy variable, because it is not particularly important to any of the segments, and thus it does not serve as a marker variable for any specific segment and is not associated with any of the other benefits a tourist may seek. Both adventure tourists and culture tourists may be equally interested in a change in environment when taking a vacation. We may assume that higher proportions of noisy variables in the segmentation base increase sample size requirement.

Methodology

Data Generation

Because, as opposed to power calculations for statistical hypothesis testing, there is no direct way of calculating what an adequate sample size is for any given segmentation problem, we used simulation analyses. Simulation studies have one major advantage over studies with empirical survey data: the true cluster structure is known. Consequently, whether any given segmentation solution has identified the cluster structure in the data correctly can easily be assessed.

In statistical hypothesis testing, the performance of a statistical test is measured by its power. Segmentation analysis

Table 1. Overview of Factors Used in the Full-Factorial Design Simulation Study.

Factors	Levels
Total number of variables *d*	10, 16, and 22
Numbers of respondents *n* where *d* is the total number of variables	10·*d*, 20·*d*, 30·*d*, 40·*d*, 50·*d*, 60·*d*, 70·*d*, 80·*d*, 90·*d*, 100·*d*
Number of clusters *k*	3, 4
Proportion of noisy variables *m*	0, 1/4, and 1/2 of the total number of variables *d*
Separation between the clusters *s*	−0.1, 0, 0.1
Total of 3 · 10 · 2 · 3 · 3 = 540 different data settings	

results in a partition of the data, and a natural performance measure is the correctness of the predicted grouping compared to the true grouping. Therefore, correctness represents the key performance criterion and dependent variable in the present study. It is computed as follows: for each simulated respondent, the segment membership resulting from the clustering algorithm is compared to the true membership. Thus, the criterion is how well the original partition of the data is revealed by the clustering algorithm. As noted by Ben-David, Pál, and Simon (2007) the solution that is returned as the "best" one by the clustering algorithm does not need to correspond to the best solution with respect to the original partition. However, the use of internal criteria for selecting a solution is unavoidable in clustering where the true partition is unknown. A cross-tabulation of the assignments is used to determine the adjusted Rand index (Hubert and Arabie 1985), a measure of agreement between two partitions of a data set. The Rand index was introduced by Rand (1971) and is defined as the number of pairs of objects that are either consistently assigned to the same or different clusters across two different partitions. The Rand index does not correct for agreement by chance (Hubert and Arabie 1985), while the adjusted Rand index does. Values of the adjusted Rand index lie between −1 and 1, where 1 indicates that the exact same solution is identified across repeated computations. See the appendix for details on the Rand index and the adjusted Rand index.

To ensure that recommendations about adequate sample size are valid across a range of data circumstances encountered in tourism research, artificial data sets with different characteristics are generated by drawing data from different finite mixtures of multivariate Gaussian distributions, where the settings are selected to cover the range of situations as described in the review on previous segmentation studies in tourism given in Dolnicar (2002). Specifically, the settings differ in (1) the number of variables in the segmentation base, (2) the number of respondents, (3) the number of clusters, (4) the level of separation between clusters, and (5) the proportion of noisy variables in the segmentation base.

Where possible, we chose the exact parameters for the above variations in artificial data sets in order to model as closely as possible the characteristics of empirical tourism data sets used in previous segmentation studies in tourism. Information about typical tourism data characteristics used in segmentation studies was taken from the Dolnicar (2002) review article. As shown in Table 1, artificial data sets include 10, 16, or 22 variables in the segmentation base. These values represent the midpoint (16) and the borders of the interval containing the middle two-thirds of prior tourism segmentation studies: 10 and 22. Because 64% of prior tourism segmentation studies grouped respondents into three or four segments, these two numbers of clusters have been chosen.

With respect to noisy variables, no guidance is available based on the results of previous segmentation studies where noisy variables were not explicitly accounted for. However, we may assume situations exist where no noisy variables are present (when, e.g., segmentation variables were carefully selected in advance), as well as situations where a substantive amount of variables is not relevant for the clustering structure. Therefore, we created artificial data sets with levels of contamination covering an extensive range: some contained no noisy variables (0% contamination), some one-quarter (25% contamination), and in some cases half of all variables were noisy (50% contamination). Noisy variables were generated by drawing from a multivariate Gaussian distribution of similar variation to the non-noisy variables and independent of the non-noisy variables.

The degree of cluster separation was controlled using the so-called separation index, as described by Qiu and Joe (2006). The separation index measures the amount of space between two clusters by determining the optimal projection direction for the data and then defining the distance between groups based on the lower and upper $\alpha/2$ percentiles of the projected clusters. We use a value of 0.05 for α. Separation values can lie between −1 and 1, with values close to 1 indicating the maximum distance between neighbouring groups. Again, no guidance on the separation between clusters is available from previous segmentation studies, so we selected the levels to cover clusters ranging from well separated to overlapping, but which were still distinguishable. Three levels of cluster separation are simulated: a positive separation value of 0.1 leads to well-separated clusters, a separation value of 0 leads to segments that touch one another, and a negative separation value of −0.1 leads to overlapping clusters. A separation index of −0.1 corresponds to a misclassification rate of 5% for two equally sized clusters that were drawn from univariate normal distributions with the same variance and where the true distributions were known; that is, the classification problem has a Bayes risk of 5%. All

these scenarios ensure that clusters are sufficiently separated in order to be able to distinguish them, conditional on enough observations being provided.

The number of respondents were chosen using the rule given by Qiu and Joe (2009) implying that they linearly depend on the number of variables leading to $10{\cdot}d$, $20{\cdot}d$, $30{\cdot}d$, $40{\cdot}d$, $50{\cdot}d$, $60{\cdot}d$, $70{\cdot}d$, $80{\cdot}d$, $90{\cdot}d$, and $100{\cdot}d$ respondents simulated in different data scenarios, where d is the number of variables. This range for the number of respondents allows analyzing how an increase in sample size impacts on the performance, because they vary from rather small to very high numbers covering the range of insufficient information from the data to a rather large amount, where hardly any improvement is achieved by a further increase in sample size.

The full-factorial design of all independent variables led to 540 data settings (see Table 1). For each setting, 50 data sets were created, which were subsequently clustered using the *k*-means algorithm. Results from the best solution from 30 random initializations were used. This analysis led to 27,000 adjusted Rand index values that were used to determine a rule of thumb for adequate sample size in data-driven segmentation studies.

The artificial data sets were generated using the statistical software environment R and the package cluster Generation.

Generalized Additive Model

In order to systematically investigate the association between the adjusted Rand index value and sample size, we used a generalized additive model (Hastie and Tibshirani 1986). Generalized additive models are regression models that enable the analyst to flexibly model the functional relationship between the independent and dependent variables. When using a linear regression approach, the functional relationship is restricted to being linear. When approximating a more flexible functional relationship, expansions of the independent variables can be used as predictors. In additive models, the functional relationship is assumed to be given by a smooth function, and spline bases are used for approximation, while flexibility and wiggliness of the function are penalized. The amount of penalization is controlled by a hyperparameter. We use generalized additive models because the influence of sample size is not expected to be linear. Generalized additive models allow the estimation of the effect of sample size on the dependent variable (adjusted Rand index) as a nonlinear, smooth function while controlling for the remaining covariates (number of variables, number of clusters, number of noisy variables, and separation value). The hyperparameter controlling the smoothness of the function was automatically selected using generalized cross-validation. The model was estimated in R using the package mgcv (Wood 2006, 2012). Thin plate regression splines were used as the spline basis. The control variables were added as categorical variables with a fully saturated design. This led to the following model:

$$\text{RIadj} = f(\text{sample size}) + g(\text{separation, clusters, variables, noisy}) + \varepsilon,$$

where RIadj is the dependent variable corresponding to the adjusted Rand index value, f is an unknown smooth function of the covariate sample size, and g is a linear function capturing the average adjusted Rand index values for all different combinations of levels of the four covariates and $\varepsilon \sim N(0, \sigma^2)$. The assumed distribution of the response variable is the normal distribution with an identity-link function. For details on estimation, see the appendix.

Results indicate that the number of variables does not affect the adjusted Rand index significantly, either alone or in interaction with other independent variables. The R-squared measure of goodness of fit is 0.754 for both the models with and without the covariate "number of variables." However, this does not mean that the number of variables has no effect at all. We generated samples of size $10{\cdot}d$, $20{\cdot}d$, etc. The nonsignificance of d in the model above lends strong support for having a multiplicative relationship between sample size and number of variables. As a consequence of this finding, we excluded the number of variables as an independent variable and modified the model accordingly:

$$\text{RIadj} = f(\text{sample size}) + g(\text{separation, clusters, noisy}) + \varepsilon,$$

where f is an unknown smooth function of the covariate sample size, and g is a linear function capturing the average adjusted Rand index values for all different combinations of levels of the three covariates and $\varepsilon \sim N(0, \sigma^2)$.

We consider the adequate sample size to be the smallest sample size where the adjusted Rand index values do not significantly differ from adjusted Rand index values of higher sample sizes (i.e., $100.d$) with $\alpha = 5\%$.

Results

Simulation results are displayed in Figure 2 and Figures A1 to A3 (appendix), which show the smooth function f resulting from the generalized additive model. In all figures, the x axis displays the sample size while the y axis represents the effect of the sample size on the adjusted Rand index. In order to easily detect the optimal sample size, we added a 95% confidence interval to the whole function. The recommended sample size is marked with a cross and its confidence bands are drawn using dashed lines.

Figure 2 depicts the function f aggregated over all levels of the linear predictors. Figures A1 to A4 show the effects separately for all combinations of levels of number of noisy variables, the number of clusters, and separation values. Thus, the interaction of the single stages of the linear predictors and the sample size on the adjusted Rand index values

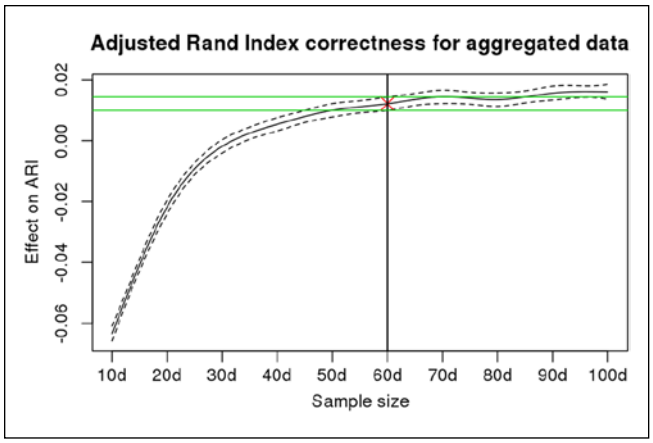

Figure 2. Smooth function of sample size and its effect on adjusted Rand index values with 95% confidence interval (dashed lines) resulting for the aggregated data over all levels of number of clusters, noisy variables, and separation indices. The recommended sample size is marked with a cross as well as a vertical line and the corresponding confidence interval using dashed lines.

can be investigated. Figure 2 gives an aggregate overview of the effect of the sample size on clustering performance averaged over the different scenarios, while Figures A1 to A3 indicate these effects for each specific scenario.

In the aggregated data setting, the degree of freedom of the smoothing term for sample size is 7.49 (p value $< 2 \cdot 10^{-16}$), indicating that the influence of the sample size on the adjusted Rand index values is not linear. Investigating all combinations of factor levels indicates that for very simple segmentation tasks (which contain no noisy variables and separation values of 0.1 or 0), smoothed functions are almost linear and horizontal. For more difficult data situations, the function f is not linear, confirming our choice of using a generalized additive model for analysis.

The coefficients of the three factors (number of clusters, number of noisy variables, and separation index) are significant (p value $< 2 \cdot 10^{-16}$), as are the coefficients of most of the two-way interactions (p value < 0.01; see Table A1 in the appendix). None of the three-way interactions are significant. This means that all investigated factors, as well as their two-way interactions, have an influence on the adjusted Rand index value. A further analysis also indicated that the interactions between sample size and a combination of the linear predictors are significant (p value < 0.01), except for the two easiest segmentation tasks with $0 \cdot d$ noisy variables, separation value of 0.1 and three or four clusters.

The first key result emerging from the simulation study is that depending on the data situation, correctness of the segmentation solution can suffer substantially if the sample size for analysis is insufficient. The fitted smooth function shows a clear increase, which means that the adjusted Rand index values are higher when higher sample sizes are available. This effect is more obvious for difficult segmentation problems. As mentioned above, smoothed functions are almost linear and horizontal in data situations with a clear group structure. In these cases, the effect of additional samples on the correctness of cluster results is not very strong. In the remaining cases, the additional samples have a considerable positive effect on the adjusted Rand index values, which increases with the difficulty of the task. This means that for harder segmentation tasks, higher sample sizes can result in considerably improved results. In the case of 50% contamination by noisy variables, a separation index of −0.1, and four clusters, the improvement is 0.24 when comparing the adjusted Rand index values for sample sizes of $10 \cdot d$ and $60 \cdot d$.

In the aggregated data setting the adequate sample size is $60 \cdot d$, which means an improvement of approximately 0.08 from the worst-case scenario with a sample size equal to $10 \cdot d$. Sample sizes higher than $60 \cdot d$ did not, for the data scenarios studied, lead to a significant improvement of the adjusted Rand index, although the index continues to increase. A separate analysis of the results when holding one of the covariates (separation, clusters, noisy) fixed indicates that the minimum of the adequate numbers observed in our study is $10 \cdot d$, the maximum $70 \cdot d$, and the mean is $40 \cdot d$.

The fact that most optimal sample sizes range between $30 \cdot d$ and $40 \cdot d$ provides some support for the rule proposed by Qiu and Joe (2009). Their idea to investigate the optimal sample size as a function of the number of variables d seems to be a good choice, although our results indicate that for some data situations, approximately double the sample size recommended by Qiu and Joe (2009) is required to ensure a reliable result.

Overall, our results indicate that sample size-to-variable ratios currently used in tourism segmentation studies cannot be considered to be adequate. Based on the review by Dolnicar (2002), the median ratio is 22.4, meaning that half of the segmentation studies published in tourism use samples that are lower than 22.4 times the number of variables in the segmentation base, and half use samples sizes that are higher than 22.4 times the number of variables. A Wilcoxon test ($V = 162$, p value < 0.001) indicates that the adequate numbers of respondents are significantly higher than $22.4 \cdot d$ across all different data scenarios.

The complication when attempting to derive sample size requirements in dependence of the number of variables in the segmentation base is that each empirical data set is different and the degree of cluster structure is not known. We therefore recommend using at least $70 \cdot d$ respondents (70 times the number of variables in the segmentation base) for data-driven segmentation studies. This represents the most conservative sample size requirement for data sets typical for tourism studies, as simulated in the present study. The sample size requirement of $70 \cdot d$ allows for the possibility that the empirical data under study has a high degree of difficulty.

Conclusions

The aim of the present study is to determine the required sample size for data-driven market segmentation studies in tourism. An extensive simulation study using artificial data sets of varying structure and difficulty was conducted, and the effect of a number of typical factors relating to data structure was examined. Our results indicate that in most cases, the correctness of segmentation analyses can be significantly improved by increasing the sample size. This effect is stronger for more difficult segmentation tasks. Only in the case of data with a very clear segment structure—a situation yet to be encountered by the authors—does increasing the sample size *not* improve the segmentation solution.

Because it is impossible in the case of empirical survey data to know the true data structure, we must by default assume that the segmentation task is complex, and consequently, the most conservative rule for sample size requirement resulting from the simulation study should be used: *at least 70 times the number of variables*. Note that the required adequate sample sizes determined in this study, though substantially lower than the requirement proposed by Formann (1984), are significantly higher than the sample sizes used in most published tourism segmentation studies.

Another conclusion that we may draw from the present study is that noisy variables in the segmentation base increase the complexity of the segmentation task substantially. It is therefore worthwhile carefully selecting variables to be included in the segmentation base, rather than including an entire question battery by default. Noisy variables in the segmentation base can be avoided by (1) identifying and removing them after data collection (Brusco and Cradit 2001; Carmone, Kara, and Maxwell 1999; Steinley and Brusco 2008) or by (2) ensuring, before data collection, that survey questions are only included if they contain relevant information, as advocated by Rossiter (2002, 2011). Methods for identifying and removing can either be employed before the clustering using characteristics of the distribution of the single variables (Steinley and Brusco 2008) or simultaneously during clustering, by taking into account the concordance and agreement between cluster solutions implied by different variables (see Brusco and Cradit 2001, who directly build on and improve Carmone, Kara, and Maxwell 1999).

Future work that would further add to our understanding of sample size requirements could investigate the degree to which sample size requirements vary across scale formats and clustering algorithms.

Appendix

Rand Index and Adjusted Rand Index

The Rand index was introduced by Rand (1971), and the adjusted Rand index by Hubert and Arabie (1985). A further overview on these indices is also given in Steinley (2004).

Assume we have two partitions of the data with N observations; that is, partition one assigns each observation uniquely to R groups and partition two to C groups. The cross-tabulation of these assignments gives a $R \times C$ table with entries n_{ij} indicating the number of observations that are assigned to group i in partition one and j in partition two. From these entries, the following quantities can be computed

$$a = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 - N}{2},$$

$$b = \frac{\sum_{i=1}^{R} n_{i.}^2 - \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2}{2},$$

$$c = \frac{\sum_{j=1}^{C} n_{.j}^2 - \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2}{2},$$

$$d = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 + N^2 - \sum_{i=1}^{R} n_{i.}^2 - \sum_{j=1}^{C} n_{.j}^2}{2},$$

where $n_{i.}$ and n_{j} denote the number of observations assigned to group i by partition one and to group j by partition two, respectively. a corresponds to the number of pairs where the observations are placed in the same clusters by both partitions, b corresponds to the number of pairs where observations are placed in the same cluster by partition one but not by partition two; c analogously corresponds to the number of pairs, where observations are placed in the same cluster by partition two but not by partition one, and finally d corresponds to the number of pairs where observations are placed in different clusters by both partitions. The total number of pairs equals $\binom{N}{2} = N(N-1)/2.$

This implies that the Rand index (RI), which is defined by the sum of pairs, where the two partitions agree on putting observations in the same or different groups, divided by the total number of pairs, is given by

$$RI = \frac{a+d}{a+b+c+d}.$$

The adjusted Rand index (RIadj) is determined by subtracting the expected value of agreement by chance from the nominator as well as the denominator

$$RIadj = \frac{a+d-E}{a+b+c+d-E},$$

where

$$E = \frac{(a+b)(a+c)+(c+d)(b+d)}{\binom{N}{2}}.$$

Separation Index

The separation index, as used in Qiu and Joe (2006), measures the magnitude of the gap between a pair of clusters. In the following, L_i and U_i are the lower and upper $\alpha/2$ percentiles of cluster i. The separation index (SI) is defined as:

$$SI = \frac{L_2 - U_1}{U_2 - L_1}.$$

An illustration is given in Figure A1.

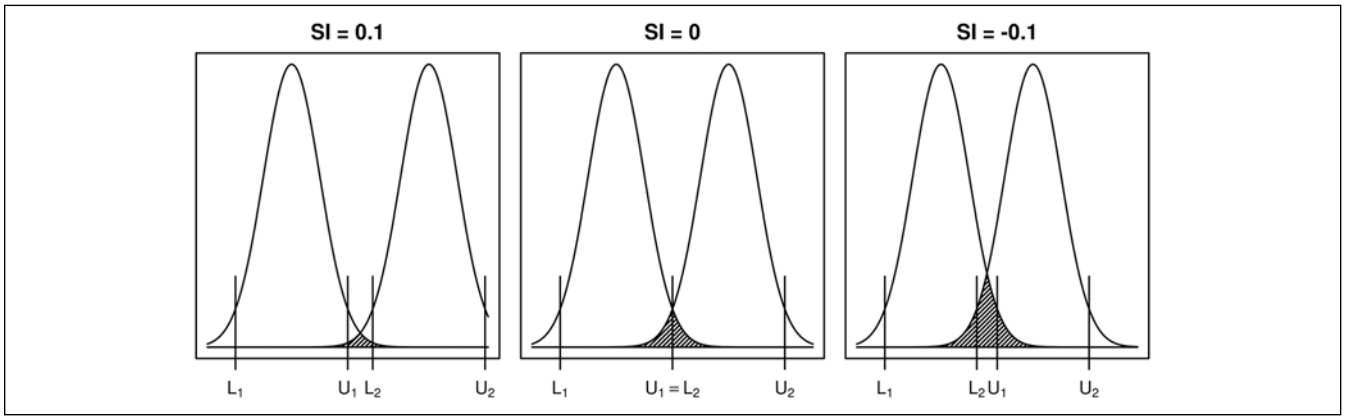

Figure A1. Illustration of the concept of the separation index (SI) for univariate normally distributed data. L_i and U_i are the lower and upper $\alpha/2$ percentiles of the clusters, where α is set to 5%. The grey-shaded areas indicate the proportion of observations that will be misclassified.

Generalized Additive Model

The function for the dependent variable in the generalized additive model can be represented as a linear function:

$$RIadj = X\alpha + Z\beta + \varepsilon,$$

where X is the covariate matrix of the control variables with corresponding coefficients α, which is used to represent the linear function g, and Z the covariate matrix of the spline basis for sample size with penalized coefficients β, which is used to represent the smooth function f. The coefficients are estimated by minimizing with respect to $\hat{\alpha}$ and $\hat{\beta}$:

$$|| RIadj - X\hat{\alpha} - Z\hat{\beta} ||_2^2 + \lambda J(\hat{\beta}),$$

where λ is a given hyperparameter controlling the penalization, and $J(\hat{\beta})$ measures the smoothness of the function; that is, it is smaller the smoother the function is as implied by the second derivative. The optimal value for λ is determined using the generalized cross-validation score, which equals the average squared prediction error for each observation when fitting the model without this observation and employing a rotation, such that each observation has the same leverage. For further details see the introduction to GAMs given in Wood (2006).

Results—Additional Figures

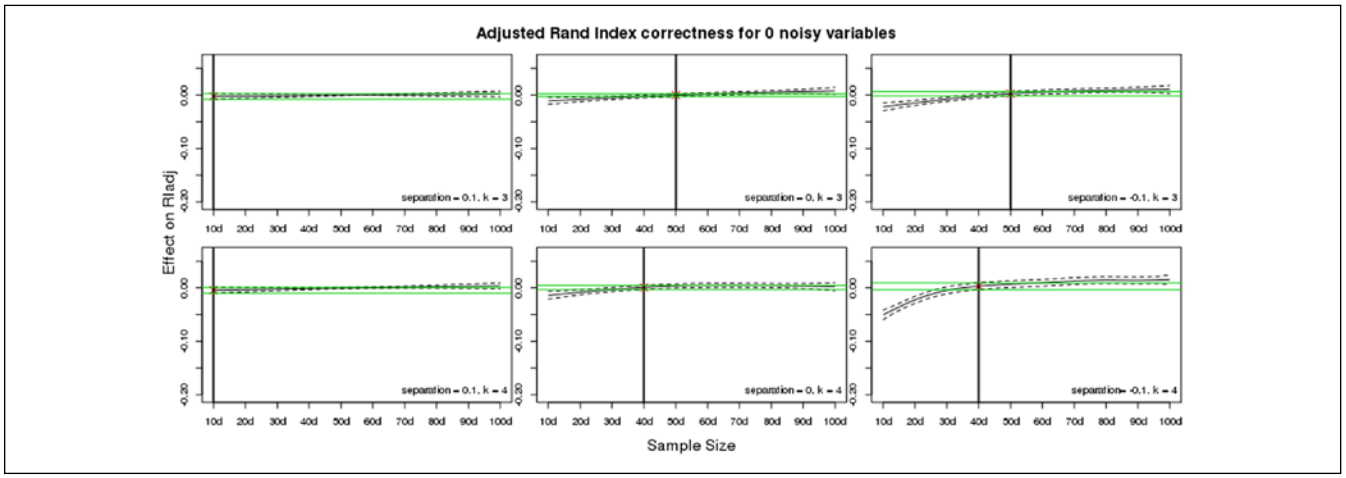

Figure A2. Smooth function of sample size and its effect on adjusted Rand index values with 95% confidence interval (dashed lines) resulting separately for all combinations of covariate levels of number of clusters k, separation indices, and noisy variables m = 0d. The recommended sample size is marked with a cross as well as a vertical line and the corresponding confidence interval using dashed lines.

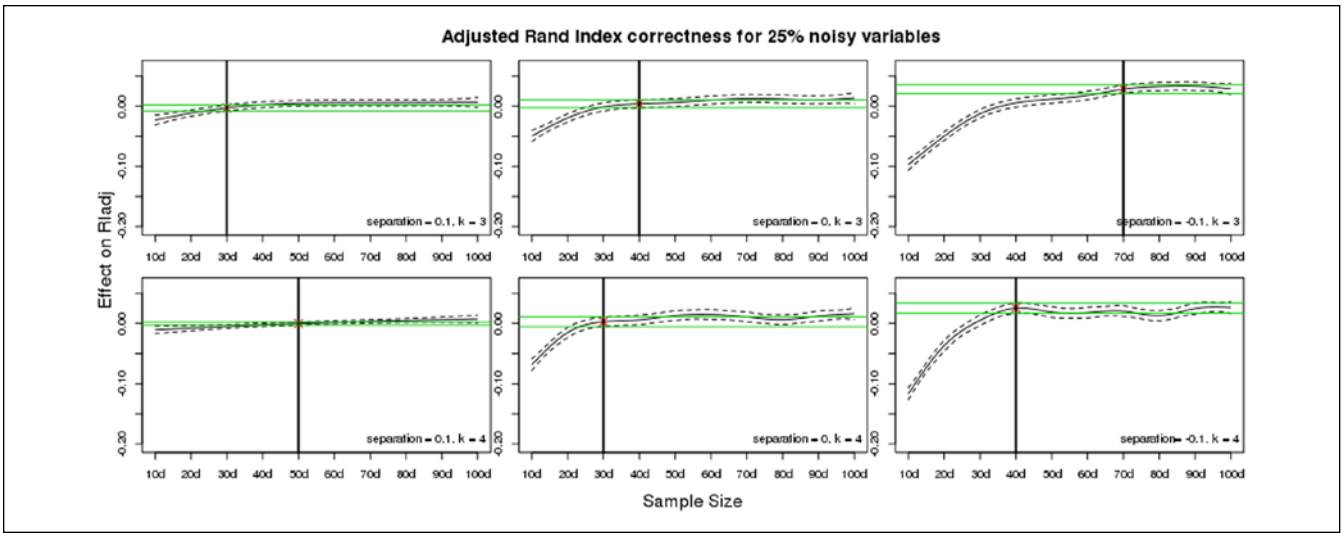

Figure A3. Smooth function of sample size and its effect on adjusted Rand index values with 95% confidence interval (dashed lines) resulting separately for all combinations of covariate levels of number of clusters *k*, separation indices and noisy variables $m = 1/4d$. The recommended sample size is marked with a cross as well as a vertical line and the corresponding confidence interval using dashed lines.

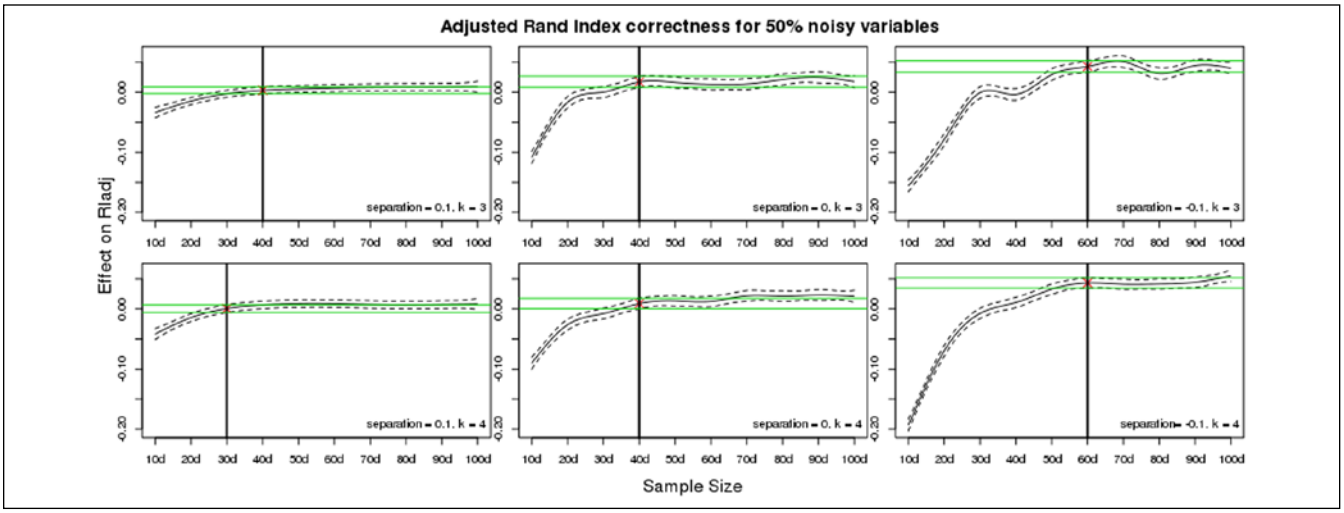

Figure A4. Smooth function of sample size and its effect on adjusted Rand index values with 95% confidence interval (dashed lines) resulting separately for all combinations of covariate levels of number of clusters *k*, separation indices and noisy variables $m = 1/2\ d$. The recommended sample size is marked with a cross as well as a vertical line and the corresponding confidence interval using dashed lines.

Table A1. Overview of Regression Coefficient Estimates or Degrees of Freedom and *p* Values of the Generalized Additive Model with Linear Predictors Separation (Reference Class −0.1), Clusters (Reference Class 3) and Noisy (Reference Class 0) in the Aggregated Data Setting.

Parametric Coefficients	Estimate	*p*-Value
Intercept	0.678	$<2\cdot10^{-16}$
Cluster 4	−0.032	$<2\cdot10^{-16}$
Separation 0	0.153	$<2\cdot10^{-16}$
Separation 0.1	0.257	$<2\cdot10^{-16}$
Noisy 2	−0.043	$<2\cdot10^{-16}$
Noisy 4	−0.030	$<2\cdot10^{-16}$
Cluster 4 * Separation 0	0.014	$8.09\cdot10^{-5}$
Cluster 4 * Separation 0.1	0.023	$1.11\cdot10^{-10}$
Cluster 4 * Noisy 2	−0.003	0.4475
Cluster 4 * Noisy 4	0.010	0.0073
Separation 0 * Noisy 2	0.025	$3.84\cdot10^{-12}$
Separation 0.1 * Noisy 2	0.034	$<2\cdot10^{-16}$
Separation 0 * Noisy 4	0.019	$5.43\cdot10^{-8}$
Separation 0.1 * Noisy 4	0.023	$6.99\cdot10^{-11}$
Cluster 4 * Separation 0 * Noisy 2	0.003	0.6031
Cluster 4 * Separation 0.1 * Noisy 2	0.004	0.4579
Cluster 4 * Separation 0 * Noisy 4	−0.008	0.1056
Cluster 4 * Separation 0.1 * Noisy 4	−0.006	0.2518
Smooth Term	Degrees of Freedom	*p*-Value
Sample size	7.489	$<2\cdot10^{-16}$

Note: Coefficients that are significant on an α-level of 5% are marked in bold face.

Acknowledgments

We thank the Australian Research Council and the Austrian Science Fund (FWF) for contributing to the funding of this study.

Authors' Note

This project was undertaken during Kathrin Schmidt's visit at the University of Wollongong; Kathrin Schmidt was primarily responsible for data analysis presented in this paper. Authors are listed in alphabetical order.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Australian Research Council (ARC) grants LX0881890 and DP110101347 and Austrian Science Fund (FWF) Elise-Richter grant V170-N18.

References

Bejarano, J., K. Bose, T. Brannan, A. Thomas, K. Adragni, N. K. Neerchal, and G. Ostrouchov. (2011). "Sampling within K-Means Algorithm to Cluster Large Datasets." Technical Report HPCF-2011-12, UMBC High Performance Computing Facility, University of Maryland, Baltimore County.

Ben-David, S., D. Pál, and H. U. Simon. (2007). "Stability of k-means Clustering." In *Learning Theory (20th Annual Conference on Learning Theory, COLT 2007)*, edited by N. H. Bhsouty and C. Gentile. Berlin: Springer, pp. 20-34.

Brusco, M. J., and J. D. Cradit. (2001). "A Variable-Selection Heuristic for K-means Clustering." *Psychometrika*, 66 (2): 249-70.

Carmone, F. J., Jr., A. Kara, and S. Maxwell. (1999). "HINoV: A New Model to Improve Market Segment Definition by Identifying Noisy Variables." *Journal of Marketing Research*, 36: 501-9.

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

Dolnicar, S. (2004). "Beyond 'Commonsense Segmentation'—A Systematics of Segmentation Approaches in Tourism." *Journal of Travel Research*, 42: 244-50.

Dolnicar, S., and F. Leisch. (2010). "Evaluation of Structure and Reproducibility of Cluster Solutions Using the Bootstrap." *Marketing Letters*, 21: 83-101.

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

Dolnicar, S., and K. Lazarevski. (2009). "Methodological Reasons for the Theory/Practice Divide in Market Segmentation." *Journal of Marketing Management*, 25: 357-74.

Dolnicar, S., K. Grabler, B. Grün, and A. Kulnig. (2011). "Key Drivers of Airline Loyalty." *Tourism Management*, 32 (5): 1020-26.

Dolnicar, S., S. Kaiser, K. Lazarevski, and F. Leisch. (2012). "Biclustering—Overcoming Data Dimensionality Problems in Market Segmentation." *Journal of Travel Research*, 51 (1): 41-49.

Formann, A. K. (1984). *Die Latent-Class-Analyse: Einführung in die Theorie und Anwendung* [Latent class analysis: Introduction to theory and application]. Weinheim: Beltz.

Hastie, T., and R. Tibshirani. (1986). "Generalized Additive Models." *Statistical Science*, 1: 297-318.

Hubert, L., and P. Arabie. (1985). "Comparing Partitions." *Journal of Classification*, 2: 193-218.

Legoherel, P., and K. K. L. Wong. (2012). "Market Segmentation in the Tourism Industry and Consumer Spending." *Journal of Travel & Tourism Marketing*, 20 (2): 15-30.

Masiereo, L., and J. L. Nicolau. (2012). "Tourism Market Segmentation Based on Price Sensitivity: Finding Similar Price Preferences on Tourism Activities." *Journal of Travel Research*, 51 (4): 426-35.

Mazanec, J. A. (2000). "Market Segmentation." In *Encyclopedia of Tourism*, edited by J. Jafari. London: Routledge.

Needham, M. D., R. B. Rollins, R. L. Ceurvorst, C. J. B. Wood, K. E. Grimm, and P. Dearden. (2011). "Motivations and Normative Evaluations of Summer Visitors at an Alpine Aki Area." *Journal of Travel Research*, 50 (6): 669-84.

Nicolau, J. L. (2012). "Asymmetric Tourist Response to Price: Loss Aversion Segmentation." *Journal of Travel Research*, 51 (5): 568-76.

Qiu, W., and H. Joe. (2006). "Generation of Random Clusters with Specified Degree of Separation." *Journal of Classification*, 23: 315-34.

Qiu, W., and H. Joe. (2009). clusterGeneration: Random Cluster Generation (with Specified Degree of Separation), R package version 1.2.7.

R Development Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rand, W. M. (1971). "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association*, 66: 846-50.

Rossiter, J. R. (2002). "The C-OAR-SE Procedure for Scale Development in Marketing." *International Journal of Research in Marketing*, 19: 305-35.

Rossiter, J. R. (2011). *Measurement for the Social Sciences. The C-OAR-SE Method and Why It Must Replace Psychometrics*. New York: Springer.

Steinley, D. (2004). "Properties of the Hubert–Arabie Adjusted Rand Index." *Psychological Methods*, 9 (3): 386-96.

Steinley, D., and M. J. Brusco. (2008). "A New Variable Weighting and Selection Procedure for K-Means Cluster Analysis." *Multivariate Behavioral Research*, 43 (1): 77-108.

Van Middelkoop, M., A. Borgers, and H. Timmermans. (2003). "Inducing Heuristic Principles of Tourist Choice of Travel Mode: A Rule-Based Approach." *Journal of Travel Research*, 42: 75-83.

Weaver, D. B., and L. J. Lawton. (2011). "Visitor Loyalty at a Private South Carolina Protected Area." *Journal of Travel Research*, 50 (3): 335-46.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction*. Boca Raton, FL: Chapman and Hall/CRC.

Wood, S. N. (2012). mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation. R package version 1.7-18.

Zins, A. (2008). "Market Segmentation in Tourism: A Critical Review of 20 Years' Research Efforts." In *From the "OLD" to the "NEW" Tourism: Managing Change in the Tourism Industry*, edited by S. M. C. Kronenberg, M. Peters, B. Pikkemaat, and K. Weiermair. Berlin: Erich Schmidt Verlag, pp. 289-301.

Author Biographies

Sara Dolnicar is Research Professor of Tourism at the University of Queensland. Her primary research interests are measurement and segmentation methodology in the social sciences. Most of Sara's work has been tested and applied in the areas of tourism and social marketing.

Bettina Grün is a Senior Research Fellow in Statistics at the Johannes Kepler University Linz. Her research interests include finite mixture modeling, statistical computing, and quantitative methods in marketing research.

Friedrich Leisch is Professor of Statistics at the University of Natural Resources and Life Sciences, Vienna, Austria. His primary research interests are statistical computing, cluster analysis, finite mixture models and biostatistics and their application in life and business sciences.

Kathrin Schmidt has completed a Masters Degree in Statistics at the Ludwig Maximilian University of Munich. After completing her degree she spent some time doing an internship at the University of Wollongong in Australia, where she worked on the project reported in this paper. Kathrin now works as a data analyst for a European telecommunication company.

Dolnicar, Grün and **Leisch** have over the past 15 years jointly conducted a range of research studies investigating methodological aspects of market segmentation which has been enabled by financial support of the Austrian and Australian government research funding bodies.