Journal of Business Research xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Business Research

Increasing sample size compensates for data problems in segmentation studies

Sara Dolnicar [a,*], Bettina Grün [b,1], Friedrich Leisch [c,2]

[a] *University of Queensland, St Lucia, Brisbane 4072, Australia*
[b] *Johannes Kepler University, Altenbergerstraße 69, 4040 Linz, Austria*
[c] *University of Natural Resources and Life Sciences, Gregor-Mendel-Straße 33, 1180 Wien, Vienna, Austria*

ARTICLE INFO

Article history:
Received 29 January 2015
Received in revised form 1 September 2015
Accepted 8 September 2015
Available online xxxx

Keywords:
Market segmentation
Sample size
Survey data
Response bias

ABSTRACT

Survey data frequently serve as the basis for market segmentation studies. Survey data, however, are prone to a range of biases. Little is known about the effects of such biases on the quality of data-driven market segmentation solutions. This study uses artificial data sets of known structure to study the effects of data problems on segment recovery. Some of the data problems under study are partially under the control of market research companies, some are outside their control. Results indicate that (1) insufficient sample sizes lead to suboptimal segmentation solutions; (2) biases in survey data have a strong negative effect on segment recovery; (3) increasing the sample size can compensate for some biases; (4) the effect of sample size increase on segment recovery demonstrates decreasing marginal returns; and—for highly detrimental biases—(5) improvement in segment recovery at high sample size levels occurs only if additional data is free of bias.

1. Introduction

Market segmentation "is essential for marketing success: the most successful firms drive their businesses based on segmentation" (Lilien & Rangaswamy, 2002, p. 61) and "tools such as segmentation [...] have the largest impact on marketing decisions" (Roberts, Kayande, & Stremersch, 2014, p. 127).

Despite the importance of market segmentation and its widespread use in industry, segmentation experts have repeatedly raised concerns about discrepancies between academic progress in the field and practical application challenges (Dibb & Simkin, 1997, 2001; Greenberg & McDonald, 1989; Wind, 1978; Young, Ott, & Feigin, 1978) pointing to an overemphasis on the data analytic aspect at the expense of developing solutions for conceptual and implementation challenges. This is particularly true for data-driven segmentation studies which construct segments by applying a statistical algorithm to several variables in the segmentation base as opposed to commonsense segmentation studies where segments result from dividing the population according to prior knowledge (Dolnicar, 2004).

One key implementation challenge companies face every time they conduct a segmentation study is that of data quality. Recent work by Coussement, Van den Bossche, and De Bock (2014) studies the extent to which data accuracy problems in databases affect the performance of direct marketing actions and segmentation solutions and investigate the robustness of different segmentation algorithms against inaccurate data. Despite the extensive body of work on survey data quality, targeted investigations of the effect of data quality on segmentation solutions have only recently started to emerge: Dolnicar and Leisch (2010) speak to the issue of segmentability of the market, first raised by Wind (1978) and Young et al. (1978), and offer a framework for data structure analysis before constructing segments.

The present study contributes a novel solution to the data quality challenge in data-driven market segmentation by investigating whether increasing the sample size can compensate for typical survey data quality problems. Specifically, the study investigates (1) the extent of the detrimental effect of data characteristics typical for survey data on the correctness of market segmentation solutions, (2) the general potential of increasing sample sizes to improve the correctness of market segmentation solutions, and (3) the potential of increased sample sizes to improve the correctness of market segmentation solutions when encountering typical survey data challenges. While it is to be assumed that larger sample sizes improve data analysis, the present study aims at deriving recommendations about the extent of required sample size increase to counteract specific kinds of survey data problems. Increasing the sample size to the required level represents—in times where survey data is increasingly collected online—a simple and affordable measure. The results of this study, therefore, will generate managerial recommendations which can easily be implemented.

2. Literature review

The potentially detrimental effect of bad data on market segmentation solutions has been discussed in the earliest studies on market

[*] Corresponding author. Tel.: +61 7 3365 6702.
E-mail addresses: s.dolnicar@uq.edu.au (S. Dolnicar), bettina.gruen@jku.at (B. Grün), friedrich.leisch@boku.ac.at (F. Leisch).
[1] Tel.: +43 732 2468 6829.
[2] Tel.: +43 1 47 654 5061.

http://dx.doi.org/10.1016/j.jbusres.2015.09.004

segmentation: Claycamp and Massy (1968) point to the challenge of measuring response elasticities for segments; Young et al. (1978) argue that each segmentation problem is unique, and consequently, it is critical to select carefully who is interviewed, which questions are asked, and in which way. Wind (1978) discusses shortcomings related to segmentation bases typically used, and calls for increased efforts in determining the unit of analysis, the operational definition of dependent and independent variables, sample design, and checking of data reliability.

Several data characteristics that can reduce the validity of segmentation solutions have been known for a long time with no generally accepted solutions to reduce their impact available to date. For example, masking variables in the segmentation base which "hide or obfuscate the true structure in the data" (Brusco, 2004, p. 511) and consequently lead to inferior segmentation solutions (Carmone, Kara, & Maxwell, 1999; DeSarbo, Carroll, Clark, & Green, 1984; DeSarbo & Mahajan, 1984; Milligan, 1980) led to the development of a range of different variable selection and weighting approaches (Maugis, Celeux, & Martin-Magniette, 2009a, 2009b; Raftery & Dean, 2006; Steinley & Brusco, 2008a, 2008b).

Survey data are also known to be susceptible to response styles. Response styles result from response tendencies regardless of the content (Paulhus, 1991) and can manifest in extreme or acquiescence response styles (Baumgartner & Steenkamp, 2001). Again, different approaches have been proposed to address this problem, such as standardization of the data prior to the analysis (Schaninger & Buss, 1986) or a joint segmentation approach (Grün & Dolnicar, in press) which allows for response style and content-driven segments simultaneously.

Many other survey data characteristics—the effect of which on market segmentation analysis has not been studied to date—can also reduce the ability of a segmentation algorithm to identify naturally existing market segments or to construct managerially useful segments: sampling errors due to the decreasing willingness of people to participate in survey studies (Bednell & Shaw, 2003); respondents not answering survey questions carefully (Krosnick, 1999) or in a socially desirable way (Goldsmith, 1988; Tellis & Chandrasekaran, 2010); respondents interpreting survey questions differently, respondent fatigue leading to some low-quality responses (Johnson, Lehmann, & Horne, 1990); questionnaire items not being selected carefully (Rossiter, 2002, 2011); and the provision of binary or ordinal answer options to respondents where continuous measures could be used, which leads to less information available for data analysis (Kampen & Swyngedouw, 2000). An overview of these challenges affecting the quality of survey data is provided in Table 1.

These factors are, to some degree, in the control of the firm, because good item and scale development, questionnaire design, and fieldwork administration can reduce the incidence of survey data contamination. General recommendations for developing good survey questions are offered by Converse and Presser (1986), who recommend the use of short, simple, intelligible, and clear questions which employ straightforward language. An overview on survey sampling is given in Kalton (1983) who emphasizes the importance of reducing nonresponse because of the limitations of statistical procedures based on weighting to account for or remove nonresponse bias. Respondent fatigue, for example, can be reduced by employing procedures requiring fewer judgments from the respondents (Johnson et al., 1990). Yet, all these quality issues can never be totally excluded because, for example, some respondents always fail to take the task of completing the survey seriously. In some cases, statistical methods can be employed to account for data contaminations in the analysis, as is the case for response styles (see Grün & Dolnicar, in press; Schaninger & Buss, 1986). As a pre-processing tool to remove delinquent respondents Barge and Gehlbach (2012), for example, suggest determining the amount of satisficing of each respondent and to then assess the influence of exluding these respondents from the subsequent analysis.

Furthermore, all of the data issues discussed above can occur in situations where market characteristics already complicate the task for segmentation algorithms. For example, segment recovery is more complicated for segments of unequal size (De Craen, Commandeur, Frank, & Heiser, 2006) and for segments which overlap (Steinley, 2003) and depends on the number of segments. Such factors are entirely out of the control of the firm.

One aspect of segmentation analysis is usually *in* the control of the firm: the sample size. If shown to be effective in counteracting the detrimental effects of survey data problems, adjusting the sample size represents a simple solution. Increased sample sizes should improve solutions because market segmentation studies typically use data sets containing large numbers of variables and are thus subject to the so-called "curse of dimensionality" (Bellman, 1961).

Little research has been conducted to date to understand the effect of sample size on the correctness of segment recovery, although researchers as early as in the late 1970s noted that increasing sample size "can increase the confidence in a particular structure" and that reducing "the dimensionality can have the effect of increasing sample size" (Dubes & Jain, 1979, p. 240). Sample size recommendations for segmentation studies have, until recently, not been available at all, and the issue of sample size requirements has not been discussed as being critical—not even by authors who emphasize the importance of data quality. Only three discussions of sample size in the context of market segmentation analysis exist, none of which represent generalizable recommendations: (1) Formann (1984), in a monograph on latent class analysis, provides a sample size recommendation in the context of goodness-of-fit testing using the Chi-squared test for binary data: a minimum of two to the power of the number of variables in the segmentation base and preferably five times this number; (2) Qiu and Joe (2009) recommend that, for the purpose of generating artificial data for clustering simulations, the sample size should be at least ten times the number of variables in the segmentation base, times the number of clusters in the simplest case where clusters are of equal size; and (3) Dolnicar, Grün, Leisch, and Schmidt (2014) simulate

Table 1
Sources of quality issue problems in survey data

Problem	Description	Reference
Sampling error	Nonresponse bias occurring due to nonresponse or noncontacts, i.e., a subset of the population is not covered by the survey	e.g., Bednell and Shaw (2003)
Delinquent respondents	Satisficing respondents minimizing effort involved or respondents giving socially desirable answers	e.g., Goldsmith (1988), Krosnick (1999), Tellis and Chandrasekaran (2010)
Respondent fatigue	Respondents becoming tired of the survey task leading to a deterioration of data quality	e.g., Johnson et al. (1990)
Construct measurement and scale development	Surveys can use either single or multiple questions to measure constructs, where multi-item scales often lead to answers being highly correlated	e.g., Rossiter (2002, 2011)
Response alternatives	Choices provided to the respondents determining the measurement scale, i.e., metric, ordinal, or binary	e.g., Kampen and Swyngedouw (2000)
Response style	A systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content (Paulhus, 1991, p. 17)	e.g., Baumgartner and Steenkamp (2001)