Journal of Business Research 66 (2013) 1267–1271

Contents lists available at SciVerse ScienceDirect

Journal of Business Research

A combinatorial optimization based sample identification method for group comparisons ☆

Robyn L. Raschke [a,*], Anjala S. Krishen [b], Pushkin Kachroo [c], Pankaj Maheshwari [d]

[a] University of Nevada, Las Vegas, Department of Accounting, 4505 S. Maryland Parkway, Las Vegas, NV 89154, United States
[b] University of Nevada, Las Vegas, Department of Marketing, United States
[c] University of Nevada, Las Vegas, Department of Electrical and Computer Engineering, United States
[d] University of Nevada, Las Vegas, Department of Civil and Environmental Engineering, United States

ARTICLE INFO

Article history:
Received 1 April 2011
Received in revised form 1 September 2011
Accepted 1 November 2011
Available online 25 February 2012

Keywords:
Sample identification
Sample selection
Sample location identification
Nonprobability samples

ABSTRACT

Researchers often face having to reconcile their sample selection method of survey with the costs of collecting the actual sample. An appropriate justification of a sampling strategy is central to ensuring valid, reliable, and generalizable research results. This paper presents a combinatorial optimization method for identification of sample locations. Such an approach is viable when researchers need to identify sites from which to draw a nonprobability sample when the research objective is for comparative purposes. Findings indicate that using a combinatorial optimization method minimizes the population variation assumptions based upon predetermined demographic variables within the context of the research interest. When identifying the location from which to draw a nonprobability sample, an important requirement is to draw from the most homogeneous populations as possible to control for extraneous factors. In comparison to a standard convenience sample with no identified location strategy, results indicate that the proposed combinatorial optimization method minimizes population variability and thus decreases the cost of sample collection.

1. Introduction

Academics conducting both survey and experimental research must often weigh the costs and benefits of their sampling strategy. Because sampling can have an impact on the validity of research results, a defensible strategy is necessary (Ferber, 1977). The collection method for the data determines the classification of the sample as either a probability or a nonprobability sample. Probability sampling (e.g., simple random, stratified, or systematic) indicates that every element in the population has a known probability of being chosen in the sample for that survey. Thus, a key benefit of probability sampling is the ability to generalize the results, which allows for an estimate of the sampling error. However, probability sampling can require significant resources in both time and money. Unlike probability sampling, nonprobability sampling (e.g., convenience, quota, or judgmental) indicates that every element in the population does not have a known probability of being chosen in the sample for that survey. Therefore, the results are not as generalizable and the sampling error cannot be estimated. But, nonprobability sampling generally is less costly.

The differences between probability and nonprobability sampling are very clear and allow researchers an evaluation criterion to determine an appropriate sampling method. When faced with limited time and money, researchers usually choose the nonprobability sampling method. However, even when a nonprobability sample is the choice, the relation between variability and precision remains. Therefore, if a nonprobability sample comes from a highly variable population, the precision of the results can be in question. If the purpose of the research is for comparison (i.e., to examine the differences between two or more diverse groups of people), homogeneity of the different groups is of utmost importance. Thus, researchers need to minimize demographic differences as much as possible.

The purpose of this paper is to demonstrate a combinatorial optimization method for identifying potential data collection locations for a nonprobability sample. The substantive context of this method comes from a research project aimed at understanding the differences between urban and rural residents and their perceptions of a potential transportation tax policy. The next section of the paper describes the importance of an appropriate sampling strategy when handling targeted group comparisons. Following this, the paper presents the sample identification location problem in a substantive context that details the results of the combinatorial optimization method and demonstrates that this method provides a reasonable strategy as opposed to simply selecting a convenient location for a nonprobability sample. Next, the paper concludes with a discussion of the sampling strategy considerations necessary and the practical implications of this method.

☆ The authors thank Myla Bui-Nguyen, Loyola Marymount University; Angeline Close, University of Texas Austin; Nadia Pomirleanu, University of Nevada Las Vegas, and JBR reviewers for reading and comments of an early version of this article.

* Corresponding author. Tel.: +1 702 895 5756; fax: +1 702 895 4306.
E-mail addresses: robyn.raschke@unlv.edu (R.L. Raschke), anjala.krishen@unlv.edu (A.S. Krishen), pushkin@unlv.edu (P. Kachroo), pankaj47@gmail.com (P. Maheshwari).

0148-2963/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbusres.2012.02.024

2. The importance of sampling strategy

2.1. Sampling strategies for targeted group comparisons in survey research

For decades, social science researchers have debated the tradeoffs associated with obtaining accurate data, setting up valid experiments, and achieving reliable measures. For a research study to be accurate, the findings must be both reliable and valid. Reliability means that the findings are consistently the same even if researchers repeat the study; and validity refers to the truthfulness of the findings, which means that the study actually measures the intended elements (Calder, Phillips, & Tybout, 1982). Although many different threats to validity as well as reliability exist, internal validity is an important early consideration. Internal validity refers to the choice of the most appropriate research design for the topic of study (i.e., experimental, quasi-experimental, survey). This study deems survey research as an appropriate method, and finds that the external validity threat of selection bias impacts how well inferences from the results of the research generalize to the target population and how confident this generalization is. An issue specifically arises when nonprobability sampling is the most cost effective and realistic choice of the researcher, which makes the sampling strategy a difficult decision. Thus, researchers often face tradeoffs to achieve validity and reliability in their studies while trying to justify their choices.

To further explain the differences between selecting probability and nonprobability sampling strategies, a comparison to statistical theory is most appropriate. For probability sampling, the expectation is that, if researchers repeat the sample, then they achieve similar results and make the same sampling inferences. The only difference between the selection and the non-selection of units in the sample is the start of the random number generator. Because the sample is a finite set, probability determines the selection of a unit in the sample. This model is a design based sampling model that uses a randomization theory approach that does not need distributional assumptions. In contrast, in nonprobability sampling, the probability does not determine the selection of the units. This sampling method is a model based approach where, if the model is not true, then sampling estimates might be severely biased (Lohr, 1999).

Justification of a valid sampling method becomes even more critical when researchers seek to compare the subjective or objective characteristics of two or more homogeneous groups (Mullen, Budeva, & Doney, 2009). For example, in the cross-cultural domain, sampling strategies include convenience student sampling for experimental designs (Mikhailitchenko, Javalgi, Mikhailitchenko, & Laroche, 2009; Ueltschy, Laroche, Zhang, Cho, & Yingwei, 2009), multi-stage random sampling for descriptive survey research (Rojas-Méndez, Davies, & Madran, 2009), convenience sampling with locals for descriptive survey research (Chang & Hsieh, 2006), and restricted student sampling for descriptive survey research (Lopez, Babin, & Chung, 2009). Under ideal conditions, to achieve a sample that is representative of the comparative groups of interest, researchers must divide the population into meaningful subpopulations, or strata, that coincide with the domain context of the study. For example, if the purpose of a study is to compare educational workforce experiences of female and male engineering graduates, the basis for the strata is gender (McIlwee & Robinson, 1992). Identification of the strata specifically makes the sampling strategy more efficient if the populations of male and female engineering graduates are not equal, because random sampling from each subgroup allows the researchers to obtain more precision for their comparative groups. Thus, precision means that the variance within each subgroup is more likely to be lower than when compared to the variance in the whole population.

However, a probability sampling strategy can be too costly or impractical, leaving researchers with no choice but to select a nonprobability sampling strategy for comparing groups. The strategy is similar to probability sampling in that, initially, the strategy identifies meaningful subgroups where the variance is minimal. The specific context in this study compares the perceptions among rural and urban residents of a potential transportation tax policy, but a random selection of multitudes of locations throughout the state to draw the sample from is too costly and is not feasible. Therefore, the first challenge is to consider the optimal number of sample locations while minimizing cost and, secondly, to identify those locations that are most representative of the criteria for data collection.

2.2. Sample location identification problem: rural and urban residents in a state

The overall objective of this study is to determine what perceptual differences exist, if any, between constituents residing in urban versus rural counties within the state of Nevada with regard to a potential transportation tax policy. Prior social sciences research that uses inter-group analysis within the socio-cultural context indicates that communication and technological innovations significantly polarize rural and urban residents (Penz, 2006). This research supports the concept that these groups should also be targeted in different ways with regard to these innovations. A basis for a reasonable method is to create strata from urban and rural regions in Nevada to sample. However, the issue remains to identify which locations in both urban and rural areas are potentially representative enough for data collection. The fact that Nevada is a geographically dispersed state complicates this problem. The southern and northern regions of the state both include large populations of urban and rural residents. The evidence of these populations is the location of the two interstates that go through the northern and southernmost regions of the state (I-80 and I-15 respectively). However, no interstate connects the northern regions of the state to the southern regions. Because of the dispersion within the state, the problem centers on the extent to which a convenience sample can truly represent the population groups of interest for comparison.

To achieve reliable results for comparative purposes and to control for extraneous factors, identification of urban and rural locations within the state that accurately represent these two homogeneous groups is important. The use of a combinatorial optimization method helps to determine the total number of locations and the specific locations that are most representative of the population for the comparison groups. Operations research and engineering use combinatorial optimization with the primary goal of selecting the optimum from a set of finite objects (Schrijver, 2005).

3. A combinatorial optimization method for sample location identification

The sampling method that this paper provides is a nonprobability bi-level stratified cluster sampling technique. The first level of stratification is the division into rural and urban areas. The second level of stratification comes from dividing the urban and rural areas into their various counties. This project for data collection has limited funding, so managing costs is a major criterion. Therefore, a simple random sampling that involves collecting data at distributed geographical locations is not feasible. In light of this, the project uses census data from 2008 for the different counties to help identify sampling locations. Because the research context relates to individual perceptions of a potential transportation tax policy (Krishen, Raschke, & Mejza, 2010), the project collects the following variables in relation to working populations for each county: average travel time to work, mean household income, percentage of high school graduates, and percentage of population between the ages of 18 and 65.

To identify the appropriate county and locations, the project formulates a combinatorial optimization method to show the representation factor and cost factor in the analysis. The state represents a fixed number of locations, and the objective is to minimize the

resulting function to obtain the optimal number of sample locations (representation) with the least amount of population variation (cost) for each comparison group. The total number of locations is chosen to manage travel and sampling costs.

Based on the project objective of comparing perspectives between urban and rural residents, the combinatorial optimization method examines locations for each group in Nevada. Because collecting data from every location is cost prohibitive, this method chooses a subset for each comparison group with the best mix of cost savings and representativeness. To clarify the approach, consider the two farthest extremes: (1) collect data from every location, maximizing cost and representation; or (2) collect data from one location, greatly reducing the cost and representation. Thus, a balance between these two extremes is best. The two key decisions are: (1) ascertain how many sites to take; and (2) find out which specific sites those should be. These two decisions must balance maximized representation at minimized cost. For example, to find two locations, the combinatorial optimization allows for the comparison of all possible combinations using the value computed from the objective function and determines the best two location pair results.

To demonstrate the reasoning behind this heuristic, panel A of Fig. 1 illustrates an example that assumes five random variables representing sample selection locations. However, due to costs, only one sample location can be selected.

For illustrative purposes, panel A of Fig. 1 assumes normal distributions with different means, but the same variance for all of the random locations. The objective is for the sample location selection to provide the mean of the entire population when cluster samples come from that location. In this case, only one sample location provides the population mean, X_3. Hence,

$$E\left(\frac{\sum_{i=1}^{5} X_i}{5}\right) = E(X_3), \tag{1}$$

where $E(\cdot)$ is the expectation operator. This location also has the least variance for the population. When selecting two samples, the only set of locations that give the same mean as the population is the two sets of symmetric pairs ((X_2,X_4), and (X_1,X_5)). The location pair that is variance minimizing is (X_2,X_4). An ideal process would sample every location for good results, but that is prohibitively costly and

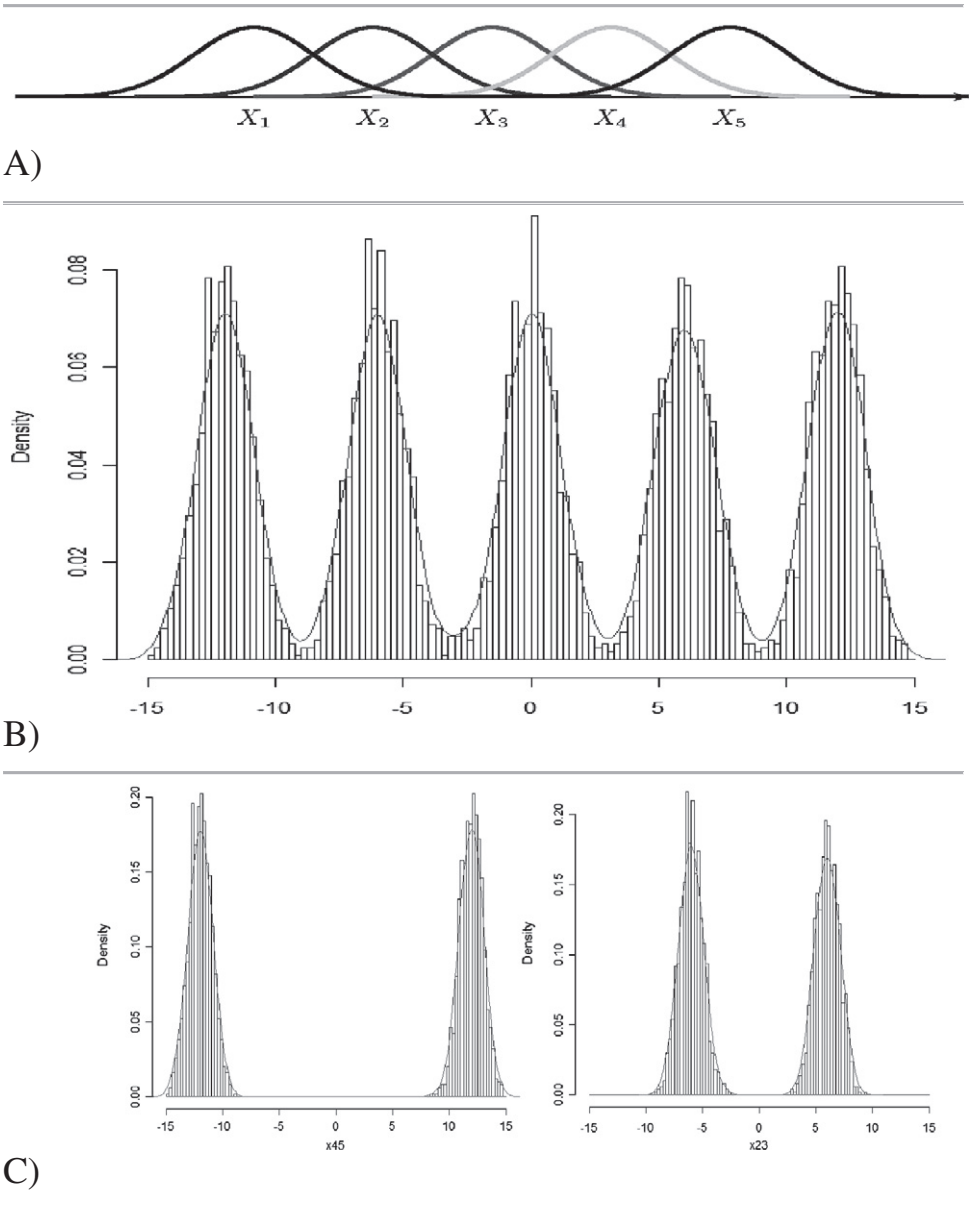

Fig. 1. Sample location identification simulation. A: heuristic illustration. B: simulation of normally distributed dataset. C: simulation of sample location selection.

not feasible; so, instead, the project samples one subsample to provide representativeness for all locations. Thus, the project conducts a similar analysis for n locations out of N. To illustrate this design heuristic further, panel B of Fig. 1 shows a simulated normally distributed dataset over five different means creating 1000 points each.

As demonstrated in panel C of Fig. 1, samples from X_4 and X_5 have a mean of −0.02, and a standard deviation of 11.99. On the other hand, samples from X_2 and X_3, have a mean of −0.01, and a standard deviation of 6.09. Because of a finite population from which to select locations within rural and urban areas of Nevada, the objective is to select the optimal number of locations. In addition, the goal requires a location from the most homogeneous groups possible by using additional information on transportation related attributes from the census.

As mentioned previously, collecting data from a large number of randomly selected locations is not feasible due to cost so the objective is to maximize representation (performance) and minimize cost. Therefore, the project develops a formula that provides the performance cost as a function of the number of sampling locations and the variance. The performance cost function, $F(p,n)$, solves for the optimal number of locations with the minimal amount of variance based on a matrix of variables associated with the population characteristics that potentially impact the proposed transportation tax policy. The use of the performance cost function commonly controls for system development with the goal of finding optimal performance at minimal cost (Levine, 1999). For the case described in this paper, the goal is to find the optimal locations with the most homogeneity (minimum variance). The performance cost function $F(p,n)$ is a function of two components. The first part is the representation function and the second part is the cost function. The first variable p is a vector that contains n elements. These n elements correspond to the n samples that are used as inputs. The variable p_1, for instance, is the first sample and is equal to some x_j, which indicates that the first sample is location j and that the variable p_1 is the same as the data x_j at location j. A finite total of N data points corresponds to data vector x, which is $x_1,x_2,\ldots x_N$, and calculates the performance cost associated with choosing n samples $p_1,p_2,\ldots p_n$. Cost increases as the number of sample locations increases. An important differentiation in this formulation is that variance comes from the whole sample space instead of only from the sample mean. The cost function has the flexibility to calculate optimization by taking one sample at a time, or two samples at a time, etc. The axioms that function $F(p,n)$ must satisfy are:

1. The function has additive cost with respect to the two variables, because that is a reasonable structure for a cost function: $F(p, n) = f(p) + g(n)$.
2. Function $g(n)$ is a monotonically increasing function of n, because increasing the number of samples increases the cost of sampling.
3. When all the choices for locations have been made, the entire location set exists, and hence the variance from the entire set is zero; if $n = N$, then $f(p) = 0$.
4. The distance between a sample and other data points must be used for weighting in variance calculation. This weighting is a monotonically increasing function of distance that makes the assumption that the locations that are near each other are automatically representative of similar populations. Eq. (2) presents a specific function that satisfies these conditions. This function uses a quadratic function for $g(n)$ and a linear distance weighting function in $f(n)$.

$$F(p,n) = \left[\sum_{j=1}^{n} \cdot \sum_{i=1}^{N_j} \left(d_{ij}\right) \frac{\left(x_{ij} - p_j\right)^2}{w_{ij}}\right] + \beta n^2 \tag{2}$$

where:

d_{ij}	distance between counties i and j
w_{ij}	percent population average of i and j counties $= \frac{w_i + w_j}{2}$
x_{ij}	independent variable
p_j	independent variable corresponding to x_i for some i
β	constant
n	number of samples
N	number of locations.

Because the size of this sample space is finite, differential calculus-based optimization methods are not applicable for this combinatorial optimization method. As n is also a variable and the method performs optimization over all the possible values of n, the cardinality of the search space becomes

$$\sum_{n=1}^{N} \frac{N!}{n!(N-n)!}. \tag{3}$$

Optimization over this set is also combinatorial, and the solution comes from the use of an explicit computation technique. The method uses a Matlab program to solve Eq. (2) that indicates that the optimal value of the function is $n = 2$. Out of twelve urban counties, Clark County and Washoe County have the minimum values after the combination of all of the independent variables. This result means that these two urban counties have the least amount of population variation on the basis of the desired census variables.

Once the solution to the upper level strata exists and identifies the urban and rural counties of interest, the lower level strata need to identify the zip codes within the selected counties. Clark County has 58 zip-code locations and Washoe County has 19. Table 1 shows the Matlab results for Washoe County that identify the most homogeneous locations within this county (zip codes 89502 and 89431). Once the researchers identify the locations, they focus on collecting their nonprobability sample data.

This method provides a strategy for targeting sample locations for a convenience sample. The alternative is no specific location identification strategy where the researcher selects a location completely at random (or by convenience) to collect data. The results in Fig. 2 illustrate a comparison of these strategies. The choices of zip codes for a county are random and the table shows the calculations and averages of the cost function over the number of selected locations. The combinatorial optimization method calculates a comparison of the outcome to the variance function. Fig. 2 depicts the results that show that the combinatorial optimization method produces less variance than selecting a location completely at random. These results mean that the locations that the combinatorial optimization method identifies

Table 1
Combinatorial optimization output for Washoe County, Nevada.

Zip code	Income	School	Age	Sum
89405	353.03	584.66	108.62	1046.32
89424	500.50	160.43	231.87	892.81
89431	10.97	5.90	1.49	**18.35**
89433	15.44	10.04	4.49	29.97
89434	14.58	17.07	2.60	34.26
89436	50.06	26.79	4.75	81.60
89439	300.28	295.95	108.82	705.05
89442	241.38	123.84	59.81	425.03
89451	218.14	106.76	20.84	345.74
89501	175.11	43.49	113.02	331.62
89502	10.46	5.00	1.25	**16.71**
89503	11.29	12.45	4.60	28.34
89506	11.71	10.96	2.43	25.09
89509	10.18	14.17	3.14	27.49
89510	143.71	98.87	25.18	267.76
89511	32.97	27.53	2.67	63.17
89512	27.22	12.14	2.07	41.43
89523	35.51	44.11	3.12	82.74
89704	245.47	138.58	30.05	414.09

Bold items signify the most homogenous locations.

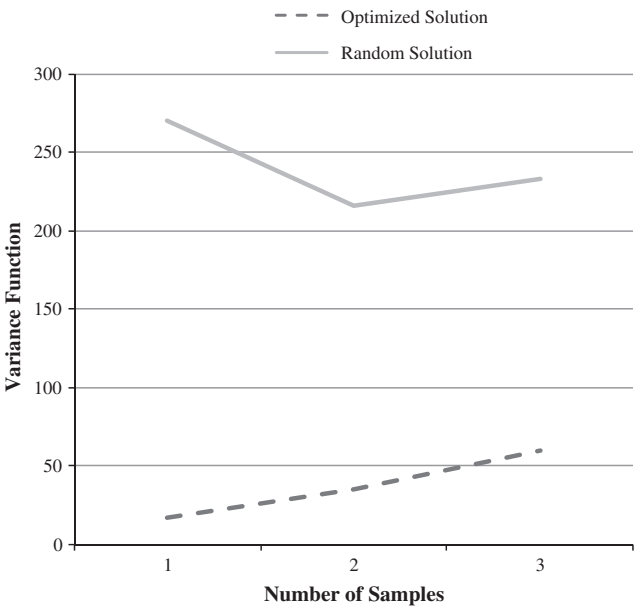

Fig. 2. Comparison of variance function for locations identified using combinatorial optimization solution and random solution.

are much more homogeneous populations. The graph in Fig. 2 demonstrates that as the number of locations to sample increases, so does the variance; however, under the optimization approach, the variance remains less than the random location selection.

4. Discussion and conclusion

The proposed combinatorial optimization method demonstrates a reasonable sample identification strategy for researchers when their research objective is to identify one or more homogeneous group samples using a nonprobability sample. Many studies use nonprobability sampling and indeed in some domains high criticism of the use exists (e.g., international and cross-cultural research) (Reynolds, Simintiras, & Diamantopoulos, 2003), while other domains, such as marketing research, deem the sampling useful (Deville, 1991). The minimization of concerns about nonprobability sampling requires a reasonable justification for the method.

Although the combinatorial optimization method can be useful for researchers electing to use a convenience sample and can provide a reasonable justification for selecting the locations for the sample, researchers should consider the following steps prior to using this approach. First, the sampling objective should be for comparative purposes where a need exists to identify distinctive groups in which the desired sampling attributes are to eliminate extraneous factors and maintain homogeneous groups as much as possible. In essence, the homogeneity within the sample reduces the likelihood that the differences among groups are due to extraneous variables and instead are the result of differences between the constructs of interest. Secondly, researchers need to use reasonable demographic variables available within their research context for the combinatorial optimization method to identify the most representative population locations, because the purpose of this method is to minimize the variance for each comparison group. Within the context of the problem this paper presents, the first objective distinguished between rural and urban comparative groups on the basis of government designated commuting codes. Additionally, as identified from the census data, demographic information relating to those who are most likely to work and drive is useful in finding the most homogeneous areas to sample. If the wrong variables are considered, the groups might be homogeneous for the wrong reasons and weaken the interpretation of the results.

An appropriate sampling strategy is crucial to the validity of the results. Internal validity can be threatened if any extraneous variables affect or influence the dependent variable. In relation to sampling, two methods for controlling or minimizing the threat of extraneous variables are suggested: (1) the selection of homogeneous groups; or (2) the random selection of a sample (Kerlinger, 1986). The combinatorial optimization method presented provides a strategy for justifying the sample locations selected to encourage homogeneity.

This paper demonstrates a solution for identifying sample locations when the research is for comparative purposes, and a probability sample is prohibitively costly and impractical. Much marketing and social sciences research uses nonprobability samples; however, the approach this paper presents helps strategize sample location identification and provides additional justification as to the technique and method for increasing the homogeneity of the comparative samples.

References

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1982). The concept of external validity. *Journal of Consumer Research*, *9*, 240–244.
Chang, J., & Hsieh, A. -T. (2006). Leisure motives of eating out in night markets. *Journal of Business Research*, *59*, 1276–1278.
Deville, J. -C. (1991). A theory of quota surveys. *Survey Methodology*, *17*, 163–181.
Ferber, R. (1977). Research by convenience: Editorial. *Journal of Consumer Research*, *4*, 57–58.
Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). New York: Holt Rinchart and Winston.
Krishen, A. S., Raschke, R. L., & Mejza, M. (2010). Guidelines for shaping perceptions of fairness of transportation infrastructure policies: The case of the vehicle mileage tax. *Transportation Journal*, *49*, 24–38.
Levine, W. S. (1999). *Control system applications.* Boca Raton: CRC Press.
Lohr, S. L. (1999). *Sampling: Design and analysis.* Pacific Grove: Duxbury Press.
Lopez, T., Babin, B., & Chung, C. (2009). Perceptions of ethical work climate and person–organization fit among retail employees in Japan and the US: A cross-cultural scale validation. *Journal of Business Research*, *62*, 594–600.
McIlwee, J. S., & Robinson, J. G. (1992). *Women in engineering: Gender, power, and workplace culture.* Albany: University of New York Press.
Mikhailitchenko, A., Javalgi, R., Mikhailitchenko, G., & Laroche, M. (2009). Cross-cultural advertising communication: Visual imagery, brand familiarity, and brand recall. *Journal of Business Research*, *62*, 931–938.
Mullen, M., Budeva, D., & Doney, P. (2009). Research methods in the leading small business-entrepreneurship journals: A critical review with recommendations for future research. *Journal of Small Business Management*, *47*, 287–307.
Penz, E. (2006). Researching the socio-cultural context: Putting social representations theory into action. *International Marketing Review*, *23*, 418–437.
Reynolds, N. L., Simintiras, A. C., & Diamantopoulos, A. (2003). Theoretical justification of sampling choices in international marketing research: Key issues and guidelines for researchers. *Journal of International Business Studies*, *34*, 80–89.
Rojas-Méndez, J., Davies, G., & Madran, C. (2009). Universal differences in advertising avoidance behavior: A cross-cultural study. *Journal of Business Research*, *62*, 947–954.
Schrijver, A. (2005). On the history of combinatorial optimization (Till 1960). In K. Aardal, G. L. Nemhauser, & R. Weismantel (Eds.), *Handbooks in operations research and management science*, *12*, Amsterdam: Elsevier.
Ueltschy, L., Laroche, M., Zhang, M., Cho, H., & Yingwei, R. (2009). Is there really an Asian connection? Professional service quality perceptions and customer satisfaction. *Journal of Business Research*, *62*, 972–979.