The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

Stability and similarity of clusters under reduced response data

Marisciel Litong-Palima[*1], Kristoffer Jon Albers[*2] and Fumiko Kano Glückstad[*1]

[*1] Copenhagen Business School, DK-2000 Frederiksberg, Denmark [*2] Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

This study presents a validated recommendation on how to shorten the surveys while still obtaining segmentation-based insights that are consistent with the analysis of the full length version of the same survey. We use latent class analysis to cluster respondents based on their responses to a survey on human values. We first define the clustering performance based on stability and similarity measures for ten random subsamples relative to the complete set. We find foremost that the use of true binary scale can potentially reduce survey completion time while still providing sufficient response information to derive clusters with characteristics that resemble those obtained with the full Likert scale version. The main motivation for this study is to provide a baseline performance of a standard clustering tool for cases when it is preferable or necessary to limit survey scope, in consideration of issues like respondent fatigue or resource constraints.

1. Introduction and motivation

Consumer segmentation research based on survey data has benefited from the availability of more efficient and economical means to collect survey responses, like the proliferation of survey hosting platforms, like Qualtrics [https://www.qualtrics.com/], SurveyMonkey [https://www.surveymonkey.com/], Google Surveys [https://surveys.google.com/] and SurveyXact [https://www.surveyxact.dk/] as well as crowdsourcing online platforms where the surveys can be quickly deployed, like the Amazon Mechanical Turk [https://www.mturk.com/] and Prolific Academic [https://www.prolific.ac/]. As the use of these tools are likely to increase, it also becomes likely that more and more surveys will be deployed which poses a challenge for recruiting respondents and for maintaining the quality of the responses. Hence, there is a timely need to optimize the design of surveys to counterbalance these effects.

This study aims to address an aspect of survey design related to the presentation of questions in order to minimize the survey length. The intention is to maintain consistency and reliability in the gathered responses in order to provide better input to computational tools for segmentation. While having more questions in a survey intuitively provides more data for extracting insight, this study considers how the questions can be presented in order to mitigate issues like respondent fatigue, wherein the response rates and quality decreases with survey length [Burchell 1992, Rolstad 2011]. The goal is to have more data of higher quality to support the computational tools in segmentation because these algorithms perform better when data is substantial and of good quality.

In this article, we describe a procedure for comparing the relative baseline performance of a standard clustering method when applied to survey data that has been reduced or limited as compared to the performance when clustering is applied to the original and full data. There are two interpretations of data reduction considered in this study: (1) having fewer question items, i.e. 21 to 10 questions; and (2) having fewer response levels, i.e. Likert to binary. The basic motivation for this study is to provide a recommendation for designing surveys that can be completed by respondents in less time and still provide data for clustering solutions of expected stability and similarity to the full data solution, and more importantly with comparable mean responses or item profile. This work follows closely the motivations and recommendations discussed in [Dolnicar 2007, Dolnicar 2011].

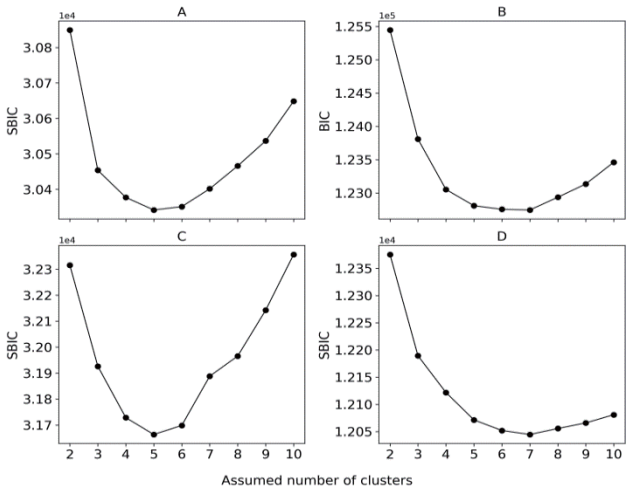

Figure 1 Based on the best, i.e. least, Bayesian information criterion (BIC) or its sample-adjusted version (SBIC), the LCA model with five clusters is a reasonable model to assume for all the cases of data sets (A) WVS Likert 10Q, (B) ESS Likert 21Q (C) PRF Likert 21Q and (D) PRF True Binary 21Q.

2. Human value survey data

As listed in Table 1, we use three samples of human value survey data: Sets 1 and 2 are secondary sources from the World Values Survey [http://www.worldvaluessurvey.org/] and the European Social Survey [http://www.europeansocialsurvey.org] respectively while Set 3 is a primary source that was disseminated through SurveyXact and Prolific Academic

Contact: Marisciel Litong-Palima, Copenhagen Business School, DK-2000, Frederiksberg, Denmark, Phone: +45 3815 3238, Email: mpa.msc@cbs.dk .

platforms. All respondents are from the United Kingdom. The questionnaires used in the three sets have 10 human value questions in common. The ESS and the PRF have 11 human value questions in addition.

Table 1 Description of data sets used. N is the number of questions in the survey set.

Set	Source	N	Sample size	Remarks
1	WVS	10	989	Secondary/Interview-based
2	ESS	21	2005	Secondary/Interview-based
3	PRF	21	516	Primary/Online

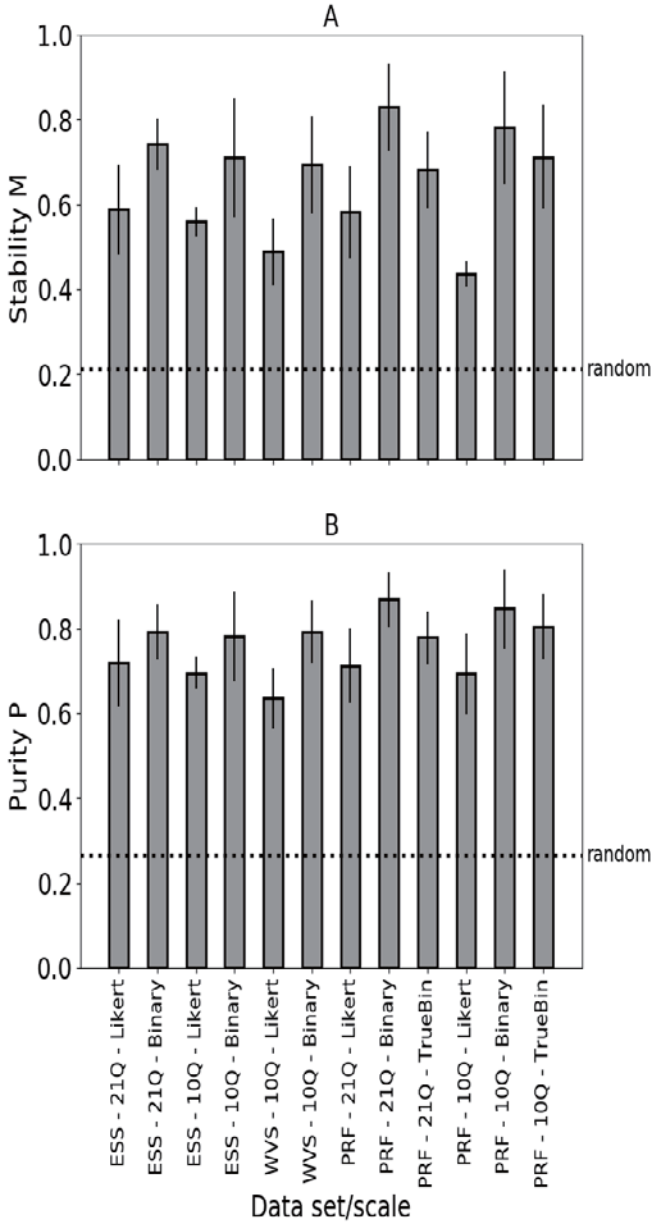

Figure 2 (A) Stability M and (B) similarity/purity P for the ESS, WVS and PRF data cases considered. The labels reflect the data case type. For example, PRF-10Q-TrueBin means the data is from the online survey Set 3 with 10 human value using the multiple-choice/tick box, i.e. true binary response scale.

In this study, we define the case of full data to correspond to a data set with 21 questions on a 6-scale Likert response scale, which consists of two negative ("Not at all like me", "Not like me") and four positive ("A little like me", "Somewhat like me", "Like me", "Very much like me") response levels, e.g. ESS. The WVS data set has at most ten value questions in common with the ESS data set and is thus an example of a data set with reduced item number. The PRF data set has the same 21 questions as the ESS but for some questions, response is limited to either a positive or a negative, i.e. reduced response levels.

Note that the clustering solutions for each of the different data sets cannot be directly compared. In order to make a meaningful comparison across the different data sets, we use a relative measure of performance that compares the clustering solution obtained with ten random subsamples to that obtained with the complete sample. To quantify performance, we use the relative measures of stability M and similarity measure P, also known as purity score.

$\mathbf{M} = \ll \delta_{\mathbf{Trc}, \mathbf{T}^{(\mu)}_{\mathbf{rc}}} \gg_{\mathbf{m}}$, where the values of $\mathbf{T_{rc}}$ and $\mathbf{T^{(\mu)}_{rc}}$ are the connectivity matrices for the complete and the resampled data, and $\ll\gg_{\mathbf{m}}$ is a two-step averaging over the neighbors in the complete set and then over the random samples, as introduced in [Levine 2001].

$\mathbf{P} = \Sigma_{\mathbf{r}}\, \mathbf{p_r}\, (\mathbf{max_c}\, (\mathbf{p_{rc}}\, /\mathbf{p_r}))$, where the values of $\mathbf{p_r}$ and $\mathbf{p_c}$ are probabilities of membership to clusters $\mathbf{r}$ and $\mathbf{c}$ obtained using the reduced and the complete data sets respectively and $\mathbf{p_{rc}}$ is the probability for membership to both.

We then compare **M** and **P** across the different cases of reduced data. Cases are labelled to indicate the following:

- 21Q – complete 21 questions on human values
- 10Q – reduced 10 questions on human values
- Likert – complete scale with six values consisting of 4 positive and 2 negative levels
- Binary – reduced scale with two levels converted from the negative and positive levels of the Likert scale
- True Bin – reduced scale, asked directly to the respondents as a multiple choice-tick box type (i.e. not converted)

3. Method to estimate relative baseline of clustering performance and results

We choose latent class analysis (LCA) [Lazarsfeld 1968] as implemented in the MPlus software [https://www.statmodel.com/] as our standard clustering method. Similar clustering approaches are used for exploring questionnaire data and examining human behavior and value priorities [Szakolczai1998, Moors2009, Magun2015]. LCA is implemented in MPlus based on a mixture model of multivariate distributions for categorical variables. In doing clustering analysis, one assumes that the respondents belong to significantly different groups but that this grouping is unknown or cannot be observed directly and can only be inferred from analyzing the data. LCA is a method to unmix the respondents into these

distinct but unobserved clusters based on the over-all structure of the entire response data.

We choose the Bayes information criterion (BIC) and its sample-adjusted version SBIC to decide how many latent clusters there are in the data set. The BIC has been shown to be a better and consistent criterion in inferring the unknown number of clusters for a given sample, even when the item profile or structure in the data is complex and even when the cluster sizes are significantly different [Nylund 2007]. For this study, we found that the related criterion SBIC gives a clearer indication of the optimal values for most of the data set cases. The use of either BIC or SBIC is enough for the purposes of this study. Results based on the BIC and SBIC, as shown in Figure 1 indicate minimum values around five to six clusters across all the different data sets considered. For this study, we choose models with five clusters.

Figure 2 shows the stability and similarity measures across the full and reduced data cases. All solutions have stability and similarity measures that are at least twice as much as random solutions, whose corresponding measures are **M**~0.21 and **P**~0.27 on average.

The clustering solutions obtained using binarized or true binary scales consistently perform better when compared against their Likert-based counterparts. This suggests that converting the scale to binary or using a binary scale designed into the survey can give clustering solutions that are more robust with **M**~0.7 and highly similar with **P**~0.8.

When it comes to the reduction of questions, performance with just ten questions are in general less stable and with less similarity, but nevertheless stability has value **M**~0.5. Note that stability is least for the PRF-10Q-Likert case, even though the response scale is Likert, and hence not reduced.

The PRF-10Q-True Bin has performance measures that are nearly the same as the PRF-21Q-True Bin cases for both **M** and **P**. This is a useful result as it supports the question style where respondents are asked just once to account which among 21 human values they relate to, instead of being asked a list of 21 questions. This supports a shorter survey design.

We recognize that the process of converting the scale to binary or the use of a true binary scale, by definition, reduces the variations in the possible responses and reduces the information available for the LCA algorithm. Nevertheless, what these stability and similarity results suggest is that despite the reduction in the information due to a limited scale, the clustering solutions obtained are still descriptive and can still reveal the clusters that resemble those obtained in the full length or scale version. To illustrate this claim, Figure 3 shows the matching in the cluster profile for the PRF-21Q-Likert case compared with the PRF-21Q-True Bin case. Pairing the clusters from the two solutions is easily done using a visual check. The purity score computed directly between the two solutions gives **P**~0.46 and supports well the correspondence of the structure of the item profiles for both solutions. The characteristics of the clusters suggested by the solution from PRF-21Q-Likert case are well-retained in the solutions from the corresponding true binary case.

4. Summary and Conclusion

The tendencies with regards to the clustering performance of latent class analysis on random subsamples relative to the complete sample and across the cases of reduced response data are summarized as follows:

- The average **M** values across different reduced data cases is 0.55 for the Likert cases and 0.75 for the binary cases. **M** is 0.21 for a purely random solution.
- The average **P** values across different reduced data cases is 0.65 for the Likert cases and 0.81 for the binary cases. **P** is 0.27 for a purely random solution.
- The reduction of the number of questions have less impact on the decrease in **M** and **P**, although can account for more variability.
- Even under a reduction in the response levels, cluster characteristics are well maintained. In the case of the PRF-21Q-Likert and PRF-21Q-True Bin, the radars plots can even be easily matched and the purity score is at 0.46.

This study recommends to reduce the amount of response data from human values survey designs primarily by using true binary response scales, particularly when the goal is to cluster the respondents. This can be particularly useful in studies where the value questions are included only as part of a more comprehensive questionnaire, and hence reducing parts of the supporting response data may reduce the overall cost of the study and limit respondent fatigue.

Though the present study is based on LCA on human value questions, the presented approach can easily be extended to other domains and clustering models as well.

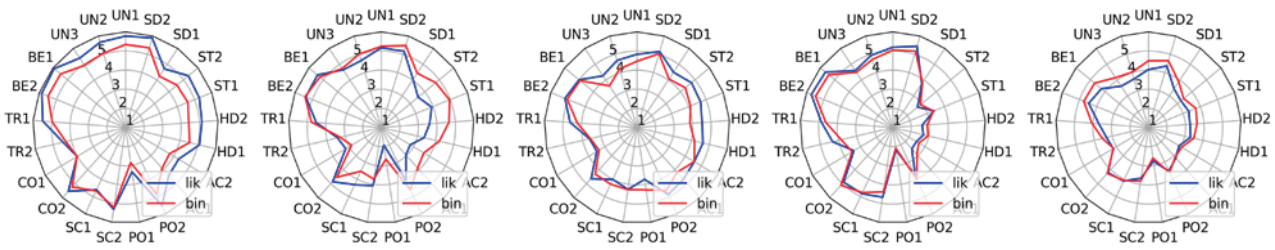

Figure 3 Comparison of the mean responses for each of the five cluster for the cases (A) PRF-21Q-Likert and (B) PRF-21Q-True Bin.

Acknowledgement

This work has been conducted as part of “UMAMI: Understanding Mindsets Across Markets, Internationally” No. 61579-00001A funded by Innovation Fund Denmark.

References

[Burchell 1992] Burchell, Brendan, and Catherine Marsh. "The effect of questionnaire length on survey response." *Quality and quantity* 26.3: 233-244, Springer, 1992.

[Dolnicar 2007] Dolnicar, Sara, and Bettina Grün. "How constrained a response: A comparison of binary, ordinal and metric answer formats." Journal of Retailing and Consumer Services 14.2: 108-122, Elsevier, 2007.

[Dolnicar 2011] Dolnicar, Sara, Bettina Grün, and Friedrich Leisch. "Quick, simple and reliable: Forced binary survey questions." International Journal of Market Research 53.2: 231, Elsevier, 2011.

[Lazarsfeld 1968] Lazarsfeld, P.F. & Henry. N.W. “Latent structure analysis.” New York: Houghton Mifflin, 1968.

[Levine 2001] Levine, Erel, and Eytan Domany. "Resampling method for unsupervised estimation of cluster validity." Neural computation 13.11: 2573-2593, MIT Press, 2001.

[Magun 2015] Magun, Vladimir, Maksim Rudnev, and Peter Schmidt. "Within-and between-country value diversity in Europe: A typological approach." European Sociological Review 32.2 (2015): 189-202.

[Moors 2007] Moors, Guy, and Jeroen Vermunt. "Heterogeneity in post-materialist value priorities. Evidence from a latent class discrete choice approach." European Sociological Review 23.5 (2007): 631-648.

[Nylund 2007] Nylund, Karen L., Tihomir Asparouhov, and Bengt O. Muthén. "Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study." Structural equation modeling 14.4 (2007): 535-569.

[Rolstad 2011] Rolstad, Sindre, John Adler, and Anna Rydén. "Response burden and questionnaire length: is shorter better? A review and meta-analysis." *Value in Health* 14.8: 1101-1108, Elsevier, 2011.

[Szakolczai1998] Szakolczai, Arpád, and László Füstös. "Value Systems in Axial Moments A Comparative Analysis of 24 European Countries." European Sociological Review 14.3 (1998): 211-229.