ESOMAR | 70

Data Confidential

How to link surveys with internal data while maintaining privacy

Diane Berry • Josef Rieder

ESOMAR
Office address:
Atlas Arena, Azië Gebouw
Hoogoorddreef 5
1101 BA Amsterdam
Phone: +31 20 664 21 41
Fax: +31 20 664 29 22

Email: events@esomar.org
Website: www.esomar.org

Publication Date: October 2017
ESOMAR Publication Series Volume S382
ISBN 92-831-0298-3

Copyright

Published by ESOMAR, Amsterdam,
The Netherlands
Edited by: Deborah S. Fellows

About ESOMAR

ESOMAR is the global voice of the data research and insights community, representing a network of 35,000 data professionals.

With more than 4,900 members from over 130 countries, ESOMAR's aim is to promote the value of market and opinion research in illuminating real issues and bringing about effective decision-making.

To facilitate this ongoing dialogue, ESOMAR creates and manages a comprehensive programme of industry specific and thematic events, publications and communications, as well as actively advocating self-regulation and the worldwide code of practice.

ESOMAR was founded in 1948.

About ESOMAR Membership

ESOMAR is open to everyone, all over the world, who believes that high quality research improves the way businesses make decisions. Our members are active in a wide range of industries and come from a variety of professional backgrounds, including research, marketing, advertising and media.

Membership benefits include the right to be listed in the ESOMAR Directories of Research Organisations and to use the ESOMAR Membership mark, plus access to a range of publications (either free of charge or with discount) and registration to all standard events, including the Annual Congress, at preferential Members' rates.

Members have the opportunity to attend and speak at conferences or take part in workshops. At all events the emphasis is on exchanging ideas, learning about latest developments and best practice and networking with other professionals in marketing, advertising and research. Congress is our flagship event, attracting over 1,000 people, with a full programme of original papers and keynote speakers, plus a highly successful trade exhibition. Full details on latest membership are available online at www.esomar.org.

Contact us

ESOMAR

ESOMAR Office:
Atlas Arena, Azië Gebouw
Hoogoorddreef 5
1101 BA Amsterdam
The Netherlands
Tel.: +31 20 589 7800

Email: customerservice@esomar.org
Website: www.esomar.org

Data Confidential

How to link surveys with internal data while maintaining privacy

Diane Berry • Josef Rieder

Introduction

Strategic primary research is a key tool for companies to understand how the public and their customers perceive them in relation to competitors. It is often desirable to link the survey results to customer databases (CRM). Due to data protection regulations, the linkage of survey insights to internal data is generally approached in an aggregated manner.

Common methods to link insights from surveys to internal data sources are either to build in survey questions that identify predefined customer segments that link directly with internal segmentation. Another approach to linkage is to segment from selected survey questions and devise tagging equations, which make use of common variables across the two data sources. In our suggested new approach, we have applied techniques from record linkage to link survey data with internal data, whilst still maintaining the individual anonymity.

With our approach, analytic practitioners can profit from several benefits. We are able to enrich the databases with more details than just a single segmentation variable, so that the combined data is more useable for tactical objectives. It is more time efficient as no pre-stage segmentation is required. If desired, a segmentation still can be built from the linked data, running it directly on the CRM.

We have applied the described approach on a real project example, where we were able to demonstrate clearly a better discrimination for match and link variables compared to a segment tagging approach.

When the foremost aim of a survey is to enrich the customer database with external data, then we recommend this approach as an alternative to segmentation-first.

Method specification

Segment tagging – the classic way to link survey data to customer databases
Companies conduct surveys to get a better understanding of their customers, their needs and attitudes, their purchase and usage behaviors, the value they contribute to the bottom line and how all these elements vary by demographics.

During the insights generation phase often customer segmentations are derived. They highlight characteristic patterns and profiles of needs, behaviours and demographics. Typical statistical methods for segmentation are cluster analysis or latent class analysis.

Companies can tailor value propositions towards segment profiles, and more specifically, new products and services, based on segment-specific price sensitivities. Often segments differ by how customers are thought to react towards marketing activities. For example, to acquire customers of a specific segment a specific mix of omni-channel capabilities are required. Likewise, cross-selling opportunities may be identified and can vary across segments. Segments can have typical critical touch points for customer service and potentially, can show different churn behaviour. Customer segmentations are also very powerful method to communicate the analysis.

For the operationalization of most of the segmentation objectives, it is highly desirable to tag the segments to the database. Usually this is accomplished by developing a segment tagging algorithm often using transactional/ behavioral data and customer demographics. There are several ways to build such segment membership predicting models, e.g. linear discriminant models, multinomial logit regression with penalization, or decision trees models. The key is that link variables exist in both the survey data and in the customer database.

When deploying the segment tagging algorithm, each customer of the database is assigned to one segment, so that we learn something about the needs and attitudes of customers, information which previously was non existent in the database. While this information is highly useable and desirable, it is only available at highly aggregated level. For example, if there are seven segments, then only seven different profiles for the match variables (e.g. needs, attitudes) exist. This might be good enough for several segmentation applications, but can still be too high level, especially for more tactical marketing objectives like assortment adjustments, marketing spend optimization or service offering improvements. This desire for more granular data matching was the origin of our research.

Data confidentiality considerations

To accomplish a more granular match it would be easiest to use a 1-to-1 match where possible. When conducting the survey for a customer segmentation the sample is often taken from the customer database. Then such a back linking of the survey responses to the database would be theoretically possible. However, the regulations of the market research industry prohibit this practice for good reasons.

In the new efamro/ESOMAR guidance on data protection, which will come into effect in all EU member states in May 2018, this is well formulated: "All research needs to be based on robust data protection measures to build trust and meet the significant regulatory and legal requirements" (efamro/ESOMAR, 2017). This is in addition to general privacy principles. In summary, these principles and guidelines (and many similar guidelines on national level) limit the use of personal data to a minimum. Needs, attitudes etc are considered as personal when this data can be linked to demographic details. Hence, the 1-to-1 match is against the regulations of the market research industry. Survey response data and any additionally matched variables may not identify an individual.

For the purpose at hand, it is also possible to work with "pseudonymized" data. We recommend this method for our analysis purposes. Pseudonymized data means that a specific customer cannot be identified any more even if some of the individual's database variables were linked to the survey data (efamro/ESOMAR, 2017). First, the link variables are categorized before matching to the survey data (e.g. rough spend ranges instead of detailed numbers). Second, a third party does the matching, so that neither client nor analytical service provider are able to link survey data and customer database directly.

Suggested approach based on record linkage

Given the need for a more granular data match and the considerations on data protection, we suggest a new approach that consists of five main steps:

1. Matching binned database variables (for example demographics/ firmographics, usage/ spend) to the survey data using the pseudonymization approach indicated above
2. Aggregating the survey data to combinations of binned variables common across database and survey
3. Dealing with over-fit by introducing some randomness in the survey, to ensure that the method is 100% data protection proof, and re-aggregating to the combinations of binned link variables
4. Matching the modified survey data (for example needs, attitudes, NPS) using direct linkage via combinations of link variables
5. Multiple imputation of missing data in match variables – usually, there are far more combinations of binned link variables in the customer database than in the survey, in such cases it is not possible to cover all records in the database by direct matching

The linkage variables, typically demographics and usage data, are appended to the survey in a pseudonymized way. This matching ensures that the linkage variables are available in both data sets. The main idea is to match needs, attitudes or similar information collected with the survey directly to the customer database using combinations of linkage variables so that no tagging model is necessary. We keep these combinations as detailed as possible. For the match variables (needs etc.) we apply the average of the respondents of a particular combination of linkage variables to the many customers in the database with the same demographic and usage characteristics.

Combinations of binned linkage variables might have only a few respondents or even only one. In such situations, the direct match would not be statistically robust. As a remedy, we have found a way to borrow information from other respondents. We create multiple sets of our survey data, each time we change the value of the linkage variables for 20% of the data towards nearby categories. For the final matching we took the average value across all these datasets.

When using several linkage variables, each with a couple of levels, the total number of possible combinations can easily go into the thousands. Not all of them exist in the customer database and even less in the survey data. With the

described approach of information borrowing, we can reduce this discrepancy a bit. However, in most cases there will be still unmatched records in the database. We used multiple imputation methods to replace these missing values. In multiple imputation, the missing values are imputed several times, and like with bootstrapping we have taken the final value to be the average of the multiple scores. After this step, all customers in the database will have an assigned value for the match variables.

To summarize, following the outlined method it is possible to match survey variables to the customer database in a more detailed granularity than with common segmentation approaches while still adhering to strict data protection standards.

Probability matching
With combinations of link variables, it is also possible to use probability matching methods for adding survey variables to the customer database. Probability matching looks at the similarity of records related to the linkage variables. Each respondent in the survey has a specific similarity to the each record in the customer database. This similarity is expressed in form of a match probability that is used to create a weighted average for the match values. Finally, this weighted average is assigned to the specific records of the customer database.

One problem with probabilistic matching is the assumption of independence between the individual linkage variables. Second, it is conceptually hard to find the ideal distribution of match probabilities so that variation in the data is kept without the risk of running into single response based matches. Hyper-parameter optimization would be needed which can be time consuming.

Example: Retail project

We applied all three described approaches (segment tagging, record linkage and probability match) to datasets from a retail project. Below is an outline of the data used, the techniques applied and a comparison of the outcomes.

Data set used and analyses applied
A cash & carry retailer with club membership program and a clear B2B focus wanted to know more about the needs of its customers. The results provided inputs for category optimization and improved service offerings, among others.

We commissioned a market research field agency to conduct a store intercept survey with a sample of ~1200 respondents. All were club members. To collect the relative importance of needs we used maximum difference scaling. There were in total 17 needs statements: 'Lowest prices', 'Very good customer service', 'Fresh and high quality products', 'Seasonal assortment' and so on.

The field agency matched binned customer database variables to the data following the pseudonmyzation approach. This combined data allowed us to develop a latent class segmentation based on needs using firmographics and usage variables as covariates. The final solution had six segments. A good mix of needs and behaviors could explain the segments.

Finally, we developed a tagging model using the binned firmographics and usage variables to assign the segments to the customers in the database. We used decision trees for the tagging model with a high tagging accuracy of 85%. Having the segment variable in the database allowed us to run further analyses for example to inform category management efforts.

Segment tagging appends needs information to the database on a very high aggregation level. For tactical marketing objectives, more details on needs are helpful. Therefore, we also linked needs directly following the above-described approaches of direct linkage and probability matching.

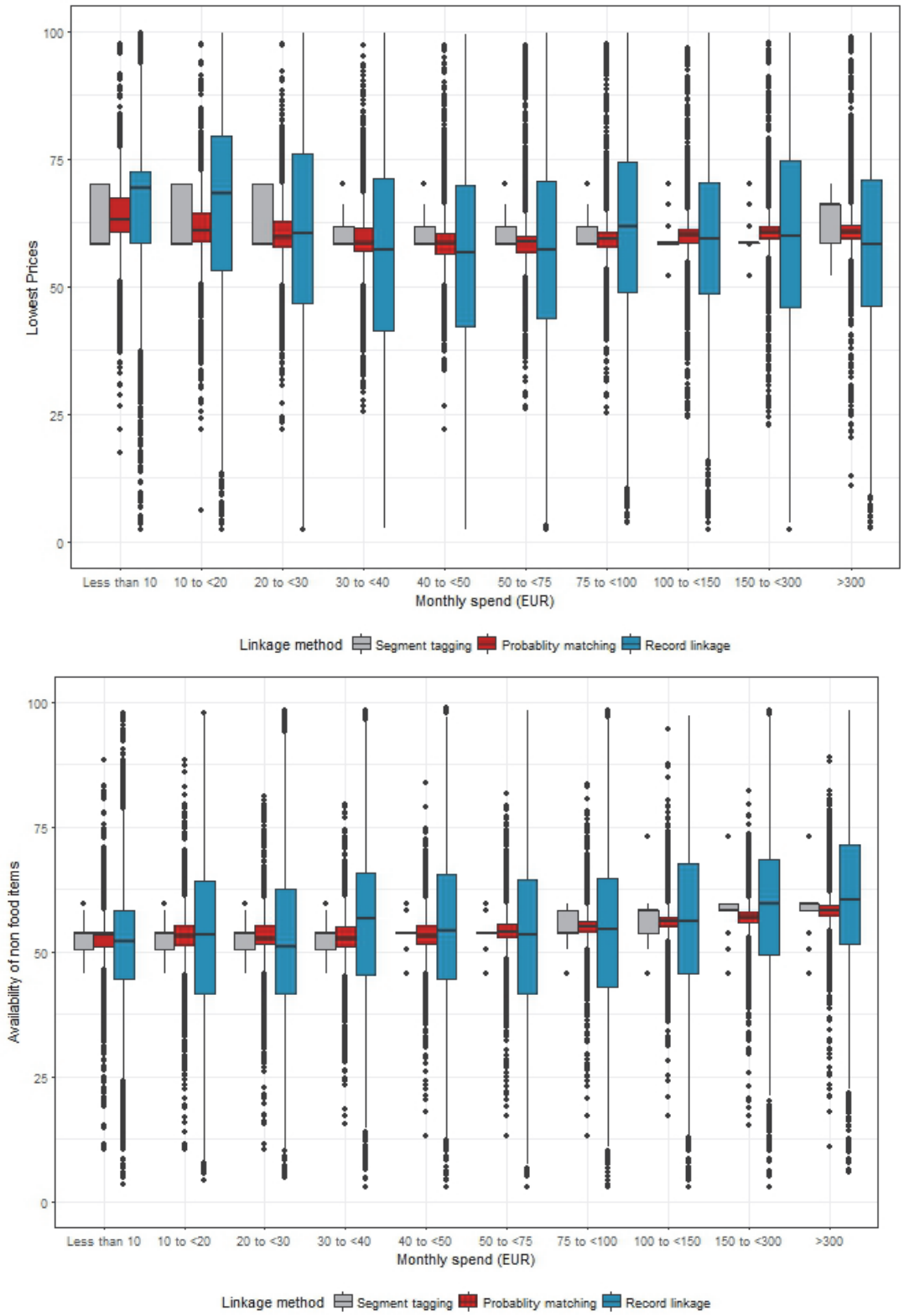
With probability matching, to create the weights we generated an Euclidean distance matrix between the binned linkage variables in the survey and CRM database. Using the distances, we investigated several weight transformations, such as inverse distance and probability from cumulative density of distance distribution. In some of the weight calculations, we set minimum distance thresholds where those records exceeding a threshold were assigned zero weight to avoid converging the needs to the mean.

For record linkage, we merged with the survey aggregated to the binned linkage variables a randomly altered survey dataset.

Comparison of results

Via the proposed linkage methods, the variance in the matched data is much higher than with traditional segment tagging. Figure 1 nicely illustrates this with box-plots, which show the distribution of two need scores in monthly spend groups for the different matching approaches we tested: segment tagging, probability matching and record linkage with multiple imputation.

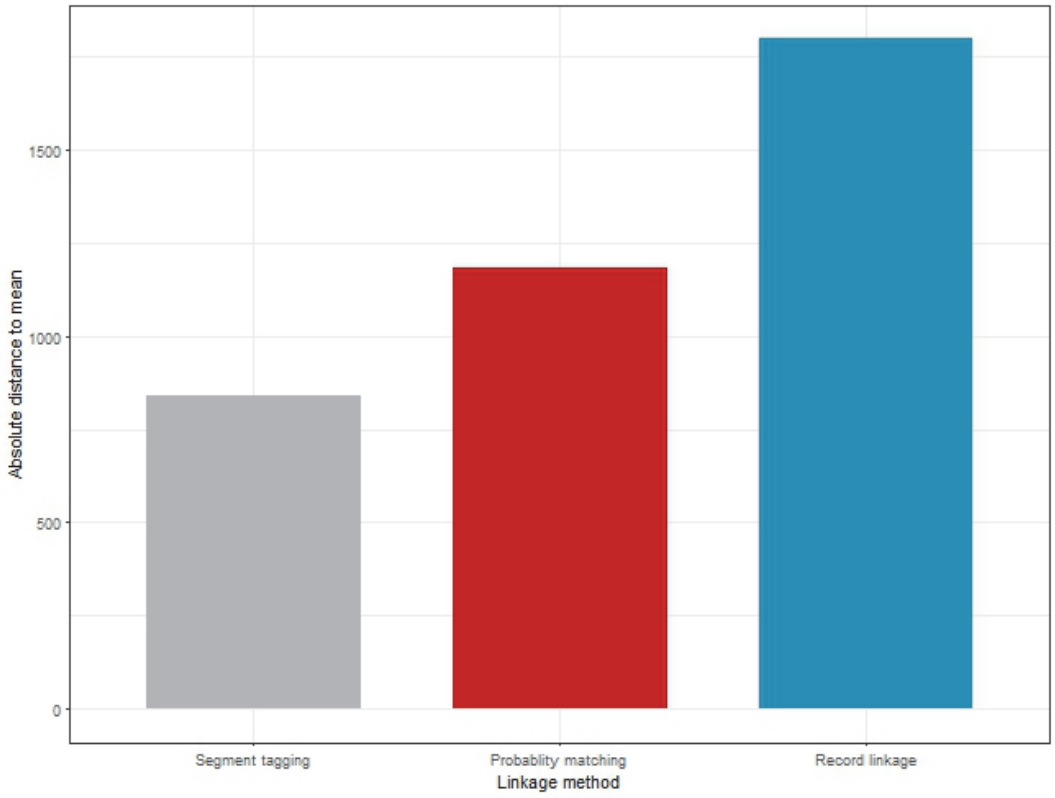
Figure 1. Distribution of needs lowest price and availability of non-food over monthly spend in the combined data, merged using segment tagging, probability matching and record linkage



More variance for matched needs in the database has clear advantages in allowing more actionable marketing initiatives. If for example a retailer wants to focus more on its most profitable customers, they can select the high-profit customers in the database and sub-select those customers who score higher on specific needs, which are relevant for the planned campaign. Generally, in our data we have found better correlations between profit and needs using the newer techniques than with segment tagging (to be noted: that profit was not used as a link variable).

Another key finding is the better discrimination of matched needs against usage variables particularly with the record linkage method. Figure 2 indicates a clearly better discrimination for record linkage compared to segment tagging, and probability matching comes out in the middle between the others. This comparison is based on the absolute differences in the needs scores between the total and the particular usage bins. A better relationship between needs and database variables is the ultima ratio of many segmentations. A stronger relationship between these two key dimensions helps with many applications of a segmentation.

Figure 2. Discrimination of needs across linkage vatiables in combined data, merged using segment tagging, probability matching and record linkage



An interesting side aspect of the record linkage approach is that it better maintains the structure of the data than the other approaches tested. Maximum difference scores of needs have very little cross-correlation. This is a general characteristic of maximum difference based scores. With record linkage also the tagged scores on the database show little cross-correlation, less so than for the other mentioned approaches.

Outlook

Our approach achieves a nuanced and more realistically varied estimation of the customers' needs than with segmentation matching, whilst maintaining the distributions and relationships across variables. This combined data gives additional opportunities to create segments with internal metrics, such as profit, that can be valuable for targeting profitable customers' needs.

With our example data, record linkage combined with multiple imputation worked best for both dimensions. Furthermore, the proposed approach for record linking with multiple imputation prohibits any direct matches. This ensures that the approach is fully in-line with the data protection regulations of the primary research industry.

While probability matching techniques have the advantage of simplicity, there is still some work to do to make them as bullet-proof in terms of data protection requirements than the described record linkage approach.

Both mentioned approaches can be still optimized via hyper-parameter optimization. In probability linkage approach, a grid search approach can used to decide on optimal distance and transformation used for weighting. With record linkage, a proportion of random noise was added to prevent over-fitting.

Starting with a direct need match towards the customer database enables us to run a segmentation directly on the database without the need for a later tagging model. These approaches also eliminate the discussion around segment tagging accuracy.

We are looking forward from the feedback of other researchers to improve this method further but also to hear from further success stories.

References

Literature

- efamro/ ESOMAR, General Data Protection Regulation (GDPR) Guidance - Note for the Research Sector, 2017 https://www.esomar.org/uploads/public/government-affairs/position-papers/EFAMRO-ESOMAR_GDPR-Guidance-Note_Legal-Choice.pdf
- Jones/ Frazier/ Murphy/ Wurst, Reverse Segmentation: An Alternative Approach, In: Proceedings of the Sawtooth Software Conference, 2006 https://www.sawtoothsoftware.com/downloadPDF.php?file=2006Proceedings.pdf
- Goldstein/ Harron, Record linkage: A missing data problem, In: Harron/ Goldstein/ Dibben, Methodological Developments in Data Linkage, 2016

R (version 3.4.1) and packages

- R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org
- Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL http://www.jstatsoft.org/v45/i03/
- Hadley Wickham, Romain Francois, Lionel Henry and Kirill Müller (2017). dplyr: A Grammar of Data Manipulation. R package version 0.7.2. https://CRAN.R-project.org/package=dplyr
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.
- Matt Dowle and Arun Srinivasan (2017). data.table: Extension of `data.frame`. R package version 1.10.4. https://CRAN.R-project.org/package=data.table

The Authors

Diane Berry is Manager Advanced Analytics Group, Bain & Company, United Kingdom

Josef Rieder is Sr. Manager Advanced Analytics Group, Bain & Company, Germany