Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 32 (2018) 253–259

Transportation Research Procedia

www.elsevier.com/locate/procedia

International Steering Committee for Transport Survey Conferences

Using inverted relative entropy to determine the representativeness of samples in mobility panels

R.G. Hoogendoorn[a], M.C. de Haas[a]*, C.E. Scheepers[a], G.M.M. Gelauff[a], S. Hoogendoorn-Lanser[a]

KiM Netherlands Institute for Transport Policy Analysis, Bezuidenhoutseweg 20, 2594AV The Hague, The Netherlands

Abstract

Travel behaviour is influenced by many factors. Changes in travel patterns over time can typically be derived from panel data. However, non-random attrition may influence the external validity of the results obtained through panel data. Consequently, it may be necessary to recruit additional respondents between waves. But how to determine how many and which respondents to recruit? In this paper we introduce a new method for determining the sample's allowed deviation from the population, as based on the concept of inverted relative entropy. We used several working examples to show how the method provides a good representation of the complexity of recruiting respondents who possess various (combinations of) characteristics. The paper concludes with a discussion and recommendations for future research.

Keywords: sample; population; representativeness; allowed deviation; inverse relative entropy

1. Introduction

Travel behaviour is influenced by many factors. Examples of such factors include life events, the use of ICT, and the attitudes of travellers. Changes in travel patterns over time, as related to these factors, can be derived from panel data. In this context, several mobility panels have been developed, including the German Mobility Panel (MOP) (Kunhimhof et al., 2006), the Longitudinal Mobility Survey (LVO) (Meurs et al., 1989), and the Netherlands Mobility Panel (MPN) (Hoogendoorn – Lanser et al., 2015).

* Corresponding author. Tel.: +31-6-112-999-20.
E-mail address: mathijs.de.haas@minienm.nl

10.1016/j.trpro.2018.10.047

The MPN is a state-of-the-art household mobility panel designed to establish the short- and long-term dynamics in the travel behaviour of households and individuals (Hoogendoorn – Lanser et al., 2015). As of 2013, members of more than 2,000 households annually recorded their travel behaviour using a three-day location based diary, providing information about all their trips (transport modes, trip purpose, delays, etc.). A screening questionnaire was used to assess their willingness to participate in the panel. In the MPN, the so-called 'Gold Standard' (MOAWeb, 2015) is applied to determine the extent to which the sample is representative of the Dutch population. The Gold Standard is comprised of personal characteristics, such as - gender, age, education and work situation - and household characteristics, including the household situation, household size, and level of urbanisation. These variables are used to assess the representativeness of the MPN sample.

Attrition may influence the representativeness of the net sample. Attrition is nearly always non-random, which means that the respondents who drop out of the sample between two consecutive waves are systematically different from the respondents who remain in the sample. Research has shown, for example, that attrition is strongly related to income, education, work situation, and transport mode use (Kitamura & Bovy, 1987; Pendyala et al., 1993; Brownstone & Chu, 1997). Attrition may therefore not only lead to a deviation from the Gold Standard, but also to a reduction in the (external) validity of results obtained via the panel data if the attrition is related to the variables of interest; the mobility of respondents.

Due to attrition, additional respondents must be recruited between waves in order to ensure the sample's longitudinal character and representativeness. This raises the question how many and which respondents should be recruited? This is further complicated by the fact that not all respondents are equally willing to participate in a panel (e.g. youngsters and people with a low education). To account for this differences in willingness to participate, some groups should be oversampled in the recruitment phase when it is expected that their willingness to participate to the panel is lower. In the MPN, the expected willingness to participate is calculated based on earlier recruitment experiences.

An additional complexity in the MPN is that the sample must be representative in terms of both personal characteristics (gender, age, etc.) and household characteristics (household situation, household size, etc.). Moreover, recruiting respondents who possess combinations of characteristics, while also varying in their willingness to participate, is extremely challenging. When the complexity of recruiting such respondents is taken into account, by how much can the net sample be allowed to deviate from the population ex post? A commonly used method to account for attrition is adding refreshment samples to the panel (Deng et al., 2013). A refreshment sample consists of new randomly (stratified) drawn respondents who are added to the panel at a subsequent wave. Another method to account for attrition is to use a rotation scheme in which respondents are asked to participate in a panel for a fixed period of time. Germany's MOP, for instance, uses a rotation scheme to limit the impact of attrition on the results' validity (Kunhimhof et al., 2006). In the MOP respondents are recruited to participate three waves, after which they naturally drop out of the panel. After the initial recruitment, no new respondents are added to the cohort, removing the need to deal with non-random attrition between waves. Certain groups, such as households with children, are, however, oversampled to account for expected attrition. Although this eases the refreshment of the panel, using such schemes places limitations on the longitudinal character of the panel.

This study's research objective is to introduce a method for determining the degree to which the net sample is allowed to deviate from the population while taking the complexity of achieving representativeness on certain characteristics into account. As far as the authors are aware, such a method is currently not available. An often used method to assess the representativeness of a sample is by performing Chi-Square tests. This does, however, not allow to account for differences in complexity between characteristics. Traditionally, with probability sampling, the fact that sampling respondents on certain characteristics is more complex than on other characteristics is not taken into account. Insights into the proposed method's applicability allow for a more accurate determination of the required gross sample size. Moreover, empirically underpinned insights into the net sample's allowed deviation can be used to correct for non-random attrition. It should be noted that this paper does not aim to improve weighting procedures for panel data. The aim with the proposed method is to account for attrition and initial non-response in the recruitment phase in order to obtain a more representative net sample. Hereby the need to calculate weights is minimized.

To determine the allowed deviation, we introduced a new method that uses the concept of inverted relative entropy, as based on Shannon's information theory (Shannon, 2001). Shannon's entropy is deemed a suitable concept for determining the allowed deviation from the population, as it provides us with information pertaining to the complexity of recruiting respondents who possess (a combination of) certain characteristics.

In the following section we provide a brief state-of-the-art regarding the concept of entropy. Moreover, we provide an extensive mathematical formulation of the proposed method and several working examples. The paper concludes with a discussion and recommendations for future research.

2. A closer look at entropy

Entropy is a measure of the unpredictability of a situation, and therefore of its average information content. Take for example a coin toss: under the assumption that the probability of the toss resulting in heads is the same as the toss resulting in tails, one can easily imagine that it is impossible to accurately predict the outcome of the toss. In terms of Shannon's entropy (Shannon, 2001), the entropy is at its maximum and therefore 1. If however a trick-coin is used (it has two tails, for example), it thus becomes very easy to predict the outcome. The entropy in that case is at its minimum and therefore 0.

Shannon defined the entropy H of a discrete variable χ with possible values $\{x_1, \ldots, x_n\}$ and probability mass function $P(\chi)$ as:

$$H(\chi) = E[I(X)] = E[-\ln(P(X))] \tag{1}$$

In this equation, E is the expected value operator, while I is the information content of X. We can rewrite this equation as:

$$H(\chi) = \sum_{i=1}^{n} P(x_i) I(x_i) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i) \tag{2}$$

Here, b is the base of the logarithm. Normally, with information entropy, b is set to *2*. With a base of 2, entropy is measured in bits. If $P(x_i)=0$ for i, the value of the corresponding summand $0 \log_b(0)$ is 0, which is consistent with the limit:

$$\lim_{p \to 0+} p \log(p) = 0 \tag{3}$$

Above, we used the example of a coin toss. A coin toss can be modelled as a Bernouilli process. We have seen that entropy is at its maximum when $Pr(X=1)=0.5$. This is illustrated in Figure 1:

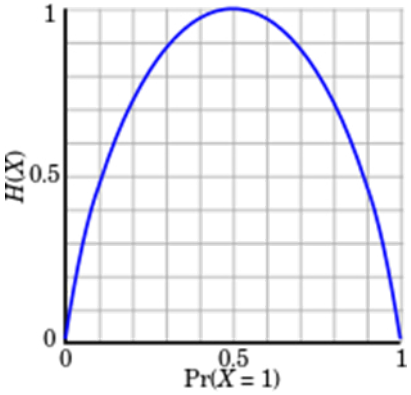

Figure 1. Entropy as a function of the level of uncertainty of an outcome.

3. Method formulation

The mathematical formulation of entropy presented above expresses that for each probability of a category, the probability must be multiplied by the binary logarithm and summed for each category. This means that the maximum entropy can be calculated by extracting this binary logarithm from the number of categories. For example, allow X to be uniformly distributed, and let us assume that |X|=n. In this case $\forall x \in X: P(X = x) = \frac{1}{n}$ and

$H(\chi) = log_2(n)$.

In order to use Shannon's entropy (Shannon, 2001) for determining the sample's allowed deviation, it is important to understand that we must use the inverse of Shannon's entropy. In other words: one would want a lower allowed deviation in the sample when entropy is high (in case of a (near) uniform distribution), while a higher deviation is allowed when entropy is low (in case of a non-uniform distribution).

For the purpose of using entropy as a measure, it is convenient that $0<H(\chi)<1$. However, Shannon's entropy increases when the number of categories increases. For example, if there are 4 categories, Shannon's entropy is $\log_2(4) = 2$. In this study we therefore use the relative entropy to arrive at an entropy between 0 and 1. Let us suppose that we calculate this relative entropy for a stochastic variable $\chi: X \to \mathbb{R}$, in which $|X| = n$:

$$H_{rel}(\chi) = \frac{H(\chi)}{H_{max}(\chi)} = \frac{-\sum_{x \in X} P_x(X = x) \log_2 2\,(P_x(X = x))}{\log_2(|X|)} \tag{4}$$

In this equation H_{max} is the maximum entropy. The relative entropy H_{rel} is an indication of how large the level of entropy of χ is. Next, we take the inverse of this relative entropy:

$$H_{inv}(\chi) = \frac{1}{H_{rel}(\chi)} \tag{5}$$

If the distribution is not (near) uniform, then $H_{inv}(\chi)$ is larger than when χ has a uniform distribution. By using inverted relative entropy, it is possible to determine the allowed deviation per stochastic variable. To do this, we introduce an upper and lower bound of the sample's allowed deviation, represented by T_{max} and T_{min}. These bounds are arbitrarily chosen. We use the following expressions to calculate the allowed deviation from the population:

$$T_c = \frac{T_{max} - T_{min}}{\max\,(H_{inv}(\chi)) - \min\,(H_{inv}(\chi))} \tag{6}$$

and

$$T_{cp} = T_{min} + (T_c(H_{inv}(\chi) - \min\,(H_{inv}(\chi)) \tag{7}$$

In the first equation, we take the difference between the upper bound T_{max} and the lower bound T_{min} of the allowed deviation from the population. In the working examples presented below, we use 20% as an upper bound and 15% as a lower bound. These bounds were chosen as they are deemed to be achievable based on the current representativeness of the MPN sample. When the sample's composition gets closer to the population's composition, the range and the limits can be lowered. Next, we divide this difference by the difference between the maximum and the minimum relative inverted entropy. Therefore, T_c reflects the extra allowed deviation from the population per unit of inverted relative entropy. In the last equation we calculate the allowed deviation per variable (T_{cp}). To do so, we add a multiplication of T_c and the difference between the relative inverted entropy of the class minus the minimum relative inverted entropy to the lower bound of the sample's allowed deviation from the population.

4. Working examples

As stated before, the representativeness of the sample in the MPN is assessed based on the Gold Standard. The Gold Standard reflects the composition of the Dutch population on several personal- and household characteristics (MOAWeb, 2015). To assess the representativeness of the MPN we use data from the Gold Standard about gender, age, educational level, working situation, household situation, household size and level of urbanization. Besides assessing representativeness on these single variables, representativeness is also assessed based on the variables in conjunction (pairs) with each other. Including combinations with more than two variables is possible for some of the used variables (e.g. gender with age and work situation) as the Gold Standard provides their distribution in the Dutch population. However, drawing a sample on a combination of three or more variables is not possible for the field work agency of the MPN. Therefore, in the proposed method, variables are only included independently and in pairs of two. The working examples below illustrate how the entropy of different variables is calculated.

To calculate the allowed deviation per variable and combination of variables, T_c should be calculated as this is a constant. To calculate T_c, the minimum and maximum of $H_{inv}(\chi)$ should be calculated first. These minimum and maximum values can change yearly, as the population's composition changes. Based on the Gold Standard it was

found that the minimum and maximum values for $H_{inv}(\chi)$ are found for gender (1.0001120) and household situation in conjunction with household size (1.5756778) respectively. As stated before, the lower and upper bound of maximum allowed deviation are set at 15% and 20% respectively. When the sample's composition gets closer to the population's composition the bounds can be lowered. It is also possible to use another range of bounds. If it turns out that the deviation on certain variables is well below the lower bound, but the deviation on other variables is close or over the upper bound, it can be decided to only lower the lower bound and vice versa, resulting, for instance, in a range from 10% to 20% or 15% to 25%. In this context, we can calculate the extra allowed deviation per unit of inverse relative entropy as follows:

$$T_c = \frac{.20 - .15}{(1.5756778 - 1.0001120)} = .0868710 \quad (8)$$

Following the same analogy, if the range of bounds would be set at 15% to 25%, T_c would amount to 0.173742. This would result in larger differences between allowed deviations of the different variables and combinations of variables.

4.1. Gender

The first working example pertains to the variable gender. According to the Gold Standard, the fraction of men in the Dutch population of 12 years and older amounts to .4937698, while the fraction of women amounts to .5062302. We start by calculating the entropy per class. For men this means:

$$H(\chi_1) = -(.4937698 * \log_2(.4937698)) = .5027019 \quad (9)$$

and for women:

$$H(\chi_2) = -(.5062302 * \log_2(.5062302)) = .4971861 \quad (10)$$

Next, we calculate the relative entropy:

$$H_{rel}(\chi) = \frac{.5027019 + .4971861}{\log_2(2)} = .9998880 \quad (11)$$

And the inverse relative entropy:

$$H_{inv}(\chi) = \frac{1}{.9998880} = 1.0001120 \quad (12)$$

using T_c, the allowed deviation for gender can be calculated:

$$T_{cp} = .15 + .0868710 * (1.0001120 - 1.0001120) = .15 \quad (13)$$

As can be observed from these calculations, the allowed deviation from the sample is equal to the lower bound of 15%. In other words: recruitment based on gender is relatively low in complexity, and this is primarily due to the fact that only two classes are present which are almost equal in size.

4.2. Work situation

In the MPN screening questionnaire, the work situation is divided into five different classes: working, housekeeping, student, unemployed/disabled, and retired. Again, we start by calculating the entropy per class. The results obtained in a similar manner for gender are displayed in the table below.

Table 1. Fractions of the population and entropy values for the variable work situation

Class	*Fraction*	*Entropy*
Working	.5028349	.4987334
Housekeeping	.0766462	.2840234
Student	.1306658	.3836409
Unemployed / disabled	.0842309	.3006628

Retired	.2056221	.4692158
Sum	1.0000000	1.9362764

Using the equations previously illustrated, we calculate that the relative inverse entropy amounts to 1.1991718. The inverse relative entropy is slightly higher than it was for gender. Next, we calculated the allowed deviation of the sample, which amounts to .167292532. From these calculations it follows that a deviation of 16.73% from the Dutch population is allowed for work situation, which is slightly higher than the allowed deviation for gender. In other words, due to the fact that it is somewhat more complex to recruit a representative sample based on work situation than on gender, the requirements for representativeness on work situation are somewhat relaxed and therefore the allowed deviation is higher.

4.3. Education x work situation

The final working example pertains to the variable education, in conjunction with the variable work situation. In the previous example, we already presented the various classes for work situation. For education, three different classes are distinguished in the MPN, namely: low, medium and high education. In this context, the combination of education and work situation consists of 15 classes. Table 2 shows the fractions of the population and the estimated entropy.

Table 2. Fractions of the population and entropy values for the variable education x work situation

Class		*Fraction*	*Entropy*
Low education	Working	.1150208	.3588687
	Housekeeping	.0308630	.1548698
	Student	.0846196	.3014881
	Unemployed / disabled	.0377479	.1784516
	Retired	.1029023	.3375867
Medium education	Working	.2242902	.4836956
	Housekeeping	.0304502	.1533899
	Student	.0426177	.1940128
	Unemployed / disabled	.0364395	.1741206
	Retired	.0659940	.2587970
High education	Working	.1628091	.4263558
	Housekeeping	.0122444	.0777734
	Student	.0067030	.0484020
	Unemployed / disabled	.0112229	.0726954
	Retired	.0360756	.1729044
Sum		1.0000000	3.3934117

Using the equation shown above, we calculated that the relative inverse entropy amounts to 1.1513164. The inverse relative entropy is higher than for gender, but lower than for work situation alone. The allowed deviation from the sample is:

$$T_{cp} = .15 + .0868710 * (1.1513164 - 1.0001120)) = .1631353 \quad (14)$$

We may conclude from these calculations that when education and work situation are combined, the allowed deviation from the population is 16.31%, as based on the chosen lower and upper bounds. This percentage is higher than gender, but slightly lower than for work situation alone. The sample's calculated allowed deviation from the population clearly proves the applicability of the method, in the sense that it provides a good representation of the complexity involved in recruiting respondents. We see that there is a low allowed deviation for gender, and a higher allowed deviation for work situation and for the combination education x work situation.

5. Discussion

Travel behaviour is influenced by many factors. Changes in travel behaviour over time can be derived from panel data. In order to attain a satisfactory level of external validity, the panel must be representative of the population.

Attrition can influence the representativeness of the sample, and because attrition is almost always non-random, this can lead to a bias in the data. Due to attrition, additional respondents must therefore be recruited between waves of a panel. But how many and which respondents must be recruited?

This study's research objective was to introduce a new method for determining the net sample's allowed deviation from the population. Insights into the applicability of the sample allow for a more accurate determination of the required gross sample size, as well as for empirically grounded insights into the net sample's allowed deviation from the population.

The proposed method was based on the concept of relative inverse entropy. We used several working examples to show how the proposed method works. Moreover, we showed that the method provides a good representation of the complexity of recruitment. We revealed for example that a lower deviation from the population was allowed for gender than for work situation. In turn, work situation had a higher allowed deviation than the combination education and work situation. Apparently, from the three presented examples, gender has, relatively speaking, the most uniform distribution, followed by the combination of education and work situation. Work situation has, relatively, the least uniform distribution.

In this paper, we did not endeavour to compare the proposed method to more traditional methods of probability sampling in which the differences in complexities are not taken into account; hence, we recommend that future research will be conducted to determine the added value of the proposed method as compared to the more traditional methods. Further, the proposed method uses an upper and a lower bound for the allowed deviation. The working examples used bounds of 15 and 20%, but these bounds were arbitrarily chosen. We therefore recommend that future research be conducted to determine the optimal upper and lower bounds.

Acknowledgements

The authors wish to thank Philip Michgelsen, MSc., for his enthusiastic contribution to the contents of this paper.

References

Deng, Y., Hillygus, D. S., Reiter, J. P., Si, Y., & Zheng, S. (2013). Handling attrition in longitudinal studies: The case for refreshment samples. Statistical Science, 28(2), 238-256.

Hoogendoorn-Lanser, S., Schaap, N. T., & OldeKalter, M. J. (2015). The Netherlands Mobility Panel: An innovative design approach for web-based longitudinal travel data collection. Transportation Research Procedia, 11, 311-329.

Kitamura, R., & Bovy, P. H. (1987). Analysis of attrition biases and trip reporting errors for panel data. Transportation Research Part A: General, 21(4-5), 287-302.

Kuhnimhof, T., B. Chlond, and D. Zumkeller, Nonresponse, selectivity, and data quality in travel surveys: Experiences from analyzing recruitment for the German mobility panel. Transportation Research Record: Journal of the Transportation Research Board, No. 1972, 2006, pp. 29–37.

Meurs, H., L. Van Wissen, and J. Visser, Measurement biases in panel data. Transportation, Vol. 16, No. 2, 1989, pp. 175–194.

MOAWeb, Gold Standard: A Unique Calibration Tool for National and Regional Samples (https://www.moaweb.nl/services/services/gouden-standaard.html), 2016.

Pendyala, R. M., Goulias, K. G., Kitamura, R., & Murakami, E. (1993). Development of weights for a choice-based panel survey sample with attrition. Transportation Research Part A: Policy and Practice, 27(6), 477-492.

Shannon, C. E. (2001). A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review, 5(1), 3-55.