ESTIMATION METHODS FOR THE INTEGRATION OF ADMINISTRATIVE SOURCES – DELIVERABLE D4

Template T4-2: Estimating classification errors under edit-restrictions in combined register-survey data

Responsible person at Commission: Fabrice Gras, Eurostat – Unit B1

Written by: Nicoletta Cibella, Ton de Waal, Marco Di Zio, Mauro Scanu, Sander Scholtus, Arnout van Delden, Tiziana Tuoto, Li-Chun Zhang
FOREWORD

Task 4: Literature review presenting one actual example in the NSIs for each of the type of use of administrative sources and for the steps that have been previously identified

This activity aims at providing examples of actual usages of statistical estimation methods when using administrative sources based on NSIs experiences. The examples are referred to the usages and steps identified in Tasks 1 and 2.

For each identified example is provided an executive summary presenting the method, the contextual framework and the main results.

ISTAT experts: Marco Di Zio, Nicoletta Cibella, Mauro Scanu, Tiziana Tuoto

CBS experts: Ton de Waal, Arnout van Delden and Sander Scholtus

with help of Li-Chun Zhang

Deliverable 4. An in-depth inventory of the data needs of EUROMOD for each country. Specific contract n°000052 ESTAT N°11111.2013.001-2016.038 under framework contract Lot 1 n°11111.2013.001-2013.252
## Title of the case study:
**Estimating classification errors under edit-restrictions in combined register-survey data**

<table>
<thead>
<tr>
<th>Executive summary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Usages (Deliverable 1)</strong></td>
</tr>
<tr>
<td><strong>Statistical Tasks (Deliverable 2)</strong></td>
</tr>
<tr>
<td><strong>Data configurations (Deliverable 1)</strong></td>
</tr>
<tr>
<td><strong>Methods (Deliverable 5)</strong></td>
</tr>
</tbody>
</table>

## Presentation of the Case study

<table>
<thead>
<tr>
<th>Agency – country</th>
<th>Statistics Netherlands</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic</td>
<td>Classification errors in admin and survey variables on home ownership</td>
</tr>
</tbody>
</table>

### Description of the problem

Statistics Netherlands has access to a registration of addresses and buildings (BAG). This registration can be linked to the Dutch population register and the information in the BAG can be used to derive whether a person owns or rents a home. The variable home ownership is useful for many analyses in social research. These analyses may be affected by classification errors in the BAG, i.e., cases where a person who in reality rents a home is classified as owning a home in the BAG, or vice versa.

To investigate the size of the classification error problem in the BAG, Statistics Netherlands has linked the BAG data to survey data from the so-called LISS panel (Longitudinal Internet Studies for the Social sciences). The LISS panel is a web panel, administered by CentERdata, based on a random sample from the Dutch population register. Respondents in the LISS panel were asked whether they own or rent a home. Thus, by using the overlapping units in the BAG and LISS data, two observed variables that measure home ownership were obtained for a sample of the population.

The data were analysed using a latent class model to account for classification errors. In addition, the LISS survey contained questions on marital status and whether the respondent received rent benefit. These were included as covariates in the model, to aid the estimation of classification errors. The fact that only people who rent a home can receive rent benefit implies an edit rule for the data. It is useful to incorporate this edit rule into the model, as it helps to identify some of the classification errors.

### Input data

Administrative data (BAG) on home ownership, survey data (LISS panel) on home ownership, marital status and rent benefits. All data refer to the year 2013.
Expected output

Estimated classification error probabilities for home ownership in both data sources, as well as a multiply-imputed data set with estimated true home ownership status values.

Technical summary

Boeschoten et al. (2016) used a latent class model to analyse the combined register-survey data. In this model, the true classification of each unit with respect to the variable of interest is denoted by a latent (unobserved) variable. The model describes, for each unit, the probability that it is classified in a particular category according to each observed variable, given its true category according to the latent variable. Thus, the model describes the probability of correct classification as well as the probability of each possible classification error. A more detailed introduction to latent class models is given in the module “Variable harmonisation based on latent variable models” in Deliverable 5.

Let $X$ denote the true home ownership status of a person, with two possible categories: “own” and “rent”. This variable is unobserved (latent). The observed home ownership status in the BAG and LISS panel is denoted by $Y_1$ and $Y_2$, respectively. Finally, the covariates are denoted by $Q$ (which may be a vector).

In the latent class model, the marginal probability of observing a particular response pattern $P(Y_1 = y_1, Y_2 = y_2)$ is modelled as the sum of the joint probabilities $P(Y_1 = y_1, Y_2 = y_2, X = x) = P(X = x)P(Y_1 = y_1, Y_2 = y_2 | X = x)$ over all latent classes $x \in \{\text{own}, \text{rent}\}$. Furthermore, two assumptions are made:

- The classification errors in different observed variables are independent, given the score on the true variable (latent class).
- The classification errors in the observed variables are independent of the covariates, given the score on the true variable.

This leads to the following model:

$$P(Y_1 = y_1, Y_2 = y_2 | Q = q) = \sum_{x \in \{\text{own}, \text{rent}\}} P(X = x | Q = q) \prod_{l=1}^{2} P(Y_l = y_l | X = x).$$

Here, $P(X = x | Q = q)$ denotes the probability that an individual has true home ownership status $x$, given his/her scores on the covariates; also, $P(Y_l = y_l | X = x)$ denotes the probability of observing a home ownership status $y_l$ on the $l$-th observed variable, given that the true home ownership status is $x$. Thus, the classification error probabilities for the BAG under this model are given by $P(Y_1 = \text{rent} | X = \text{own})$ and $P(Y_1 = \text{own} | X = \text{rent})$, and similarly for LISS.

As mentioned above, people who own their home cannot receive rent benefit. Therefore, if the rent benefit indicator is included as one of the covariates in the model, say $Q_1$, then it must hold that

$$P(X = \text{own} | Q_1 = \text{receives rent benefit}) = 0.$$
From the estimated model, posterior membership probabilities for each latent class $x \in \{\text{own}, \text{rent}\}$, given the observed values of $Y_1$, $Y_2$, and $Q$, can be obtained by applying Bayes’ rule. Boeschoten et al. (2016) propose to use these posterior probabilities to impute an estimated true home ownership status for each person.

To reflect the uncertainty in these estimated true values, they apply a multiple imputation on the basis of the estimated posterior probabilities. This leads to a data set with, in addition to the original observed variables, $M$ imputed variables with predicted true home ownership status values. These imputed variables can be used in further statistical analyses to correct for classification errors; see Boeschoten et al. (2016) for more details.

After combining the BAG registration with the LISS panel data, Boeschoten et al. (2016) were left with 3011 individuals for which both $Y_1$ and $Y_2$ were available. The model was fitted to this combined data set. Values of marital status were also available from the LISS panel for all these persons. The indicator whether a person receives rent benefit was available from the LISS panel for only 779 individuals, due to the routing in the questionnaire. Missing values on this covariate where taken into account by applying Full Information Maximum Likelihood to estimate the model. Estimation of the model was done using the Latent Gold software (Vermunt and Magidson, 2013).

We discuss the results for two different latent class models. Both models included marital status and rent benefit as covariates. In the first model, the above restriction that people who receive rent benefit cannot own a home was not explicitly included in the model (unrestricted model). In the second model, this restriction was included (restricted model).

The following table shows the estimated classification error probabilities in the two sources, for both models. It is seen that, according to the model, the two sources are not error-free. However, the proportion of misclassified persons appears to be relatively small in both sources. In other words, the quality of measurement is good. In the LISS panel, errors where persons who rent a home classify themselves as home owners appear to be more frequent than misclassifications in the reverse direction. In the BAG, both types of misclassification occur about equally often according to the restricted model.

<table>
<thead>
<tr>
<th>Data source</th>
<th>Probability</th>
<th>Unrestricted</th>
<th>Restricted</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAG</td>
<td>$P(Y_1 = \text{rent}</td>
<td>X = \text{own})$</td>
<td>0.0251</td>
</tr>
<tr>
<td></td>
<td>$P(Y_1 = \text{own}</td>
<td>X = \text{rent})$</td>
<td>0.0500</td>
</tr>
<tr>
<td>LISS</td>
<td>$P(Y_2 = \text{rent}</td>
<td>X = \text{own})$</td>
<td>0.0003</td>
</tr>
<tr>
<td></td>
<td>$P(Y_2 = \text{own}</td>
<td>X = \text{rent})$</td>
<td>0.1062</td>
</tr>
</tbody>
</table>

The next table shows the estimated proportion of home owners in the population $P(X=\text{own})$ according to both sources separately and according to
both latent class models. The model-based estimates were obtained by applying multiple imputation as described above and pooling the results.

<table>
<thead>
<tr>
<th></th>
<th>Point estimate</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed in BAG</td>
<td>0.6450</td>
<td>[0.6448; 0.6451]</td>
</tr>
<tr>
<td>Observed in LISS</td>
<td>0.6830</td>
<td>[0.6829; 0.6832]</td>
</tr>
<tr>
<td>Unrestricted model</td>
<td>0.6505</td>
<td>[0.6503; 0.6506]</td>
</tr>
<tr>
<td>Restricted model</td>
<td>0.6591</td>
<td>[0.6590; 0.6593]</td>
</tr>
</tbody>
</table>

The original estimated proportions based on a single data source differ significantly. With both models, the estimated proportion of home owners after correction for misclassification lies closer to the BAG estimate than to the LISS estimate. See Boeschoten et al. (2016) for more results, including an illustration of the use of the imputed data set in a logistic regression of home ownership on marital status.

Main findings (lessons learnt)

### Advantages
- The latent class model can be used to correct for classification errors in observed variables, by estimating and imputing a harmonised variable. The uncertainty in this harmonised variable can be taken into account through multiple imputation.
- The latent class model can account for classification errors in all observed variables (both administrative and survey variables). It is not necessary to take one source as the ‘gold standard’ a priori.
- The method can also take edit rules into account in the form of restrictions on the model parameters.

### Disadvantages
- The model is based on two strong assumptions: that classification errors are independent across observed variables and do not depend on covariates, conditional on the latent (true) class. In practice, these assumptions may not always hold and then the model-based results may be biased.
- Classification errors in the covariates are not explicitly accounted for.

### Gap analysis
- To make the method more suitable for use in practice, the above disadvantages could be addressed.
- Estimated relations of the latent class variable to covariates that are not included in the model may be biased. It is not always possible or desirable to include all potentially relevant covariates in the model. It would be good to have a method that can correct estimated relations between the latent variable and covariates not included in the latent class model.

### Other remarks
/