Problem of Missing Data in Census - Who Are the Non-Response Respondents?

Jan Hora [1], Partially supported by the project GAČR No. 102/07/1594 of Czech Grant Agency and by the projects 2C06019 ZIMOLEZ and MŠMT 1M0572 DAR.

1 Introduction

In the modern networked society there is an increasing demand on dissemination and sharing of statistical information. To meet the expectations of users the statistical agencies release two major forms of statistical data: the traditional tabular data and the sets of individual respondent records called microdata. The advantage of releasing microdata instead of specific pre-computed tables and statistics is the increased flexibility and availability of information for the users. With appropriate microdata the users may examine unusual hypotheses and find new issues beyond the usual scope of data providers.

In any case the fundamental obligation of data providers is to protect the privacy of respondents. For this reason the explicit identifiers such as names, addresses and phone numbers are commonly removed. However, anonymous respondents may by re-identified by combining other data such as birth date, sex, ZIP code which uniquely pertain to specific individuals. Different statistical disclosure control (SDC) methods have been proposed to protect the confidentiality of data. With tabular data a disclosure can occur if a cell corresponds to a very small group of respondents. This problem can be disabled by suppressing cells, aggregating values, removing sensitive variables or by other techniques. In case of microdata the easily identifiable quantitative variables may be transformed to discrete intervals and sensitive qualitative variables may be combined to produce more general categories. Rare data can be suppressed, swapped, modified or simulated. Obviously, disclosure limitation procedures are connected with some information loss. There is a trade-off between disclosure protection and the accuracy of data.

[1]Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic

There is an extensive literature on SDC techniques (for the most frequent references see, e.g., Dalenius 1977, Bethlehem et al. 1990, Winkler 1998, Fienberg et al. 1998, Willenborg and de Waal 2001) available but the underlying problem is still to be considered open. Let us recall that a disclosure may be inferential (Duncan and Lambert 1989) without any actual re-identification of a record and, in a special context, even a modified erroneous value may harm the re-identified respondent. To produce safe and analytically valid public-use files much extra work is needed and, simultaneously, there is always a rest of disclosure risk. Thus, many statistical agencies release microdata for research purposes only, usually under special licence agreements and through secure data archives. In general the non-disclosure policy becomes a serious limitation of information dissemination.

In the last years we have developed an alternative approach to presentation of survey results based on interactive statistical models (Grim 1992, Grim and Boček 1996, Grim et al. 2001, Grim et al. 2004). We estimate the joint probability distribution of the original microdata in the form of a multivariate distribution mixture with product components using the EM algorithm (Dempster et al. 1977). The estimated product mixture can be used directly as a knowledge base of the probabilistic expert system PES (Grim 1990, Grim 1994) and, in this way, we can derive the statistical information from the mixture model without any further access to the original database. The statistical model provides flexibility and comfort of information access which is comparable or even better than in the case of microdata subsets. The resulting software product does not contain original microdata and therefore it can be freely distributed without any limitation. The interactive model-based re-identification of respondents is disabled by the decreasing accuracy of the estimated distribution mixture at low probability levels. The balance between the accuracy of useful statistical information and protection of anonymity of rare data is automatically controlled by the underlying maximum-likelihood criterion. According to our best knowledge, in recent literature there are no similar approaches proposed by other authors.

In this paper we describe the application of the proposed method to the individual microdata records from the Czech Census in 2001. The statistical model has been computed in the framework of a special cooperation project between the Czech Statistical Office, Prague University of Economics and the Institute of Information Theory and Automation. The aim of the project is to verify the applicability of the interactive statistical model to the

next Czech Census in 2011. To illustrate a general possibility of information fusion from different sources we have combined two originally separately treated databases of persons and households. In particular, for every of the responding persons we have combined ten variables from the database of individuals with fourteen variables from the corresponding household.

The resulting source database contained 10 230 060 records, with about 1.5 millions incomplete records including nearly three millions of non-response (missing) values. As the primary purpose of the project has been to demonstrate the accuracy of the method in case of ideal complete data, we decided first to estimate the model parameters from the incomplete records and then to use the resulting distribution mixture to estimate and substitute missing values. The final statistical model has been computed from the set of complete microdata. The accuracy of the final model has been verified by comparing the model probabilities with the relative frequencies of all statistically relevant combinations of responses. We have found that the accuracy of model probabilities is comparable with that of the relative frequencies computed from a randomly chosen 1 million subset of the original microdata (without anonymization). The preliminary version of the final interactive software product is to be offered free at *http://ro.utia.cas.cz/dem.html*.

The paper is organized as follows: In Section 2 we describe the choice of variables for the statistical model, the EM algorithm and its properties. Section 3 deals with the problem of missing data and in Section 4 we evaluate the accuracy of the estimated mixture. In the concluding section we summarize advantages and different application aspects of the proposed method.

2 Statistical Model of Census Data

The primary purpose of the considered statistical model is to reproduce the statistical relationship of a set of discrete variables as exactly as possible. The number of variables and number of their values should be kept in reasonable bounds because of the well known trade-off between the complexity of the estimated probability distribution and its accuracy. For the sake of estimating the statistical model of the Czech Census 2001 we have chosen 24 categorial variables (questions) as listed in Table 1. In order to decrease the formal complexity of the model we have applied less detailed coding of some variables (regional localization,

age intervals). Simultaneously we have omitted too unambiguous variables which are less informative and unproductive in combination with other variables. To illustrate a general possibility of information fusion from different sources we have combined two originally separate databases of individuals and households. In particular, the first ten variables from the database of individuals have been merged with fourteen variables of the corresponding household. Note that in the resulting database the household-related response frequency has a different meaning, namely the number of respondents living in such households. Thus, instead of the properties of flats, we may analyze the housing conditions of respondents.

For every respondent we have a record of 24 variables. The third column in Table 1 contains the number of possible responses for the respective questions and the fourth column contains the frequency of missing values in percent. The total number of non-response is 2933427. The uncertainty of variables expressed in percent of maximum Shannon entropy is given in the last column.

Formally, we consider the source database to be a set of independent and identically distributed observations of a random vector of 24 discrete finite valued random variables:

$$\boldsymbol{v} = (v_1, v_2, \ldots, v_{24}) \in \mathcal{X}, \quad \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_{24}. \tag{1}$$

We assume, that the unknown multivariate discrete probability distribution $P^*(\boldsymbol{x})$ of the random vector $\boldsymbol{v}$ can be approximated by a finite distribution mixture of product components:

$$P(\boldsymbol{x}) = \sum_{m=1}^{M} w_m F(\boldsymbol{x}|m), \qquad F(\boldsymbol{x}|m) = \prod_{n=1}^{24} p_n(x_n|m), \quad \boldsymbol{x} \in \mathcal{X}, \quad \sum_{m=1}^{M} w_m = 1. \tag{2}$$

Here $w_m \geq 0$ is the a priori weight of the m-th component, $p_n(x_n|m)$ are the conditional (component specific) univariate distributions of the variables v_n and M is the number of components.

The standard way to estimate the parameters of the distribution mixture (2) is to use the EM algorithm, which converges monotonously to a possibly local maximum or a saddle point of the log-likelihood criterion (Schlesinger 1968, Dempster et al. 1977, Grim 1982, Grim 1992, Grim and Boček 1996, Grim et al. 2001, Grim et al. 2004)).

We recall that any marginal distribution of the mixture (2) is easily obtained by ignoring superfluous terms in the products. In view of this property, the discrete distribution mixture

(2) is directly applicable as a knowledge base of the Probabilistic Expert System (PES) (cf. Grim 1994, Grim and Boček 1996). The inference mechanism of PES can derive the statistical information from the estimated model without any access to the original data. In particular, considering a given input sub-vector

$$\boldsymbol{x}_C = (x_{i_1}, x_{i_2}, \ldots, x_{i_k}) \in \mathcal{X}_C, \quad C = \{i_1, i_2, \ldots, i_k\} \subset \{1, 2, \ldots, 24\},$$

and an output variable $x_n, (n \notin C)$, we can write directly Eqs. for the related marginal

$$P_C(\boldsymbol{x}_C) = \sum_{m=1}^{M} w_m F_C(\boldsymbol{x}_C|m), \quad F_C(\boldsymbol{x}_C|m) = \prod_{i \in C} p_i(x_i|m), \quad \boldsymbol{x}_C \in \mathcal{X}_C, \tag{3}$$

and for the corresponding conditional distribution

$$P_{n|C}(x_n|\boldsymbol{x}_C) = \frac{P_{n,C}(x_n, \boldsymbol{x}_C)}{P_C(\boldsymbol{x}_C)} = \sum_{m=1}^{M} W_m(\boldsymbol{x}_C) p_n(x_n|m), \quad (P_C(\boldsymbol{x}_C) > 0). \tag{4}$$

Here $W_m(\boldsymbol{x}_C)$ are the conditional component weights for the given subvector $\boldsymbol{x}_C \in \mathcal{X}_C$:

$$W_m(\boldsymbol{x}_C) = \frac{w_m F_C(\boldsymbol{x}_C|m)}{\sum_{j=1}^{M} w_j F_C(\boldsymbol{x}_C|j)}. \tag{5}$$

Let us note that the conditional distributions $P_{n|C}(x_n|\boldsymbol{x}_C)$ (conditional histograms) describe the statistical properties of the sub-population specified by the sub-vector $\boldsymbol{x}_C$ in terms of all variables x_n not included in $\boldsymbol{x}_C$. For a given input $\boldsymbol{x}_C$ the formula (4) is applicable to different variables $n \notin C$ with identical weights $W_m(\boldsymbol{x}_C)$. Thus, for any fixed subvector $\boldsymbol{x}_C$, we obtain a set of histograms which characterize the corresponding subpopulation. We can store extensive lists of sub-populations efficiently in terms of defining sub-vectors. In this way different sub-populations can be quickly compared and characterized, e.g., by the most apparent differences from the whole population. In addition, the analytical simplicity of the statistical model suggests some new possibilities of information analysis.

3 Model Based Information Analysis

Another possibility to utilize the latent information potential of the statistical model is to analyze the properties of sub-populations (cf. Grim et al., 2004). A natural basis of information analysis is a suitably chosen list $\mathcal{A}$ of statistically relevant sub-populations which can be specified by combining variables (cf. (16)). The general scheme of the considered information

analysis can be summarized as follows: we order the virtual list $\mathcal{A}$ of statistically relevant sub-populations (combinations of responses) according to a chosen statistical criterion and display the initial part of the ordered list to the user. In some cases also the ascending ordering of sub-populations (instead of descending one) could be of interest. In this section we suggest some criteria which may be useful for different purposes.

A very simple criterion to order the sub-populations $\mathcal{A}$ is the conditional probability of a specific value $x_n \in \mathcal{X}_n$. We can order the sub-populations $\mathcal{S}(\boldsymbol{x}_C)$ from the list $\mathcal{A}$ according to the highest conditional probability $P_{n|C}(x_n|\boldsymbol{x}_C)$ (cf. (4)). By displaying the initial part of the ordered sub-population list we can identify, e.g., social groups or sub-populations which are particularly hit by unemployment if the variable x_n defines unemployed respondents. Obviously, we should exclude from evaluation the "trivial" sub-populations $\mathcal{S}(\boldsymbol{x}_C)$ for which $n \in C$ since in these cases the probability $P_{n|C}(x_n|\boldsymbol{x}_C)$ is trivially either 1 or 0.

A simple modification of the conditional distribution $P_{n|C}(x_n|\boldsymbol{x}_C)$ is to use the unconditional probability

$$P_{nC}(x_n, \boldsymbol{x}_C) = P_{n|C}(x_n|\boldsymbol{x}_C)P(\boldsymbol{x}_C) = \sum_{m=1}^{M} w_m p_n(x_n|m)F_C(\boldsymbol{x}_C|m). \tag{6}$$

The preceding criterion can be easily generalized to a pair of specified values $x_n \in \mathcal{X}_n, x_r \in \mathcal{X}_r$:

$$P_{nr|C}(x_n, x_r|\boldsymbol{x}_C) = \sum_{m=1}^{M} W_m(\boldsymbol{x}_C)p_n(x_n|m)p_r(x_r|m). \tag{7}$$

In this way the sub-populations can be ordered with respect to the highest relative frequency of a pair of values, for example we can identify sub-populations with a high unemployment of young people. Analogously a natural alternative to this criterion is to use the unconditional probability

$$P_{nrC}(x_n, x_r, \boldsymbol{x}_C) = P_{nr|C}(x_n, x_r|\boldsymbol{x}_C)P(\boldsymbol{x}_C) = \sum_{m=1}^{M} w_m p_n(x_n|m)p_r(x_r|m)F_C(\boldsymbol{x}_C|m) \tag{8}$$

which corresponds to the estimated frequency $|\mathcal{S}|P_{nrC}(x_n, x_r, \boldsymbol{x}_C)$ of the values $x_n, x_r, \boldsymbol{x}_C$. Again, in the evaluation process we should exclude the combinations $\boldsymbol{x}_C$ for which $n, r \in C$ because the corresponding probabilities $P_{nrC}(x_n, x_r, \boldsymbol{x}_C)$ equal to 1 or 0.

In some cases we could be interested in sub-populations where the conditional distribution of a variable concentrates on an arbitrary single value (or small subset of values). For example

we could look in general for sub-populations having a typical (prevailing) type of occupation. In such a case a suitable choice would be to use the minimum entropy criterion

$$H_{\boldsymbol{x}_C}(\mathcal{X}_n) = \sum_{x_n \in \boldsymbol{\mathcal{X}}_n} -P_{n|C}(x_n|\boldsymbol{x}_C) \log P_{n|C}(x_n|\boldsymbol{x}_C). \tag{9}$$

In other words, in the sub-populations characterized with a low entropy $H_{\boldsymbol{x}_C}(\mathcal{X}_n)$ the answer to the n-th question is almost unique. Note that it would be rather difficult to identify such sub-populations by other means, e.g., by counting the relative frequencies.

The statistical model also provides a general possibility to identify dependence between categorial variables. Recall that the standard tool to characterize relation between two real random variables is the correlation coefficient computed by means of the expected value of the normalized product of the involved variables. Unfortunately, in case of discrete nominal variables like eye color, profession, marital status etc., the product of two variables is undefined and there is no generally acceptable way to introduce a reasonable definition.

One possibility to analyze the statistical dependence between nominal (qualitative) random variables is to use the statistical information. If $X_n, X_r,\ n, r, \in \mathcal{N}$ are two discrete random variables then their mutual statistical information can be expressed by means of the Shannon formula

$$I(X_n, X_r) = H(X_n) + H(X_r) - H(X_n, X_r) \tag{10}$$

where $H(X_n), H(X_r), H(X_n, X_r)$ are the respective Shannon entropies:

$$H(X_n) = \sum_{x_n \in X_n} -P_n(x_n) \log P_n(x_n), \qquad P_n(x_n) = \sum_{m=1}^{M} w_m p_n(x_n|m), \quad n \in \mathcal{N}, \tag{11}$$

$$H(X_n, X_r) = \sum_{x_r \in X_r} \sum_{x_n \in X_n} -P_{nr}(x_n, x_r) \log P_{nr}(x_n, x_r), \quad n, r \in \mathcal{N}, \tag{12}$$

$$P_{nr}(x_n, x_r) = \sum_{m=1}^{M} w_m p_n(x_n|m) p_r(x_r|m). \tag{13}$$

The Shannon information is zero if the two variables X_n, X_r are statistically independent and it is maximum if one of the two variables uniquely defines the value of the other one. The information criterion (10) can be used, e.g., to order the subpopulation list $\mathcal{A}$ according to the statistical dependence between two chosen variables.

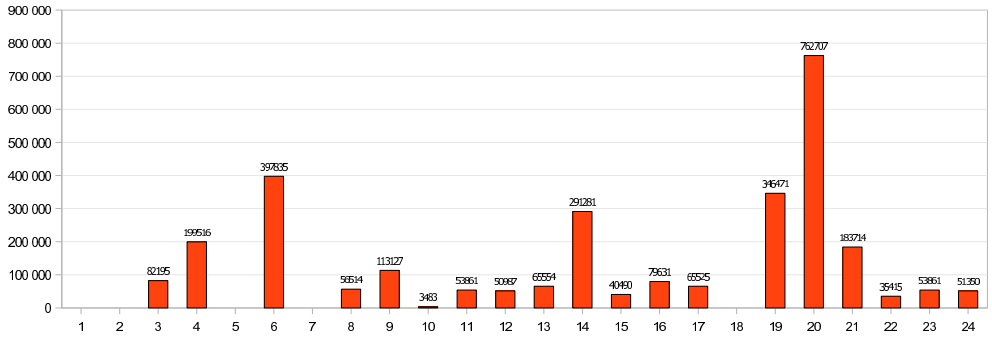

Figure 1: Non-response frequency for individual questions. The number of incomplete records is 1524240, the total number of missing values is 2933427.

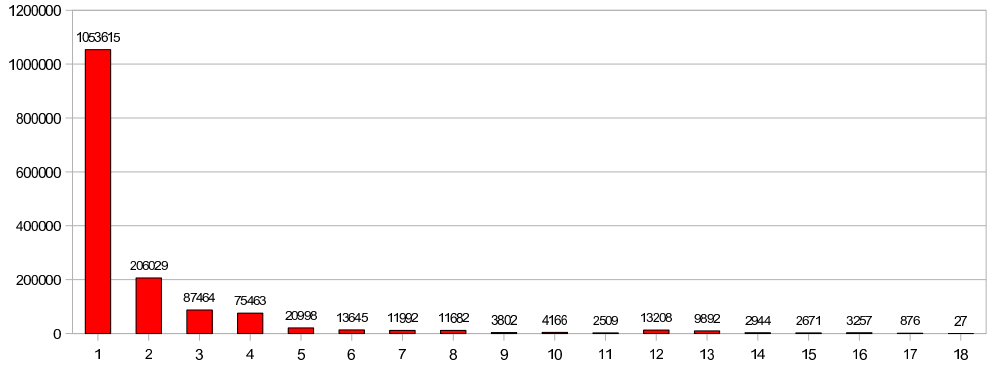

Figure 2: Distribution of incomplete records according to the number of non-response. The total number of missing responses is 2933427.

4 Missing Data - Non Response

A typical feature of census data is the presence of incomplete records. The census database considered in this paper (cf. Table 1) included 1524240 incomplete records containing up to eighteen missing values. The distribution of non-response according to variables is given in the Fig. 1. The next Fig. 2 displays the distribution of non-response by the number of missing values. The total number of missing values in our database was 2933427.

The problem of missing data is traditionally an important area of mathematical statistics because most statistical methods cannot be applied to incomplete data. One can see that, by simply omitting the incomplete records we would lose about 15% of records in our database. Similarly, only five questions would remain should we ignore incomplete variables.

In particular, there are two ways of handling the problem. First we can extend the

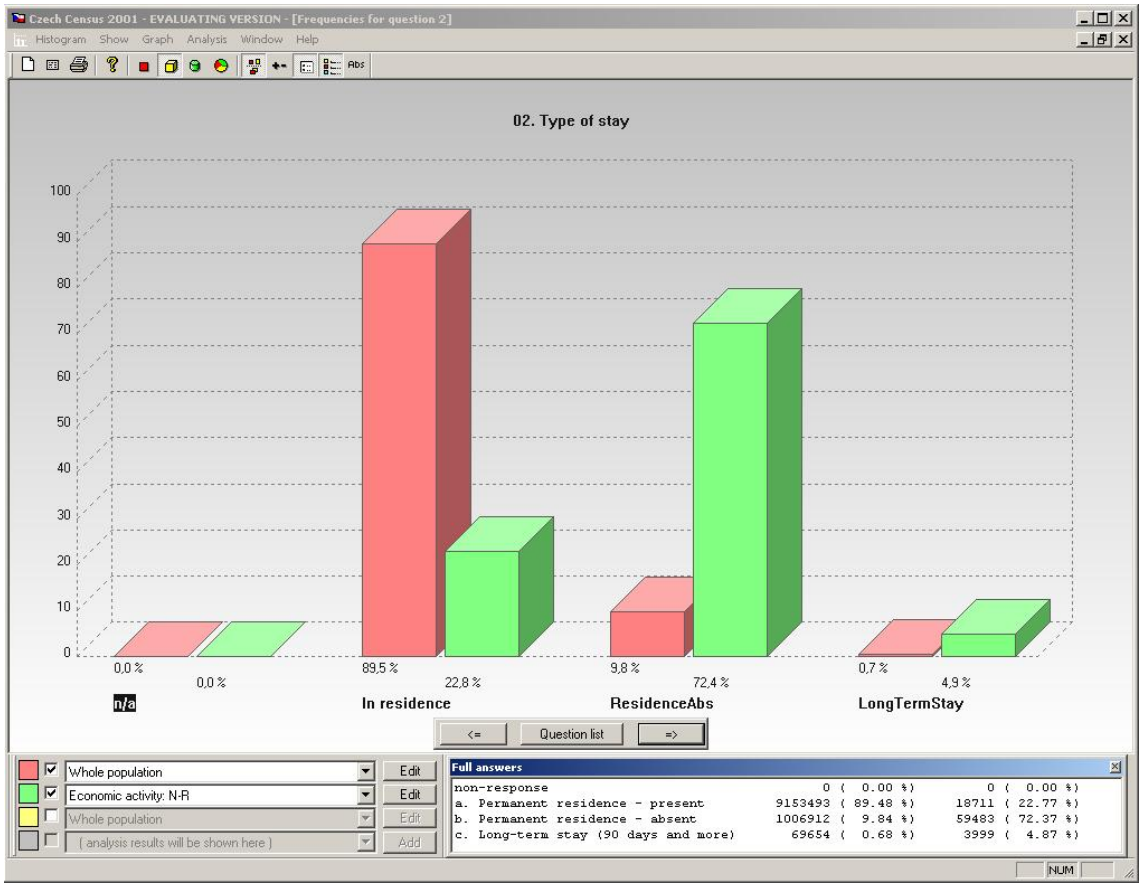

Figure 3: Frequencies for the question "Type of stay" for subpopulation of Non-response for the economic activity question (green) compared to whole population (red). Over 70% of non-responding respondents have permanent residence but were absent.

statistical model by including the "non-response" as an additional response alternative. This allows us to analyze the statistical properties of the "non-response" respondents.

Figure (3) shows a simple example of evaluating properties of the subpopulation of respondents who did not answer the question 3 - Economic activity (green row) compared to the whole population (red row). More than 70% of the green subpopulation are people who have permanent residence but were currently absent.

5 Substitution of Missing Data

An important feature of estimating product mixtures is the possibility to modify the EM algorithm to be directly applicable to incomplete data. In this case there is no necessity

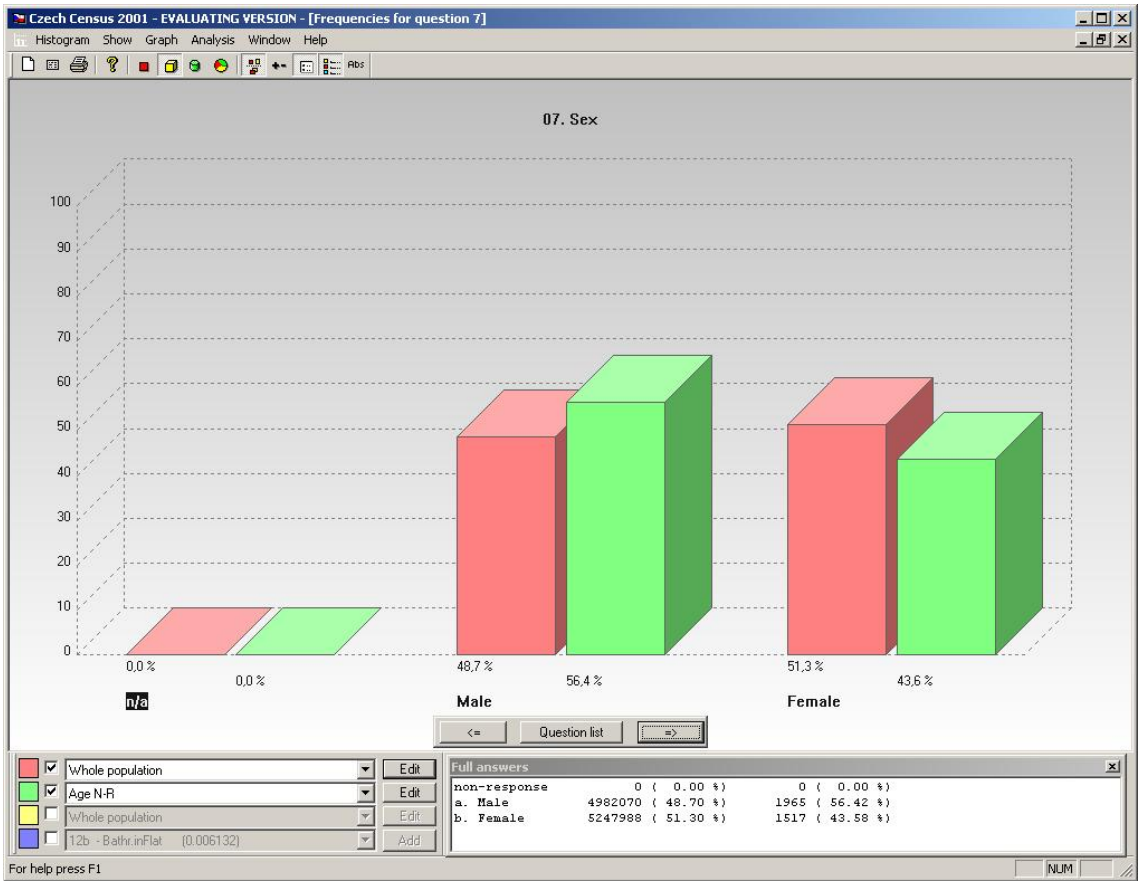

Figure 4: Properties of the respondents who had not filled in their age. It can be seen that more often it was a man who had not told his age.

to substitute for the missing values, we estimate the mixture parameters from the available data only. Formally, the type of missing values is irrelevant, the estimated model is capable to utilize all statistical information available in the data (cf. Grim at al 2009 ...).

However, it appears that the mixture model estimated from incomplete data is biased by a considerable error already at the level of unconditional marginals. In additional experiments we have found that the accuracy of a model obtained from incomplete data is approximately twice worse than that of the comparable model computed from complete data.

Let us recall that the main purpose of our project has been to verify the possibility to reproduce the statistical properties of a large set of microdata and therefore the model accuracy is of primary importance. For this reason we decided to solve the estimation problem in two steps. First we estimated the distribution mixture (2) from incomplete data by means of the modified EM algorithm. The resulting mixture (M=10000) has been used

to replace missing values by estimates. And in the second step we have used the completed database to estimate the final distribution mixture.

It is obvious that the imputation of missing values may affect the final model accuracy. There is no direct possibility to verify if the replacement of missing values has been done correctly but we can simulate an analogous situation by estimating known values. In particular, for each variable separately, we have chosen randomly 10^5 records with available value of the tested variable. Then we computed the corresponding estimate of this value and compared it with the true original. The results of the imputation test are summarized in the Table 2, which provides an additional information about the accuracy of the final statistical model (M=15000). The third column contains the number of non-response and in the fourth column we list the percentage of the correctly estimated values. The number in parentheses corresponds to the trivial imputation of the most frequent response. Expectedly, the imputation accuracy is variable-dependent. In some cases the success of global imputation of the most frequent value is comparable with the statistical model (Nos. 2, 4, 5, 12, 14, 20, 22) but the improvement achieved by the maximum-likelihood estimate is often considerable (Nos. 1, 3, 8, 9, 10, 13, 15, 17, 19, 23). In the mean 73% of missing values would be correctly identified by the maximum-likelihood estimates. The last column contains the number of the probably correctly replaced non-response from the third column.

6 Accuracy of the Statistical Model

Let us recall that the primary purpose of the estimated model is to reproduce the statistical properties of the original data. In the domain of statistical surveys we usually specify the properties of sub-populations by combining responses. Therefore the statistical model should reproduce the empirical frequencies of different properties as exactly as possible. In particular, in order to verify the model accuracy, we compare the empirical frequency of different combinations of responses with the estimates derived from the statistical model. Considering an elementary property defined by a sub-vector of responses $\boldsymbol{x}_C$, we denote

$$\mathcal{S}(\boldsymbol{x}_C) = \{\boldsymbol{y} \in \mathcal{S} : \boldsymbol{y}_C = \boldsymbol{x}_C\}, \quad N(\boldsymbol{x}_C) = |S(\boldsymbol{x}_C)|, \quad \boldsymbol{x}_C = (x_{i_1}, \ldots, x_{i_k}) \in \mathcal{X}_C \tag{14}$$

where $\mathcal{S}(\boldsymbol{x}_C)$ is the subset of respondents (a subpopulation) with the property $\boldsymbol{x}_C$ and $N(\boldsymbol{x}_C)$ is the (empirical) frequency of the property $\boldsymbol{x}_C$ in the census population $\mathcal{S}$. Obviously, the

frequency $N(\boldsymbol{x}_C)$ can be estimated from the statistical model (2) as the product of the probability $P(\boldsymbol{x}_C)$ and the population size $|\mathcal{S}|$:

$$\hat{N}(\boldsymbol{x}_C) = |\mathcal{S}|P(\boldsymbol{x}_C), \quad P(\boldsymbol{x}_C) = \sum_{m=1}^{M} w_m \prod_{j=1}^{k} p_{i_j}(x_{i_j}|m) \tag{15}$$

It appears that, ideally, we should compare the estimated frequency $\hat{N}(\boldsymbol{x}_C)$ with the empirical value $N(\boldsymbol{x}_C)$ for all possible elementary combinations of values $\boldsymbol{x}_C$. However, there are two important limitations.

Recall first, that we are not interested to reproduce small frequencies. On the contrary, the decreasing accuracy of the model at low probabilities is an important confidentiality protecting property. For this reason we decided to evaluate the accuracy of the estimates $\hat{N}(\boldsymbol{x}_C)$ only for the empirical frequencies $N(\boldsymbol{x}_C)$ greater than a suitably chosen threshold N_ϵ. In order to specify the threshold frequency N_ϵ we confine ourselves only to "statistically relevant" properties $\boldsymbol{x}_C$, the frequency of which may differ from the assumed "true" unknown frequency $N^*(\boldsymbol{x}_C)$ by less than $\epsilon = 5\%$ (cf. Appendix II). In particular, if we confine ourselves to the properties $\boldsymbol{x}_C$ satisfying the inequality $N(\boldsymbol{x}_C) > 1612$ (i.e., $N_\epsilon = N_{0.05} = 1612$), then, according to the central limit theorem of probability theory, the empirical frequency $N(\boldsymbol{x}_C)$ of the property $\boldsymbol{x}_C$ in the population $\mathcal{S}$ may differ from the unknown "true" frequency $N^*(\boldsymbol{x}_C)$ by less than 5% (at the confidence level 0.95).

The second limitation has a computational reason. The number of all properties $\boldsymbol{x}_C$ specified by all possible combinations of responses is too high and the evaluation would be too time-consuming. For this reason we decided to verify the model accuracy by considering combinations of at most five responses. As a result we obtained a list $\mathcal{A}_5$ of about 26 millions "statistically relevant" properties $\boldsymbol{x}_C$ along with the corresponding empirical frequencies

$$\mathcal{A}_5 = \{\boldsymbol{x}_C = (x_{i_1}, \ldots, x_{i_5}) : N(\boldsymbol{x}_C) > 1612\}, \qquad |\mathcal{A}_5| = 26425727. \tag{16}$$

A natural way to measure the accuracy of the statistical model (2) is to compute the mean absolute error E_a of the estimated frequencies $\hat{N}(\boldsymbol{x}_C)$ for the properties $\boldsymbol{x}_C \in \mathcal{A}_5$:

$$E_a = \frac{1}{|\mathcal{A}_5|} \sum_{x_C \in \mathcal{A}_5} |P(\boldsymbol{x}_C)|\mathcal{S}| - N(\boldsymbol{x}_C)|, \qquad P(\boldsymbol{x}_C) = \sum_{m=1}^{M} w_m \prod_{j=1}^{5} p_{i_j}(x_{i_j}|m) \tag{17}$$

where $P(\boldsymbol{x}_C)$ is the probability of the combination $\boldsymbol{x}_C$ computed by means of the mixture model (2). However, as it can be seen, the criterion E_a does not differentiate between errors

of large and small estimates. For this reason we have introduced the following mean relative error criterion

$$E_r = \frac{100}{|\mathcal{A}_5|} \sum_{x_C \in \mathcal{A}_5} \frac{|P(\boldsymbol{x}_C) - \frac{N(\boldsymbol{x}_C)}{|\mathcal{S}|}|}{\frac{N(\boldsymbol{x}_C)}{|\mathcal{S}|}} = \frac{100}{|\mathcal{A}_5|} \sum_{x_C \in \mathcal{A}_5} \frac{|P(\boldsymbol{x}_C)|\mathcal{S}| - N(\boldsymbol{x}_C)|}{N(\boldsymbol{x}_C)} \tag{18}$$

which is more sensitive in this respect since the same absolute difference of frequencies is less important if the empirical frequency $N(\boldsymbol{x}_C)$ is high and more important for lower $N(\boldsymbol{x}_C)$.

We have used the criteria E_a and E_r to evaluate the accuracy of the final distribution mixture (2). Table **??** contains the results obtained by applying both criteria to the list of properties $\mathcal{A}_5$ (third column) and, for comparison, to the list $\mathcal{A}_4$ of properties specified by maximally four responses (second column). For both of the considered tests Table **??** shows the mean relative and mean absolute error and the corresponding standard deviations. In addition we have computed the maximum relative and absolute error and also the number of relative errors exceeding 100%. The mean relative error was 4.2% in case of the list $\mathcal{A}_5$ and 4.1% in case of the list $\mathcal{A}_4$, the corresponding absolute error was 338 and 460 respondents respectively. Since all other results are comparable, too, we may assume that a more extensive test experiment would not yield essentially different values. We recall that by combining more than five responses we would obtain mostly very small frequencies $N(\boldsymbol{x}_C)$ that would fall below the threshold N_ϵ and therefore the resulting list would not be much longer than $\mathcal{A}_5$.

Let us recall that the relative error in the criterion E_r is invariant with respect to arbitrary norming. Consequently, the mean error of any displayed histogram column is 4.2%. In order to illustrate the distribution of relative errors in more detail we include Table **??**. As it can be seen, for very small empirical frequencies ($1612 < N(\boldsymbol{x}_C) < 3000$) the mean relative error is 6.12% and quickly decreases for greater values of $N(\boldsymbol{x}_C)$ (larger sub-populations). Our interactive software disables evaluation of sub-populations $S(\boldsymbol{x}_C)$ smaller than the threshold value $N_{0.05} = 1612$ and indicates any histogram column which corresponds to a sub-threshold frequency.

Obviously, the results in Table **??** strongly depend on the chosen sub-population threshold N_ϵ. It is therefore unclear whether the achieved mean relative error 4.2% is to be considered too high or low enough. To answer this question we have compared the accuracy of our mixture model with the reproduction accuracy of a randomly chosen subset of 1 million

individual microdata records (10% of the original data set). Note that the standard way of statistical information dissemination by means of subsets of anonymized microdata provides similar comfort and flexibility in evaluating the statistical properties of survey data as the interactive statistical model. In particular, we can estimate the empirical frequencies $N(\boldsymbol{x}_C)$ by using a representative subset of microdata. For the sake of comparison with the statistical model we have evaluated the accuracy of the microdata subset in the same way as in the Table **??**. As it can be seen in Table 3, at reproducing the empirical frequencies the accuracy of the 10% microdata subset is marginally better than the statistical model. In practical situation the results in the Table 3 would be most likely worse due to errors introduced by the necessary anonymization process which has been omitted in our test case.

7 Interactive Data Presentation

OBRÁZKY Z APLIKACE

8 Concluding Remarks

The so-far most informative way to disseminate statistical information is to release representative subsets of anonymized microdata. With appropriate microdata the users have the full freedom to examine arbitrary hypotheses and issues beyond the usual scope of data providers. Unfortunately, both the choice of a subset of the original microdata (typically about one million of individual records) and the indispensable anonymization procedures may negatively influence the statistical validity of the contained information. Moreover, there is always some residual disclosure risk and for this reason the statistical agencies release microdata for research purposes only, usually under special licence agreements and through secure data archives.

In view of these facts the primary purpose of the considered statistical model has been to make the census results freely available in a new user-friendly way with a well guaranteed confidentiality of data. The resulting interactive software provides flexibility and user comfort analogous to the sets of anonymized microdata at a comparable or even higher level of accuracy. In addition, the analytical simplicity of the underlying distribution mixture opens new possibilities of information oriented data analysis (data mining) based on efficient eval-

uation of a virtual list of several hundreds of thousands of sub-populations. The statistical model does not contain the original data and therefore the final interactive software product can be distributed without any confidentiality concerns.

The representation accuracy of the statistical model has been analyzed in detail. We have shown that the resulting distribution mixture can be used to estimate the probability of complete or incomplete records with a high reliability. This property can be used to estimate missing values but also to identify unusual or possibly incorrect records. The identification of untypical records is a crucial step of most of the anonymization algorithms, but the application of SDC techniques becomes superfluous in case of statistical models.

References

[1] Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990). Disclosure Control of Micro-data. Journal of the American Statistical Association, 85 (409), 38-45.

[2] Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc., B 39, 1-38.

[3] Duncan, G., Lambert, D. (1989). The risk of disclosure for micro-data. Journal of Business & Economic Statistics, 7, 207-217.

[4] Dalenius, T. (1977). Towards a methodology for statistical disclosure control. Statistik Tidskrift, 15, 429-444.

[5] Feller, W. (1962). An Introduction to Probability Theory and Its Applications. I, John Wiley & Sons, New York, London.

[6] Fienberg, S.E. (1994). Conflicts between the needs for access to statistical information and demands for confidentiality. Journal of Official Statistics, 10, 115-132.

[7] Fienberg, S.E., Makov, U.E., Steel, R.J. (1998). Disclosure limitation using perturbation and related methods for categorical data, with discussion. Journal of Official Statistics, 14, 485-502.

[8] Grim, J. (1982). On numerical evaluation of maximum - likelihood estimates for finite mixtures of distributions. Kybernetika, 18, (3), 173-190.

[9] Grim, J. (1990). Probabilistic expert systems and distribution mixtures. Computers and Artificial Intelligence, 9, (3), 241-256.

[10] Grim, J. (1992). A dialog presentation of census results by means of the probabilistic expert system PES. In Proceedings of the Eleventh European Meeting on Cybernetics and Systems Research, Vienna 21-24 April 1992, (Ed. R. Trappl), 997-1005, World Scientific, Singapore 1992.

[11] Grim, J. (1994). Knowledge representation and uncertainty processing in the probabilistic expert system PES. Int. Journal of General Systems, 22, (2), 103-111.

[12] Grim, J., Boček, P. (1996). Statistical model of Prague households for interactive presentation of census data. In SoftStat'95. Advances in Statistical Software 5, 271-278, Lucius & Lucius: Stuttgart.

[13] Grim, J., Boček, P., Pudil, P. (2001). Safe dissemination of census results by means of interactive probabilistic models. In: Proceedings of the ETK-NTTS 2001 Conference. (Nanopoulos P., Wilkinson D. eds.) European Communities, Rome 2001, 849-856.

[14] Grim, J., Hora, J., Boček, P., Somol, P., Pudil, P. (2004). Information Analysis of Census Data by Using Statistical Models. In: Proceedings: Statistics - Investment in the Future, Praha.

[15] Grim, J., Hora, J., Pudil, P. (2004). Statistical Model for Interactive Presentation of Census Results under Protection of Confidentiality. (in czech) Statistika, 40, (5), 400-414.

[16] McLachlan, G.J., Peel, D. (2000). Finite Mixture Models. John Wiley & Sons: New York, Toronto.

[17] Schlesinger, M.I. (1968). Relation between learning and self-learning in pattern recognition (in Russian). Kibernetika, (Kiev), (2), 81-88.

[18] Willenborg, L.C.R.J., de Waal, A.G. (2001). Elements of statistical disclosure control. Springer Verlag, New York.

[19] Winkler, W.E. (1998). Re-identification methods for evaluating the confidentiality of analytically valid microdata. Research in Official Statistics, 2, 87-104.

[20] Grim, J., Somol, P.,Boček, P.,Hora, J., Pudil, P. (2009). Interactive statistical data model from Czech census 2001 (in czech), Statistika, 4-2009, 285-299.

[21] Grim, J., Hora, J., Boček, P., Somol, P., Pudil, P. (submitted). Statistical Model of the Czech Census 2001 for Interactive Presentation. Journal of Official Statistics

	Text of question (name of variable)	Number of values	Non-response in %	Shannon entropy in %
1.	Region of residence	14	0.00	96.88
2.	Type of residence	3	0.00	32.92
3.	Economic activity	10	0.80	67.80
4.	Birth place (relatively)	6	1.95	74.65
5.	Religion	6	0.00	60.57
6.	Occupation type	14	3.89	68.33
7.	Sex	2	0.00	99.95
8.	Marital status	4	0.55	81.01
9.	Education	14	1.11	78.04
10.	Age	9	0.03	96.09
11.	Category of flat	5	0.53	27.81
12.	Bathroom	5	0.59	14.02
13.	Size of flat	7	0.64	80.62
14.	Internet and PC	4	2.85	49.11
15.	Legal relation to flat	9	0.39	72.43
16.	Gas supply	3	0.78	64.54
17.	Number of rooms over $8m^2$	7	0.64	80.57
18.	Number of cars in household	4	3.39	71.32
19.	Number of persons in flat	6	0.00	93.79
20.	Vacational property	6	7.45	42.10
21.	Telephone in flat	5	1.80	80.88
22.	Water supply	4	0.35	8.02
23.	Type of heating	6	0.53	74.81
24.	Toilet	6	0.50	16.73

Table 1: List of questions included in the statistical model of the Czech Census 2001. The third column contains the number of possible responses, percentage of missing values (non-response) is given in the fourth column. There are 1524240 incomplete records, the total number of non-response is 2933427. Uncertainty of variables in % of maximum Shannon entropy is given in the last column.

	Text of question (name of variable)	Number of non-response	Successful imputation in %	Successful imputation
1.	Region of residence	0	27.49 (12.41)	0
2.	Type of residence	0	90.35 (89.48)	0
3.	Economic activity	82195	88.02 (44.08)	72348
4.	Birth place (relatively)	199516	56.36 (53.52)	112447
5.	Religion	0	66.27 (59.04)	0
6.	Occupation type	397835	67.64 (50.62)	269096
7.	Sex	0	67.91 (51.30)	0
8.	Marital status	56514	82.91 (46.63)	46856
9.	Education	113127	48.36 (19.29)	54708
10.	Age	3483	59.22 (16.71)	2063
11.	Category of flat	53861	97.48 (89.37)	52504
12.	Bathroom	50987	98.90 (95.91)	50426
13.	Size of flat	65554	63.22 (38.48)	41443
14.	Internet and PC	291281	81.12 (79.15)	236287
15.	Legal relation to flat	40490	63.49 (39.70)	25707
16.	Gas supply	79631	75.94 (63.84)	60472
17.	Number of rooms over 8m^2	65525	63.48 (38.76)	41595
18.	Number of cars in household	346471	66.97 (51.77)	232032
19.	Number of persons in flat	0	49.48 (29.27)	0
20.	Vacational property	762707	80.39 (78.11)	613140
21.	Telephone in flat	183714	57.36 (43.93)	105378
22.	Water supply	35415	99.39 (98.08)	35199
23.	Type of heating	53861	76.90 (41.45)	41419
24.	Toilet	51350	97.98 (94.32)	50313
	Total:	**2 933 427**	**73.06 (61.35)**	**2 143 326**

Table 2: Accuracy of the estimation of missing values. The third column contains the number of non-response. In the fourth column we list the percentage of correctly estimated responses. The numbers in parentheses correspond to the trivial imputation of the most frequent response. In the mean 73% of missing values would be correctly identified by the maximum-likelihood imputation procedure. The last column lists expected numbers of the correctly replaced non-response from the third column. The total number of non-response is 2933427.

Used model	Model with substituted values	Extended model with missing values
Mean relative error in %:	**4.07**	**4.10**
Standard deviation of the relative error:	6.33	5.83
Maximum relative error of the model in %:	240.84	250.90
Number of relative errors exceeding 100%:	925	1037
Mean absolute error:	**470**	**459**
Standard deviation of the absolute error:	951	791
Maximum absolute error of the model:	45779	56808
Number of combinations tested:	**3503448**	**3895873**

Table 3: Mean relative and mean absolute error of the two statistical models with M=15000 components - one computed using original data with missing values, the second using data where missing data were substituted. It can be seen that the models are of comparable quality.