Statistical Model of the 2001 Czech Census for Interactive Presentation

Jiří Grim [1], *Jan Hora* [2], *Pavel Boček* [1], *Petr Somol* [1], *Pavel Pudil* [3]

This paper describes the application of a recently developed method of interactive statistical database presentation to the 2001 Czech Census. The method is based on estimating the multivariate probability distribution of the original microdata which are supposed to be discrete or discretized continuous. The estimated statistical model in the form of a distribution mixture of product components can be used as a knowledge base of a probabilistic expert system. By means of the probabilistic inference mechanism we can derive conditional distributions of variables for each subpopulation interactively without any further access to the source database. The conditional distributions (histograms) describing the properties of subpopulations represent the basic form of user information. The statistical model does not contain the original data and therefore can be distributed without any confidentiality concerns. The accuracy achievable by the statistical model is comparable with that of the anonymized subsets of microdata.

Keywords: Interactive statistical model, census data presentation, distribution mixtures, data modeling, EM algorithm, incomplete data, data reproduction accuracy, data mining

1. Introduction

In our modern networked society there is an increasing demand for dissemination and sharing of statistical information. To meet the expectations of users, statistical agencies release two major forms of statistical data: 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 follows from 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, 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, and 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 eliminated by suppressing cells, aggregating values, removing sensitive variables or by other techniques. In a case of microdata, easily identifiable quantitative variables

[1] Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic
[2] Faculty of Nuclear Science and Physical Engineering, Czech Technical University
[3] Faculty of Management, Prague University of Economics, Czech Republic

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

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 extensive literature on SDC techniques available (for the most frequent references see, e.g., Dalenius 1977, Bethlehem et al. 1990, Winkler 1998, Fienberg 1994, Fienberg et al. 1998, Willenborg and de Waal 2001), 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 be harmful for the re-identified respondent. Much extra work is needed to produce safe and analytically valid public-use files and, on the other hand, there is always a remaining 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 on information dissemination.

In recent 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 discrete microdata in the form of a multivariate distribution mixture with product components using the EM algorithm (Dempster et al. 1977). We assume the variables to be discrete (ordinal or categorical, qualitative, nominal). The mixture model is not suitable to reproduce the statistical properties of continuous variables and discrete variables with extensive number of values (small area identifiers, detailed age groups). Continuous variables have to be discretized by introducing intervals. The estimated product mixture can be used directly as a knowledge base of a probabilistic expert system PES (Grim 1990, Grim 1994). We can thus 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 analysis which is comparable or even better than microdata subsets, in some respects. The resulting software product does not contain original microdata and therefore it can be distributed freely, without any limitation. The interactive model-based re-identification of respondents is prevented 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 since the mixture model "fits" better to more probable properties than, e.g., to outliers.

The method of interactive statistical model can improve the information offer of statistical offices but does not provide exact values and it is not suitable to reproduce the statistical properties of continuous variables and of discrete variables with extensive number of values (small area identifiers, detailed age groups). Nevertheless, we assume that the limited applicability of the method is well counterbalanced by the possibility of unlimited distribution of the final software product. According to our best knowledge, in recent literature there are no similar results on statistical data models based on distribution mixtures.

In this paper we describe an application of the proposed method to the individual microdata records from the 2001 Czech Census. The statistical model has been computed within 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 the databases of persons and households that were originally treated separately.

The resulting source database contained 10230060 records, with about 1.5 million incomplete records including nearly three million non-response (missing) values. As the primary purpose of the project has been to demonstrate the accuracy of the method in the 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 the 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 established that the accuracy of model probabilities is comparable with that of the relative frequencies computed from a randomly chosen one-million subset of the original microdata (without anonymization). The preliminary version of the final interactive software product can be downloaded at our web-page *http://ro.utia.cas.cz/dem.html*, (Data Mining: Interactive Presentation of Census Results by Probabilistic Models).

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 Section 5 we discuss some tools of information analysis and 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 relationships within a given finite 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 2001 Czech Census we have chosen 24 categorial variables (questions) as listed in Table 1. We have applied less detailed coding of some variables (regional localization, age intervals) to decrease the formal complexity of the model. Simultaneously we have also omitted 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 the respective 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 as a percentage. The total number of non-response is 2933427. The uncertainty of variables expressed in percentage of maximum Shannon entropy is given in the last column. As it can be seen, the highest uncertainty comes with question 7 (respondent sex implies two nearly equal response frequencies). On the contrary, question 22 (water supply) implies the most unambiguous response.

Formally, we consider the source database to be a set of independent and identically

	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 2001 Czech Census. 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.

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:

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

Here $w_m \geq 0$ is the prior weight of the m-th mixture 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. Note that the product components do not imply the involved variables to be independent. In this sense the mixture model (2) is not restrictive. It is easily verified (cf. Grim, Bocek 1996) that by increasing the number of components we can describe any discrete probability distribution in the form (2).

The standard way to estimate the parameters of the distribution mixture (2) is to use the EM algorithm (Schlesinger 1968, Dempster et al. 1977, Grim 1982, Grim 1992, Grim and Boček 1996, Grim et al. 2001, Grim et al. 2004)). In particular, let $\mathcal{S}$ be a set of K data vectors as obtained, e.g., in census:

$$\mathcal{S} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(K)}\}, \quad \boldsymbol{x}^{(k)} \in \mathcal{X}, \quad (K = |\mathcal{S}|). \tag{3}$$

To estimate the unknown mixture parameters we maximize the log-likelihood function

$$L = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log P(\boldsymbol{x}) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log \left[\sum_{m=1}^{M} w_m F(\boldsymbol{x}|m) \right] \tag{4}$$

by means of the following EM iteration equations ($m = 1, 2, \ldots, M,\ n = 1, 2, \ldots, 24$):

$$q(m|\boldsymbol{x}) = \frac{w_m \prod_{n=1}^{24} p_n(x_n|m)}{\sum_{j=1}^{M} w_j \prod_{n=1}^{24} p_n(x_n|j)}, \qquad w_m^{'} = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q(m|\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{S}, \tag{5}$$

$$p_n^{'}(\xi|m) = \frac{1}{\sum_{x \in \mathcal{S}} q(m|\boldsymbol{x})} \sum_{x \in \mathcal{S}} \delta(\xi, x_n) q(m|\boldsymbol{x}), \quad \xi \in \mathcal{X}_n. \tag{6}$$

Here the apostrophe denotes the new parameter values and $\delta(\xi, x_n)$ is the delta-function in the usual sense ($\delta(\xi, x_n) = 1$ for $\xi = x_n$ and $\delta(\xi, x_n) = 0$ for $\xi \neq x_n$).

The EM algorithm monotonously converges to a local or global maximum or to a saddle point of the log-likelihood function L in the sense that the corresponding sequence of values $\{L^{(t)}\}_{t=0}^{\infty}$ is nondecreasing (cf. Dempster et al. 1977, Grim 1982). The existence of local maxima makes the procedure starting-point dependent. In this respect a well known difficulty is to specify the number of mixture components and to choose the initial parameter values (cf. e.g., McLachlan and Peel 2000). However, the problem becomes less relevant in high-dimensional spaces and with increasing number of components since the values of different local maxima are similar and therefore the quality of the corresponding mixture estimates is comparable. For the same reason in all of our experiments the mixture parameters have been initialized randomly.

As it can be expected, the accuracy of the model increases with the model complexity but, on the other hand, the number of components is the main limiting feature from the computational point of view. For this reason in the following the choice of the number of mixture components M is mainly influenced by practical hardware-specific considerations.

As the main purpose of the mixture model (2) is to reproduce the statistical properties of the original data, it is important that, in each iteration of the EM algorithm, the univariate marginal distributions of the estimated mixture are identical with the global marginal frequencies of the data. In particular, by using Eqs. (5), (6), we can write

$$P_n^{'}(\xi) = \sum_{m=1}^{M} w_m^{'} p_n^{'}(\xi|m) = \sum_{m=1}^{M} \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q(m|\boldsymbol{x})\ p_n^{'}(\xi|m) = \tag{7}$$

$$= \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \delta(\xi, x_n) \sum_{m=1}^{M} q(m|\boldsymbol{x}) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \delta(\xi, x_n), \quad \xi \in \mathcal{X}_n, \quad n = 1, 2, \ldots, 24.$$

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

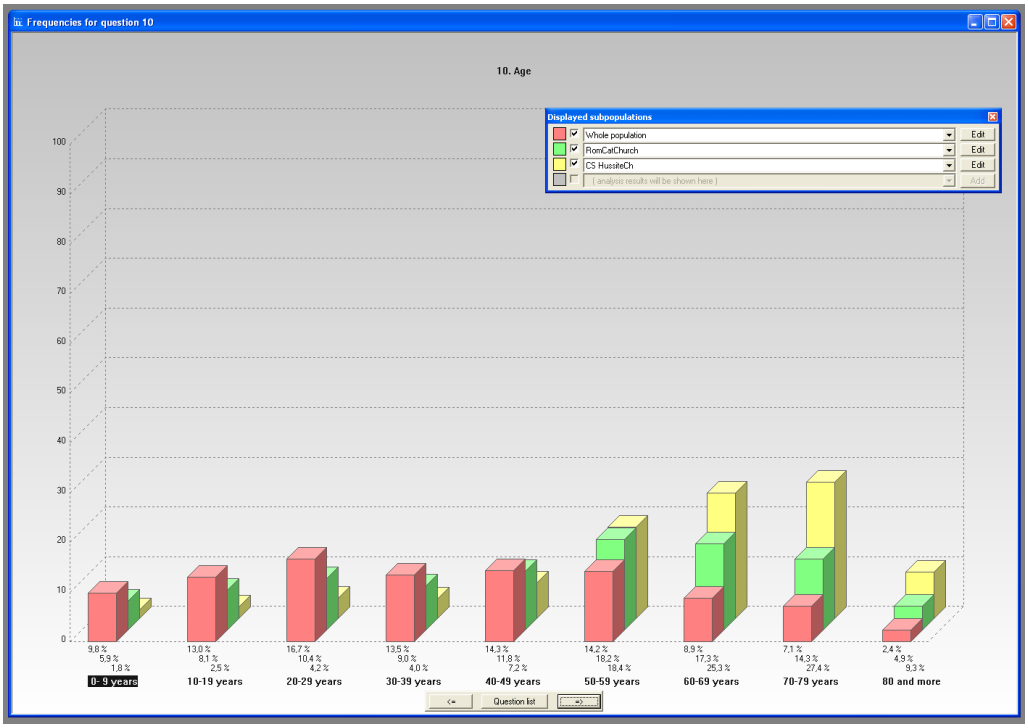

Figure 1: Comparison of age distribution in three different sub-populations (cf. http://ro.utia.cas.cz/dem.html).

mixture (2) is directly applicable as a knowledge base of the Probabilistic Expert System (PES) (cf. Grim 1994, Grim and Boček 1996). In this way the inference mechanism of PES (cf. Appendix I) derives the statistical properties of different sub-populations directly from the estimated model without any access to the original microdata and without using any synthetic data.

From the point of view of a user the conditional histograms describing the properties of different sub-populations represent basic form of statistical inference. In particular, considering a given input sub-vector

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

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

$$P_C(\boldsymbol{x}_C) = \sum_{m=1}^{M} w_m F_C(\boldsymbol{x}_C|m), \qquad F_C(\boldsymbol{x}_C|m) = \prod_{i \in C} p_i(x_i|m), \quad \boldsymbol{x}_C \in \mathcal{X}_C, \tag{8}$$

and for the corresponding conditional distribution

$$P_{n|C}(x_n|\boldsymbol{x}_C) = \frac{P_{n,C}(x_n, \boldsymbol{x}_C)}{P_C(\boldsymbol{x}_C)} = \sum_{m=1}^{M} W_m(\boldsymbol{x}_C) p_n(x_n|m), \qquad (P_C(\boldsymbol{x}_C) > 0). \tag{9}$$

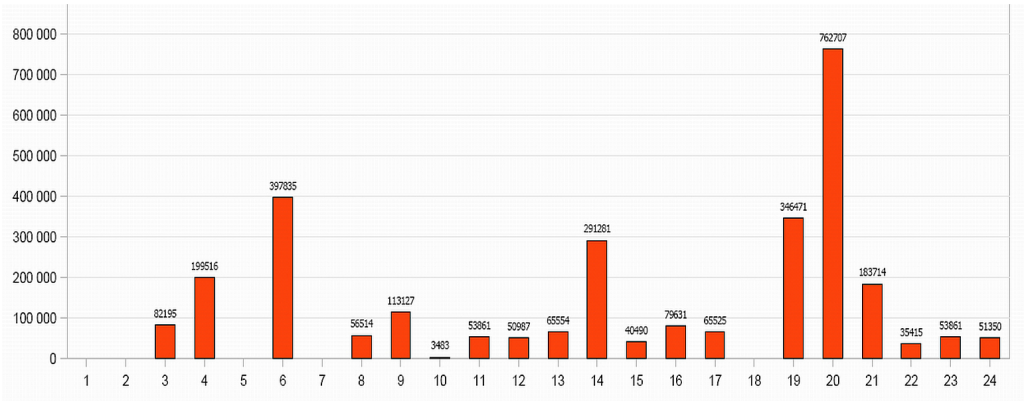

Figure 2: Non-response frequency for individual questions. The number of incomplete records is 1524240, the total number of missing values is 2933427.

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

$$W_m(\boldsymbol{x}_C) = \frac{w_m F_C(\boldsymbol{x}_C|m)}{\sum_{j=1}^{M} w_j F_C(\boldsymbol{x}_C|j)}. \tag{10}$$

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 (9) 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 efficiently store extensive lists of sub-populations in terms of the defining sub-vectors. In this way different sub-populations can be quickly compared and characterized in terms of conditional histograms (cf. Figure 1), 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 (cf. Sec. 4).

3. Missing Data

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 Figure 2. The next, Figure 3, 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 loose about 15% of records in our database. Similarly, only five questions would remain should we ignore incomplete variables. Since the missing values are denoted as non-response in all records, they could always be treated as a specific additional response. However, in most cases, the information value of non-response is limited because of its latent dependence on the context (there are multiple modes of non-response, e.g. don't know, refuse). In this sense the additional value would cause superfluous increase of relationship complexity; it means

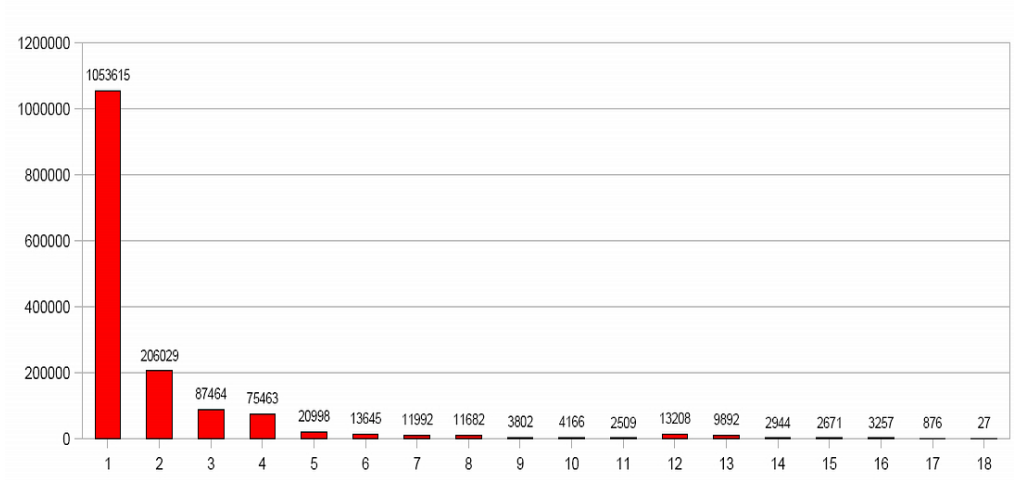

Figure 3: Distribution of incomplete records according to the number of non-response.

we would force the model to describe many meaningless statistical relations.

An important feature of estimating product mixtures is the possibility to modify the EM algorithm to be directly applicable to incomplete data. In this sense there is no necessity to substitute for the missing values; we estimate the mixture parameters from the available data only. The type of missing values is nearly irrelevant since the estimated model can utilize all statistical information in the data. In particular, denoting by $\mathcal{N}(\boldsymbol{x})$, the subset of indices of the available variables of $\boldsymbol{x}$, and $\mathcal{S}_n \subset \mathcal{S}$, the subset of vectors with the available variable x_n:

$$\mathcal{N}(\boldsymbol{x}) = \{n : x_n \text{ available in } \boldsymbol{x}\}, \quad \mathcal{S}_n = \{\boldsymbol{x} \in \mathcal{S} : n \in \mathcal{N}(\boldsymbol{x})\}, \tag{11}$$

we can write the modified EM iteration equations in the form $(m = 1, 2, \ldots, M,\ n = 1, 2, \ldots, 24,\ \boldsymbol{x} \in \mathcal{S})$:

$$q(m|\boldsymbol{x}) = \frac{w_m \prod_{n \in \mathcal{N}(\boldsymbol{x})} p_n(x_n|m)}{\sum_{j=1}^{M} w_j \prod_{n \in \mathcal{N}(\boldsymbol{x})} p_n(x_n|j)}, \qquad w_m^{'} = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q(m|\boldsymbol{x}), \tag{12}$$

$$p_n^{'}(\xi|m) = \frac{1}{\sum_{x \in \mathcal{S}_n} q(m|\boldsymbol{x})} \sum_{x \in \mathcal{S}_n} \delta(\xi, x_n) q(m|\boldsymbol{x}), \quad \xi \in \mathcal{X}_n. \tag{13}$$

Roughly speaking, we calculate the values $q(m|\boldsymbol{x})$ a $p_n^{'}(\xi|m)$ in Eqs. (12), (13) only for the variables currently available in $\boldsymbol{x}$.

From the theoretical point of view we would prefer the direct estimation from incomplete data, since by replacing the missing values we generally use some typical values and, in this way, the natural variability of data decreases. Unfortunately, unlike the standard EM algorithm, the modified iteration equations (12), (13) do not guarantee that the univariate marginal distributions of the estimated mixture are identical with the global marginal frequencies of the data (cf. (7)):

$$P_n^{'}(\xi) = \sum_{m=1}^{M} w_m^{'} p_n^{'}(\xi|m) = \sum_{m=1}^{M} \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q(m|\boldsymbol{x})\ p_n^{'}(\xi|m) = \tag{14}$$

	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:	**2933427**	**73.06 (61.35)**	**2143326**

Table 2: Accuracy of the estimation of missing values. The third column contains the number of non-responses. In the fourth column we list the percentage of correctly estimated responses. The numbers in parentheses correspond to the trivial global 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.

$$= \frac{1}{|\mathcal{S}|} \sum_{m=1}^{M} \frac{\sum_{x \in \mathcal{S}} q(m|\boldsymbol{x})}{\sum_{x \in \mathcal{S}_n} q(m|\boldsymbol{x})} \sum_{x \in \mathcal{S}_n} \delta(\xi, x_n) q(m|\boldsymbol{x}) \neq \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \delta(\xi, x_n), \quad \xi \in \mathcal{X}_n,$$

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 established that a model obtained from incomplete data is approximately twice as inaccurate as 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 (12), (13). The resulting mixture (M=10000) has been used to replace missing values by estimates. In particular, we

have replaced each missing value by the response x_n, which is most probable in the sense of the conditional distribution $p_n(x_n|\boldsymbol{x}_C)$ (cf. (9)):

$$x_n = \arg\max_{\xi \in \mathcal{X}_n} \{p_n(\xi|\boldsymbol{x}_C)\}. \tag{15}$$

In other words, we have replaced each non-response by the value $x_n \in X_n$, which is the most probable response with respect to the known part $\boldsymbol{x}_C$ of the record. In the second step we have used the completed database to estimate the final distribution mixture.

Generally, the computing time of the EM algorithm is proportional to the model complexity; therefore, in case of the given large database, the number of mixture components is the most relevant practical limitation. We have chosen M=15000 components for the final experiment. Again, we generated the initial parameters randomly and the computation has been stopped at the point of sufficient convergence after approximately thirty iterations. It took about eight hours per iteration (on a standard PC) which corresponds to the total computing time of about ten days. As a stopping rule for EM iterations we have used the threshold $\epsilon = 10^{-4}$ for the relative increment of the log-likelihood criterion (4). Nevertheless, in final stages of convergence the model accuracy does not change very much and the algorithm can be stopped manually.

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 the known values (cf. Williams 2005). In particular, for each variable separately, we have randomly chosen 10^5 records for which the value of the tested variable was available. Then for each record we have 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-responses and in the fourth column we list the percentage of the correctly estimated values. The number in parentheses corresponds to the trivial global imputation of the most frequent response. As one would expect, 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 using the maximum-likelihood estimate is often considerable (Nos. 1, 3, 8, 9, 10, 13, 15, 17, 19, 23). On the average 73% of missing values would be correctly identified by the maximum-likelihood estimates. The last column contains the number of non-responses from the third column that are expected to be replaced correctly.

4. 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 precisely as possible. In particular, in order to verify the model accuracy, we compare the empirical frequencies 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

$$\mathcal{S}(\boldsymbol{x}_C) = \{\boldsymbol{y} \in \mathcal{S} : \boldsymbol{y}_C = \boldsymbol{x}_C\}, \quad N(\boldsymbol{x}_C) = |S(\boldsymbol{x}_C)|, \quad \boldsymbol{x}_C = (x_{i_1}, \ldots, x_{i_k}) \in \mathcal{X}_C \tag{16}$$

where $\mathcal{S}(\boldsymbol{x}_C)$ is the subset of respondents (a subpopulation) with the property $\boldsymbol{x}_C$ and

List of test combinations:	$\mathcal{A}_4$	$\mathcal{A}_5$
Mean relative error in %:	**4.07**	**4.17**
Standard deviation of the relative error:	6.33	5.80
Maximum relative error of the model in %:	240.84	240.84
Number of relative errors exceeding 100%:	925	4092
Mean absolute error:	**470**	**348**
Standard deviation of the absolute error:	951	655
Maximum absolute error of the model:	45779	45779
Number of combinations tested:	**3468134**	**26425727**

Table 3: Mean relative and mean absolute error of the statistical model with M=15000 components. Results obtained by applying both criteria to the list of properties $\mathcal{A}_5$ (third column) and to the list $\mathcal{A}_4$ of properties specified by a maximum of four responses (second column)

.

$N(\boldsymbol{x}_C)$ 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 $|\mathcal{S}|$:

$$\hat{N}(\boldsymbol{x}_C) = |\mathcal{S}|P(\boldsymbol{x}_C), \quad P(\boldsymbol{x}_C) = \sum_{m=1}^{M} w_m \prod_{j=1}^{k} p_{i_j}(x_{i_j}|m) \tag{17}$$

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 in reproducing small frequencies. On the contrary, the decreasing accuracy of the model at low probabilities is an important confidentiality-protecting property. We decided for this reason 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 a maximum of five responses. As a result we obtained a list $\mathcal{A}_5$ of about 26 million "statistically relevant" properties $\boldsymbol{x}_C$ along with the corresponding empirical frequencies

$$\mathcal{A}_5 = \{\boldsymbol{x}_C = (x_{i_1}, \ldots, x_{i_5}) : N(\boldsymbol{x}_C) > 1612\}, \qquad |\mathcal{A}_5| = 26425727. \tag{18}$$

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 $\boldsymbol{x}_C \in \mathcal{A}_5$:

$$E_a = \frac{1}{|\mathcal{A}_5|} \sum_{x_C \in \mathcal{A}_5} |P(\boldsymbol{x}_C)|\mathcal{S}| - N(\boldsymbol{x}_C)|, \qquad P(\boldsymbol{x}_C) = \sum_{m=1}^{M} w_m \prod_{j=1}^{5} p_{i_j}(x_{i_j}|m) \tag{19}$$

where $P(\boldsymbol{x}_C)$ is the probability of the combination $\boldsymbol{x}_C$ computed by means of the mixture model (2). However, as can be seen, the criterion E_a does not differentiate between errors of large and small estimates. Therefore we have introduced the following mean relative error criterion

$$E_r = \frac{100}{|\mathcal{A}_5|} \sum_{x_C \in \mathcal{A}_5} \frac{|P(\boldsymbol{x}_C) - \frac{N(\boldsymbol{x}_C)}{|\mathcal{S}|}|}{\frac{N(\boldsymbol{x}_C)}{|\mathcal{S}|}} = \frac{100}{|\mathcal{A}_5|} \sum_{x_C \in \mathcal{A}_5} \frac{|P(\boldsymbol{x}_C)|\mathcal{S}| - N(\boldsymbol{x}_C)|}{N(\boldsymbol{x}_C)}, \tag{20}$$

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 3 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 3 shows the mean relative and mean absolute error and the corresponding standard deviations. In addition we have computed the maximum relative and absolute errors and also the number of relative errors exceeding 100%. The mean relative error was 4.2% in the case of the list $\mathcal{A}_5$ and 4.1% in the case of the list $\mathcal{A}_4$; the corresponding absolute error was 338 and 460 respondents, respectively. Since all other results are also comparable, we may assume that a more extensive test would not yield substantially different values. We recall that by combining more than five responses we would mostly obtain 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 4.17%. In order to illustrate the distribution of relative errors in more detail we include Table 4. As it can be seen, for very small empirical frequencies ($1612 < N(\boldsymbol{x}_C) < 3000$) the mean relative error is 6.10% and quickly decreases for greater values of $N(\boldsymbol{x}_C)$ (larger sub-populations). In the case of estimates based on the 10%-subset of microdata (last column) the first value 5.16% is smaller but the distribution of relative errors is similar. 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 that corresponds to a sub-threshold frequency.

Obviously, the results in Table 3 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 $\mathcal{S}$). Note that the standard method of disseminating statistical information by means of subsets of anonymized microdata provides similar comfort and flexibility compared to 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 3. As

Interval	Lower bound	Upper bound	Number of combinations	Mean relative error in %	Mean relative error in %
1	1612	3000	7688027	**6.10**	**5.16**
2	3000	5000	5011625	**4.88**	**3.86**
3	5000	7500	3220931	**4.04**	**3.07**
4	7500	10000	1906156	**3.50**	**2.58**
5	10000	15000	2213787	**3.04**	**2.17**
6	15000	30000	2695817	**2.38**	**1.67**
7	30000	50000	1296118	**1.80**	**1.23**
8	50000	100000	1075615	**1.37**	**0.94**
9	100000	150000	372570	**1.03**	**0.70**
10	150000	300000	358112	**0.78**	**0.55**
11	300000	500000	125103	**0.55**	**0 39**
12	500000	1000000	71104	**0.39**	**0.28**
13	1000000	1500000	15324	**0.29**	**0.20**
14	1500000	3000000	8511	**0.22**	**0.14**
15	3000000	5000000	1349	**0.12**	**0.08**
16	5000000	10300000	200	**0.02**	**0.04**

Table 4: Distribution of relative errors of estimates according to the empirical frequency $N(\boldsymbol{x}_C)$ (sub-population size). Comparison of the statistical model and 10%-subset of microdata. In the first two columns we specify the lower and upper bounds of the frequency intervals, respectively. The third column contains the number of properties falling into the given interval of empirical frequencies. The last two columns contain the corresponding mean relative errors for the statistical model and subset of microdata, respectively.

we can see in Table 5, the accuracy of the 10% microdata subset is marginally better than the statistical model at reproducing the empirical frequencies. However, if the randomly chosen subset of microdata records should be used as a public-use file then it would be necessary to apply some anonymization procedure to the data. Probably, after anonymization, the accuracy of the resulting public use file would decrease due to errors introduced by the anonymization process. In our comparison experiment the anonymization of microdata has been omitted.

5. Model-Based Information Analysis

Another possible way 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 virtual list $\mathcal{A}$ of statistically relevant sub-populations, which can be specified by combining variables (cf. (18)). The general scheme of the considered information analysis can be summarized as follows: we order, e.g., the virtual list $\mathcal{A}_4$ of 3.5 million 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 the ascending ordering of sub-populations (instead of the descending one) could also be of interest. In this section we suggest some criteria which may be useful for different purposes.

A very simple criterion applicable to ordering 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. (9)). By

List of test combinations:	$\mathcal{A}_4$	$\mathcal{A}_5$
Mean relative error in %:	**2.94**	**3.60**
Standard deviation of the relative error:	3.00	3.23
Maximum relative error of the model in %:	35.19	36.34
Number of relative errors exceeding 100%:	0	0
Mean absolute error:	**307**	**409**
Standard deviation of the absolute error:	450	1913
Maximum absolute error of the model:	12348	59815
Number of combinations tested:	**3503448**	**26425727**
Number of microdata in the subset:	1022666	1022666

Table 5: Relative and absolute accuracy of the estimates computed from the randomly chosen subset of 10% of microdata from the original database $\mathcal{S}$. The test has been obtained by applying both criteria to the list of properties $\mathcal{A}_5$ (third column) and to the list $\mathcal{A}_4$ (second column) in the same way as in the Table 3.

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 only 1 or 0.

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

$$P_{nC}(x_n, \boldsymbol{x}_C) = P_{n|C}(x_n|\boldsymbol{x}_C)P(\boldsymbol{x}_C) = \sum_{m=1}^{M} w_m p_n(x_n|m) F_C(\boldsymbol{x}_C|m). \tag{21}$$

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

$$P_{nr|C}(x_n, x_r|\boldsymbol{x}_C) = \sum_{m=1}^{M} W_m(\boldsymbol{x}_C) p_n(x_n|m) p_r(x_r|m). \tag{22}$$

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

$$P_{nrC}(x_n, x_r, \boldsymbol{x}_C) = P_{nr|C}(x_n, x_r|\boldsymbol{x}_C)P(\boldsymbol{x}_C) = \sum_{m=1}^{M} w_m p_n(x_n|m) p_r(x_r|m) F_C(\boldsymbol{x}_C|m) \tag{23}$$

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 conditional probabilities $P_{nr|C}(x_n, x_r|\boldsymbol{x}_C)$ equal to 1 or 0.

In some cases we could be interested in sub-populations where the conditional distribution of a variable is concentrated on an arbitrary single value (or a 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

$$H_{\boldsymbol{x}_C}(\mathcal{X}_n) = \sum_{x_n \in \mathcal{X}_n} -P_{n|C}(x_n|\boldsymbol{x}_C) \log P_{n|C}(x_n|\boldsymbol{x}_C). \tag{24}$$

In other words, in the sub-populations characterized by 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 calculating the relative frequencies.

The statistical model also provides a general possibility of identifying dependence between categorial variables. Recall that the standard tool to characterize a relationship 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 cases of discrete nominal variables like eye color, profession, marital status, etc., the product of two variables is not defined and there is no generally acceptable way to introduce a reasonable definition.

One possibility available for analyzing 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

$$I(X_n, X_r) = H(X_n) + H(X_r) - H(X_n, X_r) \tag{25}$$

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

$$H(X_n) = \sum_{x_n \in X_n} -P_n(x_n) \log P_n(x_n), \qquad P_n(x_n) = \sum_{m=1}^{M} w_m p_n(x_n|m), \quad n \in \mathcal{N}, \tag{26}$$

$$H(X_n, X_r) = \sum_{x_r \in X_r} \sum_{x_n \in X_n} -P_{nr}(x_n, x_r) \log P_{nr}(x_n, x_r), \quad n, r \in \mathcal{N}, \tag{27}$$

$$P_{nr}(x_n, x_r) = \sum_{m=1}^{M} w_m p_n(x_n|m) p_r(x_r|m). \tag{28}$$

The value of 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 (25) can be used, e.g., to order the subpopulation list $\mathcal{A}$ according to the statistical dependence between two chosen variables.

5. Concluding Remarks

To date, one of the most informative known ways 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 individual records) and the indispensable anonymization procedures may negatively influence the statistical validity of the contained information. Moreover, there is always some residual risk of disclosure and, for this reason, statistical agencies release microdata for research purposes only, usually under special license 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 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 evaluation of a virtual list of several hundred thousand 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.

References

[1] Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990). Disclosure Control of Microdata. 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] Williams, T. (2005). The Development of Truth Decks for the 2010 Census Count Imputation Research. In: Proceedings for the Section on Survey Research Methods, American Statistical Association.

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

Appendix I. Probabilistic Inference Mechanism

Note that any marginal distribution of the mixture (2) is easily obtained by deleting superfluous terms in the products $F(\boldsymbol{x}|m)$. Actually, in view of this property, the discrete distribution mixture (2) can be used as a knowledge base of the Probabilistic Expert System (PES). Considering a basic situation, we assume $v_{i_1}, v_{i_2}, \ldots, v_{i_k}$ to be a subset of input variables. Then, for any given input vector

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

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

$$P_C(\boldsymbol{x}_C) = \sum_{m=1}^{M} w_m F_C(\boldsymbol{x}_C|m), \qquad F_C(\boldsymbol{x}_C|m) = \prod_{i \in C} p_i(x_i|m), \quad \boldsymbol{x}_C \in \mathcal{X}_C, \tag{29}$$

$$P_{n,C}(x_n, \boldsymbol{x}_C) = \sum_{m=1}^{M} w_m F_{n,C}(x_n, \boldsymbol{x}_C|m), \qquad F_{n,C}(x_n, \boldsymbol{x}_C|m) = \prod_{i \in C \cup \{n\}} p_i(x_i|m), \tag{30}$$

and for the desired conditional distribution

$$P_{n|C}(x_n|\boldsymbol{x}_C) = \frac{P_{n,C}(x_n, \boldsymbol{x}_C)}{P_C(\boldsymbol{x}_C)} = \sum_{m=1}^{M} W_m(\boldsymbol{x}_C) p_n(x_n|m), \qquad (P_C(\boldsymbol{x}_C) > 0). \tag{31}$$

Here $W_m(\boldsymbol{x}_C)$ are the component weights corresponding to the given input vector $\boldsymbol{x}_C \in \mathcal{X}_C$:

$$W_m(\boldsymbol{x}_C) = \frac{w_m F_C(\boldsymbol{x}_C|m)}{\sum_{j=1}^{M} w_j F_C(\boldsymbol{x}_C|j)}. \tag{32}$$

The conditional distributions $P_{n|C}(x_n|\boldsymbol{x}_C)$ (conditional histograms) describe the statistical properties of the subpopulation specified by the sub-vector $\boldsymbol{x}_C$ in terms of all variables x_n not included in $\boldsymbol{x}_C$.

The interactive inference mechanism of the expert system PES can be extended to a more general case when each of the input variables x_{i_j} is confined to a subset of values

$$x_{i_j} \in \mathcal{D}_{i_j} \subset \mathcal{X}_{i_j}, \quad i_j \in C. \tag{33}$$

Thus, instead of an input sub-vector $\boldsymbol{x}_C \in \mathcal{X}_C$, we are given a subset of input vectors $\mathcal{D}_C \subset \mathcal{X}_C$:

$$\mathcal{D}_C = \mathcal{D}_{i_1} \times \mathcal{D}_{i_2} \times \ldots \times \mathcal{D}_{i_k} \subset \mathcal{X}_C. \tag{34}$$

Analogously to Eqs. (29)–(30), we can write

$$F_C(\mathcal{D}_C|m) = \sum_{x_C \in \mathcal{D}_C} F_C(\boldsymbol{x}_C|m) = \sum_{x_{i_1} \in \mathcal{D}_{i_1}} \ldots \sum_{x_{i_k} \in \mathcal{D}_{i_k}} \prod_{i_j \in C} p_{i_j}(x_{i_j}|m) = \prod_{i_j \in C} p_{i_j}(\mathcal{D}_{i_j}|m), \tag{35}$$

and finally we can compute the conditional probability $P_{n|C}(x_n|\mathcal{D}_C)$ by Eq.

$$P_{n|C}(x_n|\mathcal{D}_C) = \frac{P_{n,C}(x_n, \mathcal{D}_C)}{P_C(\mathcal{D}_C)} = \sum_{m=1}^{M} W_m(\mathcal{D}_C) p_n(x_n|m). \tag{36}$$

Here $W_m(\mathcal{D}_C)$ are the component weights corresponding to the given subset of input vectors $\mathcal{D}_C \subset \mathcal{X}_C$:

$$W_m(\mathcal{D}_C) = \frac{w_m F_C(\mathcal{D}_C|m)}{\sum_{j=1}^{M} w_j F_C(\mathcal{D}_C|j)}. \tag{37}$$

The conditional distribution (31) represents an exact response to the definite input $\boldsymbol{v}_C = \boldsymbol{x}_C \in \mathcal{X}_C$. In a case of uncertain input information which is generally described by the probability distribution $\tilde{P}_C(\boldsymbol{x}_C)$ on the subspace of input variables $\mathcal{X}_C$, we obtain

$$\tilde{P}_n(x_n) = \sum_{\mathbf{x}_C \in \mathbf{X}_C} P_{n|C}(x_n|\mathbf{x}_C) \tilde{P}_C(\mathbf{x}_C) = \sum_{m=1}^{M} \tilde{W}_m p_n(x_n|m), \tag{38}$$

where

$$\tilde{W}_m = \sum_{\mathbf{x}_C \in \mathbf{X}_C} W_m(\mathbf{x}_C) \tilde{P}_C(\mathbf{x}_C). \tag{39}$$

Let us remark that the formula (38) realizes the so-called memoryless information channel with noise – a well known object of information theory.

The processing of uncertain information is a typical feature of expert systems as decision-supporting tools. In the present context, the uncertain information on input

can be used to analyse the properties of hypothetical subsets or sub-populations. Note that the choice of input and output variables is affected only by their meaning or availability – without any formal restrictions implied by the knowledge base. Thus, unlike the rule-based systems, the expert system PES is fully symmetrical with respect to the role of variables.

Appendix II. Statistical Validity of Census Data

The standard census is a statistical investigation of extreme extent that includes the entire population. All individuals have to answer a set of questions and for any combination of possible answers we can specify the exact number of respondents having the corresponding property. This number is fixed and uniquely given by the current state of the population. Nevertheless, for theoretical reasons, any statistical property of the census population specified by means of empirical frequency has a limited validity.

Obviously, a census is unique and cannot be repeated as a random experiment under identical conditions. On the other hand, the census questionnaire can be viewed formally as a vector of discrete finite-valued random variables

$$\boldsymbol{v} = (v_1, v_2, \ldots v_N), \quad v_n \in \mathcal{X}_n,$$

and every respondent can be assumed to provide an independent observation $\boldsymbol{x}$ of this random vector

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_N) \in \mathcal{X}; \quad \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_N.$$

From a theoretical point of view, statistical properties of a random vector can be described in full generality by a joint probability distribution $P^*(\boldsymbol{x})$ of the involved random variables. We can assume that for a given property specified by a combination of values $\boldsymbol{x}_C = (x_{i_1}, \ldots, x_{i_k})$ there is an unknown "hypothetical" probability

$$p^* = P^*(\boldsymbol{x}_C), \quad \boldsymbol{x}_C = (x_{i_1}, \ldots, x_{i_k}) \in \mathcal{X}_{i_1} \times \mathcal{X}_{i_2} \times \ldots \times \mathcal{X}_{i_k} = \mathcal{X}_C \tag{40}$$

of this property. The probability $P^*(\boldsymbol{x}_C)$ represents a statistical property of the sub-population characterized by $\boldsymbol{x}_C$ and can be estimated by means of the corresponding relative frequency. Typically, speaking about the statistical properties, we refer to large sub-populations; let us recall that a statistical property becomes questionable if the related sub-population is small.

We can formally view the number of respondents s_n satisfying the considered property $\boldsymbol{x}_C$ as a random result of n independent observations of a random event (so-called Bernoulli trials) with probability $p^* = P^*(\boldsymbol{x}_C)$. We recall that, in the case of a census, the number s_n is fixed and uniquely defined and $P^*(\boldsymbol{x}_C)$ is a hypothetical and unknown probability which cannot be verified by repeated experiments. Despite these facts, the relationship between the observed number of respondents s_n with the property $\boldsymbol{x}_C$, the population size n and the probability $P^*(\boldsymbol{x}_C)$ can be used to characterize the reliability of the empirical frequency s_n. In other words we can decide whether the statistical property is statistically significant or not. We recall first the well known De Moivre–Laplace limit theorem of probability in the following simple form (cf. Feller, 1962):

Theorem. Let s_n stands for the number of successes in n Bernoulli trials with probability $p^*, (0 < p^* < 1)$ for success. Then for each fixed $\alpha, (0 < \alpha < 1)$ the probability

$$P\{|\frac{s_n}{n} - p^*| < \alpha\} = P\{|\frac{s_n - np^*}{\sqrt{np^*(1-p^*)}}| < \alpha\sqrt{\frac{n}{p^*(1-p^*)}}\} \tag{41}$$

satisfies the relation

$$\lim_{n\to\infty}\left[P\{|\frac{s_n}{n}-p^*|<\alpha\}-\frac{2}{\sqrt{2\pi}}\int_0^{\alpha\sqrt{\frac{n}{p^*(1-p^*)}}}\exp\left(-\frac{z^2}{2}\right)\,dz\right]=0. \tag{42}$$

•

This Theorem provides an asymptotic approximation of the probability that the observed relative frequency $\frac{s_n}{n}$ of a random event and the related unknown probability p^* differ from each other by less than a fixed upper bound $\alpha, (0<\alpha<1)$. From the Theorem it follows that for large values of n the error bound probability (41) can be approximated by the integral expression:

$$P\{|\frac{s_n}{n}-p^*|<\alpha\}\approx\frac{2}{\sqrt{2\pi}}\int_0^{\alpha\sqrt{\frac{n}{p^*(1-p^*)}}}\exp\left(-\frac{z^2}{2}\right)\,dz. \tag{43}$$

In particular, if we intend to clarify the conditions implying that the probability (41) is high enough, e.g., at the confidence level 0.95, we have to guarantee the inequality

$$P\{|\frac{s_n}{n}-p^*|<\alpha\}\geq 0.95,\quad (0<\alpha<1). \tag{44}$$

In other words, from inequality (44) it follows that (at the confidence level 0.95) the unknown true probability p^* belongs to the interval

$$\frac{s_n}{n}-\alpha<p^*<\frac{s_n}{n}+\alpha \tag{45}$$

or, equivalently, the relative frequency $\frac{s_n}{n}$ falls between the bounds

$$p^*-\alpha<\frac{s_n}{n}<p^*+\alpha. \tag{46}$$

In view of the limit theorem we can analyze, for large values of n, the inequality

$$\frac{2}{\sqrt{2\pi}}\int_0^{\alpha\sqrt{\frac{n}{p^*(1-p^*)}}}\exp\left(-\frac{z^2}{2}\right)\,dz\geq 0.95. \tag{47}$$

The integral on the left hand side (sometimes called the error function) is an increasing function of the upper bound. It can be proven that inequality (47) is satisfied if it holds:

$$\alpha\sqrt{\frac{n}{p^*(1-p^*)}}\geq 1.96. \tag{48}$$

Now, as we are interested in the relative accuracy of estimates, we choose the threshold α in the dependence of p^*, e.g., by setting $\alpha=ap^*, (0<a<1)$. Here $a=\frac{\alpha}{p^*}$ is the required relative accuracy of the probability p^*. Substituting for α in (46) we can write

$$\frac{|\frac{s_n}{n}-p^*|}{p^*}=\frac{|s_n-np^*|}{np^*}<a \tag{49}$$

or equivalently

$$(1-a)np^*<s_n<(1+a)np^*. \tag{50}$$

Making substitution for α in (48), we obtain

$$a\sqrt{\frac{np^*}{(1-p^*)}} \geq 1.96 \tag{51}$$

and further we obtain the lower bound for the probability p^* as a function of its required accuracy a:

$$p^* \geq \frac{(1.96)^2}{(1.96)^2 + a^2 n}. \tag{52}$$

In other words, if we want to guarantee the relative accuracy $a, (0 < a < 1)$ then the estimated probability p^* must be greater than the expression on the right hand side. In particular, for the czech census 2001 we can fix the number of respondents to $n = 10230060$ and, by making substitution, we can describe the underlying relation between p^* and a by Table 6.

a	0.01	0.02	0.05	0.10
$p^* >$	0.0037412	0.0009379	0.0001502	0.0000376
$np^* >$	38272	9595	1536	384

Table 6: Relationship between the relative accuracy a of the probability p^* of a specific property of respondents, and of the related subpopulation size np^* (at the confidence level 0.95), in the case of a census population of size 10230060.

If we choose the "admissible" relative error $a = 0.05$ then, for the true unknown frequency satisfying the inequality

$$np^* > 1536, \tag{53}$$

the corresponding empirical frequency s_n is bounded by the inequality (cf. (49))

$$0.95np^* < s_n < 1.05np^*. \tag{54}$$

From inequality (54) it follows that, in turn, the unknown frequency np^* is bounded by means of the empirical frequency s_n:

$$\frac{s_n}{1.05} < np^* < \frac{s_n}{0.95}. \tag{55}$$

However, the last inequality holds only if $np^* > 1536$. As the true unknown frequencies np^* are not available, we consider only the empirical frequencies s_n which guarantee the above condition (53). In particular, we confine ourselves to the empirical frequencies s_n satisfying the inequality

$$\frac{s_n}{1.05} > 1536 \tag{56}$$

which implies the condition (53)

$$np^* > \frac{s_n}{1.05} > 1536. \tag{57}$$

Consequently, if the empirical frequency s_n is greater than 1612:

$$s_n > 1.05 * 1536 = 1612.8, \tag{58}$$

then the unknown true frequency np^* is greater than 1536 (cf. (53)) and cannot differ from s_n by more than 5%. We can conclude that, by using the lower bound $N_{0.05} = 1612$, we guarantee the 5% accuracy of the empirical frequencies s_n (at the confidence level 0.95) for the census population size $|\mathcal{S}| = 10230060$.