Interaktivní statistický model dat ze sčítání lidu v České republice v r. 2001

Jiří Grim, Jan Hora, Pavel Boček, Petr Somol, Pavel Pudil
Ústav teorie informace a automatizace AV ČR,
Fakulta jaderná a fyzikálně inženýrská ČVUT, Fakulta managementu VŠE

1. Problém publikace výsledků sčítání lidu

Sčítání lidu se provádí opakovaně v desetiletém intervalu ve většině zemí EU. Význam sčítání lidu je často předmětem diskusí vzhledem k vysokým nákladům a omezené dostupnosti výsledků. Podle zákona o ochraně dat je stát povinen chránit zjištěné osobní údaje občanů. Základní identifikační údaje osob jsou proto odděleny od dotazníků již v počáteční fázi zpracování. Nicméně, v některých případech by respondent přesto mohl být identifikován podle svých odpovědí s využitím obecně známých informací [5], [20]. Vzhledem k potenciální možnosti reidentifikace respondentů je databáze dotazníků uložena za přísných bezpečnostních opatření a přístup k původním datům je velmi omezen. Český statistický úřad obsáhle publikuje výsledky sčítání lidu v tištěné podobě i na svých internetových stránkách, zpravidla ve formě tabulek pro dvojice ukazatelů, případně podmíněných dalšími údaji. Celkový počet tabulek, které lze odvodit z původních dat, bohužel rychle roste s počtem podmiňujících údajů. Bez ohledu na rozsah publikačních možností nutně zůstane velká část tabulek nedostupná i když by mohla být pro některé uživatele zajímavá.

Existují různé možnosti zpracování a využívání uložených statistických dat. Všeobecně se předpokládá, že nejdokonalejší přístup k výsledkům sčítání lidu umožňují tzv. anonymizované soubory mikrodat. Jde o náhodně vybrané podsoubory dotazníků, které při velikosti zhruba 10^6 záznamů s dostatečnou přesností reprodukují statistické vlastnosti původních dat a pomocí vhodného databázového programu z nich mohou být odvozeny stejné tabulky, jako z původního úplného souboru. Výhodou mikrodat je rychlá dostupnost různorodých informací. Pomocí reprezentativního souboru mikrodat může uživatel snadno ověřovat neobvyklé hypotézy a vyhledávat nová témata mimo rámec obvyklé nabídky statistických agentur. Soubory mikrodat je třeba před vydáním uživateli upravovat pomocí různých technik [20], [19] a s využitím speciálních metod [3],[6], s cílem znemožnit případnou identifikaci respondentů. Vytvoření anonymizovaného souboru mikrodat je pracné přičemž přesnost údajů, které z něj lze odvodit, přirozeně klesá s jeho velikostí a závisí také na míře znehodnocení dat, způsobené ochrannými anonymizačními postupy. Anonymizované soubory mikrodat jsou běžnou součástí nabídky statistických úřadů [15] nicméně, s ohledem na zbytkové bezpečnostní riziko, jsou obvykle poskytovány pouze pro výzkumné účely na základě speciálních licenčních podmínek.

[0]S částečnou podporou grantu GAČR 102/07/1594 a projektu MŠMT 1M0572 DAR.

Interaktivní reprodukce výsledků sčítání lidu pomocí statistického modelu nabízí v této souvislosti alternativní publikační možnost s dokonale zabezpečenou ochranou anonymity dat. Základem metody je odhad statistického modelu původní databáze ve tvaru diskrétní součinové směsi, která je následně použita jako báze znalostí pravděpodobnostního expertního systému [8], [10]. Interaktivní statistický model umožňuje odvozování libovolně podmíněných tabulek resp. histogramů s uživatelským komfortem, který je srovnatelný nebo lepší než v případě souboru mikrodat. Odhadnutá distribuční směs neobsahuje původní data, takže výsledný interaktivní softwarový produkt může být zpřístupněn všem uživatelům bez jakéhokoli omezení.

Metoda interaktivní reprodukce statistických vlastností dat pomocí směsového modelu je výsledkem více než patnáctiletého vývoje v rámci spolupráce mezi Ústavem teorie informace a automatizace AV ČR a Fakultou managementu VŠE, Jindřichův Hradec. Princip metody byl poprvé publikován na konferenci "Eleventh European Meeting on Cybernetics and Systems Research, Vienna 1992"v práci [9] odměněné cenou "F. de P. Hanika Memorial Award". Navržený postup se opírá o původní výsledky v oblasti pravděpodobnostních expertních systémů [8], [9], [10], byl prakticky ověřován na databázi pražských domácností ze sčítání lidu v roce 1991 [11] a úspěšně prezentován na mezinárodních konferencích [12], [13].

V této práci je popsán výpočet statistického modelu dat ze sčítání lidu, domů a bytů v České republice v roce 2001. Výsledný produkt je volně k dispozici na internetové adrese *http://ro.utia.cas.cz/dem.html*[1]. Projekt byl realizován na základě smlouvy o spolupráci, uzavřené v červnu 2008 mezi předsedou ČSÚ J. Fischerem a rektorem VŠE R. Hindlsem v rámci příprav sčítání lidu v roce 2011.

2. Statistický model dat ve tvaru součinové směsi

Sčítání lidu představuje unikátní jednorázové statistické šetření, které není opakovatelné jako náhodný experiment. Na druhé straně, statistický dotazník vyplněný respondentem můžeme do značné míry oprávněně považovat za nezávislou realizaci N-tice náhodných proměnných $\boldsymbol{v} = (v_1, v_2, \ldots, v_N)$, které nabývají konečný počet hodnot podle otázek dotazníku. Vyplněný dotazník může být zapsán ve tvaru diskrétního datového vektoru

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_N) \in \mathcal{X}, \quad x_n \in \mathcal{X}_n, \quad \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N, \tag{1}$$

kde $\mathcal{X}_n$ označuje konečnou množinu možných odpovědí x_n na n-tou otázku.

Pro účely výpočtu statistického modelu dat ze sčítání lidu z r. 2001 jsme zvolili 24 otázek, tak jak jsou uvedeny v Tabulce 1. Výběr otázek byl veden snahou modelovat v co nejširší míře sociologicky a ekonomicky zajímavé vztahy při zachování přijatelné formální složitosti statistických závislostí. Z tohoto důvodu jsme v některých případech použili méně podrobné kódování otázek (např. územní členění, věkové skupiny) a vynechali jsme otázky s příliš jednoznačným rozložením odpovědí (např. národnost). Prvních deset otázek bylo převzato z dotazníku pro sčítání osob

[1] Zjednodušená verze programu (s jednodušším modelem) je k dispozici ve formě appletu v jazyce "Java"na adrese: http://simu0292.utia.cas.cz/census/

a zbývajících čtrnáct otázek se vztahuje k dotazníku domácnosti, v níž respondent žije. Četnosti údajů o bytech mají v takto vzniklém souboru nový smysl, udávají počet respondentů v bytech resp. domácnostech s danou vlastností. Z výsledného modelu tak lze odvodit řadu nových závěrů, protože vlastnosti obou databází byly dosud publikovány většinou odděleně.

	Text otázky	Počet odpovědí	Nezjištěné údaje v %	Entropie odpovědí v %
1.	Kraj	14	0.00	96.88
2.	Druh pobytu	3	0.00	32.92
3.	Ekonomická aktivita	10	0.80	67.80
4.	Místo narození - relativně	6	1.95	74.65
5.	Náboženské vyznání	6	0.00	60.57
6.	Odvětví práce	14	3.89	68.33
7.	Pohlaví	2	0.00	99.95
8.	Rodinný stav	4	0.55	81.01
9.	Stupeň vzdělání	14	1.11	78.04
10.	Věk	9	0.03	96.09
11.	Kategorie bytu	5	0.53	27.81
12.	Koupelna	5	0.59	14.02
13.	Obytná plocha bytu	7	0.64	80.62
14.	Osobní počítač a internet	4	2.85	49.11
15.	Právní důvod k užívání bytu	9	0.39	72.43
16.	Plyn v bytě	3	0.78	64.54
17.	Počet místností nad 8m	7	0.64	80.57
18.	Počet aut v domácnosti	4	3.39	71.32
19.	Počet osob v bytě	6	0.00	93.79
20.	Rekreační objekt	6	7.45	42.10
21.	Telefon	5	1.80	80.88
22.	Vodovod	4	0.35	8.02
23.	Způsob vytápění	6	0.53	74.81
24.	Záchod	6	0.50	16.73

Tabulka 1: Seznam otázek zvolených pro výpočet statistického modelu dat ze sčítání lidu, domů a bytů v České republice z r. 2001 s počtem možných odpovědí, procentem chybějících údajů a celkovou neurčitostí (entropií) odpovědí.

V třetím sloupci tabulky 1 je uveden počet možných odpovědí pro danou otázku. Četnost chybějících údajů v procentech je uvedena ve čtvrtém sloupci. Připomeňme, že v případě bytů se údaj "nezjištěno"objeví v dotaznících všech členů domácnosti. Celkový počet chybějících údajů je 2933427. Poslední sloupec udává neurčitost otázky v procentech maximální entropie, tj. jako poměr entropie rozložení četnosti odpovědí a maximální entropie příslušného rovnoměrného rozložení. Nejvyšší neurčitost 99.95% má otázka č.7 udávající dvě téměř stejně četné možnosti pohlaví respondenta. Nejmenší neurčitost má velmi jednoznačná odpověď na otázku č. 22 o vybavení bytu vodovodem.

Sčítání lidu představuje rozsáhlé statistické šetření zahrnující celou populaci.

Konkrétně, databáze osob tvořící základ souboru $\mathcal{S}$, obsahovala dotazníky celkem 10230060 respondentů.

$$\mathcal{S} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(K)}\}, \quad \boldsymbol{x}^{(k)} \in \mathcal{X}, \quad (K = 10230060). \tag{2}$$

Připomeňme, že veškerá informace o statistických vztazích mezi náhodnými proměnnými (otázkami dotazníku) je popsána mnohorozměrným rozložením pravděpodobnosti $P^*(\boldsymbol{x}) = P\{\boldsymbol{v} = \boldsymbol{x}\}$ diskrétního náhodného vektoru $\boldsymbol{v}$. Pro účely odhadu statistického modelu předpokládáme neznámé rozložení pravděpodobnosti $P^*(\boldsymbol{x})$ ve tvaru diskrétní distribuční směsi součinových komponent

$$P(\boldsymbol{x}) = \sum_{m=1}^{M} w_m F(\boldsymbol{x}|m), \qquad F(\boldsymbol{x}|m) = \prod_{n=1}^{N} p_n(x_n|m), \quad \boldsymbol{x} \in \mathcal{X}, \quad \sum_{m=1}^{M} w_m = 1, \tag{3}$$

kde M je počet komponent, w_m jsou pravděpodobnostní váhy komponent a $p_n(x_n|m)$ označují příslušné jednorozměrné diskrétní distribuce. Pro danou proměnnou n a komponentu m je distribuce $p_n(x_n|m)$ určena vektorem pravděpodobností jednotlivých údajů z množiny $\mathcal{X}_n$, tj. platí

$$\sum_{\xi \in \mathcal{X}_n} p_n(\xi|m) = 1, \quad n = 1, 2, \ldots, N, \quad m = 1, 2, \ldots, M. \tag{4}$$

Distribuční směs (3) umožňuje jednoduché odvození libovolného marginálního rozložení pravděpodobnosti. Konkrétně, nechť $C = \{i_1, i_2, \ldots, i_k\}$ je nějaká podmnožina indexů proměnných. Potom pro odpovídající část vektoru $\boldsymbol{x}$

$$\boldsymbol{x}_C = (x_{i_1}, x_{i_2}, \ldots, x_{i_k}) \in \mathcal{X}_C$$

můžeme přímo zapsat výraz pro příslušné marginální rozložení pravděpodobnosti:

$$P_C(\boldsymbol{x}_C) = \sum_{m=1}^{M} w_m F_C(\boldsymbol{x}_C|m), \qquad F_C(\boldsymbol{x}_C|m) = \prod_{i \in C} p_i(x_i|m), \quad \boldsymbol{x}_C \in \mathcal{X}_C, \tag{5}$$

a také pro podmíněné rozložení pravděpodobnosti (podmíněný histogram) libovolné proměnné $x_n, (n \notin C)$ (srv. [14]):

$$P_{n|C}(x_n|\boldsymbol{x}_C) = \sum_{m=1}^{M} q(m|\boldsymbol{x}_C) p_n(x_n|m), \qquad q(m|\boldsymbol{x}_C) = \frac{w_m F_C(\boldsymbol{x}_C|m)}{\sum_{j=1}^{M} w_j F_C(\boldsymbol{x}_C|j)}. \tag{6}$$

Poznamenejme, že možnost jednoduchého výpočtu podmíněného histogramu $P_{n|C}(x_n|\boldsymbol{x}_C)$ je dána součinovým tvarem komponent distribuční směsi (3).

Standardní metodou odhadu parametrů distribuční směsi je EM algoritmus [1], [7], [16], [17], který maximalizuje věrohodnostní funkci

$$L = \frac{1}{K} \sum_{k=1}^{K} \log P(\boldsymbol{x}^{(k)}) = \frac{1}{|S|} \sum_{x \in \mathcal{S}} \log[\sum_{m=1}^{M} w_m F(\boldsymbol{x}|m)] \tag{7}$$

pomocí iteračních rovnic ($m = 1, \ldots, M,\ n = 1, \ldots, N,\ \boldsymbol{x} \in \mathcal{S},\ t = 0, 1, \ldots$):

$$q^{(t)}(m|\boldsymbol{x}) = \frac{w_m^{(t)} \prod_{n=1}^{N} p_n^{(t)}(x_n|m)}{\sum_{j=1}^{M} w_j^{(t)} \prod_{n=1}^{N} p_n^{(t)}(x_n|j)}, \qquad w_m^{(t+1)} = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} q^{(t)}(m|\boldsymbol{x}), \tag{8}$$

$$p_n^{(t+1)}(\xi|m) = \frac{1}{\sum_{x\in\mathcal{S}} q^{(t)}(m|\boldsymbol{x})} \sum_{x\in\mathcal{S}} \delta(\xi, x_n) q^{(t)}(m|\boldsymbol{x}), \quad \xi \in \mathcal{X}_n, \tag{9}$$

kde $|\mathcal{S}| = K$ je počet nezávislých pozorování (datových vektorů) a $\delta(\xi, x_n)$ je tzv. delta-funkce. [2] Počet komponent směsi M není předmětem výpočtu a musí být zadán předem, stejně jako počáteční hodnoty parametrů směsi $w_m^{(0)}$ a $p_n^{(0)}(x_n|m)$. Iterační vztahy (8)-(9) generují neklesající posloupnost hodnot kriteria (7), která konverguje k lokálnímu nebo globálnímu maximu (resp. sedlovému bodu). Podrobnou diskusi různých výpočetních aspektů EM algoritmu lze nalézt např v monografii [16].

Hlavním účelem statistického modelu ve tvaru distribuční směsi (3) je reprodukce (relativních) četností platných v původním souboru $\mathcal{S}$. Důležitou předností EM algoritmu (8) - (9) je v této souvislosti shoda jednorozměrných marginál odhadovaného modelu se skutečnými relativními četnostmi. Konkrétně, v každé iteraci platí:

$$P_n^{(t+1)}(\xi) = \sum_{m=1}^{M} w_m^{(t+1)} p_n^{(t+1)}(\xi|m) = \sum_{m=1}^{M} \frac{1}{|\mathcal{S}|} \sum_{x\in\mathcal{S}} q^{(t)}(m|\boldsymbol{x})\; p_n^{(t+1)}(\xi|m) = \tag{10}$$

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

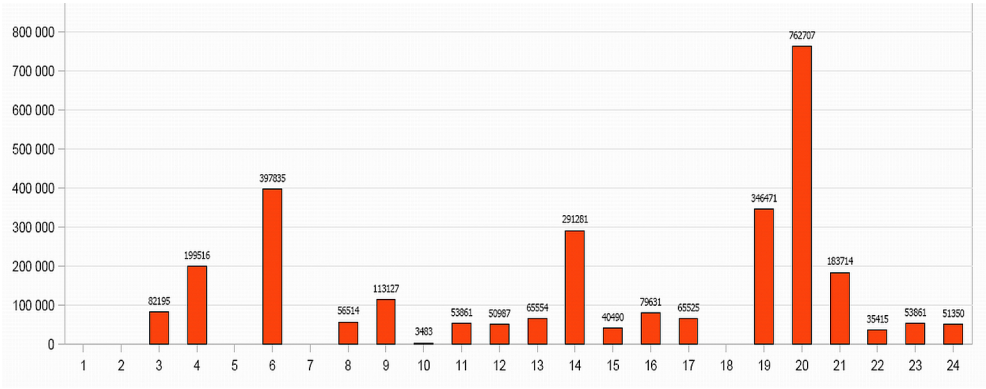
3. Problém chybějících údajů

Typickou vlastností výsledků sčítání lidu jsou neúplná data. Zákon o sčítání lidu se týká celé populace bez možnosti výběru a v jednotlivých případech lze očekávat různou míru dostupnosti potřebných údajů. V oficiálních publikacích ČSÚ je proto v tabulkách vždy uváděn počet chybějících údajů s označením "nezjištěno", tzn. chybějící údaj by bylo možné u každé otázky formálně považovat za další alternativní možnost odpovědi. Na druhé straně odpověď "nezjištěno"je obtížně interpretovatelná, protože může mít řadu různých i když zajímavých důvodů. Při výpočtu modelu by zvýšený počet odpovědí zbytečně zvyšoval složitost statistických vztahů a proto byla odpověď "nezjištěno"vždy interpretována pouze jako chybějící údaj (s vyjímkou náboženského vyznání, kde lze odpověď "nezjištěno"do značné míry považovat za odmítnutí).

Datový soubor vytvořený na základě otázek z Tab. 1 obsahoval celkem 1524240 neúplných dotazníků, tj. přibližně 15%. Obr. 1 udává počty chybějících odpovědí pro jednotlivé otázky. Celkový počet chybějících údajů je 2933427. Rozložení četnosti dotazníků podle počtu chybějících údajů ukazuje Obr. 2.

Problematika zpracování resp. doplňování vícerozměrných dat s chybějícími údaji představuje důležitou oblast matematické statistiky, protože většina statistických metod se nedá použít na neúplná data. Připomeňme, že např. prostým vynecháním neúplných datových vektorů by se v našem případě zmenšil soubor o 15% a po vynechání proměnných s chybějícími údaji by se jejich počet zredukoval na pět (viz Tab. 1).

[2] $\delta(\xi, x_n) = 1$ pro $\xi = x_n$ a $\delta(\xi, x_n) = 0$ pro $\xi \neq x_n$



Obrázek 1: Četnost chybějících odpovědí pro jednotlivé otázky. Celkový počet neúplných dotazníků: 1524240.

Významnou předností EM algoritmu pro odhad parametrů součinové distribuční směsi je možnost přímého zpracování neúplných dat. Označíme-li $\mathcal{N}(\boldsymbol{x})$ podmnožinu indexů proměnných, pro které jsou definovány složky vektoru $\boldsymbol{x}$ a $\mathcal{S}_n \subset \mathcal{S}$ podmnožinu vektorů s definovanou složkou x_n

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

potom upravenou verzi EM algoritmu pro neúplná data můžeme zapsat ve tvaru ($m = 1, 2, \ldots, M,\ n = 1, 2, \ldots, N,\ \boldsymbol{x} \in \mathcal{S},\ t = 0, 1, \ldots$):

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

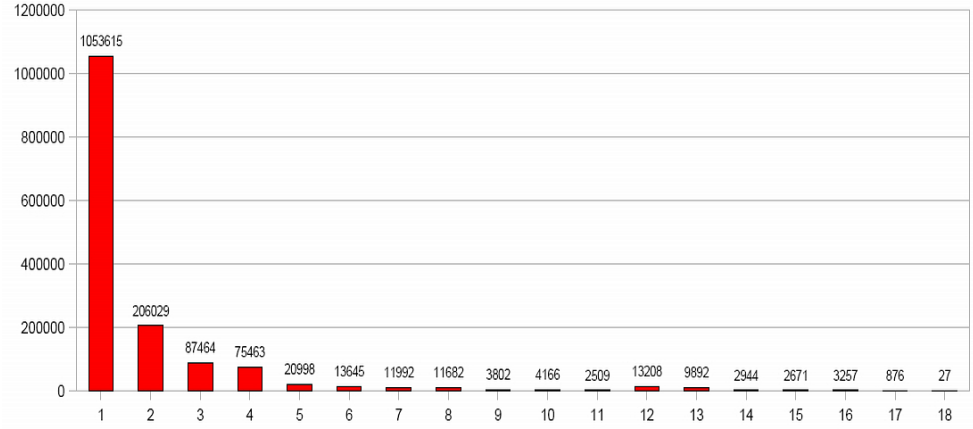
$$p_n^{(t+1)}(\xi|m) = \frac{1}{\sum_{x \in \mathcal{S}_n} q^{(t)}(m|\boldsymbol{x})} \sum_{x \in \mathcal{S}_n} \delta(\xi, x_n) q^{(t)}(m|\boldsymbol{x}), \quad \xi \in \mathcal{X}_n, \tag{13}$$

Jednoduše řečeno, výpočet hodnot $q^{(t)}(m|\boldsymbol{x})$ a $p_n^{(t+1)}(\xi|m)$ se ve vzorcích (12), (13) provádí vždy pouze pro složky vektoru $\boldsymbol{x}$, které jsou v daném případě k dispozici.

Uvedený způsob zpracování neúplných dat se z teoretického hlediska jeví jako lepší než nahrazování chybějících údajů pomocí odhadů, protože využívá pouze informace, která je v datech k dispozici. Doplňování chybějících údajů jakkoli dokonalým způsobem nutně snižuje původní variabilitu dat, protože odhadované údaje mají vždy z principu vyšší pravděpodobnost a menší rozptyl, než chybějící skutečná data.

Bohužel, nevýhodou modifikovaného EM algoritmu je porušení shody jednorozměrných marginál s příslušnými relativními četnostmi. Snadno lze ověřit, že iterační rovnice EM algoritmu upravené pro zpracování neúplných dat (srv. (12), (13)) nesplňují podmínky (10):

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



Obrázek 2: Četnost neúplných dotazníků podle počtu chybějících údajů. Celkový počet chybějících údajů: 2933427.

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

takže model odhadnutý z neúplných dat je zřejmě zatížen značnou chybou již na úrovni nepodmíněných jednorozměrných marginál. Výpočetní experimenty ukázaly, že model odhadnutý z úplného datového souboru je asi dvakrát přesnější než model odhadnutý z neúplných dat pomocí upraveného EM algoritmu.

Je zřejmé, že problém chybějících údajů má z hlediska přesnosti statistického modelu základní význam. Připomeňme, že cílem projektu bylo především ověření možnosti reprodukce statistických vlastností "ideální", tj. úplné databáze, chybějící údaje představují z tohoto hlediska samostatný problém. Z těchto důvodů byl výpočet modelu řešen ve dvou krocích. Parametry distribuční směsi byly nejprve odhadnuty z neúplných dat pomocí modifikovaného EM algoritmu a vypočtený model získaný z neúplných dat byl dále použit pro odhad a doplnění chybějících údajů. Doplňován byl vždy nejpravděpodobnější údaj na základě podmíněného rozložení pravděpodobnosti $p_n(\xi|\boldsymbol{x}_C$ pro známou část dotazníku $\boldsymbol{x}_C$ (srv. (6)):

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

Správnost nahrazení chybějících údajů samozřejmě není možné ověřit. Nicméně, průměrnou přesnost doplňování lze poměrně přesně vyhodnotit v samostatném experimentu odhadováním známých údajů. Pro danou proměnnou náhodně vybíráme dotazníky, pro které je tento údaj znám, vypočítáme odhad tohoto údaje podle vztahu (15) a výsledek porovnáme se skutečnou hodnotou. Uvedeným způsobem můžeme postupně odhadnout přesnost nahrazování pro všechny proměnné. V tabulce 2 je u každé proměnné uveden skutečný počet chybějících údajů, procento úspěšnosti při nahrazování a odhadovaný počet správně doplněných údajů z třetího sloupce. Údaj v závorce udává úspěšnost globálního jednotného nahrazení

N	Text otázky	Nezjištěné údaje	Úspěšnost odhadu v %	Správně odhadnuto
1.	Kraj	0	27.49 (12.41)	0
2.	Druh pobytu	0	90.35 (89.48)	0
3.	Ekonomická aktivita	82195	88.02 (44.08)	72348
4.	Místo narození - relativně	199516	56.36 (53.52)	112447
5.	Náboženské vyznání	0	66.27 (59.04)	0
6.	Odvětví práce	397835	67.64 (50.62)	269096
7.	Pohlaví	0	67.91 (51.30)	0
8.	Rodinný stav	56514	82.91 (46.63)	46856
9.	Stupeň vzdělání	113127	48.36 (19.29)	54708
10.	Věk	3483	59.22 (16.71)	2063
11.	Kategorie bytu	53861	97.48 (89.37)	52504
12.	Koupelna	50987	98.90 (95.91)	50426
13.	Obytná plocha bytu	65554	63.22 (38.48)	41443
14.	Osobní počítač a internet	291281	81.12 (79.15)	236287
15.	Právní důvod k užívání bytu	40490	63.49 (39.70)	25707
16.	Plyn v bytě	79631	75.94 (63.84)	60472
17.	Počet místností nad 8m	65525	63.48 (38.76)	41595
18.	Počet aut v domácnosti	346471	66.97 (51.77)	232032
19.	Počet osob v bytě	0	49.48 (29.27)	0
20.	Rekreační objekt	762707	80.39 (78.11)	613140
21.	Telefon	183714	57.36 (43.93)	105378
22.	Vodovod	35415	99.39 (98.08)	35199
23.	Způsob vytápění	53861	76.90 (41.45)	41419
24.	Záchod	51350	97.98 (94.32)	50313
	Celkem	**2 933 427**	**73.06 (61.35)**	**2 143 326**

Tabulka 2: Odhad úspěšnosti nahrazovaní chybějících údajů. V třetím sloupci je uveden počet chybějících odpovědí. Celkový počet chybějících údajů: 2933427. Procento úspěšnosti nahrazování chybějících údajů pro jednotlivé otázky je uvedeno ve čtvrtém sloupci. Poslední sloupec udává odhadovaný počet správně doplněných údajů z třetího sloupce.

všech chybějících hodnot dané proměnné příslušnou nejčastější odpovědí. Přesnost globálního nahrazení je v některých případech srovnatelná s predikcí podle modelu (č. 2,4,5,12,14,20,22), ale u řady otázek je maximálně věrohodný odhad podstatně přesnější (č. 1,3,8,9,19,13,15,17,19,23). V případě Tab. 2 by v průměru 73% údajů bylo podle modelu nahrazeno správně (při globálním nahrazování 61.3%).

Při dané velikosti souboru je doba výpočtu úměrná počtu komponent distribuční směsi a představuje hlavní výpočetní omezení. Konečný výpočet statistického modelu s počtem komponent M=15000 byl proveden pomocí souboru s doplněnými údaji. Počáteční hodnoty parametrů byly generovány náhodně. Výpočet (30 iterací) trval na běžném PC asi 10 dní při 8 hodinách na jednu iteraci.

4. Přesnost reprodukce statistických vlastností dat

V případě dat ze sčítání lidu jsou statistické vlastnosti respondentů zpravidla určovány kombinací několika odpovědí. Hlavním účelem statistického modelu je co nejpřesnější reprodukce empirické četnosti různých kombinací hodnot proměnných v původním datovém souboru. Při ověřování přesnosti modelu proto porovnáváme empirické četnosti různých kombinací odpovědí s příslušnými odhady odvozenými z modelu.

Označíme-li $\boldsymbol{x}_C$ elementární vlastnost definovanou nějakou kombinací odpovědí a $\mathcal{S}(\boldsymbol{x}_C)$ příslušnou podmnožinu respondentů (subpopulaci) s touto vlastností

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

potom empirická četnost této vlastnosti $N(\boldsymbol{x}_C)$ současně udává velikost příslušné subpopulace $\mathcal{S}(\boldsymbol{x}_C)$. Empirickou četnost $N(\boldsymbol{x}_C)$ můžeme odhadnout jako součin velikosti celé populace a pravděpodobnosti $P(\boldsymbol{x}_C)$ odvozené z modelu:

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

Ideálně by odhad $\hat{N}(\boldsymbol{x}_C)$ měl být porovnán s příslušným empirickým údajem $N(\boldsymbol{x}_C)$ pro všechny možné elementární vlastnosti $\boldsymbol{x}_C$ definované kombinací několika odpovědí, nicméně je třeba vzít v úvahu dvě důležitá omezení.

Především, cílem metody není přesná reprodukce malých četností. Naopak, klesající přesnost modelu na úrovni málo pravděpodobných jevů představuje důležitý mechanizmus ochrany anonymity respondentů. Z tohoto důvodu se při ověřování přesnosti modelu omezíme na empirické četnosti větší než nějaký vhodně zvolený práh N_ϵ. Při určování prahové hodnoty vyjdeme z centrální limitní věty teorie pravděpodobnosti (podrobněji viz např. [4], [14]). Vzhledem k tomu, že spolehlivost empirické četnosti klesá s její velikostí, omezíme se pouze na "statisticky významné"četnosti, u kterých lze očekávat méně než pětiprocentní chybu. Konkrétně, v případě uvažovaného sčítání lidu z r. 2001 s počtem respondentů K=10 230 060, je chyba empirického údaje $N(\boldsymbol{x}_C)$ jako odhadu skutečné (neznámé) četnosti $N^*(\boldsymbol{x}_C)$ menší než 5% (na hladině spolehlivosti 95%), pokud neznámá skutečná četnost $N^*(\boldsymbol{x}_C) = |\mathcal{S}|P^*(\boldsymbol{x}_C)$ je větší než 1536, resp. pokud příslušná neznámá pravděpodobnost $P^*(\boldsymbol{x}_C)$ je větší než 0.0001502 (srv. [14]). Bohužel, informace o skutečné četnosti $N^*(\boldsymbol{x}_C)$ nejsou k dispozici. Chceme-li zaručit aby zjištěné empirické četnosti $N(\boldsymbol{x}_C)$ odpovídaly předpokládaným skutečným četnostem $N^*(\boldsymbol{x}_C)$ s chybou menší než 5% (na úrovni spolehlivosti 95%), musíme zvolit práh N_ϵ s určitou rezervou $N_{0.05} = 1612$ aby byla zajištěna platnost výše uvedené podmínky $N^*(\boldsymbol{x}_C) > 1536$. [3]

Druhé omezení je praktické povahy. Počet všech možných kombinací odpovědí je příliš velký z hlediska reálných výpočetních možností a je proto nutné omezit maximální počet odpovědí, které určují jednotlivé kombinace. V případě otázek uvedených v Tab.1 jsme výpočtem příslušných četností zjistili, že jednou odpovědí je

[3] Připomeňme, že uvažovaná nepřesnost 5% má čistě teoretický důvod plynoucí z limitní věty a nijak nesouvisí s praktickou realizací sčítání lidu. Případné chyby spojené s vyplňováním dotazníků mohou pouze zatížit odhady dodatečnou nepřesností.

Kriterium přesnosti	Test $\mathcal{A}_4$	Test $\mathcal{A}_5$
Průměrná relativní chyba modelu v %:	**4.07**	**4.17**
Standardní odchylka relativní chyby:	6.33	5.80
Maximální relativní chyba modelu v %:	240.84	240.84
Počet relativních chyb nad 100%:	925	4092
Průměrná absolutní chyba četnosti:	470	348
Standardní odchylka absolutní chyby:	951	655
Maximální absolutní chyba četnosti:	45779	45779
Počet testovacích subpopulací	**3 468 134**	**26 425 727**

Tabulka 3: Relativní a absolutní chyba statistického modelu s počtem komponent M=15000. Test přesnosti na základě souboru statisticky relevantních subpopulací (tj. větších než 1612 respondentů) určených kombinací nejvýše pěti odpovědí ($\mathcal{A}_5$, třetí sloupec) a na základě souboru relevantních subopulací určených kombinací nejvýše čtyř odpovědí ($\mathcal{A}_4$, druhý sloupec).

určeno 159 různých statisticky relevantních subpopulací, 10060 je určeno kombinací dvou nezávislých odpovědí, 270443 je určeno kombinací tří nezávislých odpovědí a 3503448 je určeno kombinací čtyř nezávislých odpovědí. Celkový počet statisticky relevantních vlastností, které lze určit kombinací nejvýše pěti nezávislých odpovědí je 26425727. Další zvyšování počtu otázek (tj. šest a více) při konstrukci testovacích vlastností značně prodlužuje výpočet a navíc nemá podstatný význam, protože příslušné subpopulace mají většinou podprahovou velikost. Z tohoto důvodu použijeme při ověřování přesnosti statistického modelu následující seznam vlastností definovaných pomocí statisticky významných kombinací nejvýše pěti odpovědí

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

Přirozeným kriteriem přesnosti statistického modelu je, ve smyslu předchozích úvah, průměrná absolutní chyba odvozovaných četností ve tvaru

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

kde $N(\boldsymbol{x}_C)$ je empirická četnost dotazníků s vlastností $\boldsymbol{x}_C$ v původním souboru $\mathcal{S}$ a $\mathcal{A}_5$ je třída všech statisticky relevantních subpopulací, které lze určit kombinací nejvýše pěti nezávislých odpovědí.

Hodnocení přesnosti statistického modelu pomocí kriteria (19) vychází zpravidla velmi příznivě, protože absolutní rozdíl není příliš vhodnou mírou přesnosti odhadů $\hat{N}(\boldsymbol{x}_C) = |\mathcal{S}|P(\boldsymbol{x}_C)$ odvozených z modelu. Většinou jde o poměrně malo rozdílné četnosti a je zřejmé, že pro malé hodnoty $N(\boldsymbol{x}_C)$ může i malý absolutní rozdíl $|P(\boldsymbol{x}_C)|\mathcal{S}| - N(\boldsymbol{x}_C)|$ odpovídat relativní chybě řádově v desítkách procent. Následující kriterium založené na relativní přesnosti počítá průměrnou hodnotu absolutní chyby ve vztahu k empirické četnosti $N(\boldsymbol{x}_C)$

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

Kriterium (20) (dále jen průměrná relativní chyba) představuje mnohem náročnější míru přesnosti, protože v případě malých hodnot $N(\boldsymbol{x}_C)$ i malé chyby odhadu $\hat{N}(\boldsymbol{x}_C)$ mohou značně zhoršit výsledný průměr E_r. Relativní chyba je také invariantní vůči normování, tzn. platí i pro sloupce zobrazovaných podmíněných histogramů. Jinými slovy, jestliže průměrná relativní chyba statistického modelu je $E_r\%$, potom sloupce libovolných histogramů jsou v průměru rovněž zobrazovány s chybou $E_r\%$.

Interval subpopulace	Dolní mez intervalu	Horní mez intervalu	Průměrná rel.chyba v %	Počet subpopulací
I= 1	1612	3000	**6.10**	7688027
I= 2	3000	5000	**4.88**	5011625
I= 3	5000	7500	**4.04**	3220931
I= 4	7500	10000	**3.50**	1906156
I= 5	10000	15000	**3.04**	2213787
I= 6	15000	30000	**2.38**	2695817
I= 7	30000	50000	**1.80**	1296118
I= 8	50000	100000	**1.37**	1075615
I= 9	100000	150000	**1.03**	372570
I= 10	150000	300000	**0.78**	358112
I= 11	300000	500000	**0.55**	125103
I= 12	500000	1000000	**0.39**	71104
I= 13	1000000	1500000	**0.29**	15324
I= 14	1500000	3000000	**0.22**	8511
I= 15	3000000	5000000	**0.12**	1349
I= 16	5000000	10300000	**0.02**	121

Tabulka 4: Rozložení četnosti relativních chyb podle velikosti zkoumané subpopulace (pro testovací soubor $\mathcal{A}_5$).

Tabulka 3 shrnuje výsledky testu přesnosti výsledného statistického modelu (M=15000 komponent). Přesnost byla ověřována pomocí souboru statisticky relevantních subpopulací (tj. větších než 1612 respondentů) určených kombinací nejvýše pěti odpovědí ($\mathcal{A}_5$, třetí sloupec) a pro srovnání též pomocí souboru relevantních subopulací určených kombinací nejvýše čtyř odpovědí ($\mathcal{A}_4$, druhý sloupec). Soubor $\mathcal{A}_5$ zahrnuje přibližně 26 mil. subpopulací, soubor $\mathcal{A}_4$ přibližně 3.5 mil. subpopulací. V tabulce je uvedena průměrná relativní a absolutní chyba (srv. (19), (20)), jejich standardní odchylky a maximální hodnoty a dále počet relativních chyb přesahujících 100%. Porovnání obou testů naznačuje, že použití šesti otázek při konstrukci statisticky relevantních subpopulací by patrně podstatným způsobem neovlivnilo výsledek testu přesnosti modelu.

Tabulka 4 ukazuje rozložení relativních chyb v závislosti na velikosti příslušné subpopulace (pro testovací soubor $\mathcal{A}_5$). Tabulka 5 ukazuje rozložení relativních chyb v závislosti na jejich velikosti (pro test $\mathcal{A}_5$). Z tabulky vyplývá, že většina (73%) odhadů odvozených ze statistického modelu (tj. sloupců v histogramech) se zobrazuje s relativní chybou menší než 5%. Průměrná relativní chyba reprodukce nejmenších relativních četností (subpopulace od 1612 do 3000 respondentů) je 6.1% a rychle klesá s rostoucí velikostí subpopulace. Z tabulky je zřejmé, že většina relevantních subpopulací z testovacího souboru $\mathcal{A}_5$ (téměř 94%) má pod 60000 respondentů.

Interval	Dolní mez chyby v %	Horní mez chyby v %	Relativní četnost
0	0	5	0.729425
1	05	10	0.178543
2	10	15	0.052805
3	15	20	0.019376
4	20	25	0.008664
5	25	30	0.004211
6	30	35	0.002399
7	35	40	0.001493
8	40	45	0.000875
9	45	50	0.000621
10	50	55	0.000415
11	55	60	0.000256
12	60	65	0.000204
13	65	70	0.000162
14	70	75	0.000097
15	75	80	0.000083
16	80	85	0.000027
17	85	90	0.000034
18	90	95	0.000075
19	95	100	0.000075
20	100	a více	0.000217

Tabulka 5: Rozložení relativních chyb podle jejich velikosti (pro test $\mathcal{A}_5$).

Připomeňme, že klesající přesnost reprodukce malých relativních četností je důležitá z hlediska ochrany osobních údajů. Jakákoli identifikace respondentů pomocí statistického modelu je znemožněna sníženou spolehlivostí histogramů odvozených pro extrémně malé části populace nebo dokonce pro jednotlivce (viz [12]). V interaktivním režimu je výpočet histogramů pro malé subpopulace zablokován, jednotlivé sloupce histogramů odpovídající příliš malému počtu respondentů (méně než 1613) jsou zobrazovány s příslušným varováním.

Metoda interaktivní reprodukce vlastností dat pomocí statistického modelu je svojí obecností srovnatelná s možnostmi, které nabízí distribuce tzv. anonymizovaných souborů mikrodat. V tabulce 6 jsou uvedeny výsledky testu přesnosti relativních četností vypočítaných z podsouboru mikrodat, který byl vytvořen náhodným výběrem 10% dotazníků z původního souboru (s doplněnými údaji). Přesnost odhadů byla testována stejným způsobem jako v případě tabulky 3 pomocí souborů statisticky relevantních subpopulací. Anonymizační procedury nebyly použity, tzn. dosažená přesnost odpovídá ideální situaci bez nutnosti jakékoli manipulace s mikrodaty. Na základě porovnání tabulek 3 a 6 lze říci, že přesnost reprodukce relativních četnosti pomocí souboru mikrodat je srovnatelná se statistickým modelem. Relativní chyba 3.6%, je menší než v případě modelu (4.2%), rovněž rozptyl relativní chyby je menší. Při opakování testu s jinými náhodnými podsoubory byly výsledky velmi podobné. Nevýhodou souboru mikrodat jsou omezené možnosti distribuce (pouze

Kriterium přesnosti	Test $\mathcal{A}_4$	Test $\mathcal{A}_5$
Střední relativní chyba v %:	**2.94**	**3.60**
Standardní odchylka relativní chyby:	3.00	3.72
Maximální relativní chyba v %:	35.20	42.62
Počet relativních chyb nad 100%:	0	0
Střední absolutní chyba četnosti:	307	409
Standardní odchylka absolutní chyby:	450	1913
Maximální absolutní chyba četnosti:	12348	59815
Počet testovacích kombinací	**3 468 134**	**26 425 727**
Velikost vybraného podsouboru mikrodat	1022666	1022666

Tabulka 6: Relativní a absolutní přesnost odhadů relativních četností vypočítaných z náhodně vybraného souboru mikrodat (asi 10% dotazníků). Test přesnosti by proveden stejně jako v případě tabulky 3 pomocí souboru statisticky relevantních subpopulací.

pro výzkumné instituce) a nutnost dodatečné anonymizace dotazníků.

7. Možnosti využití a přínos metody

Předmětem článku je výpočet statistického modelu dat ze sčítání lidu v České republice v roce 2001 pro účely interaktivní reprodukce statistických vlastností výchozí databáze. Model ve tvaru diskrétní distribuční směsi je bezprostředně použit jako báze znalostí pravděpodobnostního expertního systému PES [8], [9], [10]. Expertní systém modifikovaný pro specifické potřeby sčítání lidu poskytuje uživateli formálně stejné možnosti, jako přímý kontakt s datovým souborem prostřednictvím databázového systému. Pomocí inferenčního mechanismu expertního systému může být libovolná statistická informace odvozena přímo z odhadnutého modelu - ve zlomcích vteřiny a bez dalšího přístupu k výchozím datům. Interaktivní modul nabízí také řadu nástrojů pro usnadnění informační analýzy proměnných a pro rychlou analýzu vlastností různých částí populace, jako je např. možnost vícenásobného porovnávání histogramů, vyhledávání informativních otázek a pod. Výsledný softvarový produkt neobsahuje původní chráněná data a může být distribuován bez jakéhokoli omezení, ať již ve formě interaktivní internetové aplikace nebo jako soubor, který je volně k dispozici ke stažení. Statistická informace obsažená v původním chráněném datovém souboru tak může být zpřístupněna širokému okruhu uživatelů bez rizika porušení ochrany osobních údajů.

Hlavní pozornost je v článku věnována výpočtu modelu z nekompletních dat a především výsledné přesnosti modelu ve srovnání s možnostmi tzv. anonymizovaných souborů mikrodat. Chybějící údaje byly nejprve nahrazeny maximálně věrohodnými odhady pomocí modelu vypočítaného z neúplných dat a výsledný doplněný soubor byl použit pro výpočet konečného modelu. Možnost zpracování neúplných dat vytváří předpoklady pro výpočet statistického modelu již v průběhu počáteční časově náročné fáze formální kontroly dat. Z výsledků testu je zřejmé, že i při velkém počtu chybějících údajů by statistický model mohl být užitečným zdrojem předběžné in-

formace pro identifikaci a kontrolu chybných resp. nepravděpodobných údajů.

Přesnost modelu byla ověřována pomocí souboru tzv. statisticky relevantních subpopulací. Testovací soubor zahrnoval cca 26 mil. statisticky významných relativních četností. Průměrná relativní chyba reprodukce testovacích údajů pomocí modelu se blíží 4%, tzn. sloupce libovolných histogramů jsou v interaktivním režimu zobrazovány s průměrnou přesností 4%. Dosažená přesnost statistického modelu je srovnatelná s přesností odhadů, které lze odvodit z náhodně vybraného podsouboru 10^6 dotazníků (3.6%), nicméně rychlost a složitost informační analýzy, kterou nabízí interaktivní režim expertního systému, jednoznačně přesahují možnosti současných databázových systémů. Metoda interaktivního statistického modelu může významně zlepšit dostupnost výsledků sčítání lidu a tím přispět k lepšímu zhodnocení vynaložených prostředků.

Reference

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

[2] *European Plan for Research in Official Statistics.* Report of the Statistical Office of the European Communities (EUROSTAT). Ed. Deo Ramprakash, Luxembourg, 1998

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

[4] Feller, W.: *An Introduction to Probability Theory and Its Applications, Vol. I*, John Wiley & Sons, New York, London 1962

[5] Fienberg, S.E.: *Conflicts between the needs for access to statistical information and demands for confidentiality*, Journal of Official Statistics, Vol. 10, pp. 115-132, 1994

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

[7] Grim, J.: *On numerical evaluation of maximum - likelihood estimates for finite mixtures of distributions*, Kybernetika, Vol. 18, No. 3, pp. 173-190, 1982

[8] Grim, J.: *Probabilistic expert systems and distribution mixtures*, Computers and Artificial Intelligence, Vol. 9, No. 3, pp. 241-256, 1990

[9] Grim, J.: *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), pp. 997-1005, World Scientific, Singapore 1992

[10] Grim, J.: *Knowledge representation and uncertainty processing in the probabilistic expert system PES*, Int. Journal of General Systems, Vol. 22, No. 2, pp. 103-111, 1994

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

[12] Grim, J., Boček, P., Pudil, P.: *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, pp. 849-856

[13] Grim, J., Hora, J., Boček, P., Somol, P., Pudil, P.: *Information Analysis of Census Data by Using Statistical Models.* Sborník z mezinárodní konference Statistics - Investment in the Future, Praha, 6. - 7. září 2004

[14] Grim, J., Hora, J., Pudil, P.: *Interaktivní reprodukce výsledků sčítání lidu pomocí statistického modelu se zaručenou ochranou anonymity dat.* Statistika. Vol. 40, No. 5 (2004), pp. 400-414.

[15] Kincl, T.: *Metody publikace výsledků sčítání lidu v zemích Evropské unie.* Výzkumná zpráva č. 1/2001 FM VŠE, Jindřichův Hradec, duben 2001

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

[17] Schlesinger, M.I.: *Relation between learning and self-learning in pattern recognition (in Russian)*, Kibernetika, (Kiev), No. 2, pp. 81-88, 1968

[18] Titterington, D.M., Smith, A.F.M., Makov, U.E.: *Statistical analysis of finite mixture distributions*, John Wiley & Sons, Chichester, New York, Singapore 1985

[19] Willenborg, L.C.R.J., de Waal, A.G.: *Elements of statistical disclosure control*, Springer Verlag, New York, 2001.

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

Jiří Grim, Petr Somol, Pavel Boček, Ústav teorie informace a automatizace AV ČR, Pod vodárenskou věží 4, 18208 Praha 8, (též za Fakultu Managementu, Jindřichův Hradec)

Jan Hora, Fakulta jaderná a fyzikálně inženýrská ČVUT, Trojanova 13, 120 00 Praha 2

Pavel Pudil, Fakulta managementu VŠE, Jarošovská 1117/II, 37701 Jindřichův Hradec

Abstract

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. 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. In this way we can derive the statistical properties of data interactively without any further access to the source database. 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 anonymised subsets of microdata.

Keywords: Interactive statistical model, distribution mixtures, EM algorithm, incomplete data, probabilistic expert systems, data reproduction accuracy