O STATISTICKÉM VYHODNOCENÍ REGIONÁLNÍCH DEMOGRAFICKÝCH UKAZATELŮ

Jaromír Běláček, SEÚ ČSAV

Klasická analýza základních demografických procesů (*porodnosti, úmrtnosti, sňatečnosti, rozvodovosti* ev. *potratovosti* a *migrací*) je veskrze založena na přepočtech odpovídajících demografických událostí (tj. četností narozených, zemřelých, sezdaných atd.) vůči analogicky vytříděným demografickým specifikacím jiným (počátečním, středním stavům obyvatelstva apod.). Většina používaných ukazatelů je konstruována ve formě *podílu*, který (v závislosti na metodické volbě čitatele, jmenovatele a způsobu třídění) může mít interpretaci *pravděpodobnosti*, *míry* intenzity nebo poměrového *indexu* (viz [7],str.91). Ukazatele vypočtené v rámci předem vymezeného regionálního systému jsou obvykle srovnávány mezi sebou a kvantitativní rozdíly interpretovány v souladu s jejich vnitřním demografickým obsahem. Na úrovni okresů a nižších jednotek však dosažené hodnoty nelze kvalifikovat jako přesné, neboť se výrazněji uplatňuje stochastický charakter sledovaných událostí.

V některých případech lze odhadovat směrodatnou odchylku každého jednotlivého ukazatele prostřednictvím binomického rozdělení a na tomto základě testovat hypotézy o teoretických *hodnotách* (viz [4] nebo [8]). Pro nejobvyklejší situace, kdy se jedná o tzv. *hrubé* nebo *specifické míry* (počty událostí na 1000 obyvatel středního stavu žijících) výše zmíněné pojetí přímo aplikovat nelze. Nezbývá, než vyhodnocovat regionální variabilitu *ad poste*. Formální analýza založená na prostých průměrech a konstantních směrodatných odchylkách je však v demografickém i statistickém smyslu nevhodná (viz odstavec 1). V rámci tohoto příspěvku je východisko hledáno v aplikacích *zobecněného (Aitkenova) lineárního modelu* při demograficky opodstatněné volbě vah (odstavec 2).

Navržená metoda byla uplatněna ve studiích [1] a [2]. Pro účely tohoto metodického shrnutí byl vybrán ukazatel *úhrnné plodnosti* odhadnutý pro 71 bilančních celků Severočeského kraje za období 1982-87. (Územní identifikace a použité datové podklady jsou součástí článku [2].) Definice úhrnné plodnosti je nepatrně složitější nežli pro *podílové* míry; umožňuje však přirozeně nahlédnout do struktury plodnosti podle věku žen, což představuje stěžejní bod tohoto směru demografické analýzy. Rozklady v rámci druhého stupně třídění (odstavec 4) poskytují i testy vztažené k hladinám tzv. *přímo standardizovaných měr*. Některé další aplikační i teoretické poznámky jsou učiněny ve stati 5.

1. ELEMENTÁRNÍ ANALÝZA

Pro *i*-tou regionální jednotku je ukazatel úhrnné plodnosti f_i definován předpisem

$$f_i = \sum_{k=1}^{m} f_{ik} = \sum_{k=1}^{m} \frac{n_{ik}}{P_{ik}} , \qquad i=1,\ldots,n , \tag{1}$$

kde věkově specifické plodnosti f_{ik} jsou vyjádřeny jako podíly všech živě narozených n_{ik} vůči všem ženám žijícím v regionu i ve věkové skupině k. Jeho demografická interpretace odpovídá průměrnému počtu dětí, které se narodí matkám jedné fiktivní generace tvořené *transverzální* (průřezovou) věkovou strukturou ke středu sledovaného období. (V případě, že jde o agregované věkové skupiny nebo o agregovaný časový interval sledování, je třeba pronásobit hodnotu f_i resp. každý člen za sumačním znaménkem v (1) poměrovou konstantou, což ovšem nemá vliv na níže navrhovaná statistická vyhodnocení.)

Obr. 1: Histogram ukazatelů úhrnné plodnosti f_i v 71 bilančních celcích Severočeského kraje

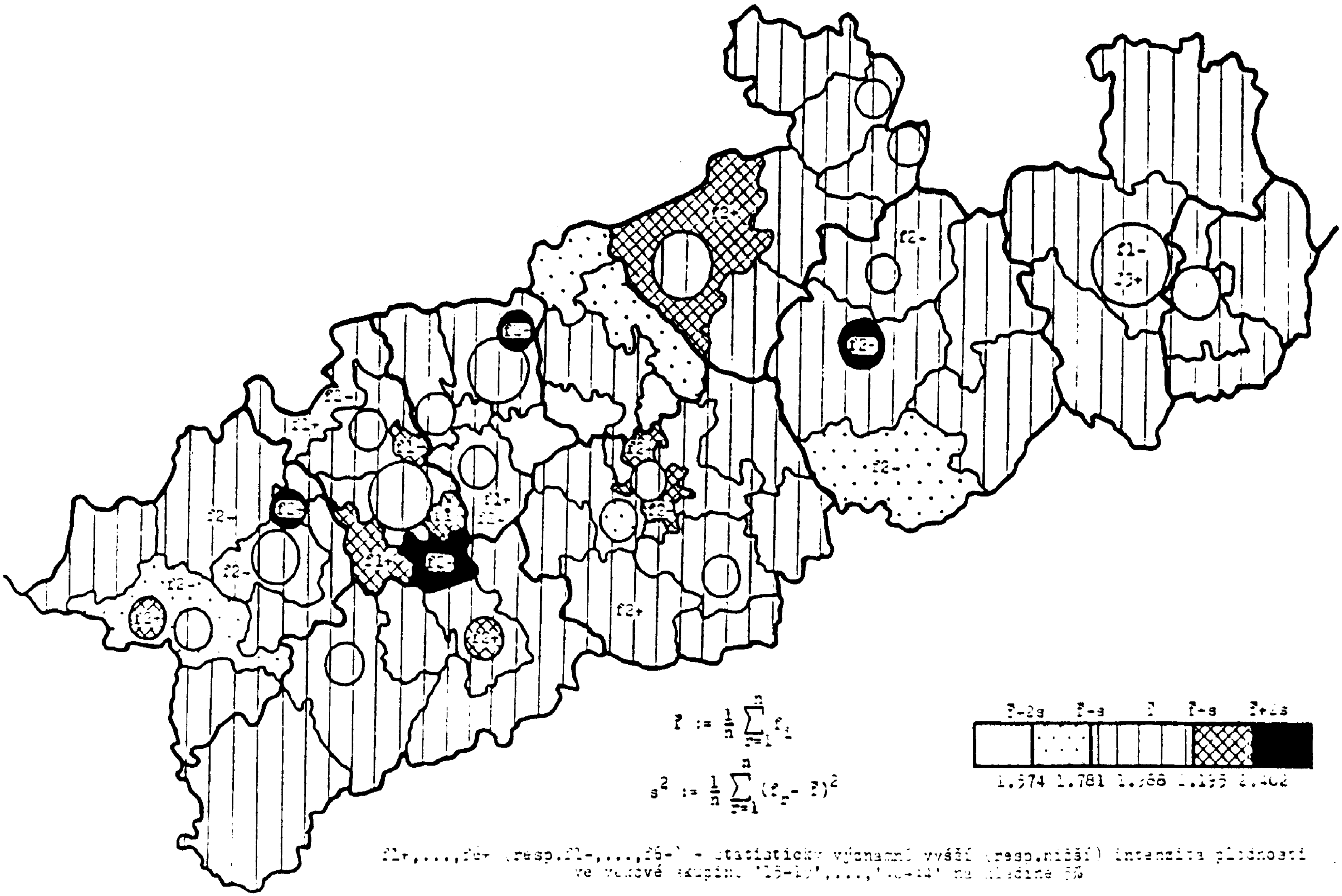
Bin	Lower	Upper	Count	Prcnt	Total	Prcnt	Histogram
1	1.423	1.498067	1	1.4	1	1.4	:*
2	1.498067	1.573133	1	1.4	2	2.8	:*
3	1.573133	1.6482	0	0.0	2	2.8	:
4	1.6482	1.723267	2	2.8	4	5.6	:**
5	1.723267	1.798333	3	4.2	7	9.9	:***
6	1.798333	1.8734	14	19.7	21	29.6	:**************
7	1.8734	1.948467	13	18.3	34	47.9	:*************
8	1.948467	2.023533	10	14.1	44	62.0	:**********
9	2.023533	2.0986	9	12.7	53	74.6	:*********
10	2.0986	2.173667	9	12.7	62	87.3	:*********
11	2.173667	2.248733	0	0.0	62	87.3	:
12	2.248733	2.3238	3	4.2	65	91.5	:***
13	2.3238	2.398867	2	2.8	67	94.4	:**
14	2.398867	2.473933	2	2.8	69	97.2	:**
15	2.473933	2.549	2	2.8	71	100.0	:**

Obvyklým metodologickým postupem geografů je zvolit (v souladu s interpretací ukazatele a zjištěnými hladinami hodnot) pevné ekvidistantní dělení a na základě této diskretizace škálovat regiony odpovídadající barvou v *kartogramu*. Výhodou takového zpracování je možnost identifikace případných regionálních souvislostí, nevýhodou naopak subjektivní volba škály, která může být nereprezentativní například z časového pohledu. Dokumentuje to analogicky zkonstruovaný histogram na obr.1 pořízený na základě 71 regionálních hodnot f_i úhrnné plodnosti v Severočeském kraji. Typicky jednovrcholový průběh může snad evokovat závěry o anomálním vývoji plodnosti v 'krajních škálovacích intervalech'. V blízkosti modu teoretické křivky však nelze pochybovat o znáhodněném přiřazení do zvolené intervalové kategorizace.

Statisticky motivované řešení nabízí volit dělení škálovacích intervalů v souladu s formální statistickou významností odchylek regionálních hodnot f_i od průměru $\bar{f}$. Regiony se 'statisticky významně' vyšší plodností (tj. při $f_i > \bar{f}+2s$, kde $s^2 = n^{-1}\Sigma(f_i - \bar{f})^2$) jsou reprezentovány na obr.2 černou podkladovou barvou (Jirkov, Česká Lípa, Krupka, Bečov), bilanční celky s nižší plodností (při $f_i < \bar{f}-2s$) jsou naopak vyznačeny bíle (Ústí n.Labem-Středohoří, Teplice-Venkov). Použitý průměr $\bar{f}$ však není demograficky reprezentativním odhadem 'parametru polohy'. Tím je přirozeně hladina úhrnné plodnosti f vypočtená pro nadřazený územní celek (v daném případě pro Severočeský kraj).

I statistické vyhodnocení odchylek $(f_i - f)$ musí být vázáno na adekvátní odhad variability. V daném případě lze využít poznatku, že ukazatel f vypočtený přesně pro nadřazené území je numericky téměř shodný s váženým průměrem $f\hat{} = \Sigma q_i f_i$, když váhy q_i (splňující podmínku $\Sigma q_i = 1$) volíme proporcionálně jmenovatelům v definičním předpisu pro f_i . (V situaci *hrubých* nebo ***specifických*** měr platí přímo $f = f\hat{}$. U ukazatelů z definice (1) je dosahováno přibližné rovnosti při volbě vah přímo úměrných počtům žijících obyvatel v regionech, počtům žijících žen nebo žen v plodném věku.) To umožňuje testovat odchylky f_i od f resp. $f\hat{}$ analogicky jako výše s odhadem s^2 nahrazeným jeho váženou alternativou $s\hat{}^2 = \Sigma q_i (f_i - f\hat{})^2$. Normalizované odchylky $(f_i - f\hat{})/s\hat{}$

Obr. 2: Škálování odchylek ukazatelů úhrnné plodnosti od (neváženého) průměru Severočeského kraje (bilanční celky 1982-7)



$$\bar{f} := \frac{1}{n}\sum_{r=1}^{n} f_i$$

$$s^2 := \frac{1}{n}\sum_{r=1}^{n} (f_r - \bar{f})^2$$

f1+,...,f6+ (resp. f1-,...,f6-) - statisticky významně vyšší (resp. nižší) intenzita plodnosti ve věkové skupině '15-19',...,'40-44' na hladině 5%

jsou vychýleny ve smyslu vyšších hodnot standardu f ($\approx f^{\wedge}$) Severočeského kraje a většího rozptylu $s^{\wedge 2}$ ($>s^2$) ; odpovídající kartogram (ve srovnání s obr. 2) by jen nepatrně zvýraznil nižší plodnost v některých periferních oblastech (podkrušnohorské části okresů Chomutov a Teplice, zázemí města Kadaň a Jablonce n. Nisou).

2. HETEROGENITA ROZPTYLŮ

Analýza představená výše přizpůsobila statistické vyhodnocení demograficky korektní interpretaci parametru polohy. Formálně byla založena na elementárním modelu

$$Ef_i = f \quad , \quad i=1,\dots,n \quad , \tag{2}$$

když rozptyl $Var(f_i)$ předpokládáme konstantní. Již rozbor podle velikostních skupin obcí poukazuje na větší variabilitu f_i u méně osídlených bilančních celků. Empiricky zjištěný rozptyl jednoletých ukazatelů vypovídá o jeho nepřímé závislosti na počtu žijících obyvatel v regionech (obr. 3). Uvědomíme-li si, že tyto početní stavy působí *de facto* jako generátory událostí kvantifikovaných v čitateli definičních předpisů typu (1), je přirozené doplnit (2) předpokladem

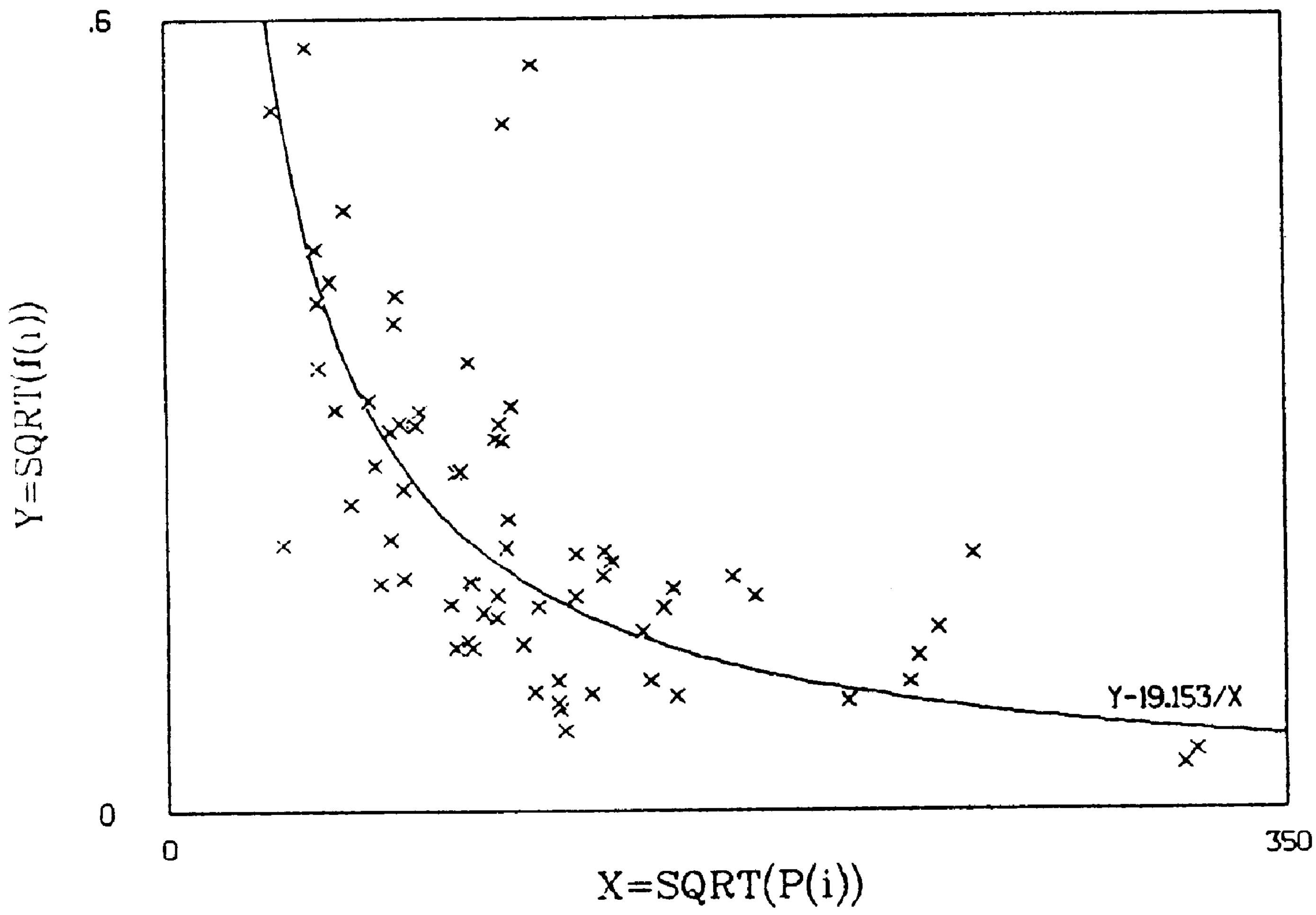
$$Var(f_i) = \sigma^2/P_i \quad , \quad i=1,\dots,n \quad . \tag{3}$$

Model (2)+(3) můžeme nyní považovat za speciální případ *zobecněného lineárního modelu*

$$E\underline{y} = \underline{X}\underline{\theta} \quad , \quad Var(\underline{y}) = \sigma^2.\underline{\Omega}^{-1} \quad , \tag{4}$$

kde $\underline{y}:=(f_1,\dots,f_n)'$, $\underline{X}:=(1,\dots,1)'$, $\underline{\theta}:=f$ a $\underline{\Omega}:=\mathrm{diag}(P_1,\dots,P_n)$.

Obr. 3: Závislost směrodatných odchylek jednoletých ukazatelů úhrnné plodnosti (bilanční celky Severočeského kraje 1982-1987) na počtech žijících obyvatel (střední stavy k 31.12.1984)



Nejlepší nestranné lineární odhady parametrů f a σ^2 získáme dosazením do obecných vzorců (viz např. v [3], str. 65)

$$\hat{\theta} = (\underline{X}'\underline{\Omega}\underline{X})^{-1}\underline{X}'\underline{\Omega}\underline{Y} = \sum_{i=1}^{n} \frac{P_i}{P_+} = \hat{f} , \qquad (5)$$

$$\hat{\sigma}^2 = \frac{1}{(n-h(\underline{X}))} (\underline{Y}-\underline{X}\hat{\theta})'\underline{\Omega}(\underline{Y}-\underline{X}\hat{\theta}) = \frac{1}{(n-1)} \sum_{i=1}^{n} P_i(f_i - \hat{f})^2 , \qquad (6)$$

kde $P_+ := \Sigma P_i$ a $h(\underline{X})$ značí hodnost matice $\underline{X}$.

Na vzorcích (3), (5) a (6) můžeme založit škálování individuálních odchylek f_i od hodnoty $\hat{f}$. Pro jednoduchost volíme dělící body jako $\hat{f}-2\hat{s}_i$, $\hat{f}-\hat{s}_i$, $\hat{f}+\hat{s}_i$, $\hat{f}+2\hat{s}_i$, když

$$\hat{s}_i^2 := \frac{1}{nP_i} \sum_{r=1}^{n} P_r(f_r - \hat{f})^2 , \quad i=1,\ldots,n , \qquad (7)$$

jsou odhady rozptylu počítané pro jednotlivé bilanční celky. Oproti předchozím situacím (viz v [2] kartogram č.4) se zvýrazní vyšší úhrnná plodnost v některých okresních městech (Louny, Chomutov, Most), sníží se formální významnost pro některé méně osídlené regiony (Bečov). Tato vlastnost je vhodná i z hlediska demografických interpretací, neboť fyzicky menší odchylky ukazatelů pro městské populace mají možnost projevit se jako více statisticky významné nežli méně stabilní hodnoty vypočtené za menší obce.

Jako problematická se může jevit interpretace v rámci vyššího územního systému (kupř. České republiky). (Srovnatelná hodnota úhrnné plodnosti pro Severočeský kraj byla ve sledovaném období vyšší než standard ČR). Rozšíříme-li původní regionální systém o 'zbytek území ČR' (tj. o $[n+1]$-ní bilanční celek s vahami úměrnými počtu žijících obyvatel), můžeme každou hodnotu f_i korektně srovnávat se standardem $f_{ČR}$ analogicky jako výše. Kartogram na obr.4 škálovaný na základě odhadů rozptylu (7) pro rozšířený regionální systém poukazuje na zvýšenou plodnost ještě v dalších urbanizovaných jednotkách (Žatec a Ústí nad Labem-město); nižší plodnost se indikuje výrazněji pouze v demograficky málo atraktivních lokalitách (České Středohoří, Doksy). Výhodou tohoto kartogramu je jeho srovnatelnost s eventuálně dalšími územními systémy (na úrovni krajů, okresů, bilančních celků či agregovaných obcí), pokud by jejich třeba i nezávislá analýza byla provedena stejnou metodikou.

3. JEDNODUCHÉ TŘÍDĚNÍ

Při regionálních úlohách je často zapotřebí vyhodnotit hladiny demografických ukazatelů pro specificky vymezené oblasti (okresy, sídelní regionální aglomerace, jinak definované bilanční celky apod.). Jsou-li tyto oblasti agregovatelné z nižších regionálních jednotek, pro které jsou k dispozici podrobnější údaje (hodnoty f_i resp. P_i za bilanční celky nebo za obce), můžeme testovat hypotézy o *oblastních* hodnotách nebo o jejich vnitřní variabilitě na základě týchž principů jako v předchozí stati. Nechť $I_1,\ldots,I_l$ značí rozklad indexové množiny $\{1,\ldots,n\}$ na l $(<n)$ disjunktních oblastí. Za předpokladu (3) použijeme model

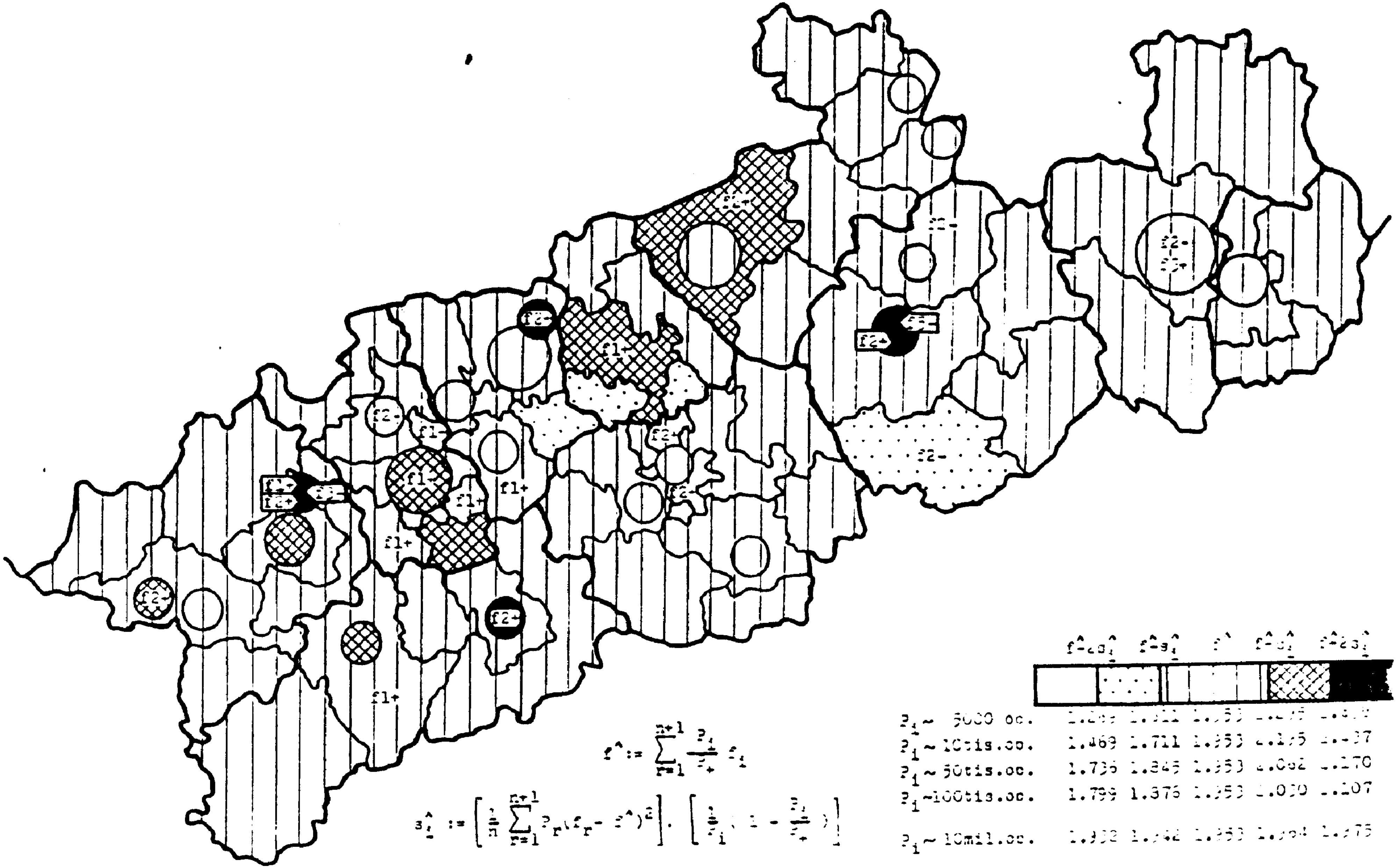
$$Ef_i = x_{ij}.f^j , \quad i=1,\ldots,n , \quad j=1,\ldots,l , \qquad (8)$$

kde

$$x_{ij} = 1 \text{ pro } i \in I_j \quad \text{a} \quad x_{ij} = 0 \text{ pro } i \notin I_j . \qquad (9)$$

Aitkenův odhad vektoru oblastních ukazatelů $(f^1,\ldots,f^l)'$ vyjádříme z levé strany obecného vzorce (5), kam dosadíme $\underline{\Omega}:=\{P_1,\ldots,P_n\}$ a za $\underline{X}$

Obr.4: Škálování odchylek ukazatelů úhrnné plodnosti od standardu (váženého průměru) České republiky (bilanční celky JČ kraje 1982-7)



$$f^{\wedge} := \sum_{r=1}^{n+1} \frac{P_r}{P_+} f_r$$

$$s_i^{\wedge} := \left[\frac{1}{n} \sum_{r=1}^{n+1} P_r (f_r - f^{\wedge})^2 \right] \cdot \left[\frac{1}{P_i} \left(1 - \frac{P_i}{P_+} \right) \right]$$

	$f^{\wedge}-2s_i^{\wedge}$	$f^{\wedge}-s_i^{\wedge}$	$f^{\wedge}$	$f^{\wedge}+s_i^{\wedge}$	$f^{\wedge}+2s_i^{\wedge}$
$P_i \sim$ 5000 ob.	[illegible]	[illegible]	1.953	[illegible]	[illegible]
$P_i \sim$ 10tis.ob.	1.469	1.711	1.953	[illegible]	[illegible]
$P_i \sim$ 50tis.ob.	1.736	1.845	1.953	2.062	2.170
$P_i \sim$ 100tis.ob.	1.799	1.876	1.953	2.030	2.107
$P_i \sim$ 10mil.ob.	1.932	1.942	1.953	1.964	1.975

matici $n\times l$ s prvky (x_{ij}). Získaný vektor odhadů

$$(f^{1\wedge},\dots,f^{l\wedge})' = \left(\frac{\sum_{i\in I_1} P_i f_i}{P^1} , \dots , \frac{\sum_{i\in I_l} P_i f_i}{P^l} \right)' \tag{10}$$

je numericky vhodnou aproximací přesně vypočtených oblastních ukazatelů s varianční maticí $\sigma^2.\mathrm{diag}\{(P^1)^{-1},\dots,(P^l)^{-1}\}$, kde $P^j = \sum_{i\in I_j} P_i$ pro $j=1,\dots,l$.

Test hypotézy H_0: $f^1 = f^2 = \dots = f^l$ o rovnosti oblastních hodnot můžeme založit na identitě

$$S_T^2 = \sum_{i=1}^{n} P_i (f_i - f^\wedge)^2 = \sum_{j=1}^{l} P^j (f^{j\wedge} - f^\wedge)^2 + \sum_{j=1}^{l} \sum_{i\in I_j} P_i (f_i - f^{j\wedge})^2 = S_A + S_e \tag{11}$$

pro $f^\wedge = \sum_{i=1}^{n} \frac{P_i}{P_+} f_i = \sum_{j=1}^{l} \frac{P^j}{P_+} f^{j\wedge}$. Hypotézu H_0 zamítneme, když

$$F_A := \frac{S_A/(l-1)}{S_e/(n-l)} > F_{l-1,n-l}(\alpha) , \tag{12}$$

kde $F_{l-1,n-l}(\alpha)$ je kritická hodnota *F-rozdělení* s $(l-1)$ a $(n-l)$ stupni volnosti na hladině významnosti α . V pozitivním případě je korektní identifikovat statistickou významnost odchylek ukazatelů každé dvojice oblastí I_j a $I_{j'}$, s použitím kritéria

$$|f^{j\wedge} - f^{j'\wedge}| > \left[\left(\frac{1}{P^j} + \frac{1}{P^{j'}}\right) \frac{(l-1)S_e}{(n-l)} F_{l-1,n-l}(\alpha)\right]^{1/2}. \tag{13}$$

Pokud hypotézu H_0 nezamítáme, lze oblastní regionalizaci $I_1,\dots,I_l$ chápat opět jako vstupní regionální systém ze stati 1 a prověřovat významnost odchylek oblastních ukazatelů vůči standardu nadřazeného území. (Za platnosti H_0 splňují ukazatele vypočtené za agregované oblasti právě tak jako jejich Aitkenovy odhady předpoklady (2) a (3)). Smyslem takového vyhodnocování by bylo získání větší průkaznosti identifikovaných anomálií (pro větší území), nedostatkem naopak nivelizace individuálních rozdílů zjištěných při nejpodrobnější možné regionalizaci. Při vymezení oblastí Severočeského kraje podle deseti okresů resp. šesti velikostních skupin ('0-5tis.', '5-10tis.', '10-20tis.', '20-50tis.','100tis. a více' žijících obyvatel) nebyly v úrovni plodnosti zjištěny statisticky významné rozdíly mezi oblastmi. Oproti analýze rozptylu s nekorektními neváženými průměry se při *zobecněném* testu zvýšila u obou regionalizací hladina významnosti F-statistiky z (12).

4. DVOJNÉ TŘÍDĚNÍ BEZ INTERAKCÍ

Pro konstrukci demografických předpovědí je podstatný rozbor struktury ukazatelů vypočtených v rámci jednotlivých třídících kategorií. V regionálním smyslu je typickým datovým podkladem dvojrozměrná tabulka *specifických měr* f_{ik} definovaných jako v (1). Poučeni předchozími zkušenostmi předpokládáme model

$$Ef_{ik} = \mu + \alpha_i + \beta_k , \quad i=1,\dots,n , \tag{14}$$

$$\mathrm{Var}(f_{ik}) = \sigma^2/Q_{ik} , \quad k=1,\dots,m , \tag{15}$$

kde μ, α_i, β_k a $\sigma^2>0$ jsou neznámé parametry a Q_{ik} známé konstanty v určitém smyslu aproximující jmenovatele v definičních předpisech pro f_{ik}. Formalizuje se tedy předpoklad, že po odstranění aditivního vlivu regionálních a věkově selektivních efektů je přirozená variabilita systému určená soustavou vah interpretačně vztažených k mechanizmu výpočtu ukazatelů. Parametr σ^2 figuruje pouze jako konstanta úměrnosti umožňující volnější specifikaci vahových koeficientů.

Aitkenovy odhady systematických parametrů jsou vyjádřitelné z *normálních rovnic* (viz [5], str. 28-9) tvaru

$$
\begin{aligned}
Q_{++}\mu &+ \sum_{r=1}^{n} Q_{r+}\alpha_r + \sum_{s=1}^{m} Q_{+s}\beta_s = \sum_{r=1}^{n}\sum_{s=1}^{m} Q_{rs} f_{rs} , \\
Q_{i+}\mu &+ Q_{i+}\alpha_i + \sum_{s=1}^{m} Q_{is}\beta_s = \sum_{s=1}^{m} Q_{is} f_{is} , \qquad i=1,\dots,n , \qquad (17)\\
Q_{+k}\mu &+ \sum_{r=1}^{n} Q_{rk}\alpha_r + Q_{+k}\beta_k = \sum_{r=1}^{n} Q_{rk} f_{rk} , \qquad k=1,\dots,m ,
\end{aligned}
$$

kde $Q_{i+}= \sum_s Q_{is}$, $Q_{+k}= \sum_r Q_{rk}$ a $Q_{++}= \sum_r\sum_s Q_{rs}$.

Za předpokladu

$$\sum_{r=1}^{n} Q_{ik}\alpha_i = \sum_{s=1}^{m} Q_{ik}\beta_k = 0 \qquad \text{pro } i=1,\dots,n , \quad k=1,\dots,m \qquad (*)$$

dostáváme explicitní vzorce

$$
\begin{aligned}
\hat{\mu} &= \sum_{i=1}^{n}\sum_{k=1}^{m} \frac{Q_{ik}}{Q_{++}} , \\
\hat{\alpha}_i &= \sum_{k=1}^{n} \frac{Q_{ik}}{Q_{++}} - \hat{\mu} =: \hat{\mu}_{i+} - \hat{\mu} , \qquad i=1,\dots,n , \qquad (18)\\
\hat{\beta}_k &= \sum_{i=1}^{m} \frac{Q_{ik}}{Q_{++}} - \hat{\mu} =: \hat{\mu}_{+k} - \hat{\mu} , \qquad k=1,\dots,m .
\end{aligned}
$$

Nestranný odhad parametru σ^2 pořídíme potom dosazením do levé strany v (6) jakožto

$$\tilde{\sigma}^2 = \frac{1}{(n-1).(m-1)} \sum_{i=1}^{n}\sum_{k=1}^{m} Q_{ik}(f_{ik} - \hat{\mu}_{i+} - \hat{\mu}_{+k} + \hat{\mu})^2 . \qquad (19)$$

Volíme-li $Q_{ik}= P_{ik}$ (tj. jako střední stavy obyvatelstva ve třídách exponované populace), jsou odhady $\hat{\mu}_{1+},\dots,\hat{\mu}_{n+},\hat{\mu}$ resp. $\hat{\mu}_{+1},\dots,\hat{\mu}_{+m}$ z (18) po řadě rovny hodnotám tzv. *hrubých* (regionálních) *měr* resp. *specifických měr standardu* (nadřízené územní jednotky). Při takové volbě vah však přesné Aitkenovy odhady nebudou splňovat (18) a podmínka (*) může být příliš omezující. Pro smysluplné demografické aplikace však postačuje struktura vah typu

$$Q_{ik}= R_i.S_k , \qquad i=1,\dots,n , \quad k=1,\dots,m , \qquad (**)$$

která již zaručuje platnost (*) (a tudíž i (18)). Jsou-li konstanty $R_1,\dots,R_n,S_1,\dots,S_m$ proporcionální marginálním (řádkovým a sloupcovým) součtům žijících obyvatel, mají odhady $\hat{\mu}_{1+},\dots,\hat{\mu}_{n+}$ resp. $\hat{\mu}$ interpretace tzv. *přímo standardizovaných měr*, volba $S_k= m^{-1}$ (pro $k=1,\dots,m$) ztotožňuje hodnoty $m.\hat{\mu}_{1+},\dots,m.\hat{\mu}_{n+}$ resp. $m.\hat{\mu}$ s typem *úhrnných měr*. V obou těchto případech jsou $\hat{\mu}_{+1},\dots,\hat{\mu}_{+m}$ dobrými *aproximacemi* specifických měr standardu (ve stejném smyslu jako $\hat{f}$ vůči f v odstavci 2).

Na explicitních odhadech z (18) lze zakládat testy hypotéz o rovnosti standardizovaných nebo úhrnných měr. Testové statistiky jsou obdobné jako ve stati 3, pouze $S_e/(n-1)$ nahrazujeme odhadem σ^2 z (19) a každé $P^j = R^j.S_+$ (pokud $\mu^{\wedge}_{i+}$ hraje roli vstupní proměnné f_i, vycházíme z předpokladu $\mathrm{Var}(\mu^{\wedge}_{i+}) = \sigma^2/(R_i.S_+)$, kde $S_+ = \Sigma\, S_k$). Z hlediska demografických interpretací má však význam zejména formální analýza reziduí. Platí-li (**), můžeme přímým výpočtem ověřit identitu

$$\mathrm{Var}(f_{ik} - \mu^{\wedge} - \alpha^{\wedge}_i - \beta^{\wedge}_k) = \sigma^2 \left(\frac{1}{R_i} - \frac{1}{R_+} \right)\left(\frac{1}{S_k} - \frac{1}{S_+} \right) . \qquad (20)$$

Odchylku f_{ik} od modelové hodnoty $f^{\wedge}_{ik} := (\mu^{\wedge}_{i+} + \mu^{\wedge}_{+k} - \mu^{\wedge})$ považujeme tedy za statisticky významnou, když

$$|f_{ik} - f^{\wedge}_{ik}| > u_{\alpha/2}.\tilde{\sigma} \left[\frac{(R_+ - R_i)(S_+ - S_k)}{R_i S_k R_+ S_+}\right]^{1/2} , \qquad (21)$$

kde $u_{\alpha/2}$ je kritická hodnota *normalizovaného normálního rozdělení* na hladině α.

Strukturální anomálie v plodnosti zjištěné reziduální analýzou v systému bilančních celků Severočeského kraje jsou vyznačeny na obr. symbolickými názvy statisticky významných věkových skupin (α=5%). Nejvýznamnější *úhrnné hodnoty* f_i (vyjádřené již dříve černou podkladovou barvou) se vysvětlují vesměs zvýšenou intenzitou plodnosti v nejdůležitější skupině žen '20-24 let'. V některých pánevních centrech (Ústí n. Labem, Most, Jirkov) je charakteristická zvýšená plodnost pro nejnižší věkovou kategorii ('15-19 let'), což ve světle demografických procesů 80.let naznačuje jisté vývojové zpoždění. Přesun intenzity plodnosti u města Liberce ve prospěch věku '25-29 let' identifikuje naopak vyspělejší demografickou strukturu (i vůči průměru-standardu České republiky).

5. POZNÁMKY

V odborné literatuře referující teoretické vlastnosti zobecněných lineárních odhadů se pro případ modelu (4) vychází z transformace

$$\underline{Y}^* = \underline{\Omega}^{1/2}\underline{Y} , \quad \underline{X}^* = \underline{\Omega}^{1/2}\underline{X} , \qquad (22)$$

která umožňuje vyjádřit Aitkenův odhad vektoru regresních koeficientů $\underline{\theta}$ jako standardní odhad *metodou nejmenších čtverců* v modelu

$$E\underline{Y}^* = \underline{X}^*\underline{\theta} , \quad \mathrm{Var}(\underline{Y}^*) = \sigma^2.\underline{I} . \qquad (23)$$

Úloha *zobecněné analýzy rozptylu* (s nula-jedničkovými maticemi $\underline{X}$) se tak transformuje na *obecný regresní model*, jehož numerické zpracování může být pochopitelně o něco pracnější. Pro váhy součinového typu ($Q_{ik} = R_i.S_k$) lze ovšem vycházet z explicitních vzorců (10), (18) resp. (19), které lze snadno napočítat v tabulkovém kalkulátoru (*spreadsheetu*).

Koeficient n^{-1} ve vzorci (7) pro odhad individuálního rozptylu každého regionálního ukazatele f_i je motivován nejen snahou o maximální výpočtové zjednodušení, ale i tím, že aproximační vážený průměr $f^{\wedge}$ na pravé straně v (6) může být fakticky interpretován jako *konstantní* hodnota $f = Ef_i$ 'známá' pro nadřazený územní celek. Pokud $f^{\wedge}$ chápeme jako *odhad* neznámého parametru f v modelu (2)+(3), je vhodnější hodnotit odchylku každé hodnoty f_i od $f^{\wedge}$ použitím přesného vzorce

$$\mathrm{Var}(f_i - \hat{f}) = \sigma^2 \frac{1}{P_i} \left(1 - \frac{P_i}{P_+} \right) . \qquad (24)$$

kam dosazujeme $\hat{\sigma}^2$ z (6) za σ^2. Při obecné volbě vah v *heteroskedastickém* modelu (14)+(15) je analogie k (20) resp. (24) vyjádřitelná jako

$$\mathrm{Var}(f_{ik} - \hat{f}_{ik}) = \sigma^2 \frac{1}{Q_{ik}} (1 - h_{ik}) , \qquad (25)$$

kde h_{ik} je (i,k)-tý diagonální prvek *projekční matice* příslušné transformovanému lineárnímu modelu (22). (Pro účely tohoto aplikačního experimentu byly hodnoty $h_{11},\ldots,h_{nm}$ generovány jako součást standardního výstupu *panelu pro mnohonásobnou regresi* programového produktu Solo - viz [6],str.73).

V souvislosti s testovými kritérii (12), (13) a (21) se implicitně předpokládá *normalita* vstupních demografických proměnných. Původně *symetrický* histogram regionálních ukazatelů *úhrnné plodnosti* (f_i) se však po transformaci (22) jeví výrazně *sešikmeným*. Tato vlastnost odpovídá geografickým zkušenostem na poli jiných empirických výzkumů, ale i intuitivní představě o skutečném charakteru teoretického rozložení (jde o nezáporné veličiny s interpretačně omezeným *nosičem*). Z teoretického hlediska to může být záminkou pro ještě složitější metodologická řešení. Pragmaticky vzato jde však o přirozenou licenci nahrazení *prostého průměru* demograficky lépe reprezentativním *průměrem váženým*.

Indikace statistické významnosti reziduí zobrazené na obr.4 odpovídají vahám proporcionálním středním stavům žen pro každé pole vstupní tabulky ($Q_{ik} = P_{ik}$). Téměř shodné výsledky dostáváme i při jiné výše zmíněné volbě vah vedoucí k lépe interpretovatelným odhadům parametrů (viz také na obr.2). Použití přesných kritických hodnot odvozených od normálního rozdělení je z hlediska finálního kartografického vyhodnocení ovšem zanedbatelné. Jeho smyslem je věcně zdůvodněná detekce signifikantních regionálních anomálií zvoleného ukazatele (*míry*) podložená statisticky unifikovanou metodikou. Analogické aplikace pro jiná časová období by měly identifikovat pouze podstatné strukturální odchylky, které při tradičním (neměnném) kartografickém škálování musí být zkresleny systematickými změnami hladin ukazatelů, vnitřní variability systému i stochastickými vlivy. Na rozdíl od jiných demografických postupů umožňuje popsaná metoda formálně zahrnovat libovolné další regresní proměnné, jejichž vliv na demografické procesy je dosud velmi nejasný.

Literatura:

[1] Běláček J.: *Struktura a vývoj základních demografických charakteristik okresů Liberec a Jablonec n.Nisou v podrobnějším územním členění*. Výzk.studie pro VÚVA Praha, SEÚ ČSAV, Ústí n.Labem, 1992

[2] Běláček J.: *Regionální diferenciace plodnosti v Severočeském kraji*. Demografie 4/34, 1992

[3] Cipra T.: *Ekonometrie*. Skripta MFF UK Praha, 1983

[4] Fiala T.: *K problému hodnocení statistické významnosti srovnávacích indexů úmrtnosti*. Sborník 'Úmrtnost a stárnutí obyvatelstva', ČSDS, Praha, 1985, str.32-37

[5] Hušková M.-Dupačová J.: *Analýza rozptylu*. Skripta MFF UK Praha, 1978

[6] *Návod k používání programového produktu SOLO*. Ed.:Škrabková Z. KPMS MFF UK Praha - SVT LF UK Hradec Králové - PřF OU Ostrava, 1992

[7] Pavlík Z.-Rychtaříková J.-Šubrtová A.: *Základy demografie*. Academia Praha, 1986

[8] Syrovátka A.-Koschin F.: *Stáří matky a stav novorozence a kojence*. Demografie 3/31, str.220-227