MOŽNOSTI VYUŽITÍ PROGRAMU LOCCONTINGENCY K ANALÝZE REGIONÁLNÍCH ROZDÍLŮ

Daniel ČERMÁK

Abstract: *LOCCONTINGENCY programme – a tool for the analysis of the regional differences LOCCONTINGENCY programme is originally a tool for completing unknown fields of contingency table with known marginal frequencies. Unknown fields are computed on the basis of another, already known, contingency table. Use of LOCCONTINGENCY programme is tested on the different data sets and its usability is verified by comparing model results with real results. This article shows possible ways how LOCCONTINGENCY programme could be used.*
Key words: *regional disparities, model, standardisation, electoral preference*

Program LOCCONTINGENCY vznikl jako součást dnes již ukončeného projektu, jenž se zabýval vlivem regionálních rozdílů na politické orientace voličů (Kostelecký 2000-2002).

Byl vytvořen pro použití v situacích, kdy se nám nedostává údajů v podrobném třídění II. stupně za malé územní celky, za tyto malé celky známe pouze výsledky třídění stupně I. Někdy se ocitneme v situaci podobné následující modelové – známe podrobně třídění obyvatel podle věku a rodinného stavu za libovolnou velkou územní jednotku, např. Českou republiku, avšak neznáme toto podrobné třídění za jednotku nižšího řádu, např. Ústecký kraj, za něj známe pouze strukturu podle věku a strukturu podle rodinného stavu, avšak neznáme vzájemný vztah mezi nimi. To je situace, kdy můžeme využít program LOCCONTINGENCY a ten nám na základě známých údajů z územní jednotky vyššího řádu a neúplných údajů z jednotky nižšího řádu dopočítá údaje chybějící.

Celý postup výpočtu je dokumentován následujícími tabulkami. Celý program LOCCONTINGENCY je naprogramován pomocí jazyka VISUAL BASIC jakožto makro programu MS Excel. Jedná se o sled několika po sobě jdoucích kontingenčních tabulek. Do první tabulky zadáme konkrétní počty[1] obyvatel ČR roztříděné podle rodinného stavu a věku.

1 *Můžeme zadávat také relativní četnosti nebo procenta.*

Mgr. Daniel Čermák
Sociologický ústav Akademie věd České republiky, Jilská 1, 110 00 Praha 1,
e-mail: daniel.cermak@soc.cas.cz

Tabulka 1: ČR 2001 – obyvatelstvo podle věku a rodinného stavu

	Single	Married	Divorced	Widowed	Unknown	Sum
0-4	448040					448040
5-9	565803					565803
10-14	651144					651144
15-19	681401	3148	81	8	3340	687978
20-24	711224	125592	8383	265	7496	852960
25-29	377771	433479	52913	1361	9298	874822
30-34	123953	476573	83038	2877	7020	693461
35-39	73859	508269	99551	5535	6194	693408
40-44	54129	503612	106588	10086	5216	679631
45-49	49405	589788	128284	20697	5674	793848
50-54	39452	612499	120729	38420	5369	816469
55-59	25292	478805	80290	54109	3705	642201
60-64	15682	341623	45225	69813	2348	474691
65-69	12746	283461	34469	103172	2045	435893
70+	31679	402889	58144	481679	6471	980862
Unknown	946	1077	205	136	1247	3611
Sum	3862526	4760815	817900	788158	65423	10294822

Zdroj: SLDB 2001, ČSÚ.

Do druhé tabulky zadáme pouze známé struktury obyvatel Ústeckého kraje podle věku a podle rodinného stavu jakožto marginální četnosti.

Tabulka 2: Ústecký kraj 2001 – obyvatelstvo podle věku a rodinného stavu (marginální četnosti)

	Single	Married	Divorced	Widowed	Unknown	Sum	Marginal
0-4							38827
5-9							47705
10-14							53659
15-19							56689
20-24							69868
25-29							73596
30-34							56550
35-39							55252
40-44							53095
45-49							64384
50-54							70713
55-59							49535
60-64							37248
65-69							32119
70+							68330
Unknown							208
Sum							827778
Marginal	321242	357802	79933	61339	7462	827778	

Zdroj: SLDB 2001, ČSÚ.

V dalším kroku program vychází z již známého vztahu mezi dvěma proměnnými z tabulky 1 a předpokládá, že stejný vztah bude existovat také mezi dvěma proměnnými na úrovni jednotky nižšího řádu a dopočítá pomocí metody minimální informační divergence[2] co nejpřesněji všechna políčka kontingenční tabulky tak, jak je uvedeno v tabulce 3.

2 *Tento výpočetní algoritmus se snaží vypočítat očekávané rozložení jež je statisticky co nejpodobnější rozložení skutečnému.*

Tabulka 3: Ústecký kraj 2001 - odhad četností podle věku a rodinného stavu

	Single	Married	Divorced	Widowed	Unknown	Sum
0-4	38827					38827
5-9	47705					47705
10-14	53659					53659
15-19	56046	242	8	1	392	56689
20-24	58464	9666	836	24	878	69868
25-29	32233	34630	5477	126	1131	73596
30-34	10247	36890	8328	258	827	56550
35-39	5954	38367	9736	483	712	55252
40-44	4268	37183	10196	861	586	53095
45-49	4038	45134	12719	1832	661	64384
50-54	3450	50149	12807	3638	669	70713
55-59	1973	34978	7599	4572	412	49535
60-64	1245	25384	4354	6000	266	37248
65-69	942	19614	3090	8257	215	32119
70+	2143	25513	4770	35280	624	68330
Unknown	48	51	13	7	90	208
Sum	321242	357802	79933	61339	7462	827778

Je zde samozřejmě vždy možnost, že vztah mezi proměnnými na úrovni územních jednotek nižšího řádu bude odlišný, než v případě jednotek řádu vyššího. Proto byly obdobně testovány závislosti řad proměnných rozličných typů – ekonomických, demografických, postojových apod. za účelem zjištění, v kterých případech můžeme výpočtu věřit a ve kterých ne. Nyní se ještě vrátíme k našemu příkladu. Jelikož je to příklad pouze modelový, tak ve skutečnosti známe rozložení obyvatel podle věku a rodinného stavu v Ústeckém kraji. Skutečné výsledky nyní srovnáme s výsledky odhadovanými a zjistíme, zda jsou statisticky významně odlišné či nejsou.

Tabulka 4: Ústecký kraj – rozdíl mezi skutečnými a očekávanými četnostmi

	Single	Married	Divorced	Widowed	Unknown	Sum
0-4	0					0
5-9	0					0
10-14	0					0
15-19	-125	36	3	-1	87	0
20-24	38	1	-93	3	51	0
25-29	750	-684	-147	17	63	0
30-34	334	-664	308	27	-5	0
35-39	30	-407	383	19	-25	0
40-44	-47	-145	127	37	29	0
45-49	-67	71	-62	74	-16	0
50-54	-151	644	-375	-37	-81	0
55-59	-142	415	-254	19	-37	0
60-64	-71	210	49	-182	-7	0
65-69	-59	165	-18	-44	-43	0
70+	-494	358	78	67	-9	0
Unknown	4	1	1	1	-8	0
Sum	0	0	0	0	0	0

V jednotlivých políčkách tabulky vidíme rozdíl mezi skutečným a očekávanými počty obyvatel. Možná se zdají tyto rozdíly velké, ale musíme si uvědomit, že v políčkách s největšími odchylkami se nachází několik desítek tisíc lidí. Podrobíme-li tabulky skutečných a očekávaných četností chí-kvadrát testu, zjistíme, že se od sebe významně neliší, naopak jsou téměř totožné.

Na programu LOCCONTINGENCY byly obdobným způsobem otestovány i další vztahy mezi demografickými strukturami a odhadované počty se vždy jen velmi málo odlišovaly od skutečnosti. A to i v případech, kdy rozložení marginálních četností u jednotek nižšího řádu bylo dosti odlišné. Je ovšem důležité podotknout, že územní jednotka nižšího řádu musí být stále dostatečně velká, aby se zamezilo vlivu sezónních výkyvů, jež mohou zásadně ovlivnit testované vztahy (např. počet zemřelých podle věku na úrovni malých obcí). Vyzkoušeli jsme také možnost dopočítat strukturu jevů u územních jednotek vyššího řádu na základě rozložení jevů u jednotek řádu nižšího. Také v tomto případě výpočet fungoval u demografických ukazatelů velmi dobře.

Samozřejmě program LOCCONTINGENCY není jediná možnost, jak získat chybějící údaje o struktuře v kontingenční tabulce. Můžeme využít také metody výběrového šetření, kdy bychom měli získat požadované údaje také. Musíme ovšem zvolit dostatečně velký reprezentativní vzorek, na němž zmíněné šetření provedeme. Pomineme-li důvody, jež ovlivňují reprezentativnost výběrových šetření (např. nízká návratnost dotazníků), nesmíme zapomenout ani na nákladnost výběrových šetření. Avšak není účelem tohoto textu zavrhnout výběrová šetření, pouze ukázat jiné cesty k řešení dané problematiky. Na dalších příkladech si ukážeme, kdy můžeme program LOCCONTINGENCY využít k analýze dat (především agregátních), ale také jako doplněk k výběrovým šetřením.

Další možnou oblastí využití programu jsou srovnávání regionů přibližně stejné velikosti (ale není to bezpodmínečně nutné). Při tomto srovnávání vycházíme z toho, že známe vzájemný vztah mezi proměnnými v obou regionech a chceme zjistit, zda vzájemné vztahy mezi proměnnými mají obdobný charakter. Pokud nám program na základě znalosti vztahů v jednom regionu dopočte očekávané četnosti v kontingenční tabulce pro druhý region, a ty se budou shodovat s četnostmi skutečnými, můžeme říci, že mezi regiony není rozdíl ve sledovaném vztahu mezi jevy. V případě, že sledovaný vztah mezi jevy není v obou regionech totožný, zjistíme významný rozdíl mezi skutečnými a očekávanými četnostmi buď ve všech políčkách kontingenční tabulky, či aspoň v některých z nich. Tento postup je vlastně jakousi obdobou standardizace, kdy jednu strukturu vztahu 2 proměnných upravujeme na základě znalosti struktury druhé.

Konkrétně si to ukážeme na následujícím příkladu, kdy srovnáváme vztah mezi věkem rodinným stavem u mužů v České republice 2002 a Slovenské republice 2001.

Tabulka 5: Odhadované rozložení mužů dle věku a rodinného stavu ve SR 2001

	Single	Married	Divorced	Widowed	Sum
15-19	219799	271	2		220071
20-24	215367	18416	510	15	234308
25-29	122780	89131	4618	95	216624
30-34	45488	123135	9951	218	178793
35-39	27208	146149	14472	561	188390
40-44	21306	158712	16682	1093	197793
45-49	17170	163537	17556	2179	200443
50-54	10539	137396	13247	3242	164424
55-59	5371	99187	7699	3849	116106
60-64	3440	81767	4620	5114	94941
65-69	2456	68938	2844	7306	81544
70-74	1800	53792	1805	9844	67241
75-79	1237	34291	998	11073	47599
80-84	447	11801	322	6462	19032
85+	284	5567	145	7038	13033
Sum	694692	1192089	95472	58089	2040342

Tabulka 6: Skutečné rozložení mužů dle věku a rodinného stavu ve SR 2001

	Single	Married	Divorced	Widowed	Sum
15-19	219199	789	27	56	220071
20-24	206207	27307	714	80	234308
25-29	110381	100670	5432	141	216624
30-34	46047	122694	9813	239	178793
35-39	32513	141927	13446	504	188390
40-44	26319	154524	15718	1232	197793
45-49	20266	160316	17587	2274	200443
50-54	12675	135326	13217	3206	164424
55-59	7382	97014	7806	3904	116106
60-64	4882	79786	5009	5264	94941
65-69	3487	67597	3140	7320	81544
70-74	2531	52983	1924	9803	67241
75-79	1681	33879	1097	10942	47599
80-84	602	11692	353	6385	19032
85+	520	5585	189	6739	13033
Sum	694692	1192089	95472	58089	2040342

Zdroj: SODB 2001, ŠÚ SR.

Tabulka 7: Rozdíl mezi skutečným a odhadovaným rozložením mužů dle věku a rodinného stavu ve SR 2001

	Single	Married	Divorced	Widowed	Sum
15-19	-600	518	25	56	**0**
20-24	-9160	8891	204	65	**0**
25-29	-12399	11539	814	46	**0**
30-34	559	-441	-138	21	**0**
35-39	5305	-4222	-1026	-57	**0**
40-44	5013	-4188	-964	139	**0**
45-49	3096	-3221	31	95	**0**
50-54	2136	-2070	-30	-36	**0**
55-59	2011	-2173	107	55	**0**
60-64	1442	-1981	389	150	**0**
65-69	1031	-1341	296	14	**0**
70-74	731	-809	119	-41	**0**
75-79	444	-412	99	-131	**0**
80-84	155	-109	31	-77	**0**
85+	236	18	44	-299	**0**
Sum	**0**	**0**	**0**	**0**	**0**

Z tabulek vyplývá, že největší odlišnost mezi Slovenskem a Českem je v počtu svobodných mužů a v počtu ženatých mezi 15 a 29 lety. Na základě znalosti vztahu mezi věkem a rodinným stavem v Česku program očekával na Slovensku větší počet svobodných mužů a samozřejmě současně s tím menší počet ženatých, než ve skutečnosti byl. Z toho plyne, že mezi 15 a 29 rokem na Slovensku muži uzavírají sňatek častěji než v Česku.

Samozřejmě nemusíme srovnávat pouze dva rozdílné regiony, je také možné srovnávat jeden a ten samý region v různých obdobích. Můžeme např. srovnat počty narozených dětí podle pořadí narození a věku matky v České republice v letech 1993 a 2003 a program nám opět ukáže, kde nestaly největší změny ve vztahu mezi těmito dvěma proměnnými. Abychom přešli od demografických proměnných k jinému typu, ukážeme si jedno srovnání v čase na volebních datech. Jedná se volby do Poslanecké sněmovny v letech 1998 a 2002, konkrétně závislost počtu odevzdaných hlasů politickým stranám na krajích. Jelikož od roku 2000 existuje v České republice nové krajské zřízení s odlišným počtem krajů, bylo nutné přepočítat hlasy z voleb 2002 na staré rozdělení krajů, aby bylo možné tyto volby porovnat. Rozdělení hlasů v krajích je velmi důležité pro rozdělení poslaneckých mandátů, neboť ty se odvozují od volebních výsledků v krajích a ne na celostátní úrovni.

Tabulka 8: Odhadované výsledky voleb 2002 na základě výsledků 1998[3]

	Praha	Středo-český	Jihočeský	Západo-český	Severo-český	Východo-český	Jiho-moravský	Severo-moravský	Sum
ODS	39,5	25,9	24,8	23,9	22,6	24,5	20,1	19,6	24,5
ČSSD	23,3	30,3	29,1	30,2	32,0	28,3	29,7	36,2	30,2
KSČM	12,5	19,2	19,3	19,4	20,9	17,0	20,7	18,8	18,5
KDU-ČSL+US+DEU	17,0	11,7	14,9	12,2	9,8	16,8	17,1	12,2	14,3
NEZ	1,7	3,3	3,1	2,9	2,9	3,3	2,3	3,2	2,8
SZ	1,7	2,3	2,4	3,0	3,0	2,7	2,1	2,4	2,4
SPR-RSČ	0,6	1,0	0,9	1,1	1,5	0,9	0,9	1,1	1,0
OTHER	3,8	6,4	5,6	7,2	7,4	6,4	7,2	6,7	6,4
Sum	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

3 *Volební výsledky KDU-ČSL, US a DEU byly pro rok 1998 sečteny, neboť v roce 2002 kandidovaly společně v koalici. Nástupnická organizace SPR-RSČ kandidovala v roce 2002 pod názvem Republikáni Miroslava Sládka.*

Tabulka 9: Skutečné výsledky voleb 2002

	Praha	Středo-český	Jiho-český	Západo-český	Severo-český	Východo-český	Jiho-moravský	Severo-moravský	Sum
ODS	33,8	26,8	25,6	25,2	25,2	24,6	20,3	20,1	24,5
ČSSD	25,9	31,5	30,5	30,0	28,7	28,5	30,2	34,4	30,2
KSČM	11,1	18,7	18,4	20,4	22,9	16,8	19,3	20,5	18,5
KDU-ČSL+US+DEU	18,5	11,9	12,6	10,8	9,2	15,7	17,9	12,9	14,3
NEZ	2,1	1,9	3,4	2,5	2,9	3,3	2,9	3,1	2,8
SZ	2,5	2,4	2,1	2,6	2,8	2,2	2,4	2,2	2,4
SPR-RSČ	0,5	1,0	0,9	1,0	1,7	1,0	0,9	1,1	1,0
OTHER	5,6	5,8	6,6	7,3	6,7	7,9	6,2	5,9	6,4
Sum	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Tabulka 10: Rozdíl skutečné minus očekávané výsledky voleb 2002

	Praha	Středo-český	Jiho-český	Západo-český	Severo-český	Východo-český	Jiho-moravský	Severo-moravský	Sum
ODS	-0,69	0,10	0,05	0,10	0,26	0,02	0,06	0,09	0,00
ČSSD	0,31	0,14	0,09	-0,02	-0,34	0,03	0,11	-0,32	0,00
KSČM	-0,17	-0,05	-0,06	0,07	0,20	-0,02	-0,29	0,31	0,00
KDU-ČSL+US+DEU	0,18	0,03	-0,16	-0,11	-0,06	-0,15	0,15	0,12	0,00
NEZ	0,06	-0,16	0,02	-0,03	0,00	-0,01	0,12	-0,01	0,00
SZ	0,10	0,01	-0,02	-0,03	-0,02	-0,06	0,06	-0,04	0,00
SPR-RSČ	-0,01	0,00	0,00	0,00	0,02	0,00	-0,01	0,00	0,00
OTHER	0,22	-0,07	0,07	0,01	-0,07	0,19	-0,20	-0,15	0,00
Sum	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Podíváme-li se na očekávané četnosti, vidíme, že program předpokládal v Praze lepší výsledek ODS a naopak horší výsledek KDU-ČSL a především ČSSD. Naopak očekával lepší výsledek ČSSD v Severních Čechách a na Severní Moravě, kde si dost na její úkor polepšili komunisté. Chí-kvadrát test považuje tabulky očekávaných a skutečných četností za shodné.

Schopnost programu LOCCONTINGENCY předpovídat výsledky voleb v krajích využil Tomáš Kostelecký (Kostelecký, 2005) ve své práci zabývající se předvolebními odhady. Zatímco agentury na výzkum veřejného mínění prováděly některé předvolební výzkumy[4] preferencí na oddělených reprezentativních vzorcích respondentů v každém kraji, Kostelecký využil reprezentativního průzkumu za celou republiku a na základě znalosti pouze marginálních četností – volebních preferencí za celou ČR a odhadované volební účasti v krajích – odhadl volební preference v krajích. Předpokládal určitou kontinuitu volebního chování mezi voliči v krajích, bez nepředvídatelné zásadní změny ve veřejném mínění před volbami, jež by dramaticky ovlivnila volební chování voličů. Pro všechny odhady výsledků voleb do poslanecké sněmovny a do krajských zastupitelstev dostal výsledky, které se přiblížily skutečnosti o něco více, než předvolební výzkumy preferencí, prováděné v jednotlivých krajích. Z tohoto pohledu se jeví jako výhodná myšlenka použít pro odhad volebních preferencí jeden velký celostátní reprezentativní výzkum a poté ho pomocí programu LOCCONTINGENCY rozpočítat do jednotlivých krajů na základě znalosti výsledků předchozích voleb a předpokladu vysoké míry kontinuity volebních preferencí voličů. Díky provedení jednoho celostátního výzkumu namísto čtrnácti oddělených regionálních, lze získat řádově levnější a přitom stále přesné výsledky. Tato problematika si však zaslouží ještě další a podrobnější testování.

4 *Prováděly však i celonárodní průzkumy volebních preferencí, ty využil Kostelecký k získání marginálních četností při odhadech volebních výsledků.*

Shrneme-li naše zkušenosti s použitím programu LOCCONTINGENCY, můžeme říci, že při odhalování vztahů uvnitř kontingenční tabulky a jejich následnému využití pro odhad neznámých hodnot se jedná o velmi užitečný nástroj. Nejvyšší přesnosti dosahuje při odhalování vztahů, jež nepodléhají snadno krátkodobým výkyvům a jsou stabilní v čase – obvykle „tvrdá" data získaná z censů, demografická data, některé postoje obecnějšího charakteru apod. Program lze také využít pro porovnávání odlišností ve vztazích mezi dvěma stejnými proměnnými ve dvou různých regionech. Další možnosti využití programu a zpřesňování stávajících postupů je předmětem dalšího zkoumání.

Tento příspěvek byl zpracován v rámci výzkumného projektu „Vliv politické kultury, sociálně ekonomických a institucionálních faktorů na rozdíly ve fungování českých regionů" podpořeným Grantovou agenturou České republiky na období 2004-2006 (č.proj.: 403/04/1300).

Literatura

KOSTELECKÝ, T., D. ČERMÁK 2003. „Výběrová šetření a analýza agregátních dat – diskuse na téma použitelnosti různých přístupů v komparativních analýzách poltického chování." *Sociologický časopis* 39 (4): 529-550.

KOSTELECKÝ, T., D. ČERMÁK 2004. „Vliv teritoriálně specifických faktorů na formování politických orientací voličů." *Sociologický časopis* 40 (4): 469-487.

KOSTELECKÝ, T. 2005. „Model predikce výsledků voleb – alternativní přístup k odhadům volebních výsledků ve volebních obvodech." *Sociologický časopis* 41 (1): 79-101.

VAJDA, I., K. VRBENSKÝ 2001. „Manuál programu LOCCONTINGENCY." Sociologický ústav, Praha.

VAJDA, I. 2001. „Adaptation of Contingency Tables to Local Marginals." Institute of Sociology, Prague.

POSSIBILITIES OF THE LOCCONTINGENCY PROGRAMME FOR ANALYSIS OF REGIONAL DISPARITIES

Summary

This article shows three possible ways how LOCCONTINGENCY programme could be used. 1. Unknown fields of contingency table can be completed for sub-region of one particular region when only marginal frequencies are known in sub-region. Unknown fields are computed on the basis of distribution of frequencies from that particular region. 2. For comparison of regional differences. It is a kind of standardisation. 3. For estimation and prediction of distribution of fields in contingency table on the basis of known distribution in the past.

Recenzovali: RNDr. Radoslav Klamár, PhD.
RNDr. Monika Ivanová