Výběrová šetření a analýza agregátních dat – diskuse na téma použitelnosti různých přístupů v komparativních analýzách politického chování

TOMÁŠ KOSTELECKÝ, DANIEL ČERMÁK*
Sociologický ústav AV ČR, Praha

Surveys and Aggregate Data Analysis – A Discussion of the Usability of Different Approaches in a Comparative Analysis of Political Behaviour

Abstract: Researchers who make comparative analyses of political behaviour in sub-national territorial units must often choose between the use of aggregate data and surveys. The use of surveys has long been considered the superior option, as it allows researchers to avoid the danger of ecological fallacy connected with the use of aggregate data, but it is also an extraordinarily expensive option. The article elaborates the pros and cons of both methodological approaches, and suggests the use of a method that seeks to combine the advantages of both. The method is based on combining the information from surveys on the national level with the aggregate data describing the sub-national territorial unit which are available from other sources, like electoral statistics or censuses. The method using the program LOCCONTINGENCY is tested on the data from Czech regions and its usability is verified by comparing model results with the results of surveys that were conducted in four model regions.
Sociologický časopis, 2003, Vol. 39, No. 4: 529–550

Úvod

Skutečnost, že se politická orientace voličů liší od místa k místu, je známa již od doby, kdy jak v západní Evropě, tak i v USA byla přijata zásada všeobecného volebního práva. Ze všech teorií pokoušejících se vysvětlit politické orientace voličů a příčiny jejich regionální proměnlivosti, můžeme rozlišit dva základní teoretické přístupy. První z nich – *kompoziční přístup* – je založen na předpokladu, že pro vysvětlení politické orientace jedince či populace nějaké teritoriálně definované jednotky je nutná především znalost strukturálních charakteristik sledovaného „objektu". Jestliže je objektem sledování jedinec, jde především o to nalézt charakteristiky definující jeho/její pozici ve společenské struktuře nebo jeho/její příslušnost k politicky specifické skupině. Pokud je objektem sledování populace regionu (města, sousedství), pak je klíčovou informací struktura dané populace. Územní rozdíly politických

* Veškerou korespondenci zasílejte na adresu: RNDr. Tomáš Kostelecký, CSc., Sociologický ústav AV ČR, Jilská 1, 110 00 Praha 1, e-mail: kostel@soc.cas.cz, resp. dcermak@soc.cas.cz

orientací jednoduše odrážejí územní proměnlivost ve složení populace. Druhý – *kontextový přístup* – zdůrazňuje důležitost prostorového kontextu před individuálními charakteristikami voličů. Místní podmínky jsou považovány za nejdůležitější faktor formující politickou orientaci voličů. Osobní postavení jedinců ve společenské sféře pouze „mírně pozměňuje" jejich politická rozhodování. V důsledku toho se kontextový přístup soustředí na nutnost zkoumat místně specifickou kulturu, vztahy mezi jednotlivcem a regionem, kde bydlí, stejně jako na vztahy mezi různými skupinami voličů, které společně s aktuálními historickými událostmi reprezentují prostředí, ve kterém jsou tato voličská rozhodnutí učiněna.

Analýzy příčin regionálních rozdílů politického chování

První pokus analyzovat příčiny regionálních rozdílů v politickém chování se objevil již na začátku 20. století ve Francii. A. Sigfried [1913] porovnal mapy ukazující prostorové rozdíly v geologii, podnebí, ekonomice a sociální struktuře s výsledky voleb a studoval jejich vzájemné korelace. V 50. a 60. letech docházelo, v důsledku kvantitativní revoluce v sociálních vědách a snazšího přístupu k výkonnější počítačové technice k intenzivnímu rozšiřování analýz tohoto typu ve všech vyspělých zemích [V. O. Key 1955, O. Rantala 1967, Capecchi – G. Galli 1969, D. E. Butler – D. E. Stokes 1969 a mnoho dalších]. Slavná práce S. M. Lipseta a S. Rokkana [1967], kteří přišli s konzistentní teorií vývoje stranických systémů v západní Evropě, je považována za průlom na poli teorie. Podle autorů může být existence politických stran, jejich vývoj a základní charakter jejich vnitřních vztahů vysvětlena jako více či méně přesná reflexe existujících společenských struktur.

Hlavním rysem všech těchto studií (možná s výjimkou amerických) je základní idea, že politická orientace voličů (obvykle měřená volebními preferencemi) je primárně ovlivněna pozicí jednotlivce ve společenské struktuře. Po mnoho let bylo obecně přijímáno tvrzení D. Butlera a D. Stokese [1969], že regionální rozdíly politických preferencí pouze odrážejí prostorové rozdíly ve složení populace a vše ostatní jsou jen „detaily a balast". Proto se výzkum orientoval hlavně na studium různých dělících linií ve společnosti (*cleavages*), které stály v pozadí odlišné politické orientace jednotlivců i regionů. Dělicí linie ve společnosti byly poznávány jak na úrovni individuální (prostřednictvím sociologických výběrových šetření), tak na úrovni agregátní (prostřednictvím analýzy volebních statistik a sčítání lidu). Výběr „vysvětlujících faktorů" se liší stát od státu, a to nejen díky historickým rozdílům, ale i specifickým rysům společenských struktur a politických systémů. Prostorové změny politických orientací byly obvykle nejvíce vztahovány k prostorovým změnám sociálních struktur. V některých zemích bylo nejdůležitějším strukturálním faktorem náboženské vyznání (Nizozemí), etnická nebo jazyková skupina (Belgie) nebo příslušnost k jednotlivým sektorům ekonomiky („zemědělství versus průmysl" – Norsko).

Situace se změnila až v 70. a 80. letech, kdy byl v mnoha zemích zaznamenán nárůst regionálních rozdílů ve volebním chování voličů [J. C. Archer – F. M. Shelley – P. J. Taylor 1990]. Zvětšování územních rozdílů ve volebních výsledcích, které ne-

mohlo být vysvětleno prostorovými změnami společenských struktur, vedlo k formulaci teorie nerovnoměrného vývoje [T. Nairn 1977]. Tato teorie tvrdí, že ekonomický vývoj je stále více územně nerovnoměrný, a proto chování voličů, reprezentující „politickou odpověď" na místní vývoj, musí být také územně nerovnoměrné. Ve stejné době bylo také zaznamenáno postupné rozvolňování vztahu mezi ekonomickým statusem jedince a jeho voličskými preferencemi, pozorování třídních struktur tak ztratilo tradiční svou primární důležitost pro identifikaci politických orientací voličů v regionech a městech [W. L. Miller 1982, S. Berglund – U. Lindström 1982]. Sám S. Rokkan, ve snaze vysvětlit důvody těchto změn, přišel s úplně odlišným vysvětlením, které chápe nárůst územních změn volebního chování jako důsledek zpolitizování periférií a „účinku skrytého teritorialismu" [S. Rokkan – D. W. Urwin 1983]. F. Nielsen [1980] tvrdil, že příčinou prostorových změn politického chování je modernizace, která vede k větší solidaritě uvnitř různých skupin. Prostorová změna je pouze sekundární a je způsobena tím, že různé skupiny přirozeně obývají různé „niky". Zcela odlišné vysvětlení naznačil J. Agnew [1987], který zdůraznil specifičnost vývoje konkrétní lokality a důležitost aktuálních historických událostí pro formování politických postojů.

V 90. letech dokázaly některé studie značnou míru politického regionalismu i na území České republiky [Jehlička a Sýkora 1991; Kostelecký 1994, 1996, 2001, 2002]. Zájem objasnit možné příčiny tohoto fenoménu vedl formulování výzkumného projektu nazvaného „Vliv územně specifických faktorů na formování politické orientace voličů", který byl řešen v Sociologickém ústavu AV ČR v Praze. Jako hlavní cíle projektu byly určeny:

– Zjistit, zda-li (a v jakém rozsahu) je politická orientace jednotlivce ovlivňována politickou, sociální a ekonomickou situací v regionu, ve kterém žije a jeho specifickou kulturou.
– Určit aktuální územně-specifické faktory ovlivňující formování politické orientace voličů a porovnat jejich vliv s vlivem „klasických strukturálních" faktorů.

Dva teoretické přístupy k testování hypotéz

Pokud chce kdokoli testovat validnost kompoziční a kontextové hypotézy v praxi, nutně potřebuje informace o voličském chování jednotlivých voličů společně s jejich osobními charakteristikami, a také, kvůli znalosti prostorového kontextu, další informace o místě regionu, ve kterém volič žije. V první fázi je nutné rozhodnout, jaká metodologie může být použita pro analýzu. Existují dvě zcela odlišné metodologické tradice, které jsou v tomto ohledu k dispozici: analýza agregátních dat a analýza sociologických výběrových šetření získaných na individuální úrovni[1].

[1] Upozorňujeme na skutečnost, že existují i výběrová šetření na úrovni agregátních dat (např. výběrová šetření vycházející se zápisů okrskových volebních komisí). U tohoto typu šetření se ovšem do značné míry kombinují metodologické nevýhody obou výše uvedených metodologií, proto se jím nebudeme v dalším textu zabývat.

Pozornost „průkopníka" na tomto poli, Andre Siegfrieda, byla věnována regionální analýze volebních výsledků ve vztahu k různým potenciálně vysvětlujícím faktorům ve Francii [Siegfried 1913]. Ve své studii se Siegfried plně spoléhal na agregátní data o voličském chování, socioekonomické struktuře a dalších potenciálních vysvětlujících faktorech, které byly shromážděny na „nadindividuální" úrovni (volební obvod, obec, kraj,...). Důvod, proč používal agregátních data, byl zcela jednoduchý: jiné údaje nebyly k dispozici. Použití agregátních dat ke studiu politického chování bylo ovšem po dlouhou dobu zcela obvyklé i mimo politickou geografii. Také Herbert Tingstein [1937] použil ve svém výzkumu politického chování analýzu agregátních dat jako nástroj pro studium chování jednotlivce. V 50. letech ovšem popularita užívání agregátních dat jako vstupních dat pro analýzu prudce poklesla. Do značné míry to byla reakce na článek „Ekologické korelace a chování jednotlivců", který byl publikován W. S. Robinsonem [1950] v *American Sociological Review*. Robinson přesvědčivě ukázal, že statistický vztah, který je významný na agregátní úrovni (například pro data z jednotlivých obcí nebo volebních okrsků), nemusí být významný na úrovni jednotlivce a naopak. Byl dokonce schopný nalézt příklady, v nichž byla pozitivní korelace na jedné úrovni doprovázena negativní korelací na úrovni jiné. Této chybě se později dostalo označení *„ecological fallacy"*. Práce s agregátními daty má však ještě celou řadu dalších omezení. Problémem je samotná dosažitelnost dat. Vědec je totiž zcela závislý na sčítání lidu nebo jiných typech oficiálních statistik, které pochopitelně neshromáždily právě ta data, které by výzkumník pro řešení svého úkolu potřeboval. Pro celý stát, resp. velké (a tudíž málo početné) územní jednotky je k dispozici nejvíce údajů, se zmenšováním územního rozsahu jednotek sledovaných dat ubývá. Jiný typ problému s agregátními daty představuje skutečnost, že počet agregátních charakteristik, které asociují s volebními preferencemi a dalšími ukazateli politického chování, je tak velký, že je složité je logicky integrovat do vnitřně konsistentních modelů. Vysvětlující proměnné jsou navíc velmi různého typu. Největší komplikaci ovšem představuje skutečnost, že jednotlivé charakteristiky, které slouží v agregátních analýzách jako vysvětlující proměnné, jsou často silně vzájemně korelovány (problém multikolinearity) a je u nich obtížně rozhodnout, co je příčinou a co následkem.

Odklon vědců od užívání agregátních dat pro politickou analýzu byl také samozřejmě uspíšen rychlým vývojem na poli výběrových šetření [Lazarsfeld et al. 1948]. Bylo vždy jasné, že výběrová šetření jsou schopná přinést údaje o jednotlivci, a jsou proto vhodná pro analýzu chování jednotlivce. Úspěšné pokusy George Gallupa a jeho následníků, kteří relativně přesně předpověděli výsledky voleb na základě rozhovorů s relativně omezeným počtem respondentů, podpořily přesvědčení, že výběrová šetření jsou také vhodná pro analýzu makrostruktur. Tento vývoj vedl ve svém důsledku k obecnému rozšíření názoru, že výběrová šetření jsou metodologicky nadřazena analýze agregátních dat.

Shora uvedené tvrzení může být považováno za pravdivé i navzdory řadě pokusů mnoha vědců vyřešit „problém ekologické inference", to je, slovy S. R. Thomsena [2000], zjistit, „jak odvodit volební chování jednotlivce z agregátních dat, jako

jsou volební výsledky nebo sčítání lidu". Po desetiletí trvající úsilí vyvinout a zlepšit techniky, které by spolehlivě odvozovaly vztahy mezi proměnnými na individuální úrovni z agregátních dat tam, kde žádná individuální data nejsou k dispozici, bylo směřováno k vyvinutí nástroje, který by sloužil jako alternativa v případech, kdy „nejlepší volba" (čti: výběrové šetření) nebyla z nějakých příčin možná. Časem bylo vytvořeno velké množství různých metod pro řešení ekologické inference, které byly inspirovány pokrokem ve statistice a umožněny narůstajícím výkonem výpočetní techniky. S. D. Withers [2001] zmiňuje nejméně deset modelů, které byly vyvinuty a testovány s lepšími či horšími výsledky. Ty zahrnovaly „klasickou" Goodmanovu techniku ekologické regrese [1953], „model sousedství" [Freedman et al. 1991], agregátní složený multinominální model [Brown a Payne 1986], ekologicko-logitový model [Thomsen 1987], metodu maximizace entropie [Johnston et al. 1982], regresi dvojité rovnice [Groffman 1997], metodu rozkladového přístupu [Lupia a McCue 1990] a kvadratické kontextové efektové modely [Owen a Grofman 1997]. V nedávné době přitáhla značnou pozornost kniha harvardského profesora veřejné správy Garry Kinga, který ve své knize *A solution to the ecological inference problem* [King 1997] popsal metodu, o níž tvrdí, že je „řešením problému ekologické inference, rekonstrukce chování jednotlivce z agregátních dat". Kingem navržené metodě se dostalo mnohých pochval pro novátorský a kreativní přístup k problému a jim navržená metoda byla brzy přijata jako standard většinou vědců ze společenských věd [W. K. Tam 1998; H. Reynolds 1998; N. L. Beck 2000; S. D. Withers 2001], kteří raději používají k řešení svých vědeckých otázek již vytvořené statistické metody, než aby se pokoušeli o vyvinutí nových. Kingova metoda se ovšem stala také předmětem velké kritiky, především z řad statistiků a dalších specialistů na metodologii [Freedman, Klein, Ostland a Roberts 1998; Freedman et al. 1999; K. F. McCue 2001].

Metodologických problémů spojených s užíváním agregátních dat pro analýzu politického chování je skutečně mnoho. Předtím ovšem, než přitakáme tvrzení, že „výběrová šetření jsou metodologicky nadřazena analýze agregátních dat", jsme se rozhodli podrobněji prozkoumat eventuální nedostatky a problémy vyplývající z použití výběrového šetření pro naplnění cílů našeho výzkumného projektu. Nejdříve jsme zhodnotili výhody a nevýhody klasické metody dotazníkových šetření pro výzkum vztahů mezi volebním rozhodováním, individuálními charakteristikami respondentů a kontextovými charakteristikami jednotlivých regionů. Zjistili jsme, že výběrová šetření, přinejmenším ve své reálné a ne „ideální" podobě, mají také poměrně značné množství nedostatků. Mezi problémy výběrových šetření, které jsou velmi dobře známé a prozkoumané, patří existence tzv. „výběrové chyby". Výběrová chyba je nevyhnutelnou součástí každého výběrového šetření, protože informace získáváme pouze od vzorku cílové populace a ne od populace celé. Výběrovou chybu můžeme odhadnout zcela přesně, jelikož závisí na několika známých parametrech – na velikosti vzorku, velikosti cílové populace a použité hladině významnosti. Pokud výzkum založený na rozhovorech s 1000 respondenty zjistil, že v některém regionu podporovalo stranu ABC 50 % voličů, můžeme počítat s výběrovou chybou ± 3,1 % na 95% hladině významnosti. Jinými slovy: můžeme si být na 95 % jisti, že voličská podpora strany ABC v populaci regionu se nachází v intervalu od 46,9 %

do 53,1 %. Jestliže je podíl stoupenců strany ABC nižší nebo vyšší než 50%, výběrová chyba je menší, ale nesnižuje se přímo úměrně ke změně velikosti podílu stoupenců strany ABC. (Výběrová chyba má hodnotu ± 2.8%, jestliže je podíl stoupenců 30% nebo 70%, a ± 1,9 %, jestliže je podíl 10% nebo 90%...).

Potíže s výběrovou chybou se zvětšují v případě, kdy potřebujeme studovat regionální rozdíly v politických preferencích. Teoreticky není použití výběrových šetření pro tento typ úlohy žádný problém, jde pouze o to, uskutečnit reprezentativní výběrová šetření se stejnou velikostí vzorku v každém ze studovaných regionů. Ve skutečnosti jde o komplikaci vážnou, neboť prudce vzrůstá počet potřebných respondentů, a tím i cena výzkumu. Prakticky jediná výběrová šetření, týkající se politického chování v různých regionech, která jsou vedena tímto nákladným způsobem, bývají předvolební průzkumy. Navzdory značnému zájmu médií a jejich štědré podpoře, pracují tyto typy předvolebních průzkumů obvykle s nižším počtem respondentů v každém regionu, než je obvyklé při „standardním předvolebním průzkumu". Například v České republice největší předvolební průzkumy zaměřené na odhad volebních výsledků ve všech 14 krajích pracují přibližně s 500 respondenty v každém kraji. Se snižujícím se počtem respondentů ze standardních 1000 na 500 výběrová chyba narůstá až na ± 4,4 %.

Ale výběrová chyba není, bohužel, jedinou chybou ovlivňující výsledky výběrového šetření. Vše, co zde bylo dosud řečeno o výběrové chybě, je pravdivé pouze v případě, že dotazovaní respondenti byli vybráni metodou náhodného (pravděpodobnostního) výběru, což je procedura, která dává všem jednotlivcům stejnou šanci, aby byli vybráni do vzorku. V praxi ale žádné vzorky používané pro výběrová šetření v České republice nevznikají prostým náhodným výběrem, protože zákon o ochraně osobních dat přísně zakazuje užívání registru populace pro komerční účely, tedy i pro výzkumy politického chování. U všech výběrových šetření, kde se vyžaduje použití náhodného výběru respondentů, se ve skutečnosti používá výběr „v mezích možností co nejbližší" prostému náhodnému výběru. Obvykle se jedná o vícestupňový pravděpodobnostní výběr, kterým se v prvním kroku vybírá náhodně domácnost, a ve druhém kroku, opět náhodně, jeden z členů domácnosti. Hlavními problémy tohoto vícestupňového výběru jsou kvalita samotné výběrové opory použitá pro výběr domácnosti (většinou totiž nezahrnuje všechny domácnosti), a pak také skutečnost, že jednotliví respondenti mají díky dvoukrokovosti výběru různou pravděpodobnost, že budou vybráni do vzorku (lidé žijící v malých domácnostech mají větší pravděpodobnost, že budou vybráni, než ti, co žijí v domácnostech větších).

I když necháme stranou problém výběrové chyby, mají výběrová šetření s politickými tématy celou řadu dalších chyb, které vycházejí ze způsobu provedení výzkumu. Možná právě proto, že je těžké o těchto nedostatcích diskutovat „jazykem" statistické teorie, je řada z nich přívrženci a uživateli výběrových výzkumů považována za něco nepodstatného a zanedbatelného. Mnoho nedostatků výběrových výzkumů jednoduše pramení z praktických problémů, a přitom mají hluboké důsledky pro kvalitu a spolehlivost výsledků. Jednu z největších potíží, které výzkumníci

čelí, představuje sama skutečnost, že výzkumy jsou velmi drahé, zvláště pak ty, které používají náhodný výběr respondentů. Reakce firem, zabývajících se výběrovými šetřeními, na rostoucí náklady byla tedy zcela logická: téměř kompletně nahradily náhodný výběr výběrem kvótním, zejména u předvolebních výzkumů, kde se očekává, že řada výsledků bude publikována již během volební kampaně. U kvótního výběru nejsou respondenti vybíráni náhodně, ale nejdříve je zkoumána struktura cílové populace (obvykle s využitím statistických údajů ze sčítání lidu o věku, pohlaví, vzdělání a regionálním rozdělení populace), a pak je vzorek respondentů vybrán tak, aby proporčně reprezentoval všechny tyto skupiny definované kvótami odvozenými z populace. Na rozdíl od metody náhodného výběru, kde je úlohou tazatele zpovídat respondenty, kteří byli pro dotazování přesně a nedvojznačně definováni, v kvótním výběru má tazatel k dispozici pouze orientační popis, že má uskutečnit následující interview v obci či městě XY, s (například) jedním mužem s univerzitním vzděláním mladším třiceti let, se dvěma vyučenými muži ve věku mezi 31 a 45 lety, jednou ženou se středním vzděláním ve věku mezi 46 a 60 lety, dvěma ženami se základním vzděláním staršími 60 let atd. I za předpokladu, že tazatelé přesně dodrží instrukce, je konečné rozhodnutí, s kterou konkrétní osobou uskutečnit rozhovor, pouze v jejich rukou. Ačkoli existuje celá řada dalších dodatečných pravidel, které umožňují kontrolovat průběh dotazování a udržet kvalitu výběru na přijatelné úrovni (například nedovolit tazatelům dotazovat se stále stejných respondentů, zpětné kontroly práce tazatelů...), v principu není možné zabránit tomu, aby tazatelé nevyužívali ty nejjednodušší a nejpohodlnější způsoby, jak nalézt vhodného respondenta. Jelikož většina tazatelů provádí dotazování pod časovým tlakem, hledá respondenty nejprve mezi svými přáteli, sousedy, „přáteli přátel", prostě a jednoduše: v rámci svých sociálních skupin či sítí. V důsledku toho ovšem mohou být výsledky zkreslené, neboť jsou ovlivněny nadreprezentací respondentů pocházejících ze stejných společenských skupin, v nichž se pohybuje tazatel. Tato systematická chyba, která by mohla být s trochou nadsázky označována jako „zkreslení způsobené přáteli tazatelů", je nejpravděpodobnější příčinou, proč mají výsledky výzkumu produkované některými agenturami, zabývajícími se výzkumem veřejného mínění a politickým výzkumem, sklon se systematicky lišit od výsledků jiných společností navzdory tomu, že používají stejné metody výběru a aplikují stejné způsoby kontroly práce tazatelů, které jsou doporučeny standardy ESOMAR/WAPOR.

Významným problémem, spojeným obecně s metodologií výběrových šetření a speciálně s výzkumy na politická témata, je snižující se ochota respondentů účastnit se průzkumů. V současné době se v České republice míry návratnosti dotazníků při výběrovém šetření s politickým obsahem, které užívají náhodného výběru, uvádějí v rozmezí 50 % až 60 % (u výběrových šetření používajících kvótního výběru se většinou míry návratnosti neuvádějí). Tak nízká míra návratnosti neznamená nic menšího, než že názory a postoje téměř poloviny populace nejsou zachyceny. Z metodologického hlediska je podstatné, že lidé, jejichž názory se nepodařilo zachytit nebo kteří odmítli odpovídat, nejsou v žádném případě náhodným vzorkem populace. Lidé z některých specifických sociálních skupin jsou pravidelně ve výběrovém souboru podreprezentováni. Často se jedná o mladé, chudé nebo žijící na okraji spo-

lečnosti, ale také o podnikatele, manažery a ostatní lidi s časově velmi náročnou prací. Jinou tvář téhož problému představuje skutečnost, že také sama volební účast je relativně nízká, a má v posledním desetiletí tendenci se neustále snižovat. Proto se agentury provádějící předvolební výzkumy stále více zajímají o to, jak co nejpřesněji odhadnout, který z respondentů se skutečně zúčastní voleb. Evidentně nestačí se respondenta jednoduše zeptat, zda má v úmyslu volit či ne. Pod tlakem veřejného mínění, které stále považuje účast ve volbách za občanskou ctnost, odpovídá mnoho respondentů na otázku po zamýšlené volební účasti kladně, ale následně se skutečných voleb nezúčastní. Metodologickým problémem je především to, že lidé, kteří se účastní průzkumů a z jejichž stranických preferencí se dělají předvolební odhady výsledků voleb, nejsou nutně ti, kteří se skutečně zúčastní voleb. Ačkoliv se tyto dvě skupiny z velké části překrývají (nevíme ovšem, do jaké míry, neboť díky autostylizaci respondentů se skutečná účast ve volbách nedá spolehlivě odhalit ani v *ex post* prováděných povolebních průzkumech), paradoxně měříme volební chování a jeho vztahy k osobním charakteristikám respondenta, jeho hodnotám, postojům a cílům na části populace, která není zcela identická s tou částí populace, která skutečně volí[2].

Existují ještě některé další obtíže spojené výlučně s průzkumy voličského chování. Volební rozhodnutí je pro mnoho respondentů tak soukromou záležitostí, že při rozhovoru s tazatelem zatají své skutečné rozhodnutí. Každý respondent může jednoduše odmítnout odpovědět na jednotlivé otázky vztahující se k volbám, případně odpovědět „nevím" nebo dát záměrně nesprávnou odpověď. Toto je pravděpodobně mnohem vážnější problém v postkomunistických zemích než v zemích se stabilní demokracií. Navzdory dekádě demokratického vývoje někteří lidé stále váhají nad tím, zda svůj politický postoj vyjádří otevřeně. Tento typ zkreslení je pak zpětně viditelný při porovnání výsledků předvolebních výzkumů se skutečnými výsledky voleb. Velmi často je výzkumy podhodnocena podpora stran, které jsou obecně považovány za extremistické, nebo stran, které jsou silně kritizovány nejvlivnějšími médii. Podobný typ zkreslení se vyskytuje i v povolebních výzkumech – podpora stran, které volby vyhrály, bývá v povolebních výzkumech vyšší než ve skutečných volbách a naopak.

[2] Potíže s odhadem, kdo se vlastně voleb účastní a kdo nikoliv, se samozřejmě netýkají specifických výzkumů uskutečňovaných v době voleb dotazováním voličů, kteří právě vycházejí z volební místnosti – tzv. „exit polls". Použitelnost těchto výzkumů pro akademickou práci je ovšem silně omezena skutečností, že jde primárně o komerční výzkumy. Tyto výzkumy se zpravidla musí soustředit na rychlost sběru a zpracování dat, používají jen krátké dotazníky s malým počtem proměnných. Data z exit polls jsou navíc často nepřístupná veřejnosti.

Analýza dat a testování hypotéz

Přehled problémů spojených jak s analýzou agregátních dat, tak s metodologií výběrových šetření naznačuje, jak těžké rozhodování čeká badatele, mají-li si pro řešení svého vědeckého problému vybrat mezi jedním či druhým metodologickým přístupem. V našem případě jsme měli to štěstí, že jsme měli k dispozici dostatek grantových peněz, abychom mohli použít obě metody. V rámci výzkumu „Region a politika" bylo v říjnu 2000 uskutečněno velké výběrové šetření s více než 4200 respondenty. Šetření se soustředilo na postižení vlivu regionálních specifik na formování politických orientací voličů. Proto byl výběr vzorku záměrně navržen tak, jako by šlo o pět paralelně probíhajících výzkumů. Prvním z nich byl representativní průzkum dospělé populace České republiky (N = 1 143), ostatní čtyři byly průzkumy vedené ve čtyřech modelových regionech (s N větším než 800 v každém regionu). Ve všech případech byl použit kvótní výběr, kvótami byly věk, pohlaví, vzdělání a míra urbanizace. Modelové regiony byly záměrně vybrány tak, aby reprezentovaly 4 regiony s rozdílnými politickými tradicemi v České republice. Region „Praha", který se skládá z města Prahy a sousedních okresů Praha-západ a Praha-východ a pokrývá pražskou aglomeraci, je nejbohatší region v České republice, má dlouhodobě nejnižší nezaměstnanost, a je také volební baštou pravicově orientovaných stran. Region „Ostrava" v severovýchodní části země, sestávající z města Ostravy a sousedícího okresu Karviná, je typický vysoce urbanizovaný průmyslový region, který prochází obtížným restrukturalizačním procesem, s vysokou nezaměstnaností, volební základna levicových stran, zejména ČSSD. Region „Zlín" v jihovýchodní části České republiky skládající se z okresů Zlín, Uherské Hradiště a Hodonín, je regionem s vysokým podílem katolíků, vysokým zastoupením venkovské populace, s průměrně fungující ekonomikou založenou na množství středních a malých firem lehkého průmyslu, tradiční bašta KDU–ČSL. Region „Louny" leží západně od Prahy, zahrnuje okresy Louny, Kladno, Beroun a Rakovník, s tradiční volební podporou komunistů, s vysokým podílem venkovské populace, těžebním průmyslem, těžkým průmyslem ve městech a vysokou mírou nezaměstnanosti. V modelových regionech byly respondentům položeny stejné otázky jako reprezentativnímu vzorku české populace. To znamená, že je možné přímo porovnávat výsledky národního průzkumu s výsledky regionálních průzkumů, neboť ve všech případech byly informace sebrány na úrovni jednotlivých respondentů.

Protože jsme si byli dobře vědomi toho, že tak velký průzkum by mohlo být v budoucnu těžké opakovat, a při vědomí toho, jaké problémy souvisejí s použitím výběrového šetření pro regionální analýzu politického chování, začali jsme přemýšlet nad metodou, která by zkombinovala ty nejlepší části obou metod: přesnost, robustnost a relativní levnost agregátních dat z volebních statistik, sčítání lidu a jiných statistických zdrojů a unikátnost a nezkreslenost informací o vztazích různých charakteristik nasbíraných na individuální úrovni rozhovory tazatelů s jednotlivými respondenty. Pokusili jsme se nalézt proceduru, která by zkombinovala data z výběrového šetření s agregátními statistickými údaji charakterizujícími strukturu populace a volebními údaji. Cílem tohoto postupu bylo odhadnout neznámé informa-

ce o politickém chování populace v regionech s použitím údajů o vztazích mezi sociální strukturou a volebními preferencemi, zjištěných národním reprezentativním výzkumem, a údajů o aktuální sociální struktuře a voličských preferencích v modelových regionech. Na tomto místě je důležité upozornit, že hledané řešení nebylo snahou o „vylepšení" klasické Goodmanovy techniky ekologické regrese nebo Kingovy metody ekologické inference. Tyto techniky, a všechny jim podobné, se totiž snaží o odvození informací o individuálním chování jednotlivce z agregátních dat za situace, kdy nejsou k dispozici žádná jiná data než agregátní. Cíl, který si vytyčil náš tým, byl skromnější (možná však realističtější): snažili jsme se nalézt metodu, která by odvodila informace o individuálním chování jednotlivce z agregátních dat popisujících určitý region a z informací o individuálním chování jednotlivce zjištěných reprezentativním výběrovým šetřením na vzorku jiné než cílové populace. Tato „jiná než cílová populace" musela být ovšem v nějakém vztahu k cílové populaci zkoumaného regionu – buď šlo o populaci velikostně nadřazené územní jednotky (celý stát), nebo o populaci jiného regionu, o které se dá předpokládat, že se podobá cílové populaci zkoumaného regionu.

Jedno z řešení popisovaného úkolu bylo vyvinuto v rámci probíhajícího výzkumného projektu [podrobnosti viz Vajda, 2001; Vajda and van der Meulen, 2001]. Použitím metody minimalizace informační divergence byl připraven základ pro první verzi programu LOCCONTINGENCY [Vajda and Vrbenský, 2001], který umožnil provést první praktické testy využitelnosti tohoto teoretického řešení. Tento program umožnil odhadnout neznámé hodnoty v jednotlivých buňkách kontingenční tabulky v situacích, kdy známe pouze marginálie tabulky (řádkové a sloupcové součty) a máme jinou kontingenční tabulku stejné velikosti, z které mohou být informace o vztazích mezi proměnnými (v řádcích a sloupcích) odvozeny. Z čistě matematického hlediska může existovat jedno řešení, nekonečně mnoho řešení nebo žádné řešení takové úlohy. Jestliže existuje právě jedno řešení, program nalezne toto řešení. Jestliže existuje nekonečný počet řešení, program vytvoří kontingenční tabulku, která je „statisticky nejpodobnější" známé kontingenční tabulce ve smyslu minimální informační divergence. Jestliže neexistuje žádné řešení, program vytvoří kontingenční tabulku, která je nejpodobnější kontingenční tabulce, která je pouze nepatrně odlišná od původní kontingenční tabulky (maximální rozdíl 1% v každé buňce).

Fakt, že máme k dispozici jak data z národního výběrového šetření, tak data z regionálních výběrových šetření v modelových regionech, nám umožní porovnat statistický odhad založený na agregátních datech, provedený programem LOCCONTINGENCY, s výsledky skutečného výběrového šetření v příslušných regionech. Následující příklady v principu dokumentují, jak program pracuje a jak statistické odhady vypadají v porovnání s výsledky regionálních průzkumů. Jedna z otázek v dotazníku zněla" *„Jakou politickou stranu jste volil v posledních parlamentních volbách v roce 1998?"*, na jiném místě byli respondenti dotazováni na svůj věk. Výsledky pocházející z výběrového šetření na národním vzorku jsou sumarizovány v následujících tabulkách.

Tabulka 1. ***„Jakou politickou stranu jste volil/a v posledních parlamentních volbách v roce 1998?“*** **Výsledky průzkumu reprezentativního pro českou populaci.**

Název strany	%
Občanská demokratická strana (ODS)	28,8
Česká strana sociálně demokratická (ČSSD)	30,7
Komunistická strana Čech a Moravy (KSČM)	14,9
Unie svobody (US)	4,3
Křesťansko-demokratická unie – Československá strana lidová (KDU-ČSL)	10,0
Ostatní strany	11,2
Celkem	100,0

Zdroj: *Výzkum „Region a politika“*, N = 745 (zbytek do celkového počtu 1141 respondentů tvoří ti, kteří na danou otázku buď odmítnuli odpovědět, nebo odpověděli „nevím“ či „netýká se mě to“).

Tabulka 2. ***„V jakém roce jste se narodil/a?“*** **Výsledky průzkumu reprezentativního pro českou populaci.**

Název strany	%
Občanská demokratická strana (ODS)	28,8
Česká strana sociálně demokratická (ČSSD)	30,7
Komunistická strana Čech a Moravy (KSČM)	14,9
Unie svobody (US)	4,3
Křesťansko-demokratická unie – Československá strana lidová (KDU-CSL)	10,0
Ostatní strany	11,2
Celkem	100,0

Zdroj: *Výzkum „Region a politika“,* N = 1141

Protože data, která máme k dispozici, byla získána na individuální úrovni, můžeme také vytvořit kontingenční tabulku kombinující věk a stranické preference respondentů (viz tabulku 3).

Předpokládejme, že známe volební podporu pro hlavní politické strany a věkovou strukturu populace v regionu Zlín (tedy řádkové a sloupcové součty v kontingenční tabulce podobné tabulce 3), ale nevíme, jaká je volební podpora jednotlivých stran u lidí různého věku v témž regionu (tedy neznáme hodnoty v buňkách

Tabulka 3. Vztah mezi věkem a stranickými preferencemi respondentů. Kontigenční tabulka byla vytvořena z dat průzkumu reprezentativního pro českou populaci (v procentech z celkového počtu validních odpovědí).

Věková skupina	ODS	ČSSD	KSČM	US	KDU-ČSL	Ostatní	Celkem
18–29	5,2	5,6	0,9	1,4	1,3	1,7	**16,1**
30–44	9,6	8,2	1,9	1,3	2,0	3,2	**26,1**
45–59	9,3	9,6	4,9	0,9	3,0	2,9	**30,6**
60+	4,7	7,3	7,3	0,7	3,7	3,4	**27,3**
Celkem	**28,8**	**30,7**	**14,9**	**4,3**	**10,0**	**11,2**	**100,0**

Zdroj: *Výzkum „Region a politika"*, N = 743 (zbytek do celkového počtu 1141 respondentů tvoří ti, kteří na danou otázku buď odmítnuli odpovědět, nebo odpověděli „nevím" či „netýká se mě to").

Tabulka 4. Známé marginální četnosti neúplné kontingenční tabulky popisující věkovou strukturu a stranické preference populace regionu Zlín.

Věková skupina	ODS	ČSSD	KSČM	US	KDU-ČSL	Ostatní	Celkem
18–29							**17,2**
30–44							**26,0**
45–59							**30,1**
60+							**28,8**
Celkem	**24,5**	**31,7**	**8,9**	**6,3**	**19,0**	**9,6**	**100,0**

Poznámka: Hodnoty použité v tabulce pocházejí z výsledků regionálního výběrového šetření reprezentativního pro region Zlín.

kontingenční tabulky) – viz tabulku 4[3]. Součty v řádcích jsou velmi podobné součtu v řádcích v tabulce 3, což znamená, že populace regionu Zlín má podobnou věkovou strukturu jako populace České republiky. Na druhé straně součty ve sloup-

[3] Jsme si vědomi toho, že řádkové a sloupcové součty použité v tabulce pocházejí z výběrového šetření v regionu Zlín, a jako taková jsou sama zatížena určitou výběrovou chybou. Na tomto místě jsme je ovšem použili zcela záměrně, neboť chceme odhady úplné kontingenční tabulky vytvořené programem LOCCONTINGENCY porovnávat s úplnými kontingenčními tabulkami pocházejícími právě z výběrového šetření ve zlínském regionu. Pro tento účel bude lépe, pokud budou v obou srovnávaných úplných kontingenčních tabulkách stejné marginální součty. Pokud bychom chtěli program LOCCONTINGENCY využít k vytvoření

Tabulka 5. Statistické odhady hodnot v kontingenční tabulce vytvořené programem LOCCONTIGENCY. Tabulka popisuje vztah mezi věkem a stranickými preferencemi respondentů z regionu Zlín.

Věková skupina	ODS	ČSSD	KSČM	US	KDU-ČSL	Ostatní	Celkem
18–29	4,5	5,9	0,5	2,1	2,5	1,5	**17,2**
30–44	8,1	8,4	1,1	1,9	3,8	2,7	**26,0**
45–59	7,9	9,9	2,9	1,2	5,7	2,5	**30,1**
60+	4,0	7,5	4,3	1,0	7,0	2,9	**28,8**
Celkem	**24,5**	**31,7**	**8,9**	**6,3**	**19,0**	**9,6**	**100,0**

Poznámka: Hodnoty řádkových a sloupcových součtů v tabulce pocházejí z výsledků regionálního průzkumu reprezentativního pro region Zlín. Hodnoty ve vnitřních buňkách kontingenční tabulky jsou odhady provedené statistickým modelem.

cích v tabulkách 3 a 4 jsou zcela rozdílné. Populace regionu Zlín volí s téměř dvakrát větší pravděpodobností KDU-ČSL než populace České republiky a je významně méně ochotná volit ODS a KSČM.

Použili jsme program LOCCONTINGENCY k odhadnutí vztahů mezi věkem a stranickými preferencemi populace regionu Zlín (tedy hodnot v jednotlivých buňkách v kontingenční tabulce). Program vytvořil takovou kontingenční tabulku, která má stejné součty sloupců a řádků jako tabulka 4 a je „statisticky nejpodobnější" k údajům prezentovaným v tabulce 3. Výsledek odhadu je v tabulce 5.

Tabulka 6 ukazuje vztah mezi věkem a stranickými preferencemi v regionu Zlín, jak byly zjištěny při výběrovém šetření reprezentativním pro populaci zlínského regionu.

Tabulka 7 ukazuje rozdíl mezi hodnotami v tabulce 5 a tabulce 6, což je rozdíl mezi údaji pocházejícími z regionálního výzkumu reprezentativního pro populaci zlínského regionu a odhady vytvořenými statistickým modelem.

Z tabulky 7 je zřejmé, že odhady vytvořené statistickým programem byly velmi blízké číslům zjištěným při regionálním šetření. Ve většině buněk byl rozdíl mezi odhady a zkoumanými výsledky menší než 1 procentní bod. V převažující většině buněk se odhad vytvořený statistickým modelem pohyboval v intervalu definovaném výběrovou chybou vypočítanou na 95% hladině významnosti. Odhlédneme-li od těch buněk v kontingenční tabulce, kde byl jen minimální počet respondentů

odhadu úplné kontingenční tabulky pro region Zlín v situaci, kdy bychom neměli takovou tabulku z šetření provedeného na Zlínsku (což je zdaleka nejběžnější situace), použili bychom nejlepší dostupné informace o marginálních součtech z jiných zdrojů. Za sloupcové součty bychom dosadili skutečné výsledky voleb a jako řádkové součty bychom použili data z censu.

Tabulka 6. Vztah mezi věkovou strukturou a volebními preferencemi respondentů. Kontingenční tabulka vychází z výsledků reprezentativního průzkumu v regionu Zlín (v procentech z celkového počtu validních odpovědí).

Věková skupina	ODS	ČSSD	KSČM	US	KDU-ČSL	Ostatní	Celkem
18–29	5,7	5,5	0,2	1,8	1,8	2,0	**17,2**
30–44	8,7	8,5	1,3	1,5	3,5	2,6	**26,0**
45–59	6,5	10,5	3,3	2,0	4,2	3,5	**30,1**
60+	3,7	7,2	4,1	0,9	9,4	1,5	**28,8**
Celkem	**24,5**	**31,7**	**8,9**	**6,3**	**19,0**	**9,6**	**100,0**

Zdroj: *Výzkum „Region a politika“*, N = 582

Tabulka 7. Rozdíl mezi výsledky získanými reprezentativním průzkumem v regionu Zlín a odhadem hodnot v buňkách pomocí programu LOCCONTINGENCY (tabulka 6–tabulka 5).

Věková skupina	ODS	ČSSD	KSČM	US	KDU-ČSL	Ostatní	Celkem
18–29	1,2	–0,4	–0,3	–0,3	–0,7	0,5	0,0
30–44	0,5	0,1	0,2	–0,4	–0,3	–0,1	0,0
45–59	–1,4	0,6	0,4	0,8	–1,4	1,1	0,0
60+	–0,3	–0,3	–0,3	–0,1	2,4	–1,4	0,0
Celkem	0,0	0,0	0,0	0,0	0,0	0,0	0,0

Zdroj: *Vlastní výpočty*.

(hodnoty v kontingenční tabulce byly blízké 0, a proto byly modelem těžko odhadnutelné), byly největší odchylky mezi modelovým řešením a dotazníkovým šetřením nalezeny u nejstarších respondentů, kteří uvedli, že by volili KDU-ČSL. Zatímco model předpověděl, že by v populaci zlínského regionu mělo být 7 % takových respondentů, výsledky průzkumu uskutečněného v tomto regionu ukázaly, že to bylo 9,4 % respondentů ze zkoumaného vzorku 843 respondentů. V tomto případě byl modelový odhad mimo interval spolehlivosti údaje získaného z výběrového šetření (7,0 ± 1,7 % na 95% hladině významnosti).

Výsledek prvního testu nás vedl k mírnému optimismu ohledně použitelnosti statistického programu pro odhad hodnot ve vnitřních buňkách kontingenční tabulky v případě, když nám jsou známy pouze její marginální četnosti. Výše prezentovaný příklad ovšem nemusí být typický, protože populace zlínského regionu a populace České republiky má podobnou věkovou strukturu. Jinými slovy: součty v řádcích v tabulkách 3 a 4 jsou si velice blízké, populace zlínského regionu a populace Čes-

Tabulka 8. Vztah mezi vzděláním a stranickými preferencemi respondentů. Použita data z průzkumu reprezentativního pro českou populaci.

Vzdělání	ODS	ČSSD	KSČM	US	KDU-ČSL	Ostatní	Celkem
Základní	2,9	4,1	5,6	0,1	2,7	2,6	**18,1**
Vyučení	9,0	14,6	5,4	1,4	2,7	4,8	**37,9**
Středoškolské	12,4	9,6	3,5	1,9	2,5	2,9	**32,7**
Vysokoškolské	4,3	2,7	0,4	0,9	2,2	0,7	**11,3**
Celkem	**28,7**	**31,0**	**14,9**	**4,3**	**10,1**	**11,0**	**100,0**

Zdroj: *Výzkum „Region a politika“*, N = 742

Tabulka 9. Statistické odhady hodnot v kontingenční tabulce vytvořené programem LOCCONTIGENCY. Tabulka odhaduje vztah mezi vzděláním a volebními preferencemi respondentů v regionu Praha.

Vzdělání	ODS	ČSSD	KSČM	US	KDU-ČSL	Ostatní	Celkem
Základní	2,1	1,8	2,3	0,1	1,0	2,2	**9,5**
Vyučení	9,2	9,2	3,1	1,6	1,4	5,8	**30,2**
Středoškolské	18,2	8,5	2,8	3,0	1,7	5,0	**39,2**
Vysokoškolské	10,0	3,9	0,6	2,2	2,4	2,0	**21,0**
Celkem	**39,6**	**23,4**	**8,8**	**6,8**	**6,5**	**14,9**	**100,0**

Poznámka: Hodnoty řádkových a sloupcových součtů v tabulce pocházejí z výsledků regionálního průzkumu reprezentativního pro region Praha. Hodnoty ve vnitřních buňkách kontingenční tabulky jsou odhady provedené statistickým modelem.

ké republiky se liší jenom v jedné dimenzi u sledované dvojdimenzionální kontingenční tabulky. Proto byla u dalšího testu metody použita jiná proměnná (vzdělání místo věku) a jiný region (Praha místo Zlína). Tabulka 8 ukazuje výsledky reprezentativního výběrového šetření na populaci České republiky, pokud jde o vztah mezi stranickými preferencemi a vzděláním.

Obdobně jako v předchozím případě byla v druhém kroku jako vstupní data využita informace z tabulky 8 společně se známou vzdělanostní strukturou obyvatel a volebními preferencemi pražského regionu (viz marginálie v tabulce 9), a následně byl použit program LOCCONTINGENCY, který statisticky odhadnul hodnoty v jednotlivých buňkách kontingenční tabulky (viz tabulku 9).

Z porovnání marginálií v tabulkách 8 a 9 je jasné, že v tomto případě se populace České republiky a pražského regionu zcela liší, a to jak ve struktuře vzdělání, tak ve stranických preferencích. Populace pražského regionu je jedna z nejméně typických v České republice, a to jak v položce vzdělání, tak i stranické preference.

Tabulka 10. Vztah mezi vzděláním a volebními preferencemi respondentů. Kontingenční tabulka vychází z výsledků reprezentativního průzkumu v regionu Praha (v procentech z celkového počtu validních odpovědí).

Vzdělání	ODS	ČSSD	KSČM	US	KDU-ČSL	Ostatní	Celkem
Základní	3,2	1,4	1,1	0,4	0,7	2,7	**9,5**
Vyučení	9,2	8,3	3,8	2,3	1,8	4,9	**30,2**
Středoškolské	19,1	7,9	2,0	3,1	1,6	5,6	**39,2**
Vysokoškolské	8,1	5,8	2,0	1,1	2,3	1,8	**21,0**
Celkem	**39,6**	**23,4**	**8,8**	**6,8**	**6,5**	**14,9**	**100,0**

Zdroj: *Výzkum „Region a politika“*, N = 577

Tabulka 11. Rozdíl mezi výsledky získanými reprezentativním průzkumem v regionu Praha a odhadem hodnot v buňkách pomocí programu LOCCONTINGENCY (Tabulka 10–Tabulka 9).

Vzdělání	ODS	ČSSD	KSČM	US	KDU-ČSL	Ostatní	Celkem
Základní	1,1	–0,4	–1,2	0,2	–0,2	0,5	**0,0**
Vyučení	–0,1	–0,9	0,7	0,7	0,4	–0,9	**0,0**
Středoškolské	0,9	–0,6	–0,9	0,1	–0,1	0,6	**0,0**
Vysokoškolské	–2,0	1,9	1,4	–1,1	–0,1	–0,2	**0,0**
Celkem	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**	**0,0**

Zdroj: *Vlastní výpočty*.

Tabulka 10 ukazuje pro srovnání výsledky reprezentativního průzkumu uskutečněného v pražském regionu.

Následující tabulka 11 ukazuje rozdíly mezi hodnotami získanými reprezentativním šetřením v pražském regionu a modelovým odhadem.

Rozdíly mezi statistickými odhady a výsledky výzkumu jsou v pražském regionu o něco vyšší než v případě regionu Zlín. Ačkoli je stále pravda, že ve většině buněk kontingenční tabulky se hodnoty předpovězené modelem pohybují v mezích intervalu spolehlivosti výběrového šetření (na 95% hladině významnosti), buněk kontingenční tabulky, pro které to neplatí, je více než v předešlém případě. Je však možné, že problém není v samotném rozdílu mezi strukturou vzdělání obyvatel České republiky a obyvatel Prahy, ale ve skutečnosti, že vztahy mezi vzděláním a volebními preferencemi mohou být v Praze a v České republice odlišné. Je nutné si připomenout základní předpoklad, na kterém je statistický model založen: model vypočítá takovou regionální kontingenční tabulku, která je „statisticky nejvíce podobná“ známé národní kontingenční tabulce. Jinými slovy, statistická procedura užívá informaci o vztazích mezi vzděláním a volebními preferencemi z národního průzku-

mu a „aplikuje“ ji na známou strukturu vzdělání a volebních preferencí populace pražského regionu. Při bližším pohledu na tabulku 11 vidíme, že největší rozdíly mezi modelem a skutečností nacházíme u vysokoškolsky vzdělaných osob a lidí se základním vzděláním. Výsledky průzkumu ukazují, že pražští absolventi vysokých škol jsou více orientováni doleva (podpora ČSSD a KSČM), než předpovídal model na základě znalosti vztahů mezi vzděláním a volebními preferencemi v české populaci. U lidí ze základním vzděláním tomu bylo právě naopak – nejméně vzdělaní Pražané podporují pravicové strany (ODS, US) více, než předpověděl model.

Experimentování s modelem – shrnutí výsledků

Odhadování hodnot ve vnitřních buňkách kontingenční tabulky pomocí modelu a porovnání výsledků modelových odhadů s výsledky výběrových šetření v jednotlivých regionech pokračovalo podobným způsobem, jaký byl popsán výše v textu. Byla využita široká škála proměnných, které jsme měli k dispozici, v různých vzájemných kombinacích. Proměnné použité k testování modelu zahrnovaly demografické a socioekonomické proměnné (jako jsou vzdělání, věk a náboženské vyznání), stejně jako „politické“ proměnné (jako volební preference, volební účast, politickou sebeidentifikaci – umístění na škále levice-pravice, úroveň důvěry v osobu Václava Klause). K testování modelu byly použity také některé obecnější postoje respondentů, např. postoje k roli státu v boji s nezaměstnaností, hodnocení důležitosti problému nezaměstnanosti v politice, identifikace respondentů s místem, kde žijí (*„Kde se cítíte nejlépe? Ve svém bytě, ulici, obci, regionu,...“*), chápání pojmu svobody (*„Co pro vás znamená svoboda? Žádná bída..., možnost dělat to, co chci..., odpovědnost...“*), přiřazení respondenta k Inglehartovým materialistickým, postmaterialistickým nebo smíšeným hodnotám. Ve vzájemných kombinacích bylo testováno 18 dvojic proměnných (obvykle demografické a socioekonomické proměnné proti jiným). Byla testována stejná kombinace proměnných na datech ze všech čtyř regionů, takže celkový počet testů dosáhl 72. Ve všech případech byl použit stejný postup. Úplná kontingenční tabulka z národního průzkumu a marginální četnosti z regionálních průzkumů sloužily jako vstupní data pro statistický program. Potom byl vytvořen modelový odhad regionální kontingenční tabulky a ten byl porovnán s kompletními kontingenčními tabulkami pocházejícími z regionálních výběrových šetření.

Výsledky série testů je možno shrnout v následujících bodech:

– Modelové odhady byly obecně blízké výsledkům získaným výběrovým šetřením v regionech. Ve většině případů nebyly na 95 % hladině významnosti rozdíly mezi tabulkami statisticky významné (měřeno chí-kvadrát testem). Ve většině případů se odhady v buňkách, předpovězené statistickým modelem, neodlišovaly od údajů získaných výběrovým šetřením o více než výběrovou chybu.
– Zkoumáme-li míru podobnosti mezi modelovými odhady a výsledky regionálních výběrových šetření, zjišťujeme některé obecné pravidelnosti. Modelové odhady mají tendenci být více podobné výsledkům šetření tehdy, pokud je statistická metoda aplikována na postoje a názory, které se zabývají obecnými otázkami, jaký-

mi je např. politická sebeidentifikace na škále levice-pravice. Méně podobné jsou v případě specifičtějších otázek, jako je například hlasování v parlamentních volbách. Zdá se, že charakter vztahů mezi obecnějšími postoji a demografickými a socioekonomickými charakteristikami jednotlivce není příliš regionálně podmíněn. Protože je tento vztah „plošný" a nikoliv regionálně specifický, modelová řešení, která aplikují informace o vztazích mezi proměnnými, získané na národní úrovni, na regionálně specifické marginálie, jsou velmi podobná výsledkům regionálních dotazníkových šetření. Na druhé straně, čím více je charakter vztahu mezi proměnnými „místně specifický", tím méně jsou modelové předpovědi podobné výsledkům průzkumu.

– Největší rozdíl mezi hodnotami předpovězeným modelem a výsledky regionálního průzkumu byly zaznamenány v případě zlínského regionu. Možnou příčinou je skutečnost, že je zde daleko vyšší podíl katolíků (61 %) v porovnání s průměrem české populace (36 %) i s populací ostatních regionů (lounský region 14%, pražský region 25 %, ostravský region 28 %). Analýza ukazuje, že proměnná „náboženské vyznání" významně intervenuje do vztahů mezi volebním chováním a politickými postoji na jedné straně a socio-demografickými proměnnými, jakými jsou věk a vzdělání na straně druhé.
– Při porovnávání kontingenčních tabulek vypočtených pomocí statistického modelu a kontingenčních tabulek vycházejících z výsledků regionálních výzkumů nebyla identifikována žádná systematická tendence, pokud jde o rozdíly hodnot v jednotlivých buňkách srovnávaných tabulek. Pouze v některých případech bylo možné vysvětlit identifikované rozdíly mezi buňkami. V regionu s dlouhodobou křesťansko-demokratickou politickou tradicí model kupříkladu podcení podíl starých lidí podporujících křesťanské demokraty, podobně jako v regionu, kde jsou tradičně silní komunisté, model podcení podíl starých lidí volících komunisty. To znamená, že regionální kontext nejvíce ovlivňuje staré lidi. Ti mají tendenci chovat se více regionálně specificky než ostatní. Podobné pravidlo platí také pro lidi s nižším vzděláním.

Závěr

První kolo testů použitelnosti modelu přineslo nadějné výsledky. Metoda je schopná dobře odvodit informace o individuálním chování jednotlivců ve zkoumaném regionu v případě, že máme k dispozici agregátní data souhrnně popisující populaci zkoumaného regionu a informaci o vztazích mezi sledovanými proměnnými získanou výběrovým šetřením na vzorku „podobné populace", jako je populace zkoumaného regionu. Onou „podobnou populací" pak v praxi může být národní populace v zemi, v níž se nachází zkoumaný region. O národní populaci máme nejčastěji k dispozici potřebné informace, protože je každoročně uskutečněno mnoho sociologických výběrových šetření reprezentativních na úrovni celého státu, ale jen výjimečně se dělají šetření, jehož výsledky by byly reprezentativní pro populaci některého regionu.

„Podobnou populací" však může být i populace jiného regionu, pokud o ní máme informace získané výběrovým šetřením na individuální úrovni, a pokud víme, že se populace obou regionů „chovají podle stejné logiky". Samo o sobě není důležité, nakolik jsou si oba regiony podobné strukturou svého obyvatelstva nebo jestli jsou si regiony podobné například výsledky voleb. Klíčovým měřítkem podobnosti regionů je stejná logika vztahů mezi proměnnými zjištěná na individuální úrovni. Nezáleží na tom, že je v prvním regionu hodně vysokoškoláků, zatímco ve druhém málo, nebo že v prvním regionu získávají pravicové strany podstatně více hlasů než v druhém. Důležité je, aby v obou regionech shodně platilo, že vzdělanější lidé hlasují spíše pro pravicové strany. Pokud bychom ovšem tuto podmínku nedodrželi a použili při modelování „jako vzor" populaci regionu, která se od zkoumané populace odlišuje samotnou „logikou chování", mohou být modelové odhady nepřesné. Zjevnou nevýhodou takového použití modelu je skutečnost, že při výběru „podobného regionu", z jehož chování se bude odvozovat chování populace námi zkoumaného regionu, je nutno vycházet z předchozích zkušeností a již dříve provedených analýz. Podobně může být modelový odhad méně přesný, pokud budeme studovat populaci regionů, u nichž je logika vztahů mezi individuálními charakteristikami jednotlivce a jeho chováním významně odlišná od toho, co je normální v běžné populaci. Použití vytvořené statistické procedury samozřejmě není vázáno jen na zkoumání regionálních populací. Model lze stejně dobře použít i pro odhady na lokální úrovni. V tomto případě může být „podobnou populací" populace regionu, v níž se zkoumané město nachází, či jiné podobné město, o jehož obyvatelstvu máme informace získané sběrem dat na individuální úrovni.

Zdá se, že statistický model lze naopak s velkou mírou spolehlivosti využít za situace, kdy potřebujeme odhadnout vztahy na individuální úrovni v časovém okamžiku, kdy se neuskutečnilo žádné výběrové šetření, máme-li z tohoto roku k dispozici alespoň agregátní údaje o obyvatelstvu a zároveň existují data ze sociologického šetření, které se ve stejném území uskutečnilo dříve. Za předpokladu, že se logika chování zkoumané populace v mezidobí radikálně nezměnila (což je velmi nepravděpodobné), budou modelové odhady velmi přesné. Model je proto nepochybně využitelný pro odhadování chybějících údajů v časových řadách. Jestliže je známo, že např. vztahy mezi věkem a volební podporou komunistické strany jsou stabilní v čase, je možné s pomocí modelu „dopočítat" údaje v kontingenční tabulce v roce, kdy není z nějakého důvodu k dispozici kompletní kontingenční tabulka z výběrového šetření. Metoda je obecně tím spolehlivější, čím jsou vztahy mezi pozorovanými proměnnými univerzálnější a čím méně jsou ovlivněny regionálně specifickými faktory.

Vedle testů, které se plně spoléhají na data z výzkumu, byl proveden i omezený počet testů, v kterých se pracovalo s daty z odlišných pramenů, například z censu nebo z volebních statistik. Jako jeden ze zdrojů vstupních dat pro model např. sloužila kompletní kontingenční tabulka popisující volební výsledky v parlamentních volbách z roku 1996 podle jednotlivých volebních krajů. Druhým zdrojem informací pro model byla data o rozdělení hlasů mezi stranami (řádkové součty neúplné kontingenční tabulky) a rozdělení platných hlasů mezi volební obvody (sloup-

cové marginálie téže tabulky) v parlamentních volbách z roku 1998. Poté byl použit statistický model k odhadu výsledků jednotlivých politických stran v jednotlivých volebních krajích v roce 1998 (jednotlivé buňky kontingenční tabulky). Nakonec byly porovnány odhadované výsledky s reálnými výsledky jednotlivých politických stran ve volebních obvodech v roce 1998. Podobně byly odhadovány vztahy různých demografických dat, jako rodinný status nebo dojížďka za prací mimo obec trvalého pobytu, a věku. Výsledky potvrdily, že v případech, kde statistická procedura používá „tvrdá" data (tvrdší než z výběrových šetření), mají modelem navrhovaná řešení tendenci být velmi blízká reálným počtům. Robustnější informace o vztazích mezi testovanými proměnnými znamenají ve svém důsledku velmi přesné předpovědi modelu. Detailnější analýze potenciálního využití modelu pro úlohy pracující s „tvrdými daty" by měl být dán v budoucnu větší prostor.

TOMÁŠ KOSTELECKÝ *je vědeckým pracovníkem Sociologického ústavu AV ČR, vedoucím oddělení Lokální a regionální problémy. Ve svém výzkumu se věnuje především studiu vlivu teritoriálních faktorů na lidské chování, politické geografii, regionálním aspektům vývoje společnosti a komparativní politice. V roce 2002 vydal v nakladatelství Woodrow Wilson Center Press monografii* „Political Parties After Communism. Developments in East-Central Europe".

DANIEL ČERMÁK *je vzděláním sociolog a demograf. Pracuje jako odborný pracovník v Sociologickém ústavu AV ČR v oddělení Lokální a regionální problémy. Současně studuje jako doktorand sociologii na Filozofické fakultě UK. Zabývá se analýzou dat, způsoby jejich publikace a studiem regionálních rozdílů.*

Literatura

Agnew, J. 1987. *Place and Politics*, Winchester: Unwin Hyman.
Archer, J. C., F. M. Shelley 1986. *American Electoral Mosaics.* Washington: The Association of American Geographers.
Berglund, S., U. Lindstrom 1982. *Regional Centers and Beyond: Geographic, Economic and Political Impacts*. Umea: International Political Science Association.
Berglund, S., S. R. Thomsen 1990. *Modern Political Ecological Analysis*. Copenhagen: Abo Academic Press.
Brown, P. J., C. D. Payne 1986. „Aggregate Data, Ecological Regression, and Voting Transitions". *Journal of American Statistical Association* 81: 452–460.
Butler, D., D. Stokes 1969. *Political Change in Britain: Forces Shaping Electoral Choice*. London: Macmillan.
Capecchi, V., G. Galli 1969. „Determinants of Voting Behaviour in Italy: a Linear Causal Model of Analysis". Pp. 235–284 in M. Dogan and S. Rokkan (eds.), *Quantitative Ecological Analysis in the Social Sciences*. Cambridge: The MIT Press.
Freedman, D. A., S. P. Klein, M. Ostland, M. R. Roberts 1998. „A Solution to the Ecological Inference Problem (book review)". *Journal of the American Statistical Association* 93: 1518–1520.

Freedman, D. A., S. P. Klein, J. Sacks, C. A. Smith, C. G. Everett 1991. „Ecological Regression and Voting Rights". *Evaluation Review* 15: 673–711.
Freedman, D. A., M. Ostland, M. R. Roberts, S. P. Klein 1999. „Response to King's Comments". *Journal of the American Statistical Association* 94: 355–357.
Goodman, L. 1953. „Ecological Regression and the Behavior of Individuals". *American Sociological Review* 18: 663–666.
Grofman, B. 1995. „New Methods for Valid Ecological Inference". Pp. 127–149 in Eagles, M. (ed.), *Spatial and Contextual Models in Political Research*. London: Taylor & Francis.
Jehlička, P., L. Sýkora 1991. „Stabilita regionální podpory tradičních politických stran v Českých zemích (1920–1990)". *Sborník ČGS* 96 (2): 81–95.
Johnston, R. J., A. M. Hay, P. J. Taylor 1982. „Estimating the Sources of Spatial Change in Election Results". *Environment and Planning A* 14: 951–961.
Johnston, R. J., F. M. Shelley, P. J. Taylor (eds.) 1990. *Development in Electoral Geography*. London: Routledge.
King, G. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton, NJ: Princeton University Press.
Kostelecký, T. 1994. „Economic, Social and Historical Determinants of Voting Patterns: 1990 and 1992 Parliamentary Elections in the Czech Republic". *Czech Sociological Review* 2: 209–228.
Kostelecký, T. 1996. „The Results of the 1990 Parliamentary Elections in a Regional Perspective". Pp. 136–149 in Gabal, J. (ed.), *The 1990 Election to the Czechoslovakian Federal Assembly.* Berli: WZB.
Kostelecký, T. 2001. „Vzestup nebo pád politického regionalismu? [Rise or fall of political regionalism?]". Praha: Sociologický ústav AV ČR, *Working Papers* 2001 (9), 100 p.
Kostelecký, T. 2002. *Political Parties after Communism: Developments in East-Central Europe*. Washington, D.C.: Woodrow Wilson Center Press.
Key, V. O. 1955. „A Theory of Critical Elections". *Journal of Politics* 17: 3–18.
Lazarsfeld, P. F., B. Berelson, B., H. Gaudet 1948. *The People's Choice; How the Voter Makes Up His Mind in a Presidential Campaign*. New York: Columbia University Press.
Lipset, S. M., S. Rokkan, S. (eds.) 1967. *Cleavage Structures, Party Systems, and Voter Alignments*: *Cross-National Perspectives*. New York: The Free Press.
Lupia, A., K. McCue 1990. „Why the 1980s Measures of Racially Polarized Voting are Inadequate for the 1990s". *Law and policy* 12: 355.
McCue, K. F. 2001. „The Statistical Foundation of the EI Method". *American Statistician* 55 (2): 106–111.
Miller, W. L. 1982. „Variations in Electoral Behaviour in the United Kingdom". Pp. 224–250 in: *The Dimension in United Kingdom Politics*. London: Macmillan,
Nairn, T. 1977. *The Break-up of Britain*. London: New Left Books.
Nielsen, F. 1980. „The Flemish Movement in Belgium after World War II: A Dynamic Analysis". *American Sociologic Review* 45: 76–94.
Owen, G., B. Grofman 1997. „Estimating the Likelihood of Fallacious Ecological Inference: Linear Ecological Regression in the Presence of Context Effects". *Political Geography* 16: 675–690.
Rantala, O. 1967. „The Political Regions of Finland". *Scandinavian Political Studies* 2: 117–42.
Reynolds, H. 1998. „A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data, by G. King (book review)". *Journal of Regional Science* 38: 195–196
Robinson, W. S. 1950. „Ecological Correlations and the Behavior of Individuals". *American Sociological Review* 15: 351–357.

Rokkan, S., D. W. Urwin 1983. *Economy, Territory, Identity: Politics of West European Peripheries*. Beverly Hills: Sage.
Siegfried, A. 1913. *Tableau politique de la France de l'Ouest sous la Troisième Republique*. Paris: A.Colin.
Tam, W. K. 1998. „A Solution to the Ecological Inference Problem (book review)". *The Journal of Politics* 60: 1244–1246.
Taylor, P. J. 1985. *Political Geography. World-Economy, Nation-State and Locality.* Essex: Longman Group.
Thomsen, S. R. 1987. *Danish Elections 1920–1979: A Logit Approach to Ecological Analysis and Inference*. Arhu: Politica.
Thomsen, S. R. 2000. *Issue Voting and Ecological Inference.* Article published on the web page of the author.
Tingstein, H. 1937. „Political Behaviour; Studies in Election Statistics". *Stockholm Economic Studies 7*. London: P.S. King.
Vajda, I. 2001. *Adaptation of Contingency Tables to Local Marginals*. Praha: Institute of Sociology of the Academy of Sciences ČR.
Vajda, I., E. C. van der Meulen 2001. „On Minimum Divergence Adaptation of Discrete Bivariate Distributions to Given Marginals". [Manuscript, sent for publication in *Transactions of IEEE on Information Theory*].
Vajda, I., K. Vrbenský 2001. *Manuál programu LOCCONTINGENCY*. Praha: Sociologický ústav AV ČR.
Withers, S. D. 2001. „Quantitative Methods: Advancement in Ecological Inference". *Progress in Human Geography* 25 (1): 87–96.