VYŠETŘENOST ÚDAJŮ O DOJÍŽĎCE VE SČÍTÁNÍ LIDU V ROCE 2011 A JEJICH REKONSTRUKCE POMOCÍ METODY JARO-WINKLER

Robert Šanda[1)]

THE RATE OF RESPONSE FOR THE TOPIC OF COMMUTING IN THE 2011 POPULATION AND HOUSING CENSUS AND THE RECONSTRUCTION OF THE RESPONSE RATE USING THE JARO-WINKLER METHOD

Abstract
The article analyses the item response rate for data on commuting from the 2011 census and seeks to discover the main reasons for the unusually large shares of unknown values. A research method using the Jaro-Winkler algorithm of probabilistic record linkage is then applied to the raw records of census forms, aiming to improve the resulting response rates by identifying answers harmed by partial inconsistencies or mistakes. As a result, the share of recognized values increased significantly. Based on the findings the article then proposes basic conceptual recommendations for the next census.

Keywords: Population and housing census, Czechia, data quality, commuting, record linkage

Demografie, 2020, **62: 27–42**

ÚVOD

Odborné veřejnosti je všeobecně známo, že kvalita, resp. úplnost dat získaných prostřednictvím dotazníků při sčítání lidu se dlouhodobě snižuje spolu s rostoucí neochotou veřejnosti sdělovat své osobní údaje. Přesto byl relativně nízký podíl zjištěných údajů při sčítání v roce 2011 u některých ukazatelů překvapující. Platí to zejména o charakteristikách dojížďky do zaměstnání a škol. Stěžejním údajem o dojížďce je obec pracoviště, resp. školy. Více než u třetiny zaměstnaných či studujících obyvatel se tuto informaci nepodařilo při sčítání 2011 zjistit. Údaje o dojížďce mají přitom zásadní význam, nejen pro geografický výzkum, ale pro řadu dalších vědních disciplín, pro veřejnou správu i komerční sféru. Sčítání lidu je v současné době jejich nezastupitelným zdrojem a přinejmenším v nejbližších letech nelze očekávat, že bude k dispozici jiný, srovnatelně podrobný a komplexní zdroj. V době příprav nadcházejícího populačního cenzu, který se uskuteční v roce 2021, je proto nutné pokusit se porozumět příčinám nízké vyšetřenosti a pokud možno identifikovat příležitosti ke zlepšení kvality dat pro příští sčítání.

Termín *vyšetřenost*, resp. přesněji *částečná* či *položková vyšetřenost*, je obdobou anglického termínu *item* (nebo také *partial*) *response rate*, definovaného jako podíl jednotek, které poskytly údaje o dané proměnné, na celkovém počtu jednotek

1) Český statistický úřad, Praha; kontakt: robert.sanda@czso.cz.

(*Eurostat*, 2015).[2] Český ekvivalent vyšetřenost je zaveden především v kontextu problematiky šetření v domácnostech (např. *ČSÚ*, 2019).

Předkládaný článek se nejprve zabývá základní analýzou vyšetřenosti dojížďky (resp. její nejdůležitější charakteristiky – obce pracoviště/školy) podle oficiálně publikovaných výsledků sčítání lidu 2011. Cílem je mimo jiné poskytnout uživatelům dat informace, které soubor standardních publikací ze sčítání nabízí jen (v omezené míře. V řadě publikovaných tabulek totiž chybí údaj o počtu nezjištěných hodnot.

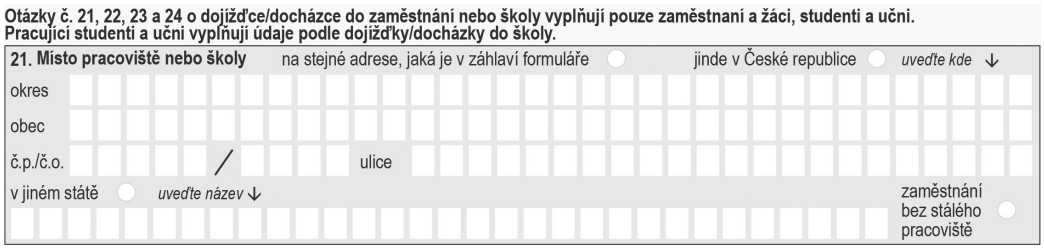
Další část je věnována pokusu o částečné doplnění údajů, které podle oficiálních výsledků sčítání lidu 2011 nebyly zjištěny. K tomu jsou využity anonymizované vstupní záznamy ze sčítacích formulářů, které jsou uchovány ve zpracovatelské databázi sčítání 2011. Na tato vstupní data jsou v článku aplikovány postupy, které při standardním zpracování výsledků sčítání 2011 nebyly realizovány. Jejich využitím se podařilo údaje o obci dojížďky částečně „rekonstruovat" a zvýšit tak celkovou vyšetřenost tohoto ukazatele. Hlavním cílem je vyhodnotit dopady aplikace prezentovaného postupu na výslednou vyšetřenost (vyšetřenost je v textu vyjádřena procentuálním podílem počtu zjištěných údajů na celkovém počtu pracujících/studujících) a pokusit se tak identifikovat některé rezervy, které je potenciálně možné při příštím cenzu využít.

Výsledky této práce mohou také pomoci alespoň v hrubších rysech posoudit, zda oficiálně publikovaná data dostatečně dobře reprezentují realitu (byť samozřejmě nikoliv v absolutních číslech), či zda po částečném doplnění údajů dochází ke změnám charakteristik, které z dat o dojížďce vycházejí (jako např. dominantní proudy, resp. celkový charakter prostorových vztahů, relativní rozmístění obsazených pracovních míst v jednotlivých odvětvích apod.). Tato témata však již přesahují rámec článku a budou předmětem navazujících prací.

ZPŮSOB ZJIŠŤOVÁNÍ A ZPRACOVÁNÍ MÍSTA PRACOVIŠTĚ NEBO ŠKOLY

Před hodnocením výsledků sčítání lidu 2011 (*SLDB*, 2011) je třeba připomenout způsob, jakým bylo místo pracoviště/školy zjišťováno, protože ten měl na úplnost výsledných údajů vliv. Otázku na místo pracoviště nebo školy měli zodpovědět všichni respondenti, kteří byli k rozhodnému okamžiku sčítání zaměstnaní či byli žáky/studenty škol. Otázkou byly získávány dva

Obr. 1: Otázka na místo pracoviště nebo školy na sčítacím listu osoby (SLDB 2011)
The question on place of work or school – 2011 census questionnaire



Pozn.: Adresou v záhlaví formuláře je míněna adresa místa sečtení, tj. adresa, na níž komisař předal respondentovi sčítací formulář (s výjimkou případů, kdy si respondenti vyzvedli formulář na pracovišti ČSÚ nebo České pošty).

Note: *The "Address in the form header" is the address of place of enumeration, i.e. the address at which the enumerator handed over the form to the respondent.*

2) Přesnější definice, lépe vystihující problematiku rozebíranou v článku, je definice doplňku k *item-response*, tedy pojmu *item (resp. partial) non-response* podle *Eurostat's Concepts* and *Definitions Database*. Podle této definice *item non-response* představují případy, kdy respondent poskytne některé, ale ne všechny požadované informace, nebo jsou některé uvedené informace nevyužitelné. Tato definice tedy zahrnuje i případy, kdy respondent sice na otázku reagoval, ale z nějakého důvodu je nebylo možné odpověď zpracovat (tj. např. nebyla relevantní nebo vykazovala takové nedostatky, že z ní nebylo možné získat využitelnou informaci).

související vstupní údaje – lokalizace místa pracoviště/školy vzhledem k místu sečtení (tj. na adrese místa sečtení, *jinde v České republice*, resp. *v jiném státě*) a konkrétní místo (viz obr. 1). Pokud se adresa místa pracoviště (školy) shodovala s adresou místa sečtení, respondent měl zaškrtnout příslušné pole. Pracoval-li či studoval na jiné adrese na území České republiky, měl označit pole *jinde v České republice* a následně uvést konkrétní adresu. Vyjíždějící do zahraničí měli zaškrtnout pole *v jiném státě* a vyplnit stát. Novinkou byla možnost zaškrtnout *zaměstnání bez stálého pracoviště*.[3)]

Řada respondentů nerespektovala strukturu otázky. Respondenti často uvedli pouze adresu pracoviště/školy, ale nikoliv lokalizaci vůči místu sečtení. Podle reportů ze zpracování lokalizace chyběla u 1 871 529 pracujících/studujících, což představuje téměř třetinu případů. U dalších 29 tisíc záznamů bylo označeno více možností. Tyto nedostatky bylo během zpracování nutné korigovat.

Prvním krokem zpracování byla digitalizace listinných formulářů. Při ní se údaje z naskenovaných formulářů nejprve automatizovaně rozpoznávaly. Nerozpoznané znaky a slova byly předány k řešení pracovníkům – validátorům. V dalším kroku se údaje kódovaly (propojovaly s číselníky), opět za podpory pracovníků, kteří kódovali položky nezakódované automaticky. V případech pracoviště/školy *jinde v ČR* bylo úkolem kódování rozpoznat adresu alespoň do určité úrovně, např. do úrovně okresu. Adresy, které se při kódování nepodařilo rozpoznat, nebylo již v dalších krocích zpracování možné využít.

V dalších fázích zpracování se údaje interpretovaly a posuzovaly ve vzájemných vazbách. Pro dojížďku byla přímo nadřazenou charakteristikou ekonomická aktivita. Pokud daná osoba nebyla po kontrolách vyhodnocena jako zaměstnaná ani studující, případné odpovědi na otázky dojížďky byly zrušeny (resp. označeny kódem nedefinováno). Kombinace údaje o lokalizaci pracoviště vzhledem k místu sečtení a případná konkrétní adresa či stát byly uvedeny do souladu, např. při vyplněné adrese, ale chybějícím označením lokalizace, bylo v databázi doplněno „zaškrtnutí" pole jinde v České republice apod. Tím se vyplněné otázky převedly do standardizované podoby a další kroky zpracování již spoléhaly na to, že otázky jsou vyplněny přesně v podobě, v jaké byly na formuláři vyžadovány. Z údajů o místě pracoviště/školy v kombinaci s místem obvyklého pobytu byl odvozen údaj o lokalizaci pracoviště/školy vzhledem k adrese obvyklého pobytu a následně byly vytvořeny dojížďkové proudy.[4)]

VYŠETŘENOST DOJÍŽĎKY VE VÝSLEDCÍCH SLDB 2011

Tab. 1 prezentuje počty pracujících či studujících podle lokalizace místa pracoviště vzhledem k místu obvyklého pobytu. U více než třetiny osob (36,7 % – viz součet podbarvených řádků v tabulce 1) se však místo pracoviště nebo školy nepodařilo určit s přesností do úrovně obce. U výrazné většiny z těchto případů (34,3 % z celkového počtu) se nepodařilo o místě pracoviště nebo školy zjistit žádné informace.

Uvedené údaje se týkají pouze obyvatel, u nichž se zjistila ekonomická aktivita (zaměstnaní, resp. žáci, studenti, učni). Dalších zhruba půl milionu obyvatel ve věku 15 a více let (571 064 obyvatel) mělo ekonomickou aktivitu nezjištěnou. Z nich 293 257 nebylo sečeno prostřednictvím formulářů, ale byli doplněni na základě údajů v evidenci obyvatel. Vzhledem k jejich věkové struktuře lze téměř s jistotou předpokládat, že velmi podstatná část z nich byli pracující nebo studující (za předpokladu stejných věkově specifických podílů by to bylo 436 tisíc zaměstnaných či studujících).

Vyšetřenost údajů o dojížďce v roce 2011 byla nízká u celé populace. Tento problém byl konstatován v řadě prací (např. *ČSÚ*, 2014; *Hampl a Marada*, 2015; *Ouředníček a kol.*, 2017; nebo *Bernard a Šimon*, 2017). Při práci s daty jsou však téměř vždy (určitou výjimkou je posledně citovaná práce) používány pouze zjištěné hodnoty a počty nezjištěných se nezohledňují, případně se pracuje s přepočty na relativní údaje. Vychází se tak de facto z předpokladu rovnoměrného rozložení vyšetřenosti dojížďky v území a v populaci.

3) Pro další informace o zjišťování dojížďky viz např. *ČSÚ* (2013).

4) Dojížďka z místa obvyklého pobytu byla při sčítání 2011 sledována poprvé, výsledky předchozích cenzů byly založeny na konceptu trvalého pobytu.

Tab. 1: Lokalizace místa pracoviště nebo školy vzhledem k místu obvyklého pobytu podle výsledků SDLB 2011
Location of the place of work or school in relation to the place of usual residence – 2011 census results

Lokalizace místa pracoviště nebo školy *Location of the place of work or school*	Absolutně / *Absolute numbers*			%		
	Pracující *Working*	Studující *Studying*	Celkem *Total*	Pracující *Working*	Studující *Studying*	Celkem *Total*
Na adrese místa obvyklého pobytu *At the address of usual residence*	580 509	170 411	750 920	12,9	11,2	12,5
Na jiné adrese v rámci obce obvyklého pobytu / *Elsewhere within municipality*	924 948	354 218	1 279 166	20,5	23,2	21,2
V jiné obci v rámci okresu obvyklého pobytu / *In another municipality within the LAU1*	596 686	184 834	781 520	13,3	12,1	13,0
V jiném okrese – obec zjištěna *In another LAU1 – municipality known*	488 688	231 529	720 217	10,9	15,2	12,0
V jiném okrese – obec nezjištěna *In another LAU1 – municipality unknown*	14 554	5 320	19 874	0,3	0,3	0,3
V jiném státě – stát zjištěn *Abroad – country known*	35 461	7 746	43 207	0,8	0,5	0,7
V jiném státě – stát nezjištěn *Abroad – country unknown*	1 787	305	2 092	0,0	0,0	0,0
V ČR – blíže nezjištěno *At some unknown place within Czechia*	98 905	29 493	128 398	2,2	1,9	2,1
Zaměstnání bez stálého místa *No fixed place*	232 986	2 076	235 062	5,2	0,1	3,9
Místo pracoviště, školy zcela nezjištěno *Unknown place*	1 526 938	539 458	2 066 396	33,9	35,4	34,3
Celkem / *Total*	**4 501 462**	**1 525 390**	**6 026 852**	**100,0**	**100,0**	**100,0**

Zdroj: Zpracovatelská databáze SLDB 2011, interní pracovní dokument ČSÚ.
Source: *2011 Population and Housing Census processing database, CZSO internal working document.*

Podrobnější výsledky SDLB 2011 však ukazují, že tento předpoklad je problematický.

Z obr. 2, znázorňujícího vyšetřenost obce dojížďky na úrovni správních obvodů obcí s rozšířenou působností (ORP), je vcelku zřetelný gradient ve směru ze severu na jih, resp. severozápadu na jihovýchod. Souvislý pás krajů podél severní hranice od Karlovarského po Královéhradecký představoval území s nejnižší vyšetřeností. Relativně nejúplnější výsledky byly získány ve Zlínském kraji (a na Vysočině, kde vyšetřenost mírně přesahovala dvě třetiny (67,6 %, resp. 67,4 %), nejnižší (57,3 %) byla v Karlovarském kraji. Na úrovni ORP se hodnoty pohybovaly mezi 51,9 % (SO ORP Lovosice) a 81,6 % (SO ORP Uherský Brod).

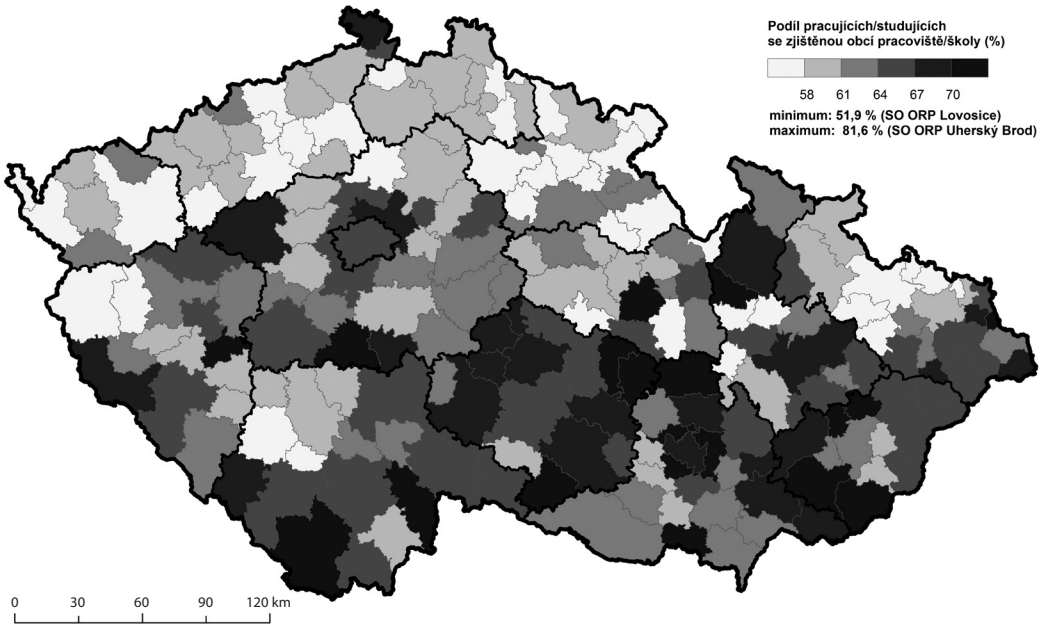
Uvedené prostorové schéma platilo spíše v hrubých rysech, vymykal se z něho např. SO ORP Blovice v Plzeňském Kraji, který měl po Uherskobrodsku druhou nejvyšší vyšetřenost, nebo SO ORP Uničov v Olomouckém kraji, kde byla vyšetřenost naopak druhá nejnižší. Za zmínku stojí také oblast Písecka a Strakonicka, kde byla vyšetřenost o poznání nižší než v okolí.

Na úrovni obcí byly rozdíly ve vyšetřenosti značné. Nejnižší byla v obci Vrátno v okrese Mladá Boleslav, kde se podařilo zjistit obec dojížďky pouze u 18 % pracujících či studujících. Celkem u 608 obcí byla vyšetřenost nižší než 50 %. Stoprocentní byla pouze ve dvou obcích, a to v Březí v jindřichohradeckém okrese a ve vojenském újezdu Březina v okrese Vyškov (zde byl však pouze jeden zaměstnaný, Březina měla v roce 2011 pouze tři obyvatele). Více než 90procentní vyšetřenost byla zjištěna u 48 obcí.

Územní diferenciace vyšetřenosti vykazovala podobné rysy s rozložením některých dalších charakteristik. Na úrovni ORP je zjevná například souvislost s rozložením podílu věřících v populaci, o něco slabší, ale patrná byla souvislost s podílem rodáků na obyvatelstvu či s účastí v komunálních i parlamentních volbách v roce 2010.[5)] Souvislost

5) Hodnoty Spearmanova koeficientu korelace byly 0,44 (náboženská víra), 0,30 (účast ve volbách do poslanecké sněmovny v roce 2010), resp. 0,27 (podíl rodáků).

Obr. 2: Vyšetřenost obce pracoviště/školy ve správních obvodech ORP
The response rates for place of work/school municipality by microregion



Note: *Ratio of the number of persons with known place of work / school municipality to the total number of working or studying persons.*
Zdroj: Zpracovatelská databáze SLDB 2011, vlastní výpočty.
Source: *2011 Population and Housing Census processing database, author's calculations.*

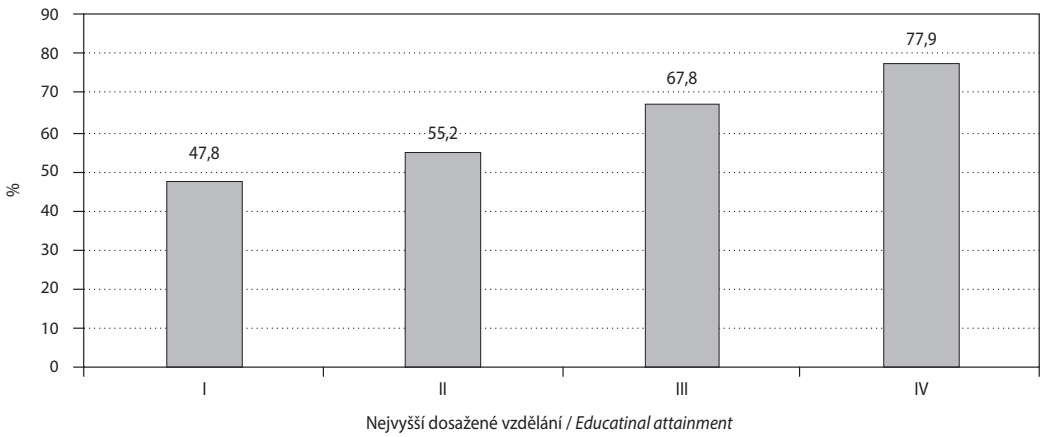
s uvedenými charakteristikami nebyla bezprostřední, např. rozdíly ve vyšetřenosti u věřících a u obyvatel bez víry byly celorepublikově i v jednotlivých regionech nevýrazné. Z uvedených podobností lze zřejmě usuzovat na souvislost mezi vztahem obyvatel k místu a společnosti (rodáci), určitým pocitem spoluzodpovědnosti za společenské dění a snad i relativně pozitivním vztahem k institucím (účast ve volbách, význam církve jako instituce v lokalitě) a ochotou věnovat úsilí korektnímu vyplnění formuláře.

Z charakteristik, s nimiž vyšetřenost dojížďky vykazovala nejen podobnost v územní diferenciaci, ale bezprostřední souvislost, výrazně dominovala úroveň vzdělanosti (ať již reprezentovaná podílem vysokoškoláků, indexem vzdělanosti, apod.).

Vzhledem k tomu, že u studujících osob nelze dosažené vzdělání a některé další důležité charakteristiky vhodně použít, je dále věnována pozornost pouze zaměstnaným a vyšetřenosti obce pracoviště (zaměstnaní představují rozhodující část dojíždějících). Uvedený výrazný vliv nevyššího dosaženého vzdělání na vyšetřenost obce pracoviště je zřejmý z grafu 1. U zaměstnaných se základním nebo neukončeným vzděláním se nepodařilo určit obec pracoviště ve více než polovině případů. S rostoucí úrovní dosaženého vzdělání úplnost výsledků rostla, u vysokoškoláků již byl podíl zaměstnaných s určenou obcí pracoviště ve srovnání se základním vzděláním o polovinu vyšší. I tak ovšem údaj chyběl u více než pětiny zaměstnaných. Specifickou kategorií byli zaměstnaní s nezjištěnou úrovní vzdělání, u nich byla vyšetřenost obce pracoviště velmi nízká, pouze 25 %.

Vzhledem k výrazné závislosti vyšetřenosti dojížďky na úrovni dosaženého vzdělání je třeba u dalších charakteristik, které často se vzdělaností souvisejí, nutné brát vzdělanostní strukturu vždy v úvahu. To je zřejmé i z tab. 2, znázorňující vliv velikosti obce bydliště v kombinaci s úrovní vzdělání na vyšetřenost obce pracoviště. Úroveň vzdělání se výrazně projevila u všech velikostních skupin obcí. Vliv velikosti obcí na vyšetřenost nebyl jednoznačný, její hodnoty s růstem velikosti

Graf 1: Vyšetřenost obce pracoviště podle úrovně dosaženého vzdělání zaměstnaných
The response rates for place of work municipality by education level



Pozn.: I – základní a neukončené; II – střední bez maturity; III – střední s maturitou; IV – vysokoškolské.
Note: *I – lower secondary or less; II – upper secondary (no GCSE – ISCED 353); III – upper secondary (GCSE – ISCED 344, ISCED 354, ISCED 4; IV – tertiary.*
Zdroj: Zpracovatelská databáze SLDB 2011, vlastní výpočty.
Source: *2011 Housing and Population Census processing database, author's calculations.*

Tab. 2: Vyšetřenost obce pracoviště podle velikostní kategorie obcí obvyklého pobytu a nejvyššího dosaženého vzdělání (v %) / The response rates for place of work municipality by the size of the municipality of usual residence and education level (%)

Velikost obce obvyklého pobytu / *Size of the municipality of usual residence*	Obec pracoviště *Place of work municipality*				Vyšetřenost podle dosaž. vzdělání *Response rate by educ. attainment*			
	Zjištěna *Known*	Nezjištěna *Unknown*	Celkem *Total*	Vyšetřenost *Resp. rate*	I	II	III	IV
499 a méně / *or less*	219 873	133 470	353 343	62,2	48,4	56,5	69,0	78,8
500–999	254 965	147 162	402 127	63,4	49,8	57,2	69,2	79,6
1 000–1 999	272 505	157 492	429 997	63,4	50,0	56,8	68,8	78,7
2 000–4 999	338 914	186 801	525 715	64,5	50,2	57,6	69,6	79,5
5 000–9 999	244 150	153 160	397 310	61,5	46,2	53,4	66,6	77,4
10 000–19 999	261 270	151 126	412 396	63,4	47,3	55,1	68,2	78,1
20 000–49 999	352 064	211 547	563 611	62,5	45,7	53,6	66,9	77,2
50 000–99 999	235 525	139 581	375 106	62,8	45,8	53,0	66,9	77,0
100 000 + (bez Prahy / *excl. Prague)*	298 491	160 792	459 283	65,0	46,0	53,0	67,7	78,8
Praha / *Prague*	383 308	199 266	582 574	65,8	47,5	52,5	66,5	76,5
Celkem / *Total*	**2 861 065**	**1 640 397**	**4 501 462**	**63,6**	**47,8**	**55,2**	**67,8**	**77,9**

Pozn.: I – základní a neukončené; II – střední bez maturity; III – střední s maturitou; IV – vysokoškolské.
Note: *I – lower secondary or less; II – upper secondary (no GCSE – ISCED 353); III – upper secondary (GCSE – ISCED 344, ISCED 354, ISCED 4; IV – tertiary.*
Zdroj: Zpracovatelská databáze SLDB 2011, vlastní výpočty.
Source: *2011 Housing and Population Census processing database, author's calculations.*

obcí kolísaly. To však do určité míry ovlivnila právě odlišná vzdělanostní struktura. Zatímco celkově byla nevyšší vyšetřenost v Praze a dalších městech nad 100 tisíc obyvatel, v rámci jednotlivých vzdělanostních kategorií byly hodnoty nejvyšší u menších obcí (do 5 tisíc).

Tab. 3: Vyšetřenost obce pracoviště podle věku
The response rates for place of work municipality by age

Obec pracoviště *Place of work municipality*	Věk / *Age*						
	15–24	25–34	35–49	50–64	65+	Nezjištěn *Unknown*	Celkem *Total*
Zjištěna / *Known*	173 447	760 661	1 178 117	698 722	48 826	1 292	2 861 065
Nezjištěna / *Unknown*	119 952	360 351	623 097	483 841	49 702	3 454	1 640 397
Celkem / *Total*	**293 399**	**1 121 012**	**1 801 214**	**1 182 563**	**98 528**	**4 746**	**4 501 462**
Vyšetřenost / *Resp. rate (%)*	59,1	67,9	65,4	59,1	49,6	27,2	63,6

Zdroj: Zpracovatelská databáze SLDB 2011, vlastní výpočty.
Source: *2011 Housing and Population Census processing database, author's calculations.*

Tab. 4: Vyšetřenost obce pracoviště podle pohlaví a nejvyššího dosaženého vzdělání
The response rates for place of work municipality by sex and education level

Obec pracoviště *Place of work municipality*	Celkem / *Total*		Z toho se zjištěnou úrovní dosaženého vzdělání *Of which with known level of educational attainment*							
			I		II		III		IV	
	muži *males*	ženy *females*	muži *males*	ženy *females*	muži *males*	ženy *females*	muži *males*	ženy *females*	muži *males*	ženy *females*
Zjištěna / *Known*	1 592 999	1 268 066	64 106	66 810	613 555	295 872	565 524	613 006	346 212	289 901
Nezjištěna / *Unknown*	861 428	778 969	61 581	81 584	446 486	293 016	249 440	309 384	94 069	86 642
Celkem / *Total*	**2 454 427**	**2 047 035**	**125 687**	**148 394**	**1 060 041**	**588 888**	**814 964**	**922 390**	**440 281**	**376 543**
Vyšetřenost / *Resp. rate (%)*	64,9	61,9	51,0	45,0	57,9	50,2	69,4	66,5	78,6	77,0

Pozn.: I – základní a neukončené; II – střední bez maturity; III – střední s maturitou; IV – vysokoškolské.
Note: *I – lower secondary or less; II – upper secondary (no GCSE – ISCED 353); III – upper secondary (GCSE – ISCED 344, ISCED 354, ISCED 4; IV – tertiary.*
Zdroj: Zpracovatelská databáze SLDB 2011, vlastní výpočty.
Source: *2011 Housing and Population Census processing database, author's calculations.*

Vyšetřenost byla diferencována i v závislosti na věku (tab. 3). Nejvyšší úrovně dosahovala u zaměstnaných ve skupině od 25 do 34 let, resp. v širší části věkového spektra zhruba od 25 do 50 let, s tendencí k poklesu vyšetřenosti s věkem. Krajní věkové kategorie zaměstnaných vykazovaly velmi nízkou vyšetřenost. Popsaná situace byla obdobná ve všech vzdělanostních kategoriích. Vliv mělo i pohlaví zaměstnaných – u mužů byla vyšetřenost místa pracoviště vyšší než u žen. Celkově nebyl rozdíl zásadní, ale v nižších vzdělanostních kategoriích se vliv pohlaví projevoval poměrně výrazně (tab. 4).

Důležitou charakteristikou související s místem pracoviště je odvětví ekonomické činnosti. I zde byla úplnost dat o dojížďce poměrně výrazně diferencovaná. V odvětví těžby a dobývání se obec pracoviště podařilo zjistit pouze u 60 % zaměstnaných. Podobně nízká byla vyšetřenost v sekci ubytování, stravování a pohostinství (61 %) a v sekci činností souvisejících s odpady a zásobováním vodou (63 %). Naopak nejúplnější informace se podařilo shromáždit u zaměstnaných v sekci informačních a komunikačních činností, která byla spolu s marginální sekcí extrateritoriálních organizací jedinou, kde byla vyšetřenost vyšší než 80 %. Zejména u některých odvětví je výsledek do značné míry opět dán vzdělanostní strukturou.

Na závěr uveďme ještě diferenciaci vyšetřenosti podle státního občanství. Z početných skupin měli nejvyšší vyšetřenost občané Slovenska, teprve poté následovali čeští občané. Výsledky u občanů Ukrajiny a zejména občanů Vietnamu byly výrazně horší. Relativně kvalitní data u Slováků byla opět zejména důsledkem jejich vzdělanostní struktury, neboť zaměstnaní Slováci se vyznačovali velmi vysokým podílem vysokoškoláků, navíc častěji zaměstnaných v oddělení informačních a komunikačních technologií a v dalších odvětvích vyznačujících se relativně vysokou vyšetřeností údajů o dojížďce.

HLAVNÍ PŘÍČINY NÍZKÉ VYŠETŘENOSTI ÚDAJŮ O DOJÍŽĎCE

Smyslem předchozího textu bylo jednoduchým přehledem poukázat na skutečnost, že úplnost údajů o místě dojížďky byla – mnohdy výrazně – diferencována v závislosti na socioekonomických charakteristikách obyvatel, na sídelní struktuře, vykazuje i vcelku zřetelnou diferenciaci prostorovou. Na celém území a napříč všemi kategoriemi však byla poměrně nízká. S přibližujícím se termínem příštího sčítání lidu (březen 2021) nabývá na aktuálnosti otázka, co bylo příčinou nízké vyšetřenosti a jak zajistit, aby byly výsledky příštího sčítání úplnější. Do určité míry lze odpověď nalézt ve zpracovatelské databázi sčítání lidu 2011. Databáze původně obsahovala celou historii procesu zpracování. Technické důvody vedly ke ztrátě informací o některých fázích zpracování, zůstaly však uchovány anonymizované vstupní údaje ze sčítacích formulářů (v podobě po digitalizaci, před kódováním).

Na základě těchto vstupních dat lze případy nezjištěného místa pracoviště nebo školy rozdělit na tři základní skupiny. První představovaly zcela chybějící odpovědi. Tyto případy lze snad do určité míry interpretovat jako důsledek všeobecně rostoucí neochoty části veřejnosti poskytovat o sobě informace úřadům. Celkově však chybějící odpovědi tvořily pouze menší část případů nezjištěné obce pracoviště/školy (konkrétně 486 tisíc případů, tj. 22 % nezjištěných).

U výrazné většiny nezjištěných odpovědí lze ve vstupních datech vypozorovat snahu na otázky k dojížďce odpovědět, odpovědi se však při kódování nepodařilo zpracovat. Část odpovědí nerespektovala logiku konstrukce jednotlivých otázek, například bylo označeno pole *„místo pracoviště/školy na adrese místa sečtení“*, zároveň však byla vyplněna adresa pracoviště jako podotázka při odpovědi *„jinde v ČR“*. Elektronické formuláře tyto chyby eliminovaly již při vyplňování, u listinných formulářů byly chyby ošetřovány logickými kontrolami při zpracování. Ne všechny tyto případy se však podařilo vyřešit.

Poslední skupinu chyb (překrývající se částečně s předchozí) tvořily případy, kdy respondenti vyplnili jednotlivé kolonky částečně či zcela chybně (např. zaměnili název obce s názvem městské části, s ulicí apod.), a to natolik, že se při zpracování výsledků nepodařilo odpovědi „rozklíčovat“. Jednalo se často o případy, z nichž by při individuálním posouzení či aplikací vhodných postupů (a dostatku času) bylo možné alespoň určité informace o dojížďce vytěžit.[6)]

Z uvedeného plyne, že s výjimkou první z výše zmíněných skupin problémů, tj. zcela chybějících odpovědí, obsahuje zpracovatelská databáze SLDB ve vstupních datech určité, byť omezeně využitelné informace. Pro účely tohoto článku byly tyto údaje v rámci možností autora zpracovány s cílem alespoň částečně zrekonstruovat údaje o obci pracoviště/školy, poskytnout možnost vytvořit úplnější obraz dojížďky a především získat podklad pro vytvoření potenciálně efektivnějších algoritmů pro zpracování příštího sčítání v roce 2021.

REKONSTRUKCE ÚDAJŮ OBCI PRACOVIŠTĚ/ŠKOLY ZE ZPRACOVATELSKÉ DATABÁZE SLDB

Při rekonstrukci údajů o obci pracoviště/školy ze vstupních dat zpracovatelské databáze SLDB nebyly provedeny všechny kroky procesu zpracování nad celým souborem všech vstupních záznamů (formulářů), neboť by to nebylo ani účelné ani technicky proveditelné. Údaje, které se do oficiálních výsledků podařilo zpracovat, byly pro tento článek pouze převzaty, nebyly nijak revidovány. Revidována nebyla ani ekonomická aktivita. Rekonstrukce se tak zaměřila jen na záznamy osob, které podle oficiálních výsledků patřily mezi zaměstnané/studující, ale u nichž nebyla obec pracoviště/školy zjištěna. Jednalo se celkem o 2 214 668 záznamů (viz tab. 2).

6) Velmi pozitivní vliv na všechny tyto problémy měly elektronické sčítací formuláře. Zatímco prostřednictvím listinných formulářů byla celková dojížďka do úrovně obce zjištěna pouze u 46,5 % pracujících/studujících obyvatel, v případě elektronických formulářů u 97,6 %. Elektronické formuláře totiž obsahovaly kontroly úplnosti a základní logiky vyplnění, název okresu respondenti vybírali ze seznamu. Rozpoznání odpovědi na elektronickém formuláři navíc nebylo závislé na čitelnosti zápisu, kvalitě naskenování, systému automatického rozpoznávání naskenovaných znaků ani kvalitě práce validátorů.

Příprava dat

Rekonstrukce byla založena na aplikaci metody pravděpodobnostního propojování dat, jejímž prostřednictvím byla měřena podobnost údajů uvedených na sčítacích formulářích (tj. respondenty zapsaných textů v podobě po jejich naskenování a digitalizaci) s oficiálními názvy v číselnících. Údaje z jednotlivých kolonek formulářů byly posuzovány ve dvojicích *okres-obec* a *obec-ulice*. Úkolem bylo nalézt ke každé dvojici z formuláře nejpodobnější referenční dvojici názvů z číselníků.

Bylo však nutné zohlednit fakt, že respondenti v odpovědích nezřídka zaměňovali části obcí za obce. Referenční seznam názvů pro dvojice *okres-obec* byl proto vytvořen spojením dvojic *okres-obec* s dvojicemi *okres-část obce*. Byly ovšem použity pouze ty dvojice o*kres-část obce*, které představovaly unikátní kombinace názvů. Referenční seznam pro dvojice *obec-ulice* rovněž vytvořen sloučením oficiálních názvů *obec-ulice* a *část obce-ulice* a byly z něho odstraněny všechny dvojice, které neumožňovaly celorepublikově jednoznačně identifikovat obec (nebylo totiž cílem ve výsledku určit místo dojížďky do úrovně ulice, ale pouze využít údaj o ulici k identifikaci obce). Názvy částí obcí byly v obou referenčních seznamech místy upraveny, protože některé názvy jsou komplikované a bylo nepravděpodobné, že by je respondenti zapsali.

Před vlastním měřením podobnosti bylo třeba provést transformaci porovnávaných textů do standardizované podoby. Všechny znaky v porovnávaných údajích z formulářů byly proto převedeny na velká písmena. Ze všech údajů byla odstraněna diakritika, mezery, interpunkční znaménka, číslice (s výjimkou ulic, u těch byly číslice ponechány) a všechny jiné znaky než písmena abecedy. Analogicky byly upraveny názvy v referenčních seznamech. Příklady provedené transformace textů z formulářů jsou v levé části tab. 5.

Propojení záznamů

Po popsaných úpravách byla vypočtena podobnost každé dvojice z polí *okres-obec* ze sčítacích formulářů s každou dvojicí *okres-obec (část obce)* z referenčního seznamu a analogicky podobnost dvojic *obec-ulice*. K měření podobnosti (často se užívá obrácený pojem *vzdálenost*) záznamů existuje v současné době relativně větší množství metod, od metod zaměřených na jednotlivé proměnné v rámci záznamů (znaky, slova) po metody vyhodnocující celé záznamy.[7] Pro rekonstrukci obce pracoviště byla využita metoda Jaro-Winkler, která porovnává textové řetězce na základě počtu shodných znaků, přičemž bere v úvahu délku řetězců a pozice jednotlivých znaků. Důraz klade především na podobnost na začátku řetězců. Metoda Jaro-Winkler je modifikací Jarovy podobnosti, vyvinuté pro porovnávání vlastních jmen při zpracování sčítání lidu 1985 na Floridě (*Jaro*, 1989). Výpočet Jarovy podobnosti (Φ_j) je následující (symboly byly s drobnými úpravami převzaty z *Winkler*, 1990):

$$\Phi_j = W_1 \cdot c/d + W_2 \cdot c/r + W_t \cdot (c - t/2) / c.$$

Symbol c představuje počet shodných znaků v obou porovnávaných řetězcích, d a r jsou délky prvního, resp. druhého řetězce. W_1 je váha určená prvnímu porovnávanému řetězci, W_2 váha druhého řetězce, W_t váha přisouzená transpozicím (standardně se hodnota všech vah stanovuje na 1/3). Transpozice – jejich počet je ve vzorci označen t – jsou znaky, které se nacházejí v obou řetězcích, ale v odlišné posloupnosti, tj. jsou „zpřeházené". Aby byly znaky považovány za shodné, musí být rozdíl v jejich pozicích v řetězci menší, než je polovina délky delšího z řetězců, tj. maximální rozdíl mezi pozicemi je roven $max(d;r)/2 - 1$.

Winkler (1990) provedl modifikaci výpočtu, spočívající v zavedení parametru zvyšujícího váhu shodných začátků řetězců.[8] Jarova-Winklerova podobnost (Φ_{jw}) se vypočítá podle vzorce:

$$\Phi_{jw} = \Phi_j + i \cdot p \cdot (1 - \Phi_j),$$

7) Problematiku automatického propojování záznamů a nejistotu pramenící z různých zápisů stejných hodnot otevřeli *Newcombe a kol.* (1959). Formulaci teorie pravděpodobnostního propojování, ze které současné algoritmy vycházejí, provedli *Fellegi a Sunter* (1969).

8) Tato úprava vychází ze zjištění, že zápisy jmen se častěji rozcházejí v středních a zadních částech než na začátku (viz i některé z příkladů v tab. 5, např. Rychnov n. Kn.).

kde *i* je počet shodných počátečních znaků v obou řetězcích (maximálně však 4), *p* je Winklerův parametr. Tento se obvykle nastavuje na hodnotu 0,1 – s touto hodnotou pracuje i *Winkler* (1990) a byla použita i při rekonstrukci dojížďky. Blíže o problematice viz také např. *Winkler* (1999; 2006).

Následně bylo nutné stanovit prahovou podobnost, kterou musí porovnávané řetězce (formulář vs. referenční seznam) splňovat. K tomu neexistuje obecně platný exaktní postup. Při rekonstrukci byly vyzkoušeny různé podmínky, z nichž jako nejvhodnější byla pro dvojice *okres-obec (část obce)* vybrána následující: *Kód obce z číselníků byl k příslušnému formuláři přiřazen (tj. obec dojížďky byla zrekonstruována), pokud dvojice obec (formulář)-obec/část obce (číselník) vykazovala podobnost alespoň 0,89 a zároveň dvojice okres (formulář)-okres (číselník) vykazovala podobnost rovněž alespoň 0,89, nebo pokud byla podobnost jedné z uvedených dvojic v intervalu 0,86–0,88 a zároveň druhá dvojice vykazovala podobnost minimálně 0,95* (viz poslední příklad v tab. 5). V případě, že tuto složenou podmínku splnilo více záznamů, hrozilo zvýšené riziko chybného přiřazení, proto v takové situaci nebyl vybrán ani jeden záznam. Příklady propojených údajů jsou uvedeny v tab. 5. Obdobně byly vypočteny podobnosti u dvojic *obec–ulice*, zde však byla vždy vyžadována minimální podobnost 0,89.

Na závěr byla v rámci možností autora provedena manuální rekonstrukce obce dojížďky. Pro ni byly vybrány často se opakující záznamy na formulářích, které se nepodařilo popsaným způsobem automaticky propojit. Manuálně byl doplněn i stát pracoviště, resp. nespecifikované místo pracoviště v zahraničí, resp. zaměstnání bez stálého pracoviště. V rámci možností byly manuálně zpracovány i automaticky nepropojené případy, kdy zápis na formuláři obsahoval unikátní názvy územních jednotek (např. název obce, který nelze zaměnit s jinou obcí ani jinou územní jednotkou) – od původního záměru využít tyto údaje k automatickému doplnění obce dojížďky bylo třeba upustit, protože výsledky byly problematické.[9)]

Tab. 5: Příklady údajů ze sčítacích formulářů a k nim přiřazených názvů z číselníků okresů a obcí
Examples of raw values from the census forms and linked official names of districts and municipalities

Údaje ze sčítacího formuláře / *Values from the census form*				**Nejpodobnější dvojice z číselníků (po transformaci)** *Most similar pair of names (transformed)*		**Podobnost** *Similarity*	
Původní (rozpoznané při digitalizaci) *Original (as recognised in the digitisation process)*		**Po transformaci** *Transformed*					
Okres *District (LAU1)*	Obec *Municipality*	Okres *District (LAU1)*	Obec *Municipality*	Okres *District (LAU1)*	Obec *Municipality*	Okres *District (LAU1)*	Obec *Municipality*
ÚSTÍ/ORLICÍ	ZÁMRSK 565 43 MŠ	USTIORLICI	ZAMRSKMS	USTINADORLICI	ZAMRSK	0,95	0,95
RYCHNOV N./KN.	DOUDLEBÝ N./ORL.	RYCHNOVNKN	DOUDLEBYNORL	RYCHNOVNADKNEZNOU	DOUDLEBYNADORLICI	0,92	0,94
BRUNTÁL	UNÁLNO	BRUNTAL	UNALNO	BRUNTAL	UVALNO	1,00	0,90
CHEB WECH	CHEB CHEB	CHEBWECH	CHEBCHEB	CHEB	CHEB	0,90	0,90
BRNO	BRNO-STŘED	BRNO	BRNOSTRED	BRNOMESTO	BRNO	0,89	0,89
KARLOVARSKÝ	VEJDEK	KARLOVARSKY	VEJDEK	KARLOVYVARY	NEJDEK	0,89	0,89
ÚSTÍ NAD LABEM	ÚSTÍ-//--//	USTINADLABEM	USTI	USTINADLABEM	USTINADLABEM	1,00	0,87

Zdroj: Zpracovatelská databáze SLDB 2011, vlastní úpravy a výpočty.
Source: *2011 Population and Housing Census processing database, author's editing and calculations.*

9) Příkladem může být několikeré uvedení názvu „*Čáslavsko*" do kolonky obec na formuláři. Čáslavsko je celorepublikově unikátní název obce v okrese Pelhřimov. Podrobnější pohled na dané záznamy však ukázal, že ani v jednom případě se o tuto obec nejednalo. V jednom případě respondent zjevně mínil okolí města Čáslav, v ostatních se s jistotou jednalo o Čáslavky, část obce Dolany v okrese Náchod. Podobných případů bylo odhaleno více.

Identifikace chybných a nevěrohodných propojení

Vyhodnocení výsledků předchozího postupu na vzorku záznamů a stejně tak ověřování metody na souboru oficiálně zjištěné dojížďky prokázaly vysokou míru spolehlivosti – počty osob, u nichž byla odhalena chybně doplněná obec dojížďky, se pohybovaly v řádu promile. Vzhledem k charakteru dat byly však i početně nevýznamné chyby někdy výrazné. Například respondent na formuláři uvedl jako místo pracoviště okres *Jesenice* (neexistující) – obec *Jesenice*, což s vysokou mírou podobnosti odpovídalo referenční dvojici okres-obec *Jeseník-Jeseník* (Φ_{jw} = 0,92). Tím byl vytvořen nepravděpodobný proud (byť reprezentovaný pouze jednou osobou) směřující ze středočeského kraje na Jesenicko. Tyto problémy by však nevyřešilo zvýšení prahové podobnosti pro propojení záznamů, resp. práh by musel být posunut natolik, že by to vedlo ke znehodnocení dosažených výsledků.

V dalším kroku byl proto vytvořen seznam „přípustných" proudů dojížďky. Ten se skládal z proudů v oficiálních výsledcích, proudů vzniklých v předchozím kroku na základě stoprocentní podobnosti (plné shody) dvojic *okres-obec*, dále z proudů vzniklých na základě nižší podobnosti, ale v případech, kdy byla obec dojížďky shodně identifikována pomocí dvojic *okres-obec* i *obec-ulice*, a konečně proudů vzniklých na základě manuálního doplnění obce dojížďky. Zbylé případy (např. pouze na základě podobnosti dvojic *obec ulice*) nesměly založit nový proud. Pokud zakládaly třeba i v jediném případě, byla zrekonstruovaná obec dojížďky zrušena i v ostatních případech, v nichž se daná kombinace údajů vyskytovala. Tímto opatřením se snížila celková chybovost a především se významně snížilo riziko vzniku chybných proudů.

V průběhu celého postupu byl také vytvářen seznam opakujících se problémů, které představovaly podstatnou část všech zjištěných nesrovnalostí.[10] Tyto typické problémy byly následně vyhledány a manuálně ošetřeny.

Posledním kontrolním krokem byl výpočet počtu obsazených pracovních míst ze zrekonstruovaných dat a jeho porovnání s oficiálními výsledky. Výrazné nárůsty počtu pracovních míst (resp. nárůsty jejich podílu na ČR), byly individuálně prověřeny.

Vyhodnocení zrekonstruovaných údajů

Výše popsaný postup umožnil chybějící údaj doplnit u 1 587 943 osob a celkovou vyšetřenost tak zvýšit z 63,3 % na 89,6 %.[11] Nezjištěná informace zůstala u 626 725 obyvatel, z nichž 478 025 (7,9 % z celkového počtu zaměstnaných/studujících) otázku zcela vynechalo. Ostatní pracující/studující respondenti odpověď uvedli, nepodařilo se ji však rozpoznat. U velké většiny těchto případů byla patrná snaha otázku zodpovědět seriózně.[12]

Z uvedených zjištění je zřejmé, že nízká vyšetřenost v publikovaných datech je převážně důsledkem rezerv v postupu zpracování výsledků v kombinaci s obtížemi, které měli respondenti se zodpovězením otázky. Neochota či laxnost na straně respondentů měla na kvalitě výsledků minoritní podíl a snižovala vyšetřenost pouze v řádu jednotek procentních bodů.

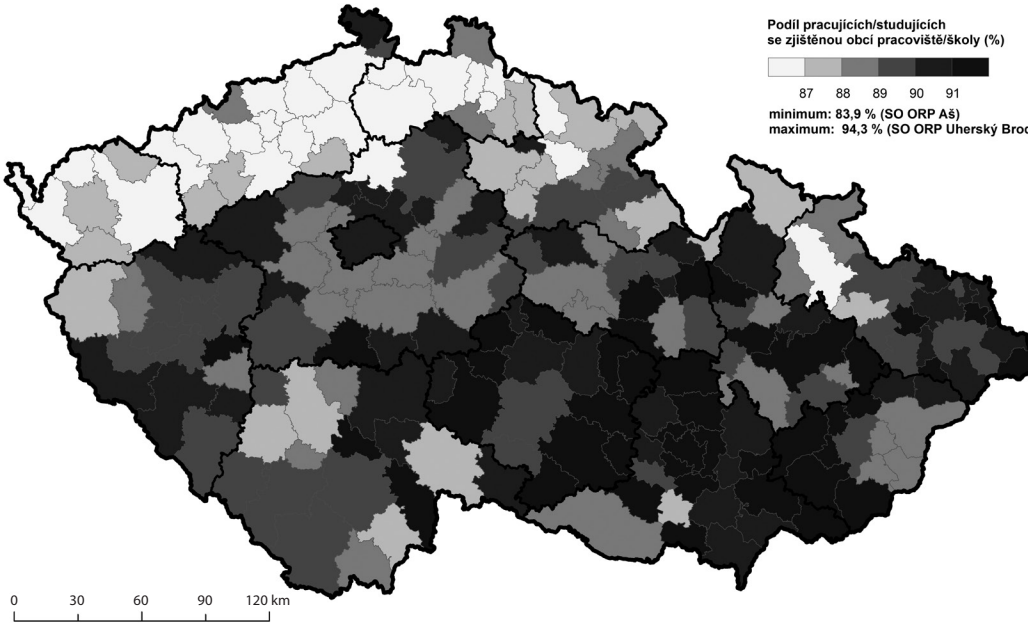
Regiony s nejúplnějšími výsledky se po rekonstrukci staly Jihomoravská kraj a Vysočina (shodně 91,3 %), teprve po nich následoval s 90,8 % Zlínský kraj (v něm byla vyšetřenost podle oficiálních

10) Často se například se na formulářích vyskytovala kombinace okres *Turnov* – obec *Turnov*. Turnov je v okrese Semily, metodou Jaro-Winkler byla však z referenčního seznamu vybrána velmi podobná dvojice *Trutnov-Trutnov*. Všeobecně problematické byly údaje týkající se Prahy a okolí (odpovědi Praha-Čakovičky byly měřením podobnosti identifikovány jako Praha-Čakovice, jednalo se ale o Mělník-Čakovičky, apod.). Při testování byly odhaleny i opakující se chyby v oficiálních datech, například Hněvkovice – část Týna nad Vltavou byly zaměňovány s obcí Hněvkovice v okrese Havlíčkův Brod.

11) U studujících obyvatel bylo toto doplnění mírně úspěšnější než u pracujících, ve výsledku tak bylo u žáků/studentů dosaženo úplnějších dat (90,0 % oproti 89,5 % u pracujících). V oficiálních datech byla přitom vyšetřenost mírně vyšší u pracujících, viz tab. 2.

12) Většinu z těchto případů představovaly odpovědi obsahující např. nestrukturované informace o zaměstnavateli, popis cesty (např. „*dálnice D1*") apod., tj. odpovědi prakticky nezpracovatelné, ale respondentem uvedené zřejmě ve snaze otázku správně zodpovědět. Pouze zanedbatelné procento představovaly vágní odpovědi typu „*to je různé*" (dva případy), resp. zjevné bojkoty (např. ve dvou případech reakce „*Co je vám do toho?*").

Obr. 3: Vyšetřenost obce pracoviště/školy ve správních obvodech ORP (zrekonstruovaná data SLDB 2011)
The response rates for place of work/school municipality by microregions (reconstructed 2011 census data)



Note: *Ratio of the number of persons with known place of work / school municipality to the total number of working or studying persons.*
Zdroj: Zpracovatelská databáze SLDB 2011, vlastní výpočty.
Source: *2011 Population and Housing Census processing database, author's calculations.*

výsledků nejvyšší – 67,6 %). Stejně jako u oficiálních dat byla nejnižší vyšetřenost v Karlovarském kraji (86,0 %). Rekonstrukce měla nejvýraznější dopad na úplnost výsledků v Moravskoslezském kraji. V něm byla v případě oficiálních výsledků vyšetřenost podprůměrná, po provedení rekonstrukce však stoupla na 90,4 %, což byla na krajské úrovni čtvrtá nejvyšší hodnota.

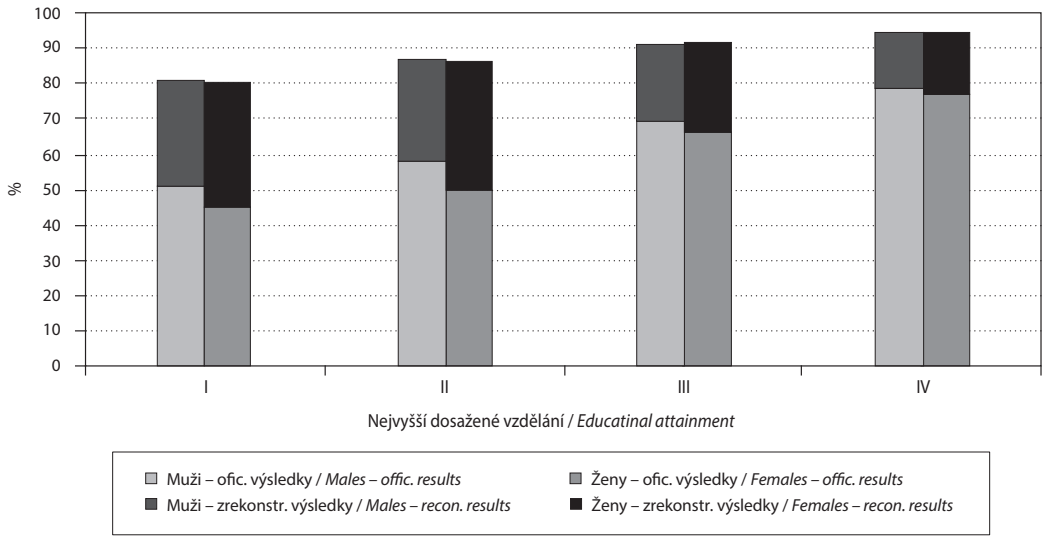
Obr. 3 znázorňuje vyšetřenost na úrovni SO ORP. Hodnoty se ve srovnání s oficiálními výsledky vyznačovaly výrazně nižší variabilitou, zato však územní rozdíly vykazovaly jasnější uspořádání. Ve srovnání s celorepublikovým průměrem zřetelněji vynikla relativně nižší vyšetřenost na severozápadě a severu území (dosídlené území Sudet) a patrnější byl zejména gradient ve směru k jihovýchodu území republiky. Ve střední, jižní a jihozápadní části Čech se vyšetřenost většinou pohybovala zhruba kolem průměru. Nevyšší vyšetřenost byla stejně jako u oficiálních výsledků v SO ORP Uherský Brod, následoval SO ORP Boskovice. Nejnižší byla v SO ORP Aš, Děčín a Cheb. Obcí s nejnižší vyšetřeností byly Třebčice v okrese Plzeň-jih (jediná obec, u níž ani po rekonstrukci nebyla zjištěna obec dojížďky u více než poloviny osob), celkem v 41 obcích se podařilo obec dojížďky zjistit ve všech případech.

V měřítku SO ORP ve srovnání s oficiálními výsledky výrazněji vynikla souvislost s diferenciací výše zmiňovaných charakteristik religiozity, volební účasti, resp. stability obyvatelstva.[13] To odpovídá zmíněnému předpokladu, že tyto charakteristiky korespondují s mírou ochoty vyplnit sčítací formuláře – po částečné nápravě (neúmyslných) chyb se výrazněji projevil vliv tohoto faktoru.

Podstatný přímý vliv na vyšetřenost pracovní dojížďky mělo i u zrekonstruovaných údajů nejvyšší dosažené vzdělání. S rostoucí úrovní vzdělání úplnost výsledků výrazně rostla, u osob se základním

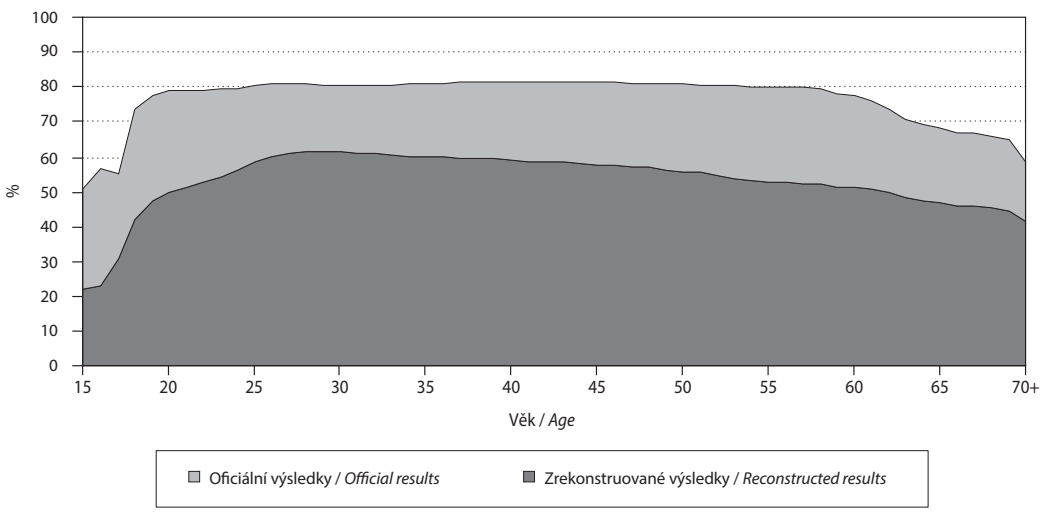
13) Hodnoty Spearmanova koeficientu korelace po rekonstrukci stouply z 0,44 na 0,60 (náboženská víra), z 0,30 na 0,39 (účast ve volbách do poslanecké sněmovny v roce 2010), resp. z 0,27 na 0,48 (podíl rodáků).

Graf 2: Vyšetřenost obce pracoviště podle úrovně dosaženého vzdělání a pohlaví (oficiální a zrekonstruovaná data SLDB 2011) / The response rates for place of work municipality by sex and educational attainment (official and reconstructed 2011 census data)



Pozn.: I – základní a neukončené; II – střední bez maturity; III – střední s maturitou; IV – vysokoškolské.
Note: *I – lower secondary or less; II – upper secondary (no GCSE – ISCED 353); III – upper secondary (GCSE – ISCED 344, ISCED 354, ISCED 4; IV – tertiary.*
Zdroj: Zpracovatelská databáze SLDB 2011, vlastní výpočty.
Source: *2011 Housing and Population Census processing database, author's calculations.*

Graf 3: Vyšetřenost obce pracoviště podle věku zaměstnaných obyvatel (oficiální a zrekonstruovaná data SLDB 2011) / The response rates for place of work municipality by sex and age (official and reconstructed 2011 census data)



Zdroj: Zpracovatelská databáze SLDB 2011, vlastní výpočty.
Source: *2011 Housing and Population Census processing database, author's calculations.*

vzděláním dosahovala 80,7 %, u vysokoškoláků již 94,4 %. V porovnání s oficiálními výsledky se však disproporce mezi vzdělanostními kategoriemi snížily.

Jestliže v případě vzdělání se mezi kategoriemi snížily rozdíly ve vyšetřenosti, v případě pohlaví se rozdíl zcela vyrovnal, resp. vyšetřenost u žen (89,7 %) byla dokonce nepatrně vyšší než u mužů (89,3 %), a to celkově i v rámci vzdělanostních skupin (obr. 5). Podle oficiálních výsledků přitom muži uvedli obec pracoviště častěji (viz tab. 4).[14)]

Poněkud odlišný charakter ve srovnání s oficiálními daty vykázaly též výsledky podle věku. Zatímco v oficiálních datech vyšetřenost intenzivně rostla od věku počátku ekonomické aktivity do maxima v 27. roku věku a poté následoval plynulý pokles vyšetřenosti s věkem, u zrekonstruovaných výsledků nehrál věk podstatnou roli, s výjimkou počátečních a koncových fází produktivního období života (graf 3).

Poměrně výrazně se ve srovnání s oficiálními výsledky proměnilo rozložení vyšetřenosti podle odvětví ekonomické činnosti. Doplněním rekonstruovaných dat o obci pracoviště se významně projevila (v původních datech nezaznamenaná) negativní korelace mezi vyšetřeností a podílem zaměstnaných bez stálého místa pracoviště na počtu zaměstnaných v daném odvětví. Mezi nejproblematičtější se tak posunula odvětví stavebnictví a zemědělství, lesnictví a rybářství. Důvodem mohly být zřejmě obtíže pracovníků těchto odvětví (práce venku, na různých místech) určit své pracoviště adresou. Podobně jako u většiny výše hodnocených charakteristik se však i v případě odvětví zmenšily rozdíly mezi jednotlivými kategoriemi ve srovnání s oficiálními výsledky.

ZÁVĚR

Z analýzy dat SLDB 2011 o místě pracoviště/školy vyplynulo, že jejich vyšetřenost nebyla v populaci rovnoměrná. Ze zkoumaných proměnných závisela především na úrovni vzdělání, na státním občanství, byla územně diferencována podle poměrně zřetelného prostorového vzorce. Podle oficiálně publikovaných výsledků měly na vyšetřenost vliv i další zkoumané charakteristiky, jako pohlaví, věk či odvětví ekonomiky. Na zrekonstruovaných datech se však vliv těchto charakteristik výrazněji neprokázal, tj. měly spíše dopad na chyby či formální nedostatky odpovědí, který lze vhodně položenou otázkou a kvalitním zpracováním zmírnit. Naopak při hodnocení prostorového rozložení vyšetřenosti jasněji vykrystalizovala souvislost s mírou religiozity, účastí ve volbách nebo podílem rodáků, tedy charakteristikami spojovanými s určitým vědomím sounáležitosti s místem a společností a pozitivnějším vnímáním institucí. To mělo dopad i na přístup ke sčítání.

Byly odhaleny určité rezervy ve vytěžení vstupních dat ze sčítacích formulářů při zpracování oficiálních výsledků. Obec dojížďky se podařilo doplnit u 1,6 milionu zaměstnaných/studujících, čímž se celková vyšetřenost zvýšila o 26 procentních bodů. Z toho je zjevné, že dominantním problémem nebyla neochota respondentů otázku zodpovědět, ale nepřesnosti v odpovědích a problémy v postupu zpracování. To je poměrně významné zjištění, které vyvolává i jistou naději, že při nalezení vhodné formulace a strukturování otázky při příštím sčítání by mohlo být dosaženo kvalitnějších výsledků než v SLDB 2011.

Opatření vedoucí ke zkvalitnění výsledků příštího cenzu je třeba hledat ve fázi sběru dat i ve fázi jejich zpracování. Za zvážení stojí např. upuštění od zjišťování ve struktuře okres-obec, která respondentům činila problémy. Poněkud problematický byl také způsob kombinování lokalizace pracoviště (zaškrtávací pole) a konkrétní adresy. V každém případě bude nutné podobu všech otázek na formuláři otestovat na vzorku veřejnosti, ještě před provedením zkušebního sčítání. Dále bude třeba klást co největší důraz na využití elektronických formulářů, které vyplnění otázek usnadňují a do určité míry mohou zajišťovat úplnost, formální správnost a konzistenci odpovědí.

Ve fázi zpracování bude nutné hledat vhodné metody vyhodnocování odpovědí s důrazem na zpracování nepřesných, neúplných či částečně

14) Ze zcela chybějících odpovědí dokonce vyplývá opak, zaměstnaní muži otázku vynechali častěji než ženy. To koresponduje se zkušenostmi z řady výsledků minulých cenzů, na nichž lze obvykle pozorovat úplnější výsledky u žen než u mužů.

chybných údajů. Sem patří např. i metody měření podobnosti textů, z nichž jedna byla využita v předkládaném článku (díky intenzivnímu vývoj v této oblasti jsou v současnosti k dispozici výrazně komplexnější metody, ovšem ne vždy jsou vyhovující pro specifické problémy sčítání). Výzvou bude definování takových postupů, které budou v průběhu zpracování maximálně modifikovatelné a opakovatelné v případě, kdy výsledky předdefinovaných postupů nebudou z důvodu neočekávaných problémů v uspokojivé kvalitě. S tím úzce souvisí nutnost odolat enormnímu tlaku ze strany uživatelů na co nejvčasnější publikování výsledků.

Hlavním cílem této práce bylo identifikovat příčiny nízké vyšetřenosti údajů o dojížďce ve sčítání 2011 a rámcově navrhnout některé možné způsoby, jak zjištěným problémům předejít v příštím cenzu. Navazovat bude práce zaměřená na porovnání oficiálních a zrekonstruovaných výsledků. Jejím cílem bude vyhodnotit, zda se částečným vyplněním „mezer" v oficiálních výsledcích významněji změní informace o rozmístění pracovních míst a o prostorových vztazích v území (podle prvotního porovnání např. u 427 obcí došlo ke změně dominantního vyjížďkového proudu, snížila se dominance Prahy vyjádřená podílem obsazených pracovních míst, mírně vzrostl význam meziobecní dojížďky, atd.). Tak bude možné odhadnout, zda publikované údaje poskytují dostatečně kvalitní obraz skutečného stavu.

Literatura

- Bernard, J. – Šimon, M. 2017. Vnitřní periferie v Česku: Multidimenzionalita sociálního vyloučení ve venkovských oblastech. *Sociologický časopis (Czech Sociological Review)*, 53(1), s. 3–28.
- ČSÚ. 2013. *Sčítání lidu, domů a bytů 2011 – pramenné dílo.* Praha: Český statistický úřad.
- ČSÚ. 2014. *Regionalizace dojížďky do zaměstnání podle výsledků sčítání lidu 2011.* Praha: Český statistický úřad.
- ČSÚ. 2019. *Příjmy a životní podmínky domácností.* Praha: Český statistický úřad.
- EUROSTAT. 2015. *ESS handbook for quality reports.* 2014 Edition. Luxembourg: Publications Office of the European Union.
- *Eurostat's Concepts and Definitions Database* [online]. Luxembourg: Publications Office of the European Union. [cit. 25.2.2020]. Dostupné z: <https://ec.europa.eu/eurostat/ramon/index.cfm?TargetUrl=DSP_PUB_WELC>.
- Fellegi, I. P. – Sunter, A. B. 1969. A Theory for Record Linkage. *Journal of the American Statistical Association*, Vol. 40, p. 1183–1210.
- Hampl, M. – Marada, M. 2015. Sociogeografická regionalizace Česka. *Geografie*, 120(3), s. 397–421.
- Jaro, M. A. 1989. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, Vol. 84, No. 406, p. 414–420.
- Newcombe, H. B. – Kannedy, J. M. – Axford, S. J. – James, A. P. 1959. Automatic Linkage of Vital Records. *Science*, Vol. 130, No. 3381, p. 954–959.
- Ouředníček, M. 2017. Dojížďka ve vybraných centrech Česka. In: Ouředníček, M. – Jíchová, J. – Pospíšilová, L. Eds. *Historický atlas obyvatelstva českých zemí*, Praha: Karolinum.
- Winkler, W. E. 1990. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In: *JSM proceedings, Survey research methods section*, Alexandria, VA: American Statistical Association, p. 354–359.
- Winkler, W. E. 1999. The state of record linkage and current research problems. *Statistical research report series*, Washington D.C.: U.S. Bureau of the Census.
- Winkler, W. E. 2006. *Overview of Record Linkage and Current Research Directions.* U.S. Bureau of the Census, Statistical Research Division Report.

ROBERT ŠANDA

Je absolventem Geografického ústavu Přírodovědecké fakulty Masarykovy univerzity, kde v současnosti pokračuje v doktorském studiu programu sociální geografie a regionální rozvoj. Působí na Českém statistickém úřadě v odboru statistiky obyvatelstva. Podílí se na přípravě sčítání lidu, domů a bytů v roce 2021.

SUMMARY

Commuting to work or school is one of the key topics traditionally covered by population and housing censuses. According to the official results of the 2011 Population and Housing Census, more than one-third of people who are employed or studying did not indicate the municipality to which they commute. The article aims to analyse the response rate and structures of the population of employed people or students who did not state the municipality to which they commute. Significant territorial differentiation in response rates and differentiation among specific sub-groups was identified. The response rate was strongly influenced by the level of educational attainment.

The Czech Statistical Office has anonymised raw microdata from the 2011 census questionnaires. These records were analysed for the purpose of this article. The analysis discovered that the question about the place to which people commute was too difficult for respondents, which resulted in numerous partly erroneous or inconsistent answers. The methods applied in official data processing addressed this issue to some extent, but a significant number of answers remained unrecognised.

The article describes a method of analysis using the Jaro-Winkler algorithm of probabilistic record linkage that was applied to the raw 2011 census records. The idea was to measure the similarity of official geographical names (e.g. districts, municipalities) with the answers given by respondents to identify the respondent's place of work/school municipality. Complementary steps following the automated linkage were then applied. Applying this approach to the raw data led to a significant increase in recognised answers.

To ensure better results in the next census, both data collection and processing should be improved. As regards data collection, the wording and structure of the questions on commuting will be modified. An electronic census form will be widely promoted because it can help respondents to answer questions properly. In data processing, methods that enabling the automated dealing with inconsistencies and errors shall be widely implemented.