

Dojížd'ka do zaměstnání a škol podle sčítání 2011

Proč byly výsledky tak neúplné?

Konference ČDS, Lednice 22. – 24. května 2019

Robert Šanda
Český statistický úřad;

Důvody prezentace

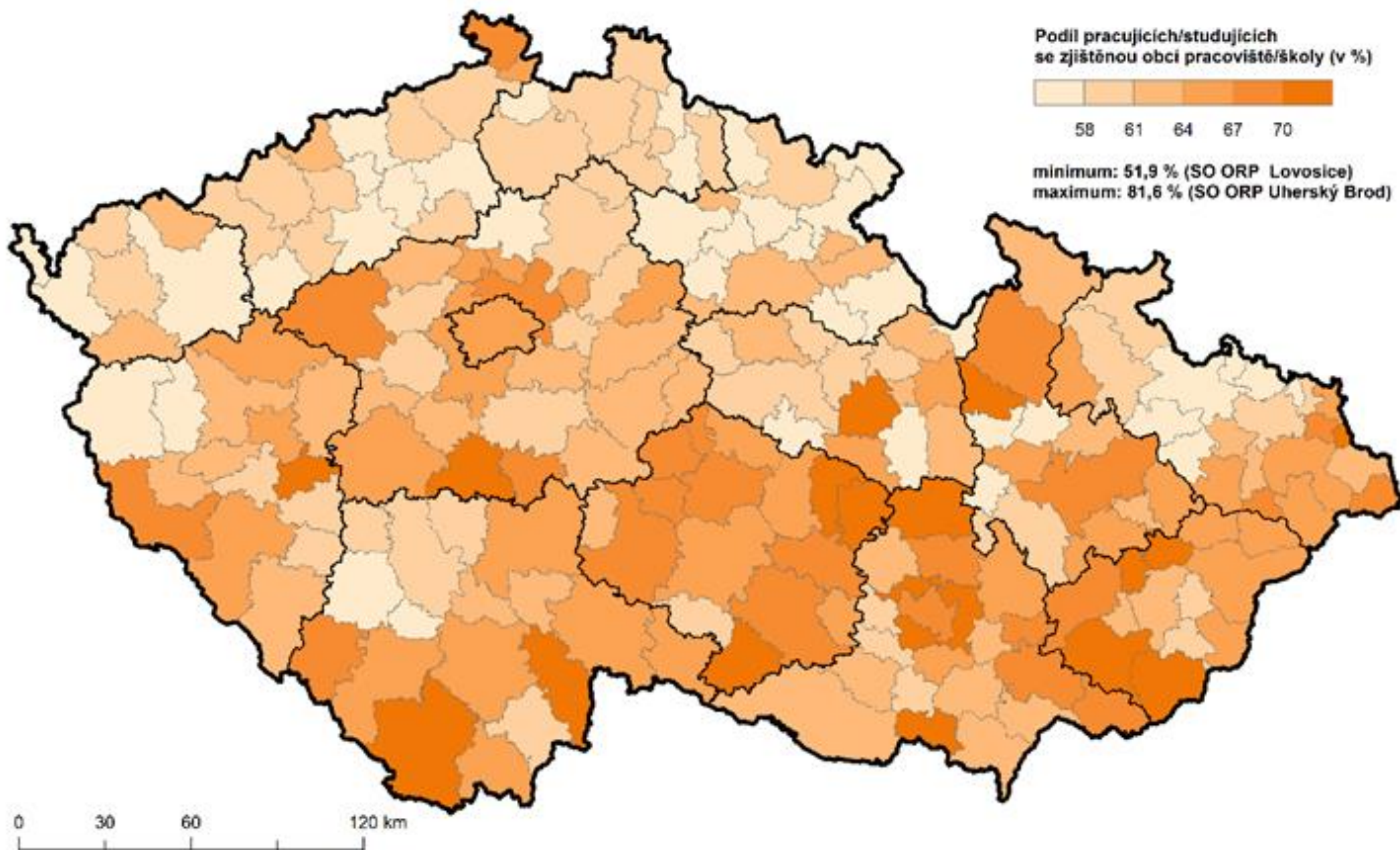
- Překvapivě nízká vyšetřenost dojížděky v roce 2011 ve srovnání s předchozími roky
- Nedostatek informací o vyšetřenosti (velikost, diferenciaci) dojížděky v datových publikacích ČSÚ
 - obvyklé řešení – předpoklad rovnoměrného rozložení
- Chybějící vysvětlení – dluh uživatelům dat
- Příprava sčítání lidu 2021

Lokalizace místa pracoviště nebo školy podle výsledků SLDB 2011

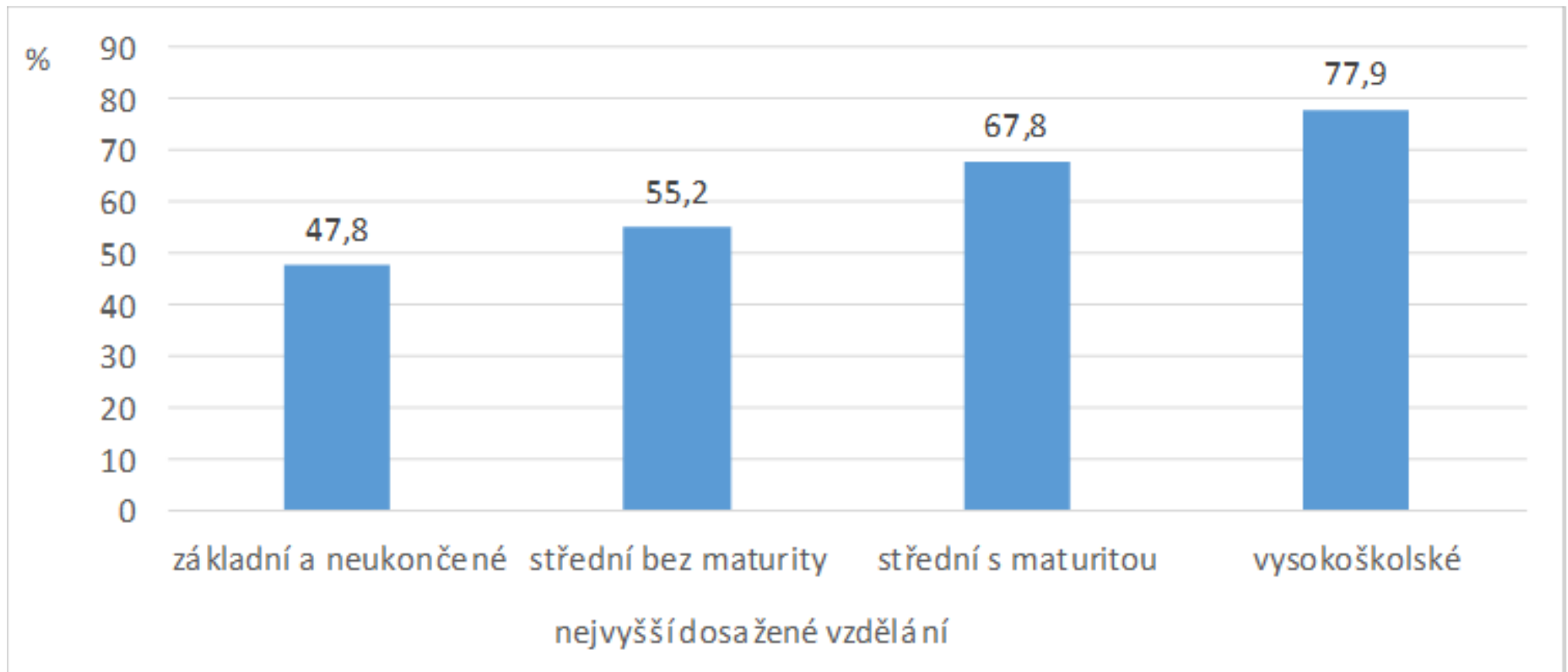
Lokalizace místa pracoviště	Absolutně			V %		
	pracující	studující	celkem	pracující	studující	celkem
na adrese místa obvyklého pobytu	580 509	170 411	750 920	12,9	11,2	12,5
na jiné adrese v rámci obce obvyklého pobytu	924 948	354 218	1 279 166	20,5	23,2	21,2
v jiné obci v rámci okresu obvyklého pobytu	596 686	184 834	781 520	13,3	12,1	13,0
v jiném okrese - obec zjištěna	488 688	231 529	720 217	10,9	15,2	12,0
v jiném okrese - obec nezjištěna	14 554	5 320	19 874	0,3	0,3	0,3
v jiném státě - stát zjištěn	35 461	7 746	43 207	0,8	0,5	0,7
v jiném státě - stát nezjištěn	1 787	305	2 092	0,0	0,0	0,0
v ČR - blíže nezjištěno	98 905	29 493	128 398	2,2	1,9	2,1
zaměstnání bez stálého pracoviště	232 986	2 076	235 062	5,2	0,1	3,9
místo pracoviště, školy zcela nezjištěno	1 526 938	539 458	2 066 396	33,9	35,4	34,3
Celkem	4 501 462	1 525 390	6 026 852	100,0	100,0	100,0

Celkový podíl pracujících/studujících s nezjištěnou obcí dojížděky: 36,7 %

Vyšetřenost obce pracoviště/školy ve správních obvodech ORP (SLDB 2011)



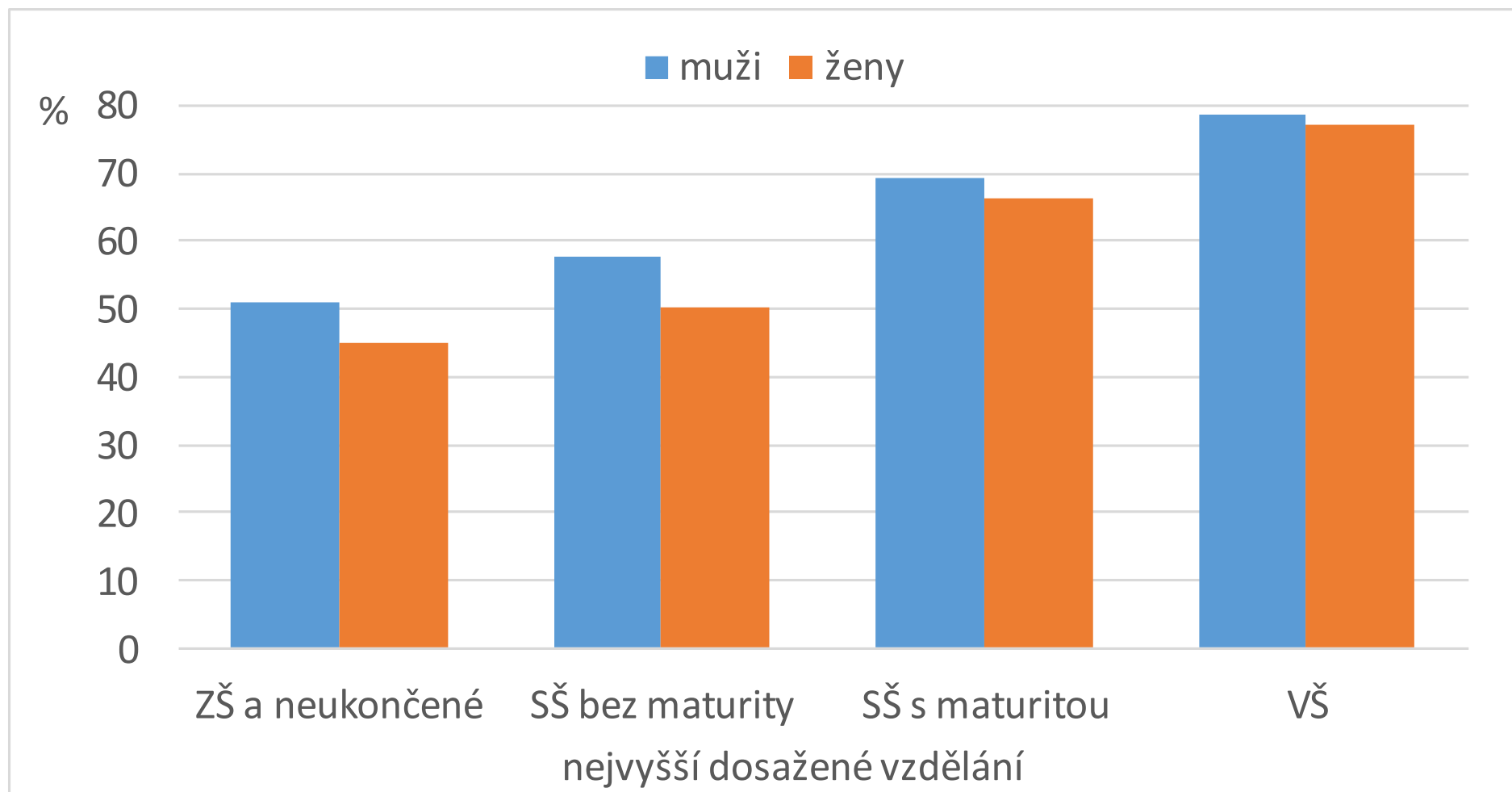
Vyšetřenost obce pracoviště podle nejvyššího dosaženého vzdělání (SLDB 2011)



Vyšetřenost obce pracoviště podle velikostní kategorie obcí a nejvyššího dosaženého vzdělání (SLDB 2011)

Velikostní kategorie obcí	Obec pracoviště				Vyšetřenost podle nejvyššího dosaženého vzdělání			
	zjištěna	nezjištěna	celkem	vyšetřenost	zákl. a neuk.	stř. bez mat.	stř. s mat.	vš
499 a méně	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
1000 – 1999	272 505	157 492	429 997	63,4	50,0	56,8	68,8	78,7
2000 – 4999	338 914	186 801	525 715	64,5	50,2	57,6	69,6	79,5
5000 - 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 a více (bez Prahy)	298 491	160 792	459 283	65,0	46,0	53,0	67,7	78,8
Praha	383 308	199 266	582 574	65,8	47,5	52,5	66,5	76,5

Vyšetřenost obce pracoviště podle pohlaví a nejvyššího dosaženého vzdělání (SLDB 2011)



Zjišťování a zpracování místa pracoviště/školy v roce 2011

Otázka na místo pracoviště nebo školy na sčítacím listu osoby

21. Místo pracoviště nebo školy

na stejné adrese, jaká je v záhlaví formuláře

jinde v České republice

uved'te kde

okres

obec

č.p./č.o.

ulice

v jiném státě

uved'te název

zaměstnání bez stálého pracoviště

Poznámka: Adresou v záhlaví formuláře je míněna adresa místa sečení.

Zpracování:

- **Digitalizace** – skenování formulářů, rozpoznání jednotlivých znaků a slov
- **Kódování** – přiřazení číselníkových kódů adrese/státu dojížděky
 - **automatické** – pro bezchybně zapsané adresy
 - **manuální** – pro adresy nezakódované automaticky
 - předmětem kódování adresa jako celek
- **Kontroly přípustnosti a logické kontroly** – uvedení komplexní otázky 21 do formálně správné podoby (lokalizace/místo v ČR/stát)
- **Odvození** údaje lokalizace místa pracoviště/školy vzhledem k místu obvyklého pobytu

Hlavní příčiny nízké vyšetřenosti dojížd'ky

(poznatky z anonymních vstupních záznamů pocházejících z digitalizace formulářů)

- **Zcela chybějící odpovědi na otázku *místo pracoviště nebo školy***
 - zhruba pětina z celkového počtu nezjištěných
- **Nepřesnosti v adrese místa pracoviště/školy „nerozluštěné“ při kódování**
 - nerespektování požadované struktury adresy při vyplňování formuláře
 - záměny krajů/okresů/obcí/částí obcí/místních názvů
- **Nedodržování instrukcí při vyplňování otázky neošetřené logickými kontrolami**
(problém listinných formulářů)

=> Nízká vyšetřenost dojížd'ky v roce 2011 byla způsobena převážně nepřesnými odpověďmi, které se nepodařilo zpracovat, nikoliv neochotou respondentů otázku zodpovědět.

- **Hlavní úskalí celkové koncepce zpracování dojížd'ky 2011**
 - pevně stanovený postup zpracování neumožňující pružně reagovat na nečekaně vysoký počet nepřesných a nekonzistentních údajů (po fázi kódování už nelze vyšetřenost pozitivně ovlivnit)
 - příliš komplikované SW řešení
 - přílišný důraz na správné zakódování jednotlivých záznamů na úkor celkové vyšetřenosti

Rekonstrukce obce pracoviště/školy v roce 2011

(jako nástin možného přístupu při příštím sčítání)

■ Cíle

- analyzovat nevyužitý potenciál vstupních dat o obci dojížděky ze sčítání 2011
- identifikovat v základních rysech možné přístupy při zpracování dat o dojížděce v příštím sčítání
- podklad pro posouzení vypovídací schopnost oficiálně publikovaných výsledků

■ Zdroj

- Zpracovatelská databáze SLDB 2011
 - vstupní záznamy ze sčítacích formulářů (po digitalizaci)
 - finální mikrodata

■ Obecný přístup

- aplikovat metody měření podobnosti textových řetězců
- zamezit podstatným chybám
- **akceptovat riziko drobných chyb na úrovni individuálních záznamů**

■ **Aplikováno na vstupní záznamy osob z formulářů, u nichž je ve výsledcích nezjištěná obec pracoviště/školy**

Rekonstrukce obce pracoviště/školy v roce 2011

Hlavní kroky (1)

■ Příprava dat

- Transformace dvojic okres-obec pracoviště (školy) z formulářů do standardizované podoby

Vstupní údaje z formuláře po digitalizaci			Údaje z formuláře - standardizované	
kolonka okres	kolonka obec		kolonka okres	kolonka obec
BRNO/MĚSTO	BRNO	→	BRNOMESTO	BRNO
ÚSTÍ NAD LABEM	ÚSTÍ-//--//	→	USTINADLABEM	USTI
RYCHNOV N./KN.	DOUDLEBÝ N./ORL.	→	RYCHNOVNKN	DOUDLEBYNORL

- Transformace oficiálních názvů v číselnících a do standardizované podoby
 - dvojice okres-obec + dvojice okres-část obce (je-li název části unikátní v rámci okresu)

Názvy z číselníků - původní			Názvy z číselníků - standardizované	
okres	obec		okres	obec
Benešov	Benešov	→	BENESOV	BENESOV
Benešov	Bernartice	→	BENESOV	BERNARTICE
...	...	→	...	...
Žďár nad Sázavou	Žďár nad Sázavou	→	ZDARNADSAZAVOU	ZDARNADSAZAVOU

Rekonstrukce obce pracoviště/školy v roce 2011

Hlavní kroky (2)

■ Propojení záznamů z předchozích kroků (referenční seznam – formulář)

■ Deterministické (přesná shoda)

Formulář (standardizované řetězce)		Číselníky (standardizované názvy)		Podobnost	
kolonka okres	kolonka obec	okres	obec	okres	obec
BRNOMESTO	BRNO	BRNOMESTO	BRNO	1,00	1,00

■ Pravděpodobnostní

- měření podobnosti řetězců (metoda Jaro-Winkler)
- výběr metody: Jaro-Winkler
- určení prahových hodnot podobnosti (Φ_{jw}):

$$\min(\Phi_{jw, \text{okres}}, \Phi_{jw, \text{obec}}) \geq 0,89$$

nebo,

$$\min(\Phi_{jw, \text{okres}}, \Phi_{jw, \text{obec}}) \geq 0,86 \text{ a zároveň } \max(\Phi_{jw, \text{okres}}, \Phi_{jw, \text{obec}}) \geq 0,95$$

Formulář (standardiz.)		Referenční seznam (standardiz.)		Podobnost	
kolonka okres	kolonka obec	okres	obec	okres	obec
USTINADLABEM	USTI	USTINADLABEM	USTINADLABEM	1,00	0,87
RYCHNOVNKN	DOUDLEBYNORL	RYCHNOVNADKNEZNOU	DOUDLEBYNADORLICI	0,92	0,94

Rekonstrukce obce pracoviště/školy v roce 2011

Hlavní kroky (3)

■ Opatření pro snížení rizika chybných propojení

■ Vytvoření kontrolního referenčního seznamu dojížděkových proudů

- proudy z oficiálních výsledků + proudy vzniklé na základě výsledků deterministického propojení
- Výsledek pravděpodobnostního propojení nesmí založit nový dojížděkový proud, ale pouze posílit některý z referenčních proudů

Formulář (standardizované řetězce)		Číselníky (standardizované názvy)		Podobnost	
kolonka okres	kolonka obec	okres	obec	okres	obec
JESENICE	JESENICE	JESENÍK	JESENÍK	0,93	0,93



■ Identifikace typických chyb v pravděpodobnostním propojení na základě:

- otestování na souboru „oficiálně“ zjištěných výsledků (tj. záznamů úspěšně zpracovaných při kódování)
- vyhodnocení „oficiálního“ vs. zrekonstruovaného podílu obsazených pracovních míst na ČR (manuální ověření případů vysokých nárůstů relativních počtů OPM)

Formulář (standardizované řetězce)		Číselníky (standardizované názvy)		Podobnost	
kolonka okres	kolonka obec	okres	obec	okres	obec
TURNOV	TURNOV	TRUTNOV	TRUTNOV	0,91	0,91



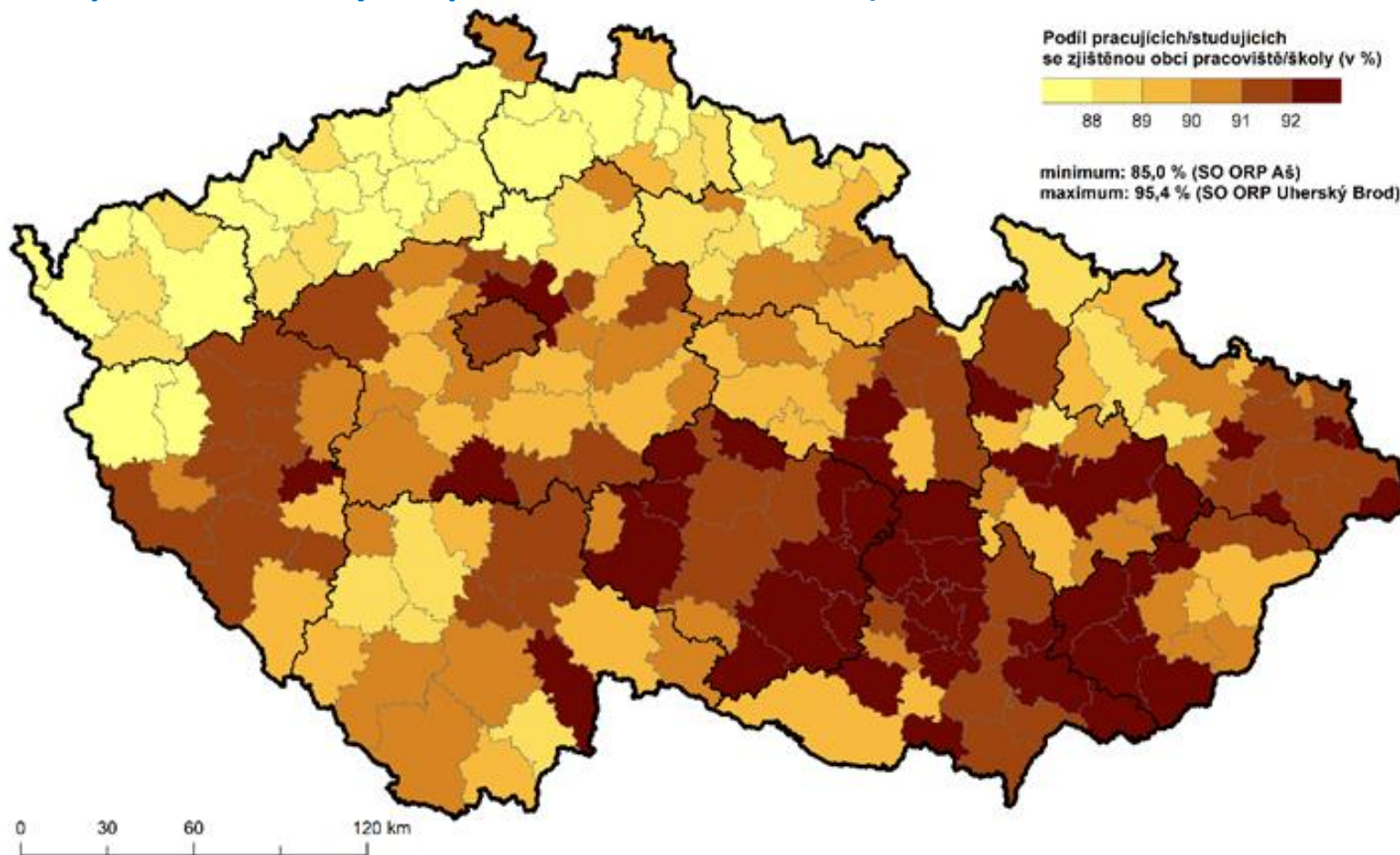
Rekonstrukce obce pracoviště/školy v roce 2011

Základní výsledky (1)

Vyšetřenost obce pracoviště/školy podle oficiálně publikovaných a podle zrekonstruovaných dat (SLDB 2011)

Zdroj	Obec dojíždky zjištěna	Obec dojíždky nezjištěna	Pracující/studující celkem	Vyšetřenost
oficiální výsledky	3 812 184	2 214 668	6 026 852	63,3
zrekonstruované výsledky	5 458 759	568 093	6 026 852	90,6

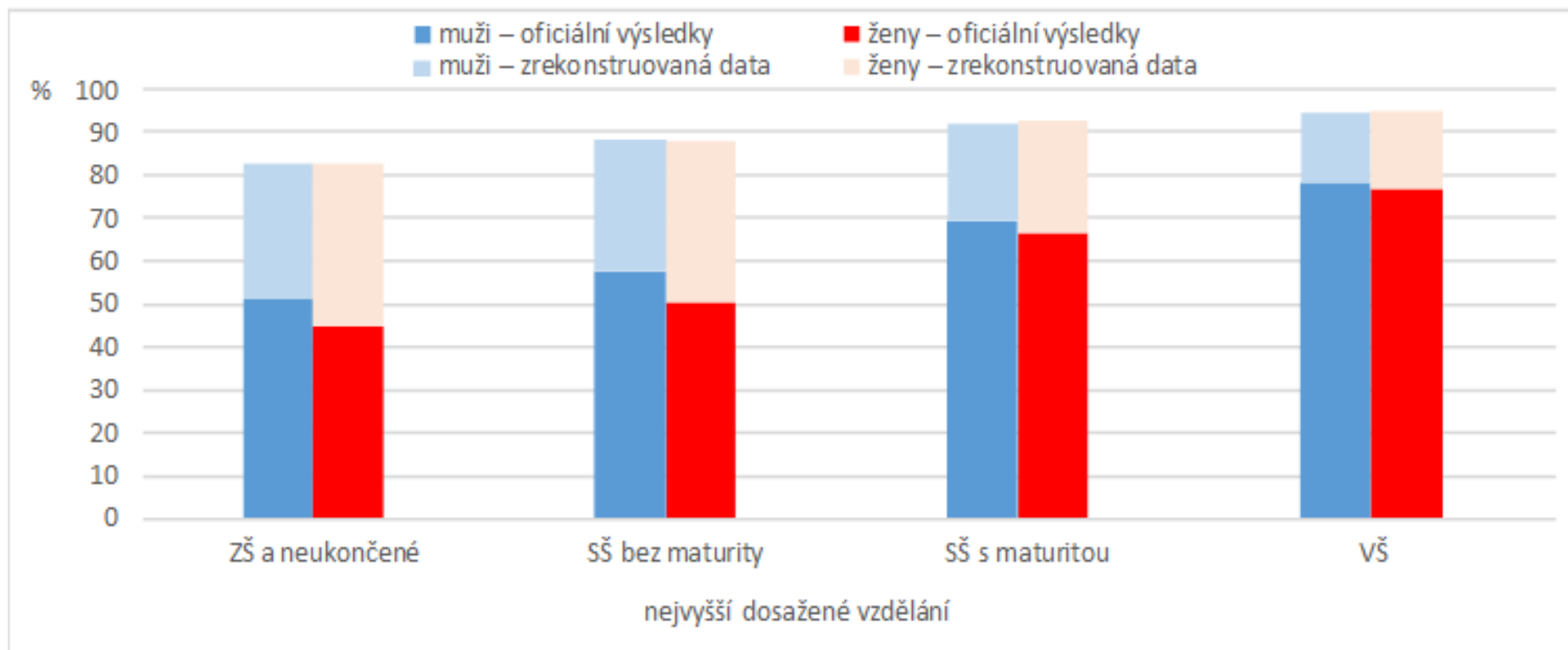
Vyšetřenost obce pracoviště/školy ve správních obvodech ORP (zrekonstruovaná data SLDB 2011)



Rekonstrukce obce pracoviště/školy v roce 2011

Základní výsledky (2)

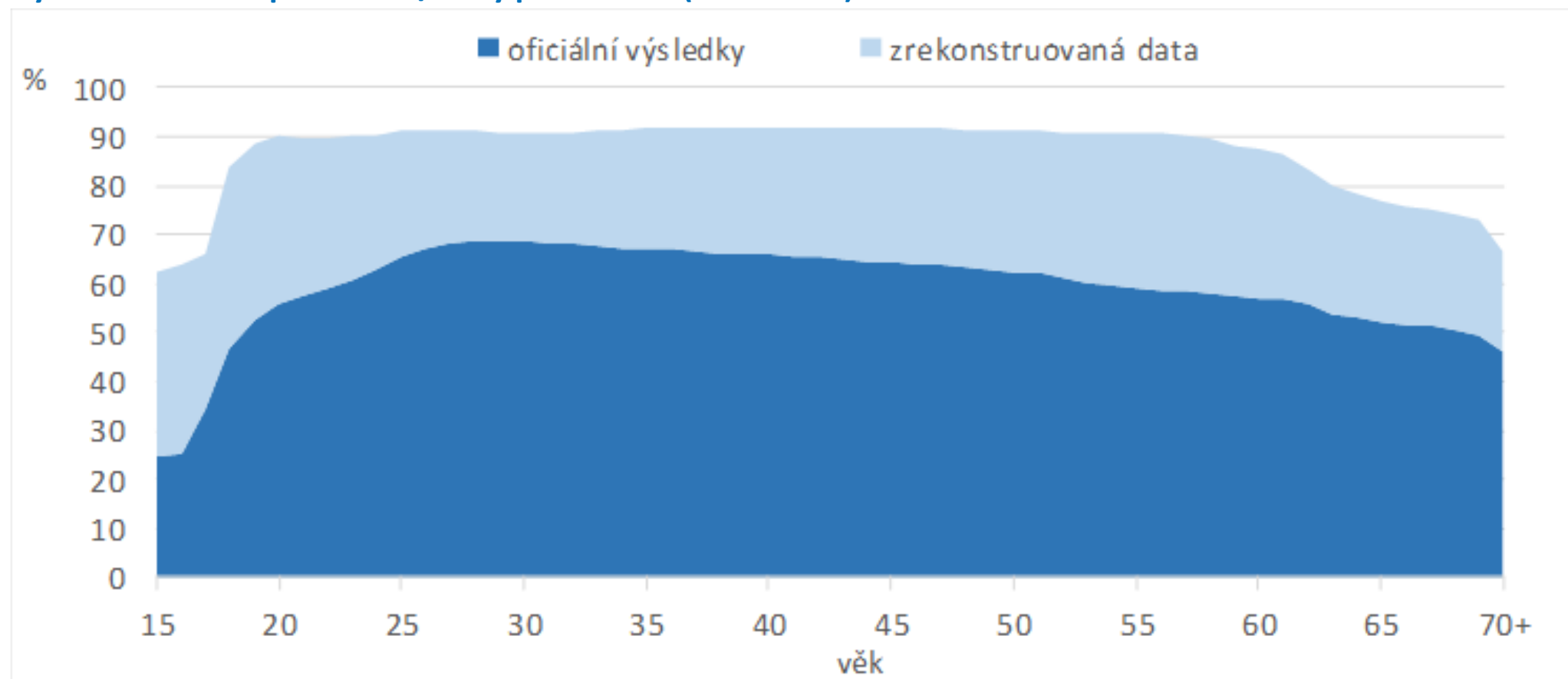
Vyšetřenost obce pracoviště/školy podle pohlaví a vzdělání (SLDB 2011)



Rekonstrukce obce pracoviště/školy v roce 2011

Základní výsledky (3)

Vyšetřenost obce pracoviště/školy podle věku (SLDB 2011)



Shrnutí

- Nízká vyšetřenost dojížděky ve výsledcích sčítání nebyla primárně způsobena neochotou respondentů, ale nepřesnostmi v odpovědích.
- Prostorový vzorec vyšetřenosti v hrubých rysech korespondoval s územní diferenciací charakteristik naznačujících vztah obyvatel k okolí, společnosti, resp. k institucím (rodáci, volební účast, religiozita). Na zrekonstruovaných datech se souvislost s těmito charakteristikami zvýraznila.
- Další zkoumané charakteristiky obyvatel (vzdělání, věk, odvětví,...) významněji ovlivňovaly vyšetřenost v oficiálních výsledcích, po provedené rekonstrukci se jejich vliv na vyšetřenost snížil.
- Informace z formulářů se při zpracování nepodařilo plně využít.
- Otázky na místo dojížděky a související charakteristiky je třeba koncipovat srozumitelněji než v roce 2011 (např. adresa ve formě běžné při korespondenci).
- Je nutné klást důraz na vysokou účast v online sčítání (výrazně kvalitnější data).
- Architektura zpracování při příštím sčítání by měla umožňovat flexibilně reagovat na průběžné výsledky jednotlivých fází zpracování.
- Při zpracování je třeba více se zaměřit na celkovou úplnost než na (příliš) důkladnou verifikaci jednotlivých záznamů (akceptování sporadických chyb neovlivňujících celkový obraz lze považovat za legitimní).
- Je žádoucí vyhnout se příliš robustnímu SW řešení při zpracování.

Děkuji za pozornost.

Robert Sanda (robert.sanda@czso.cz)