

SČÍTÁNÍ LIDU 2021

MEZINÁRODNÍ KONTEXT A VYUŽITÍ
ADMINISTRATIVNÍCH ZDROJŮ DAT

Robert Šanda (ČSÚ)

Valná hromada ČStS, 3. května 2022, ČSÚ

Významné milníky ve vývoji evropských sčítání lidu

- **Polovina 18. století** – pokusy o pravidelná sčítání zahrnující veškeré obyvatelstvo státu
- **Polovina 19. století** – počátek moderních sčítání, snahy o centralizaci sčítání a mezinárodní srovnatelnost (vznik statistických institucí, první Mezinárodní statistický kongres)
- **Konec 19. století** – centrální zpracování dat pomocí děroštitkových strojů (v Rakousku od roku 1890), umožňující vytváření řady kombinací a odvozených ukazatelů
- **Polovina 20. století** – pokusy o spojení sčítání lidu s jinými cenzy (u nás spojení se sčítáním domů a bytů, neúspěšný pokus o spojení s agrocenzem)
- **70. léta 20. století** – využívání administrativních zdrojů
- **Sčítání 2011** – závazný evropský právní rámec, masivní odklon od klasického dotazníkového šetření, snahy o integraci šetření v oblasti sociální statistiky

Koordinace sčítání na mezinárodní úrovni (sčítání kolem 2020)

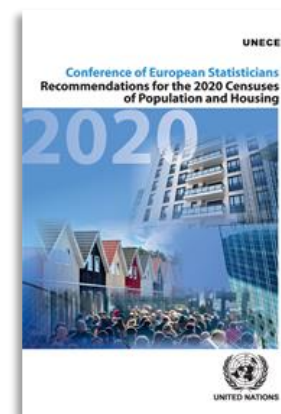
■ Rezoluce OSN Světový program sčítání lidu kolem roku 2020

- Schválena v roce 2015
- Základní parametry sčítání
- Referenční období 2015 – 2024



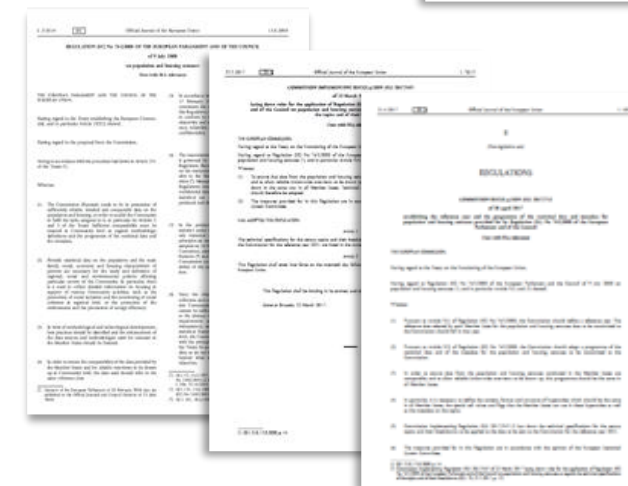
■ Principy a doporučení Konference evropských statistiků ke sčítání lidu kolem roku 2020

- Publikována v roce 2016
- Definice klíčových (core) a doplňkových (non-core) témat



■ Evropské právní předpisy

- Závazné pro členské země EU
- jednotný referenční rok
- Obvykle bydlící obyvatelstvo
- Struktury, metadata, zprávy o kvalitě sčítání



Základní rysy sčítání lidu podle Doporučení KES

- Individuální zjišťování
- Simultánnost (rozhodný okamžik, resp. rozhodné období)
- Univerzálnost (celá populace na území státu, údaje zjišťované jednotným způsobem)
- Územní detail
- Periodicita
- Mezinárodní srovnatelnost

Formy sčítání lidu

■ Formy sčítání lidu podle evropského nařízení

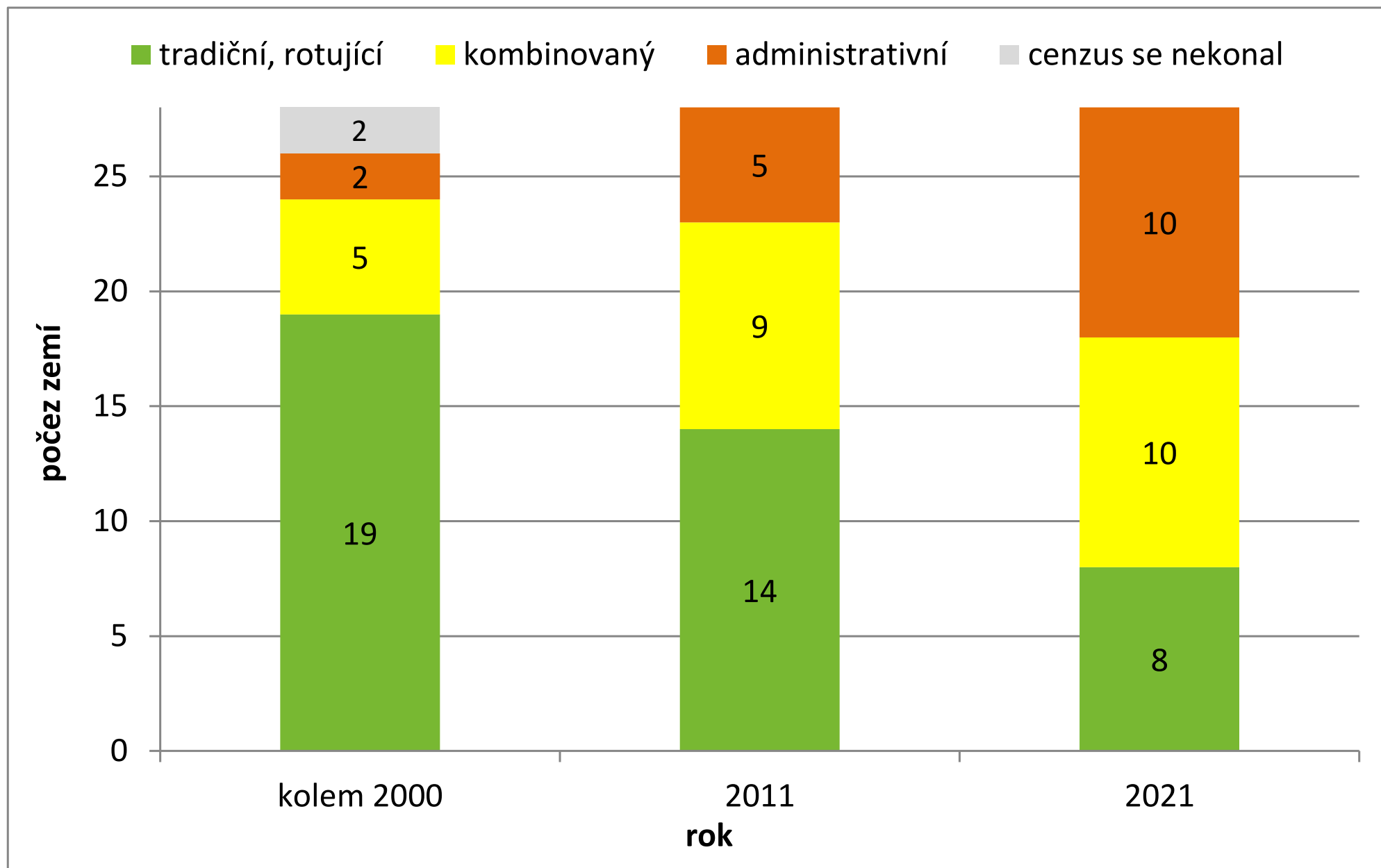
- tradiční dotazníkové šetření
- administrativní sčítání (sčítání založené pouze na registrech)
- kombinace tradičního sčítání s administrativním
- kombinace s výběrovými šetřeními
- sčítání založené na rotujících výběrech

■ Dotazníkové šetření

- papírové formuláře (sebesčítání vs. rozhovor tazatele s respondentem - PAPI)
- elektronické formuláře (CAPI, CAWI, CATI)

=> značná různorodost zdrojů

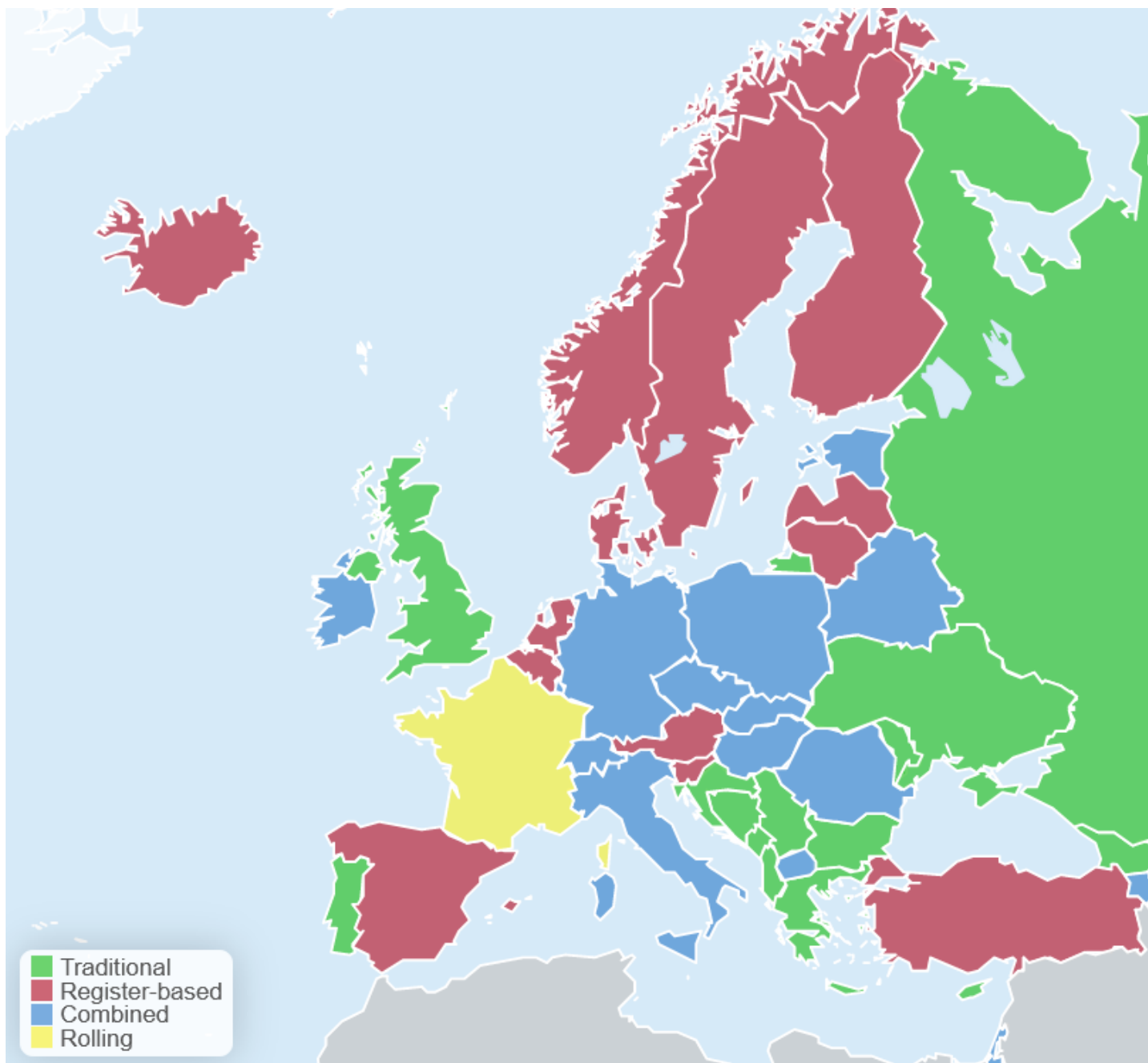
Formy sčítání členských zemí v EU v období 2000 – 2021 (včetně UK)



Základní typy sčítání lidu v zemích EU (EU-28) v letech 2011 a 2021

		Typ sčítání 2011 <i>Type of 2011 census</i>			počet zemí <i>number of countries</i>
		tradiční <i>traditional</i>	kombinované <i>combined</i>	administrativní <i>register-based</i>	
Typ sčítání 2021 <i>Type of 2021 census</i>	tradiční <i>traditional</i>	Bulharsko/ <i>Bulgaria</i> Chorvatsko/ <i>Croatia</i> Kypr/ <i>Cyprus</i> Francie/ <i>France</i> Řecko/ <i>Greece</i> Malta Portugalsko/ <i>Portugal</i> Spojené království/ <i>UK</i>			8
	kombinované <i>combined</i>	Maďarsko/ <i>Hungary</i> Irsko/ <i>Ireland</i> Itálie/ <i>Italy</i> Lucembursko/ <i>Luxembourg</i> Rumunsko/ <i>Romania</i> Slovensko/ <i>Slovakia</i>	Česko/ <i>Czechia</i> Estonsko/ <i>Estonia</i> Německo/ <i>Germany</i> Polsko/ <i>Poland</i>		10
	administrativní <i>register-based</i>		Belgie/ <i>Belgium</i> Lotyšsko/ <i>Latvia</i> Litva/ <i>Lithuania</i> Nizozemsko/ <i>Netherlands</i> Španělsko/ <i>Spain</i>	Rakousko/ <i>Austria</i> Dánsko/ <i>Denmark</i> Finsko/ <i>Finland</i> Slovinsko/ <i>Slovenia</i> Švédsko/ <i>Sweden</i>	10
počet zemí <i>number of countries</i>		14	9	5	28

Základní formy sčítání kolem roku 2020 v Evropě



SLDB 2021

Základní rysy

- **Rozhodný okamžik: půlnoc z 26. na 27. března 2021**
- **Kombinovaná forma (poprvé v roce 2011)**
 - celoplošné dotazníkové šetření
 - administrativní zdroje dat
- **Strategie**
 - maximální využití administrativních zdrojů
 - redukce obsahu formulářů na nejnutnější otázky
- **Sčítací formuláře**
 - elektronické (online vyplňování)
 - papírové formuláře
- **Vyplňování elektronických i listinných formulářů respondenty (self-enumeration)**
- **Sledovaná populace**
 - obvykle bydlící
 - „trvale“ bydlící

SLDB 2021

Zdroje dat o osobách

■ Konstitutivní zdroje

- sčítací formuláře
- základní registr obyvatel (ROB) + jeho agendové (zdrojové) systémy AISEO a AISC

■ Doplnkové administrativní zdroje

- centrální registr pojištěnců (zdravotní pojištění)
- integrovaný IS České správy sociálního zabezpečení
 - zaměstnaní, poživatelé důchodů, příjemci peněžité pomoci v mateřství
- systémy Ministerstva práce a sociálních věcí
 - uchazeči o zaměstnání, rodičovský příspěvek, příjemci pomoci v hmotné nouzi
- databáze žáků a studentů od základních škol po VŠ (VŠ včetně nedávných absolventů)
- údaje z přiznání k dani z příjmů za rok 2020

Administrativní data o osobách ve sčítání 2021

Využití ve zpracování 2021

■ Vymezení populace

- Základní registr obyvatel + agendový IS evidence obyvatel a agendový IS cizinců – konstitutivní zdroj
- Ostatní AZD – doplňkové informace (signs-of-life analýza)

■ Výhradní zdroj údajů o osobách:

- Místo registrovaného pobytu
- Státní občanství
- Rodinný stav
- Rok příchodu do země
- Postavení v zaměstnání

■ Alternativní zdroj pro údaje:

- Místo obvyklého pobytu,
- Místo obvyklého pobytu rok před sčítání a po narození
- Počet dětí
- Úroveň vzdělání
- Základní vztahy mezi osobami v domácnosti (rodinná jádra)
- (Ekonomické charakteristiky – ekonomická aktivita, odvětví)

Údaje zjišťované ve sčítání, neobsažené v admin. zdrojích

- Údaje o bytech
- Vazba obyvatel na byt (adresa pobytu do úrovně bytu – pro tvorbu domácností a charakteristiky bydlení)
- Vztahy mezi osobami v bytě (typy domácností)
- Úroveň vzdělání za většinu obyvatel
- Pobyt po narození (dříve bydliště matky v době narození) za podstatnou část obyvatel
- Místo pracoviště
- Charakteristiky dojížděky do zaměstnání/školy (frekvence, doba každodenní dojížděky, prostředek)
- Zaměstnání
- Sociokulturní charakteristiky
 - Mateřský jazyk
 - Národnost
 - Náboženská víra

Hlavní etapy zpracování dat

- **Digitalizace listinných sčítacích formulářů**
 - celkem bylo do elektronické podoby převedeno 816,3 tisíc formulářů
- **Kódování (zařazení údajů z formulářů do kategorií)**
 - 27,5 mil. údajů kódováno automaticky, 2,4 manuálně
 - největší a časově nejnáročnější část manuálně kódovaných údajů představují údaje o odvětví ekonomické činnosti a zaměstnání
- **Propojování formulářů s administrativními zdroji, „deduplikace“**
- **„Signs-of-life“ analýza**
- **Vymezení obvykle a „trvale“ bydlícího obyvatelstva**
- **Zařazení všech sečtených osob do domácností, domácností do bytů, bytů do domů**
- **Anonymizace**
- **Logické kontroly, odvozování dalších charakteristik**
- **Tvorba agregovaných výsledků**

Vybrané etapy zpracování dat - kódování

- Navazuje na příjem online formulářů a na skenování a digitalizaci listinných formulářů
- Všechny volně zapisované odpovědi (na LSF více položek než na ESF)
- Automatické kódování
 - Nad očekávání úspěšné u adres (adresa obvyklého pobytu, adresa pracoviště)
 - Nejnáročnější položky: odvětví (pouze u LSF) a zaměstnání
- Manuální kódování
 - Pracovníci manuálního kódování zejména z řad studentů demografie a geografie PŘF UK
 - Ad hoc vyvinutá aplikace (ICZ)
 - k dispozici seznamy, slovníky atd, funkce našeptávače
- Prioritní význam kódování adresy obvyklého pobytu

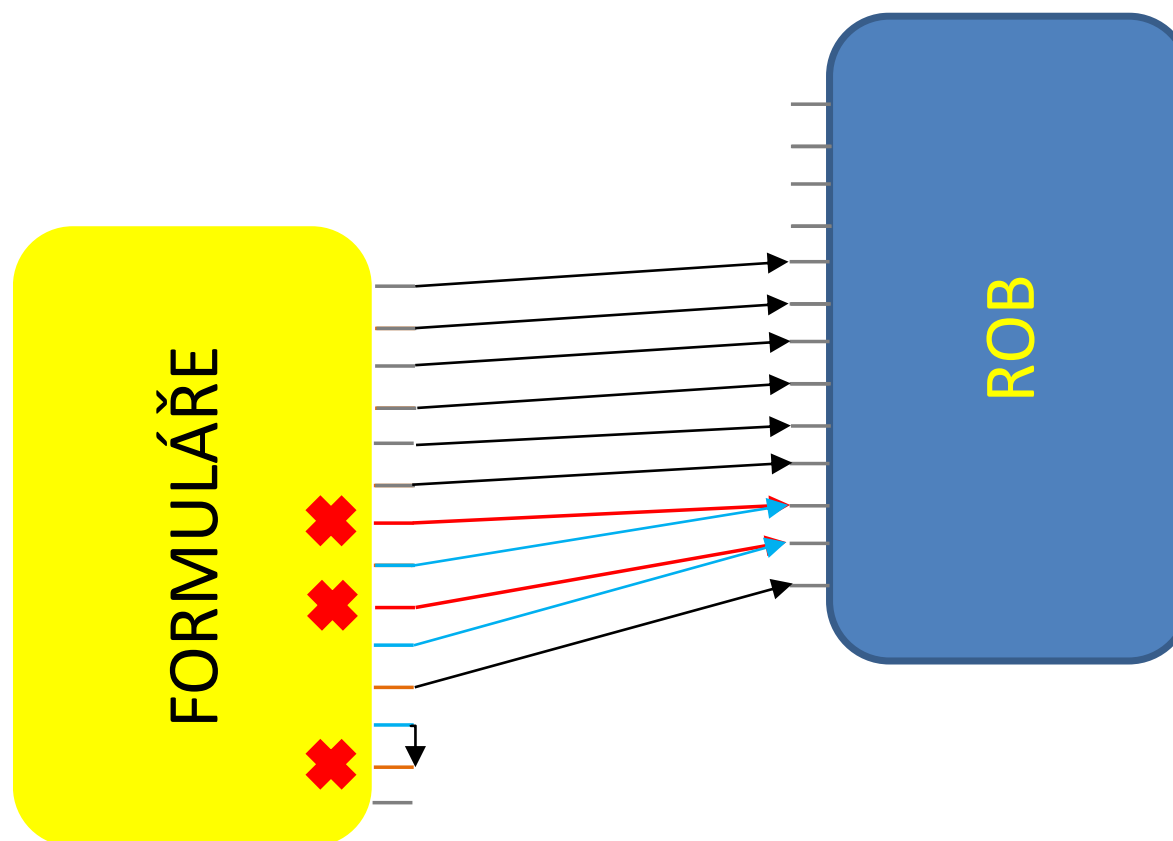
Vybrané etapy zpracování dat – napojování osob z formulářů na záznamy ROB

- Výrazný dopad na přesnost výsledného počtu obyvatel a konzistence výsledků
- Hlavní požadavek: minimalizace chyb, „vybalancování“ rizik obou druhů chyb (chybné propojení vs. chybné nepropojení)
- Postup
 - A) Standardizace identifikačních údajů (na straně formulářů i ROB)
 - např. rodná čísla 530512 / 118 -> 530512118
 - několik způsobů standardizace jmen:
 - původní záznam (ilustrační příklad)
 - jméno: Ing. Anna-Marie
 - příjmení: Horáková, Ph.D.
 - Standardizace 1: příjmení: **HORAK**
 jméno: **ANNA MARIE**
 - Standardizace 2: příjmeníjméno: **HORAKANNAMARIE**
 - Standardizace 3: celé jméno abecedně : **ANNA~HORAKOVA~MARIE**

Vybrané etapy zpracování dat – napojování osob z formulářů na záznamy ROB

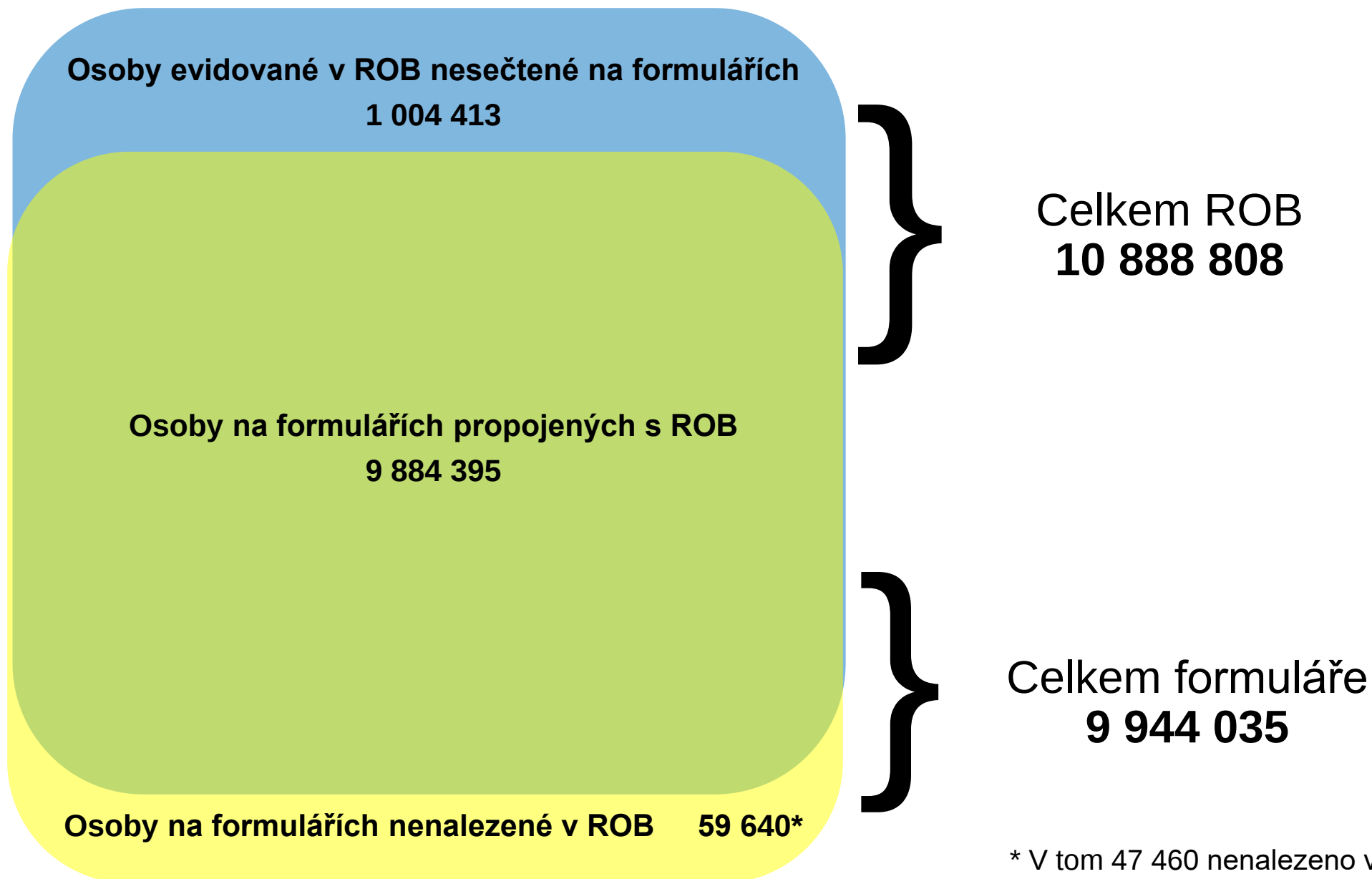
- **B) Vytvoření „black listů“**
 - neunikátní kombinace jméno-příjmení-datum narození v ROB
 - neunikátní rodná čísla v ROB
 - „půjčované“ doklady na formulářích
 - ...
- **C) Vlastní propojování podle série pravidel**
 - Deterministické propojování
 - Pravděpodobnostní propojování
 - **Levenshteinova vzdálenost** (počet rozdílů ve znacích)
 - **Jaro-Winkler** (počet shodných znaků v polovině řetězce, porovnání jejich pozic, bonifikace shodných začátků řetězců)
 - **„symetrická difference“** – počet shodných a počet rozdílných slov
- Zhruba dvacet pravidel, v každém vyžadována přesná shoda alespoň jednoho identifikačního údaje (např. rodné číslo, id datové schránky,...)

Vybrané etapy zpracování dat – prioritizace a deduplikace



- Duplicita (multiplicita): více formulářů napojených na jeden záznam ROB
- Duplicity poprvé řešeny v roce 2001
- V roce 2021 poměrně velké množství duplicit (zejména elektronické formuláře)
- Stanovena sada pravidel pro výběr prioritního záznamu z formuláře

Výsledek napojování formulářů na ROB a deduplikace



* V tom 47 460 nenalezeno vůbec, 12 283 nalezeno mezi „neplatnými“ záznamy

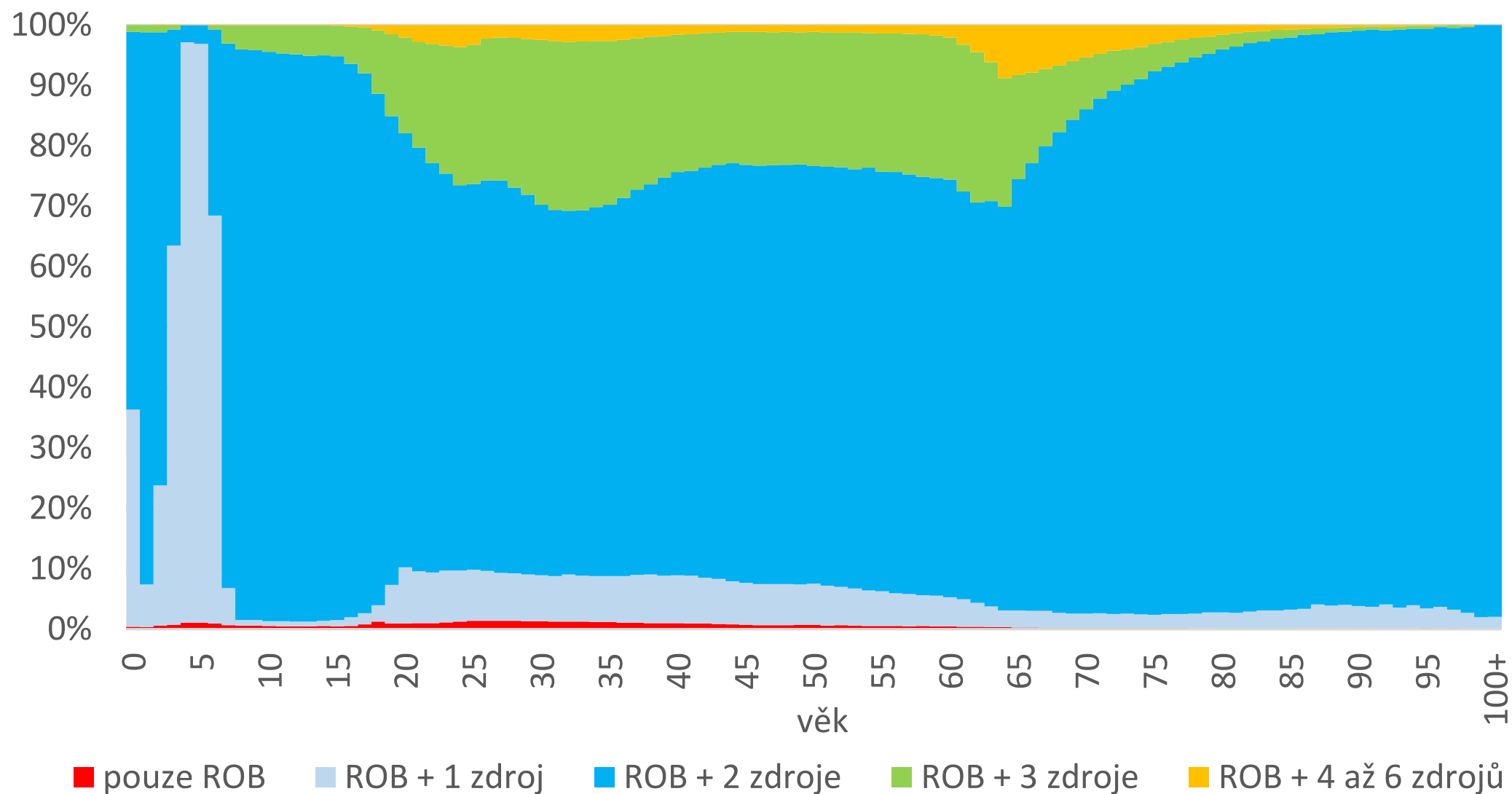
Signs-of-life analýza

- Posouzení faktické přítomnosti osob evidovaných v populačním registru na základě administrativních dat (populační registry běžně obsahují záznamy osob, které na území daného státu již nežijí)
- V různých podobách aplikována v řadě zemí provádějících kombinované nebo čistě administrativní sčítání (např. Rakousko, Švédsko, Španělsko, Estonsko,...)
- Nutný přístup k co největšímu počtu administrativních zdrojů
- Ve sčítání 2011 přístup pouze k evidenci obyvatel (ISEO), signs-of-life analýza proveditelná pouze ve velmi omezené míře

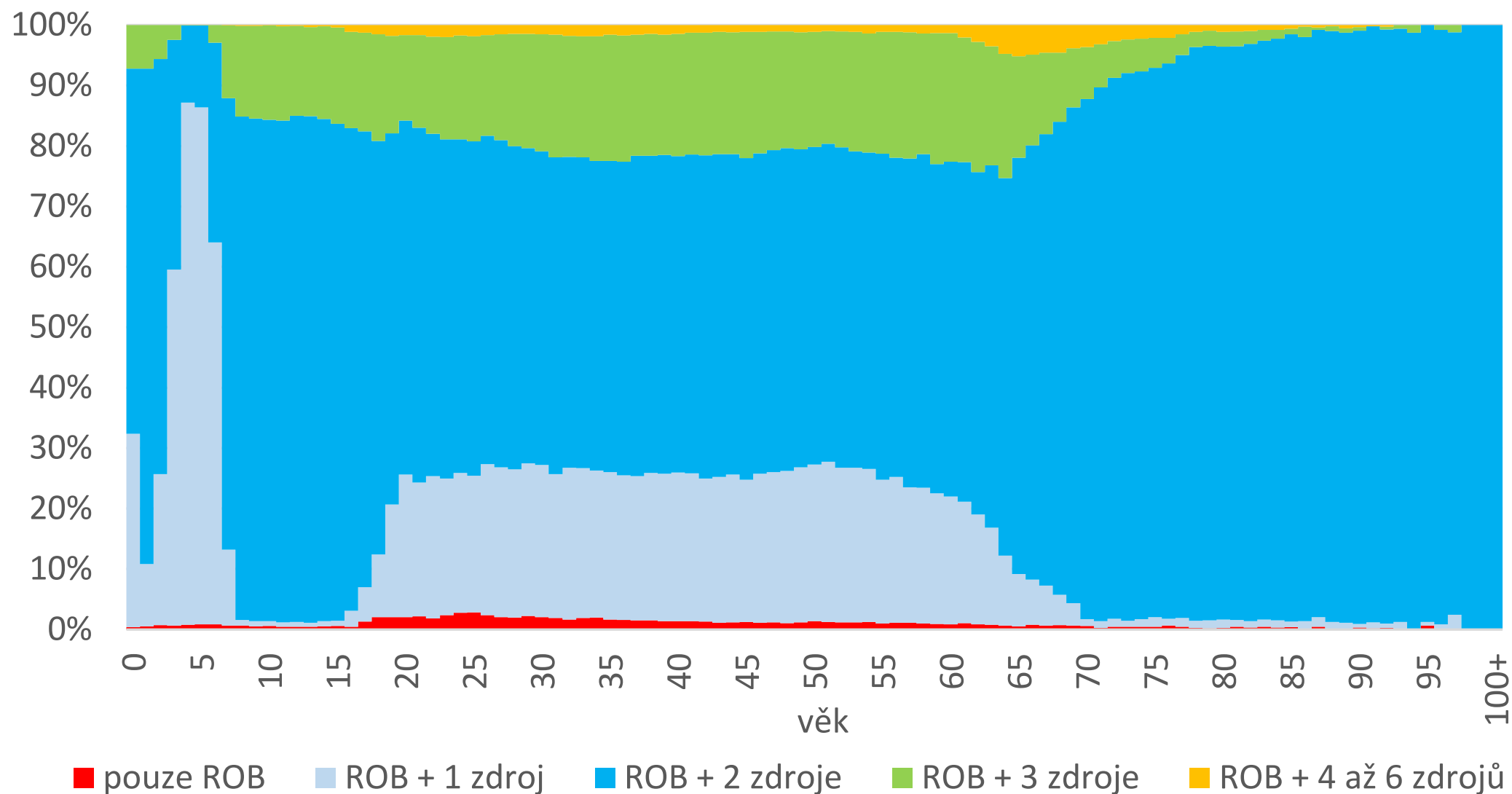
Signs-of-life analýza

- Aplikována na osoby evidované v ROB, nesečené na formulářích (rozhodnutí, zda zařadit/nezařadit mezi obvykle bydlící obyvatelstvo)
- Hlavní východisko – „chování“ osob evidovaných v ROB a zároveň sečtených na formulářích
- Nástin hlavních pravidel
- Osoba je vedle ROB vedena v některém ze zdrojů evidujících osoby po omezenou dobu a indikujících přítomnost na území ČR (evidence uchazečů o zaměstnání, pomoc v hmotné nouzi, peněžitá pomoc v mateřství)
- Osoba má platné zdravotní pojištění a zároveň
 - je evidována v některém ze zdrojů „typických“ pro jejich věkovou kategorii – např. děti 7 – 15 let v databázích MŠMT, osoby 70+ v ČSSZ-důchody, osoby 16 – 69 v kterémkoliv zdroji kromě ZŠ
 - Nebo změna v evidenci v posledním roce (sňatek na území ČR, narození dítěte na území ČR,...)
 - nebo zdravotní pojištění za osobu neplatil stát (!)
 - ...

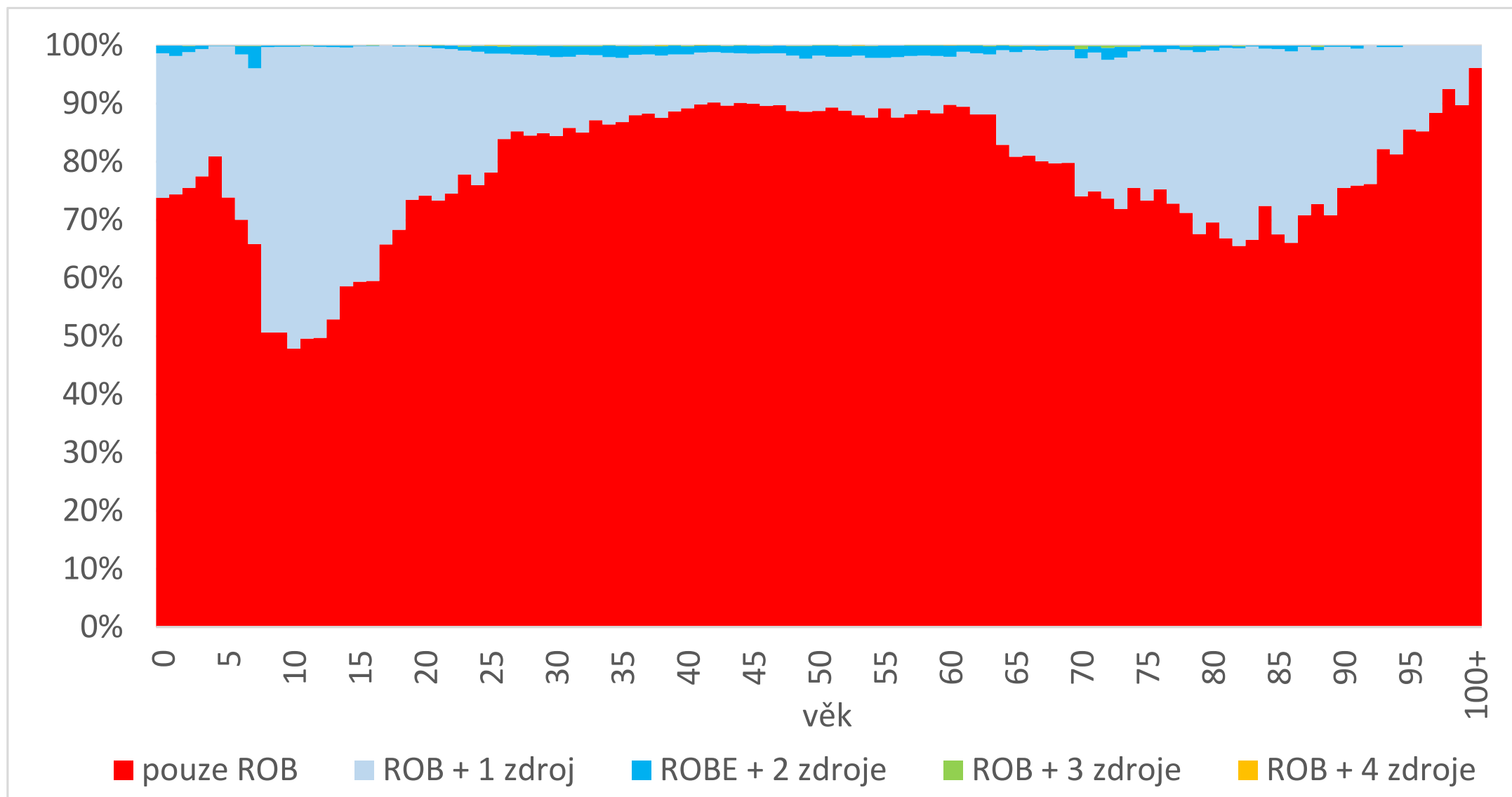
Obyvatelstvo evidované v ROB a zároveň sečtené na formulářích, podle věku a počtu doplňkových AZD



Osoby v ROB, nesečtené na formulářích, zařazené do obvykle bydlícího obyvatelstva podle věku a počtu doplňkových AZD

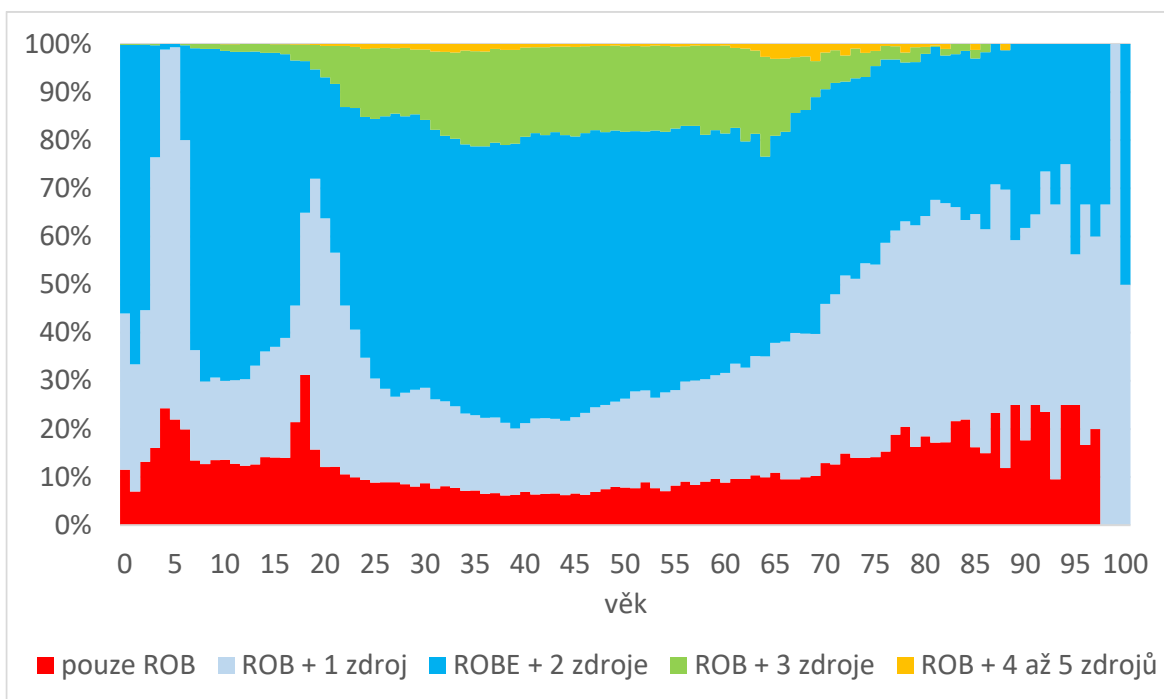


Osoby v ROB, nesečtené na formulářích, nezařazené do obvykle bydlícího obyvatelstva podle věku a počtu doplňkových AZD

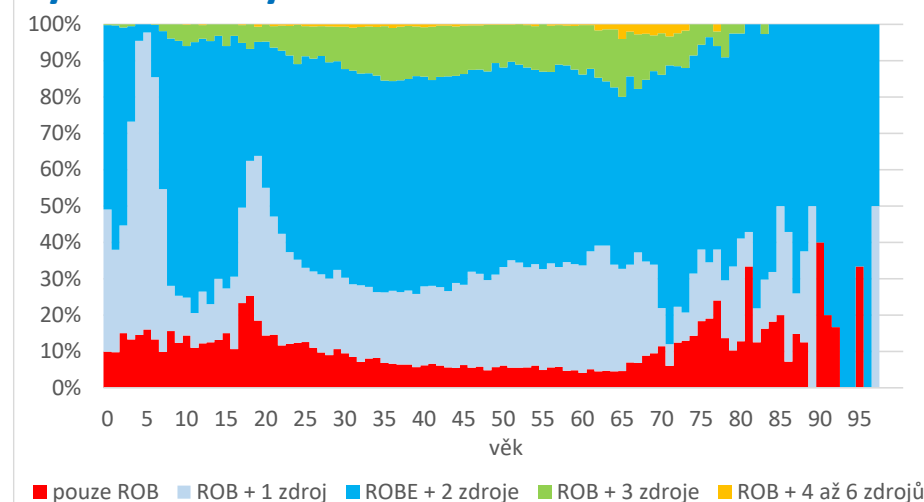


Cizinci v ROB, podle sečtenosti na formulářích a výsledku signs-of-life analýzy podle věku a počtu doplňkových AZD

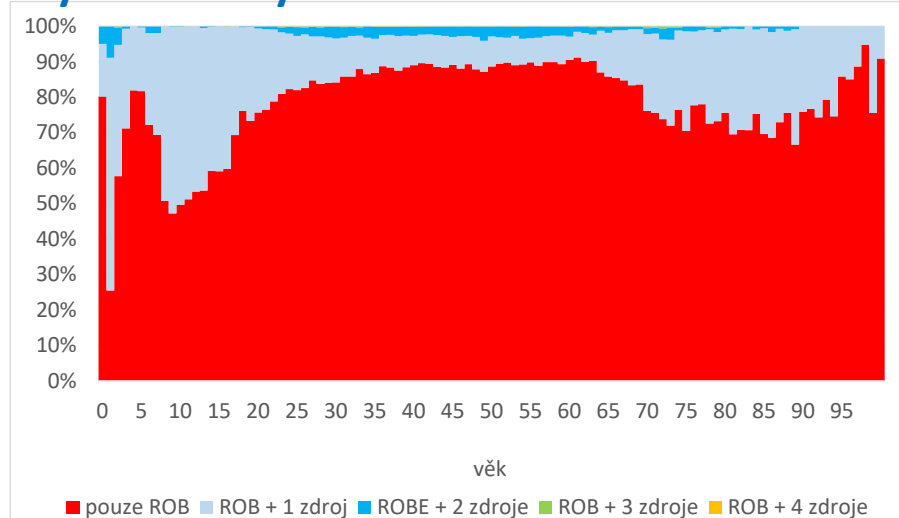
Evidování v ROB a sečtení na formulářích



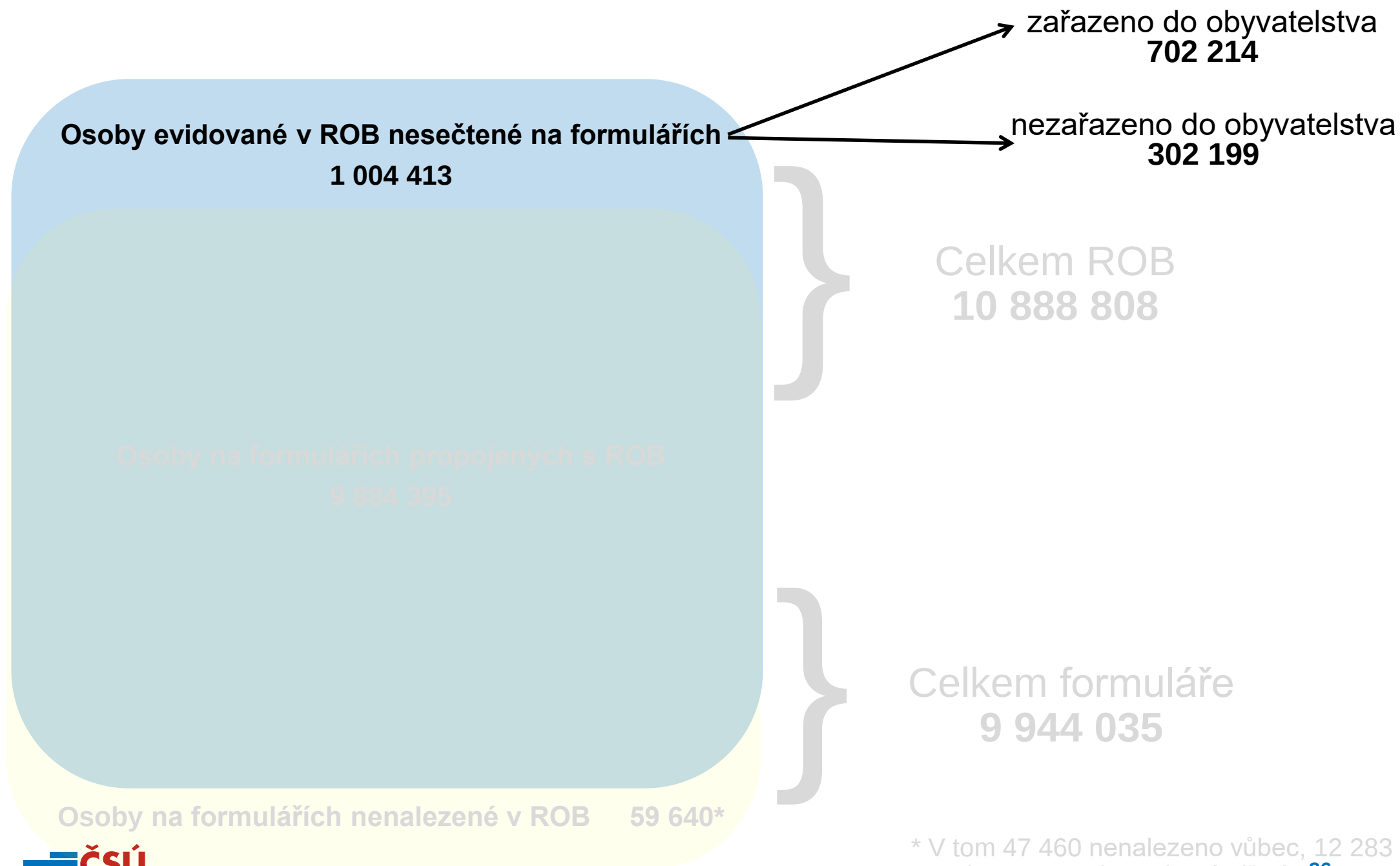
Evidování v ROB, zařazení do obvykle bydlícího obyvatelstva



Evidování v ROB, nezařazení do obvykle bydlícího obyvatelstva



Výsledek signs-of-life analýzy



Vymezení obyvatelstva v SLDB 2021

Data z registru obyvatel
propojená s dalšími administrativními zdroji

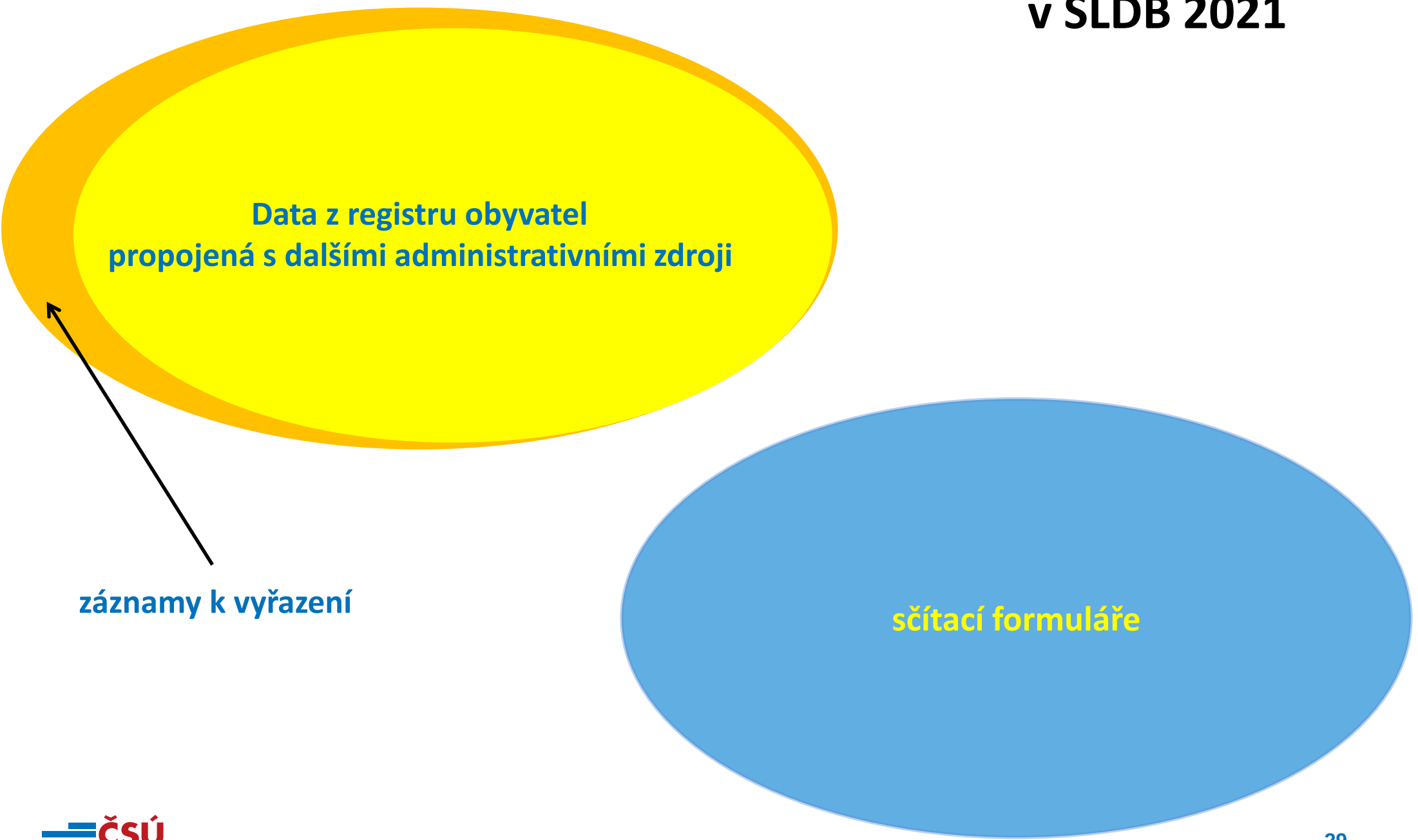
Vymezení obyvatelstva v SLDB 2021



Data z registru obyvatel
propojená s dalšími administrativními zdroji

záznamy k vyřazení

Vymezení obyvatelstva v SLDB 2021

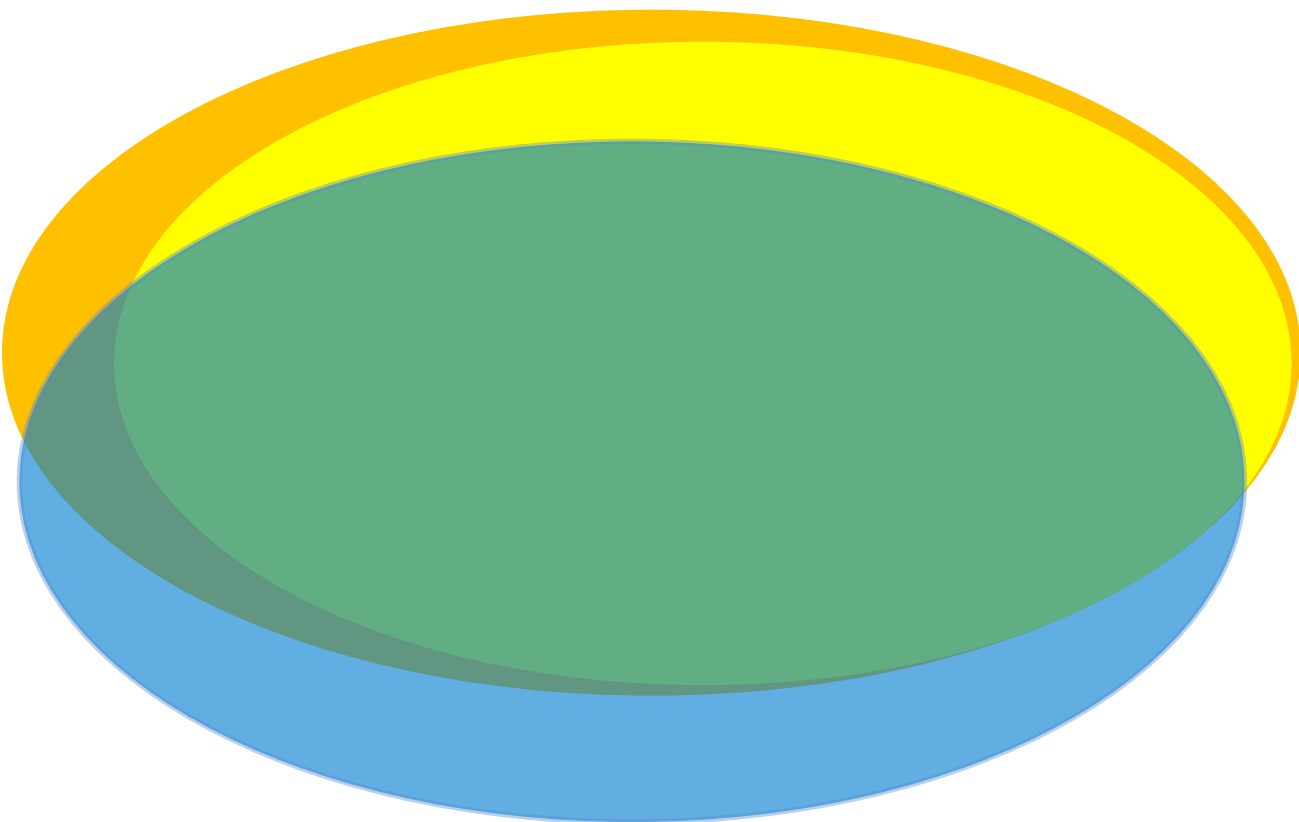


Data z registru obyvatel
propojená s dalšími administrativními zdroji

záznamy k vyřazení

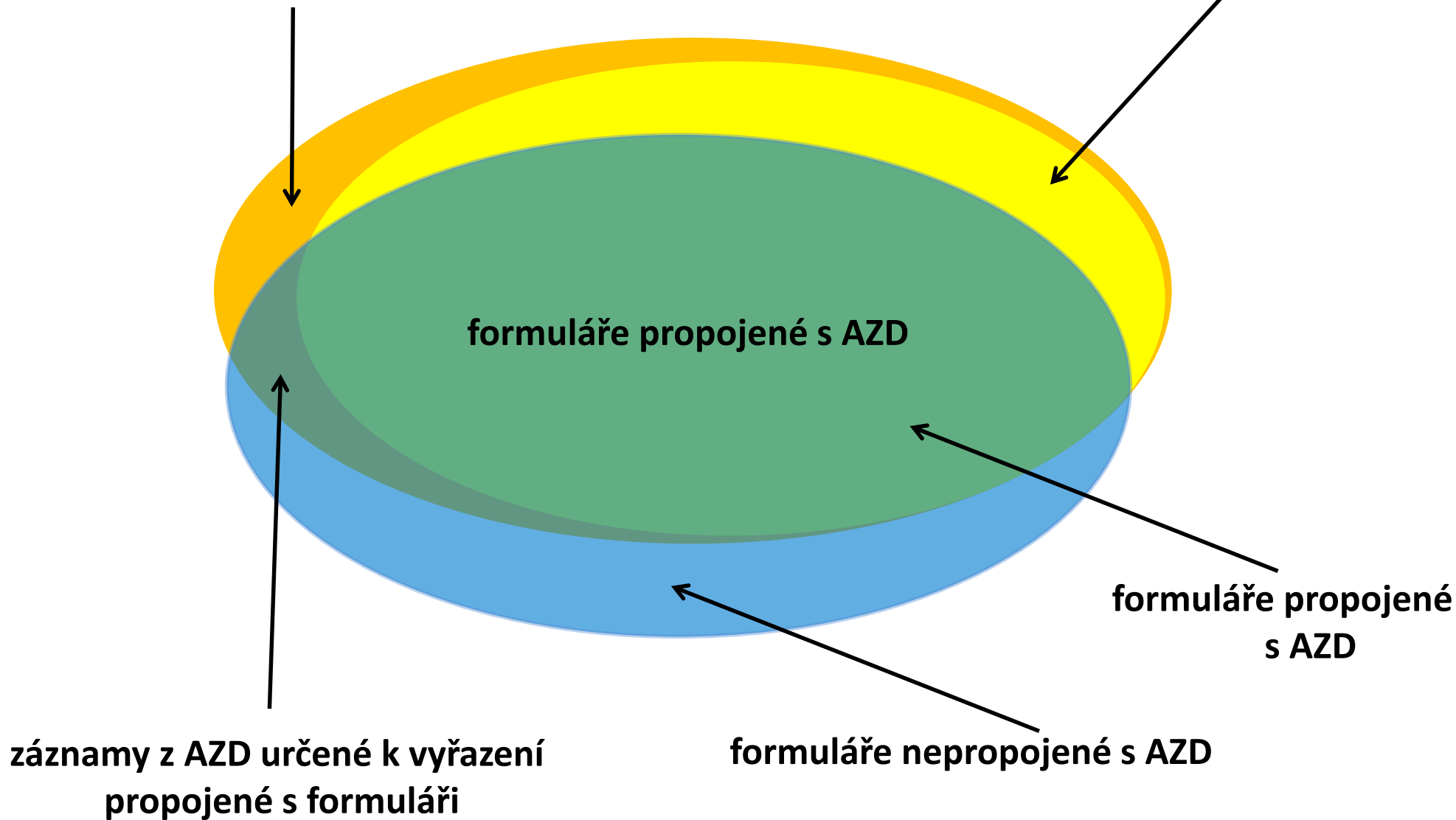
sčítací formuláře

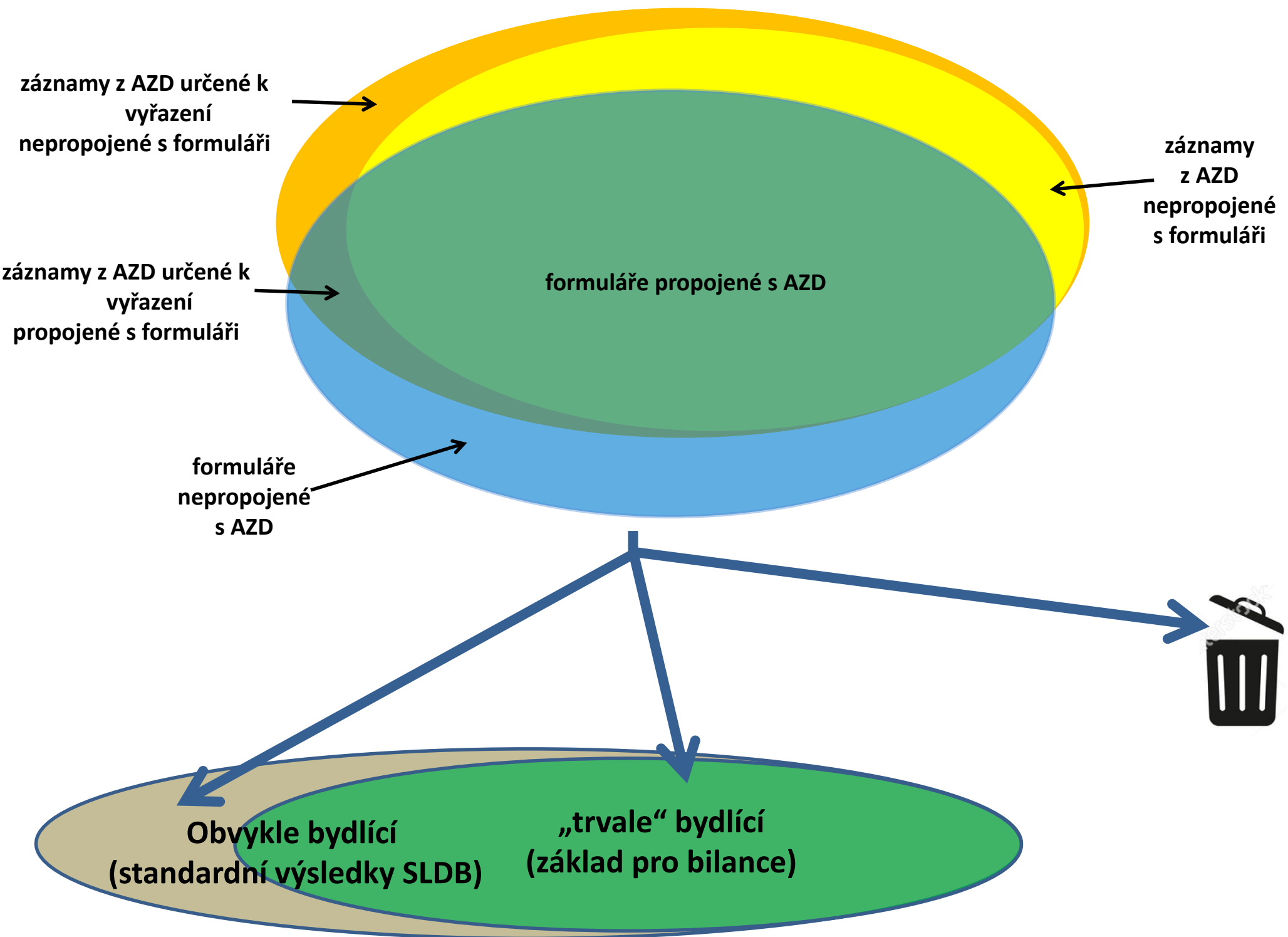
Vymezení obyvatelstva v SLDB 2021



**záznamy z administrativních zdrojů (AZD)
určené k vyřazení
nepropojené s formuláři**

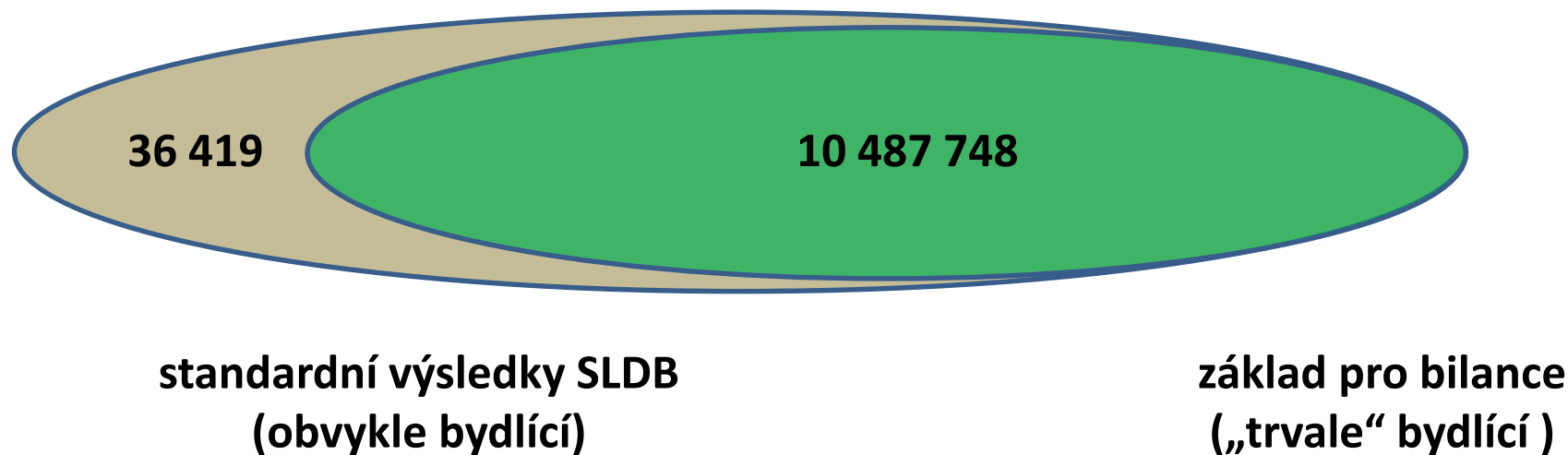
**záznamy z AZD
nepropojené s formuláři**





Obvykle a trvale bydlící obyvatelstvo podle SLDB 2021

- Počet obvykle bydlících obyvatel (OP): 10 524 167
- Počet „trvale“ bydlících obyvatel (TP): 10 487 748

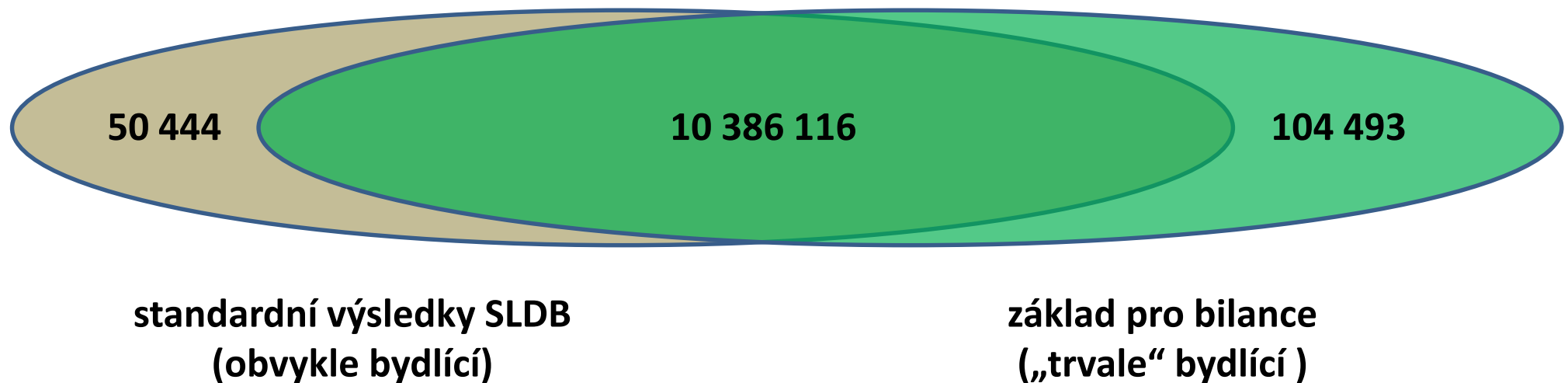


$$\text{míra shody} = \frac{OP \cap TP}{OP \cup TP} * 100 = 99,7 \%$$

na celorepublikové úrovni v roce 2021 shodné s $TP/OP * 100$

Obvykle a trvale bydlící obyvatelstvo podle SLDB 2011

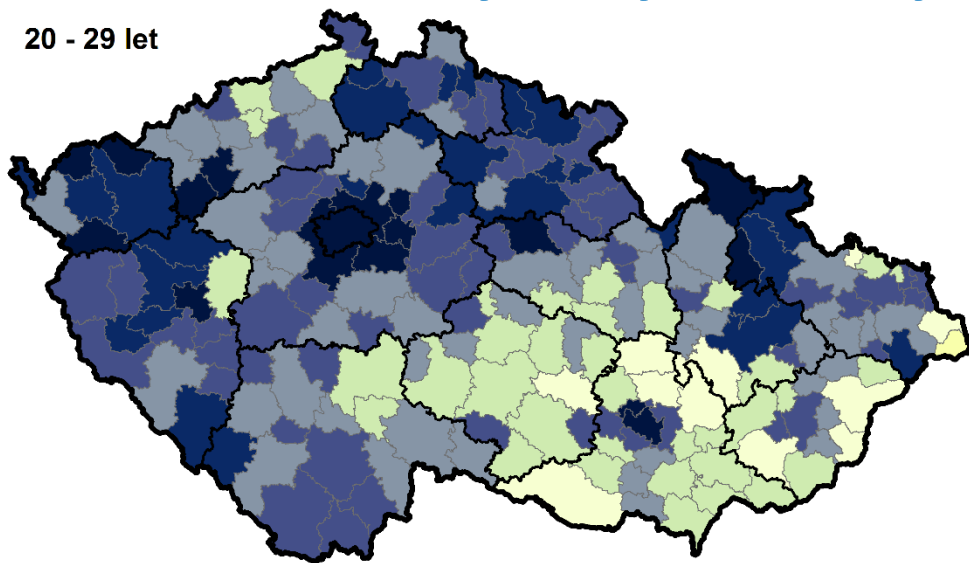
- Počet obvykle bydlících obyvatel (OP): 10 436 560
- Počet „trvale“ bydlících obyvatel (TP): 10 490 609



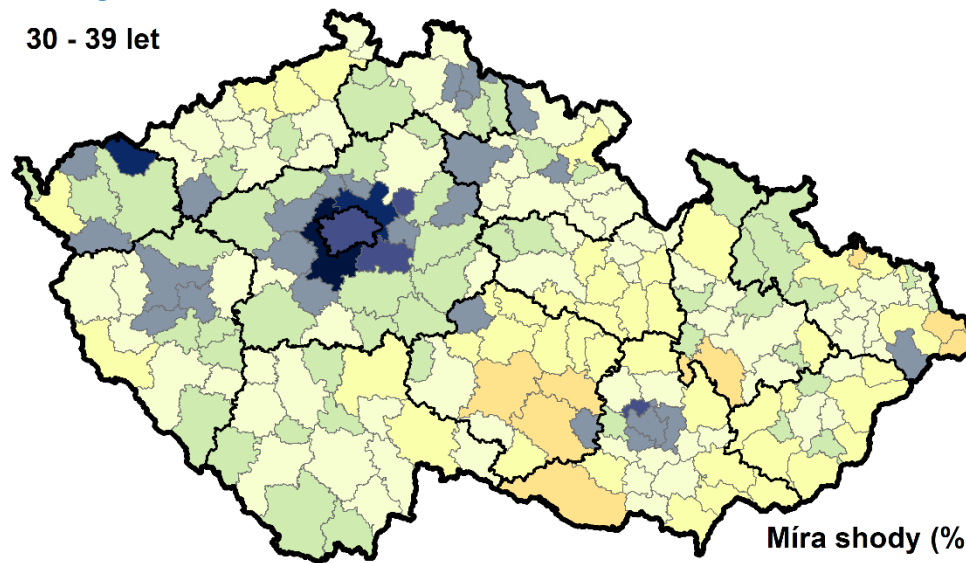
$$\text{míra shody} = \frac{OP \cap TP}{OP \cup TP} * 100 = 98,5 \%$$

Míra shody (rovnosti množin) obvykle a trvale bydlicího obyvatelstva na úrovni ORP ve vybraných věkových skupinách v roce 2011

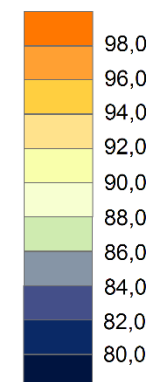
20 - 29 let



30 - 39 let

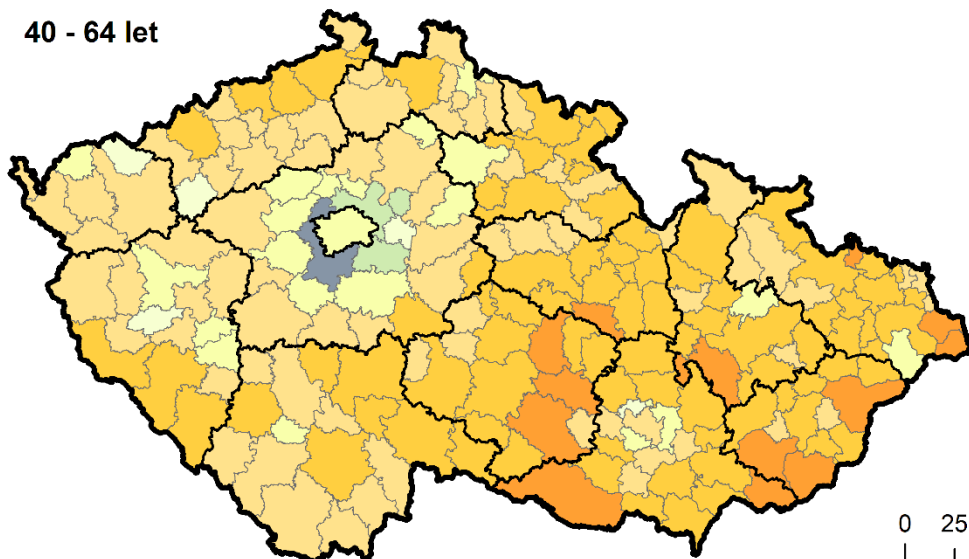


Míra shody (%)

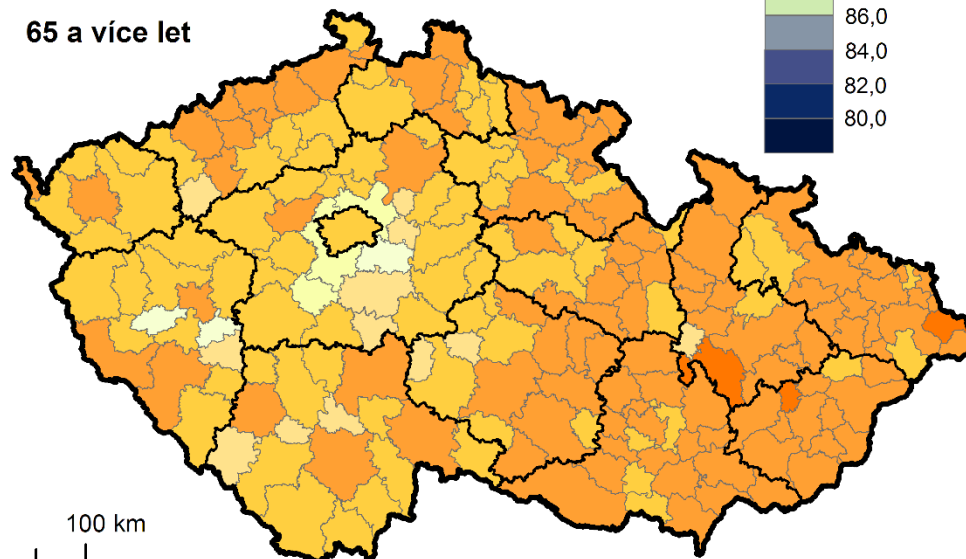


— státní hranice
— hranice kraje
— hranice obvodu ORP

40 - 64 let

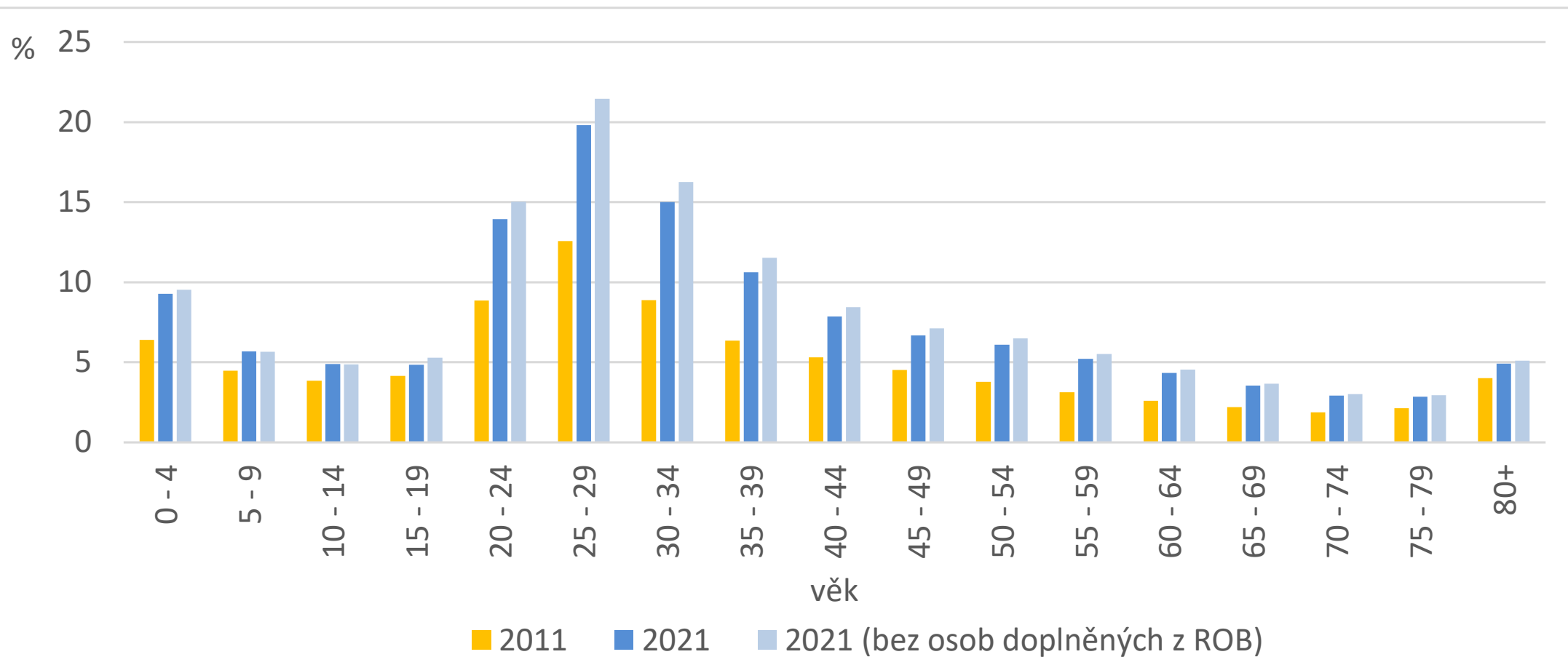


65 a více let



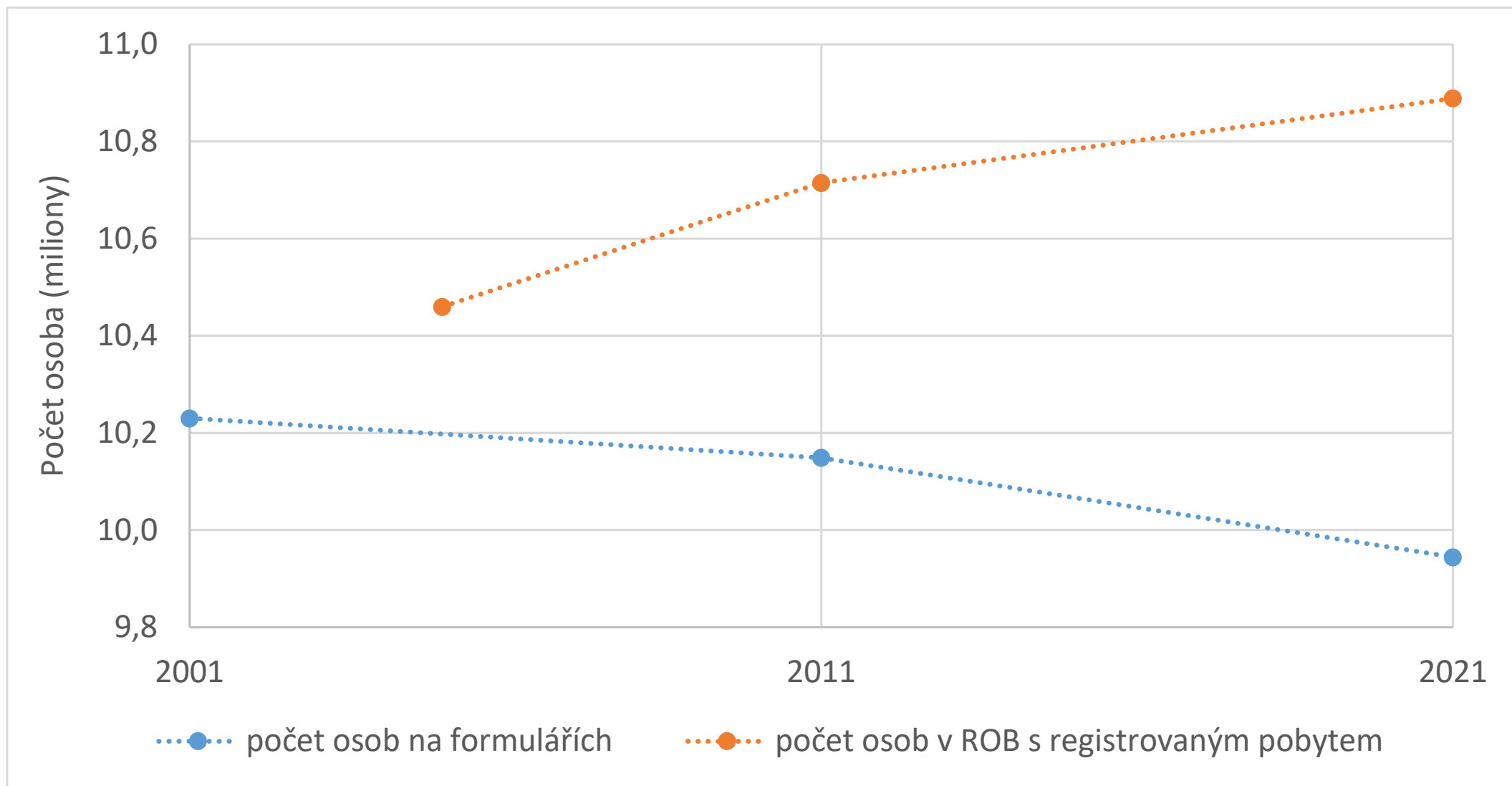
0 25 50 100 km

Podíly obyvatel obvykle i trvale bydlících na území Česka s rozdílnými obcemi obvyklého a trvalého pobytu podle věku



100 % - počet obyvatel v dané věkové skupině s obvyklým i trvalým pobytem v Česku

Počty osob v evidenci obyvatel (ISEO, resp. ROB) a počty osob na sčítacích formulářích v období 2001 – 2021



Poznámka: Údaj z roku 2001 nevyjadřuje přímo počet osob sečtených na formulářích, ale výsledný počet („trvale“ bydlících) obyvatel.

DĚKUJI ZA POZORNOST.

Robert Šanda
robert.sanda@czso.cz