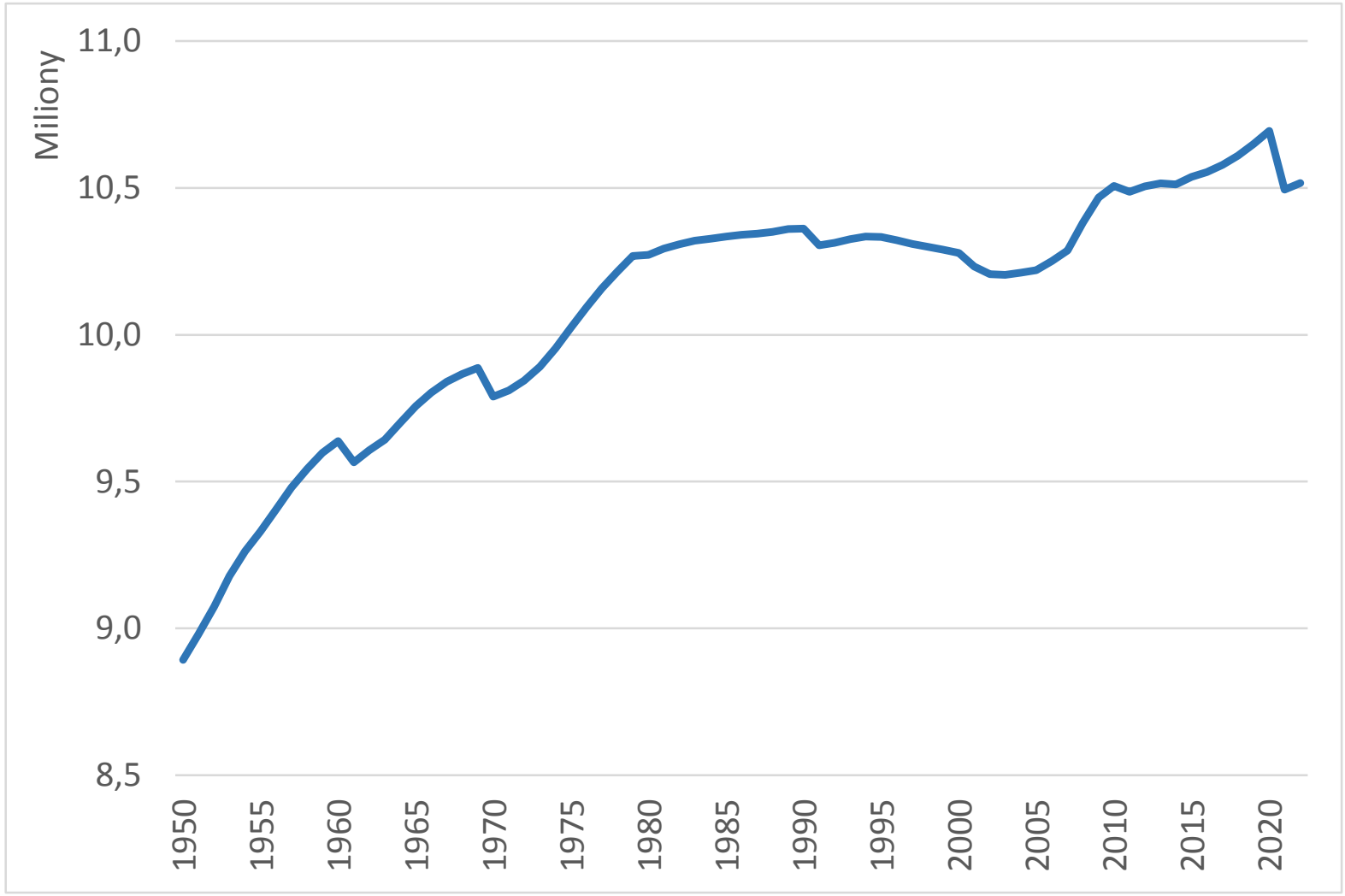
Sčítání lidu 2021

Zdroje dat a hlavní otázky zpracování výsledků

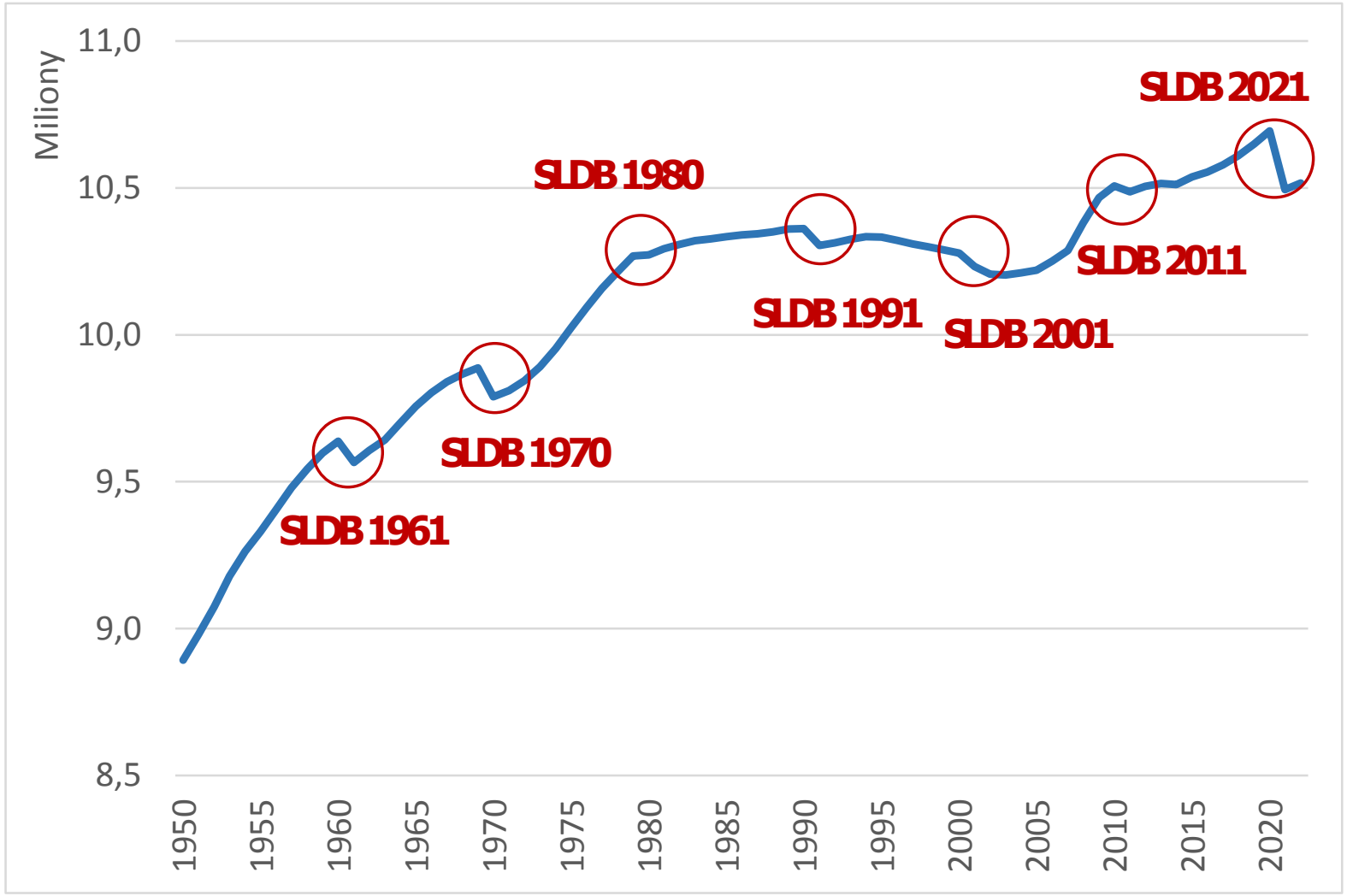
Robert Šanda

20. 4. 2023

Počet obyvatel Česka v období 1950 – 2022 (stavy k 1. 1.)



Počet obyvatel Česka v období 1950 – 2022 (stavy k 1. 1.)



Od dotazníků k registrům

Formy sčítání lidu v Evropě

Formy sčítání podle nařízení EU:

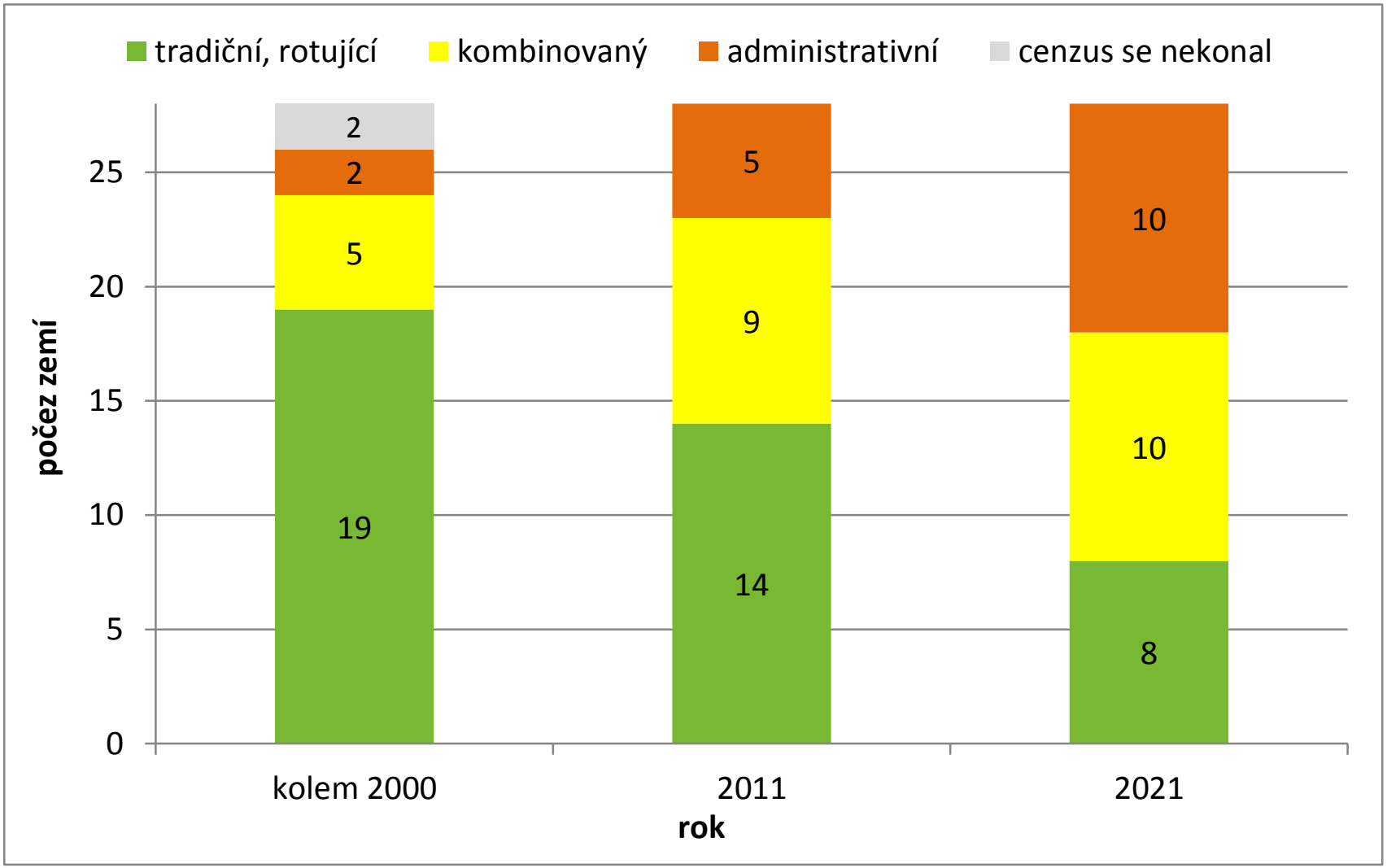
- tradiční dotazníkové šetření
- administrativní sčítání (sčítání založené pouze na registrech)
- kombinace tradičního sčítání s administrativním
- kombinace s výběrovými šetřeními
- sčítání založené na rotujících výběrech

Dotazníkové šetření

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

=> značná různorodost zdrojů

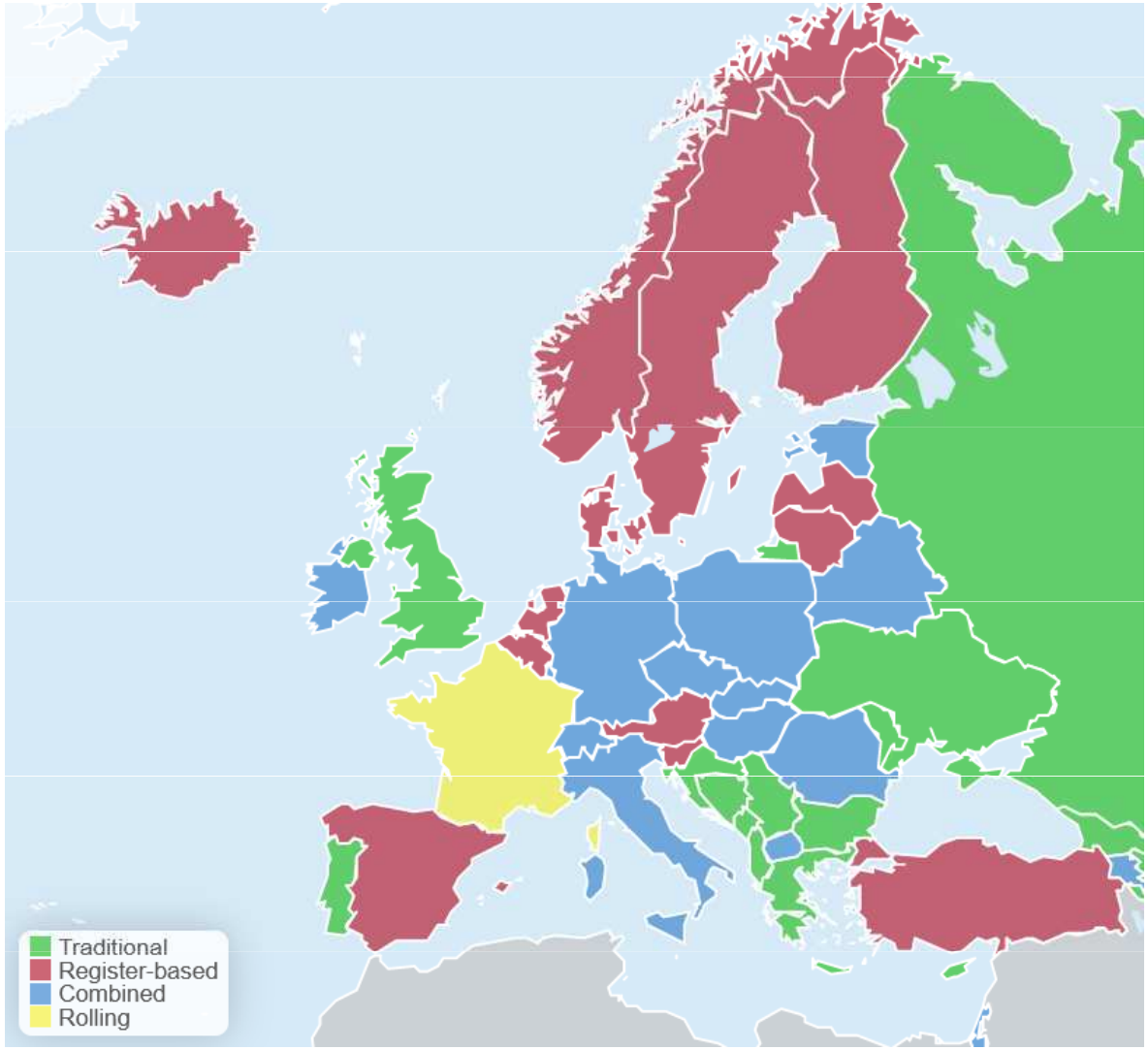
Formy sčítání členských zemích v EU v období 2000 – 2021 (včetně UK)



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

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

Základní formy sčítání kolem roku 2020 v Evropě



Převzato z: https://statswiki.unece.org/display/censuses/Censuses+of+the+2020+round

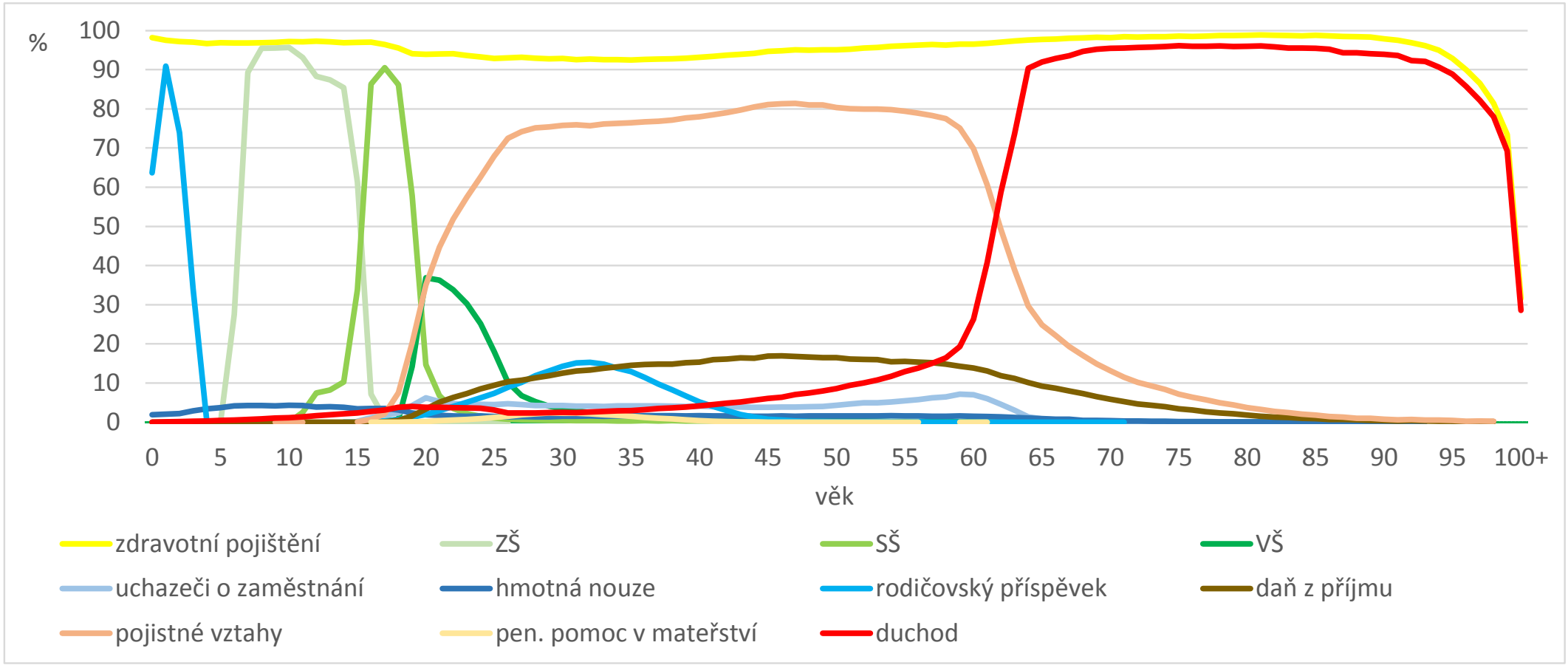
SLDB 2021 – ZDROJE DAT

Zdroje dat o osobách

- **Konstitutivní zdroje**
  - sčítací formuláře (listinné, elektronické)
  - základní registr obyvatel (ROB) + jeho agendové (zdrojové) systémy AISEO a AISC

- **Doplňkové administrativní zdroje**
  - centrální registr pojištěnců (zdravotní pojištění)
  - integrovaný IS České správy sociálního zabezpečení
  - 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

Podíly osob evidovaných v ROB nalezených ve vybraných dalších zdrojích podle věku (100 % - počet osob v ROB v daném věku)



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 administrativní zdroje – 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 neobsažené v admin. zdrojích

- **Údaje o bytech**
- **Vazba obyvatel na byt (adresa pobytu do úrovně bytu – pro tvorbu domácností a charakteristiky bydlení)**
- **Úroveň vzdělání za většinu obyvatel**
- **Pobyt po narození (dříve bydliště matky v době narození) za podstatnou část obyvatel**
- **Místo pracoviště/školy**
- **Charakteristiky dojížďky do zaměstnání/školy (frekvence, doba každodenní dojížďky, prostředek)**
- **Zaměstnání**
- **Sociokulturní charakteristiky**
  - mateřský jazyk
  - národnost
  - náboženská víra

ZPRACOVÁNÍ

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 – propojování záznamů

- **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**

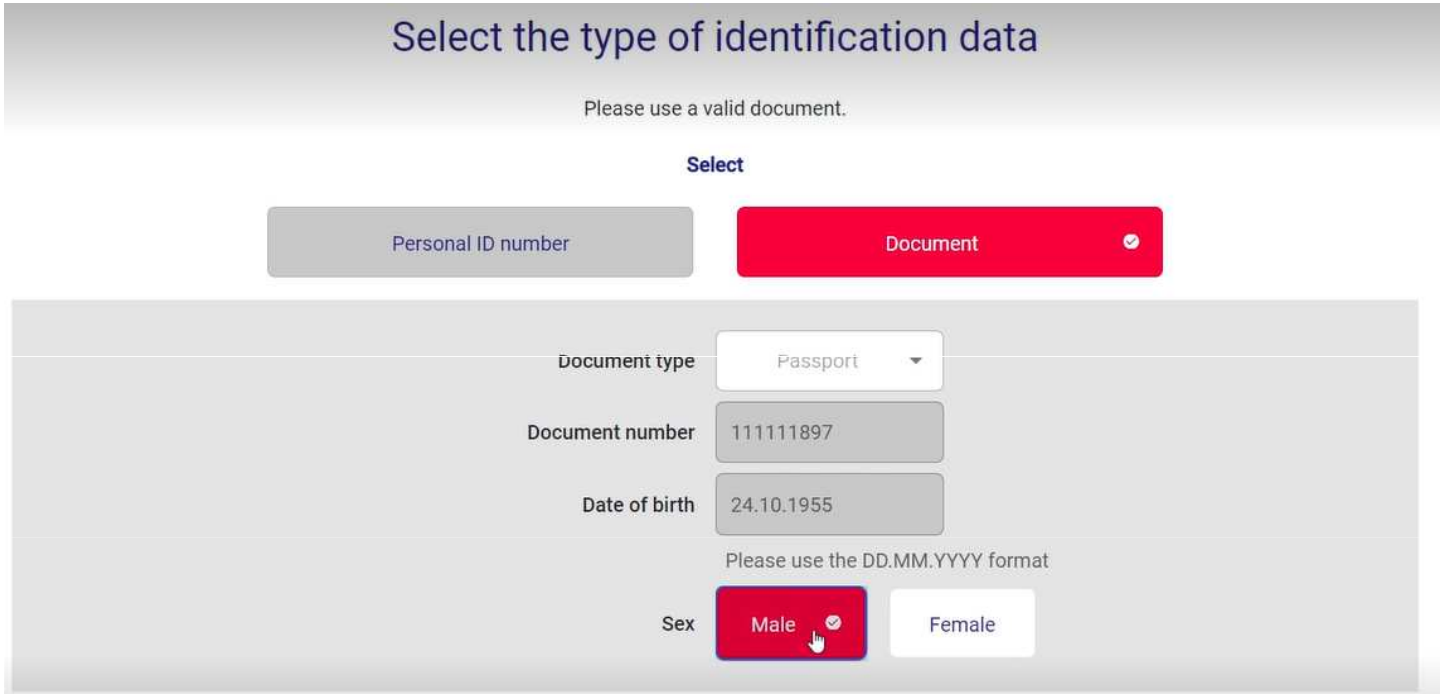
      příjmení: **Horáková**, Ph.D.
    - standardizace 1: jméno: **ANNA MARIE**

      příjmení: **HORAK**
    - standardizace 2: příjmeníjméno: **HORAKANNAMARIE**
    - standardizace 3: celé jméno abecedně : **ANNA~HORAKOVA~MARIE**

Vybrané etapy zpracování – propojování záznamů

- **B) Vytvoření „black listů“**
  - neunikátní kombinace jméno-příjmení-datum narození v ROB
  - neunikátní rodná čísla v ROB
  - „půjčované“ doklady na formulářích

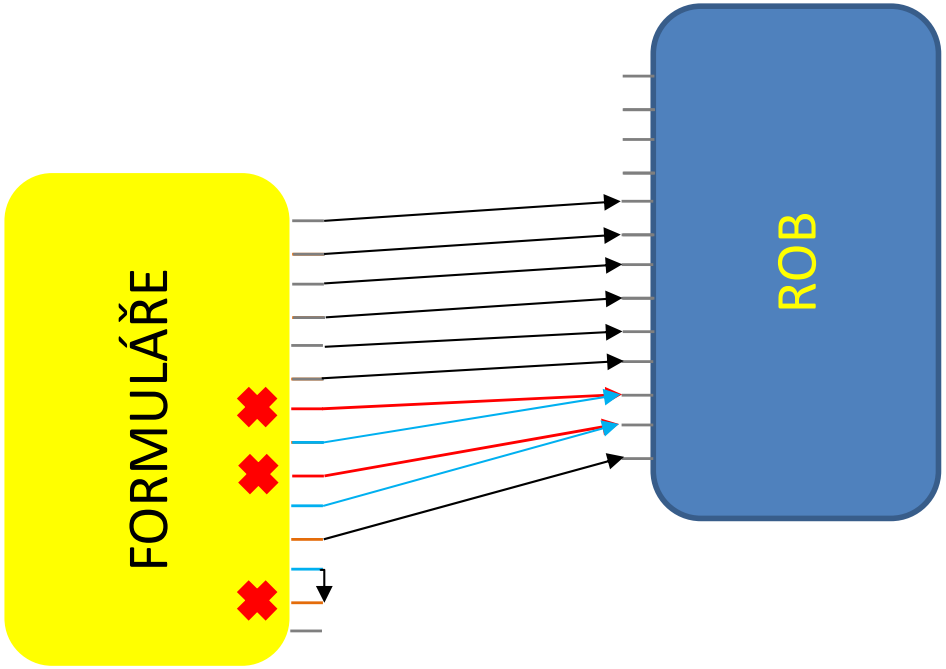
  …



Vybrané etapy zpracování – propojování záznamů

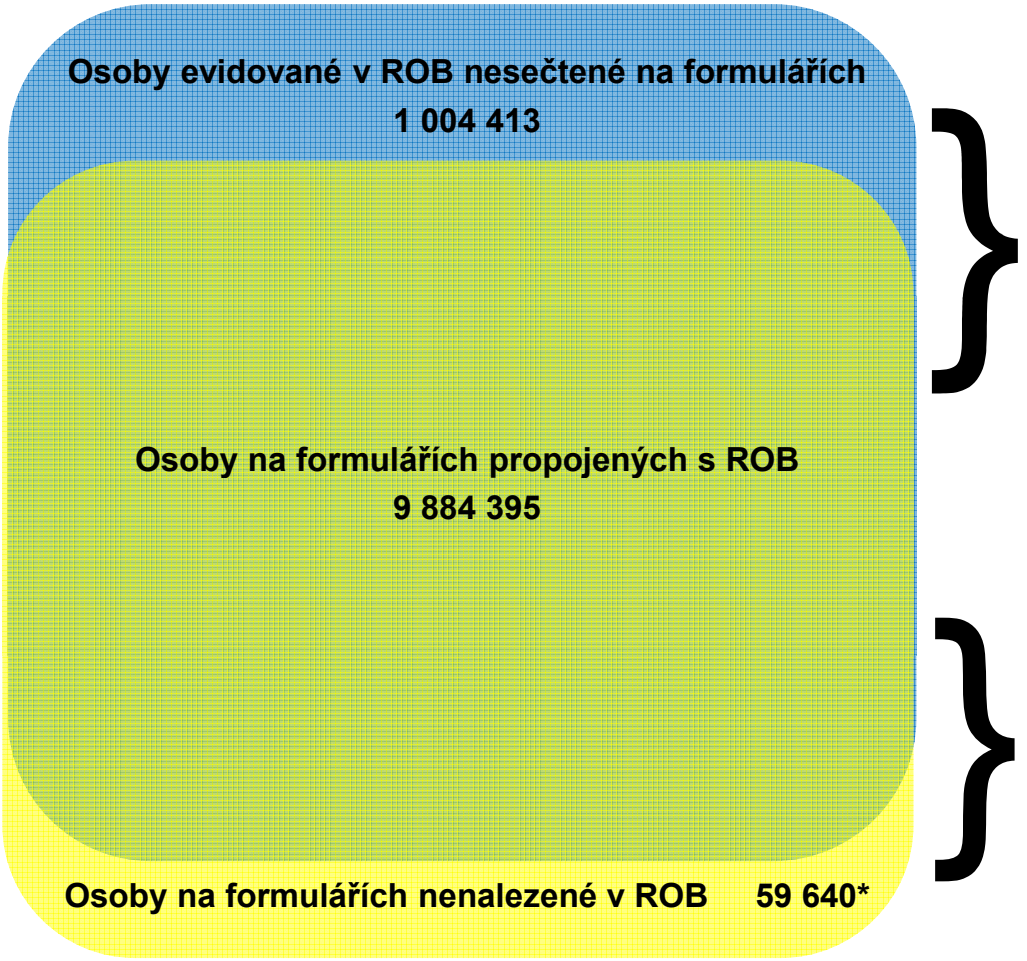
- **C) Vlastní propojování**
  - série zhruba 20 pravidel
  - deterministické propojování
    - hierarchie identifikátorů podle spolehlivosti (id datové schránky…jméno…doklad
  - pravděpodobnostní propojování
    - **Levenshteinova vzdálenost** (počet rozdílů ve znacích)
    - **Jaro-Winkler** (počet shodných znaků v polovině řetězce, porovnání jejich pozic, bonifikace shodných začátků řetězců)
    - **„symetrická diference“** – počet shodných a počet rozdílných slov

Vybrané etapy zpracování – prioritizace, deduplikace



- duplicita (multiplicita): více formulářů napojených na jeden záznam ROB
- duplicity poprvé řešeny v roce 2001
- poměrně velké množství duplicit (861 tisíc osob na více než jednom formuláři)
- stanovena sada pravidel pro výběr prioritního záznamu z formuláře

Výsledek napojování formulářů na ROB a deduplikace

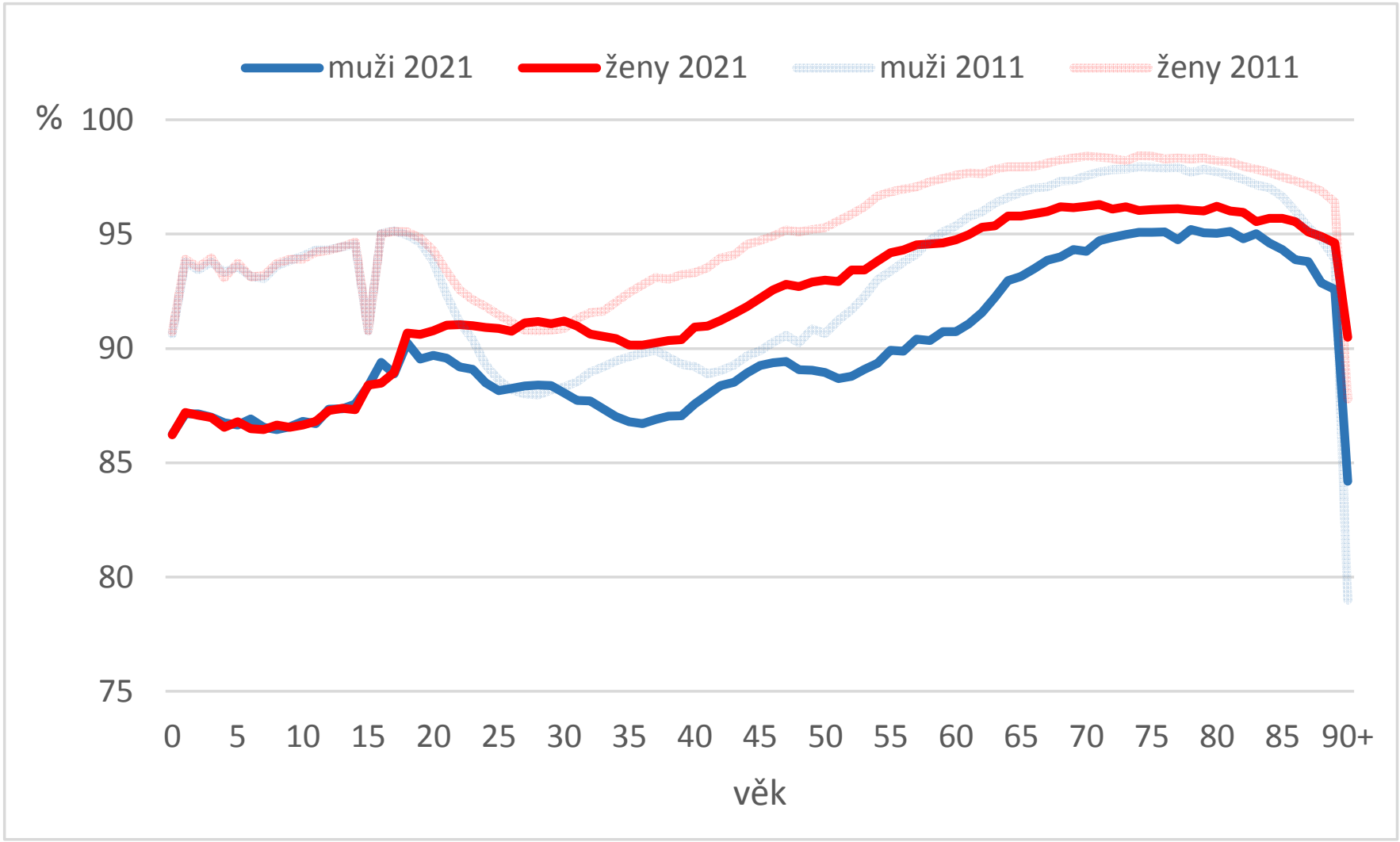


Celkem ROB
10 888 808

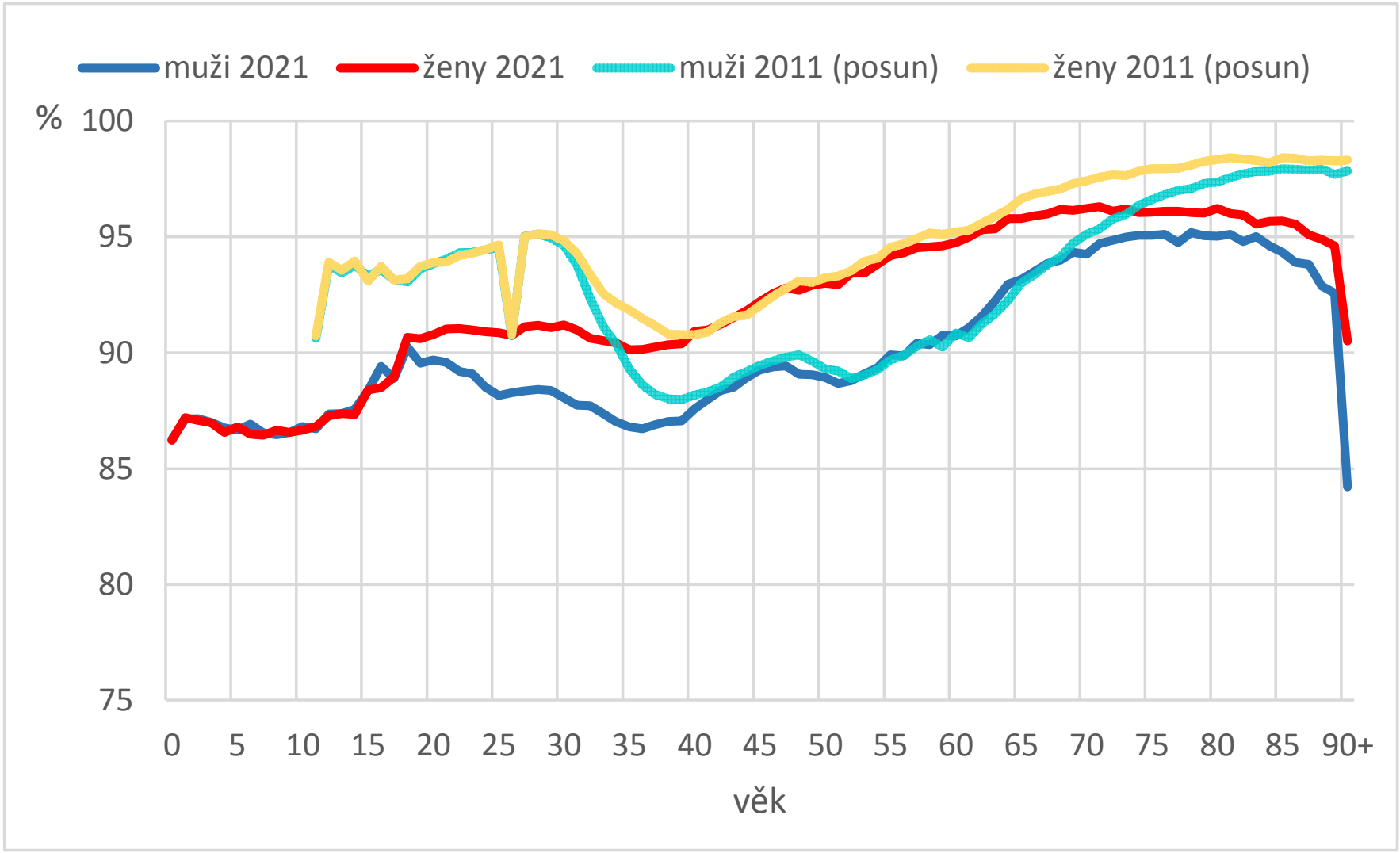
Celkem formuláře
9 944 035

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

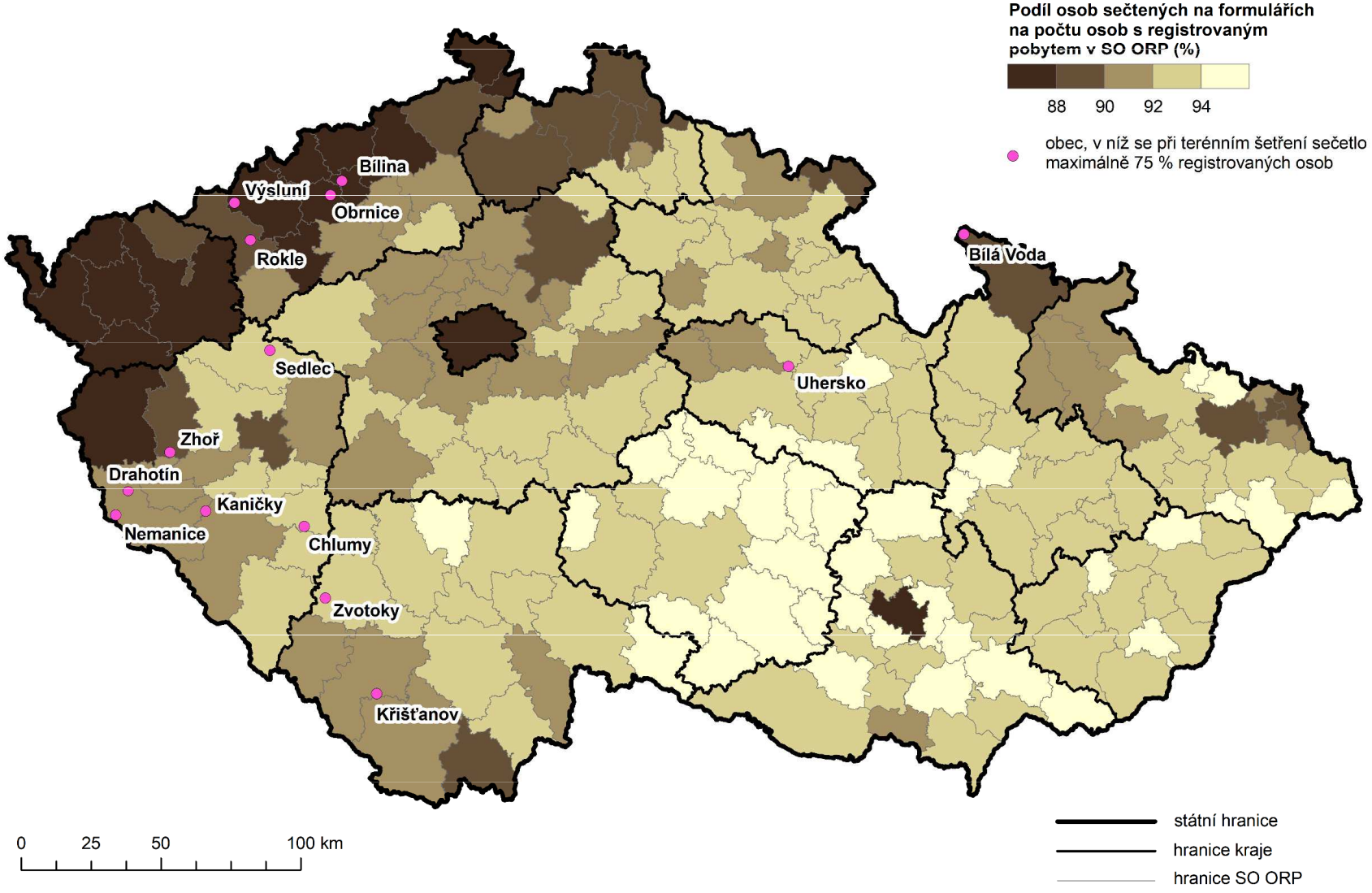
Podíly osob sečtených na sčítacích formulářích na osobách s evidovaným pobytem v registru obyvatel podle věku a pohlaví v letech 2011 a 2021



Podíly osob sečtených na sčítacích formulářích na osobách s evidovaným pobytem v registru obyvatel podle věku a pohlaví v letech 2011 a 2021



Pokrytí osob vedených v registru obyvatel terénním šetřením SLDB 2021

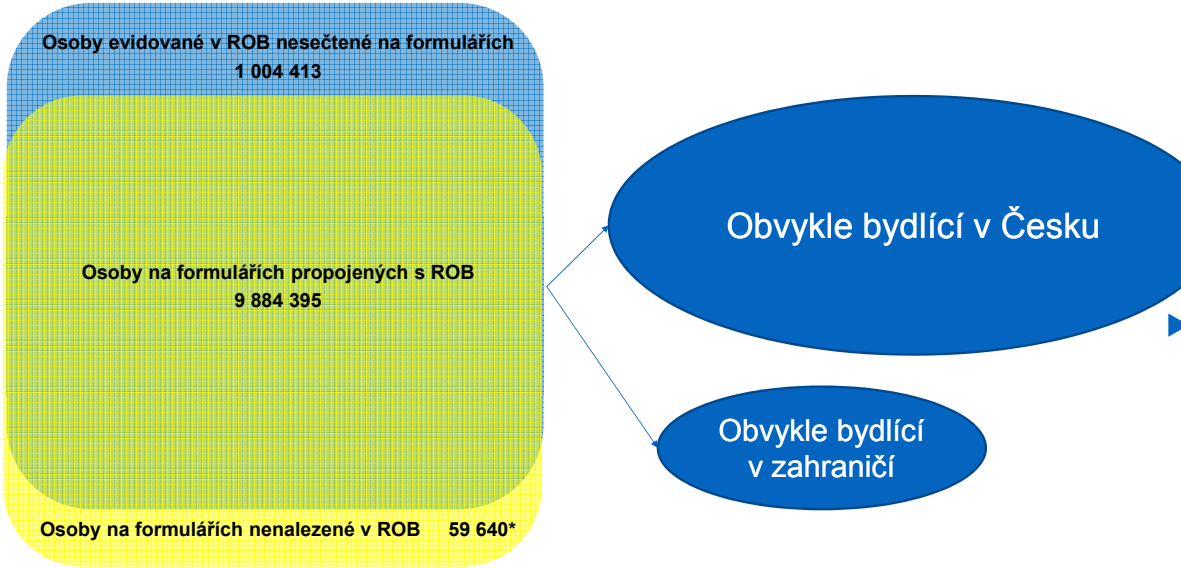


Signs of life analýza (SOL)

- 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 v SLDB 2021

- **Založena na „chování“ referenční populace v administrativních zdrojích**

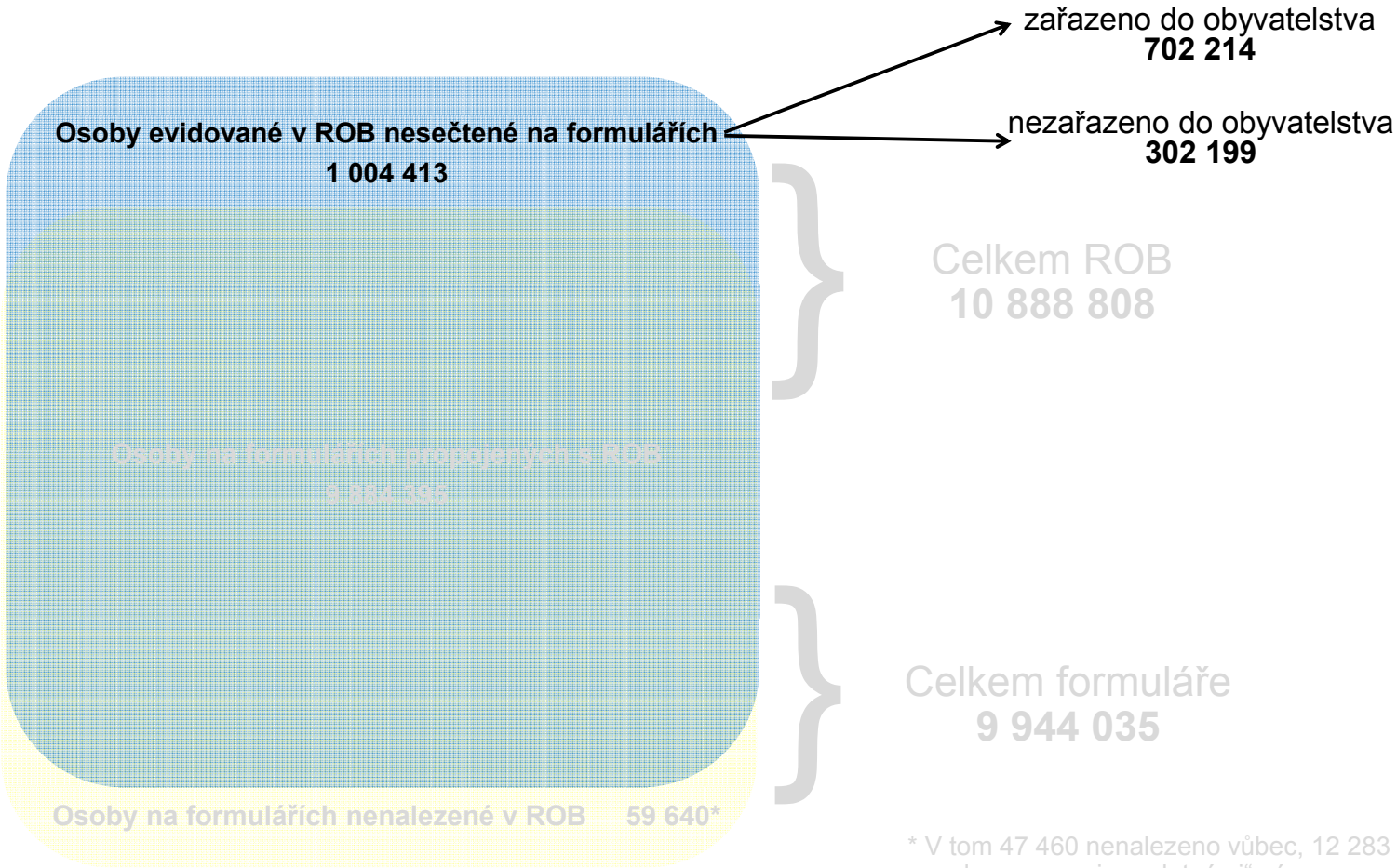


Referenční populace pro SOL

- **Z referenční populace vyplynulo např.**
  - Občané ČR téměř vždy evidování v registru pojištěnců
  - Osoby 70+ navíc téměř vždy v evidenci příjemců důchodu
  - Děti 7 – 15 téměř vždy v databázích MŠMT…

- **Navíc pravidla založená na „selském rozumu“, např.**
  - Osoby vedené Úřadem práce jako uchazeči o zaměstnání byly vždy považovány za přítomné v Česku
  - Bez ohledu na výskyt v jiných zdrojích nyla nedávná změna záznamu v ROB považována za důkaz přítomnosti na území Česka
  - …

Výsledek signs-of-life analýzy



zařazeno do obyvatelstva
702 214

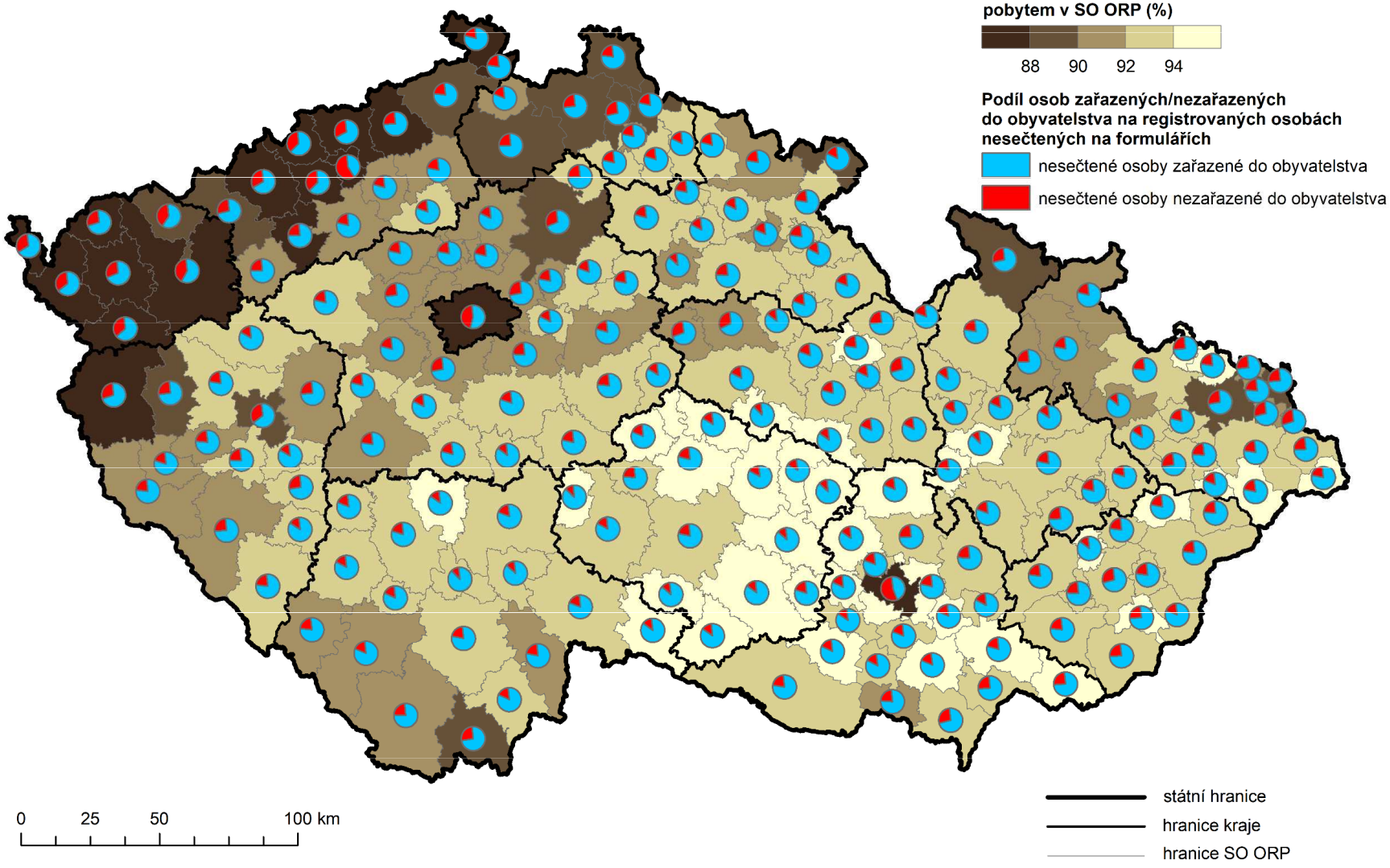
nezařazeno do obyvatelstva
302 199

Celkem ROB
10 888 808

Celkem formuláře
9 944 035

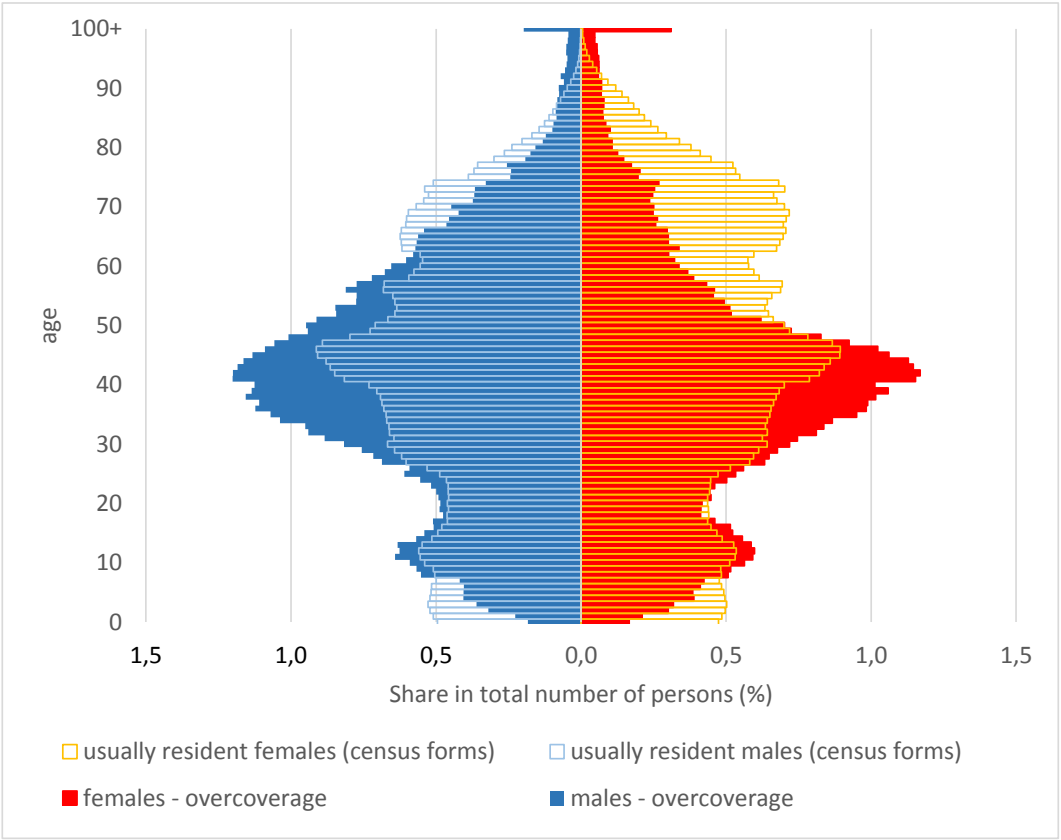
* V tom 47 460 nenalezeno vůbec, 12 283 nalezeno mezi „neplatnými“ záznamy

Pokrytí osob vedených v registru obyvatel terénním šetřením SLDB 2021 a zahrnutí nesečtených osob do obyvatelstva

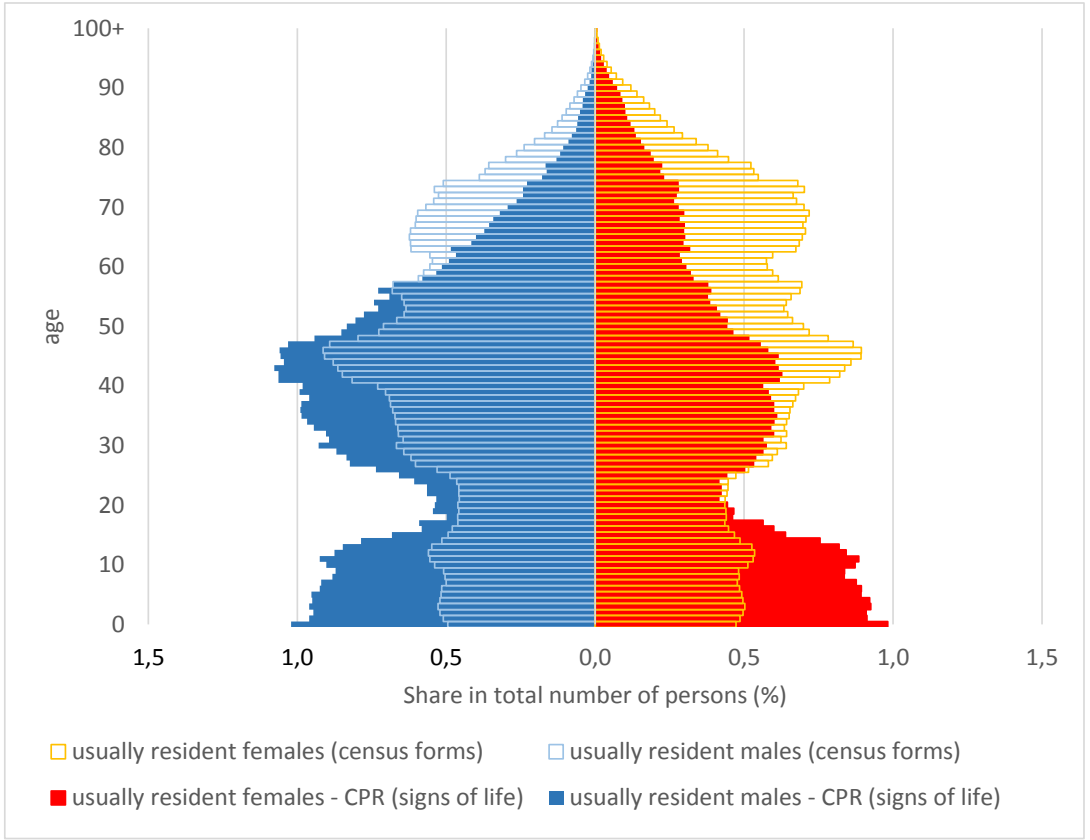


Výsledek signs-of-life analýzy

Vyřazení - 302 199 osob



Zařazení do obyvatelstva - 702 214 osob



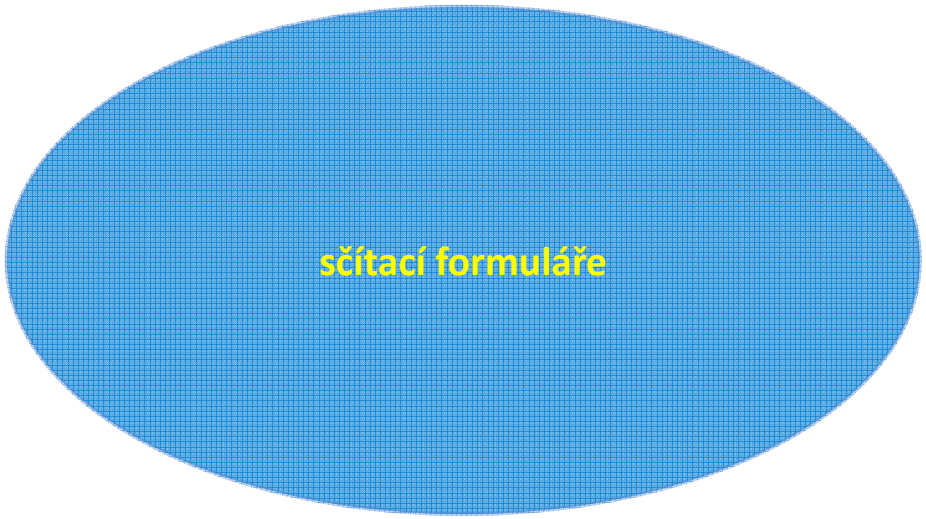
Výsledek signs-of-life analýzy

státní občanství	registrovaní v ROB	registrovaní, sečtení na formuláři		registrovaní, nesečtení na formuláři		podíl sečtených na fromuláři na registrovaných	podíl nesečtených na formuláři na výsledném obyvatelstvu	podíl vyřazených na registrovaných
		zařazení do obyvatelstva	nezařazení do obyvatelstva	zařazení do obyvatelstva	nezařazení do obyvatelstva			
ČR	10 249 602	9 393 679	87 856	603 452	164 615	92.5	6.0	2.5
cizinci celkem	639 206	391 855	11 005	98 762	137 584	63.0	20.1	23.2
Německo	20 764	4 217	265	1 331	14 951	21.6	24.0	73.3
Polsko	20 563	11 226	180	3 095	6 062	55.5	21.6	30.4
Rusko	42 229	31 243	708	4 349	5 929	75.7	12.2	15.7
Slovensko	123 877	73 148	918	20 775	29 036	59.8	22.1	24.2
Vietnam	63 841	48 770	2 160	5 055	7 856	79.8	9.4	15.7
Ukrajina	168 554	113 112	2 793	35 679	16 970	68.8	24.0	11.7
ostatní a nezj.	199 378	110 139	3 981	28 478	56 780	57.2	20.5	30.5
celkem	**10 888 808**	**9 785 534**	**98 861**	**702 214**	**302 199**	**90.8**	**6.7**	**3.7**
podíl občanů ČR	*94.1*	*96.0*	*88.9*	*54.5*	*85.9*	*x*	*x*	*x*

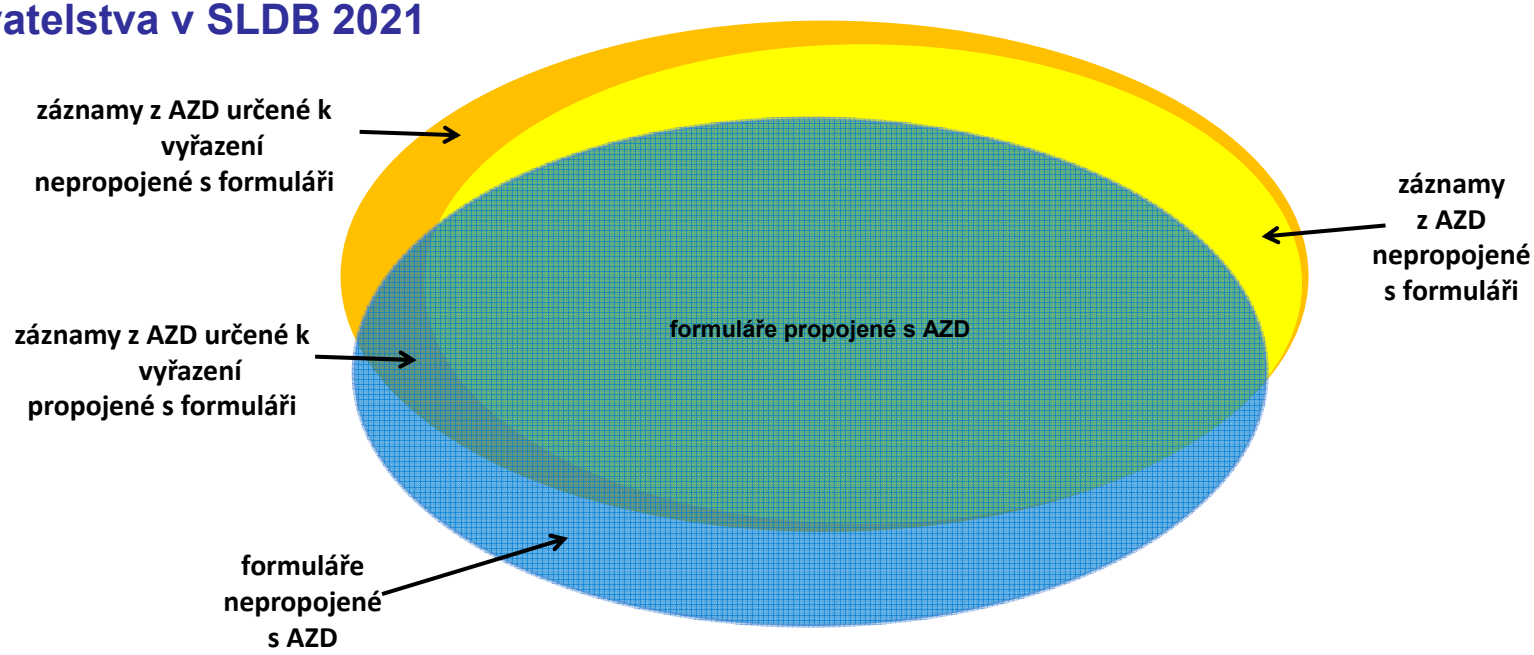
Vymezení obyvatelstva v SLDB 2021

Vymezení obyvatelstva v SLDB 2021

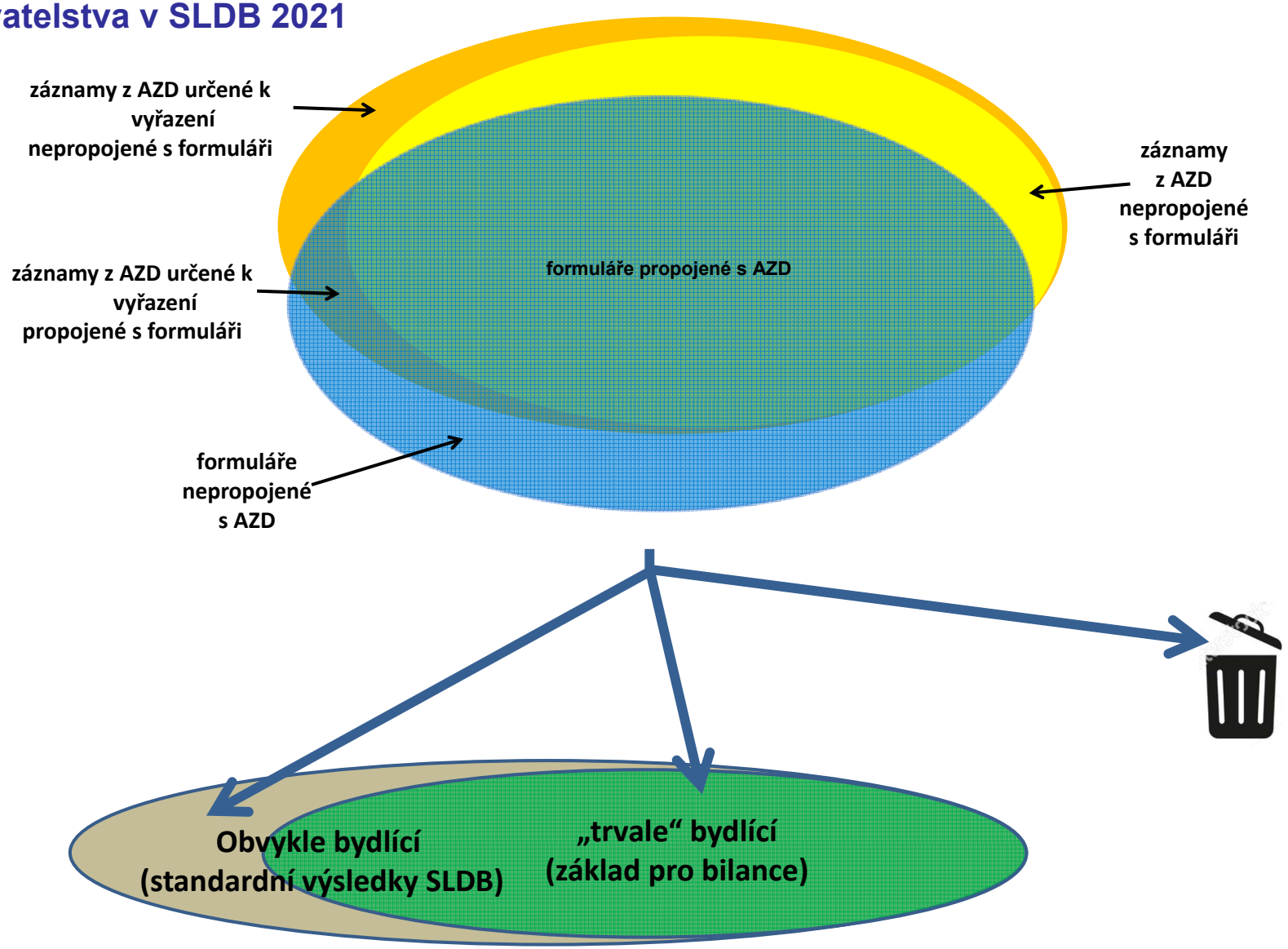
Vymezení obyvatelstva v SLDB 2021



Vymezení obyvatelstva v SLDB 2021

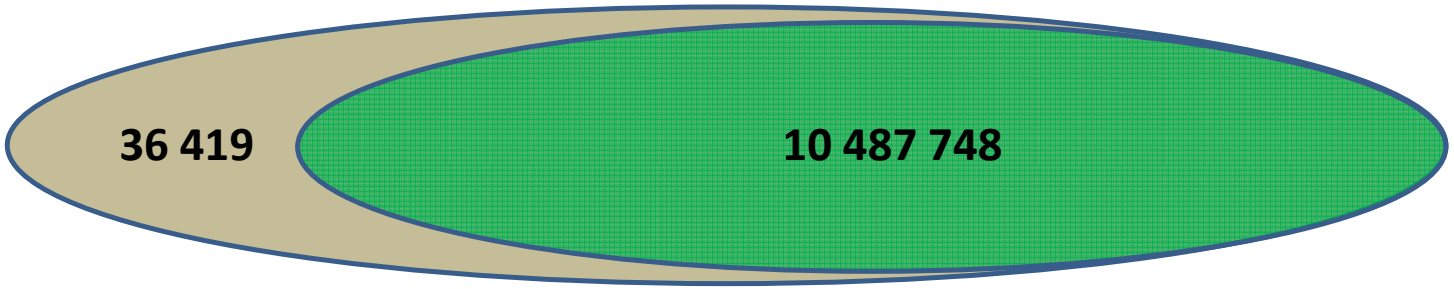


Vymezení obyvatelstva v SLDB 2021



Obvykle a trvale bydlící obyvatelstvo podle SLDB 2021

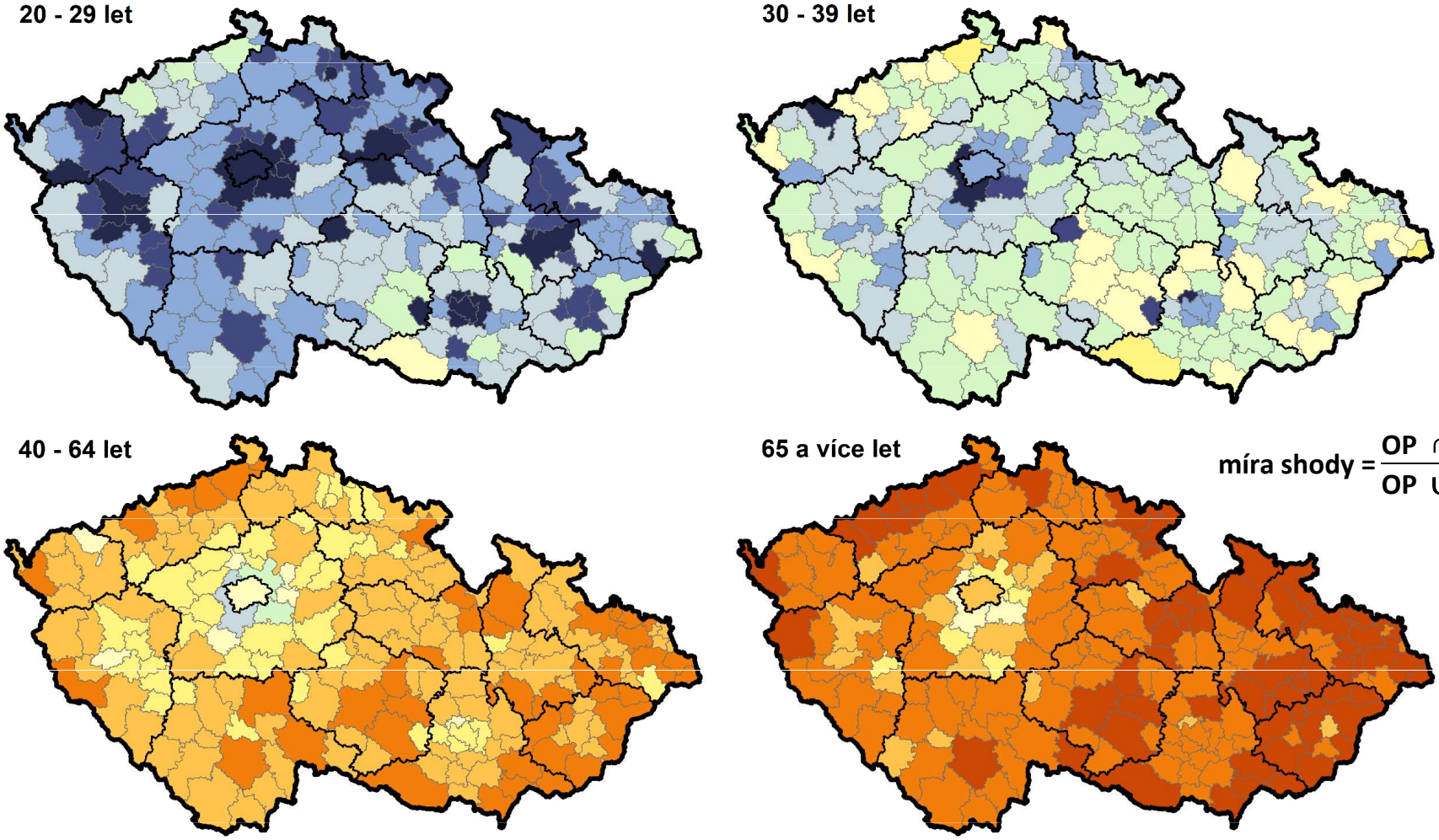
- Počet obvykle bydlících obyvatel (OP): 10 524 167
- Počet „trvale“ bydlících obyvatel (TP): 10 487 748



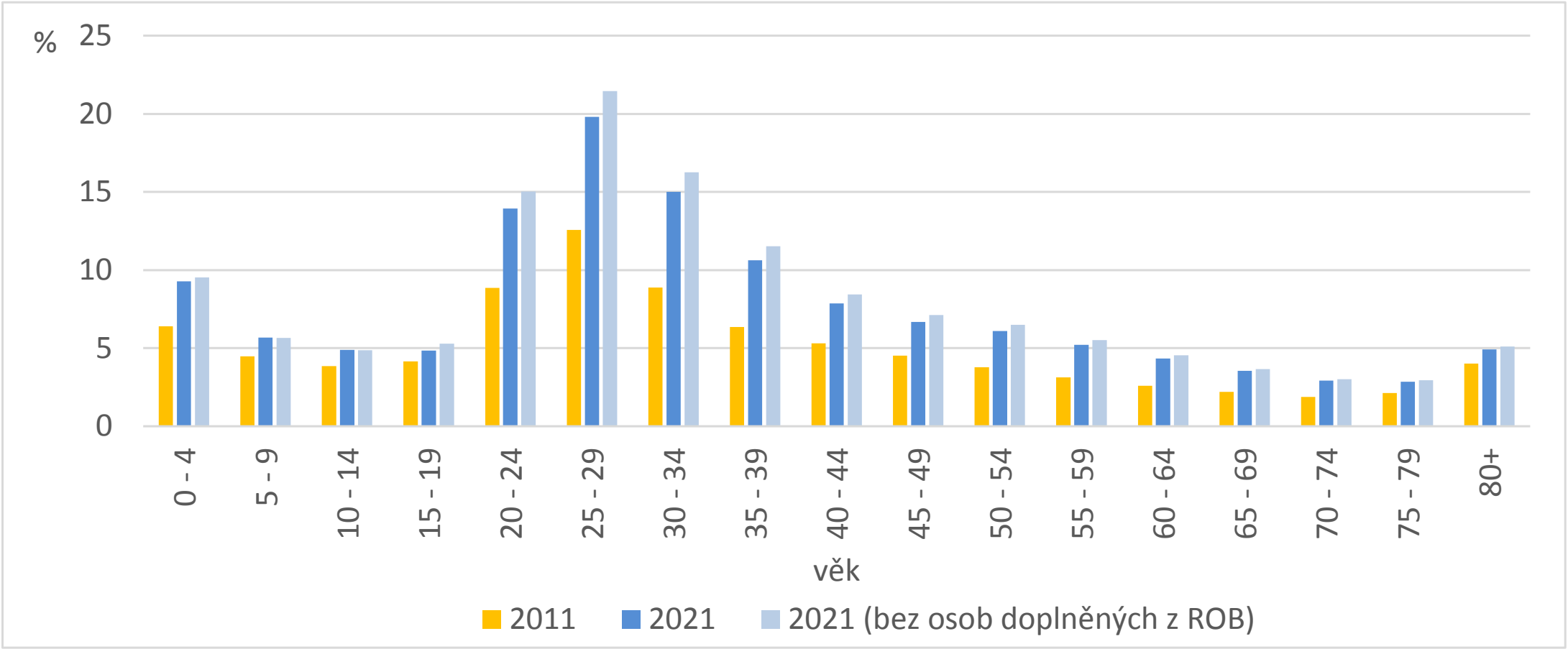
standardní výsledky SLDB
(obvykle bydlící)

základ pro bilance
(„trvale“ bydlící)

Míra shody obvykle a "trvale" bydlícího obyvatelstva podle věku ve srprávních obvodech ORP (SLDB 2021)

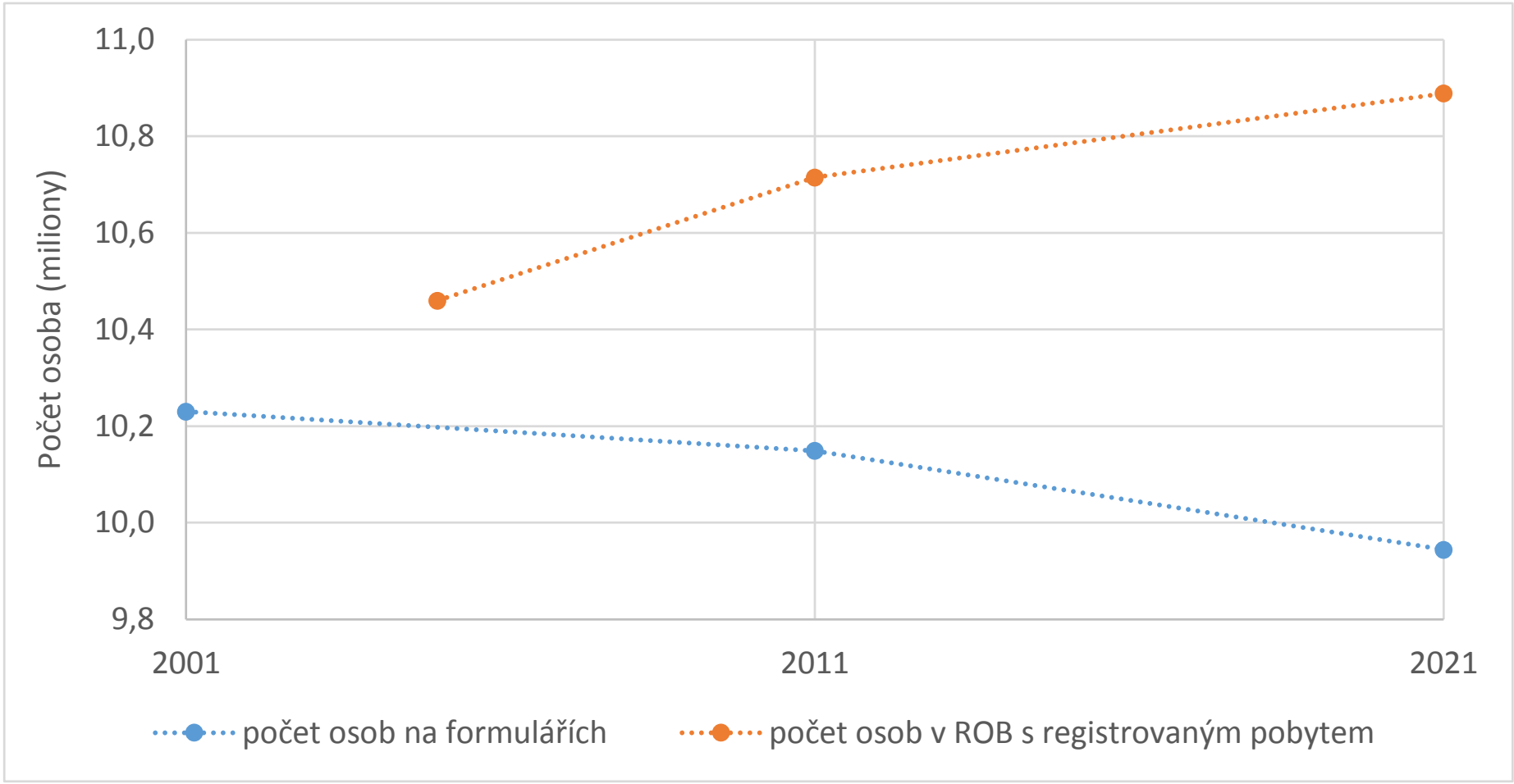


Podíly obyvatel obvykle i trvale bydlících na území Česka s rozdílnými obcemi obvyklého a trvalého pobytu podle věku



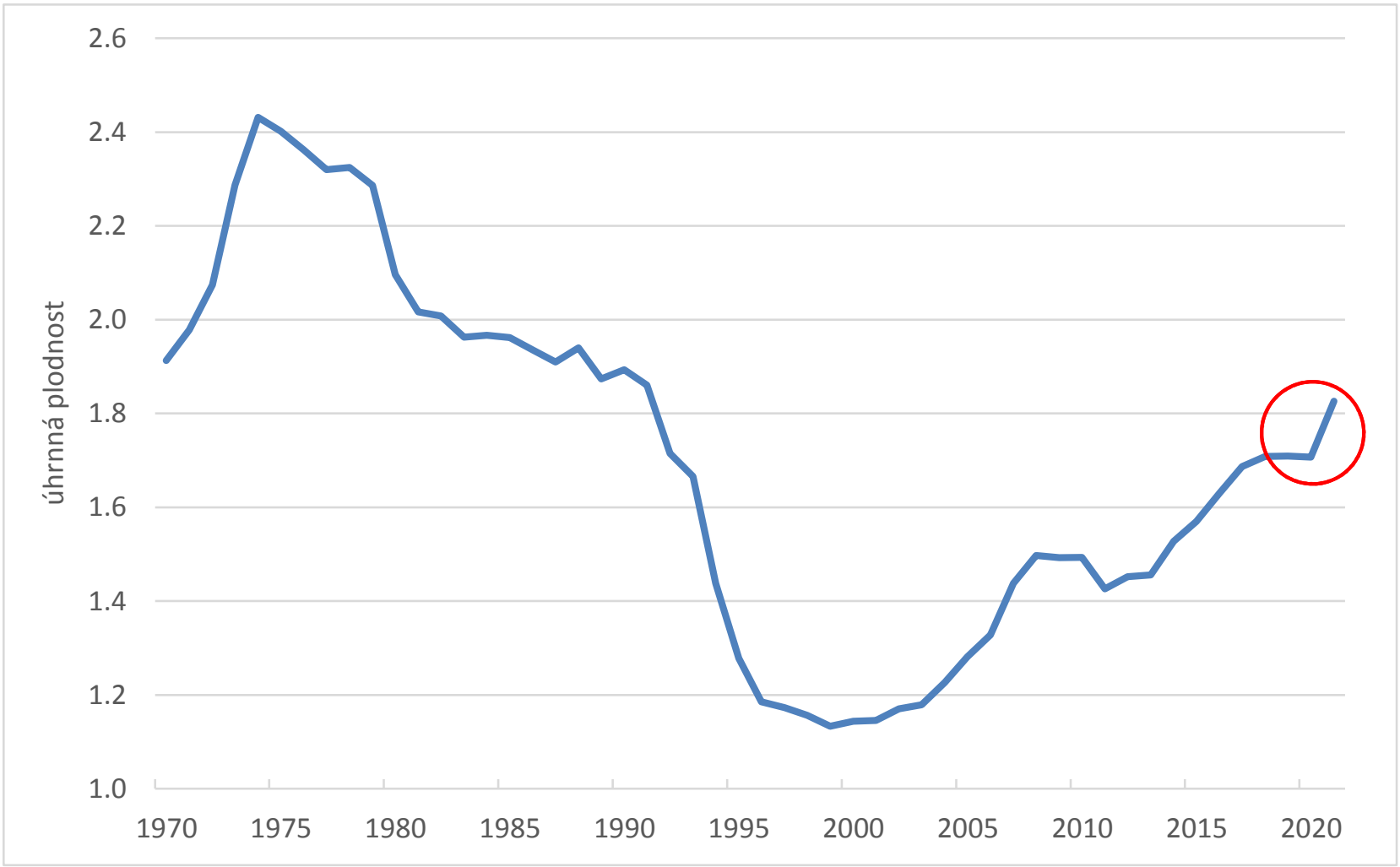
100 % - počet obyvatel v dané věkové skupině s obvyklým i trvalým pobytem v Česku

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

Úhrnná plodnost v období 1970 – 2021



Děkuji za pozornost.