Tehnična dokumentacija za bazo `produktivnost` na PostgreSQL strežniku na `umar-bi`

mz

11.June 2024

Vsebina

1 Pregled **1**

2 R skripte **1**
2.1 `01_eurostat_produktivnost.R` 1

3 Baza `produktivnost` **2**
3.1 Dostopanje do baze 2
3.1.1 Excel Data Connection / Power Query 2
3.1.1.1 Pred prvo povezavo 2
3.1.1.2 Vsaka nadaljna povezava 2
3.1.2 R 2
3.2 Vzpostavitev baze 3
3.3 Tabele 3
3.3.1 Tabela `produktivnost_makro` 3
3.3.1.1 Spremenljivke 3

4 Arhiv **4**

5 Priloga - razlike med eurostat agregati in ročno izračunanimi - Tabela `produktivnost_makro` **4**

6 Priloga II - razlike med uporabo prebivalstva iz tabele `nama_10_pe` in prebivalstva iz `demo_pjan` tabele - Tabela `produktivnost_makro` **7**

1 Pregled

Dokument vsebuje tehnično dokumentacijo glede podatkovnih tokov za zajem, obdelavo in zapis podatkov o produktivnosti v bazo `produktivnost` na PostgreSQL strežniku na `umar-bi`

Gre za selitev podatkovnih tokov, ki so pred tem temeljili na Katarininih skriptah in se zapisovali v MS Access datoteke na M: (glej Arhiv spodaj)

2 R skripte

2.1 `01_eurostat_produktivnost.R`

originalna verzija - `arhiv/1_Macro_PROD.R`

opis:

1. zajem podatkov iz evrostata iz tabel `nama_10_gdp` (BDP), `nama_10_a10_e` (zaposlenost) in `nama_10_pe` (prebivalstvo)
2. izračun agregatov za osem skupin držav (glej @ref(Tbl_pm) za detajle in združitev tabel
3. izračun 12 novih spremenljivk (glej @ref(Tbl_pm) za detajle
4. zapis v tabelo `produktivnost_makro`

3 Baza `produktivnost`

Na PostgreSQL strežniku na `umar-bi` (PostgreSQL 15) je več podatkovnih baz, za namene centralnega skladiščenja in dostopa do podatkov je bila postavljena nova baza z imenom `produktivnost`

Za začetek so vse tabele znotraj `public` sheme (najvišja strukturna raven znotraj baze), po potrebi lahko dodamo več shem in razdelimo tabele v vsebinske sklope.

3.1 Dostopanje do baze

Dostop do baze je mogoč samo z uporabniškim imenom in geslom, ki ga lahko dodeli administrator Postgres strežnika (trenutno mz).

3.1.1 Excel Data Connection / Power Query

3.1.1.1 Pred prvo povezavo Pred prvo uporabo povezave je potrebno inštalirati ODBC driver za postgres (Open Database Connectivity), ki ga dobiš na https://www.postgresql.org/ftp/odbc/releases/. Izbereš zadnjo verzijo in znotraj mape izbereš 64 bitno verzijo `.msi` datoteke in jo preneseš. **Za namestitev rabiš admin pravice, zato rečeš Petru, da ti on uredi!**

Naslednji korak: `Control panel` / `Admin Tools` / `ODBC Data Sources (64 bit)`, na prvem zavihku izbereš `Add` in iz seznama izbereš `PostgreSQL Unicode(x64)` in potem `Finish`. Potem izpoloniš setup polja:

- `Data Source`: to je ime, po katerem boš povezavo spoznal, tako da recimo `produktivnost baza` ali kaj podobnega
- `Database`: produktivnost
- `Server`: 192.168.38.21
- `Port`: 5432
- `User Name`: svoje uporabniško ime (dobiš od Maje)
- `Password`: svoje geslo (dobiš od Maje)

In še zadnji korak: odpri Excel, `Blank document` in izberi `Data`, `Get Data`, `From other sources`, `From ODBC` in iz seznama DNS izberi vir, ki si ga ravnokar poimenoval (torej "produktivnost baza" vz zgornjem primeru). Prvič, ko to narediš, te spet vpraša za uporabniško ime in geslo, kasneje pa tega ne bo več.

3.1.1.2 Vsaka nadaljna povezava Vsakič, ko hočeš dobiti podatke iz baze uporabiš sledeči postopek:

- odpreš Excel, in izbereš `Data` / `Get Data` / `From other sources` / `From ODBC`
- iz seznama `Data source names (DNS)` izbereš vir, kot si ga poimenoval in klikneš OK
- odpre se `Navigator`, kjer izbereš na katero tabelo se hočeš povezati in potem klikneš `Transform Data`, da se odpre Power Query (če namesto tega klikneš na `Load`, se ti bo prenesla cela tabela, česar ponavadi nočeš).

3.1.2 R

Za povezovanje iz `R`-ja potrebuješ knjižnjici `DBI` in `RPostgres`, za lažje delo s poizvedbami pa tudi `dbplyr`, ki se integrira z `dplyr`-jem:

```
# install.packages("DBI")
# install.packages("RPostgres")
# install.packages("dbplyr")
```

Povezavo vsakič vzpostaviš z naslednjo kodo, kjer vstaviš uporabniško ime in geslo (uporabi narekovaje okoli obeh):

```
con <- DBI::dbConnect(RPostgres::Postgres(),
                      dbname = "produktivnost",
                      host = "192.168.38.21",
                      port = 5432,
                      user = <uporabniško ime>,
                      password = <geslo>)
```

Takole pa potem dostopaš do podatkov: npr. najprej poglej katere tabele so na voljo in potem naredi poizvedbo. Uporabljaš lahko običajne "pipe", samo na začetku moraš uporabiti `tbl()` da se povežeš in na koncu `collect()` da ti vrne tabelo:

```
# poglej najprej, če je kaka tabela tam
DBI::dbListTables(con)

# naredi poizvedno na tabeli:
tbl(con, "produktivnost_makro") |>
  filter(geo == "EU27_2020") |>
  collect()
```

3.2 Vzpostavitev baze

3.3 Tabele

3.3.1 Tabela `produktivnost_makro`

geo:

- 28 EU držav (vključno z UK) - uporabljene so dvomestne ISO oznake (`SI` za Slovenijo...)
- štirje originalni agregati iz eurostata [1]
  - `EU28` (samo do leta 2019),
  - `EU15` (samo do leta 2019),
  - `EA19`,
  - `EU27_2020`
- osem dodatnih agregatov, izračunanih iz podatkov posameznih držav:
  - `EU13`,
  - `EU14`,
  - `EU27`,
  - `EU27noIE` - EU brez Irske
  - `EA20` - Evro območje
  - `EAnoIE` - Evro območje brez Irske
  - vodilne `inovatorke` (trenutno BE, DK, SE, FI in NL) ter
  - `V4` - Višegrajske 4 (CZ, HU, SK, PL)

3.3.1.1 Spremenljivke

v spodnji tabeli so razdeljene v tri skupine:

- originalne spremenljivke, nespremememnjene iz eurostata (v zadnjem stolpcu je ime izvorne tabele)
- preračunane spremenljivke (v zadnjem stolpcu je nakazan preračun glede na zaporedne številke spremenljivk)
- preračunani indeksi za EU27 = 100 [2]

Posebna pozornost je potrebna glede naslednjih parov spremenljivk

[1] Glej prilogo I., ampak na kratko: priporočljivo je, da se uporabljajo naši izračunani agregati in ne eurostatovi, kar je v praksi relevantno samo za EU27.

[2] Pri tem je uporabljen izračunan agregat in ne originalen Eurostatov

#	Oznaka	Opis	Vir / Preračun
(1)	CP_MPPS_EU27_2020_B1GQ	BDP, tekoče cene, v mio SKM (EU27 od 2020)	nama_10_gdp
(2)	CLV10_MEUR_B1GQ	BDP, stalne cene leta 2010, v mio EUR	nama_10_gdp
(3)	THS_PER_EMP_DC	Skupna zaposlenost, domači koncept, v 1000 oseb	nama_10_a10_e
(4)	THS_HW_EMP_DC	Delovne ure, domači koncept, v 1000	nama_10_a10_e
(5)	THS_PER_POP_NC	Št. Prebivalcev, nacionalni koncept, v 1000	nama_10_pe
(6)	NR_20_64	Št. običajnih prebivalcev starosti 20-64 let, v 1000	demo_pjan
(7)	NR_TOTAL	Št. običajnih prebivalcev, v 1000	demo_pjan
(8)	GDP_PC_PPS	BDP na prebivalca v SKM	(1) / (5)
(9)	GDP_PC_PPS_pjan	BDP na običajnega prebivalca v SKM	(1) / (7)
(10)	PROD_PPS	Produktivnost v SKM, na zaposlenega	(1) / (3)
(11)	PROD_PPS_HW	Produktivnost v SKM, na delovno uro	(1) / (4)
(12)	PROD_real	Realna produktivnost, na zaposlenega	(2) / (3)
(13)	PROD_real_HW	Realna produktivnost, na delovno uro	(2) / (4)
(14)	EMP_RATE	Delež zaposlenih v celotnem prebivalstvu	(3) / (5)
(15)	HW_EMP	Delovne ure na zaposlenega	(4) / (3)
(16)	EMP_W_AGE_PROP	Delež zaposlenih v preb (20-64)	(3) / (6)
(17)	W_AGE_PROP	Delež preb 20-64 v skupnem prebivalstvu	(6) / (5)
(18)	W_AGE_PROP_pjan	Delež preb 20-64 v skupnem prebivalstvu	(6) / (7)
(19)	GDP_PC_PPS_EU27_100	BDP pc v SKM (indeks EU27 = 100)	
(20)	GDP_PC_PPS_pjan_EU27_100	BDP pc v SKM (indeks EU27 = 100)	
(21)	PROD_PPS_EU27_100	Produktivnost v SKM (indeks EU27 = 100), na zaposlenega	
(22)	PROD_PPS_EU27_100_HW	Produktivnost v SKM (indeks EU27 = 100), na delovno uro	
(23)	EMP_RATE_EU27_100	Stopnja zaposlenosti (indeks EU27 = 100)	
(24)	HW_EMP_EU27_100	Delovne ure na zap. (indeks EU27 = 100)	
(25)	EMP_W_AGE_PROP_EU27_100	Delež zaposlenih v preb (20-64) (indeks EU27 = 100)	
(26)	W_AGE_PROP_EU27_100	Delež preb 20-64 v skupnem prebivalstvu (indeks EU27 = 100)	
(27)	W_AGE_PROP_pjan_EU27_100	Delež preb 20-64 v skupnem prebivalstvu (indeks EU27 = 100)	

Table 1: Spremenljivke v tabeli 'produktivnost_makro'

- BDP na prebivalca: spremenljivki 8 in 9 (oz. 19 in 20 za indekse): prva uporablja prebivalstvo po nacionalnem konceptu in je tako *pravilna*, druga uporablja vsoto "običajnih prebivalcev" je v tabeli dodana samo za primerjavo
- Delež prebivalcev 20-64 v skupnem prebivalstvu: spremenljivki 17 in 18 (oz. 26 in 27): tukaj je *pravilna* druga spremenljivka, kjer sta v imenovalcu in števcu števili običajnih prebivalcev (iz tabele `pjan`, medtem ko je v prvi definicija prebivalstva v imenovalniku po nacionalnem konceptu, kar ni OK, ampak je dodana samo za primerjavo.

4 Arhiv

V mapi `arhiv` se nahajajo predvsem originalne Katarinine skripte na podalagi katerih so bile narejene nove skripte za na bazo `produktivnost`. Gre za 5 oštevilčenh skript, zaporedne številke so enake tudi pri novih skriptah (

5 Priloga - razlike med eurostat agregati in ročno izračunanimi - Tabela `produktivnost_makro`

Eurostat ima v tabeli `produktivnost_makro` štiri agregate že izračunane: EU28, EU15, EA19 in EU27_2020. Ti agregati se ne ujemajo popolnoma z agregati, ki jih lahko sami izračunamo iz podatkov posameznih držav - kar delamo za ostalih 7 agregatov.

Spodaj so prikazane razlike med Eurostat agregatom in ročno izračunanim agregatom za 5 glavnih spremenljivk v tej tabeli za EU19 (rdeče) in EU27_2020 (črno).

Zakaj do teh razlik prihaja, ni čisto jasno, sploh ker se dinamika odstopanja razlikuje glede na spremenljivko. Razlike niso velike (največja je 0.03% pri delovnih urah), ampak vseeno se priporoča, da se uporablja naše izračunane agregate, namesto Eurostatovih - sploh če se prikazuje več agregatov hkrati, da so vsi konsistentno izračunani na isti način.

Figure 1: delta za EA19 - BDP, tekoče cene, v mio SKM

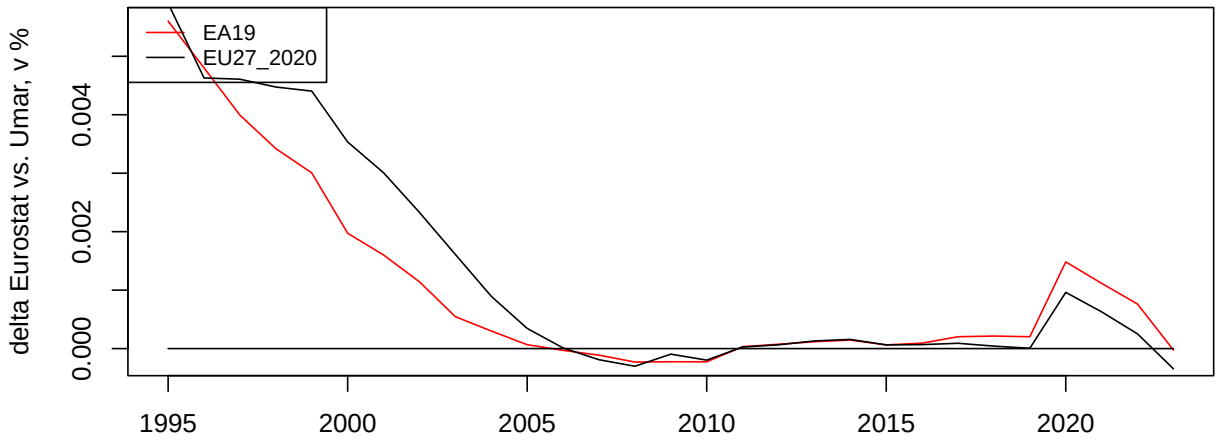


Figure 2: delta za EA19 - BDP, stalne cene leta 2010, v mio EUR

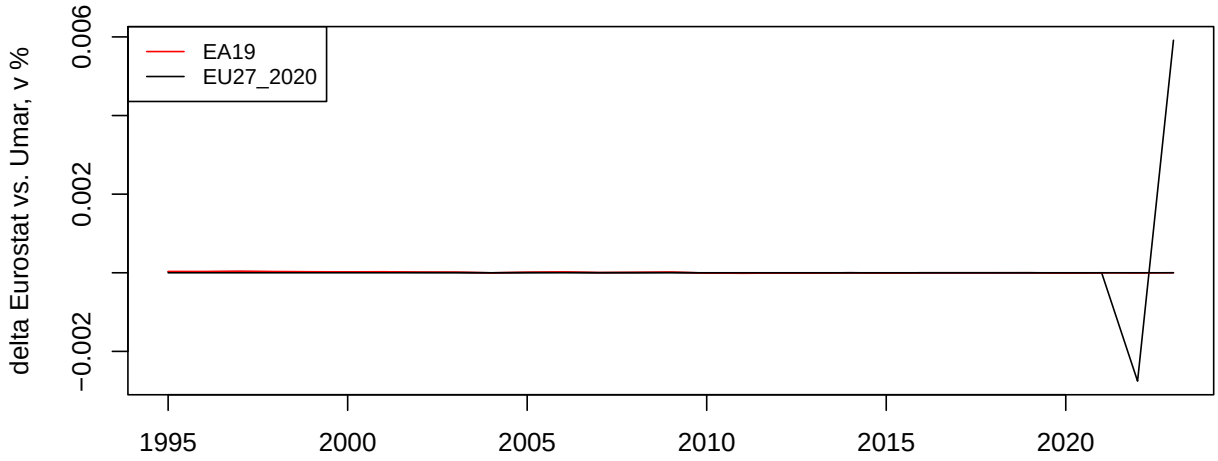


Figure 3: delta za EA19 - Skupna zaposlenost, domači koncept, v 1000 oseb

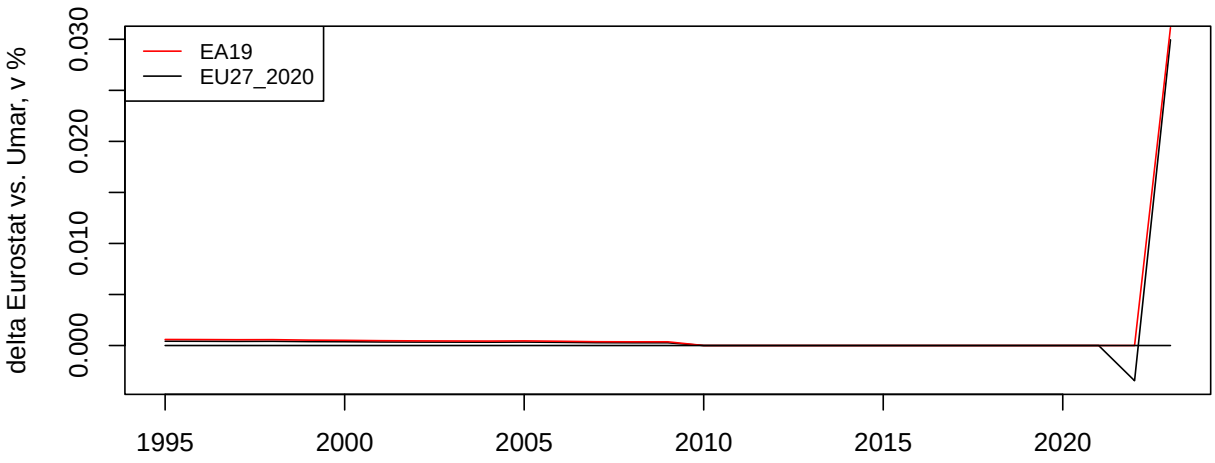


Figure 4: delta za EA19 - Delovne ure, domači koncept, v 1000

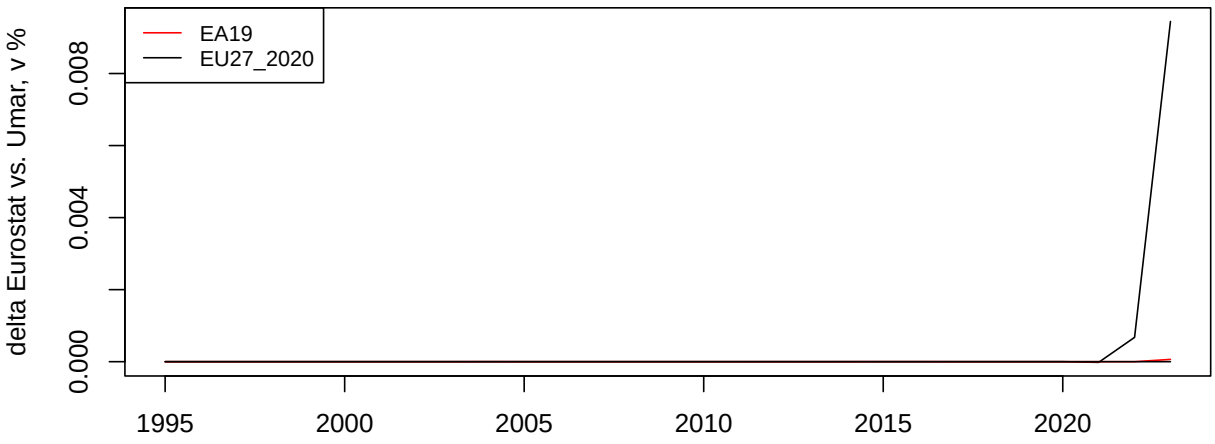


Figure 5: delta za EA19 - Št. Prebivalcev, nacionalni koncept, v 1000

6 Priloga II - razlike med uporabo prebivalstva iz tabele `nama_10_pe` in prebivalstva iz `demo_pjan` tabele - Tabela `produktivnost_makro`

Za izračun spremenljivke `BDP na prebivalca v SKM` se uporablja prebivalstvo po nacionalnem principu iz tabele `nama_10_pe`, kjer pa ni podatkov po starostnih skupinah. Za dekompozicijo spremenljivke po naslednji formuli:

$$\frac{BDP}{preb} = \frac{BDP}{zap.} \times \frac{zap.}{preb._{20-64}} \times \frac{preb._{20-64}}{preb.}$$

oz. opisno: BDP per capita `GDP_PC_PPS` je produkt:

- produktivnosti – `PROD_PPS`
- deleža zaposlenih v prebivalstvu starosti 20-64 – `EMP_W_AGE`
- deleža prebivalstva starosti 20-64 v celotnem prebivalstvu – `W_AGE_PROP`

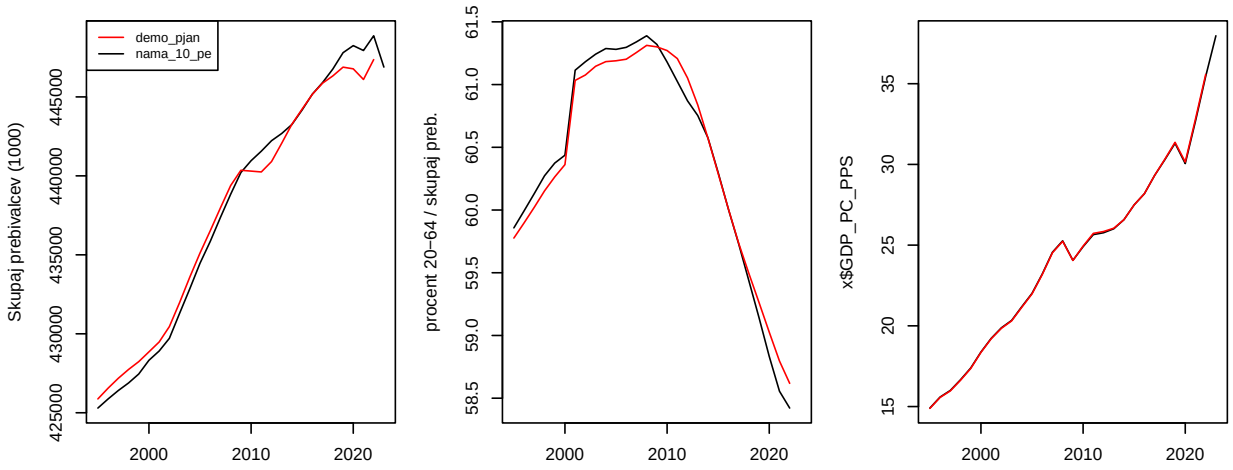
moramo uporabiti podatke o prebivalstvu iz Eurostatove tabele `demo_pjan`, ker so samo tam podatki po starostnih skupinah, ki so potrebni za drugi dve komponenti. Definicija prebivalstva v tej tabeli je "*Usually resident population which represents the number of inhabitants of a given area on 1 January of the year in question (or on 31 December of the previous year).*"

medtem ko je definicija v tabeli `nama_10_pe`: "*all persons, nationals or foreigners, who are permanently settled in the economic territory of the country, even if they are temporarily absent from it, on a given date. A person staying or intending to stay at least one year is considered to be settled on the territory. By convention, the total population excludes foreign students and members of foreign armed forces stationed in a country.*" Ker gre za pomožne indikatorje nacionalnim izračunom, so podatki na letni ravni oz. gre za vrednosti na sredini leta za razliko od demografskih podatkov v tabeli `demo_pjan`, ki veljajo na prvi dan leta.

Za uporabo podatkov o številu prebivalstva v starostni skupini 20-64 torej moramo uporabiti podatke iz `demo_pjan`, kjer vzamemo povprečje dveh zaporednih let, da dobimo vrednost na sredini leta. V praksi to žal pomeni, da podatek za zadnje leto še ni dostopen, dokler ni objavljen podatek za tekoče leto.

Problem pa se pojavi pri vprašanju katere podatke uporabiti za skupno število prebivalcev. Za BDP per capita (`GDP_PC_PPS`) se zdi smiselno uporabiti podatke, ki so skupaj objavljeni, torej prebivalstvo po nacionalnem konceptu sredi leta. Ampak za dekompozicijo, natančneje zadnjo komponento deleža prebivalstva starosti 20-64 v celotnem prebivalstvu (`W_AGE_PROP`), je seveda smiselno uporabiti prebivalstvo iz tabele `demo_pjan`, da sta v imenovalcu in številcu enaki definiciji prebivalstva.

Spodaj je za Slovenijo, EU27, Nemčijo in Italijo prikazano kakšne so razlike med obema možnostima: na levi razlika med skupnim številom prebivalstva (rdeče linije so iz `demo_pjan`), na serdini razlika med zadnjo komponento, torej delež working age v skupnem prebivalstvu in na desni BDP per capita.



Pri Nemčiji so razlike pred letom 2011 precej dramatične, kar je povezano s tem, da je Nemčija leta 2011

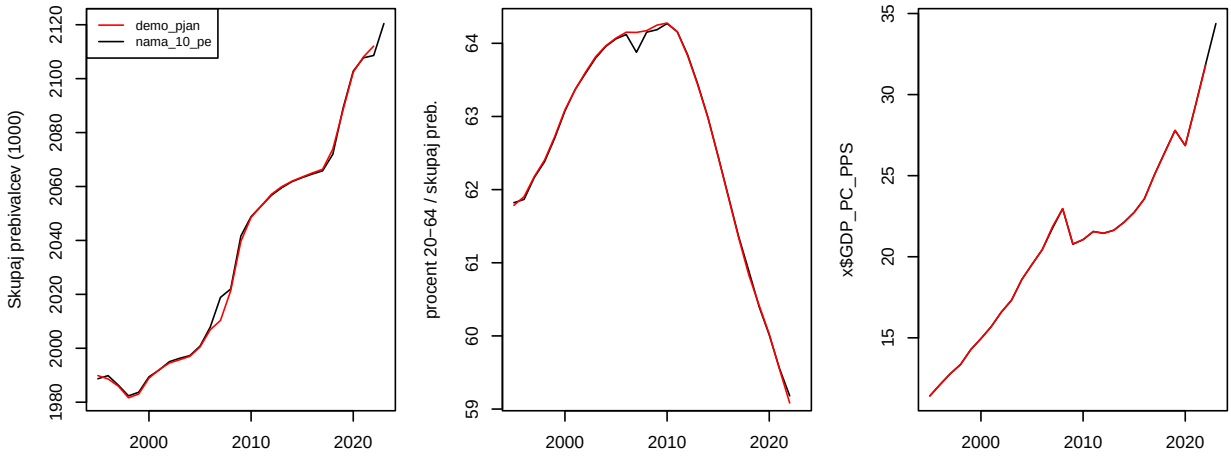
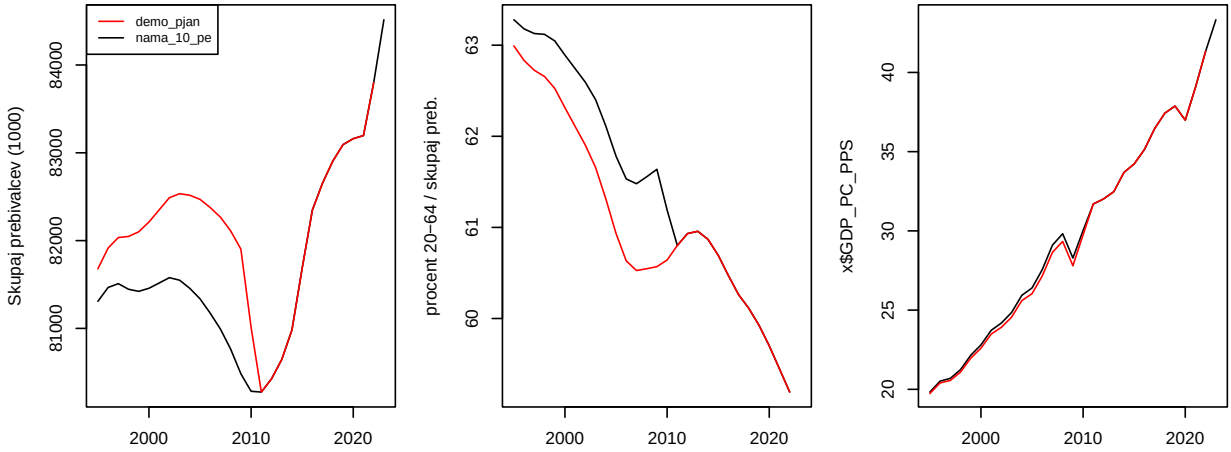


Figure 6: Slovenija

izvedla prvi popis na podlagi centralnih registrov, do takrat pa so ocene temeljine na podatkih popisa iz leta 1987. Ta popis 2011 je pokazal, da je prebivalcev v resnici cca 1.5 milijona manj, kot so pred tem ocenjevali. "Due to the long inter-censal period, the Federal Statistical Office of Germany decided not to produce backward-adjusted population estimates by single-year ages and sex for the whole period." [3] Skratka očitno so popravke za nazaj naredili samo na skupnih podatkih (in so torej v `nama_10_pe` tabeli), ne pa na podatkih po starostih, zato je vsota v `demo_pjan` še vedno narobe, ker je niso nikoli popravili.



Za Italijo (spodaj) nisem našla kake elegantne razlage, sem pa vprašala Ale in čakam odgovor.

[3] https://www.bib.bund.de/Publikation/2018/Adjusting-inter-censal-population-estimates-for-Germany-1987-2011.html?nn=1219476

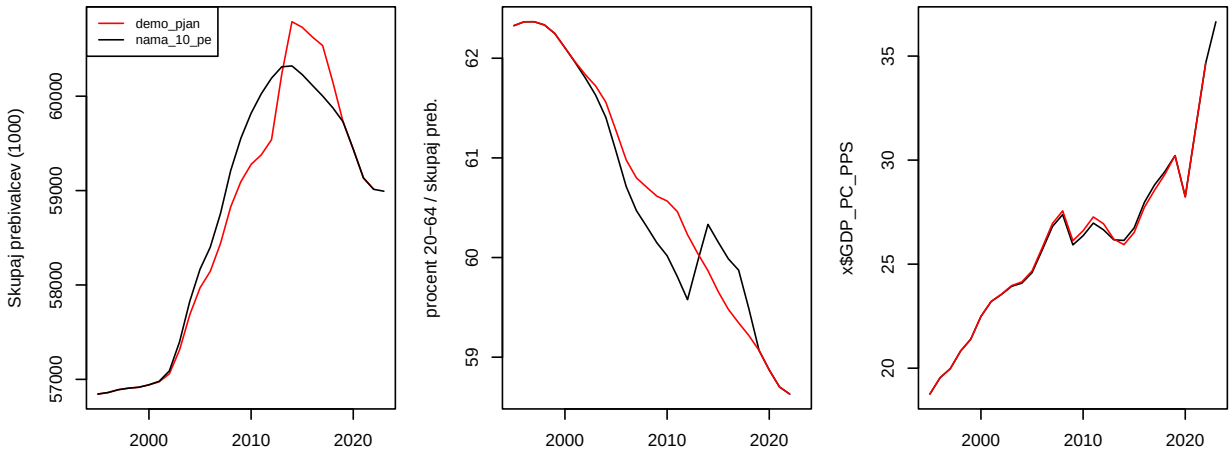


Figure 7: Italija