REGACC Production Process: a practical guide, 2022

Luis Biedma

2022-11-26

Table of Contents

# About

This is a soft but comprehensive guide detailing in some detail the steps to follow during the production process of Regional Accounts in 2022. It complements and tries to avoid to repeat what was explained, mainly regarding coding, in the 2021 guide (*U:/03\_Regional Accounts/03I\_Documentation/REGACC\_process/codebook\_2021/index.html*). Coding aspects are restricted to some novelties and instead the guide tries to illustrate the logic of why some things are done in a specific way and the sequential steps of the process Some practical examples of what to look and not miss on each step are included. It will additionally detail certain improvements implemented this year, which in many cases are not evident to the end user without digging into the code.

## Pre-requisites

For the production process we have minimised IT dependencies. There are things that are needed for the production process (FameFeed, Matis, etc) but for most steps the only things we would need are a network connection to the server with the data in the first step and R and Rstudio with some packages installed for the remaining steps.

## Tacit knowledge

Some tacit knowledge on the process is helpful. The validation procedure is quite exhaustive and using all the possibilities can lead to devote to a counterproductive amount of time to deal with one country. The following table classifies countries according to my perceived level of difficulty in dealing with them in the past. Problems can be due to recurring technical issues in the files provided, quality issues of the data, previous experiences in receiving several updates, number of regions, etc.

Time will be better employed dealing with more problematic countries, and they should be dealt with priority over the other countries if we are time constrained.

* For **Cyprus** and **Luxembourg** we just need to check if the data is equal to the one reported in NAMA as they do not have regions. If by chance we find something strange in the data we should inform our NAMA colleagues. **Estonia**, **Malta** and **Latvia** only have regional data in table 12. However these group of countries (5) tend to do some mistakes when generating the files (not including NUTS 3 or including NUTS 3 where they shouldn’t)
* **Austria**, **Czechia**, **Denmark**, **Finland**, **Italy**, **Portugal** and **Slovenia** send the data early, pass all the checks and whenever in the past we have asked questions they have provided reasonable explanations relatively fast.
* **Belgium**, **Bulgaria**, **Hungary**, **Netherlands**, **Spain**, **Lithuania** and **Slovakia** send normally plausible figures but some further attention is needed compared to the previous group.
* **Croatia**, **France**, **Germany**, **Greece**, **Ireland**, **Poland** and **Sweden** atre the countries that will deserve most attention for a number of reasons.

## Most important changes compared to 2021

* A more logical workflow. All steps are included in a single project and more data is shared across processes.
* Further automatisation of some processes.
* Calculation of GVA in previous year prices for Matis.
* No Matis extractions are needed. New data comes directly from xml files and previously validated data from {dataregacc}.
* Use of internal {regacc} and {dataregacc} packages.
* Harmonising the classification and codes used at different steps. You will use this codes consistently: **country, ref\_area, NUTS, sto, activity, accounting entry, unit\_measure, time\_period and obs\_value**.
* Different threshold by unit measure and encoding of some country exceptions to avoid showing inconsistencies we are already aware of.
* (Possible) outliers for not revised periods are excluded.
* Handling of NA values in NACE additivity checks.
* {data.table} is used in a few more more places than in the past.
* More consistent look of the charts.
* A new predefined scatterplot to better detect outliers in the exploratory shiny app.
* Improve shiny re-activity to select regions.

All the needed files are in the folder in *U:/03\_Regional Accounts/03D\_Data Production/2022/R/regacc* and should be copied to a hard drive or usb stick, not the H:/ or any other folder location if we want them to run fast.

# Steps before validation

The script UI.R provides the sequential steps to be followed when analysing incoming data. It starts by loading some libraries and the user needs to type the country it is going to analyze. It will bring the xml files from the INPUTfolder if the data has not been loaded with FameFeedof from the DONEfolder if they have been loaded. It is preferable generally to do all these steps before loading data in Matis as it is a few seconds faster to do it from the INPUTfolder and we will need to wait for FameFeed to finish, which can take a long time for countries with many regions. The user is also suggested to copy locally the country metadata file from the server.

## Basic information and building the files

This step is done with:

rmarkdown::render("01\_basic\_info.Rmd",  
 params = list(report = country\_sel),  
 output\_file = paste0("basic\_info/",country\_sel,"\_",format(Sys.time(),"%Y-%m-%d"),"\_report.html"))

This will produce a day-stamped html file in the folder /basic\_info and prepare the file for the next steps in /data/csv. The html report contains some basic information about the file which is difficult to gather in a nicely presented manner in other existing reports.:

### Metadata

We have prepared (darft) metadata files according to the template discussed in the 2022 EG RA meeting.

The xml files are read and the following information is presented:

### Periods reported by table

Please note that if a file included 2000-2008 and another 2009-2021 the report will show 2000-2021 as the reported period. If only 2009-2021 is reported in the files that is what it will show.

### Embargo dates

Not in general a very relevant issue for us as we publish the data two months later but some countries (**Greece**, **Croatia**, **Ireland**) might include a wrong date or forget to include it or it might require that we negotiate the publication date with them.

| country | table\_identifier | embargo\_date |
| --- | --- | --- |
| IE | T1001 | 2022-04-14T12:00:00 |

### Comments

Comments in general are not used by countries but in some cases provide useful information.

| country | table\_identifier | comment\_ts |
| --- | --- | --- |
| AL | T1001 | GVA data for the year 2020 are semi finale. |
| AL | T1001 | GVA data for the year 2021 are not available. |

### Data not for publication

We should pay attention to the few cases where data is sent as not publishable.

| country | table\_identifier | sto | activity | unit\_measure | conf\_status |
| --- | --- | --- | --- | --- | --- |
| MT | T1300 | B5N | \_Z | XDC | N |
| MT | T1300 | B6N | \_Z | XDC | N |
| MT | T1300 | B7N | \_Z | XDC | N |
| MT | T1300 | D4 | \_Z | XDC | N |
| MT | T1300 | D5 | \_Z | XDC | N |
| MT | T1300 | D61 | \_Z | XDC | N |
| MT | T1300 | D62 | \_Z | XDC | N |
| MT | T1300 | D63 | \_Z | XDC | N |
| MT | T1300 | D7 | \_Z | XDC | N |

### Flags

We also need to be attentive to flags, specially B and D. P, and to a more limited extent E, flags are frequently used. Special attention should be put on M flags as they will show up as reported in the “official” NQR completeness (not in ours).

| country | table\_identifier | sto | unit\_measure | obs\_status | time\_period |
| --- | --- | --- | --- | --- | --- |
| AL | T1001 | B1G | PC | P | 2020 |
| AL | T1001 | B1G | XDC | P | 2020 |
| AL | T1001 | POP | PS | E | 2020 |
| AL | T1001 | POP | PS | E | 2021 |

### Building the file with data

This is a necessary step to build a file with all the information we need for the all subsequent checks and reports in the validation process.

#### Temporary series

It starts by reading (again) the xml files for the country in data/xml. However, in this occasion we will ignore all the contextual information (flags, comments, etc) and get only the dimensions needed to identify the numbers. It also takes one single value per observation. This will not be normally needed but occasionally we received transmissions like this:

| value | ref\_area | sto | time\_period | obs\_value |
| --- | --- | --- | --- | --- |
| D:/regacc/doc/NAREG\_T1001\_A\_DE\_2020\_0000\_V0002.xml | DED2 | POP | 2020 | 1593.227 |
| D:/regacc/doc/NAREG\_T1001\_A\_DE\_2020\_0000\_V0002.xml | DED | POP | 2020 | 4064.456 |
| D:/regacc/doc/NAREG\_T1001\_A\_DE\_2020\_0000\_V0002.xml | DED5 | POP | 2020 | 1051.173 |
| D:/regacc/doc/NAREG\_T1001\_A\_DE\_2020\_0000\_V0002.xml | DED4 | POP | 2020 | 1420.056 |

Germany corrected population for a particular year and regions (DED2, DED4) and did not provide the unchanged data. This is quite an important difference between REGACC and other domains with several implications for the use of common tools which we will no describe here.

While building the csv file the data is consolidated with previous transmissions and the latest value is kept.

| date | ref\_area | sto | time\_period | obs\_value |
| --- | --- | --- | --- | --- |
| 2021-12-21 12:30:52 | DED | POP | 2020 | 4064.456 |
| 2021-12-21 12:30:52 | DED2 | POP | 2020 | 1420.056 |
| 2021-12-21 12:30:52 | DED4 | POP | 2020 | 1593.227 |
| 2021-12-21 12:30:52 | DED5 | POP | 2020 | 1051.173 |
| 2022-01-11 11:06:24 | DED2 | POP | 2020 | 1593.227 |
| 2022-01-11 11:06:24 | DED4 | POP | 2020 | 1420.056 |

The created file will include the newest values for DED2 and DED4 and the previously reported values for the other regions.

#### Validated series

For validated we will use the series internally available in {dataregacc}. That means that they are static and we will not be including in the file the most recent validated data (after October 2022). In the case that we have validated some series and we receive updates, the validated series will still be the same ones as before.

We need to do some transformations in the series to create a table t1001 which takes the data up to 2019 from t1200 and only keeps the validated data from t1001 for 2020, as there might be outdated (validated two or three years ago) t1001 tables. This happens if the country usually reports only the latest year in t1001. We will need to make an exception for volume data that is only reported in the table.

val<- dataregacc::validated %>%   
 filter(country==country\_sel) %>%   
 mutate(type="V")  
  
gvagr<- val %>%   
 filter(unit\_measure=="PC")  
  
lasty<- val %>%   
 filter(table\_identifier=="T1001" & time\_period>= 2020 & unit\_measure=="XDC")# remove GVA in PYP  
  
prevy<- val %>%   
 filter(table\_identifier=="T1200" & time\_period< 2020 & NUTS!="3" & activity %in% c("\_T","\_Z")) %>% # remove GVA in PYP  
 mutate(table\_identifier="T1001")  
  
no\_t1001<- val %>%   
 filter(table\_identifier!="T1001")  
  
val <- bind\_rows(gvagr, lasty, prevy,no\_t1001)

#### Combining them

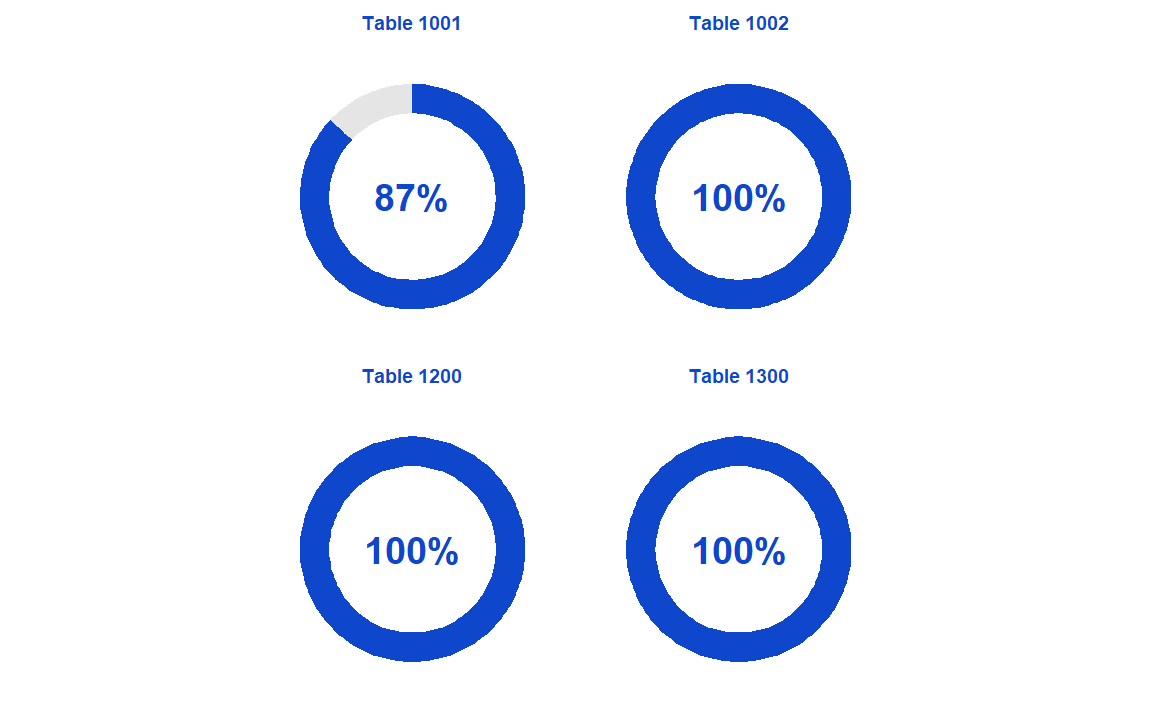
A necessary final step is to first fill in the temporary series and then combine them in a single data frame which is exported as a date-stamped csv file in the folder data/csv.

By filling we mean that we take the validated value and impute it to the temporary value when the temporary series are empty for earlier periods. We need to do that to handle cases when:

* We have not received all the tables.
* We have not received complete time series (only updates for the most recent years).

### Completeness

We will get also some information on the completeness of the files. If we see that is not 100% we will see why not in the following sections or we can use regacc::report\_completeness() with the option verbose = TRUE.



## Basic checks

The next step is to perform some basic checks (additivity, consistency, etc). A day stamped excel file in the folder basic\_checks/ is generated by executing:

# Report in excel. Mind the thresholds inside the file  
source("04\_basic\_checks.R")

We have moved them ahead in the sequence compared to last year as we were using mainly the excel file output instead of the report output. Contrary to the report they should also work on countries not sending particular tables or incomplete tables, However, not all cases might be covered and may not run in all occasions. It will have to be fixed manually if that happens.

Another advantage is that we can easily change the thresholds at the beginning of the script from the default ones if we want.

We can reuse the excel file generated to communicate to the country the problems. We could use the complete file or remove the worksheets we do not consider relevant.

We may need from time to time to update the NAMA and NFSA data used for checking the external consistency as some countries send annual data in November/December (Czechia, Romania).

If a country fails the basic check we should **stop** the analysis of that country. We communicate the problem to the country and move on to the next one. Only if we do not have more countries to deal with, we may continue to the following steps.

### External consistency

The consistency of the reported data with NAMA (tables 1 a 3) and table 8 is checked. There are three different absolute thresholds:

* 0 for table 1001.
* 1 for variables measured in persons and monetary terms in tables 1002,1200 and 1300.
* 100 for variables measured in hours worked in table 1002.

and a single relative threshold of 0.1% in tables 1002, 1200 and 1300. The 2, 100 and 0.2% threshold can be changed at the beginning of the file and the 0 down inside the script.

We have also included some known exceptions in table 1 (NACE activities for Poland) and table 8 (Italy and Poland).

| ref\_area | sto | unit\_measure | activity | time\_period | obs\_value | nama | diff | diffp |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CY | EMP | HW | \_T | 2021 | 760010.135 | 796671.603 | -36661.5 | -4.6 |
| CY | EMP | HW | BTE | 2021 | 73257.435 | 77231.730 | -3974.3 | -5.1 |
| CY | EMP | HW | C | 2021 | 62847.215 | 66744.514 | -3897.3 | -5.8 |
| CY | EMP | HW | F | 2021 | 67143.138 | 73416.523 | -6273.4 | -8.5 |
| CY | EMP | HW | GTI | 2021 | 217037.470 | 231501.134 | -14463.7 | -6.2 |
| CY | EMP | HW | J | 2021 | 27685.953 | 29177.858 | -1491.9 | -5.1 |
| CY | EMP | HW | K | 2021 | 33133.641 | 33540.555 | -406.9 | -1.2 |
| CY | EMP | HW | L | 2021 | 4873.753 | 5034.354 | -160.6 | -3.2 |
| CY | EMP | HW | M\_N | 2021 | 80808.752 | 84980.815 | -4172.1 | -4.9 |
| CY | EMP | HW | OTQ | 2021 | 137199.302 | 139660.067 | -2460.8 | -1.8 |
| CY | EMP | HW | RTU | 2021 | 75097.572 | 78263.628 | -3166.1 | -4.0 |

### NACE aggregation

The same thresholds are used for the NACE aggregation. It uses regacc::check\_NACE().The additivity of data in volume which requires a different calculation is done in a different section.

An improvement implemented this year is to handle the very specific case of extra-regio. The NUTS additivity checks are done with summarise(sum(...,na.rm = TRUE)), that means that NA values are de facto considered zeroes (if all are NA, NA is returned and if there are some NAs they are considered as zero). However NACE additivity checks are done with mutate(). In that case, if there is an NA value the sum is not done. This serve well the case when we have GTJ but not GTI and J in NUTS 3 but Extra-regio is a particular case for some countries.

| ref\_area | sto | unit\_measure | \_T | \_Z | A | BTE | C | F | GTI | GTJ | J | K | KTN | L | M\_N | OTQ | OTU | RTU |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| BEZ | D1 | XDC | 226.2 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | 226.2 | 226.2 | NA |
| BEZ | EMP | HW | 3949.5 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | 3949.5 | 3949.5 | NA |
| BEZ | P51G | XDC | 28.0 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | 28.0 | 28.0 | NA |
| BEZ | SAL | HW | 3949.5 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | 3949.5 | 3949.5 | NA |
| BEZZ | D1 | XDC | 226.2 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | 226.2 | 226.2 | NA |

This will not make any calculation in the report because of the missing values in some columns. In the basic checks we do a different calculation for extra-regio and fill with zeroes the empty values and then we check the consistency.

country\_sel<- "BE"  
temp<- files %>%   
 pivot\_wider(names\_from=activity,  
 values\_from=obs\_value,  
 values\_fill = 0) %>%   
 filter(type=="V" & str\_detect(ref\_area,paste0("^",country\_sel,"Z")) & time\_period== 2019 & table\_identifier %in% c("T1002","T1200")) %>%   
 select(ref\_area,sto,unit\_measure,10:25) %>%   
 head(5)  
  
knitr::kable(temp)

| ref\_area | sto | unit\_measure | \_T | \_Z | A | BTE | C | F | GTI | GTJ | J | K | KTN | L | M\_N | OTQ | OTU | RTU |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| BEZ | D1 | XDC | 226.2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 226.2 | 226.2 | 0 |
| BEZ | EMP | HW | 3949.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3949.5 | 3949.5 | 0 |
| BEZ | P51G | XDC | 28.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 28.0 | 28.0 | 0 |
| BEZ | SAL | HW | 3949.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3949.5 | 3949.5 | 0 |
| BEZZ | D1 | XDC | 226.2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 226.2 | 226.2 | 0 |

### Household Accounts rules

We use three different rules depending on the country.

* B\_B6N = B\_B5N + C\_D61 + C\_D62 + C\_D7 - D\_D5 - D\_D61 - D\_D62 - D\_D7, for **Austria**, **Belgium**, **Cyprus**, **Denmark**, **Estonia**, **Finland**, **Hungary**, **Latvia**, **Malta**, **Poland**, **Portugal**, **Romania** and **Slovakia**.
* B\_B6N = B\_B5N + C\_D62 + C\_D7 - D\_D5 - D\_D61 - D\_D7 for **Belgium**, **Bulgaria**, **Czechia**, **Germany**, **Greece**, **Spain**, **France**, **Ireland**, **Italy**, **Lithuania**, **Luxembourg**, **Netherlands**, **Norway**, **Serbia**, **Sweden** and **Slovenia**.
* B\_B6N = B\_B5N + C\_D61 + C\_D62 + C\_D7 - D\_D5 - D\_D61 - D\_D7 for **Croatia**.

### NUTS aggregation

Works the same way as the NACE aggregation and uses regacc::check\_NUTS().

### Negative values and NACE C < BTE

Check that values are not negative (except for volume change of Value Added). Some exceptions for Value Added and Gross Fixed Capital Formation can be expected (and hopefully documented). It is also more common in NACE C or A.

We found several cases mainly due to rounding issues for the variable self-employed which is not reported by countries but calculated by us. In such cases we should make an effort to show the issue to the countries in and understandable way.

| table\_identifier | ref\_area | NUTS | unit\_measure | time\_period | sto | BTE | C | check |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| T1200 | DK | 0 | PS | 2016 | SELF | 7 | 8 | -1 |
| T1200 | DK0 | 1 | PS | 2014 | SELF | 7 | 8 | -1 |
| T1200 | DK011 | 3 | PS | 2004 | SELF | 0 | 1 | -1 |
| T1200 | DK011 | 3 | PS | 2006 | SELF | 0 | 1 | -1 |
| T1200 | DK011 | 3 | PS | 2007 | SELF | 0 | 1 | -1 |
| T1200 | DK011 | 3 | PS | 2012 | SELF | 0 | 1 | -1 |

### Volume additivity

In order to check the volume additivity, in the past we were deriving previous year prices from growth rates but that was a bit difficult to communicate to countries in case of problems. We have started to use weighted growth rates, which are conceptually equivalent, but easier to communicate.

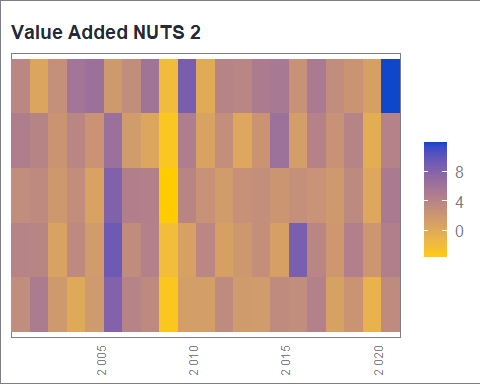
## Report

We have removed many elements of the full report which have been added to basic report. When we arrive at this step we should have a complete transmission that satisfies the minimum requirements and we start a more analytic validation process. The report will provide some, mainly graphical, basic elements to detect in which areas we should, if needed, investigate further. It creates a day stamped html file in report/ when executing.

rmarkdown::render("03\_report.Rmd",   
 params = list(report = country\_sel),  
 output\_file = paste0("report/",  
 country\_sel,"\_",  
 format(Sys.time(),"%Y-%m-%d"),  
 "\_report.html"))

### Heat map charts

Heatmap charts allow us to verify visually the reporting and detect some possible outliers. The heat maps are different for table 1001 than for the other tables. For table 1001 it shows a percentage change compared to the previous year and we should mainly look at the latest period reported.



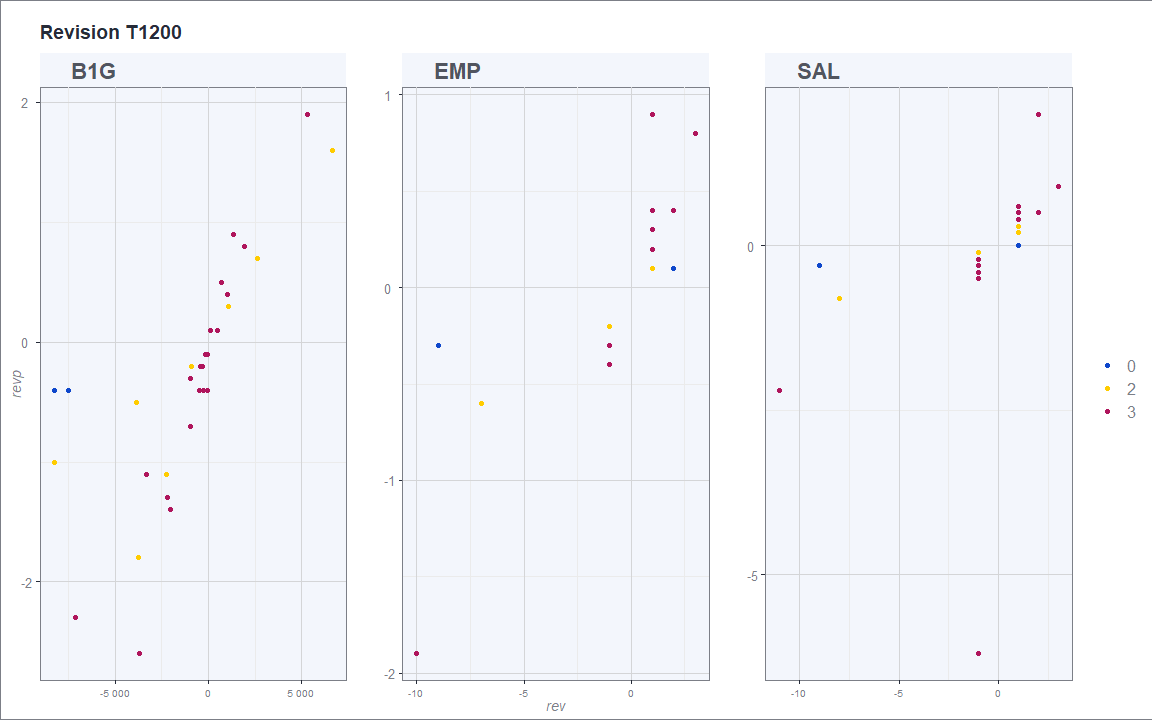
As for the extra regio the amounts are small, growth rates and revisions in percentages are quite big and therefore the extra regio is not included in the charts to avoid distortions in the scales.

while the other tables show the share of the regions in the country.



### Revisions

The report includes and aggregated view of the main revisions. It includes some charts showing the revisions for the totals (NACE TOTALS and B5N and B6N for table 13).



### Outliers

We have already seem that a possible method to detect the existence of outliers when showing the heat maps with the values as % of the national totals. Another one is to see unusual changes (positive or negative) from year to year (more than 30%).

| ref\_area | sto | activity | unit\_measure | time\_period | NUTS | obs\_value | growth | country | label |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DK01 | B1G | GTJ | XDC | 2021 | 2 | 322908 | 30.2 | DK | Hovedstaden |
| DK01 | B1G | GTI | XDC | 2021 | 2 | 252839 | 38.7 | DK | Hovedstaden |
| DK011 | B1G | GTJ | XDC | 2021 | 3 | 186121 | 36.0 | DK | Byen København |
| DK011 | B1G | GTI | XDC | 2021 | 3 | 139317 | 51.3 | DK | Byen København |
| DK02 | B1G | A | XDC | 2021 | 2 | 3086 | 37.1 | DK | Sjælland |
| DK022 | B1G | A | XDC | 2021 | 3 | 2795 | 36.9 | DK | Vest og Sydsjælland |
| DK05 | B1G | A | XDC | 2021 | 2 | 2578 | 56.7 | DK | Nordjylland |
| DK050 | B1G | A | XDC | 2021 | 3 | 2578 | 56.7 | DK | Nordjylland |
| DK021 | B1G | A | XDC | 2021 | 3 | 291 | 38.6 | DK | Østsjælland |

## Revisions

An excel file in the folder revisions/ is created by executing:

source("04\_revision.R")

All revisions are included in the same worksheet while in the past there was a worksheet by table. We should not devote too much time on revisions of normal size as they are a normal feature of updating data. We must anyway be attentive to revisions that go beyond the normal 2-3 years time span. Also revisions in which we observe the exact same percentage across regions should be considered suspicious.

## Auxiliary checks

These is a set of 4 auxiliary checks that produce html reports in others/ with:

source("05\_auxiliary.R")

### D1 place of work/ place of residence

The difference between compiling variables (employment, CoE) according to the place of work or residence is quite important in Regional Accounts and one area when we have found some poor practices in the latest years. It cannot be discarded that we find more as we saw with D.61 in the latest June EG RA meeting.

The script compares D.1 in tables 1002 (place of work) and table 13 (place of residence). We expect to find differences between both and be stable.

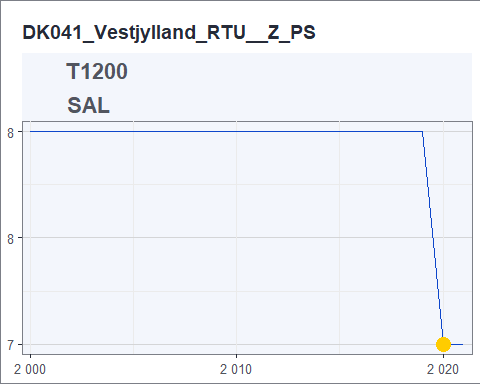
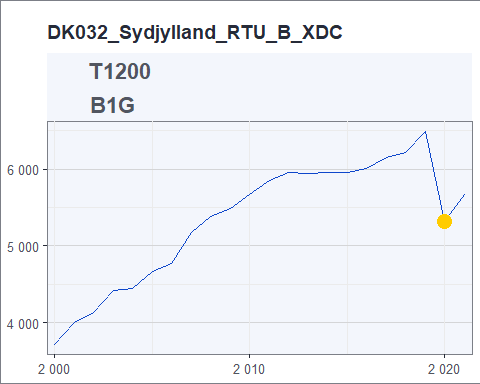
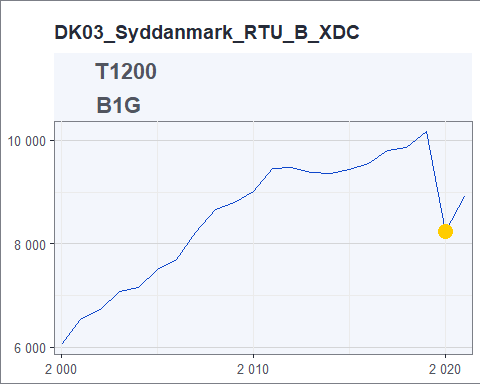
### Outliers

The z\_score argument indicates the threshold we will apply. This year the z\_score is applied to changes (t - (t-1)) instead of levels like in previous years, which should increase the number of detected cases. 3 is a standard value in the literature but if we get many results with 3 we could increase it in the file auxiliary\_scripts.R.

We also restrict the results to the new and revised data and ignore outliers in already reported values.

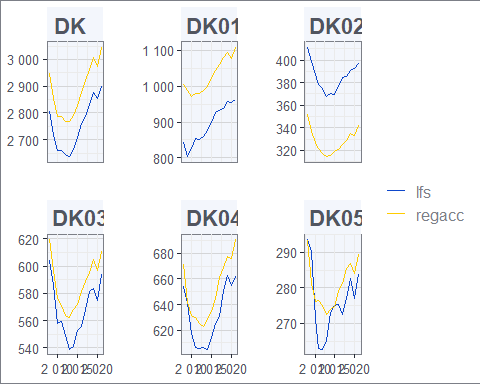
Most of the charts we will get, especially now that we will get years affected by COVID, will not be relevant. The output just restrict to a manageable level the observations we may have to investigate, together with the different approach we saw in the report and another more visual approach we will see on the data explorer.

For example, setting the threshold to 4 (to restrict the output) will output the following.



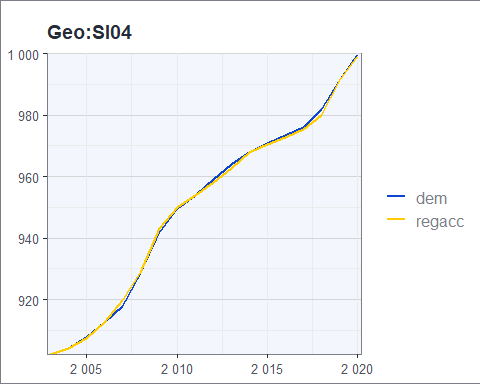
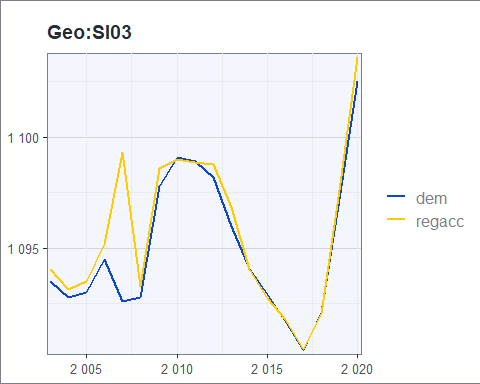
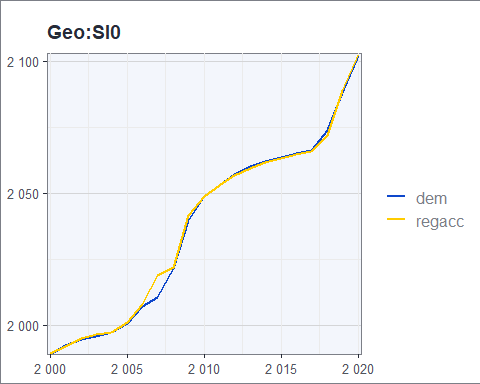
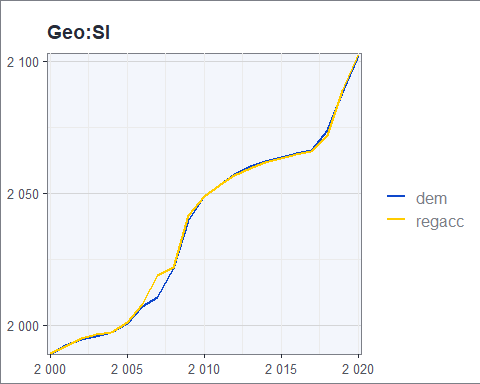
### Labour Force Survey

We usually give a quick look to compare the employment data provided to us to the one of the LFS. There should be a bit or a lot different depending on the regions. If they are the same it would be dig Also if the difference between the two series shows a recent jump we may investigate it.



### Population

We do also a comparison with population data from demography statistics. In most cases there is a perfect match or very small differences. With results of the Census 2021 coming in we should be attentive to see if there are unusual differences in the latest years.



## Visual help and analytical ratios

A set of three shiny apps should help us in investigating further some specific topics (revisions, outliers, last year data) and get some information for some analytical ratios (combination of variables). A brief look at t1001 is recommended to check the plausibility of the latest year and to the indicators that will later show up in publications (D1 per hours worked and person, income per capita, productivity, etc).

It is recommended to use Chrome (or alternatively Firefox) and do not forget to stop the shiny app when finish. Do not forget to run first the code defining the function and only later run the function.

This year we have introduced some improvements in the reactivity for the application with NUTS 3 data (t1002\_1200) and a default scatterplot view show levels and growth rates that can help us with the outlier detection.

shiny\_t1001 <- function(country\_sel) {  
 .GlobalEnv$country\_sel <- country\_sel  
 shiny::runApp("app\_t1001.R", launch.browser = TRUE)  
}  
  
# only after having run the previous  
shiny\_t1001(country\_sel)

# Data for Matis

Once the country data can be validated (do not forget to update the country metadata file if needed) we need to first bring some additional elements to Matis before computing all derived data.

## National GDP

National GDP should be extracted from NAMA in February when we know that EU aggregates are the sum of the countries and stored in the TNAMA table of Matis. For historical reasons they are stored at: *U:/0C\_Folder for shared files/01C\_regional accounts*. In the past we did some extra calculations to center (=make equal) EUR and PPS but now we stick to NAMA data as differences are minimal (see *U:/0C\_Folder for shared files/01C\_regional accounts/2022 exercise - stopped*). We normally load a preliminary version after the PPP release in December as changes in NAMA between December and February are very small.

## Regional GVA in Previous Year Prices

This year in order to calculate GDP volume growth rates we need to load in Matis regional value added in previous year prices. This relies on a data extracted with denodo a day after the data has been validated stored in data/denodo/all\_primary.parquet, national data extracted from Matis and regacc::calculate\_gva\_pyp(). The script ggva\_pyp/gva\_pyp will create a csv file in gva\_pyp/output that can later be loaded in Matis. As the denodo extraction is quite heavy it is better if only one person creates the files.

## Regional GDP for France and Portugal

For Portugal and France we need to load separately the GDP, which is directly provided by Portugal, while for France the GDP of the DOMS is provided and we need to calculate the other french regions. We store the files with the calcuations and loaded in Matis in *U:/03\_Regional Accounts/03P\_Data Releases/Data Releases GDP/XXX*.

# Matis operations

## Validation in Matis

As soon as the validation task is finished (no need to wait for additional data needed for Matis) we can validate the data.

## Computations in Matis

We will only execute the computations when all data needed is available (national data and regional previous year prices). A guide about the different computations in Matis will be prepared in the near future.

## Update Eurobase

This moves the computed and validated table to C tables. At some stage we do for all countries because it is difficult to check which countries tables are ready, what has been done, etc.

## Disseminate to Eurobase

We use it in two ways. First to produce the files needed for checking the output and create summaries for the annual metadata file. In such cases it should only be **extracted** (not send). Some days before the release we send tables in **staging** because it happens quite often that something needs to be fixed. Only the day before the release we send it to **Eurobase**.

# NQR

I am not very familiar with the technical details so Giota will need to explain them.

We provide three different metrics for the NQR.

## Timeliness

We fill an excel file (REGACC\_timeliness) in *U:/03\_Regional Accounts/03D\_Data Production/2022* with the reception and validated (date final file was provided). If done just after finishing a country is not too difficult to do it manually. If we want to do several countries in one go we can use regacc::report\_transmission(). I assume that later we put that file on the fame server and execute a Matis script.

## Completeness

We have seen the completeness and we know that in very few cases (Poland, historical series for France and Belgium) is not 100%. Everything is done in Matis with the commands …

## Revisions

We need to update an excel file with the results and store that file on the fame server and execute a Matis script. We do not do the calculation for NUTS 2 mono-regional countries. We share before the calculations with the countries by e- mail. To create the file we need to update the NQR dataset available in dataregacc with the validated data this year, which will be in data/denodo/all\_primary.parquet and run regacc::report\_NQR\_revision() which will create the file in NQR/. A csv file to update the internal NQR dataset next year is sent to NQR/data.

## Data for consistency with Main aggregates

Another script extracts data for EMP and B1G for the last four years and sends them to the NQR Fame domain in order to compare them with main aggregates.

# Communications with countries

We will contact Member States in case of problems with the data, to provide the NQR results (when applicable) and some weeks before the release to provide the so called confirmation file. The e-mails should be adressed to the names in the metadata file of the country.

## Problems with the data

It is recommended, but sometimes we forgot, to send the e-mail using the functional mailbox and also putting it in cc. It is preferable to use the functional mailbox because sometimes they do not reply to all and just the sender. To keep the mailbox clean processed e-mails should be moved to the corresponding country folder.

Emails should clearly direct the country to the problem. Apart from describing the problem with words we should include as attachment excel files and if appropriate charts.

The best way to acquire the data for excel (if what we want to include does not correspond to what out standard validation tools output) or charts is to use the country csv file.

data<- fread("data/csv/regacc\_DK\_2022-11-17.csv")

The charts should be preferably prepared using the ra\* functions of regacc. A copy of the charts and the code for the charts should be kept in a working folder (no need to copy them on the server).

In case we detect a problem with an indicator we should provide the underlying data that we have used to produce the indicator. The code to create the indicators is located at the very beginning of the shiny apps. For example, this is the code to calculate labour productivity in persons and hours.

# B1G per HW and PS  
B1G <- copy(data) %>%  
 .[sto %in% c("B1G", "EMP", "EMPhw") & activity != "\_Z",] %>%   
 dcast(...~sto, value.var = "obs\_value") %>%   
 .[,B1G\_EMP := round(B1G \* 1000 / EMP, digits = 1)] %>%   
 .[,B1G\_EMPhw := round(B1G \* 1000 / EMPhw, digits = 1)] %>%   
 .[,!c("B1G", "EMPhw", "EMP"),with=FALSE] %>%   
 melt(measure.vars = c("B1G\_EMP", "B1G\_EMPhw"),  
 variable.name = "sto",  
 value.name = "obs\_value") %>%   
 na.omit()

## NQR revision

We can provide the NQR revision file as soon as the data is validated. Normally we will combine both subjects in the e-mail.

## Confirmation file

The confirmation file should be sent some weeks before the release and after the NAMA data is released in mid February because if we would do it earlier we could get in the meantime some updates to population data. The excel file is created with a script in /confimation which uses the denodo extraction data/denodo/all\_primary.parquet.

# Checking eurobase files