# Bookpiracy Data Description

Daniel Antal, CFA

4/14/2020

## Data handling

Let's take a purely hypothetical but easy to understand example for data available in Limousin, France. In this hypothetical data frame, data is coded according to the old region codes till 2016, and in 2017 only for the region larger Aquitaine-Limousin-Poitou-Charentes.

| geo | time | code13 | code16 | change | resolution | nuts_level | name |
|---|---|---|---|---|---|---|---|
| FR63 | 2010-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |
| FR63 | 2011-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |
| FR63 | 2012-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |
| FR63 | 2013-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |
| FR63 | 2014-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |
| FR63 | 2015-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |
| FR63 | 2016-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |
| FRI | 2017-01-01 | FR6 | FRI | recoded | FR6=FRI | 1 | AQUITAINE-LIMOUSIN-POITOU-CHARE |
| FRI2 | 2018-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |
| FRI2 | 2019-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |

In this case, the Limousin region's boundaries did not change, but Limousin got a new NUTS2 code, "FRI". The data for the year 2013 is available in the dataset, but under the earlier code "FR63".

| geo | time | code13 | code16 | change | resolution | nuts_level | name |
|---|---|---|---|---|---|---|---|
| FRI1 | 2010-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |
| FRI1 | 2011-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |
| FRI1 | 2012-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |
| FRI1 | 2013-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |
| FRI1 | 2014-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |
| FRI1 | 2015-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |
| FRI1 | 2016-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |
| FRI | 2017-01-01 | FR6 | FRI | recoded | FR6=FRI | 1 | AQUITAINE-LIMOUSIN-POITOU-CHARE |
| FRI1 | 2018-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |
| FRI1 | 2019-01-01 | FR63 | FRI1 | recoded | FR63=FRI2 | 2 | Limousin |

While the year 2017 is of no interest to our models, for simpler demonstration we remain with this example. In this case, we do not have actual data for Limousine (FRI1), but we have data for the NUTS1 level larger region Aquitaine-Limousin-Poitou-Charentes.

In reality, this problem is unlikely to present itself for one region. In the case of some statistics based on Eurobarometer and other relatively small sample surveys, for the larger member states the statistics is only calculated at NUTS1 (larger region) level. If the larger region statistic is an unknown weighted averages of

the smaller constituent NUTS2 regions, we can safely impute the larger regions's value to the smaller regions.

In this hypothetical example, the value of Aquitaine-Limousin-Poitou-Charentes is in fact an average value of Aquitaine, Limousin, Poitou and Charentes.

| geo | time | code13 | code16 | nuts_level | name | values | method |
|---|---|---|---|---|---|---|---|
| FRI1 | 2010-01-01 | FR63 | FRI1 | 2 | Limousine | NA | missing |
| FRI1 | 2011-01-01 | FR63 | FRI1 | 2 | Limousine | NA | missing |
| FRI1 | 2012-01-01 | FR63 | FRI1 | 2 | Limousine | NA | missing |
| FRI1 | 2013-01-01 | FR63 | FRI1 | 2 | Limousine | 51 | actual |
| FRI1 | 2014-01-01 | FR63 | FRI1 | 2 | Limousine | 52 | actual |
| FRI1 | 2015-01-01 | FR63 | FRI1 | 2 | Limousine | 55 | actual |
| FRI1 | 2016-01-01 | FR63 | FRI1 | 2 | Limousine | NA | missing |
| FRI1 | 2017-01-01 | FR63 | FRI1 | 1 | Limousine | 57 | imputed from NUTS1 actual |
| FRI1 | 2018-01-01 | FR63 | FRI1 | 2 | Limousine | 56 | actual |
| FRI1 | 2019-01-01 | FR63 | FRI1 | 2 | Limousine | NA | actual |

We have used the zoo package to approximate the regional time series. In the case of (linear) interpolation, the zoo package refers back to the basic stats::approx function. We used zoo because of its better interface for programatic use.

```
## values = NA,NA,NA,51,52,55,NA,57,56,NA
```

```
## approximated = NA,NA,NA,51,52,55,56,57,56,NA
```

```
## nocb = 51,51,51,51,52,55,56,57,56,NA
```

```
## locf = 51,51,51,51,52,55,56,57,56,56
```

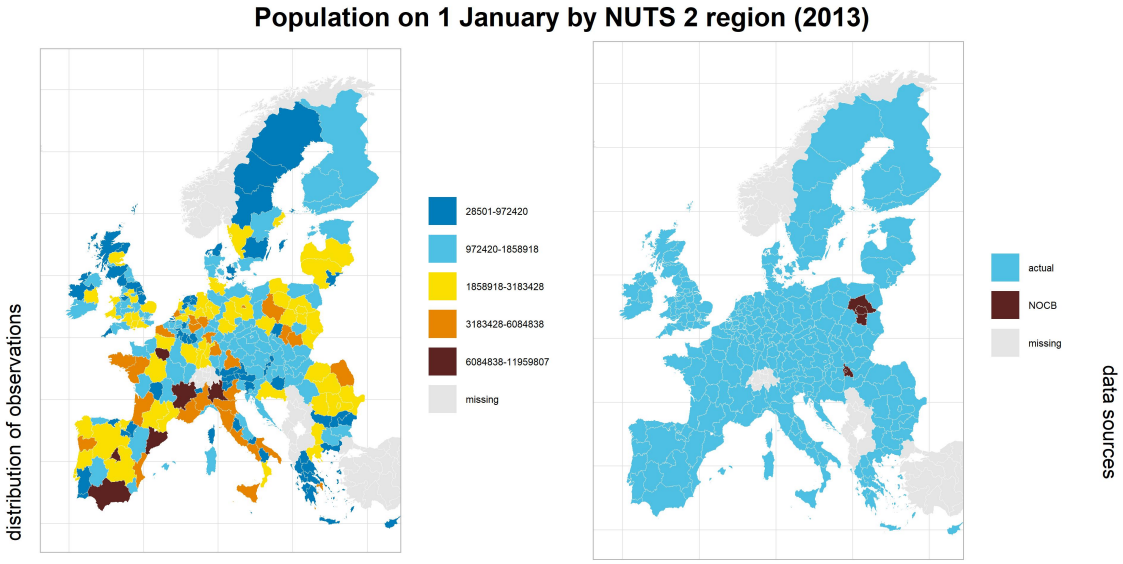
To sum up, these are the actual changes in the hypothetical example.

| geo | time | code13 | code16 | name | values | method |
|---|---|---|---|---|---|---|
| FRI1 | 2010-01-01 | FR63 | FRI1 | Limousine | 51 | nocb |
| FRI1 | 2011-01-01 | FR63 | FRI1 | Limousine | 51 | nocb |
| FRI1 | 2012-01-01 | FR63 | FRI1 | Limousine | 51 | nocb |
| FRI1 | 2013-01-01 | FR63 | FRI1 | Limousine | 51 | actual |
| FRI1 | 2014-01-01 | FR63 | FRI1 | Limousine | 52 | actual |
| FRI1 | 2015-01-01 | FR63 | FRI1 | Limousine | 55 | actual |
| FRI1 | 2016-01-01 | FR63 | FRI1 | Limousine | 56 | imputed from NUTS1 actual |
| FRI1 | 2017-01-01 | FR63 | FRI1 | Limousine | 57 | actual |
| FRI1 | 2018-01-01 | FR63 | FRI1 | Limousine | 56 | actual |
| FRI1 | 2019-01-01 | FR63 | FRI1 | Limousine | 56 | locf |

## Demography

Our data is environmental data and not microdata. We do not know who downloaded books or journal articles, therefore we cannot directly search a relationship between demographic variables, and we do not even know how many individuals used the same IP address to download books. However, we know that environments with more people are more likely to attract downloads, because, apart from robots that we tried to exclude from the dataset, the demand for books is driven by people.

We have found that more people are likely to attract more downloads, which is a very trivial relationship. To overcome this trivial affect, we normalized the download count data with various demographic variables. We started with the total population of the regions. This data is available for almost all NUTS2 regions, with the exception of newly created regions in Hungary and Poland. In these few cases we used the first available population data for the new regions, in other words, we used the 'next observation carry backwards' filling algorithm.

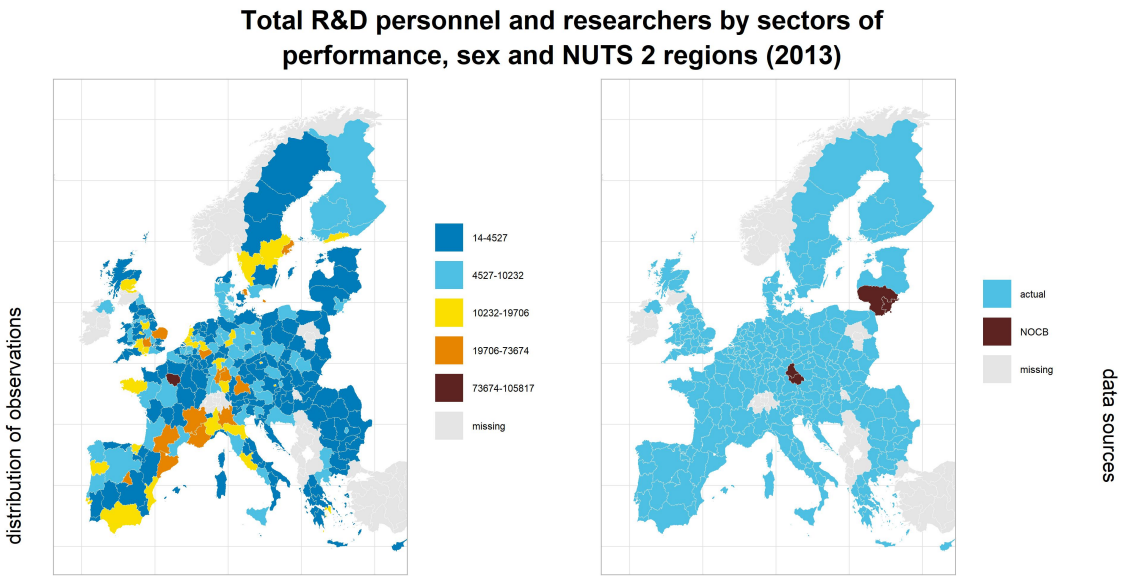



| | |
|---|---|
| title | Population on 1 January by NUTS 2 region |
| unit name | number |
| last update of data | 2020-02-20 |
| retrieve date | 2020-04-11 |
| time | 2013-01-01 |

| data source | observations |
|---|---|
| actual | 276 |
| nocb actual | 5 |
| locf actual | 0 |
| interpolated | 0 |
| imputed from nuts1 actual | 0 |
| nocb imputed from nuts1 actual | 0 |

### Researchers

Normalizing with the total population gives an equal chance for attributing downloads to uneducated old people, newborn babies, and adult scientific researchers. In order to try to get a better picture, we also normalized download count data with the researcher population of the regions.




| | |
|---|---|
| title | Total R&D personnel and researchers by sectors of performance, sex and NUTS 2 regions |
| unit name | full-time equivalent (fte) |
| last update of data | NA |
| retrieve date | NA |
| time | 2013-01-01 |

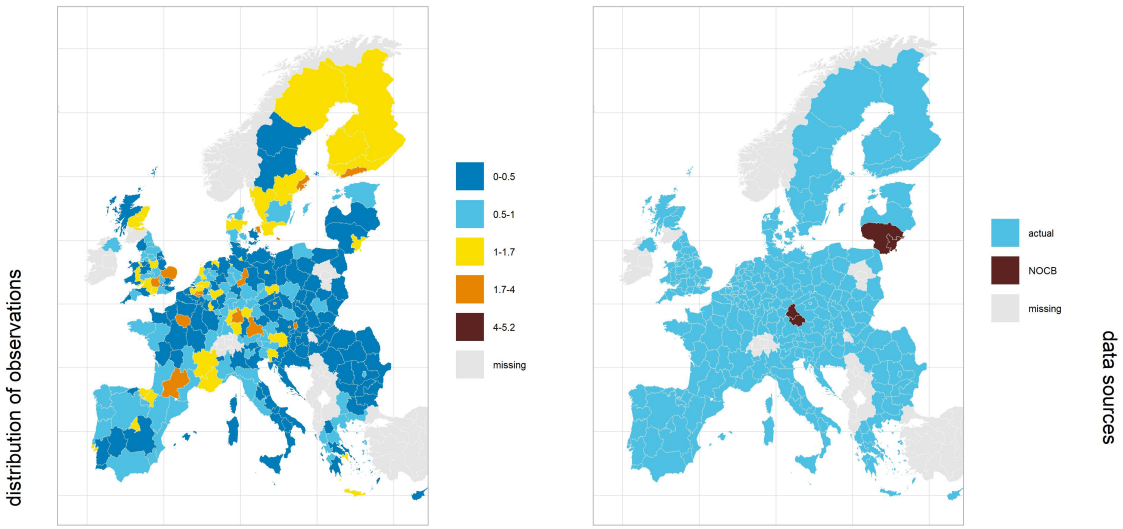
| data source | observations |
|---|---|
| actual | 265 |
| nocb actual | 4 |
| locf actual | 0 |
| interpolated | 0 |
| imputed from nuts1 actual | 0 |
| nocb imputed from nuts1 actual | 0 |

This statistic is available with many indicators in the Eurostat data warehouse, for example, in some regions we can find the total researchers employed in the business sector, the non-profit sector and by the government. However, such interesting breakdown is not reported by all countries, so we used the most available headline indicator that contains all researchers regardless of sector and sex.

### Researchers Alternative

An alternative indicator is the full-time equivalent version of researchers in all sectors and all sexes. Given that we work with environmental data, correcting for part-time workers is very unlikely to add more detail to or analysis. Because this indicator is very highly correlated with the alternative researcher count, both indicators should not be used paralel in our models.

## Researchers, all sectors by NUTS 2 regions (2013)




| | |
|---|---|
| title | Researchers, all sectors by NUTS 2 regions |
| unit name | percentage of total employment - numerator in full-time equivalent (fte) |
| last update of data | 2020-03-18 |
| retrieve date | 2020-04-11 |
| time | 2013-01-01 |

| data source | observations |
|---|---|
| actual | 263 |
| nocb actual | 4 |
| locf actual | 0 |
| interpolated | 0 |
| imputed from nuts1 actual | 0 |
| nocb imputed from nuts1 actual | 0 |

## Internet Environment

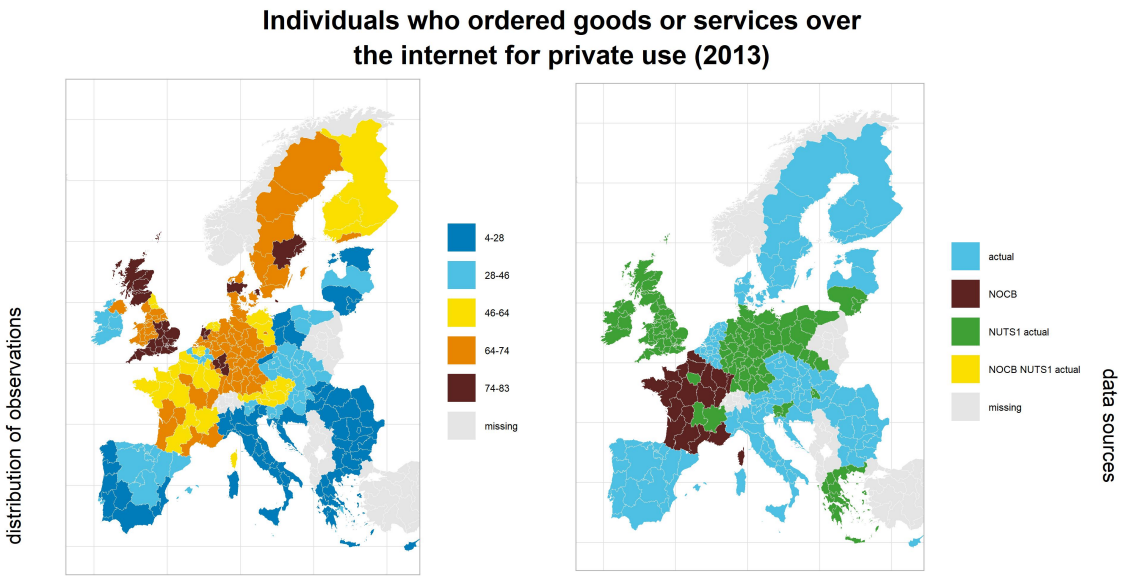
Book piracy is an online activity, so we naturally sought a relationship with the territorial differences of internet availability and internet use. Internet use data is available from various pan-European surveys, which are not made with large enough sample sizes to cover all NUTS2 regions of Europe. Because of the limited national sample size and the sample selection method, for larger countries, like in the case of France and Germany, Eurostat only publishes the data on the higher NUTS1 level. The United Kingdom has a mixed surveying, which results in more detailed data for Northern Ireland than for Great Britain, i.e. England, Scotland and Wales.

For a NUTS2 level analysis, we imputed the NUTS1 level data to the NUTS2 regions within each larger NUTS1 region. This imputation is valid for the statistic, because it is a territorial average value of survey responses. The imputation simply uses the larger area average values for the smaller areas contained in the broader region.

Projecting the NUTS1 level data to NUTS2 level data will not increase the amount of information present in our models. If we combine this data with other, more granular NUTS2 level information, than the information content of our models may increase. Another approach is to use NUTS1 level data for all variables, in effect, discarding the NUTS2 regions of these countries (country parts) and adding NUTS1 regions into the analysis. This is a valid approach, too. Given that NUTS1-NUTS2 level regions do not have a strictly set size and homogeneity criterion, there is a relatively wide level of variety in the different NUTS2 level environments. The mixed use of NUTS1 and NUTS2 level information would be problematic with indicators of dispersion, which are sensitive to aggregation level, or the number of observations that are considered. (For example, the standard deviation of income is likely to be different in a larger NUTS1 region than its constituent NUTS2 regions.) This problem is not present with survey response averages or total summaries of gross domestic product.

### Individuals who ordered goods or services

Individuals who ordered goods or services is a very important environmental variable, because it shows the environments ability to conduct online transactions, and to contact legal, purchasing transactions. We can test various hypothesis regarding environments that have a less developed internet purchasing environment; we can test if unpaid, illegal transactions are likely to substitute for paid transactions, or if the general inability to make purchases may be caused by a general inability for functional internet use.



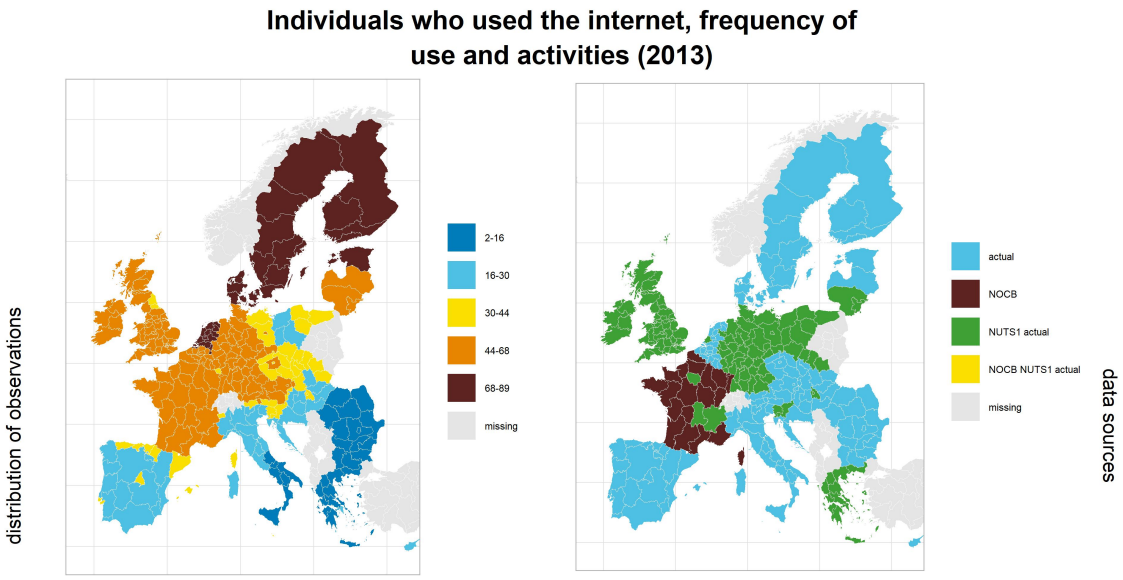

| | |
|---|---|
| title | Individuals who ordered goods or services over the internet for private use |
| unit name | percentage of individuals |
| last update of data | 2020-04-08 |
| retrieve date | 2020-04-11 |
| time | 2013-01-01 |

| data source | observations |
|---|---|
| actual | 135 |
| nocb actual | 19 |
| locf actual | 0 |
| interpolated | 0 |
| imputed from nuts1 actual | 114 |
| nocb imputed from nuts1 actual | 1 |

In this case NUTS1 level data was used in the case of Lithuania and Greece, which had changes in smaller NUTS2 regional boundaries, and the data for the year 2013 is not available for the old smaller regions.

### Frequency of Internet Use & Activities

The frequency of use and activities statistics are many indicators that are available for daily internet users, less frequent users, and people who had not used the internet for at least a year. There are various activities that are measured for these population groups, for example, the use of social networks or internet banking.




| | |
|---|---|
| title | Individuals who used the internet, frequency of use and activities |
| unit name | percentage of individuals |
| last update of data | 2020-04-08 |
| retrieve date | 2020-04-11 |
| time | 2013-01-01 |

| data source | observations |
|---|---|
| actual | 132 |
| nocb actual | 19 |
| locf actual | 0 |
| interpolated | 0 |
| imputed from nuts1 actual | 117 |
| nocb imputed from nuts1 actual | 1 |

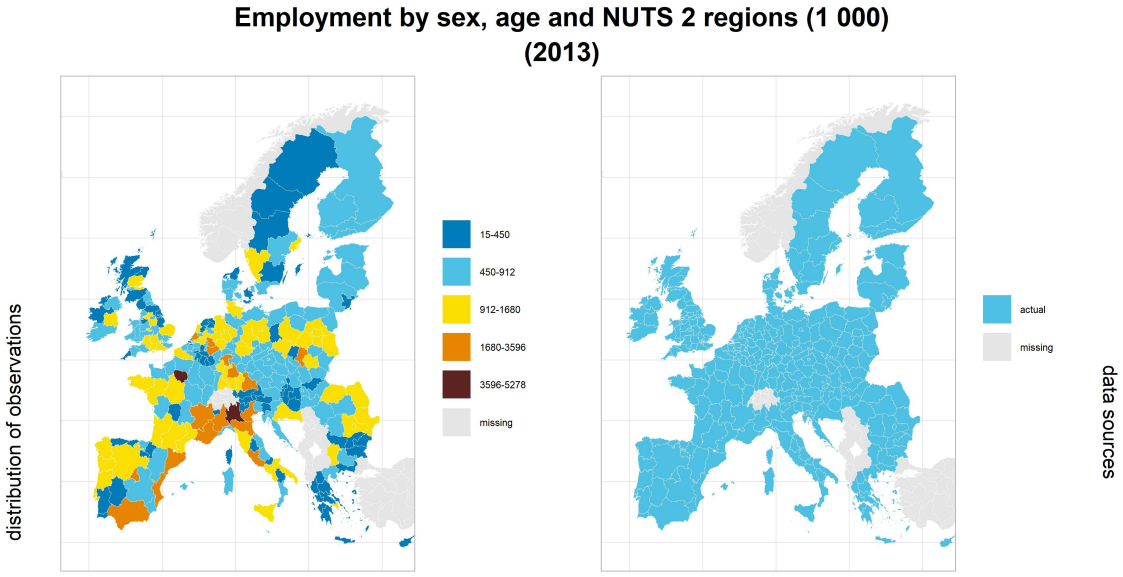
Internet banking use is an important environmental variable, because it shows both the ability to participate in legal purchasing transactions, and the social skills to use the internet for more complex activities that need more trust. The use of social networks is also important, because book torrent sites live in the shadowy parts of the internet, and information on their availability is mainly communicated via personal networking channels, and not, for example, via mainstream internet portals or via internet advertisement.

## Economic Environment

The economic environment proved to be a very important indicator in determining illegal book downloading intensities, though it is impossible to give a clear casual link between the gross domestic product, disposable income and download counts. Richer regions have higher and better employment conditions, and they can usually support more research and development activities. Researchers are themselves are more concentrated in richer regions. Rich regions also tend to have a better research and internet infrastructure, and a more skilled internet user population, given that people in rich regions had a far longer history of using the internet on a daily basis than in poor regions.

### Employment

Employment is related to the GDP, because a large part of the regional GDP is the regional income of employees in the region. It is also a demographic variable. Higher level of employment usually correlates with a more active population, and a higher likelihood of research and development activities.

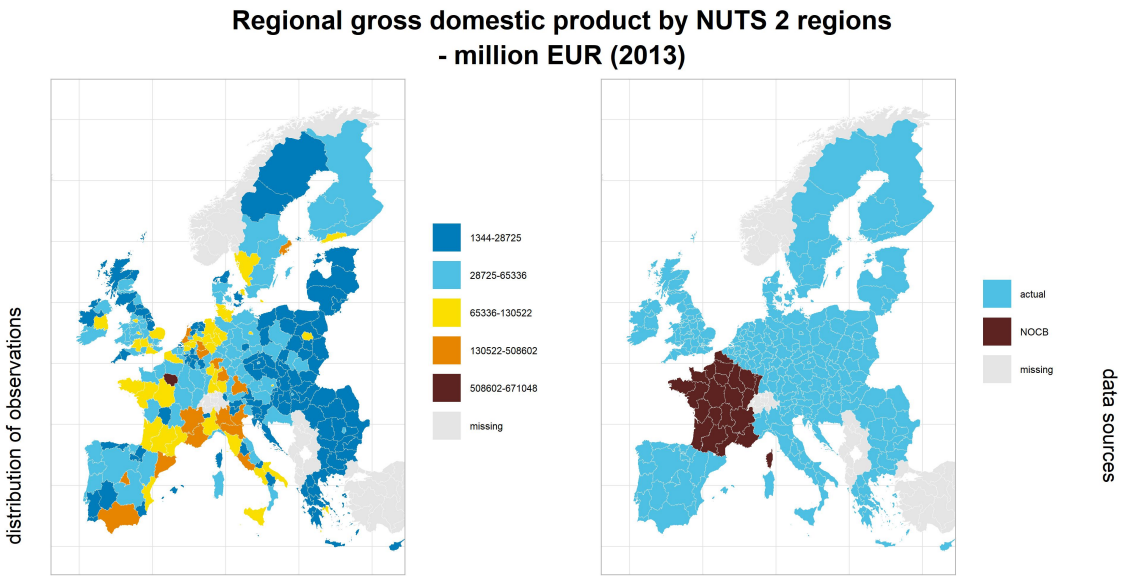



| | |
|---|---|
| title | Employment by sex, age and NUTS 2 regions (1 000) |
| unit name | thousand |
| last update of data | NA |
| retrieve date | NA |
| time | 2013-01-01 |

| data source | observations |
|---|---|
| actual | 281 |
| nocb actual | 0 |
| locf actual | 0 |
| interpolated | 0 |
| imputed from nuts1 actual | 0 |
| nocb imputed from nuts1 actual | 0 |

### Regional Gross Domestic Product in Million Euros

The regional GDP is a very strong environmental variable, because it effects a region's ability to support research institutions, educational institutions, research jobs. Very poor regions are also losing active age population and therefore the potential base of illegal book downloaders.

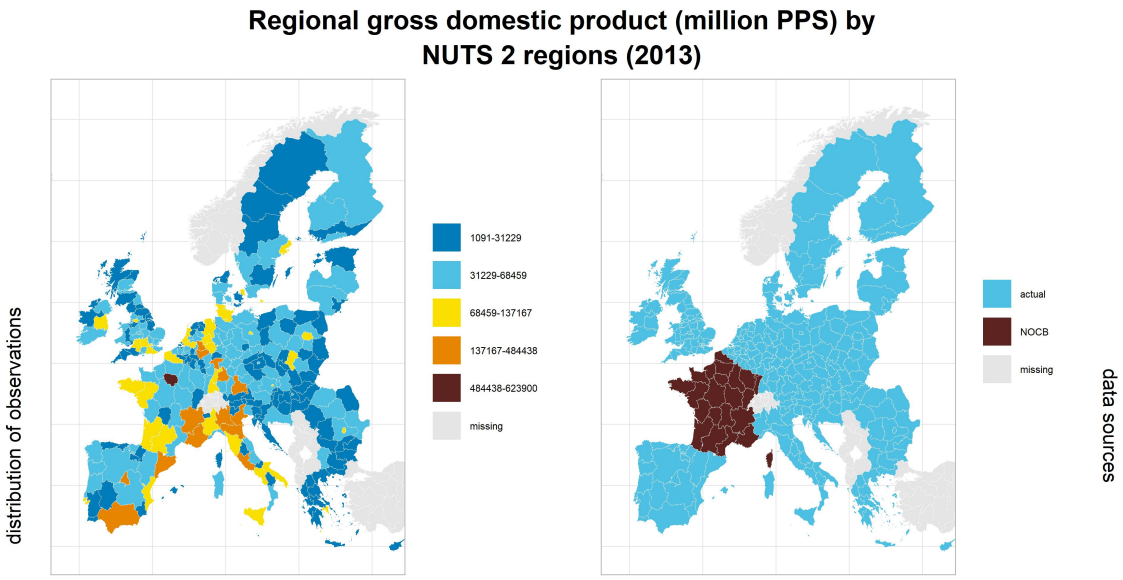



| | |
|---|---|
| title | Regional gross domestic product by NUTS 2 regions - million EUR |
| unit name | million euro |
| last update of data | 2020-03-23 |
| retrieve date | 2020-04-11 |
| time | 2013-01-01 |

| data source | observations |
|---|---|
| actual | 254 |
| nocb actual | 27 |
| locf actual | 0 |
| interpolated | 0 |
| imputed from nuts1 actual | 0 |
| nocb imputed from nuts1 actual | 0 |

### Regional Gross Domestic Product in Million PPS

The regional GDP is not only available in million euros, but also in PPS values, which are price-adjusted versions of the GDP indicator on purchasing power standards.

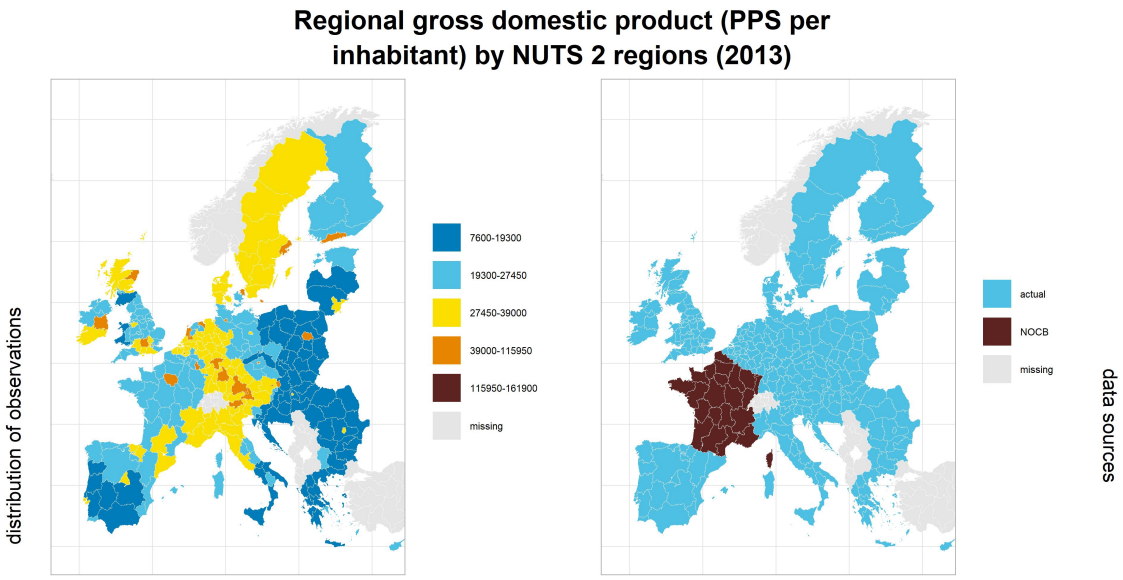



| | |
|---|---|
| title | Regional gross domestic product (million PPS) by NUTS 2 regions |
| unit name | million purchasing power standards (pps) |
| last update of data | 2020-03-23 |
| retrieve date | 2020-04-11 |
| time | 2013-01-01 |

| data source | observations |
|---|---|
| actual | 254 |
| nocb actual | 27 |
| locf actual | 0 |
| interpolated | 0 |
| imputed from nuts1 actual | 0 |
| nocb imputed from nuts1 actual | 0 |

### Regional Gross Domestic Product in Million PPS Per Capita

The regional GDP is not only available in absolute million euros and PPS, but on a per capita basis, too. Depending on the model type we use, this formulation may be more or less handy to use.



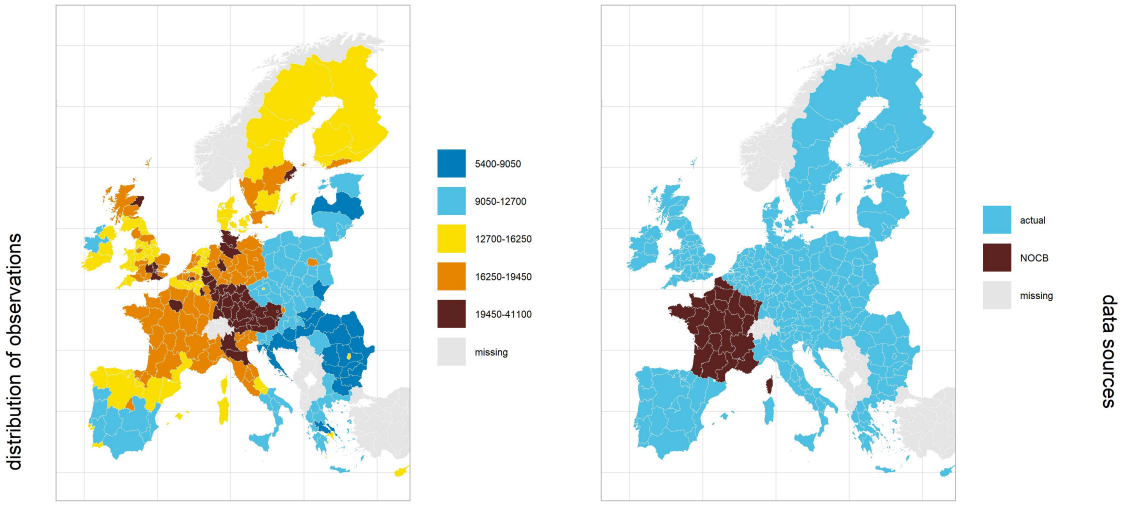

| | |
|---|---|
| title | Regional gross domestic product (PPS per inhabitant) by NUTS 2 regions |
| unit name | purchasing power standard (pps) per inhabitant |
| last update of data | 2020-03-23 |
| retrieve date | 2020-04-11 |
| time | 2013-01-01 |

| data source | observations |
|---|---|
| actual | 254 |
| nocb actual | 27 |
| locf actual | 0 |
| interpolated | 0 |
| imputed from nuts1 actual | 0 |
| nocb imputed from nuts1 actual | 0 |

### Disposable Income of Private Households

The disposable income of private households is related to their ability to purchase computers and other devices, subscribe to internet access, and the ability to purchase books via legal channels. However, because disposable income is a part of employee compensation in the region, which is an important part of the GDP aggregate, this variable is very highly correlated with the GDP variables, and may not be used together with those indicators in the same models.

## Disposable income of private households by NUTS 2 regions (2013)




| | |
|---|---|
| title | Disposable income of private households by NUTS 2 regions |
| unit name | purchasing power standard (pps) per inhabitant |
| last update of data | 2020-03-23 |
| retrieve date | 2020-04-11 |
| time | 2013-01-01 |

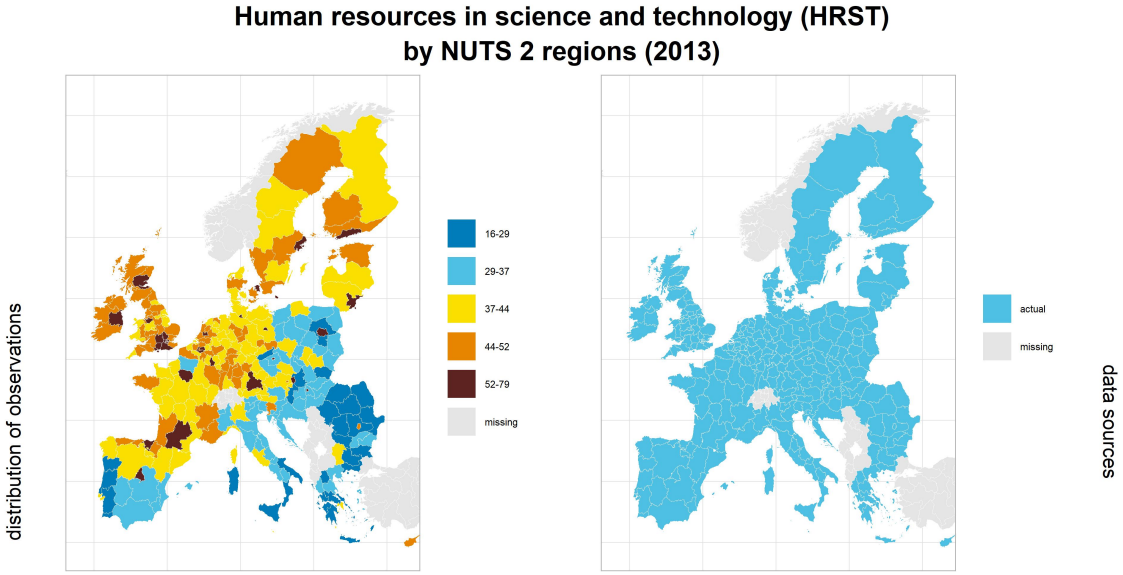
| data source | observations |
|---|---|
| actual | 253 |
| nocb actual | 27 |
| locf actual | 0 |
| interpolated | 0 |
| imputed from nuts1 actual | 0 |
| nocb imputed from nuts1 actual | 0 |

## Research Environment And Education Level

The book torrent websites are mainly focusing on scientific literature, and we believe that environments with a higher research and development output are more likely to attract downloads.

### Human Resources in Science & Technology

The human resources of science and technology are indicators that may add further detail to our main demographic variable, the research count. The statistic is available for many subpopulations, but we do not use sectoral or sex breakups, because on an environmental level they do not add more useful detail to our analysis.

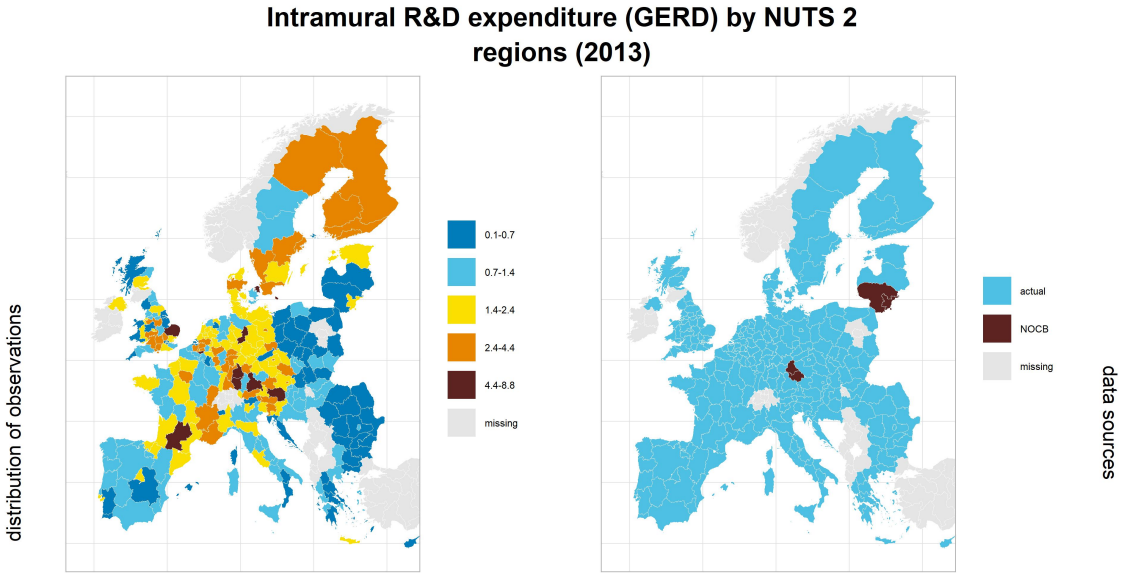



| | |
|---|---|
| title | Human resources in science and technology (HRST) by NUTS 2 regions |
| unit name | percentage of active population |
| last update of data | 2020-03-26 |
| retrieve date | 2020-04-11 |
| time | 2013-01-01 |

| data source | observations |
|---|---|
| actual | 281 |
| nocb actual | 0 |
| locf actual | 0 |
| interpolated | 0 |
| imputed from nuts1 actual | 0 |
| nocb imputed from nuts1 actual | 0 |

### Intramural R&D Expenditure

Intramural R&D Expenditure is related to the regional GDP, because it is a small but not insignificant part of it, but it is a more focused indicator for R&D activities. It may add detail to our models if illegal downloads are not related to the general regional economy but to R&D activities.



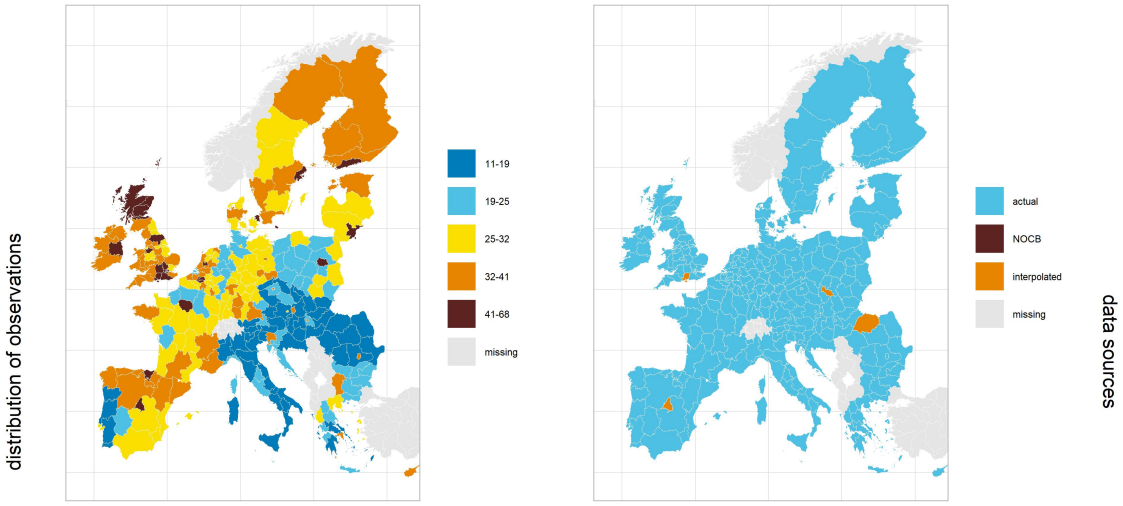

| | |
|---|---|
| title | Intramural R&D expenditure (GERD) by NUTS 2 regions |
| unit name | percentage of gross domestic product (gdp) |
| last update of data | 2020-03-18 |
| retrieve date | 2020-04-11 |
| time | 2013-01-01 |

| data source | observations |
|---|---|
| actual | 263 |
| nocb actual | 4 |
| locf actual | 0 |
| interpolated | 0 |
| imputed from nuts1 actual | 0 |
| nocb imputed from nuts1 actual | 0 |

### Tertiary Educational Attainment (age group 25-64)

Tertiary Educational Attainment is an auxilliary variable to researchers. We believe that regions where the people have higher educational levels read more, and are more likely to read in foreign languages. The most widely used book language in the torrent sites was English, followed by Russian, which are foreign language in most European regions. Regions with a more educated population attract more downloads.

## Tertiary educational attainment, age group 25-64 by sex and NUTS 2 regions (2013)




| | |
|---|---|
| title | Tertiary educational attainment, age group 25-64 by sex and NUTS 2 regions |
| unit name | percentage |
| last update of data | 2020-01-31 |
| retrieve date | 2020-04-11 |
| time | 2013-01-01 |

| data source | observations |
|---|---|
| actual | 272 |
| nocb actual | 4 |
| locf actual | 0 |
| interpolated | 4 |
| imputed from nuts1 actual | 0 |
| nocb imputed from nuts1 actual | 0 |