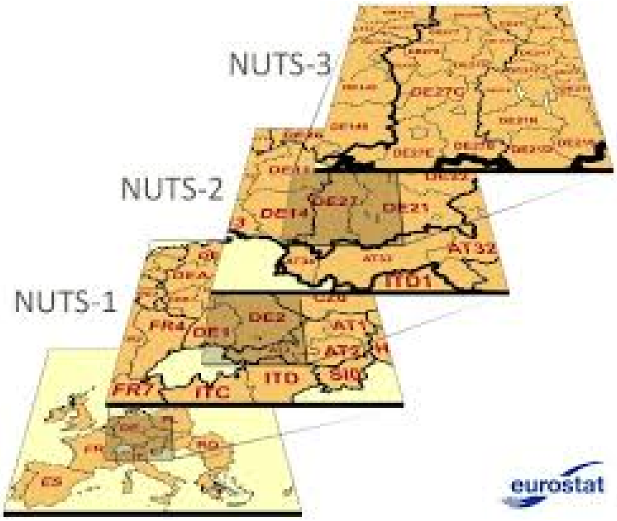
NUTS Geocoding
– project report –

July 2, 2018



Christine Rimmert

1 The goal of the project

Geographical information concerning publications with a lower level of aggregation than just the country level is a topic of interest in many projects as more detailed information is needed. In many cases, however, the city level is also not appropriate as this is too finely grained. For EU countries, geographical evaluations are frequently based on NUTS ('Nomenclature of Territorial Units for Statistics')[1] codes, so there is a need for the assignment of author addresses (and therefore publications) from the Web of Science (WoS) to NUTS codes from Eurostat[2].

WoS addresses normally provide a city and a country attribute, but no further geographical information or aggregation. Country and city information is provided as a string, not compulsory sufficiently standardized - a preceding project was concerned with the standardization of the geographical information on the community level for Germany[3].

The aim of this project is the assignment of author addresses to NUTS codes. This was done for WoS addresses but the procedure developed here may be applied to other data sources as well (on the condition of having similar structured address strings and attributes).

The project was conducted in the context of the German Competence Centre for Bibliometrics[4] in 2017/2018.

[1] https://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistics

[2] http://ec.europa.eu/eurostat/home

[3] Rimmert C, Winterhager M. Geokodierung von Autorenadressen in Publikationsdatenbanken. Abschlussbericht einer Untersuchung für das Kompetenzzentrum Bibliometrie. Bielefeld: Universität Bielefeld, Institute for Interdisciplinary Studies of Science (I^2SoS); 2017 (https://pub.uni-bielefeld.de/publication/2909586).

[4] http://www.forschungsinfo.de/Bibliometrie/en/index.php?id=home

2 Data & methods

2.1 Data

2.1.1 NUTS codes

NUTS codes are a standard for referencing geographical units on different aggregation levels within countries of the EU, provided by Eurostat. For each country, there are three aggregation levels where NUTS1 is the highest aggregation (major socio-economic regions) followed by NUTS2 (basic regions) and NUTS3 (small regions).
E.g., for Germany, NUTS1 represent states (Bundesländer), NUTS2 Government regions (Regierungsbezirke) and NUTS3 districts (Kreise). Below NUTS3 there are two further levels of local administrative units called LAUs ('LAU1' and 'LAU2', 'NUTS4' and 'NUTS5' until 2003).

Example:

- NUTS1 = DEA: Nordrhein-Westfalen (state)
- NUTS2 = DEA4: Regierungsbezirk Detmold (region)
- NUTS3 = DEA44: Kreis Höxter (district)

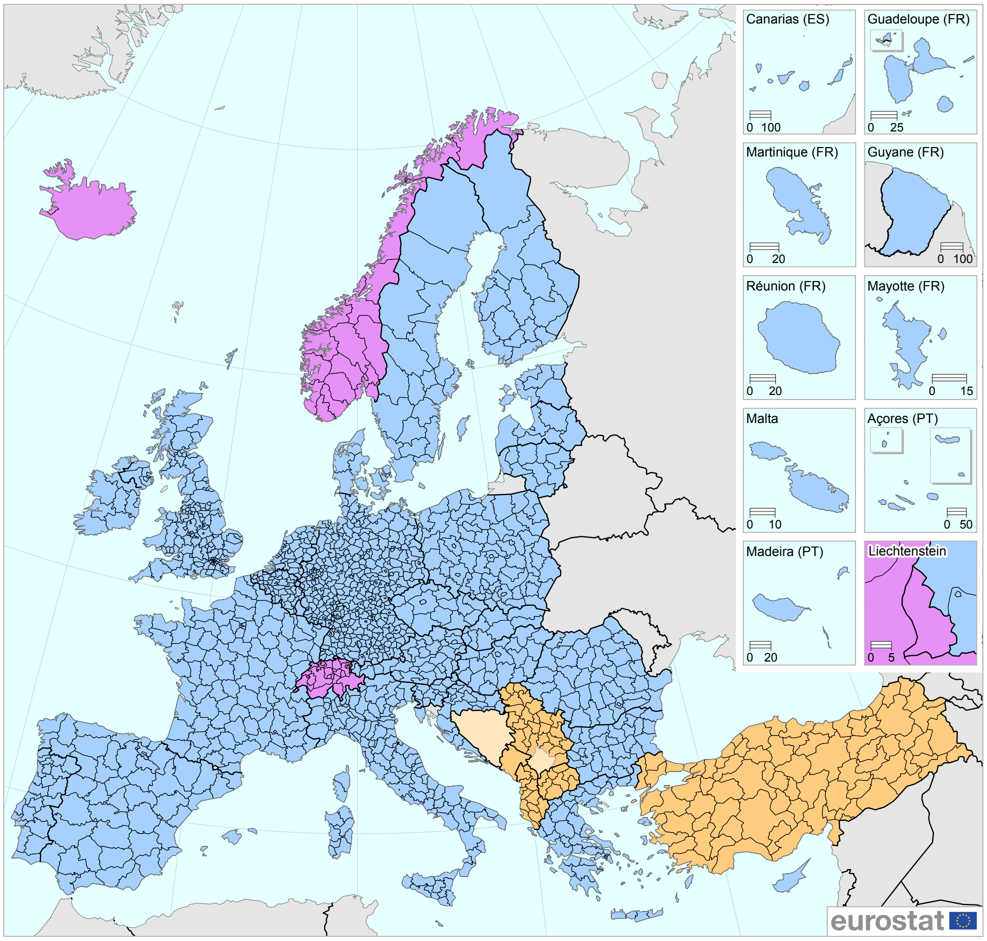
As visible in the example, higher aggregation levels (NUTS2 and NUTS1) can be obtained easily by using substrings of the NUTS3 code. Therefore, the assignment to NUTS3 codes provides all aggregation options. Figure 1 shows an overview of NUTS3 regions[5].

2.1.2 WoS address data

Concerning address data, author addresses from licenced WoS raw data were used here. Addresses are given as full address strings and, in addition, e.g., country, city, postal code and the address part defining the organization are provided in string format as separate attributes. These strings are preprocessed (e.g., removing of special characters, a kind of stemming) but not yet standardized.

[5] Figure from Eurostat: http://ec.europa.eu/eurostat/documents/345175/7451602/NUTS3-2013-EN.pdf (last visited 06.06.2018) Maps and lists for single countries are available on this website as well.

NUTS 3 regions in the European Union (EU-28), with corresponding statistical regions in EFTA countries, candidate countries and potential candidates



🟦 **Member States of the European Union (EU-28)**

🟪 **EFTA countries**

🟧 **Candidate countries**

🟨 **Potential candidates**

Administrative boundaries: © EuroGeographics © UN-FAO © INSTAT © Turkstat

Cartography: Eurostat - GISCO, 03/2018

0 200 400 600 800 km

Note: Regions in the Member States of the European Union (EU-28) according to NUTS 2013. Statistical regions in EFTA countries, candidate countries and potential candidates according to latest available bilateral agreement. The designation of Kosovo is without prejudice to positions on status, and is in line with UNSCR 1244/1999 and the ICJ Opinion on the Kosovo Declaration of Independence.

Figure 1: NUTS3 regions.

2.2 Matching Methods

The assignment of author addresses to NUTS3 codes is a typical classification task with addresses as input data and NUTS3 codes as targets: addresses have to be matched to the corresponding NUTS3 codes.
In order to get results as complete and reliable as possible, several single matching methods were applied, flowing into a procedure using a combination of different matching methods verifying and supplementing each other.

Figure 2 provides an overview of the different matching methods checked for inclusion. At the bottom, WoS attributes of addresses are displayed in orange: country, city, postalcode and organization information – this is the input for the classification task. The NUTS codes on top, provided by Eurostat, form the classification targets.
Below the NUTS, also displayed in dark green, other data sets provided by Eurostat and useful for the classification task, are presented: there is an assignment of postal codes to NUTS codes, an assignment of LAU (representing lower aggregation levels than NUTS3, e.g. cities) to NUTS codes and Eurostat also provides shapefiles for NUTS codes.

Each of these additional data sets/assignment tables requires different input: for using the postal-code-to-NUTS assignment table one has to know postal codes while for the LAU matching, city and country attributes may deliver an appropriate input.
For using the shapefiles, input in the form of geographical coordinates is needed in order to decide if a certain point lies in a polygon (the calculation of this was done with a python library here). As geographical coordinates are not directly contained in the WoS address data, an assignment of WoS attributes to geographical coordinates is needed. For this, three different ways were applied: one possibility is using an existent geocoder – in this case, the OpenStreetMap (OSM) API[6] was chosen as it is freely available.
Another approach of gaining geographical coordinates from city names is using raw data from GeoNames[7] which provides tables with city names and respective geographical coordinates.
Also wikidata[8] may serve as a supplier for geographical coordinates: for many

[6] https://wiki.openstreetmap.org/wiki/API

[7] http://www.geonames.org/

[8] https://www.wikidata.org/wiki/Wikidata:Main_Page

entities, geographical coordinates are provided. Wikidata was used in two different ways here: on the one hand, the WoS city string was matched to wikidata labels – resulting in geographical coordinates of the city in case of existence – and, on the other hand, the WoS organization string (substring until the first comma, usually containing information on the organization in WoS) was matched to wikidata labels – resulting in geographical coordinates for the respective organization in case of existence.

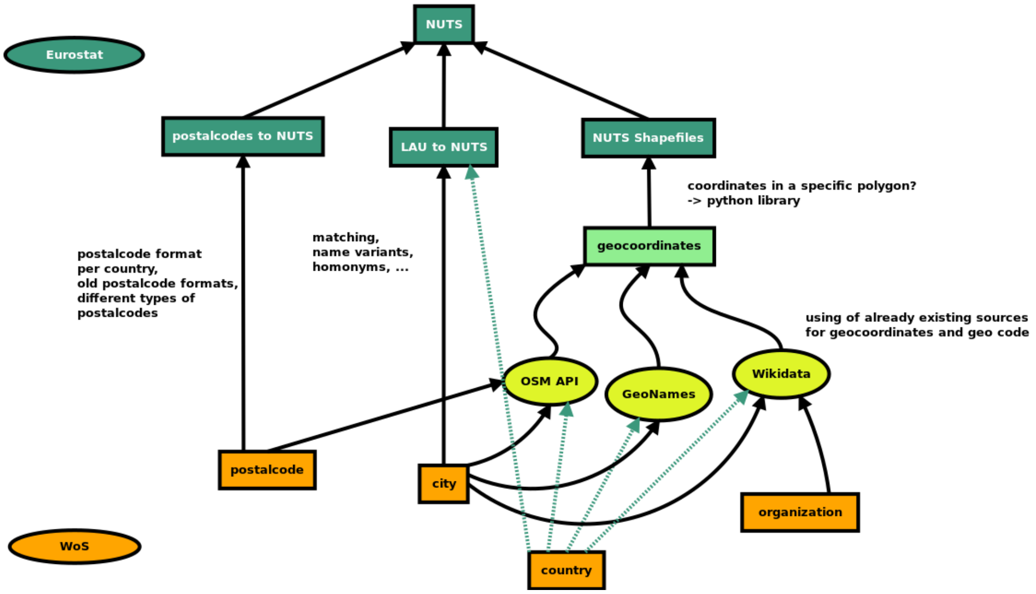


Figure 2: Overview: matching methods.

In the following each matching method is described in more detail with the corresponding advantages, disadvantages/problems concerning WoS address data.

2.2.1 Postal codes to NUTS

Eurostat provides matching tables with postal codes and belonging NUTS3 codes for NUTS-2010 and NUTS-2013[9]. For preparation, postal codes from WoS were extracted and cleaned with regular expressions. The matching was performed exclusively on the basis of exact matchings of postal codes without admitting any similarities.

Postal codes are unambiguous (in contrast to e.g., city names) and therefore provide a secure assignment option when used without admitting similarities but only exact matches.

As this method is exclusively based on postal codes, there is no chance of assignment in several cases:

- In case of changes of postal code systems, the assignment table refers to the actual situation – consequently, addresses with old postal codes cannot be assigned (this concerns countries with changes in postal code systems, e.g., CYP in 1994, in DEU in 1993, in IRL in 2015, in MLT in 1991, in PRT in 1994, in ROU in 2003).
- Of course, addresses without postal codes cannot be assigned.
- Special types of postal codes (e.g., special postal codes for major clients or post boxes) are not included in the assignment table and cannot be assigned.

Therefore, recall depends on the country of interest and may be low.

In addition, this method may produce errors due to the fact that postal code regions do not match the system of NUTS codes exactly: while NUTS are territorial units in the sense of subdivisions of countries, postal code regions are districts of postal service that do not necessarily depend on territorial units in the sense of NUTS codes. Therefore, there are postal code regions that are covered by more than one NUTS region – this fact is ignored in the assignment table provided by Eurostat, resulting in just one NUTS region per postal code. Nevertheless, these are exceptional cases – in the majority of cases postal codes regions may be assigned to one single NUTS region.

[9] http://ec.europa.eu/eurostat/tercet/flatfiles.do

2.2.2 NUTS labels

City values from WoS were matched to NUTS labels, admitting only exact matches. For preparation, both NUTS labels as well as city values were prepared using a transformation step (for this, the transformation step for the institutional disambiguation procedure[10] was used, including e.g., the removal of special characters, application of abbreviations, correction of typing errors).

Due to just one NUTS label per NUTS code, name variants prevent assignments. In addition, the city level does not match the NUTS level and therefore the NUTS label does not necessarily match a city name but more frequently the name of a region, for example, the NUTS code DE11B (labeled as 'Main-Tauber-Kreis') includes the cites 'Bad Mergentheim', 'Creglingen' and 'Freudenberg'. The other way around, large cities may have more than one NUTS-code (EL301-EL304 for Athens).

As for every matching based on city names, homonyms may lead to errors. As there is only one label per NUTS code, name variants cannot be assigned here via exact matching. Therefore, another matching was performed admitting the WoS city value to have a Jaro-Winkler-Similarity[11] of at least 95 to the NUTS label.

2.2.3 LAU to NUTS

Unlike NUTS, the LAU units match the city level, therefore, these are more suitable for a matching with WoS city values (here again, transformed labels were matched to transformed city values admitting only exact matches. In addition, another matching was performed based on a Jaro-Winkler-Similarity of at least 95).

An assignment table of LAU to NUTS is provided by Eurostat[12]. Here again, only one LAU label is provided for a LAU code preventing assignments or

[10] Rimmert C, Schwechheimer H, Winterhager M. Disambiguation of author addresses in bibliometric databases - technical report. Bielefeld: Universität Bielefeld, Institute for Interdisciplinary Studies of Science (*I^2SoS*); 2017 (https://pub.uni-bielefeld.de/publication/2914944).

[11] https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance

[12] http://ec.europa.eu/eurostat/de/web/nuts/local-administrative-units

other name variants. In addition, LAU labels often contain country specific additives to city names such as e.g., ‘Stadt’, ‘Local commune of’, ‘Sogn’or ‘Citta’, preventing an exact match to a WoS city value. Again, homonyms (that are more likely on LAU than on NUTS level) may lead to errors.

2.2.4 OSM API

A comma-separated combination of postal code, city and country name (standardized, English) was passed to the OSM-API in order to receive geographical coordinates which then were assigned to NUTS codes via shapefiles (python libraries[13] were used to query the OSM API as well as to check if a given geographical coordinate is located in a certain polygon, respectively the corresponding NUTS region).
For the matching only the most likely match was requested (other possibilities are conceivable but lead to ambiguous results).

Of course, the matching is very easy as the OSM does all the work – at the price of having no control about the matching process and errors. The WoS preprocessing (abbreviations, removing of special characters and so on) may lead to errors as OSM does not assume this preprocessing. The OSM API deals with only one request per second which leads to long processing times in case of a large amount of addresses.

2.2.5 GeoNames

GeoNames provides names and alternate name variants for cities together with their geographical coordinates. Transformed city values from WoS were matched to names and alternate names from the respective country admitting only exact matches again. This matching provided geographical coordinates gained from GeoNames which again were matched to NUTS codes via shapefiles.

The existence of name variants is an advantage compared to e.g., LAU label matching (e.g., for Köln there are amongst others the alternate names Koeln, Cologne, Cologna). Of course, here, too, homonyms can cause multiple assignments and errors.

[13] https://github.com/geopy/geopy, https://github.com/gka/pyshpgeocode

2.2.6 Wikidata

Two matching methods were performed with wikidata[14]: on the one hand, a matching of transformed WoS city values to transformed wikidata labels (only wikidata entities with geographical coordinates) provided geographical coordinates from wikidata which could be assigned to NUTS codes via shapefiles.

On the other hand, the (again transformed) first part of the address (usually containing the description of the main institution such as e.g., the university) was matched to wikidata labels – with a restriction to wikidata entities with geographical coordinates and a country attribute matching the country value of the address (country values from WoS were assigned to wikidata ids for this purpose). Here – in contrast to the wikidata city matching – targets of the matching are wikidata entities corresponding to research institutions, not cities. Also this matching provided geographical coordinates (for research institutions). This matching method is referred to as 'wikidata orga1' in the following.
Sources of errors are e.g., research institutions with several locations or ambiguity in the first part of the address (e.g., 'UNIV HOSP' while the city name is mentioned in another part of the address).

[14] A full dump of wikidata from May 2017 (excluding some classes (P31 values) assumed to be of no interest in this context, e.g., 'human') was used here.

3 Procedure

3.1 Pre-test : AUT, DEU, GRC, MLT

In order to get further insights on the performance parameters of the different matching methods, random samples of 150 addresses each were taken and labeled manually from four example countries: Austria (AUT), Germany (DEU), Greece (GRC) and Malta (MLT). This choice was made due to different numbers of addresses per country and possible problems caused by transliterations in case of GRC.

For each country and matching method, performance parameters were calculated – results are presented in figure 3 and table 1 where figure 3 provides an overview while table 1 contains the exact numbers and the f-score in addition.

Precision, recall and f-score are defined as follows in this context:

$$\begin{aligned}
\text{precision} &:= \frac{\#\text{ correct assignments}}{\#\text{ assignments (total)}}, \\
\text{recall} &:= \frac{\#\text{ addresses assigned}}{\#\text{ addresses (total)}}, \\
\text{f-score} &:= 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}.
\end{aligned}$$

Depending on the matching methods there may be more than one assignment per address. In 11 cases of the sample set of 600 addresses from the four sample countries, it was not possible to assign any NUTS code on level 3 – these cases were excluded.

In figure 3 countries are represented by symbols while the different matching methods are displayed via colors. For example the red triangle shows precision and recall for the wikidata city method applied to addresses with countrycode MLT. As visible, there are differences concerning countries as well as concerning the methods. There are cases with high precision and recall values (which is the best option) – upper right corner – as well as methods providing only little recall but high precision – upper left – and also methods with low values for both precision and recall. The latter group contains especially the NUTS and LAU label matches (for which the low recall values have already been mentioned in the previous section). For GRC there are even no matches for the direct NUTS label matching methods.

For NUTS and LAU label matchings two variants were tested each as already described above: the exact matching on the one hand – dark blue respectively dark green in the figure – and a matching with at least high string similarity (in terms of Jaro Winkler, similarity ≥ 95) – light blue/light green. It turned out that the use of similarities instead of exact matches had no effect on AUT and MLT (in these cases only one of the green, respectively blue symbols is visible in the figure as the other one is hidden behind). For DEU there is a slight gain of recall with a slight loss of precision, on the other hand, and for GRC there is a gain of recall in terms of factor 3 but an extreme decrease of precision. For this reason, allowing similarities instead of exact matches seems to be a bad deal – therefore, these matching methods were excluded.

country	**method**	**precision**	**recall**	**f-score**
AUT	geonames	0.967	0.973	0.97
AUT	lau match	1	0.4	0.571
AUT	lau, sim≥95	1	0.4	0.571
AUT	nuts label match	0.818	0.273	0.409
AUT	nuts label, sim≥95	0.818	0.273	0.409
AUT	osm	0.993	0.98	0.986
AUT	postalcode	0.952	0.7	0.807
AUT	wikidata city	0.987	0.98	0.983
AUT	wikidata orga1	0.971	0.46	0.624
DEU	geonames	0.576	0.979	0.725
DEU	lau match	0.899	0.646	0.752
DEU	lau, sim≥95	0.833	0.674	0.745
DEU	nuts label match	0.264	0.382	0.312
DEU	nuts label, sim≥95	0.261	0.389	0.312
DEU	osm	0.853	0.944	0.896
DEU	postalcode	0.985	0.465	0.632
DEU	wikidata city	0.471	0.986	0.637
DEU	wikidata orga1	0.901	0.486	0.631
GRC	geonames	0.725	0.98	0.833
GRC	lau match	1	0.075	0.14
GRC	lau, sim≥95	0.405	0.245	0.305
GRC	osm	0.877	0.939	0.907
GRC	postalcode	0.971	0.476	0.639
GRC	wikidata city	0.758	0.925	0.833
GRC	wikidata orga1	0.934	0.415	0.575
MLT	geonames	0.968	0.824	0.89
MLT	lau match	1	0.696	0.821
MLT	lau, sim≥95	1	0.696	0.821
MLT	nuts label match	0.25	0.007	0.014
MLT	nuts label, sim≥95	0.25	0.007	0.014
MLT	osm	1	0.723	0.839
MLT	postalcode	1	0.007	0.014
MLT	wikidata city	0.96	0.818	0.883
MLT	wikidata orga1	1	0.669	0.802

Table 1: Performance parameters of different matching methods

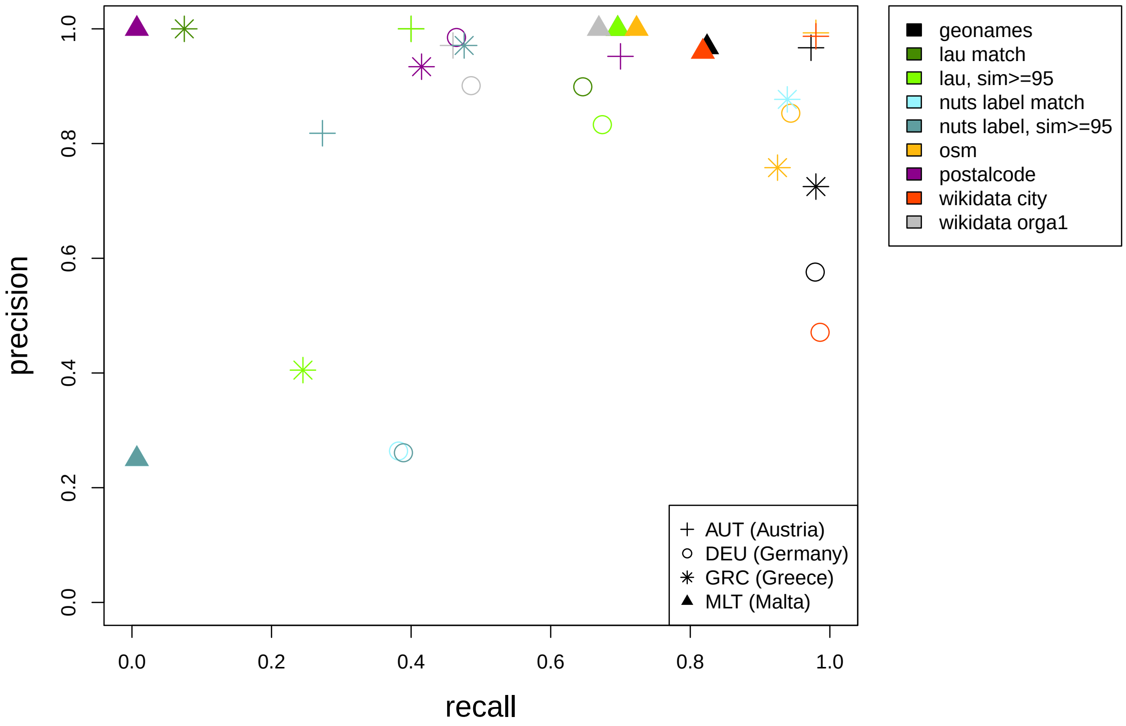


Figure 3: Precision and recall for different methods

method	**precision**	**recall**	**f-score**
geonames	0.769	**0.939**	0.846
lau match	0.963	0.453	0.616
lau, sim≥95	0.870	0.503	0.637
nuts label match	0.418	0.165	0.237
nuts label, sim≥95	0.415	0.166	0.237
osm	0.928	0.896	**0.912**
postalcode	0.967	0.413	0.579
wikidata city	0.723	0.927	0.812
wikidata orga1	**0.957**	0.508	0.664

Table 2: Performance parameters of different matching methods (AUT, DEU, GRC and MLT)

Table 2 shows the performance parameters for the whole sample set (the union of all four countries). While the geonames matching method provides the best recall, wikidata orga1 is best in precision. If one had to choose one single matching method among the ones presented, osm would be the best compromise between precision and recall (highest f-score).
In the context of this project, it is not necessary to restrict to one single method – instead, we would like to take advantage of possible combinations among the methods to achieve a better performance compared to single methods.

The simplest way of combining methods is, of course, putting them all together. As expected, this does not lead to suitable results. While the recall (0.978) is, of course, the maximal achievable one (while using the methods described here) in case of simply combining all methods, precision (0.567) is very low and even single methods achieve higher f-scores (the f-score would be 0.709).
Of course, the number of methods covering a certain assignment (an assignment address → NUTS3 can be done by just one method or simultaneous by two or more methods) may give hints for its correctness. Figure 4 shows, how many assignments are covered by a certain number of methods. As the nuts and lau exact label matching results are a subset of the results of the corresponding similarity matchings, in both cases only the similarity matchings are included, leading to a maximum number of methods of 7 (instead

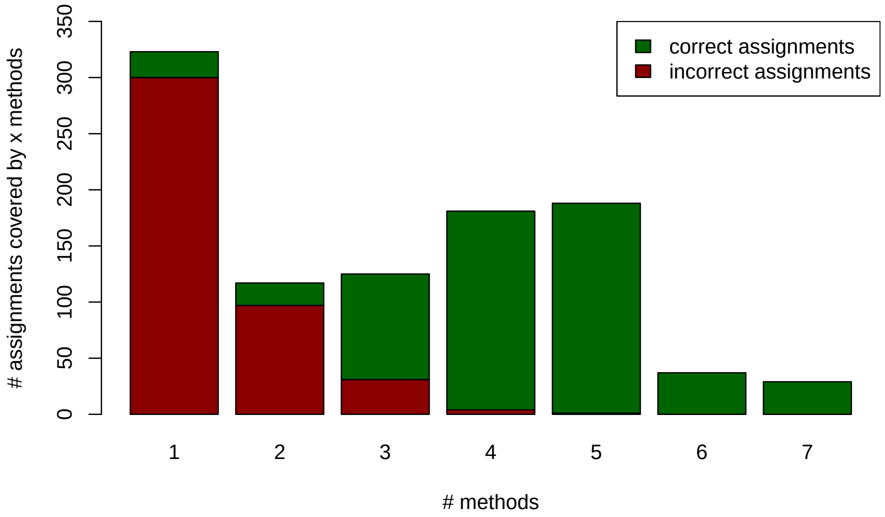


Figure 4: Number of assignments covered by a certain number of methods.

of 9). As expected, assignments covered by more methods are more likely to be correct. Nevertheless there is a large amount of assignments covered by two or three methods which are not correct and there are even assignments covered by 4 methods that are incorrect (these are just two and both concerning different NUTS3-codes within Athens). A restriction to assignments covered by at least four methods would lead to an unacceptable loss of recall – with good precision of course.

3.2 Method groups

Therefore, it is not sufficient to consider exclusively the number of methods covering a certain assignment but decisions for combinations and priority rules should be oriented towards the concrete methods involved. Figures 5 and 6 provide a closer look at combinations of two, respectively three methods. The method combination is given as a label while the y-axis shows the number of (correct in green and incorrect in red) assignments covered by exactly the given methods.

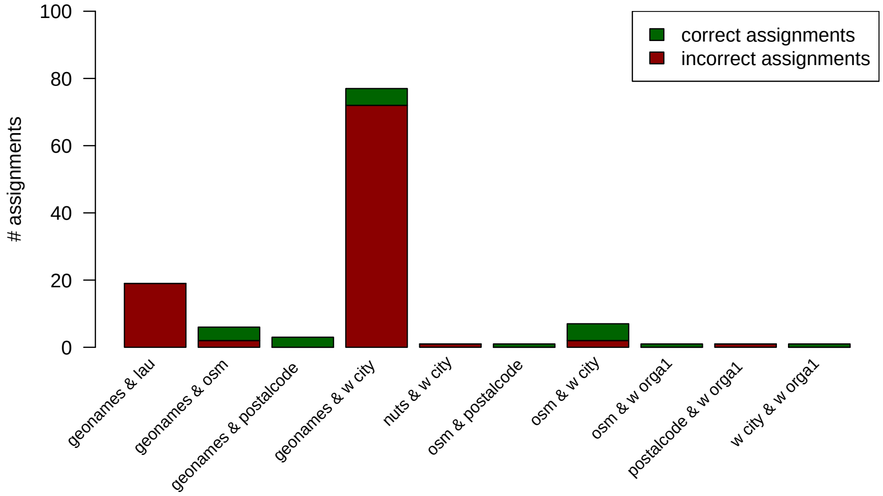


Figure 5: Combinations of 2 methods.

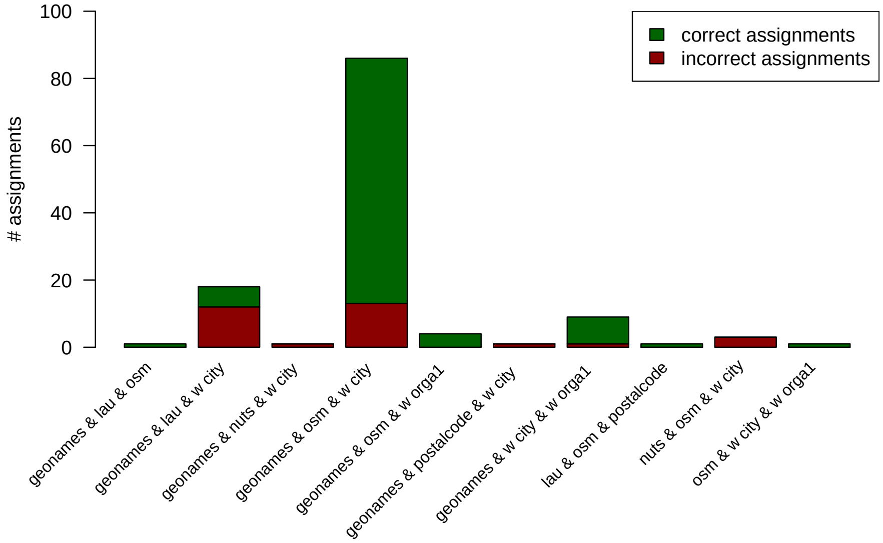


Figure 6: Combinations of 3 methods.

As visible in the figures, some method combinations seem to be more likely to produce correct results than others: while geonames and lau cover only incorrect assignments in common, geonames and postalcode cover only correct results in common, geonames and osm as well as osm and wikidata city cover more correct than incorrect assignments. While geonames and wikidata city have primarily incorrect assignments in common, adding osm or wikidata orga1 as a third method leads to much better results.

Good 'partners' for method combinations are 'independent' methods (meaning methods using different features) as the probability of a correct result is expected to be higher if it is proposed by different matching methods using different features. The matching methods can be grouped as follows:

1. postal code and city-based string matches
   (osm)

2. city-based string matchings
   (LAU and NUTS label matches, wikidata city, geonames)

3. postalcode-based string matchings
   (postalcode)

4. institution string-based matches
   (wikidata orga1).

The observations from figures 5 and 6 fit together in this: 'good partners' are independent (meaning in different groups): e.g., geonames and lau are in the same group and cover incorrect results (meaning they make the same mistakes) while geonames and postalcode is a combination with a good probability of correct results.

Matching methods from different groups can, on the one hand, be used to confirm results of each other. On the other hand they can be used to increase recall when used in addition as it is expected that methods from different groups are more likely to be able to process different addresses so that the union of results retrieves higher recall.

Figures 7 and 8 show the numbers, respectively percentage of correct and incorrect results for method group combinations (in contrast to method combinations as dealt with above) in the sample set – so this covers the aspect

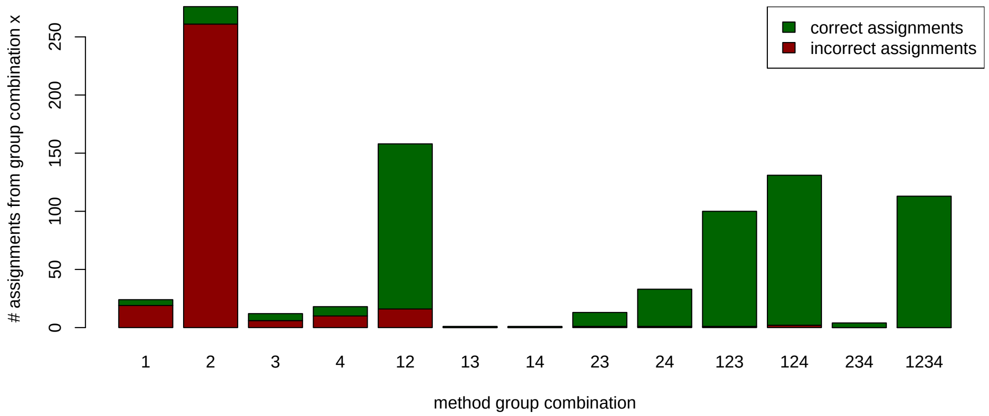


Figure 7: Correct and incorrect results for group combinations, absolute.

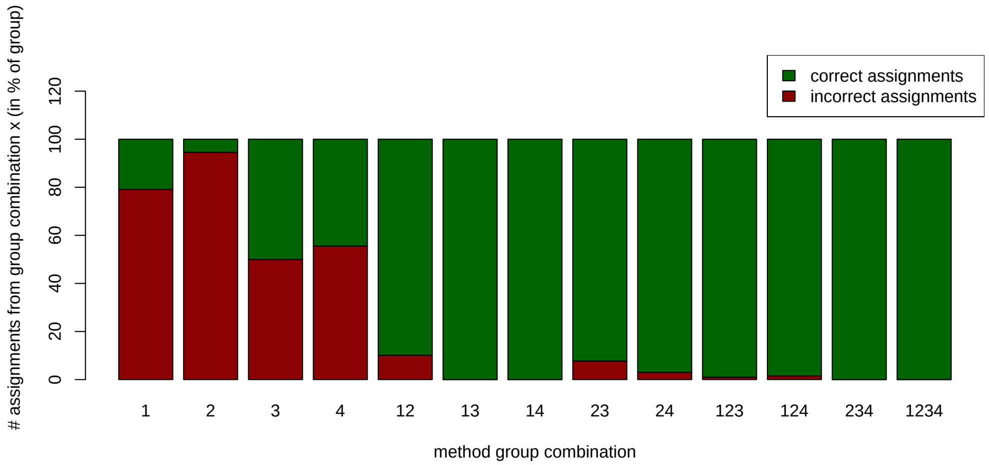


Figure 8: Correct and incorrect results for group combinations, percentage.

of different method groups used as confirmation for assignments. It can be stated that using different groups in combinations leads to better results than just looking at any combination of an arbitrary choice of e.g., three methods.

Of course, the sample is too small to obtain already reliable statements for the overall dataset (some combinations are quite rare, 134 does not even appear), but serves to get hints on appropriate method combinations.

In addition to the more content-related observations above, intersections of assignments among the methods were calculated in order to identify the similarity of the matching methods considered from a statistical point of view. This can be used as a verification of the content-based considerations concerning sensible method combinations, and such calculations may as well provide a hint on methods that are superfluous as e.g., their results are already completely contained in the results of another method.
As stated before, NUTS and LAU labels matches based on similarity were excluded. Table 3 shows the results in terms of absolute numbers while in table 4 results are displayed as relative numbers (the absolute numbers divided by the total number of assignments per (row) method). So the number in row i and column j is calculated as:

$$n_{i,j} = \frac{\text{assignments method}_i \cap \text{assignment method}_j}{\text{assignments method}_i},$$

giving the amount of joint allocations of methods i and j with respect to the number of allocations of method i. Therefore, table 4 is not symmetric ($n_{i,j} \neq n_{j,i}$) in contrast to table 3 where the total number of assignments (and therefore the denominator per row) in the sample set per method can be found on the diagonal.
Thus, looking at the column provides information concerning the share of assignments of other methods already covered by the method of interest (column label) while the row of a method leads to information on the share of assignments covered by other methods. As an example, geonames assignments cover 92.9% of the osm assignments, 93.9% of the postalcode assignments and 41.2% of the assignments done by nuts labels (information from the geonames column), while 69.7% of the geonames assignments are also covered by osm, 39.2% are also covered by wikidata org1 and so on.

It can be stated that no method is superfluous in terms of being completely contained in another method. Furthermore, the intersections between methods from different groups seem to be sufficient – if these were too low there will be no sense in using methods for verifying each others' results.

	geonames	**osm**	**postalcode**	**w_org1**	**w_city**	**nuts**	**lau**
geonames	**712**	496	229	279	627	68	276
osm	496	**534**	214	245	496	70	223
postalcode	229	214	**244**	118	226	45	97
w_org1	279	245	118	**302**	277	36	157
w_city	627	496	226	277	**750**	72	274
nuts	68	70	45	36	72	**165**	58
lau	276	223	97	157	274	58	**277**

Table 3: Intersections, absolute numbers

	geonames	**osm**	**postalcode**	**w_org1**	**w_city**	**nuts**	**lau**
geonames	**1**	0.697	0.322	0.392	0.881	0.096	0.388
osm	0.929	**1**	0.401	0.459	0.929	0.131	223
postalcode	0.939	0.877	**1**	0.484	0.926	0.184	0.398
w_org1	0.924	0.811	0.391	**1**	0.917	0.112	0.488
w_city	0.836	0.661	0.301	0.369	1	0.096	0.365
nuts	0.412	0.424	0.273	0.218	0.436	**1**	0.352
lau	0.996	0.805	0.350	0.567	0.989	0.209	**1**

Table 4: Intersections, relative numbers

3.3 Composed procedure

With the lessons learned from the previous section, we can compose a procedure using different matching methods for complementing as well as verifying respective matching results. Led by precision values for different group combinations, the priority is chosen as follows:

1. Assignments with group combination 1234
2. Assignments with group combination 234
   (for addresses not yet assigned by group combination above)
3. Assignments with group combination 124
   (for addresses not yet assigned by group combinations above)
4. Assignments with group combination 134
   (for addresses not yet assigned by group combinations above)
5. Assignments with group combination 123
   (for addresses not yet assigned by group combinations above)
6. Assignments with group combination 14
   (for addresses not yet assigned by group combinations above)
7. Assignments with group combination 13
   (for addresses not yet assigned by group combinations above)
8. Assignments with group combination 34
   (for addresses not yet assigned by group combinations above)
9. Assignments with group combination 24
   (for addresses not yet assigned by group combinations above)
10. Assignments with group combination 23
    (for addresses not yet assigned by group combinations above)
11. Assignments with group combination 12
    (for addresses not yet assigned by group combinations above)
12. Assignments with group combination 3
    (for addresses not yet assigned by group combinations above)
13. Assignments with group combination 4
    (for addresses not yet assigned by group combinations above)
14. Assignments with group combination 1
    (for addresses not yet assigned by group combinations above)
15. Assignments with group combination 2
    (for addresses not yet assigned by group combinations above)

After the application of these method group combinations, multiple assignments were reduced in three steps based on statistics concerning the results received. For a given address a with more than one NUT3 code assigned, a_o is defined as the organization1[15] string and a_c is used for the city string of a in the following.

1. **Organization1-city combination (nearly) unambiguous among the results with unique assignments:**
   Among all unique assignments in the results with organization1=a_o and city=a_c (both case insensitive), frequencies for NUTS3 codes were calculated.
   If there is one NUTS3 code assigned to at least 95% of addresses from this set, this NUTS3 code is assigned to a while all other assignments to a were deleted.

2. **Addresses with unique assignments having Jaro-Winkler-Similarity of ≥ 95 to the respective address:**
   All addresses with unique assignments in the results and city=a_c (case insensitive) having a Jaro-Winkler-Similarity of $\geq$ 90 to a were extracted. In case of the existence of at least one address with these properties, only the belonging NUTS3 code(s) were left in the results (removal of all other assignments).
   This does not necessarily lead to a unique assignment – there may be more than one NUTS3 code assigned to the set of 'similar addresses'.

3. **City is (nearly) unambiguous among the results with unique assignments excluding rather 'insecure' method groups:**
   Analogous to step 1, a_c was tested concerning uniqueness, this time based on only rather secure method groups (exclusion of 1,2,3,4,12).

Finally, addresses assigned to more than three NUTS3 codes were removed from the results.

[15] First part of WoS address, usually containing institutional information like e.g. 'Univ Bielefeld'.

3.4 Special cases

It turned out that in two cases of more than one NUTS3 code per city (London and Athens), assignments on NUTS3 level are not affordable in a satisfying manner (sometimes even hard or impossible with manual effort), while assignments for Paris – which has more than one NUTS3 code, too – could be handled better due to manually checked samples for precision evaluation described below.

Therefore, assignments for London and Athens were done on NUTS2 level. However, it can be stated that the two cases differ: while for London in most cases multiple assignments (to different NUTS3 codes for parts of London) occur, assignments for Athens could be done in a unique way and correctly in much more cases. To evaluate this further, special checks for Athens were conducted with the following results: from 114 assignments to EL301-EL304 (NUTS3 codes for Athens), 79 were correct, one is ambiguous, 9 could only be assigned on NUTS2 level and 25 are incorrect. Furthermore, assignments done on the basis of specific method groups were exclusively correct (1234, 123).

These results are clearly not sufficient for the use in analyses on NUTS3 level without further preparation, but may serve as a basis for manual post-processing and, of course, they can be used on NUTS2 level.

4 Results

Table 5 shows the distribution of assignments (on the basis of distinct addresses) over the method group combinations used. Luckily, a large amount of assignments could be done with method group combinations assumed (from the pretest) to be secure ones, such as 1234 as the best option. Nevertheless, a large part is also handled by combination 12 which is expected to be less secure than the ones mentioned before and some addresses have to rely on the use of a single method group.

method group combination	**# assignments**	**in % of all assignments**
1234	1982562	15.54
234	457772	3.59
134	37407	0.29
124	1399785	10.97
123	3562583	27.92
34	35557	0.28
24	315088	2.47
23	754121	5.91
14	8242	0.06
13	113567	0.89
12	3515919	27.56
4	37408	0.29
3	91521	0.72
2	331108	2.6
1	115755	0.91

Table 5: Distribution of assignments over method group combinations.

4.1 Precision & recall

For investigation performance parameters, manually checked random samples of 250 addresses (and all belonging assignments, may be more than one per address) each were created – resulting in 7,002 manually checked assignments overall.

The random choice was made considering the frequency for addresses, so addresses may appear more than once if used in different documents. The assignments were flagged as correct, incorrect, ambiguous (in case of ambiguous addresses, e.g., more than one city mentioned in the address, homonyms or lack of clarity) or country error (in case of country code errors in the address, e.g., ‘Mercy Hosp Women, Dept Pathol, E Melbourne, Austria’ for AUT). As mentioned above, Athens and London were evaluated on NUTS2 level.

Table 6 shows the distribution of assignments (based on distinct addresses) for the whole data set compared to the random sample. With some exceptions (especially 124 and 123) the distribution of the sample seems roughly similar to the distribution of the whole set.

method group combination	**whole set**	**sample**
1234	15.54	15.89
234	3.59	3.81
134	0.29	0.08
124	10.97	19.63
123	27.92	19.2
34	0.28	0.24
24	2.47	2.43
23	5.91	4.05
14	0.06	0.05
13	0.89	0.24
12	27.56	31.15
4	0.29	0.41
3	0.72	0.27
2	2.59	1.95
1	0.91	0.59

Table 6: Distribution of assignments over method group combinations.

The samples were drawn with equal numbers for all countries. But while the countries differ in their number of addresses, this sample cannot be used as an evaluation for the overall address set of all EU countries in a strict sense. Nevertheless, table 7 shows the performance parameters for the whole set of EU countries based on this sample (in case of precision – recall was calculated based on the whole set of addresses).
Having the random sample at hand it is now also possible to evaluate precision and f-score of the single methods (not method groups) on a broader basis than in the pretest to check if not only recall is better for the combined procedure but also the f_score. Results are also presented in table 7 where the combined procedure exceeds every single method in terms of not only recall but also f-score values.

method	**# assignments**	**# correct**	**precision**	**recall**	**f_score**
osm	6360	6049	0.951	0.874	0.911
geonames	8679	6893	0.794	0.985	0.879
wikidata city	8947	6949	0.777	0.980	0.867
postalcode	3083	2994	0.971	0.576	0.723
lau	3368	3161	0.939	0.502	0.654
wikidata orga1	3217	3064	0.952	0.476	0.635
combined procedure	7002	6928	0.989	0.998	0.993

Table 7: Performance parameters of the combined procedure.

Nevertheless, a consideration on the basis of single countries is also of interest as differences among countries showed up already in the pretest. Table 8 therefore provides performance parameters on the country level. Recall values differ per country but are high for all countries (all above 0.99 except MLT). Precision values are all above 0.95.

In addition to performance parameters of the combined procedure, the maximal f-value achieved by a single method, the concerning single method (column 'max method') and the gain in using the combined procedure compared to the country-specific best single option are given. Here again, significant differences among countries show up. First, the 'best option' differs among countries – while, e.g., osm performs best for DEU, the best option is lau for SWE and geonames for EST. Thus, when using a single method for this task, a country specific choice would be of value.
While there is a gain of f-score in almost all cases, there are indeed two cases with a loss of f-score compared to the best option (where values for LVA are nearly equal). For LVA, errors occur due to country errors in WoS and problematic addresses (postal code wrong, value in city field is not a city).
For EST, errors occur exclusively for one city (Tartu) which is assigned to EE007 instead of EE008 due to method group 1 (only osm used). For other method groups, addresses with city Tartu were assigned correctly. As this is an error that can be handled very easily as an individual case, addresses with city Tartu and countrycode EST are set to NUTS3 code EE008 (698 distinct addresses affected while 8,113 addresses with city Tartu were already assigned correctly). With this handling of Tartu addresses, precision turns to 1.000 for EST and therefore the f-score to 0.999, which leads to a gain of 0.004 compared to the maximal f-score of a single method.

country	# addr	recall	precision	f-score	max single f	max method	gain
AUT	688691	0.999	0.992	0.995	0.988	osm	0.007
BEL	1002106	0.999	0.996	0.997	0.979	wikidata city	0.018
BGR	125569	0.998	1.000	0.999	0.995	wikidata city	0.004
CYP	25334	0.992	1.000	0.996	0.992	osm	0.004
CZE	495472	0.997	0.992	0.994	0.980	wikidata city	0.014
DEU	6056627	0.998	0.980	0.989	0.899	osm	0.090
DNK	758884	0.998	0.996	0.997	0.957	osm	0.040
ESP	2521596	0.998	0.988	0.993	0.950	osm	0.043
EST	57401	0.999	0.972	0.985	0.995	geonames	-0.010
FIN	714087	0.999	0.996	0.997	0.947	wikidata city	0.050
FRA	4745049	0.999	0.952	0.975	0.923	osm	0.052
GBR	6656261	0.999	0.984	0.991	0.928	osm	0.063
GRC	524670	0.994	0.984	0.989	0.962	osm	0.027
HRV	150974	0.998	0.996	0.997	0.991	osm	0.006
HUN	374640	0.999	1.000	0.999	0.997	lau	0.002
IRL	308649	0.998	0.996	0.997	0.995	osm	0.002
ITA	3930918	0.998	1.000	0.999	0.887	osm	0.112
LTU	80065	0.997	1.000	0.998	0.991	geonames	0.007
LUX	18779	0.998	0.996	0.997	0.993	wikidata city	0.004
LVA	33568	0.998	0.972	0.985	0.986	geonames	-0.001
MLT	5985	0.921	1.000	0.959	0.911	geonames	0.048
NLD	2073814	0.998	0.980	0.989	0.973	lau	0.016
POL	1058202	0.999	0.984	0.991	0.861	osm	0.130
PRT	498493	0.993	0.996	0.994	0.976	wikidata city	0.018
ROU	326612	0.997	0.992	0.994	0.979	osm	0.015
SVK	152232	0.998	1.000	0.999	0.995	wikidata city	0.004
SVN	139414	0.998	0.988	0.993	0.979	wikidata city	0.014
SWE	1259994	0.999	0.972	0.985	0.903	lau	0.082

Table 8: Performance parameters per country.

5 Conclusion & indications

The classification task of assigning addresses to NUTS3 codes was handled by combining different matching methods based on city and organization name strings, postal codes, geographical coordinates from different sources and the OpenStreetMap geocoding API. In case of Athens and London the level had to be dropped to NUTS2 due to difficulties and ambiguities with more than one NUTS3 code per city.
The resulting procedure shows good performance parameters for the whole data set where significant differences among countries show up.

A new version of the NUTS classification ('NUTS-2016'[16]) has been released within the project term (valid since 01.01.2018). A list of changes between NUTS-2013 (used here) and NUTS-2016 is provided by Eurostat[17]. Matching tables for postal codes are not yet available for the new NUTS release, therefore the procedure could not yet be applied to the new NUTS system.
A transformation of NUTS-2013 into NUTS-2016 could be done according to tables with change information provided by Eurostat. This is easy in most cases (recodings or mergers) but due to border changes in some NUTS3 codes this is difficult in these special cases.

6 Acknowledgements

Thanks to Marc Weingart for the english language support and to Andre Bruns and Christopher Lenke for their support in checking all the lists.

[16] http://ec.europa.eu/eurostat/de/web/nuts/background

[17] http://ec.europa.eu/eurostat/de/web/nuts/history