Swiss-SEP 2.0 index
Report 1.10 - data preparation

Radoslaw Panczak *et al.*

May 16, 2022

Contents

1 SNC - buildings

1.1 Eligible buildings

Origin buildings are defined as all buildings for which index is going to be calculated. These buildings need to:

1. Be present at least once in the **period of 2010-2014** in the SNC dataset.
2. Have valid 2010+ **building ID**.
3. Have valid 2010+ **geographical coordinates**.
4. Belong to category of 'normal' **residential buildings** (ie. no prisons, churches or nursing homes; see Appendix).

Buildings are selected from the `snc2_std_pers_90_00_14_all_206_full` dataset and processed as follows:

1. All buildings that have an ID and coordinates on any year from **2010** onward are selected
2. Submeter coordinates are rounded to 1m
3. **Newest** coordinates are always used when several are available under the same building ID
4. **Non-residential** buildings (see above) are excluded
5. Buildings having different ID but **same cordinates** are groupped together using synthetic 'GIS ID' (for instance 153 (sic!) different buidling IDs pointing to the same coordinates on a caravan site?)

These coordinates become **n'hood centres** for network analysis and construction of an index.

1.2 Results

Distribution of years from which coordinates of a building are taken:

```
(SSEP 2.0 - ´origin´ SNC buildings for network analysis)

year — Year of coordinates
--------------------------------------------------------------
               |      Freq.    Percent      Valid       Cum.
---------------+----------------------------------------------
Valid   10     |       9550       0.62       0.62       0.62
        11     |      10426       0.68       0.68       1.30
        12     |      13118       0.85       0.85       2.15
        13     |      22880       1.49       1.49       3.63
        14     |    1484614      96.37      96.37     100.00
        Total  |    1540588     100.00     100.00
--------------------------------------------------------------
```

Note the distinction between IDs (ie. small amount of buildings with different ID but same coordinates):

```
         |      Observations
         |      total   distinct
---------+----------------------
 buildid |    1540588    1540588
   gisid |    1540588    1527177
```

2 SE

2.0.1 Eligible persons & households

Destination households are defined as all household that can provide information for calculation of the index. They need to be present in at least one Structural Survey (SE) during the period of 2012-2015. Surveys of 2010 and 2011 do not provide information about m2 area of the flat which is needed for calculation of standardised rent and were therefore excluded. Additionally, there are some reservations as to quality of the 2010 data.

In order to be included, SE personal record must (sequentially):

1. Link to household record.
2. Link to full SNC for `buildid`.[1]
3. Link to valid coordinates (from ORIGINS dataset, see previous section).

Key variables[2] needed are then selected from each of the sources:

1. `sncid, hhyid, age, sex, educ_agg, educ_curr, occup_isco, workstatus` from the `SEyy_pers_full` dataset.
2. `hhyid, hhtype, hhpos, hhpers, flatrooms, typeowner, rentnet` from the `SEyy_hh_full` dataset (linked via `hhyid`)
3. `buildid` from the `snc2_std_pers_90_00_14_all_206_full` dataset (linked via `sncid`)
4. `geox, geoy` from the `ORIGINS` dataset (linked via `buildid`)

At next stage, individuals are excluded if:

1. Are younger than 19 at the time of SE.
2. Have one of the 'unusual' types of residence permit (Cross-border commuter (G), Short stay (L), Asylum seeker (N), People in need of protection (S), Person required to notify (Meldepflichtige), Diplomat/internat. official with diplomatic immunity, Internat. official without diplomatic immunity, Not classified elsewhere)
3. If individual participated in more than one SE, the latest record is kept.

For remaining individuals and their households, the following data are prepared:

1. Individuals are flagged if they work in **manual or unskilled occupations** (BUT only if they are in **paid employment** at the time of SE; see below).
2. Individuals are flagged if they have **no formal or have only compulsory education** AND are not currently pursuing any further education.
3. Households have their **crowding** (number of persons per room) calculated.
4. Households are flaged if they have **three to five rooms and are rented**.

[1] Apart from 2015 SE data that are not yet included in the full SNC; egid identifier of the building was kindly provided by the SNC team

[2] Where 'yy' in the name stands for the year of the SE

2.1 Exclusions

2.1.1 Eligibility criteria

Exclusion	Year			
	2012	2013	2014	2015
Start	270654	266803	272966	255969
Age <19	14791	14463	14184	12929
Permit	570	724	692	611
No household link	41319	40275	42175	35900
No building ID	38	7	4	0
Excluded building	1334	1297	1410	3965
End	227963	225224	229377	216104

The explanation of substantial amount of individuals not linked to households came from BfS:

The reference person has to fill out a form for all household members. As the FSO "calibrate" the structural survey using the information from STATPOP they decided to not include the information for the additional household members if the household structure (number of hh members, gender information) given on the SE household form didn't match the household information in STATPOP. This always applies for around 14% of the SE reference persons.

2.1.2 Multiple SE

In cases when one person participated in more than one SE only newer records were kept.

```
Duplicates in terms of sncid

--------------------------------------
   Copies | Observations       Surplus
----------+---------------------------
        1 |       885591             0
        2 |        13074          6537
        3 |            3             2
--------------------------------------
```

2.2 Results

Distribution of SE individuals over years:

```
(SSEP 2.0 - ´destination´ SE 2012-15 data for SwissSEP 2.0)

SE — Survey year
-------------------------------------------------------------------
              |      Freq.     Percent        Valid          Cum.
--------------+----------------------------------------------------
Valid   2012  |     222305       24.92        24.92         24.92
        2013  |     224516       25.17        25.17         50.08
        2014  |     229204       25.69        25.69         75.78
        2015  |     216104       24.22        24.22        100.00
        Total |     892129      100.00       100.00
-------------------------------------------------------------------
```

Note the distinction between individuals, households, buildings and `gisid`, ie. individual and two spatial resolutions:

```
         |     Observations
         |      total   distinct
---------+----------------------
   sncid |     892129     892129
   hhyid |     892129     892129
 buildid |     892129     581256
   gisid |     892129     575955
```

2.3 Limitations

1. Major limitation is that, compared to SEP 1.0, there is no way to define **head of the household** - all respondents (see exclusions) of the SE are then used, irrespectively of their position in household.

2. 2014 SE dataset is **missing infomration on *'Sozioprofessionelle Kategorie'*** (variable sopc). It has been also signalled by BfS that this variable was of poor quality in 2010-2013 years. Therefore, it is not possible to identify individuals in manual and uskilled occupations in the same way as during construction of original index. That was mitigated by using the **ISCO-08 codes** of occupations to define manual and uskilled workers and farmers. Individuals whose occupations belong to one of the major groups 7, 8 & 9 (for manual and unskilled) and 6 (farmers) were selected.[3] Note that occupation codes are available only for people in **paid employment** so the denomintor for calculating 'employment' domain was adapted and all individuals that were not in paid employment were excluded. Also - small proportion of people eligible for calculations based on ISCO codes had them missing. Again, they were included in the study but had their profession information replaced to missing and again the denominator was adjusted to reflect that.

3. There is significant amount of individuals in SE data with **no link to household SE file** and all these records were excluded.

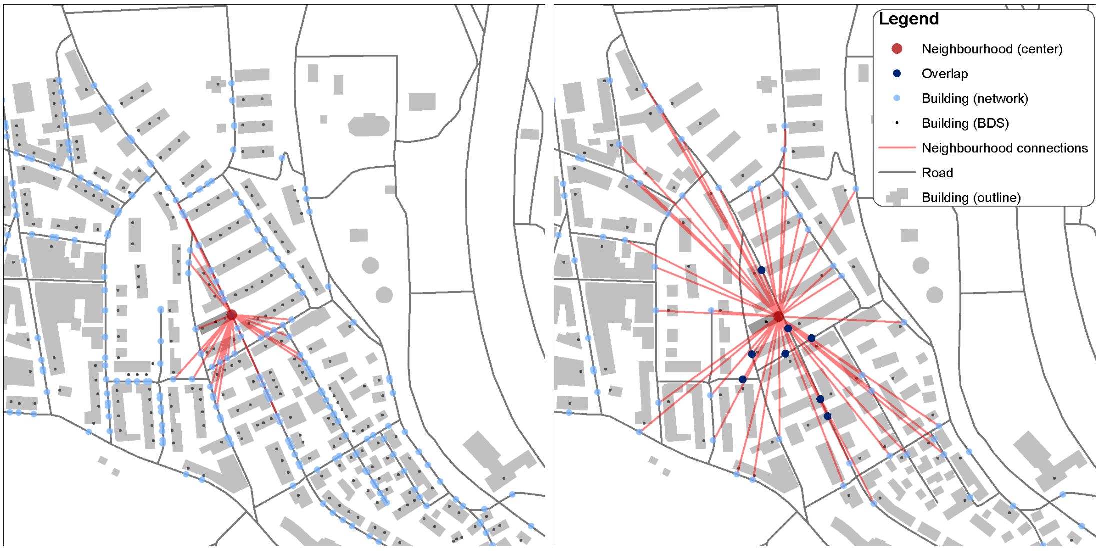
[3] Additionally, sensitivity analyses were done with more strict selection of ISCO codes (major groups 8 & 9 only) as well as by converting ISCO-08 codes to **ISEI-08 codes** to obtain continuous measure of 'International Socio-Economic Index of occupational status'and calculating summary of these vlaues in n'hood

3 Road network

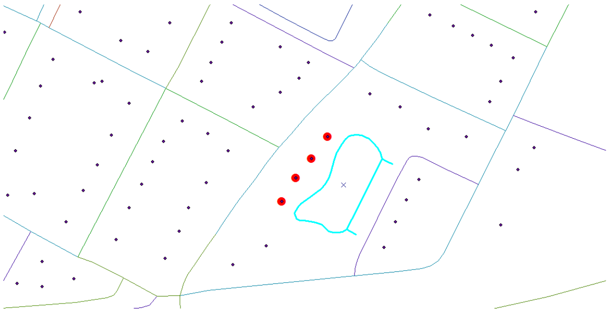
3.1 Setup

1. Network analyses were done using updated version of **swissTLM3D** data (1.5 version as compared to 1.0 vesrsion in the previous edition).
2. Network analyses were done using ArcGIS 10.5 (previously - ArcGIS 10.2).
3. Network analyses took all SNC buildings as `ORIGINS` and calculated 50 closest `DESTINATIONS` from the SE dataset. [4]
4. Treshold for n'hood construction was set up to be maximum 20 km (measured along the road network).[5]
5. As in the 1.0 index, separate n'hoods were created using rented, 3-5 bedroom flats as `DESTINATIONS`.

Schematic representation of n'hood 'search' comparing the use of all buildings to use of sample buildings could be visualized as follow:'



Small *ad hoc* corrections of the **swissTLM3D** dataset were necessary in cases where unconnected segments of the road network were found. This features were then removed:



3.2 Results - buildings

Vast majority of the SNC buildings (`ORIGINS`) have network connections to 50 SE buildings (`DESTINATIONS`) [6]:

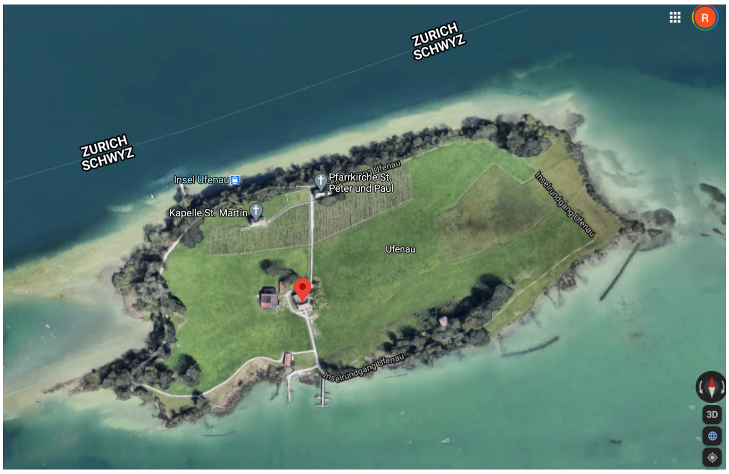
[4] In that logic, the n'hood is either constructed from one SE household and 49 SE neighbours OR 50 SE neighbours if the n'hood centre is not the SE household

[5] That was based on preliminary checks with data, results of previous analyses & common sense rationale (hard to say it's n'hood if households are more than 20km apart...

[6] Keep in mind this results will get even better when we move from buildings to households

```
b_maxdest
----------------------------------------------------------------------
                    |       Freq.    Percent      Valid        Cum.
----------------------------------------------------------------------
Valid   1           |           2       0.00       0.00        0.00
        26          |          29       0.00       0.00        0.00
        41          |           2       0.00       0.00        0.00
        44          |           8       0.00       0.00        0.00
        45          |           2       0.00       0.00        0.00
        49          |           1       0.00       0.00        0.00
        50          |     1527131     100.00     100.00      100.00
        Total       |     1527175     100.00     100.00
----------------------------------------------------------------------
```

The two cases of buildings with no neighbours are legitimate and really have no neighbours on the (highway restricted) road network: one of the buildings is located on Ufenau Island, Lake Zurich; and the other - right next to highway, on the shore of Thunersee. These two buildings were excluded from the analyses and have no index.



Similarly, buildings with n'hoods not meeting the 50 households treshold size will be flagged.

Few areas where less than 50 buildings were found in the n'hood (respecting 20km road network distance) were located in sparesely populated areas such as: Gondo (close to Simplon Pass) or Avers (Grisons) villages.

Building with the biggest (89!) number of SE households is located in Lausanne and is in fact pretty big.

3.3 Results - households

The n'hood structure of connectivity between SNC buildings & SE households changes (for better! ;) when we move from buildings to households. Keep in mind - there might be more than one SE household in a certain building and if we take that into account household n'hoods can get smaller than building n'hoods. Number of buildings (within 20km):

```
(SSEP 2.0 - household n´hood aggregated stats)
                                    -------------- Quantiles --------------
Variable        n     Mean     S.D.      Min      .25      Mdn      .75      Max
-------------------------------------------------------------------------------
  tot_bb  1527173       39        8        1       34       41       45       50
-------------------------------------------------------------------------------
```

Number of households (within 20km):

```
                                    -------------- Quantiles --------------
Variable        n     Mean     S.D.      Min      .25      Mdn      .75      Max
-------------------------------------------------------------------------------
  tot_hh  1527173       51        1       28       50       50       51       91
-------------------------------------------------------------------------------
```

Number of individuals:

```
                                        -------------- Quantiles --------------
Variable       n     Mean     S.D.      Min      .25      Mdn      .75      Max
-------------------------------------------------------------------------------
tot_hhpers 1527173     2.67     1.28     1.00     2.00     2.00     4.00    14.00
-------------------------------------------------------------------------------
```

Average distance [in meters] to the building where furthest SE household is located (within 20km):

```
                                        -------------- Quantiles --------------
Variable       n     Mean     S.D.      Min      .25      Mdn      .75      Max
-------------------------------------------------------------------------------
mean_dist 1527173      447      564        0      187      272      450    16323
-------------------------------------------------------------------------------
```

3.4 Results - households, rent

As expected, results are slightly worse when we limit network analyses to 3-5 bedroom rented flats only.

Number of rented buildings (within 20km):

```
(SSEP 2.0 - household n´hood aggregated stats - rent)
                                        -------------- Quantiles --------------
Variable       n     Mean     S.D.      Min      .25      Mdn      .75      Max
-------------------------------------------------------------------------------
tot_bb_rnt 1527173       35        8        1       31       36       41       50
-------------------------------------------------------------------------------
```

Number of rented households (within 20km):

```
                                        -------------- Quantiles --------------
Variable       n     Mean     S.D.      Min      .25      Mdn      .75      Max
-------------------------------------------------------------------------------
tot_hh_rnt 1527173       51        2        6       50       50       51      101
-------------------------------------------------------------------------------
```

Average distance [in meters] to the building where furthest rented SE household is located (within 20km):

```
                                        -------------- Quantiles --------------
Variable       n     Mean     S.D.      Min      .25      Mdn      .75      Max
-------------------------------------------------------------------------------
max_dist_rnt 1527173     1650     2051        0      492      890     2144    20000
-------------------------------------------------------------------------------
```

4 Swiss Household Panel

4.1 Setup

Combined waves I, II and III of the Swiss Household Panel (SHP) dataset were used to validate the index

1. SHP households were included if:
   (a) they provided questionarie in 2014
   (b) had complete information regarding the address
   (c) address was sucessflly geocoded[7]
2. Same variables that were used in Table 2 of original publication are extracted [8]
3. Each geocoded household was spatially linked to the colsest building from the ORIGINS dataset

4.2 Surveys & geocoding status

```
(SSEP 2.0 - SHP ´14 data for validation)

Identification of the |   Geocoding status
               survey |        no        yes |     Total
----------------------+----------------------+----------
 SHP_II (sample 2004) |        27      1,358 |     1,385
                      |      1.95      98.05 |    100.00
                      |     25.23      18.73 |     18.82
----------------------+----------------------+----------
  SHP_I (sample 1999) |        44      2,734 |     2,778
                      |      1.58      98.42 |    100.00
                      |     41.12      37.70 |     37.75
----------------------+----------------------+----------
SHP_III (sample 2013) |        36      3,160 |     3,196
                      |      1.13      98.87 |    100.00
                      |     33.64      43.57 |     43.43
----------------------+----------------------+----------
                Total |       107      7,252 |     7,359
                      |      1.45      98.55 |    100.00
                      |    100.00     100.00 |    100.00
```

[7] Geocoding was done using map.geo.admin.ch service.

[8] Note that 'Savings min. 500 SFrs monthly' has changed - it used to refer to '100 CHF'

5 SNC - mortality

5.1 Setup

Association of Swiss-SEP with mortality will be assessed using two models based on complete SNC: 'age & sex' and 'semi adjusted' (additionally taking into account: nationality, civil status, language region & level of urbanization). Setup for the analyses in this scenario:

1. Individuals who are recorded in (at least one of the) 2012 - 2018 Censuses are included
2. Individuals below age 30 on the 1.1.2012 are excluded
3. Date of entry is either 1.1.2012 or earliest census if individual was not recorderd in 2012
4. Individuals who died on or before 12.31.2011 are excluded (unless the death was cancelled in the dataset)
5. For individuals having information on one of the covariates recorded inseveral censuses the latest one is used
6. Individuals with missing civil status were excluded
7. Rhaeto-Romansch language region was merged to German
8. Individuals with no link to the index were excluded

5.2 Individuals & deaths included

```
(SSEP 2.0 - full SNC 4.0 2012-2018 data for mortality analyses)

        |     Observations
        |      total   distinct
--------+----------------------
 mortid |     304162     304162
  gisid |    5249089    1426073
```

5.3 Causes of deaths

```
    Variable |       Sum
-------------+-----------
       d_all |   304,162
        d_lc |    15,268
        d_bc |     6,068
        d_pc |     6,073
        d_re |    16,073
        d_cv |    78,859
        d_mi |     9,453
        d_st |    10,658
        d_ac |     1,025
        d_al |     2,205
        d_su |     4,237
-------------------------
```

5.4 Variables

```
Contains data from data/SNC_ALL.dta
 Observations:     5,249,089                  SSEP 2.0 - full SNC 4.0 2012-2018 data for mortality analyses
    Variables:            57                  16 May 2022 10:40
                                              (_dta has notes)
---------------------------------------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
---------------------------------------------------------------------------------------------------------------
sncid           str11   %11s                  Unique SNC ID for SNC 2.0
mortid          long    %10.0g                Mortality ID
```

```
recid3          str24   %24s                  Unique technical ID (update 17-18)
link            byte    %38.0g     linkco     2000 census records link status
link30          byte    %38.0g     linkco     2000 census records link status (update 2015 & 2016)
link40          byte    %38.0g     linkco     2000 census records link status (update 2017 & 2018)
dstart          int     %dD.N.CY              Start date
dstop           int     %dD.N.CY              Stop date
stopcode        byte    %40.0g     stopcode2_l
                                              Stop code (type of stop date)
dob             int     %dD.N.CY              Date of birth
dod             int     %dD.N.CY              Date of death
yod             int     %10.0g                Year of death
last_demig      int     %dD.N.CY              Latest emig date 00-18
sex             byte    %10.0g     sex_l      Sex
last_census_s~n int     %dD.N.CY              Date of last census seen
totweight       double  %10.0g                Linkage weight
se11_flag       byte    %12.0g     flag       Avail. in structural enquiry 2011
se12_flag       byte    %12.0g     flag       Avail. in structural enquiry 2012
se13_flag       byte    %12.0g     flag       Avail. in structural enquiry 2013
se14_flag       byte    %12.0g     flag       Avail. in structural enquiry 2014
se15_flag       byte    %12.0g     flag       Avail. in structural enquiry 2015
se16_flag       byte    %12.0g     flag       Avail. in structural enquiry 2016
zar_flag        byte    %12.0g     flag       Avail. in PETRA
death_count     byte    %10.0g                Counted in official statistics
cancelled_death byte    %8.0g                 Death set to missing (Cleaning)
m_civil         byte    %10.0g     civil_l    Marital status at death
m_ddiv          int     %dD.N.CY              (mort) Date of divorce or death of spouse
v0_buildid      long    %10.0g                v0 building number
dis_conc1_code  str1    %9s                   Concomitant disease 1, ICD8, complementary code
dis_conc2_code  str1    %9s                   Concomitant disease 2, ICD8, complementary code
cause_prim_i~10 str4    %9s                   GES-Definitive primary cause, ICD10
cause_prim_~10s str1    %9s                   GES-Definitive primary cause, ICD10 first character
cause_prim_i~2d byte    %8.0g                 GES-Definitive primary cause, ICD10 two digits
cause_prim_i~3d byte    %8.0g                 GES-Definitive primary cause, ICD10 third digit
age             float   %9.0g
nat_bin         byte    %12.0g     nat_bin_l

urban           byte    %12.0g     urban_l
lang            byte    %15.0g     lang_l   *
civil           byte    %12.0g     civil_l  *
buildid         long    %12.0g
gisid           long    %12.0g                Building ID (GIS)
geox            long    %12.0g                X coord
geoy            long    %12.0g                Y coord
year            byte    %9.0g                 Year of coordinates
dupli           int     %12.0g                Duplicate buildid
hec             byte    %9.0g                 Hectare coordinates (analytical)
d_all           byte    %9.0g                 All deaths
d_lc            byte    %9.0g                 Lung cancer
d_bc            byte    %9.0g                 Breast cancer
d_pc            byte    %9.0g                 Prostate cancer
d_re            byte    %9.0g                 Respiratory
d_cv            byte    %9.0g                 CVD
d_mi            byte    %9.0g                 MI
d_st            byte    %9.0g                 Stroke
d_ac            byte    %9.0g                 Traffic accidents
d_al            byte    %9.0g                 Alc liver disease
d_su            byte    %9.0g                 Suicide
                                            * indicated variables have notes
-----------------------------------------------------------------------------------------------------
Sorted by:
```

5.5 Last census seen

```
last_census_seen — Date of last census seen
----------------------------------------------------------------
                   |      Freq.    Percent      Valid       Cum.
-------------------+--------------------------------------------
Valid   31.12.2012 |      97123       1.85       1.85       1.85
        31.12.2013 |      99640       1.90       1.90       3.75
        31.12.2014 |      99477       1.90       1.90       5.64
        31.12.2015 |      98866       1.88       1.88       7.53
```

```
  31.12.2016 |    101768       1.94       1.94       9.47
  31.12.2017 |    105503       2.01       2.01      11.48
  31.12.2018 |   4646712      88.52      88.52     100.00
  Total      |   5249089     100.00     100.00
-----------------------------------------------------------
```

6 Appendix

6.1 Non-residential buildings

'Non-residential' buildings that were excluded from calculation of the index.

```
                  Orig. building class |      Freq.     Percent        Cum.
---------------------------------------+-----------------------------------
                   1211 - Hotel, motel |      4,906       17.69       17.69
                1220 - Office building |      3,982       14.36       32.05
1130 - Communities, home for the aged, |      3,946       14.23       46.28
    1251 - Factory, industrial building |      2,898       10.45       56.73
1212 - Short-term dwelling, youth hoste |      2,208        7.96       64.69
1271 - Farm, agricultural building, gre |      1,805        6.51       71.20
1230 - Wholesale, retail, shopping mall |      1,721        6.21       77.40
1274 - Prison, barrack, bus stop, publi |      1,707        6.16       83.56
1264 - Hospital, nursing home, institut |      1,473        5.31       88.87
1263 - School building, college, univer |      1,443        5.20       94.07
1261 - Cinema, theatre, concert hall, a |        455        1.64       95.71
          1272 - Church, chapel, morgue |        356        1.28       97.00
    1242 - Parking ramp, parking garage |        306        1.10       98.10
        1241 - Railway station, airport |        182        0.66       98.76
  1265 - Sports hall, gym, tennis court |        148        0.53       99.29
1252 - Storage building, warehouse, sil |        141        0.51       99.80
                 1262 - Museum, library |         55        0.20      100.00
             1273 - Monument, memorial |          1        0.00      100.00
---------------------------------------+-----------------------------------
                                 Total |     27,733      100.00
```