biosolids_filtering

March 20, 2024

1 Biosolids data filtering

Author: Sahar H. El Abbadi

Date: March 20, 2024

Goal: to clean biosolids data, removing permits that are likely not associated with wastewater treatment plants.

[70]:

```
# Setup

import pandas as pd
import pathlib
from utilities import check_all_sic_code # for generating SIC codes for each␣
 ↪NPDES permit
from utilities import check_for_ww_permits # check SIC code and classify as␣
 ↪"sewer_system" or "other_system"
from tqdm import tqdm
import pandoc

tqdm.pandas() # for progress bars in df.progress_applhy
```

1.1 Generate SIC biosolids dataset

1. Load biosolids dataset downloaded by Christina. Saved in 02_raw_data as Data_Download_1699657092121.csv
2. For each NPDES ID, look up all associated SIC codes.
3. Save dataframe as pickle file and as CSV

[2]:

```
# Load raw data and generate datasets

### ALERT: this takes 4+ hours to run. Comment it out and load pickle file as␣
 ↪needed.

# all_biosolids = pd.read_csv(pathlib.PurePath('01_raw_data',␣
 ↪'Data_Download_1699657092121.csv'))
#
# # test on top row
# # all_biosolids = all_biosolids.head(2).copy()
# all_biosolids['sic_permit'] = all_biosolids['NPDES ID'].
 ↪progress_apply(check_all_sic_code)
# all_biosolids.to_pickle(pathlib.PurePath('05_pickle_files',␣
 ↪'biosolids_data_sic_codes.pkl'))
# all_biosolids.to_csv((pathlib.PurePath('04_results',␣
 ↪'biosolids_data_sic_codes.csv')))
```

[15]:

```
# Load pickle file
all_biosolids = pd.read_pickle(pathlib.PurePath('05_pickle_files',␣
 ↪'biosolids_data_sic_codes.pkl'))
```

1.2 List of facilities to remove

Generate a list of facilities to remove, using the following filtering criteria:

1. Does the facility have a sewer-related SIC code? If yes –> keep
2. Is the facility listed as a POTW under its reporting obligations? If yes –> keep
3. Of remaining facilities, check SIC codes. If NO MATCH –> keep.
4. Of now remaining facilities with a non-sewer SIC match, manually keep or remove based on SIC codes

[13]:

```
# Apply filter based on SIC sewer related code
# This takes ~15 minutes to run. Load pickle to save time
all_biosolids['check_sewer_permits'] = all_biosolids['NPDES ID'].
 ↪progress_apply(check_for_ww_permits)
biosolids_to_remove = all_biosolids[all_biosolids['check_sewer_permits'] ==␣
 ↪'other_system']
biosolids_to_remove.to_pickle(pathlib.PurePath('05_pickle_files',␣
 ↪'biosolids_data_sic_codes_not_sewer.pkl'))
```

```
100%|      | 4182/4182 [14:27<00:00,  4.82it/s]
```

[17]:

```
# Load pickle for biosolids that have already been filtered based on whether or␣
 ↪not they have a sewer-related code
biosolids_not_sewer = pd.read_pickle(pathlib.PurePath('05_pickle_files',␣
 ↪'biosolids_data_sic_codes_not_sewer.pkl'))
```

[71]:

```
# Apply filter based on POTW reporting obligation

potw_mask = ~biosolids_not_sewer['Reporting Obligation(s)'].str.contains('POTW')
biosolids_not_sewer_not_potw = biosolids_not_sewer[potw_mask]

display(biosolids_not_sewer_not_potw[['Facility Name', 'NPDES ID',␣
 ↪'sic_permit', 'check_sewer_permits']])
```

```
                     Facility Name   NPDES ID  \
14      AUSTIN COUNTY WSC  PLANT 3  TX0125709
15          LAKE PFLUGERVILLE WWTF  TX0132721
16    PURTIS CREEK STATE PARK WWTP  TX0082856
```

```
25              CHISOS BASIN  WWTP  TX0094684
35     LAUGHLIN AFB WWTP BLDG 1004  TX0022608
…                             …         …
4079         DAVIES MOBILE PARK LLC  COL621009
4084          MANCHESTER BY THE SEA  MAL100871
4114              COHASSET W W T P*  MAL100285
4120                 CITY OF GRANBY  MOL107581
4135      BLUE SKY RANCH AND RESORT  UTL025763

                                             sic_permit check_sewer_permits
14                                         [4941, 4941]        other_system
15                                   [1541, 4941, 4941]        other_system
16                                         [7033, 7033]        other_system
25                                         [7999, 7999]        other_system
35    [9711, 1542, 9711, 9711, 1542, 4581, 9711, 971…        other_system
…                                                   …                 …
4079                                      [NO_SIC_MATCH]        other_system
4084                                      [NO_SIC_MATCH]        other_system
4114                                      [NO_SIC_MATCH]        other_system
4120                                      [NO_SIC_MATCH]        other_system
4135                                         [7011, 7011]        other_system

[316 rows x 4 columns]
```

[74]:

```
# Keep facilities that have no match (ie remove them from our list of␣
 ↪facilities to remove)

# These facilities have a match with an SIC code that is NOT sewer-related
biosolids_not_sewer_not_potw_has_match =␣
 ↪biosolids_not_sewer_not_potw[biosolids_not_sewer_not_potw['sic_permit'].
 ↪apply(lambda x: 'NO_SIC_MATCH' not in x)]
print(f'Length of dataframe (not sewer, not POTW, has SIC match):␣
 ↪{len(biosolids_not_sewer_not_potw_has_match)}')
```

```
Length of dataframe (not sewer, not POTW, has SIC match): 226
```

[38]:

```
# Check how many facilities have no match after previous filtering

biosolids_not_sewer_not_potw_no_match =␣
 ↪biosolids_not_sewer_not_potw[biosolids_not_sewer_not_potw['sic_permit'].
 ↪apply(lambda x: 'NO_SIC_MATCH' in x)]

print(f'Length of dataframe (not sewer, not POTW, no SIC match):␣
 ↪{len(biosolids_not_sewer_not_potw_no_match)}')
```

```
Length of dataframe (not sewer, not POTW, no SIC match): 90
```

1.2.1 Remove facilities with problematic SIC codes

Remove facilities with SIC codes that are unlikely to be associated with publicly owned wastewater treatment facilities

[72]:

```
# Apply filter based on reporting obligation
sic_remove = [6515, # mobile homes
              4941, # water supply
              8211, # schools
              8221, # colleges & universities
              7033, # trailer parks / campsites
              7032, # sporting and recreation camps
              9223, # correctional facilities
              1389, # oil & gas field services
              3533, # oil and gas field machinery
              8361, # residential care
              8661, # religious orgs
              7997, # sports / recreation clubs
              7999, # amusement and recreation
              8051, # skilled nursing care
              3498, # fabricated pipe & fitting
              7011, # hotels and motels
              3171, # handbags & purses
              2491, # wood preserving
              2493, # reconsistuted wood products
              9711, # national security
              3743, # railroad equipment
              5541, # gas station services
              4911, # electric services
              5075, # heating & cooling
              7041, # membership hotels
              2011, # meat packing plants
              8063, # psychiatric hospitals
              5812, # eating places
              7999, # amusement parks
              2899, # chemical preparation (spice / food extraction)
              3331, # primary copper
              6531, # real estate agents & managers
              4011, # railroads
              6514, # dwelling operators (residential)
              2621, # paper mills
              4581, # airports
              1522, # residential construction
              ]

sic_check = [1629, # heavy construction
             9511, # air, water, solid waste management
             9199, # general government
             7299, # misc. personal services
             2819, #
             ]

# Check SIC codes for facilities that have an SIC code match
biosolids_not_sewer_not_potw_has_match_sic_removal =␣
 ↪biosolids_not_sewer_not_potw_has_match[biosolids_not_sewer_not_potw_has_match['sic_permit']
 ↪apply(lambda x: any(item in sic_remove for item in x))]
display(biosolids_not_sewer_not_potw_has_match_sic_removal[['Facility Name',␣
 ↪'NPDES ID', 'sic_permit', 'check_sewer_permits']])
```

```
                     Facility Name   NPDES ID  \
14      AUSTIN COUNTY WSC  PLANT 3  TX0125709
15          LAKE PFLUGERVILLE WWTF  TX0132721
16     PURTIS CREEK STATE PARK WWTP  TX0082856
25               CHISOS BASIN  WWTP  TX0094684
35      LAUGHLIN AFB WWTP BLDG 1004  TX0022608
…                              …         …
3189         ATK LAUNCH SYSTEMS INC  UTL024805
3800                    OAKELY CITY  UTL020061
3867     KENNECOTT UTAH COPPER, LLC  UTL000051
3870           LYSTEK INTERNATIONAL  CAL000001
4135      BLUE SKY RANCH AND RESORT  UTL025763

                                                sic_permit check_sewer_permits
14                                            [4941, 4941]        other_system
15                                      [1541, 4941, 4941]        other_system
16                                            [7033, 7033]        other_system
25                                            [7999, 7999]        other_system
35    [9711, 1542, 9711, 9711, 1542, 4581, 9711, 971…       other_system
…                                                 …                   …
3189  [7549, 3714, 3769, 3714, 3761, 3769, 3764, 754…       other_system
3800                                          [2899, 2899]        other_system
3867                            [3331, 3331, 1021, 3331, 1021]        other_system
3870  [7538, 7538, 4212, 4212, 7513, 7513, 8211, 399…       other_system
4135                                          [7011, 7011]        other_system

[221 rows x 4 columns]
```

[73]:

```
biosolids_not_sewer_not_potw_has_match_sic_check =␣
 ↪biosolids_not_sewer_not_potw_has_match[~biosolids_not_sewer_not_potw_has_match['sic_permit'
 ↪apply(lambda x: any(item in sic_remove for item in x))]
display(biosolids_not_sewer_not_potw_has_match_sic_check[['Facility Name',␣
 ↪'NPDES ID', 'sic_permit', 'check_sewer_permits']])
```

```
                              Facility Name   NPDES ID  \
554   LIVE OAK COUNTY SAFETY REST AREA WWTF  TX0129321
938                         BAYOU CLUB WWTP  TX0083933
```

```
1082                 GE PACKAGED POWER JPORT  TX0101656
1462                           SIGMAPRO WWTP  TX0138754
2856             US DOE/SAVANNAH RIVER SITE  SCL000175

                                      sic_permit check_sewer_permits
554                                 [7299, 7299]        other_system
938                                       [8641]        other_system
1082  [3511, 3511, 7699, 3511, 7699, 3511, 7699]        other_system
1462                                      [6519]        other_system
2856  [2819, 2819, 9611, 2819, 2819, 2819, 2819]        other_system
```

1.2.2 Manually check the remaining facilities

The dataset of biosolids_not_sewer_not_potw_has_sic_check contains the facilities that I'm not confident removing based solely on their SIC codes. Manually inspect facilities and decide where they should be kept based on name / information available online

1. Live Oak County Safety Rest Area WWTF - code 7299 (misc personal services) –> remove, rest area along highway
2. Bayou Club WWTP - code 8641 (civic & social associations) –> remove, dining club
3. GE Packaged Power Jport - code 3511 (turbines / turbine generators), 7699 (repair services) –> probably remove, GE and not public
4. Sigmapro WWTP - code 6519 (real property lessors) –> Sigma Pro private company WWTP, not public
5. US DOE / Savannah River Site - codes 2819 (industrial inorganic chemicals), 9611 (administration of general economic programs) –> remove, not a public wastewater treatment facility

Based on this online search, we can remove all facilities in the original subset biosolids_not_sewer_not_potw_has_match (before filtering based on the specific SIC codes of concern).

```
[69]: biosolids_to_remove = biosolids_not_sewer_not_potw_has_match

biosolids_to_remove.to_csv(pathlib.PurePath('04_results', 'biosolids_to_remove.
 ↪csv'), index=False)
biosolids_to_remove.to_pickle(pathlib.PurePath('05_pickle_files',␣
 ↪'biosolids_to_remove.pkl'))
```