Code Description

Updating time: 2021.11.13.
We revised some bugs of the codes. We have also added some more descriptions (blue words) to this file to ensure that users can use it better.

1. This code is from the paper

<Urban hotspots detection of taxi stops with local maximum density> in Computers, Environment and Urban System
https://doi.org/10.1016/j.compenvurbsys.2021.101661
If there are any questions about the works or the codes, please do not hesitate to contact Xiao-Jian Chen (cxiaojian@whu.edu.cn)

2. Modules required: (All versions are OK, except pandas)

(1) pandas 1.3.2 (The multi-index of the dataframe in Pandas seems to have changed, causing the program to fail to run.)
(2) numpy
(3) pickle
(4) scipy
(5) joblib
(6) os
(7) itertools

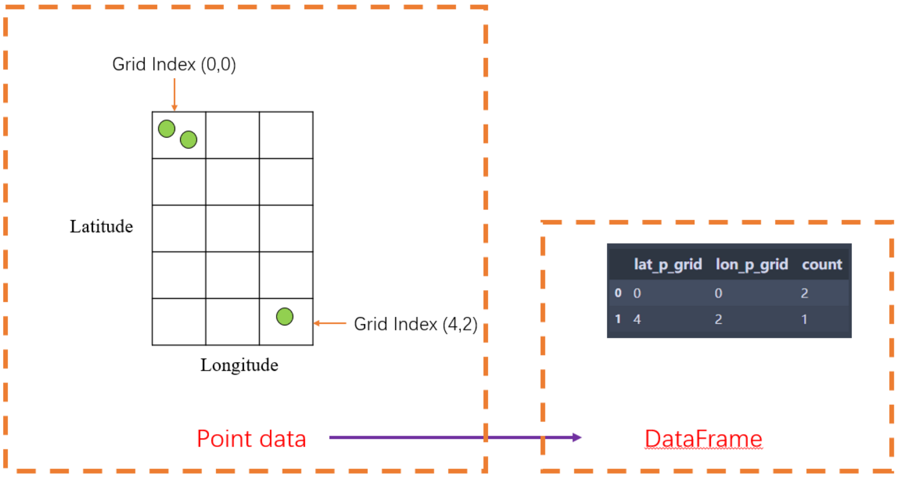
3. There are two .py programs

(1) Neighborhood_size_1.py
Description: Guidance to select grid range of a neighborhood (i.e., the radius of the neighborhood size)
(2) LMD_2.py
Description: Detect local hotspots

4. How to use it?

(1) Codes are designed to deal with square area

(2) Points data are required to be prepared by users as grid-based data before it can be processed by our codes, like below. (Notice that the coordinates of each stop are recommended to transform to projected coordinate system before transforming into grid-based data.)



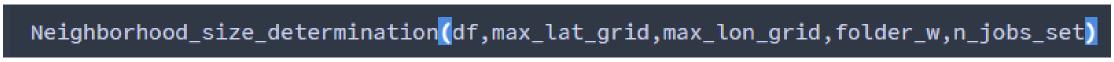
	lat_p_grid	lon_p_grid	count
0	0	0	2
1	4	2	1

(3) '.pickle' files (e.g. '1.pickle') can be loaded by pickle.load(open('1.pickle','rb'))
'.pkl' files (e.g. '2.pkl') can be loaded by pandas.read_pickle('2.pkl')

(4) How to use "Neighborhood_size_1.py"?

Function

Neighborhood_size_determination(df,max_lat_grid,max_lon_grid,folder_w,n_jobs_set)



```
Neighborhood_size_determination(df,max_lat_grid,max_lon_grid,folder_w,n_jobs_set)
```

Variable Description：

a) df

Observed grid-based data as DataFrame presented like above (2).

Note: The dataframe input by the user can only include "count! =0" grid, such as the example in (2) above. However, even if the user enters all grid data (including "count==0"), it is allowed. We have added a filter command at the beginning of the code:
df=df.query('count>0').reset_index(drop=True)

b) max_lat_grid

The maximum index in direction of latitude. For the exmaple above, the max_lat_grid=4

c) max_lon_grid

The maximum index in direction of longitude. For the exmaple above, the max_lon_grid=2

d) folder_w

Folder used to store the result.

Note: At the end of the folder, please add '/'. For example,
'/home/research/1-Neighoborhood/' (correct)
'/home/research/1-Neighoborhood' (wrong)

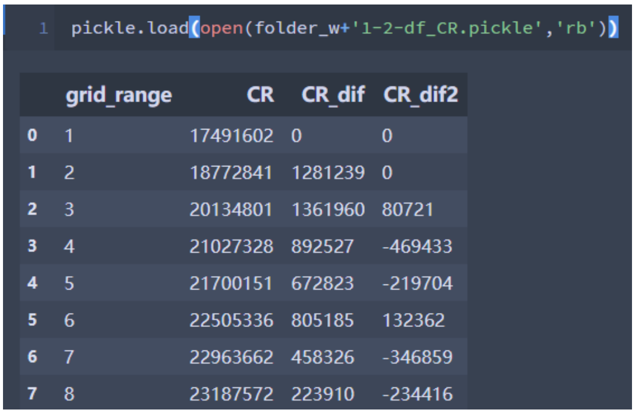
e) n_jobs_set

The number of parallelism used, "-1" means that all CPUs are used. Details can be refered in the python module "joblib".

Useful final output:

a) 1-2-df_CR.pickle

# The result of the final grid range of neighborhood is based on this detailed data (see b 1-2-result_grid_range.pickle).



```
pickle.load(open(folder_w+'1-2-df_CR.pickle','rb'))
```

	grid_range	CR	CR_dif	CR_dif2
0	1	17491602	0	0
1	2	18772841	1281239	0
2	3	20134801	1361960	80721
3	4	21027328	892527	-469433
4	5	21700151	672823	-219704
5	6	22505336	805185	132362
6	7	22963662	458326	-346859
7	8	23187572	223910	-234416

a.1) grid_range: grid range of a neighobor.

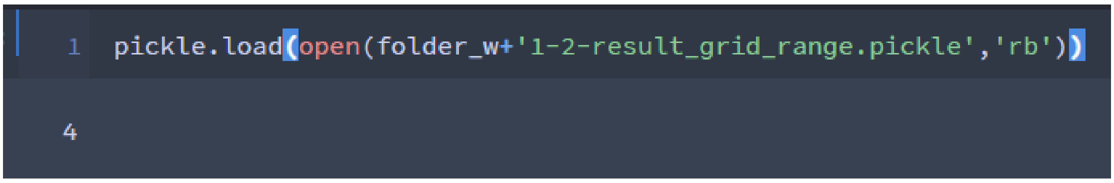
a.2) CR: number of stops covered by local hotspots (i.e., formula (3) in the paper)

a.3) CR_dif: the first derivation of CR. (Notice that grid_range=1 has no corresponding value, as such being assigned by 0)

a.4) CR_dif2: the second derivation of CR. (Notice that grid_range=1 and 2 have no corresponding value, as such being assigned by 0)

b) 1-2-result_grid_range.pickle

#The recommended grid range (i.e., the radius of the neighborhood size)



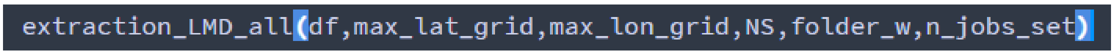
```
pickle.load(open(folder_w+'1-2-result_grid_range.pickle','rb'))

4
```

(5) How to use “LMD_2”?

Function

extraction_LMD_all(df,max_lat_grid,max_lon_grid,NS,folder_w,n_jobs_set)



```
extraction_LMD_all(df,max_lat_grid,max_lon_grid,NS,folder_w,n_jobs_set)
```

Variable Description：

a) df

Observed grid-based data as DataFrame presented like above (2).

b) max_lat_grid

The maximum index in direction of latitude. For the exmaple above, the max_lat_grid=4

c) max_lon_grid

The maximum index in direction of longitude. For the exmaple above, the max_lon_grid=2

d) NS

The radius of neighborhood size

e) folder_w

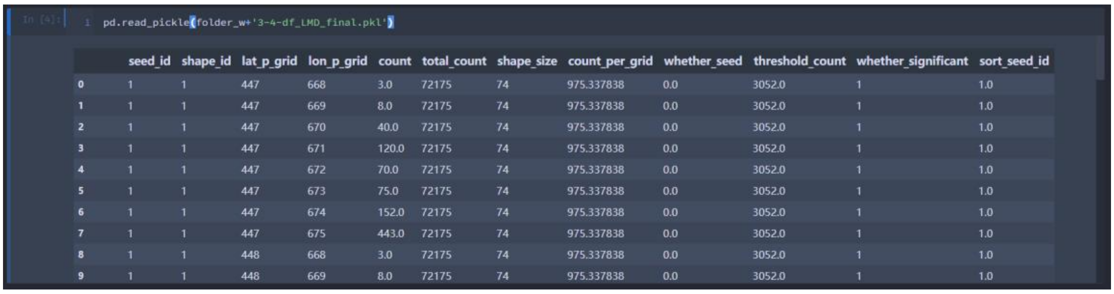
Folder used to store the result

f) n_jobs_set

The number of parallelism used, "-1" means that all CPUs are used. Details can be refered in the python module "joblib"

Useful final output:

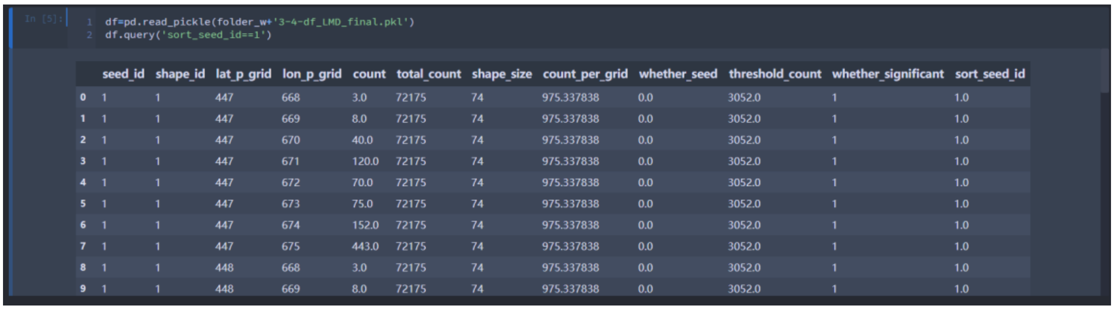
a) 3-4-df_LMD_final.pkl



```
In [4]:  pd.read_pickle(folder_w+'3-4-df_LMD_final.pkl')
```

	seed_id	shape_id	lat_p_grid	lon_p_grid	count	total_count	shape_size	count_per_grid	whether_seed	threshold_count	whether_significant	sort_seed_id
0	1	1	447	668	3.0	72175	74	975.337838	0.0	3052.0	1	1.0
1	1	1	447	669	8.0	72175	74	975.337838	0.0	3052.0	1	1.0
2	1	1	447	670	40.0	72175	74	975.337838	0.0	3052.0	1	1.0
3	1	1	447	671	120.0	72175	74	975.337838	0.0	3052.0	1	1.0
4	1	1	447	672	70.0	72175	74	975.337838	0.0	3052.0	1	1.0
5	1	1	447	673	75.0	72175	74	975.337838	0.0	3052.0	1	1.0
6	1	1	447	674	152.0	72175	74	975.337838	0.0	3052.0	1	1.0
7	1	1	447	675	443.0	72175	74	975.337838	0.0	3052.0	1	1.0
8	1	1	448	668	3.0	72175	74	975.337838	0.0	3052.0	1	1.0
9	1	1	448	669	8.0	72175	74	975.337838	0.0	3052.0	1	1.0

Notice: Each line records the information of a grid. If you want to select the local hotspot which contains the maximum number of stops, you can use the following "query" command:



```
In [5]:  df=pd.read_pickle(folder_w+'3-4-df_LMD_final.pkl')
         df.query('sort_seed_id==1')
```

	seed_id	shape_id	lat_p_grid	lon_p_grid	count	total_count	shape_size	count_per_grid	whether_seed	threshold_count	whether_significant	sort_seed_id
0	1	1	447	668	3.0	72175	74	975.337838	0.0	3052.0	1	1.0
1	1	1	447	669	8.0	72175	74	975.337838	0.0	3052.0	1	1.0
2	1	1	447	670	40.0	72175	74	975.337838	0.0	3052.0	1	1.0
3	1	1	447	671	120.0	72175	74	975.337838	0.0	3052.0	1	1.0
4	1	1	447	672	70.0	72175	74	975.337838	0.0	3052.0	1	1.0
5	1	1	447	673	75.0	72175	74	975.337838	0.0	3052.0	1	1.0
6	1	1	447	674	152.0	72175	74	975.337838	0.0	3052.0	1	1.0
7	1	1	447	675	443.0	72175	74	975.337838	0.0	3052.0	1	1.0
8	1	1	448	668	3.0	72175	74	975.337838	0.0	3052.0	1	1.0
9	1	1	448	669	8.0	72175	74	975.337838	0.0	3052.0	1	1.0

a.1) seed_id: The original id of a local hotspot used in the previous process.

a.2) shape_id: The id of local hotspot's shape which is used in the previous process.

a.3) lat_p_grid: The index of grid in the latitude direction.

a.4) lon_p_grid: The index of grid in the longitude direction.

a.5) count: The number of stops in the grid

a.6) total_count: The number of stops in the corresponding local hotspot

a.7) shape_size: The number of the grid for the corresponding shape of the local hotspot.

a.8) count_per_grid: The average number of stops in each grid for the local hotspot.

a.9) whether_seed: "1" means the local maximum grid of the local hotspot. "0" means the grid is not the local maximum of the local hotspot.

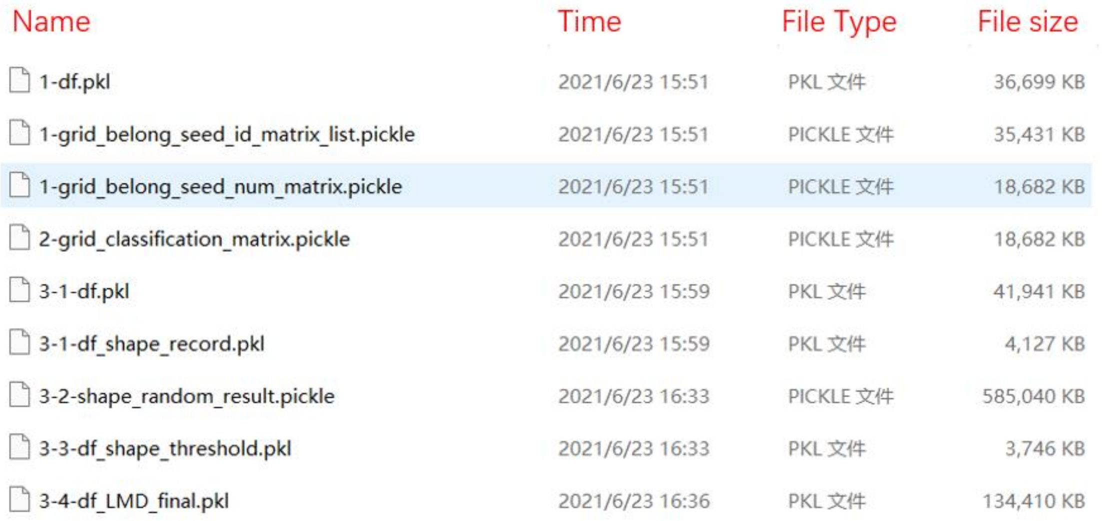
a.10) threshold_count: The threshold of count (in Step3: Popular local hotspots determination of the paper) for the corresponding shape.

a.11) whether_significant: "1" means the local hotspot is popular (i.e., "total_count>= threshold_count"). "0" means the local hotspot is not popular enough to be retained. (i.e., "total_count<threshold_count")

a.12) sort_seed_id: The id of the local hotspot. The id is sorted descending by the total_count. Therefore, the smaller of sort_seed_id the more stops the local hotspot has. For those local hotspots with "whether_significant=0", "sort_seed_id= -100"

5. Running program duration

It depends on the different size of the grid-based data. For our case here, the max_lon_grid=2239, max_lat_grid=2134. The duration is the one like following:



Name	Time	File Type	File size
1-df.pkl	2021/6/23 15:51	PKL 文件	36,699 KB
1-grid_belong_seed_id_matrix_list.pickle	2021/6/23 15:51	PICKLE 文件	35,431 KB
1-grid_belong_seed_num_matrix.pickle	2021/6/23 15:51	PICKLE 文件	18,682 KB
2-grid_classification_matrix.pickle	2021/6/23 15:51	PICKLE 文件	18,682 KB
3-1-df.pkl	2021/6/23 15:59	PKL 文件	41,941 KB
3-1-df_shape_record.pkl	2021/6/23 15:59	PKL 文件	4,127 KB
3-2-shape_random_result.pickle	2021/6/23 16:33	PICKLE 文件	585,040 KB
3-3-df_shape_threshold.pkl	2021/6/23 16:33	PKL 文件	3,746 KB
3-4-df_LMD_final.pkl	2021/6/23 16:36	PKL 文件	134,410 KB