KMeans clustering on correlations

In [9]:

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns

from sklearn.cluster import KMeans
%matplotlib inline
plt.style.use('seaborn')
```

In [19]:

```
# read saved dataset.
def get_monthly_result(month='nov', n_clusters=3):
    #month = 'nov'
    df_normalized = pd.read_csv(f'data/normalized_{month}.csv')
    df_normalized = df_normalized.drop(columns=['Unnamed: 0'])
    df_normalized.head()

    # get groups and keys on dataset.

    new_groups = df_normalized.groupby('meterid')
    new_keys = new_groups.groups.keys()  # keys: an iterable of dataids or meter ids

    new_id_list = list(new_keys)
    display('number of valid meterids:', len(new_id_list))

    #  Construct dataframe over which to run correlation analysis.

    df_total = pd.DataFrame()
    display(df_total)
    for key in new_id_list:
        df_i = new_groups.get_group(key).rename(columns={'norm_cumul_value': f'{key}'})
.reset_index()
        df_total = pd.concat([df_total, df_i[f'{key}']], axis=1)

    display(len(df_total))
    display(df_total)  # column names are meterids

    # get corr matrix

    df_total = df_total.fillna(0)
    df_corr = df_total.corr()  # get correlations between meterids for the month's cumu
lative vals.
    df_corr = df_corr.fillna(0)
    display(df_corr)

    # find top n positive correlations per meterid
    n = 10
    for meterid in new_id_list[1:2]:
        # slice is to show an example.
        df_corr_sorted = df_corr[f'{meterid}'].sort_values(ascending=False)
        sr_top5 = df_corr_sorted.iloc[1:1+n]
        df_top5 = sr_top5.to_frame()
        df_top5 = df_top5.rename(columns={f'{meterid}': 'r'})

        data = [meterid for i in range(n)]
        df_top5 = df_top5.assign(id=data)
        display(df_top5)

    # clustering

    #n_clusters = 4
    kmeans = KMeans(n_clusters=n_clusters)
    kmeans.fit(df_corr)
    labels = kmeans.labels_
    print(labels)

    origin_date = df_normalized['index'].iloc[0]

    for cluster_label in range(n_clusters):
```

```
        fig, axes = plt.subplots(nrows=1, ncols=1, figsize=(15,3))
        axes.set_title(f'cluster {cluster_label}; count: {list(labels).count(cluster_la
bel)}')
        axes.set_xlabel(f'hours elapsed since {origin_date}')
        axes.set_ylabel(f'cumul value')
        i = 0
        for key in new_id_list:
            if labels[i] == cluster_label:
                axes.scatter(df_total.index, df_total[f'{key}'])

            i += 1
```

In [20]:

```
months = ['nov', 'dec', 'jan', 'feb', 'mar']
n_clusters = 3
for month in months:
    get_monthly_result(month, n_clusters)
```

```
'number of valid meterids:'
```

```
118
```

□

```
721
```

	35	44	77	94	114	187	252	370	483	484	...	9
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0
1	12.0	0.0	0.0	0.0	30.0	0.0	2.0	8.0	0.0	0.0	...	2.0
2	12.0	0.0	8.0	0.0	30.0	0.0	2.0	8.0	0.0	0.0	...	2.0
3	14.0	0.0	8.0	0.0	30.0	0.0	2.0	10.0	0.0	2.0	...	2.0
4	14.0	0.0	8.0	2.0	30.0	0.0	2.0	10.0	0.0	2.0	...	4.0
...	...	...	...	...	...	...	...	...	...	...	...	...
716	1670.0	1588.0	854.0	2934.0	1960.0	2684.0	2878.0	948.0	2680.0	1772.0	...	33
717	1676.0	1588.0	854.0	2934.0	1960.0	2684.0	2880.0	948.0	2690.0	1772.0	...	33
718	1684.0	1588.0	860.0	2934.0	1960.0	2684.0	2920.0	948.0	2706.0	1772.0	...	33
719	1686.0	1590.0	860.0	2958.0	1960.0	2684.0	2930.0	948.0	2708.0	1776.0	...	33
720	1686.0	1590.0	864.0	2976.0	1960.0	2684.0	2946.0	948.0	2710.0	1788.0	...	33

721 rows × 118 columns

	35	44	77	94	114	187	252	370
35	1.000000	0.946914	0.983875	0.961489	0.988778	0.984140	0.952977	0.939017
44	0.946914	1.000000	0.953504	0.930343	0.940092	0.907403	0.899632	0.850495
77	0.983875	0.953504	1.000000	0.919246	0.959442	0.953459	0.897538	0.866205
94	0.961489	0.930343	0.919246	1.000000	0.988503	0.964625	0.994359	0.942711
114	0.988778	0.940092	0.959442	0.988503	1.000000	0.985998	0.983448	0.948053
...	...	...	...	...	...	...	...	...
9639	0.985468	0.941353	0.955033	0.991472	0.997169	0.983694	0.986083	0.948629
9729	0.968704	0.907196	0.916490	0.990795	0.989162	0.977188	0.996008	0.973600
9766	0.981064	0.937795	0.952122	0.991497	0.995611	0.979081	0.983186	0.938095
9849	0.994784	0.948361	0.989380	0.955921	0.982673	0.976500	0.940615	0.912749
9982	0.994337	0.955438	0.981651	0.972365	0.990544	0.980455	0.958236	0.925410

118 rows × 118 columns

	r	id
2575	0.965928	44
2980	0.959013	44
2945	0.956767	44
5484	0.956571	44
9982	0.955438	44
77	0.953504	44
7739	0.952377	44
484	0.950564	44
8084	0.948597	44
9849	0.948361	44

```
[1 1 1 0 0 0 0 2 2 0 0 0 0 2 1 0 2 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0
 2 0 0 0 1 1 1 0 0 2 0 0 0 0 0 0 0 0 2 0 0 2 0 0 0 0 0 0 2 0 0 2 1 2 0 0 0
 1 0 0 0 0 0 0 0 2 1 1 0 0 2 0 2 0 0 0 1 1 0 0 0 1 0 1 0 0 0 0 0 2 0 0 0 2
 0 0 0 0 0 1 1]
```

```
'number of valid meterids:'
```

```
119
```

□

```
745
```

	35	44	77	94	114	187	252	370	483	484	...	
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.
1	6.0	0.0	0.0	14.0	0.0	0.0	14.0	62.0	16.0	0.0	...	1(
2	38.0	4.0	8.0	22.0	0.0	0.0	18.0	64.0	16.0	4.0	...	1(
3	38.0	18.0	10.0	38.0	0.0	0.0	20.0	64.0	16.0	24.0	...	1(
4	38.0	28.0	10.0	46.0	0.0	0.0	32.0	66.0	16.0	26.0	...	1(
...	...	...	...	...	...	...	...	...	...	...	...	...
740	1968.0	1760.0	986.0	5626.0	3952.0	5290.0	4766.0	2728.0	4642.0	2578.0	...	5(
741	1968.0	1762.0	992.0	5644.0	3952.0	5290.0	4774.0	2730.0	4642.0	2606.0	...	5(
742	1968.0	1762.0	992.0	5650.0	3960.0	5290.0	4776.0	2732.0	4656.0	2624.0	...	5
743	1970.0	1770.0	994.0	5658.0	3960.0	5290.0	4776.0	2732.0	4664.0	2636.0	...	5
744	1970.0	1780.0	996.0	5682.0	3960.0	5290.0	4776.0	2734.0	4686.0	2644.0	...	5

745 rows × 119 columns

	35	44	77	94	114	187	252	370
35	1.000000	0.957005	0.975981	0.996383	0.993525	0.988584	0.994915	0.990921
44	0.957005	1.000000	0.954906	0.954778	0.958965	0.942081	0.950317	0.964298
77	0.975981	0.954906	1.000000	0.983559	0.974189	0.965539	0.971541	0.975798
94	0.996383	0.954778	0.983559	1.000000	0.996680	0.992686	0.996606	0.992827
114	0.993525	0.958965	0.974189	0.996680	1.000000	0.993535	0.997375	0.992187
...	...	...	...	...	...	...	...	...
9639	0.996081	0.944495	0.977365	0.997040	0.992907	0.990423	0.994227	0.983227
9729	0.994440	0.939653	0.955361	0.991828	0.992272	0.990091	0.995555	0.984422
9766	0.994598	0.954422	0.969726	0.993714	0.992537	0.989365	0.991251	0.984771
9849	0.978881	0.969699	0.986814	0.981692	0.979010	0.964918	0.976611	0.988710
9982	0.990150	0.934643	0.973759	0.995511	0.991595	0.989387	0.993464	0.985700

119 rows × 119 columns

	r	id
1800	0.970938	44
5484	0.969939	44
9849	0.969699	44
3635	0.967301	44
2945	0.966984	44
744	0.965922	44
8386	0.965727	44
370	0.964298	44
7682	0.964245	44
2980	0.963493	44

```
[0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 2 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 1 1 0 0 1 0 1 0 1 1 0 0 0 0 0 0
 1 0 0 0 0 0 1 0]
```

```
'number of valid meterids:'
```

```
119
```

□

```
745
```

	35	44	77	94	114	187	252	370	483	484	...	
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0
1	2.0	0.0	0.0	22.0	0.0	0.0	0.0	2.0	0.0	4.0	...	1
2	4.0	2.0	0.0	24.0	0.0	0.0	0.0	8.0	14.0	20.0	...	1
3	6.0	2.0	0.0	28.0	0.0	0.0	0.0	12.0	36.0	42.0	...	4
4	24.0	2.0	0.0	30.0	0.0	0.0	2.0	12.0	52.0	60.0	...	7
...	...	...	...	...	...	...	...	...	...	...	...	.
740	3238.0	2952.0	1138.0	7304.0	6174.0	7756.0	7814.0	4750.0	9390.0	4272.0	...	8
741	3238.0	2974.0	1138.0	7308.0	6174.0	7756.0	7814.0	4750.0	9392.0	4286.0	...	8
742	3240.0	2974.0	1138.0	7310.0	6178.0	7756.0	7816.0	4752.0	9392.0	4292.0	...	8
743	3254.0	2974.0	1142.0	7310.0	6178.0	7756.0	7818.0	4752.0	9408.0	4294.0	...	8
744	3270.0	2974.0	1142.0	7310.0	6192.0	7756.0	7822.0	4752.0	9408.0	4294.0	...	8

745 rows × 119 columns

	35	44	77	94	114	187	252	370
35	1.000000	0.996320	0.998197	0.992227	0.998622	0.993625	0.997647	0.995658
44	0.996320	1.000000	0.996913	0.994517	0.996907	0.996195	0.995531	0.994842
77	0.998197	0.996913	1.000000	0.991623	0.998441	0.994576	0.998109	0.997283
94	0.992227	0.994517	0.991623	1.000000	0.992702	0.997194	0.989391	0.990280
114	0.998622	0.996907	0.998441	0.992702	1.000000	0.995471	0.998965	0.997948
...	...	...	...	...	...	...	...	...
9639	0.998562	0.996287	0.998615	0.990581	0.999442	0.994115	0.999528	0.998439
9729	0.998498	0.995368	0.997985	0.989784	0.999053	0.993139	0.999410	0.997946
9766	0.996586	0.996552	0.997594	0.993841	0.997514	0.995653	0.996445	0.996326
9849	0.997105	0.995552	0.996963	0.993225	0.995137	0.993005	0.993450	0.992010
9982	0.998214	0.997173	0.998869	0.992848	0.999442	0.996164	0.998781	0.998635

119 rows × 119 columns

	r	id
5892	0.997839	44
5785	0.997766	44
5439	0.997738	44
7117	0.997680	44
2378	0.997565	44
7287	0.997521	44
6910	0.997516	44
2018	0.997500	44
1714	0.997465	44
9134	0.997448	44

```
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 1 0 2 0 2 0 0 0 0 0 0 0
 2 0 0 0 0 0 0 0]
```

```
'number of valid meterids:'
```

```
118
```

□

```
697
```

	35	44	77	94	114	187	252	370	483	484	...	
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	(
1	0.0	0.0	8.0	0.0	2.0	0.0	34.0	6.0	0.0	0.0	...	(
2	12.0	0.0	10.0	0.0	14.0	0.0	34.0	6.0	0.0	22.0	...	(
3	12.0	0.0	10.0	0.0	14.0	0.0	34.0	8.0	0.0	44.0	...	(
4	12.0	28.0	10.0	2.0	14.0	0.0	34.0	10.0	0.0	46.0	...	(
...	...	...	...	...	...	...	...	...	...	...	...	.
692	1886.0	2006.0	1038.0	4504.0	2520.0	4326.0	3580.0	2448.0	3868.0	2734.0	...	4
693	1886.0	2006.0	1038.0	4504.0	2520.0	4326.0	3582.0	2448.0	3868.0	2736.0	...	4
694	1886.0	2006.0	1038.0	4504.0	2520.0	4326.0	3582.0	2462.0	3888.0	2736.0	...	4
695	1886.0	2006.0	1038.0	4506.0	2520.0	4326.0	3584.0	2462.0	3898.0	2736.0	...	4
696	1888.0	2006.0	1040.0	4506.0	2528.0	4342.0	3586.0	2462.0	3898.0	2748.0	...	4

697 rows × 118 columns

	35	44	77	94	114	187	252	370
35	1.000000	0.928054	0.995099	0.993597	0.983746	0.984711	0.980459	0.991785
44	0.928054	1.000000	0.935642	0.929728	0.913702	0.904353	0.909700	0.921810
77	0.995099	0.935642	1.000000	0.989228	0.975610	0.977124	0.968558	0.990526
94	0.993597	0.929728	0.989228	1.000000	0.996072	0.995777	0.993892	0.995973
114	0.983746	0.913702	0.975610	0.996072	1.000000	0.997954	0.996659	0.992725
...	...	...	...	...	...	...	...	...
9639	0.991435	0.924835	0.987072	0.999538	0.996357	0.996252	0.995108	0.994908
9729	0.989412	0.922948	0.980041	0.997674	0.996914	0.995377	0.997808	0.991091
9766	0.986318	0.920418	0.978783	0.997180	0.997775	0.996806	0.996573	0.992104
9849	0.991105	0.923171	0.996891	0.983740	0.970642	0.973632	0.961231	0.987835
9982	0.993778	0.936351	0.993138	0.997804	0.992366	0.992899	0.986985	0.997239

118 rows × 118 columns

	r	id
8155	0.948046	44
5785	0.945293	44
9982	0.936351	44
77	0.935642	44
1415	0.935328	44
744	0.934480	44
2449	0.933765	44
4228	0.933536	44
7682	0.931510	44
1697	0.931301	44

```
[2 0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 0 2 2 2 0 2 2 2 2
 2 0 2 2 2 2 2 2 1 2 2 2 2 0 2 2 2 2 2 2 0 2 2 2 2 0 2 2 2 0 2 2 2 2 2 2 0
 2 2 2 2 2 0 2]
```

```
'number of valid meterids:'
```

```
90
```

□

```
745
```

	35	77	94	252	370	483	484	661	739	744	...	
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.
1	18.0	0.0	6.0	0.0	0.0	0.0	18.0	0.0	0.0	14.0	...	0.
2	28.0	0.0	6.0	0.0	20.0	0.0	20.0	2.0	0.0	16.0	...	2.
3	28.0	10.0	6.0	0.0	22.0	0.0	22.0	2.0	0.0	16.0	...	2.
4	28.0	10.0	20.0	2.0	24.0	0.0	22.0	2.0	2.0	18.0	...	2.
...	...	...	...	...	...	...	...	...	...	...	...	...
740	1412.0	1070.0	2324.0	2196.0	1616.0	1288.0	2290.0	2418.0	994.0	2718.0	...	1(
741	1412.0	1074.0	2324.0	2198.0	1616.0	1288.0	2302.0	2418.0	994.0	2720.0	...	1(
742	1412.0	1074.0	2326.0	2198.0	1616.0	1292.0	2302.0	2422.0	994.0	2722.0	...	1(
743	1412.0	1074.0	2326.0	2238.0	1616.0	1292.0	2302.0	2424.0	994.0	2724.0	...	1(
744	1422.0	1074.0	2326.0	2238.0	1616.0	1306.0	2304.0	2424.0	994.0	2724.0	...	1(

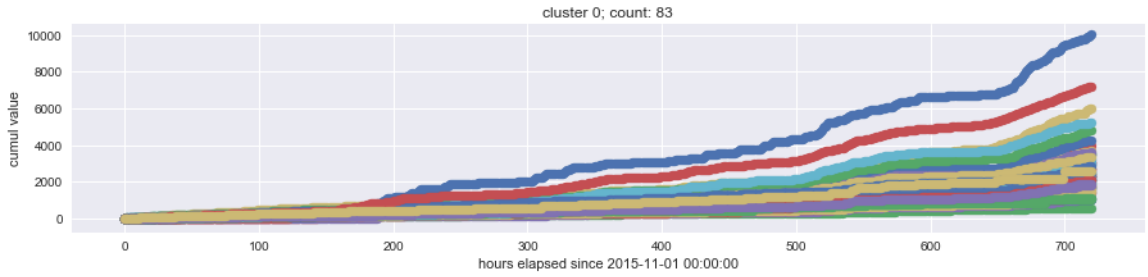
745 rows × 90 columns

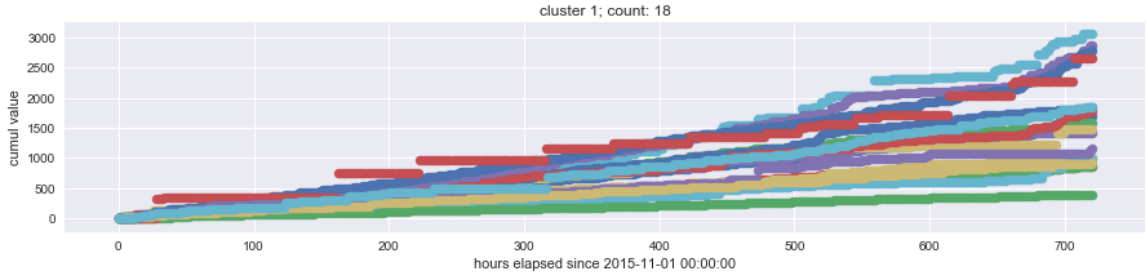
	35	77	94	252	370	483	484	661
35	1.000000	0.986832	0.994172	0.992399	0.990386	0.986098	0.986262	0.984275
77	0.986832	1.000000	0.990723	0.976826	0.999281	0.986763	0.999046	0.968092
94	0.994172	0.990723	1.000000	0.994106	0.992790	0.995710	0.990411	0.990909
252	0.992399	0.976826	0.994106	1.000000	0.981265	0.988835	0.977321	0.996945
370	0.990386	0.999281	0.992790	0.981265	1.000000	0.988162	0.998874	0.972638
...	...	...	...	...	...	...	...	...
9278	0.989821	0.988920	0.988360	0.986288	0.991590	0.980223	0.987161	0.978315
9295	0.993222	0.997032	0.993354	0.984633	0.997933	0.988018	0.997143	0.975705
9631	0.992906	0.981299	0.990810	0.989432	0.984704	0.984860	0.983445	0.982782
9729	0.991541	0.989148	0.998105	0.992761	0.991217	0.997959	0.990211	0.990494
9766	0.990540	0.991590	0.997654	0.990797	0.993138	0.996820	0.991724	0.988490

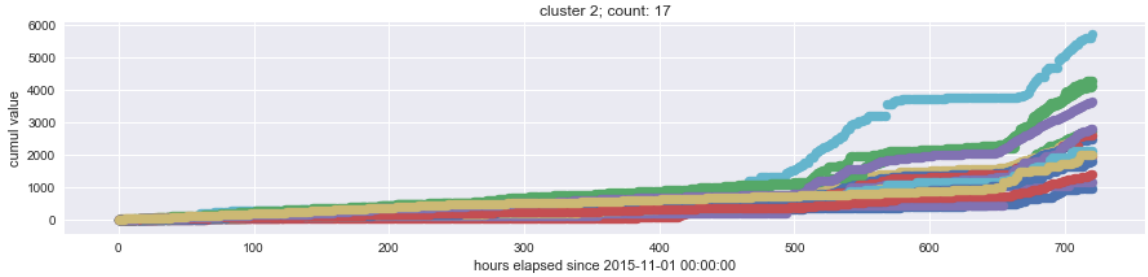
90 rows × 90 columns

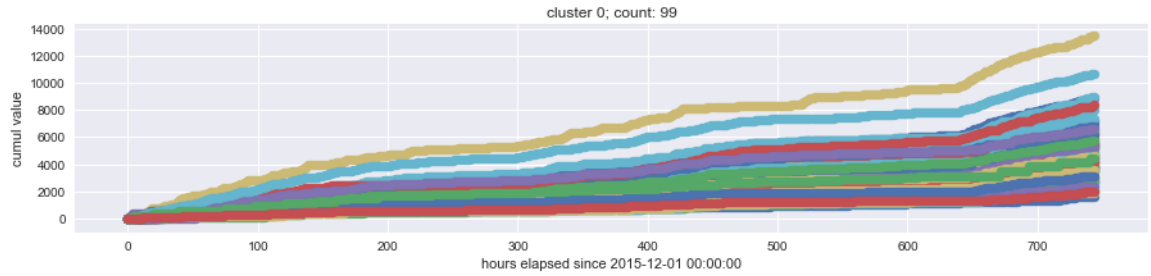
	r	id
5484	0.999547	77
1697	0.999455	77
370	0.999281	77
5810	0.999209	77
5403	0.999055	77
484	0.999046	77
744	0.998966	77
7989	0.998955	77
2449	0.998817	77
4029	0.998783	77

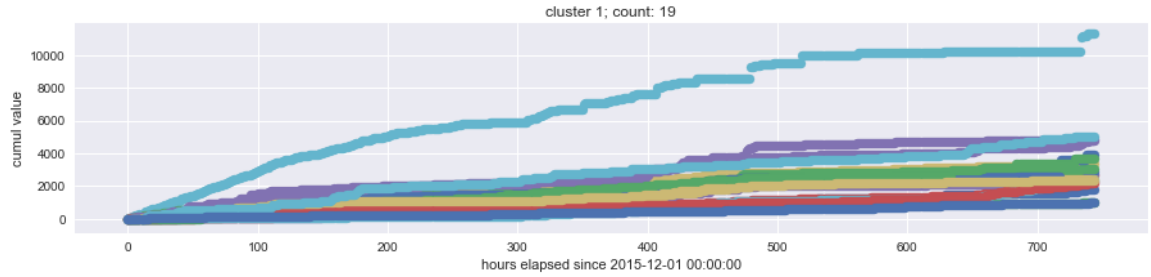
```
[0 0 0 1 0 0 0 1 0 0 1 1 0 0 0 0 0 0 1 0 1 0 0 0 0 1 1 0 0 1 0 0 0 0 0 1 2
 0 0 0 1 2 0 0 0 0 0 0 2 0 1 0 1 0 2 1 0 0 1 0 0 0 1 0 1 0 1 0 0 0 1 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
```

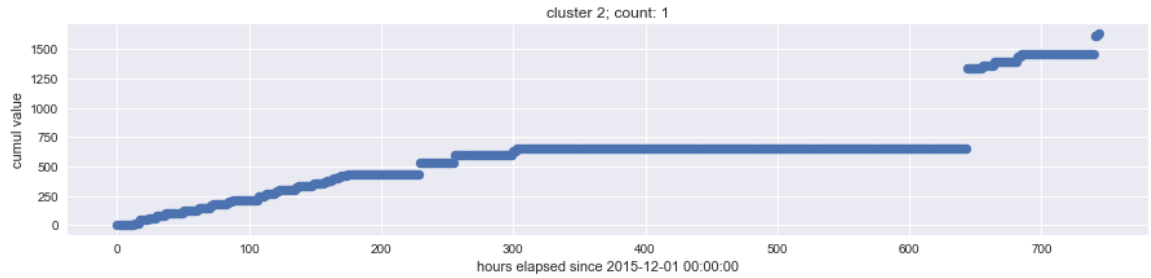


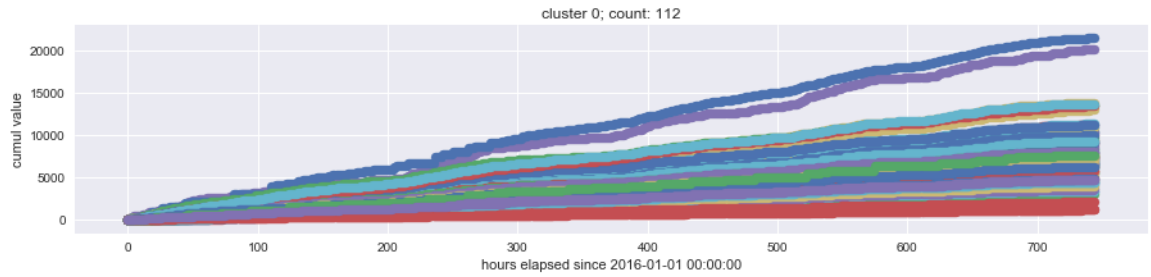


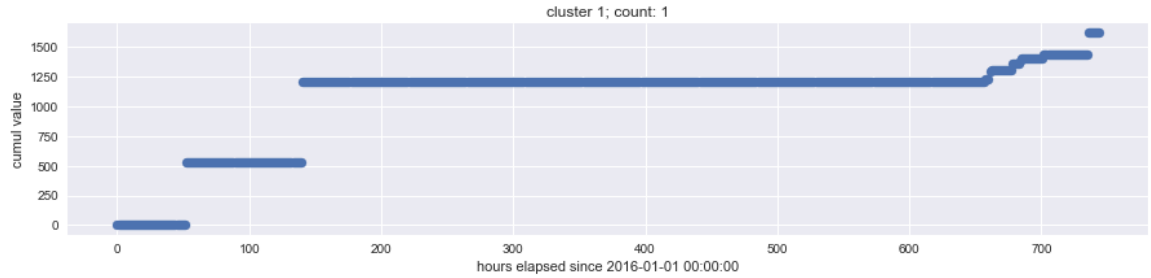


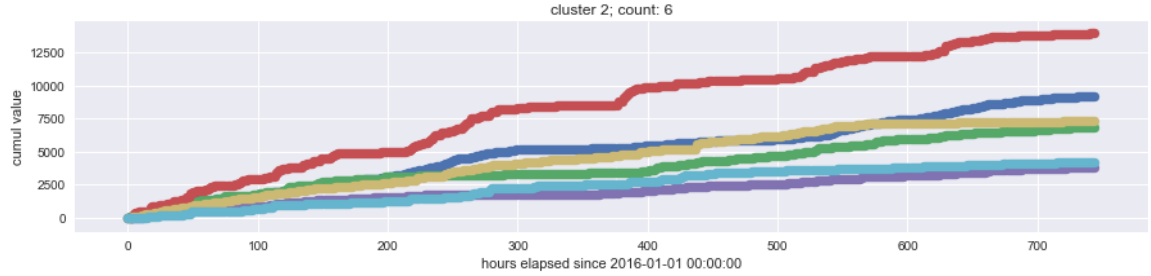

cluster 0; count: 99
cumul value
hours elapsed since 2015-12-01 00:00:00

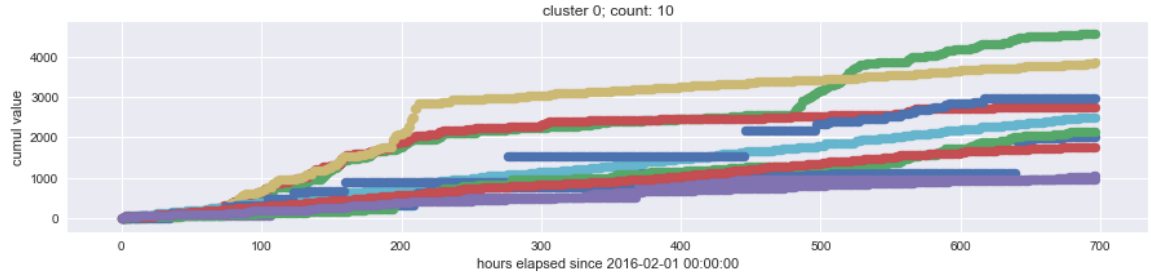

cluster 1; count: 19
cumul value
hours elapsed since 2015-12-01 00:00:00

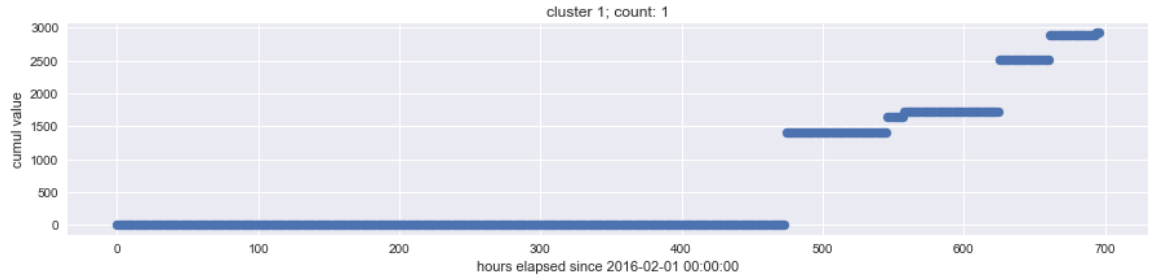

cluster 2; count: 1
cumul value
hours elapsed since 2015-12-01 00:00:00

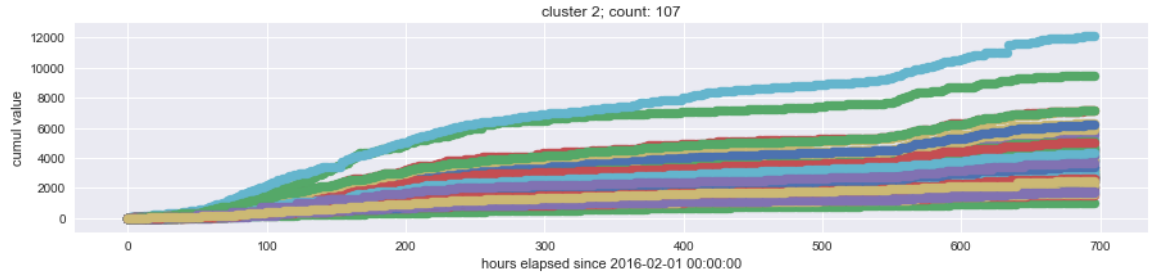

cluster 0; count: 112
cumul value
hours elapsed since 2016-01-01 00:00:00

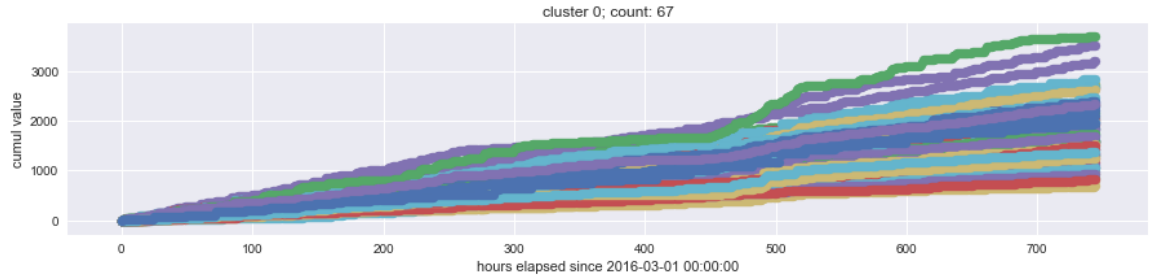

cluster 1; count: 1
cumul value
hours elapsed since 2016-01-01 00:00:00

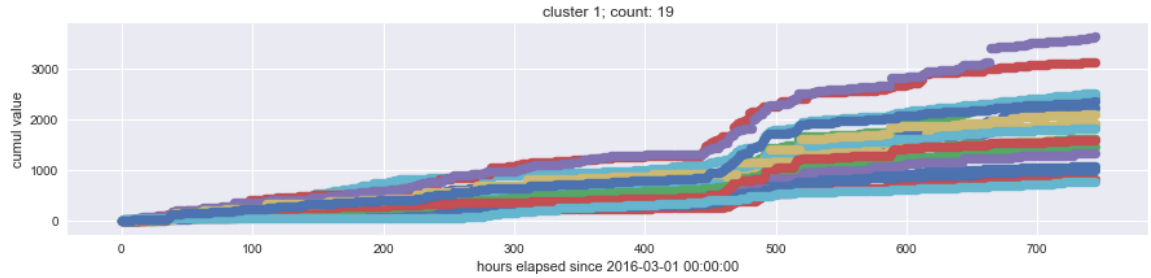

cluster 2; count: 6
cumul value
hours elapsed since 2016-01-01 00:00:00

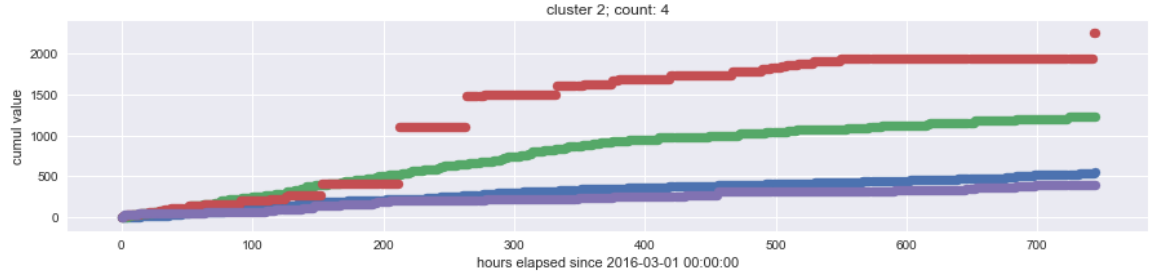

cluster 0; count: 10
cumul value
hours elapsed since 2016-02-01 00:00:00


cluster 1; count: 1
cumul value
hours elapsed since 2016-02-01 00:00:00


cluster 2; count: 107
cumul value
hours elapsed since 2016-02-01 00:00:00


cluster 0; count: 67
cumul value
hours elapsed since 2016-03-01 00:00:00


cluster 1; count: 19
cumul value
hours elapsed since 2016-03-01 00:00:00


cluster 2; count: 4
cumul value
hours elapsed since 2016-03-01 00:00:00