باسلام

دیتاست شامل ۲۴۴۷۸۳۳ قرائت یا ردیف است که از سال ۱۳۹۵ تا ۱۳۹۹ (5 شیت فایل اکسل) ازعدد شمارنده کنتور آب شهر گرگان جمع آوری شده است.

تعداد سطرها: 74113 اشتراک آب (کنتور آب)

تعداد ویژگی ها: 18 عدد

هدف اول (اجباری): خوشه بندی دیتاست و تعیین الگوی مصرف مشترکین مشکوک به مصرف غیر مجاز و همچنین ارائه شماره مشترک غیر مجاز

هدف دوم (اختیاری): تعیین مشترکین کم مصرف واقعی، مصارف صفر، خانه های خالی از سکونت، منازل مخروبه

من خودم تا الان مدل سازی رو به جایی رسوندم که اگر سرور اجاره‌ای داشته باشم خروجی می‌تونم خروجی بگیرم و تمام کنم، چون با کولب و ککل ران نشده ولی حالا روش کار تغییر کرده تا حتی الامکان به سرور اجاره ای نیاز کمتری داشته باشیم.

توضیحات ستون های داده:

ستون X و Y مختصات جغرافیایی مشترک مصرف کننده آب روی نقشه شهری میباشد.

ستون مانع، به دلیل داشتن داده های Nan (93 درصد) زیاد حذف گردد.

در ستون تعویض کنتور، فقط با داده های nan کار میکنیم و داده های دارای مقدار نیاز به تحلیل و خوشه بندی ندارند. بنابراین این داده ها، حذف میشوند و نیاز به خوشه بندی ندارند.

ستون های مهم دیتاست در انجام پروژه: مصرف، مبلغ،واحد و ظرفیت

برای خوشه بندی، الگوریتم های چگالی محور مانند dbscan و lof مدنظر است.

- در صورت استفاده از dbscan : مشخص کردن پارامترهای الگوریتم بر اساس تجربه است. بهترین مقدار حداقل همسایگی(minpts) عدد 18 تا 20

است و پارامتر اپسیلون، در بازه ای معقول حساسیت سنجی میشود (از طریق حلقه تکرار یا ابزار kneed یا زانویی پایتون) تا بهترین خوشه بندی تعیین شود. درصد معقول و منطقی داده های پرت، بین 8 تا 15 درصد است.

نمونه ای از جدول خروجی محاسبات الگوریتم dbscan در پروژه های مشابه در زیر آورده شده است.

جدول ۴-۶- بررسی نویزهای شناسایی‌شده با تغییر مقدار فاصله همسایگی در تعداد همسایگی ۱۸

faulty	noise	cluster	epsilon	minpts	
65	277	4	0.061	18	**1**
58	249	3	0.062	18	**2**
51	214	4	0.063	18	**3**
42	186	3	0.064	18	**4**
39	171	3	0.065	18	**5**
37	161	3	0.066	18	**6**
34	150	2	0.067	18	7
32	143	1	0.069	18	**8**
27	119	1	0.07	18	**9**

روش کار:

فقط کاربری مسکونی خوشه بندی میشود. با بقیه کاربریها کار نداریم. همانطور که قبلا هم توضیح داده شد داده هایی که ستون تعویض کنتور آنها مقدار دارد (تقریبا 47 درصد) حذف میشوند چون شرکت آب ا وضعیت انها مطلع است و بنابراین نیاز به بررسی ندارند.

کاربری مسکونی رو به سه دسته قطر به شرح زیر تقسیم میشود.

حالت اول: مسکونی با قطر نیم اینچ

حالت دوم: مسکونی با قطر 0.75 اینچ و یک اینچ

حالت سوم: مسکونی با قطر یک اینچ تا دو اینچ

هر کدوم از این سه حالت بالا جداگانه خوشه‌بندی شوند.

پیش پردازش داده ها:

مرحله اول: باید به ازای هر مشترک مسکونی، از فرمول زیر مصرف متوسط ماهانه تعیین شود.

{(مصرف هر دوره تقسیم بر تعداد روز بازه قرائت کنتور)} ضرب در عدد ۳۰ روز (معادل یک ماه)

این عمل برای تمامی مشترکین در طی 5 سال یا 60 ماه انجام میشود تا مجموعه داده‌ای جدید از سری زمانی‌های متوسط ماهیانه بدست آیند.

در مرحله بعد فیلتری اعمال می‌شود تا به وسیله آن، داده ها به دو دسته تقسیم شوند.

۱. مشترکینی که در بیشتر از 75 درصد ماه‌ها (48 ماه) مقدار قرائت شده دارند.

2. مشترکینی که در کمتر از 75 درصد ماه‌ها مقدار قرائت شده دارند.

فقط مشترکان دسته اول که بیشتر از 75 درصد ماه‌ها داده قرائت شده دارند خوشه بندی میشوند.

تعداد خوشه ها طبق انتظار:

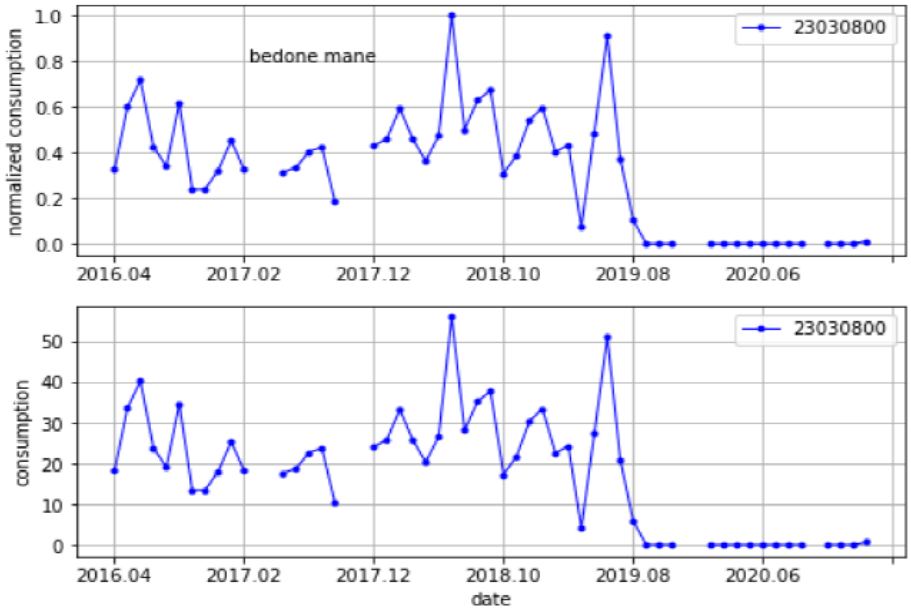
تعداد خوشه‌ها حتما بیشتر از 2 خوشه و حداکثر 5 خوشه باشد. به طوریکه بزرگترین خوشه یا همان خوشه اصلی، شامل مصارف نرمال باشد. یک خوشه نیز، شامل مشترکان با مصارف صفر است (به جز چند ماه محدود مصرف غیر صفر).

پس از تعیین داده های پرت (نویز)، الگوهای مصرف آنها شناسایی میشوند. نمونه هایی از الگوی مصرف مشترکان غیر مجاز(داده های پرت) که در پروژه های مشابه دیده شده است در زیر آورده شده است.

توضیح: تعیین 4 الی شش الگوی مصرف پر تکرار کافیست. برای تعیین الگوهای مصرف، به تغییر ناگهانی مصرف، توجه ویژه گردد.

الگوی یک

در این الگو مصارف تا یک زمان مشخص دارای یک الگوی نسبتاً متعارف هستند و از آن نقطه به بعد، به طور محسوس و قابل توجهی کمتر یا بیشتر از روند قبلی است.

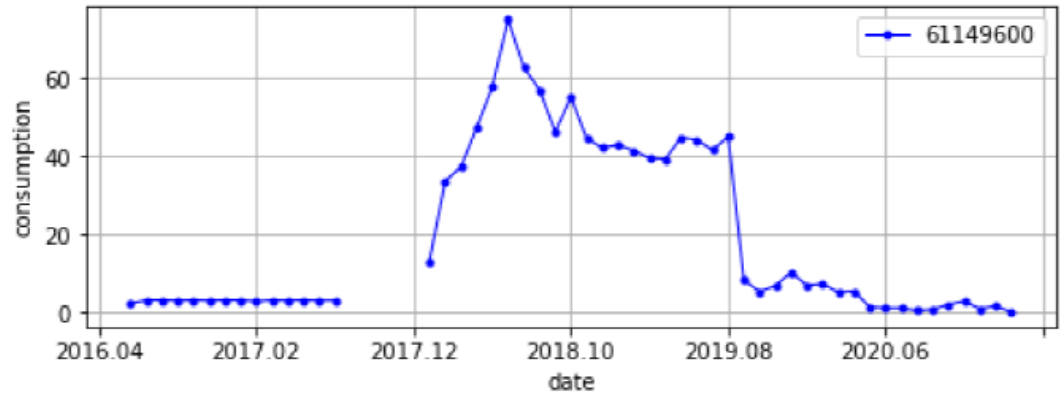


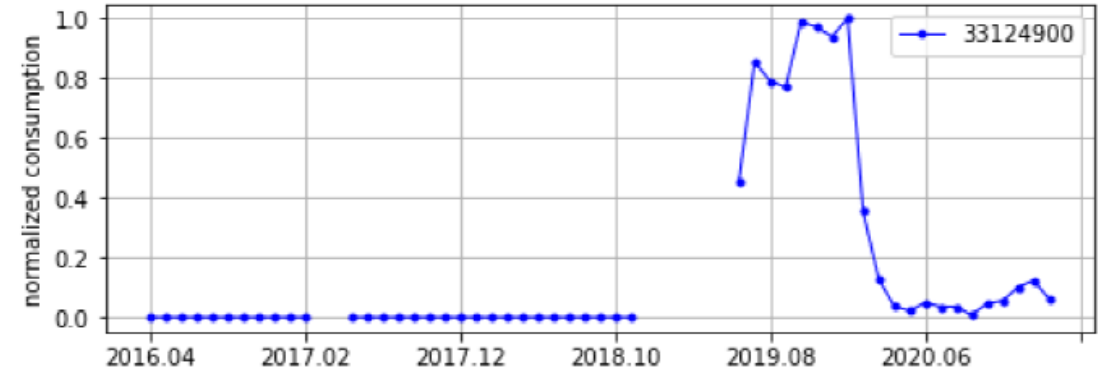
شکل ۴-۸- نمونه الگوی اول از مصارف غیرمتعارف، کاهش محسوس مصارف در طی بازه طولانی و به صفر رسیدن مصارف- مصرف نوسانی

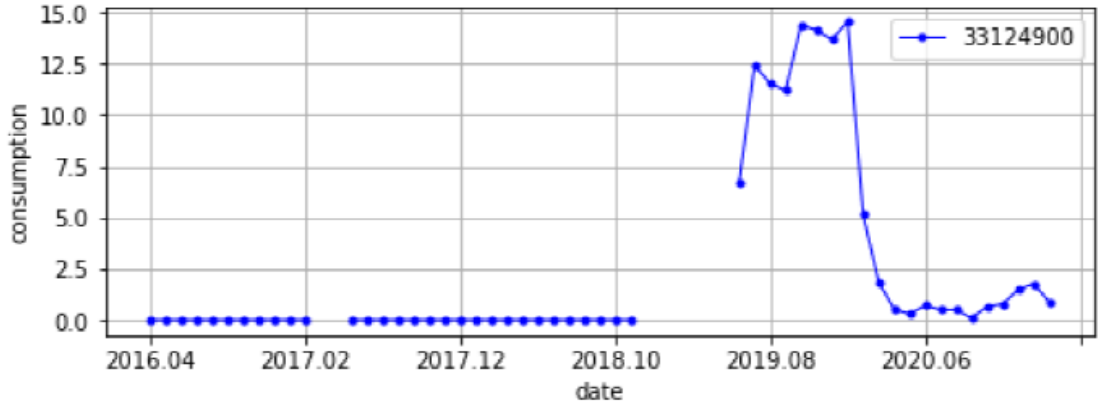
الگوی دو

تغییر مصارف ممکن است روند مثلثی داشته باشد در طول دوره‌های قرائت در بازه‌ای افزایش مصرف شدید را تجربه کند و مجدداً پس از گذشت زمان، مصرف به طور محسوسی برای مدت زیادی کاهش یابد که ممکن است مشترکی باشد که سابقه پشت کنتوری دارد و پس از اصلاح و افزایش مصرف، مجدداً به دستکاری کنتور اقدام کند. دو نمونه از الگوهای مشاهده شده، در شکل زیر نشان داده شده است.


normalized consumption
61149600
bedone mane, tak vaehd ba masrafe 60 metr mokaab dar mah!
1.0
0.8
0.6
0.4
0.2
0.0
2016.04
2017.02
2017.12
2018.10
2019.08
2020.06


consumption
61149600
60
40
20
0
2016.04
2017.02
2017.12
2018.10
2019.08
2020.06
date

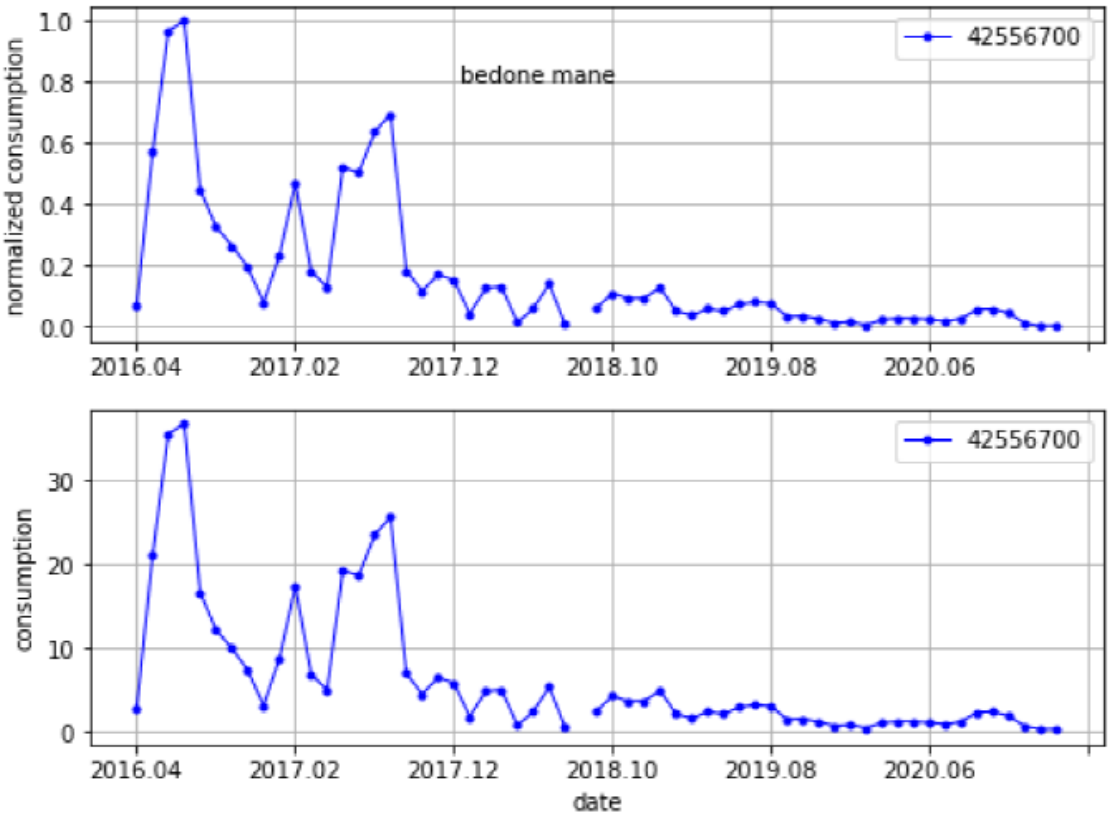

normalized consumption
33124900
1.0
0.8
0.6
0.4
0.2
0.0
2016.04
2017.02
2017.12
2018.10
2019.08
2020.06


consumption
33124900
15.0
12.5
10.0
7.5
5.0
2.5
0.0
2016.04
2017.02
2017.12
2018.10
2019.08
2020.06
date

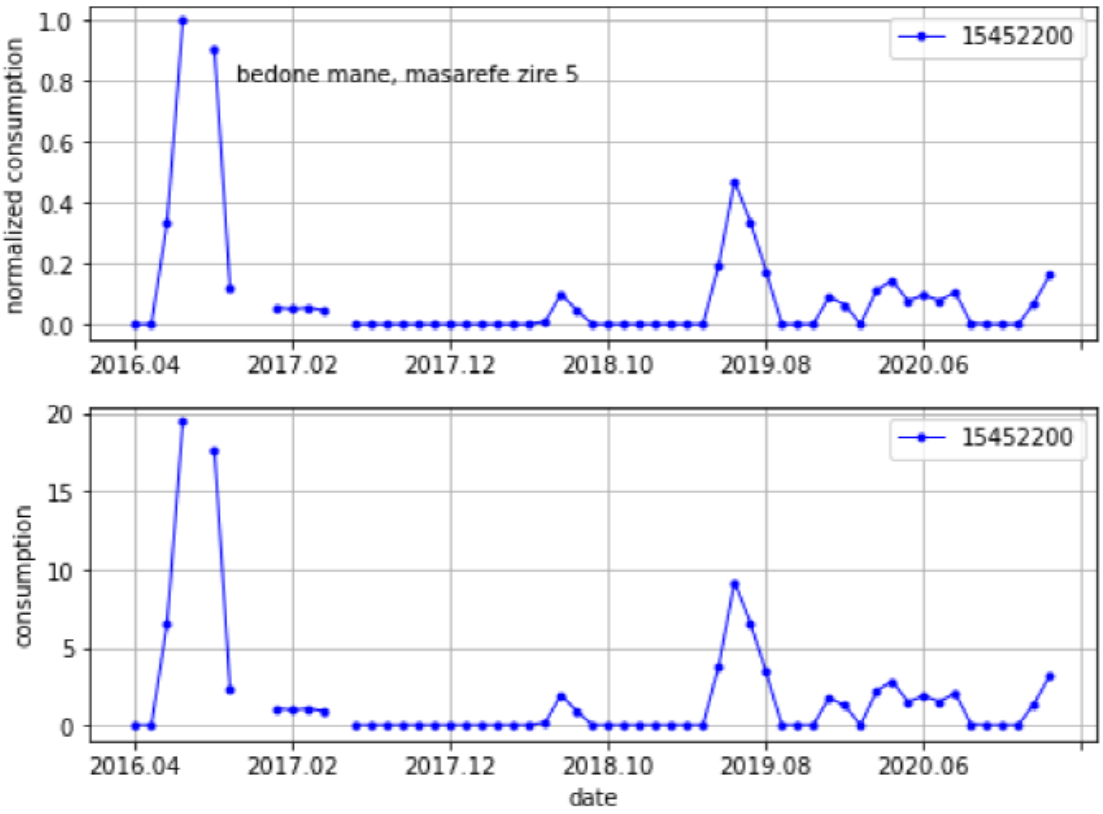
الگوی سه

اکثر شرکت‌های آب و فاضلاب برای مصرف آب ماهانه یک خانوار، حد پایین را شناسایی می‌کنند. مصرف بسیار کمتر از حد پایین هشداری برای شرکت آب و فاضلاب است. مبنا این پژوهش، عدد 10 متر مکعب در ماه است. در اشکال زیر مشاهده میشود که در یک بازه طولانی، مصرف ماهانه متوسط، کمتر از ده متر مکعب است.

الف)

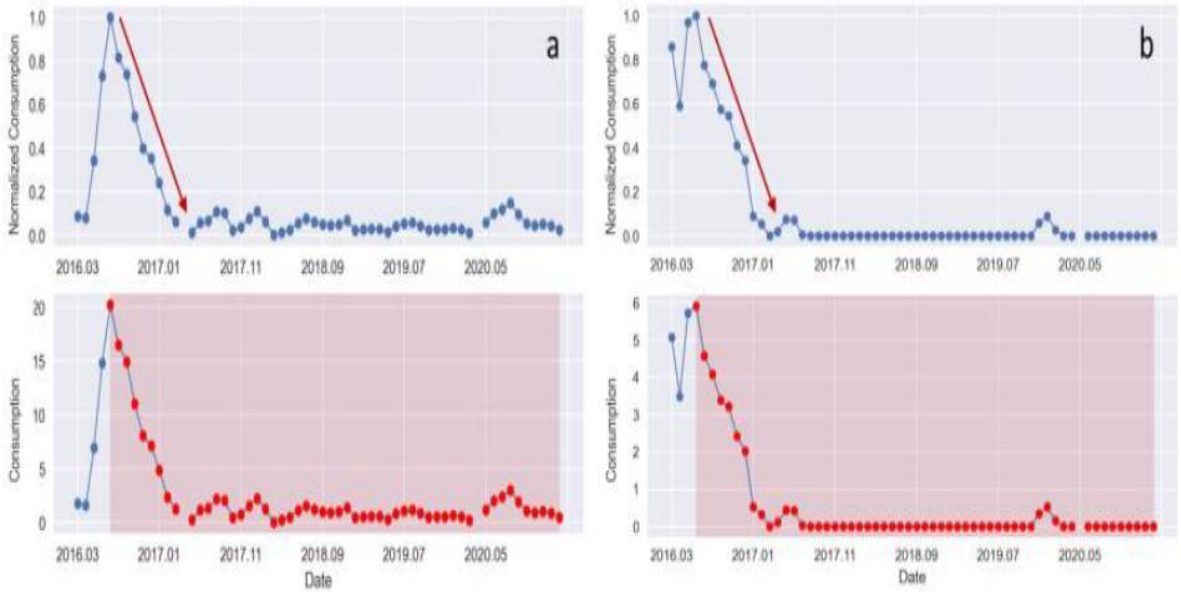


ب)



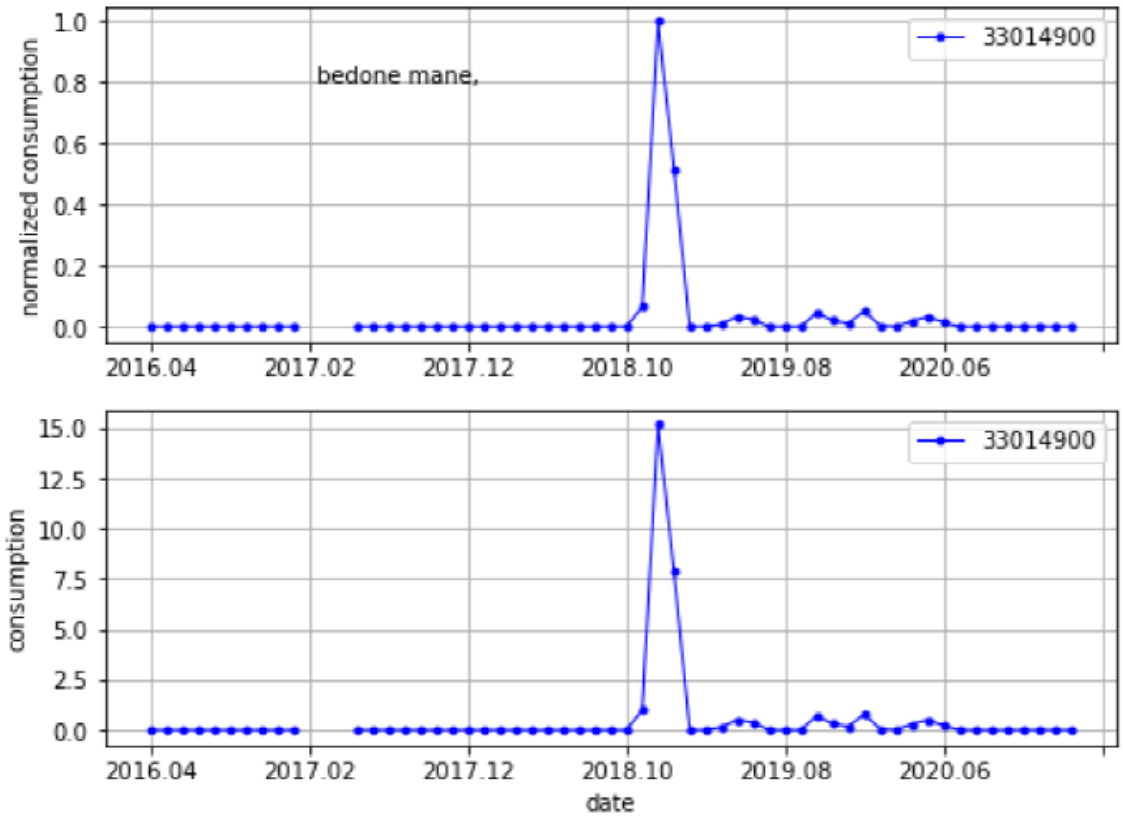
الگوی چهار

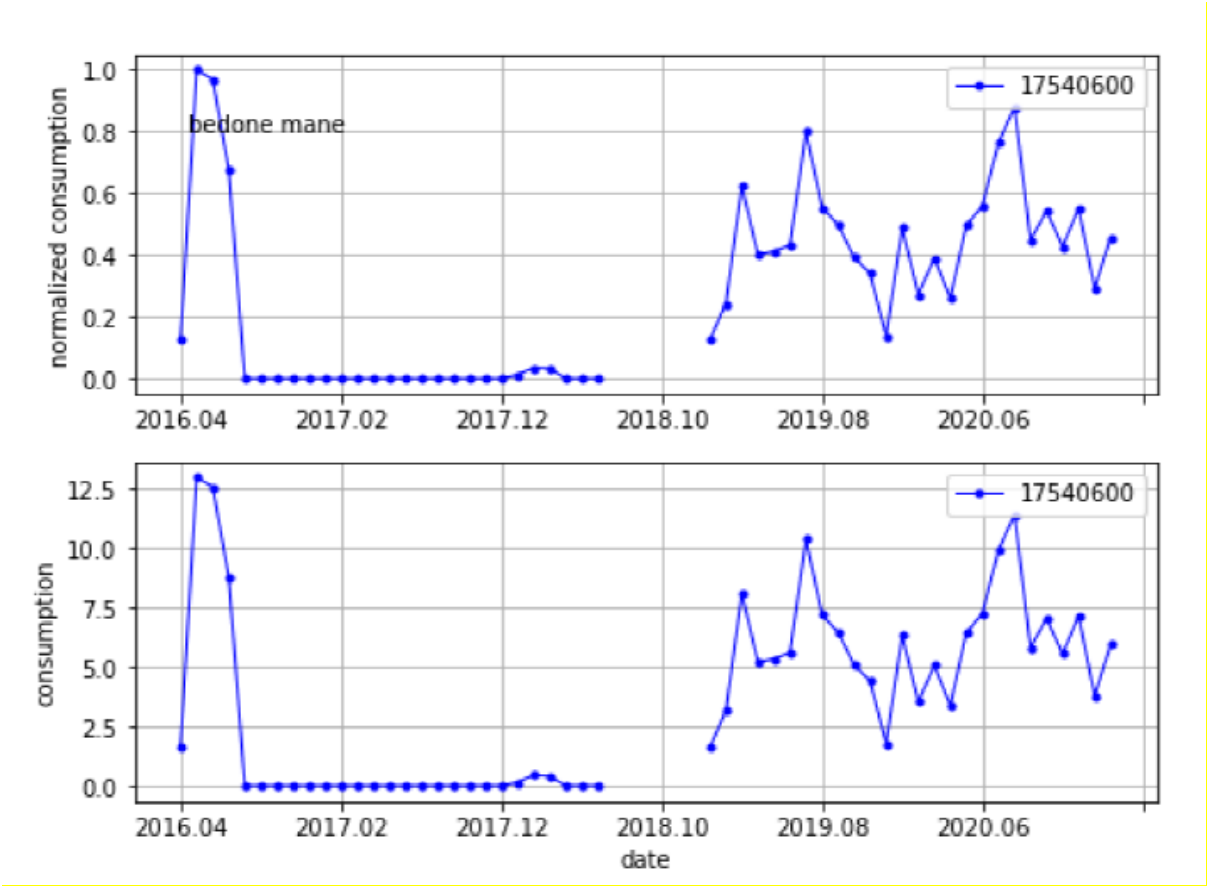
در این الگو مشترکانی مشاهده می‌شوند که مصرف آنها به تدریج کاهش پیدا می‌کند و سپس به الگوی مصرف ثابت و تکراری ختم می‌شود.



الگوی 5

در این الگو، برخی از نویزهای شناسایی شده، یک یا پند پیک مصرفی را در 60 ماه تجربه میکنند. دو نمونه در اشکال زیر آورده شده است.




normalized consumption
1.0
0.8
0.6
0.4
0.2
0.0
bedone mane
17540600
2016.04
2017.02
2017.12
2018.10
2019.08
2020.06
consumption
12.5
10.0
7.5
5.0
2.5
0.0
17540600
2016.04
2017.02
2017.12
2018.10
2019.08
2020.06
date