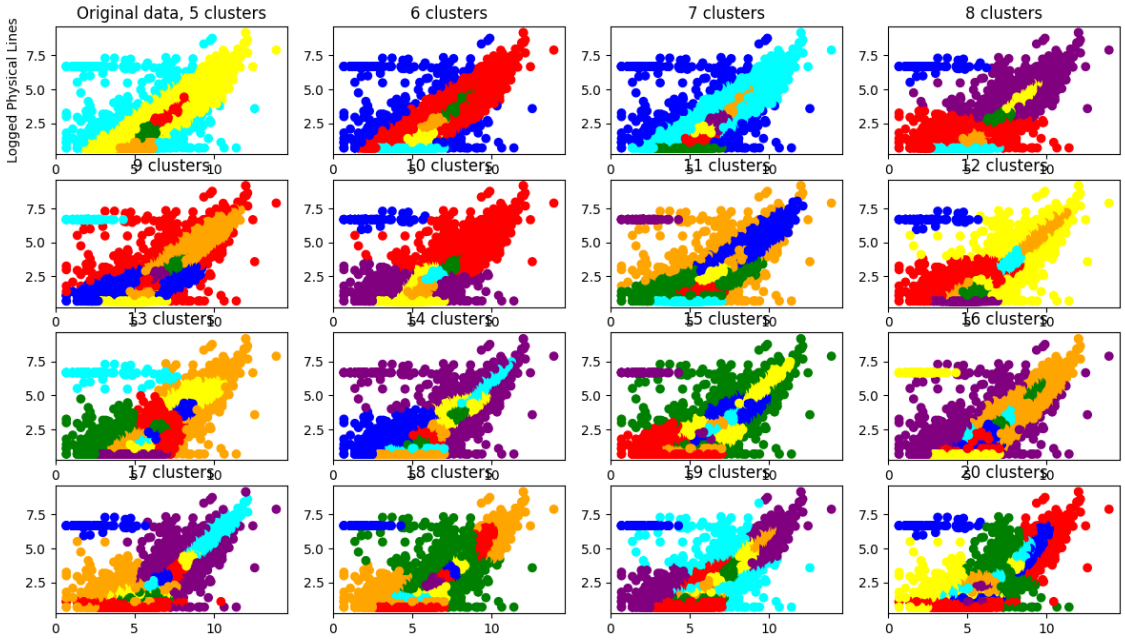
Abstract

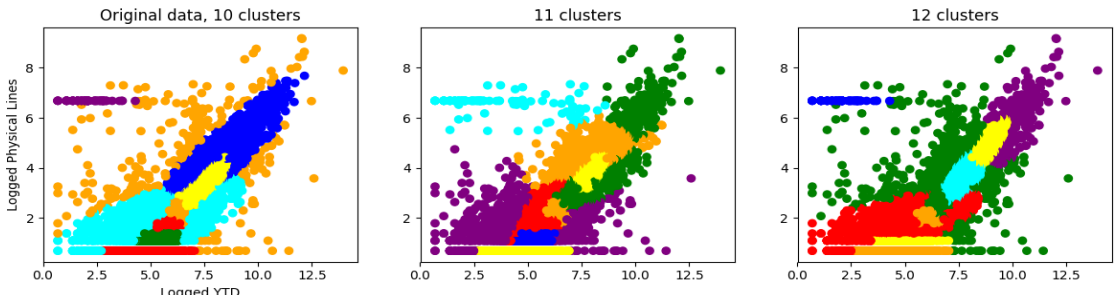
I develop a process of visualization that can help us instinctively see how good a clustering pattern is. I also create a naïve assessment algorithm based on variance. I find that my major difficulty is creating a good assessment algorithm, advise on this issue is most welcome.

Instinctive Evaluation



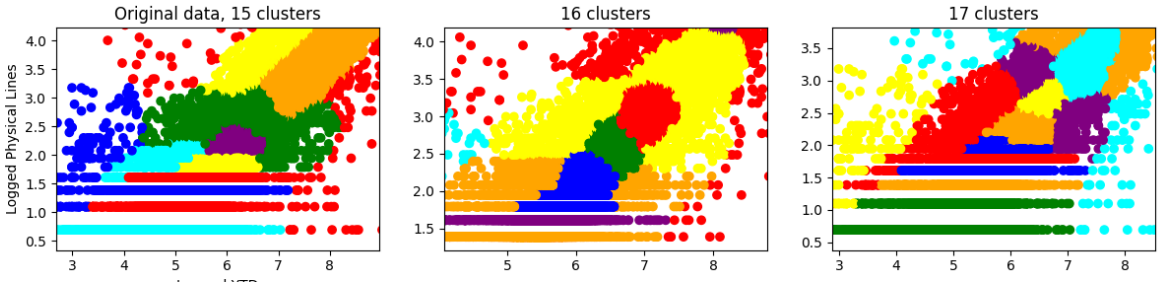
GMM on Original Data, from 5 to 20 clusters

Before starting quantitative assessment, I first cluster original data into 5 to 20 clusters, hoping for some instinctive results. I think we can see the clustering pattern grows more and more "normal" until 11 clusters, but 12 clusters do not seem much better than 11.



GMM on Original Data, from 10 to 12 clusters

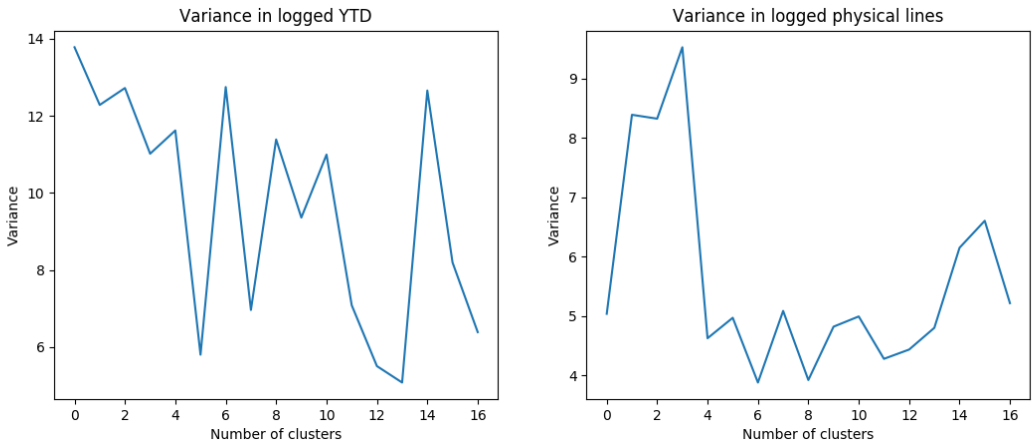
The graph above more clearly shows that 11 clusters is much better than 10, but 12 is not too obviously better than 11, so 11 clusters should be an elbow point in my assessment afterwards.



GMM on Original Data, from 15 to 17 clusters

Similarly, when we look into the detail, we can see another elbow point at 16 clusters. These two elbow points should be reflected on my assessment.

Assessment algorithms



Note that the x-axis here is incorrect, should be 5 to 20 instead of 0 to 16

Here is my first and right now only assessment algorithm: naïve assessment. Its algorithm is straightforward: returning the total variance as assessment coefficient.

With such a "naïve" algorithm, the result is not promising. I expect to see a steadily dropping line graph, but instead got such a messy shape. It might indicate that the clustering pattern does not grow better when increasing number of clusters, but without further information, I would say the assessment itself is too naive.

As for elbow points, 11 clusters (6 in the graph above) have a high variance in YTD, which is not good, and low variance in physical lines, which is good. 16 clusters (11 in graph) has low variance in both. Still, we could not draw any conclusion because of the naïveté of assessment.

Difficulty in assessment

With a mature visualization above, we can instinctively tell how good an assessment algorithm is. However, I now meet difficulty in creating such good assessment.

To create such an assessment, I need to combine four values: variance on x and y axis, maximum-minimum on x and y axis, and each of these values comes from all clusters. It seems pretty hard to me.

I now understand why we have been doing multiple steps of clustering: because we've been avoiding assess the holistic pattern, which is difficult as I just find. But I think such a holistic assessment is necessary. Without it, we would probably get lost in all those clusters inside clusters and lose the whole picture. In a word, though it is hard, I think we need to do it.

Next step

I will work on the holistic assessment algorithm I talk about above, probably apply some machine learning. I will need time with this. Any suggestion is welcome.

And I will also try the multiple step clustering based on my current naïve algorithm. I guess the result might not be good, but it is still worth to try it out. I probably can deliver the result this Thursday night.