Cell Tracking Analysis with K-Means Clustering Method

D. Askan
Clarkson University

Spring 2016

1 Introduction

This document analyzes data sets obtained cells' movements on different layers. Data sets have only velocity information of the cells. The aim of this document is to provide a classification to different layer types according to cells' movements by utilizing clustering algorithms. K-Means Clustering method has been used for clustering. This document also aims to answer the question: "Do cell's migration differ depending on the surface type?" Even though some layers were classified successfully, the algorithm has failed to classify all layers perfectly. Additionally, the paper prefers to use the term "classify" despite the fact that this is an unsupervised learning problem. Because, there are initial insights that some of the surfaces allow cells move faster than the others. The paper is organized as follows: Section II gives information about the data used. In section III, visualization and clustering analysis on the data are performed. Insights from the analysis and comments are discussed in section IV.

2 The Data

The data is provided by Professor Shantanu Sur at Department of Biology from "Cancer Cells Movement Research" at Clarkson University, Potsdam, NY. There are 5 different layer types, each type has two sample sets, each sample set has 10 cells, and each cell has 40 observations in terms of 15 minutes interval of velocity information. There are total 4000 instances.

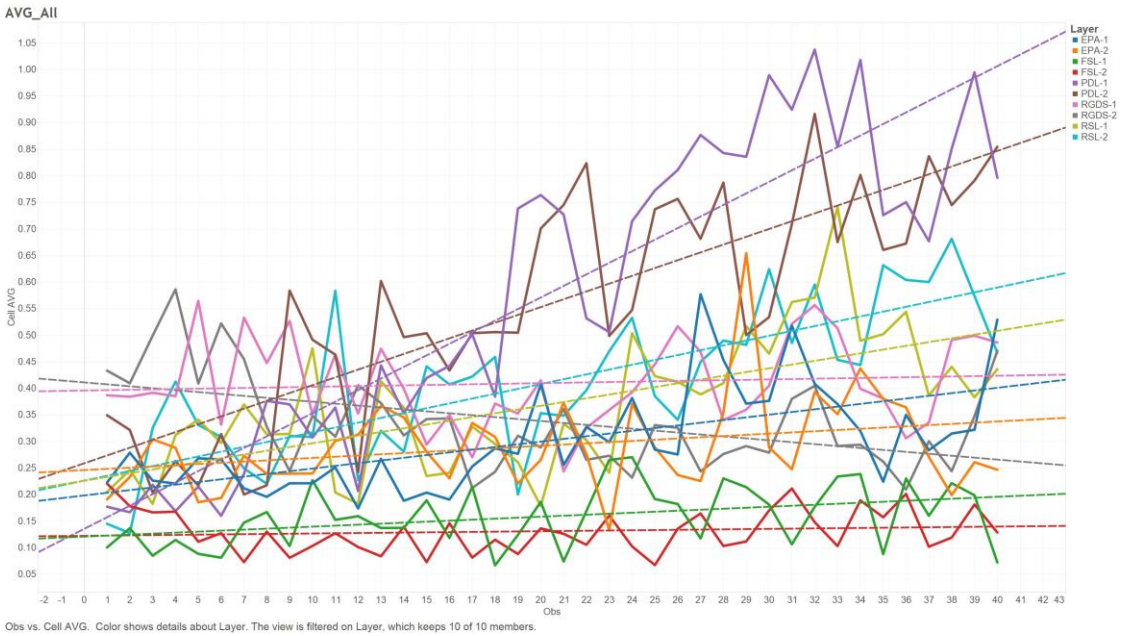
Layer types are;
FSL: Forward, strong, linear peptide amphiphile
RSL: Reverse, strong, linear peptide amphiphile
EPA: Epsilon peptide amphiphile
RGDS: Arginine, glycine, aspartate, serine
PDL

Below, a snapshot can be seen from the data sets, first 20 observations from sample set EPA-1,

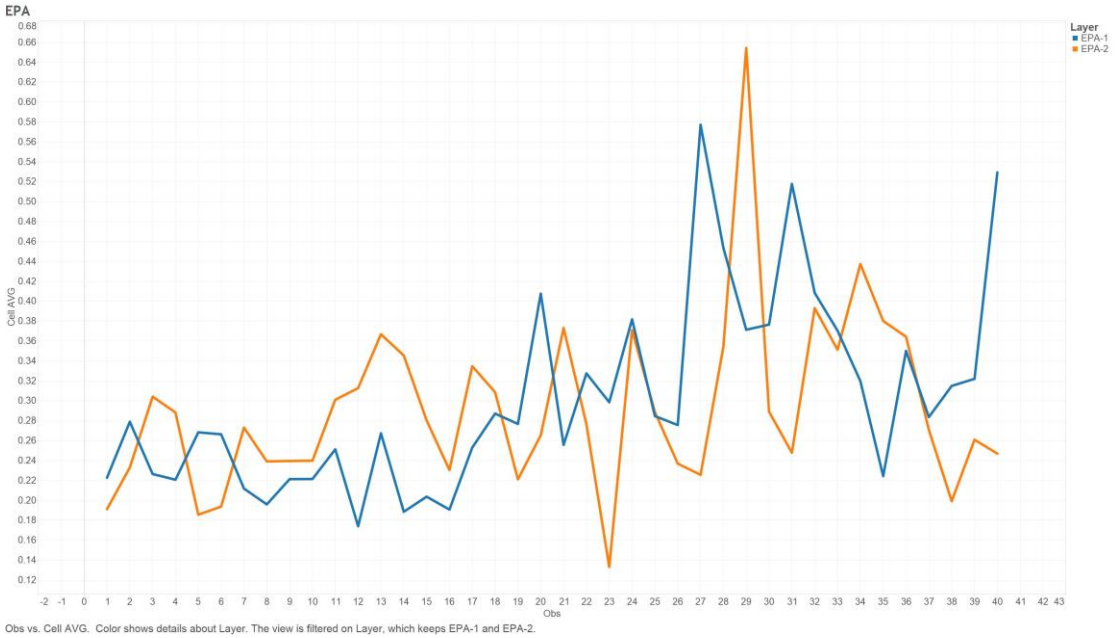
Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	Cell 7	Cell 8	Cell 9	Cell 10
0.144	0.144	0.329	0.204	0.465	0.233	0.322	0.129	0	0.258
0.658	0.266	0.193	0.193	0.288	0	0.258	0.204	0.408	0.322
0.329	0.504	0.258	0	0	0.193	0.064	0.233	0.129	0.555
0	0	0.064	0.144	0.091	0.144	0.322	0.645	0.532	0.266
0.347	0.322	0	0	0.288	0.532	0.387	0.064	0.456	0.288
0	0.387	0.258	0.064	0	0.456	0.532	0.204	0.376	0.387
0.064	0.266	0.129	0.064	0.204	0.52	0.064	0.456	0.064	0.288
0.064	0.456	0.322	0.129	0.064	0.274	0.329	0.064	0.129	0.129
0.129	0.233	0	0.204	0.288	0	0.258	0.204	0.433	0.465
0	0.144	0.144	0.322	0.129	0.451	0.064	0.288	0.182	0.491
0.064	0.376	0	0.266	0.233	0	0.387	0.322	0.288	0.577
0	0	0.182	0.347	0.129	0	0.129	0.516	0.347	0.091
0	0.064	0.465	0.258	0.129	0.392	0.376	0.144	0.433	0.413
0.266	0.274	0	0.322	0.182	0	0.376	0	0.144	0.322
0.182	0.258	0.129	0	0.288	0.451	0.274	0.392	0.064	0
0.408	0.329	0	0.329	0.193	0	0.064	0.288	0.233	0.064
0.266	0.504	0	0	0.064	0.433	0.193	0.555	0.387	0.129
0.182	0.091	0.193	1.419	0.144	0.064	0.258	0.144	0.233	0.144
0.204	0.233	0.144	0.456	0.193	0.52	0	0.288	0.274	0.456

3 Visualization and Clustering

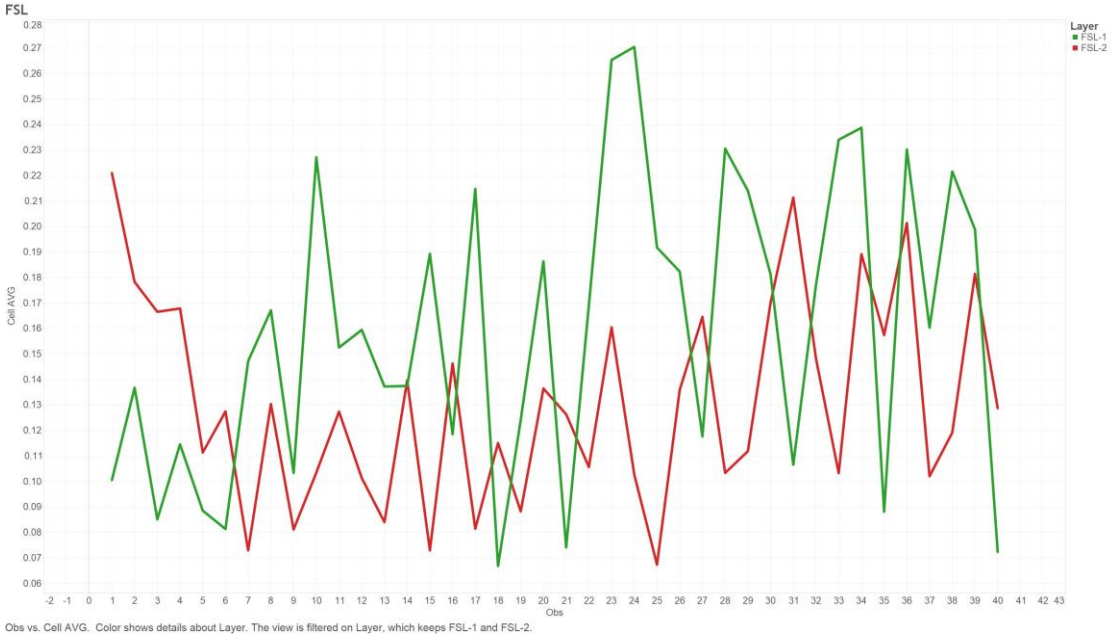
Below graph visualize the entire data set. Horizontal axis represents observations whereas vertical axis is for average velocity information of 10 cells of that particular observation. Each line represents a sample data set. As expected, similar patterns can be seen in same type layers. For instance, PDL-1 and PDL-2 or FSL-1 and FSL-2 in bottom of graph as green and red lines.



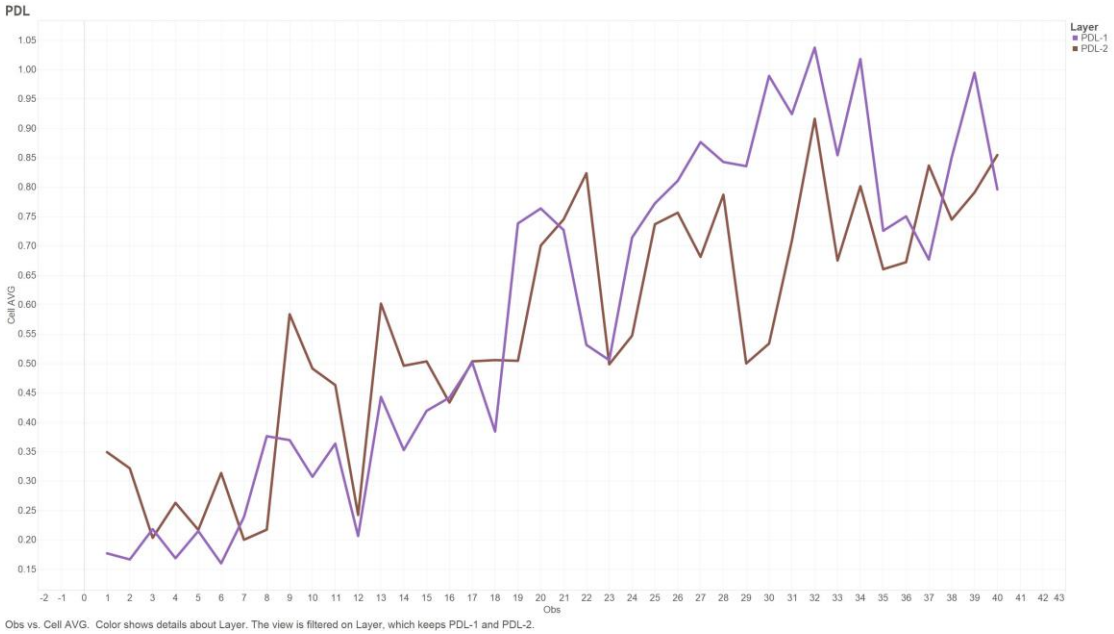
By looking same layer types individually, cells behaviors can be comprehend clearly. For EPA,



For FSL,



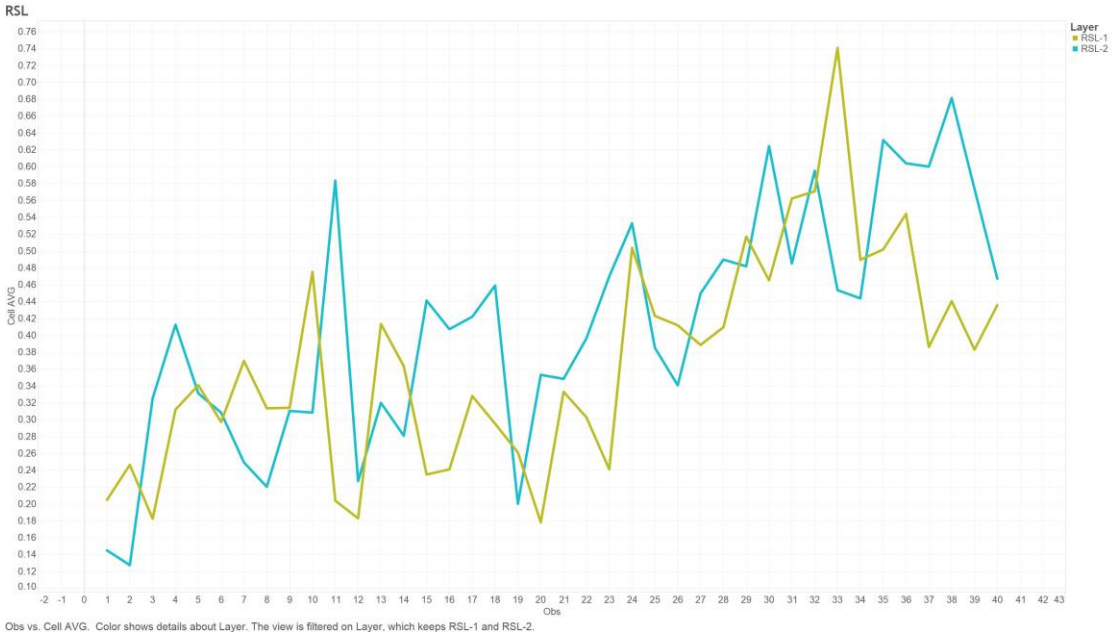
For PDL,



As seen above, average cell velocity is getting faster in time on some type of layers. And on some layers, average cell velocity is getting slower in time as in the following graph.

For RGDS,

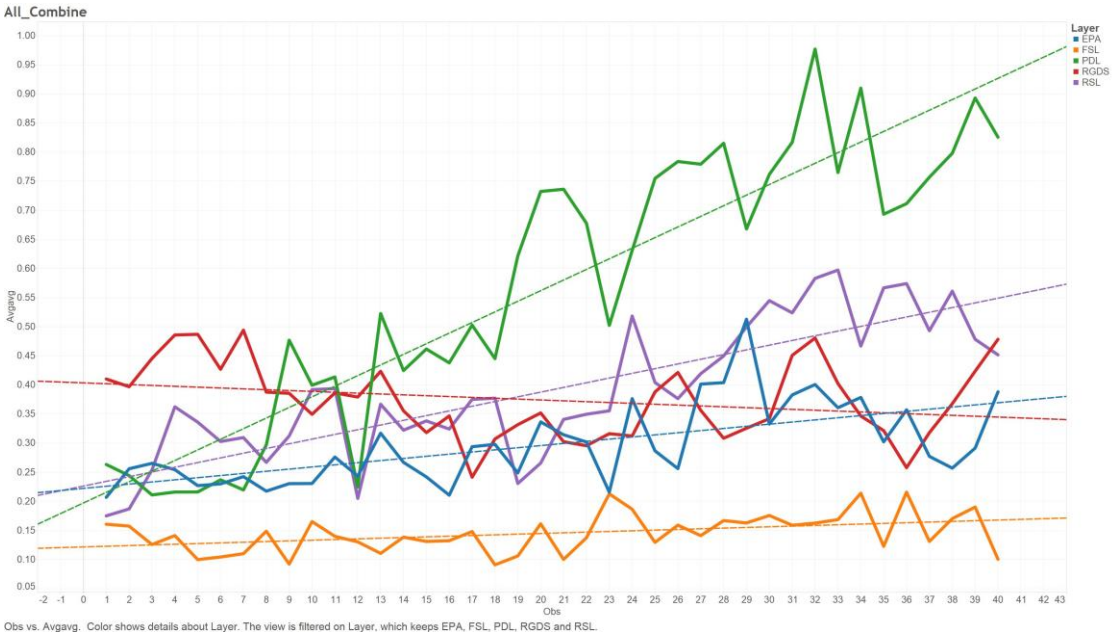
The last one for RSL,



Average cell velocity is getting faster in time in above graph too, despite the fact that the slop is relatively small.

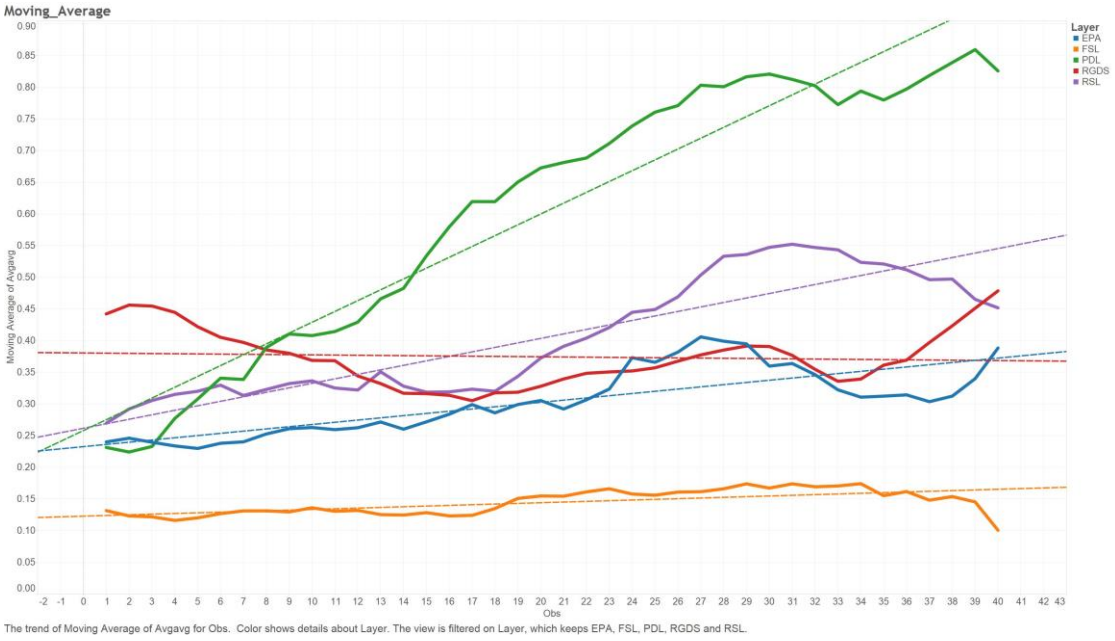
As it was indicated in the previous section, each layer has two sample data sets, and cells are expected to behave similarly on those. And, these similarity was stated in the first part of this section. So, by

combining those sample sets, the line formation can become more understandable in visualization. Below graph reflects this idea with only five different lines for 5 different layers.



However, the visualization still needs to be more understandable by smoothing to take it one step further. Moving average technique is utilized in this part by choosing next five values to get averages. It is 5 because 5 is high enough to provide smoothness to graph, and low enough to keep variability in the data. Other trials are 2, 3, 8, 10 and 5:40 ratio is the one that visualizes this graph best.

In consequence, the tendencies of lines, in other words velocities of cells in time are clear to comprehend in below graph,



Since there are different behavior trends in beginning and the maturity phases of the data sets, two different clustering have been performed. The first clustering includes the first 10 observations, in other words observations until 2.5 hours. SPSS outputs for the first clustering can be seen below,

Initial Cluster Centers

	Cluster	
	1	2
mov1	.6456	.0128
mov2	.7402	.0256
mov3	.6480	.0256
mov4	.5514	.0128
mov5	.7304	.0128
mov6	.8216	.0128
mov7	.7376	.0000
mov8	.7280	.0000
mov9	.8338	.0000
mov10	.6190	.0000

Number of Cases in each Cluster

Cluster	1	48.000
	2	52.000
Valid		100.000
Missing		.000

Final Cluster Centers

	Cluster	
	1	2
mov1	.3835	.1527
mov2	.3944	.1492
mov3	.4077	.1472
mov4	.4083	.1479
mov5	.4148	.1448
mov6	.4162	.1567
mov7	.4452	.1536
mov8	.4366	.1467
mov9	.4671	.1510
mov10	.4560	.1621

Iteration History[a]

	Change in Cluster Centers	
Iteration	1	2
1	.718	.604
2	.044	.021
3	.018	.009
4	.029	.015
5	.034	.019
6	.071	.053
7	.023	.021
8	.010	.008
9	.010	.008
10	.010	.009
11	.009	.009
12	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 12. The minimum distance between initial centers is 2.216.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
mov1	1.330	1	.018	98	73.843	.000
mov2	1.500	1	.019	98	78.382	.000
mov3	1.695	1	.020	98	85.201	.000
mov4	1.693	1	.017	98	97.458	.000
mov5	1.820	1	.019	98	96.480	.000
mov6	1.681	1	.017	98	99.652	.000
mov7	2.121	1	.015	98	145.798	.000
mov8	2.099	1	.014	98	147.118	.000
mov9	2.494	1	.018	98	140.514	.000
mov10	2.156	1	.019	98	115.125	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

In the first clustering, K-means clustering algorithm converged in 12 iterations with 90% success ratio. There are only 4 wrong classifications out of 40. Visual differences in the beginning phase is also confirmed statistically with this ratio.

Clustering membership in the first clustering can be seen below,

Cluster Membership				Cluster Membership			
Number	Layer	Cluster	Distance	Number	Layer	Cluster	Distance
41	FSL-1	2	.395	1	RGDS-1	1	.423
42	FSL-1	2	.342	2	RGDS-1	2	.316
43	FSL-1	2	.235	3	RGDS-1	1	.380
44	FSL-1	2	.224	4	RGDS-1	1	.285
45	FSL-1	2	.289	5	RGDS-1	1	.802
46	FSL-1	2	.234	6	RGDS-1	2	.246
47	FSL-1	2	.223	7	RGDS-1	1	.432
48	FSL-1	2	.116	8	RGDS-1	1	1.025
49	FSL-1	1	.384	9	RGDS-1	1	.490
50	FSL-1	2	.341	10	RGDS-1	1	.327
51	FSL-2	2	.194	11	RGDS-2	1	.474
52	FSL-2	2	.103	12	RGDS-2	1	.396
53	FSL-2	2	.207	13	RGDS-2	1	.408
54	FSL-2	2	.366	14	RGDS-2	1	.346
55	FSL-2	2	.244	15	RGDS-2	2	.129
56	FSL-2	2	.210	16	RGDS-2	1	1.097
57	FSL-2	2	.307	17	RGDS-2	1	.509
58	FSL-2	2	.142	18	RGDS-2	1	.467
59	FSL-2	2	.180	19	RGDS-2	1	.654
60	FSL-2	2	.268	20	RGDS-2	1	.474

The first clustering resulted with 1 wrong classification in FSL layer and 3 wrong classifications in RGDS layer with total of 4 miss classifications.

Next, the clustering analysis was performed again in the maturity phase, too. In this second clustering, PDL is the surface allows cells move fastest instead of RGDS in the beginning phase. This phase is between 20th and 40th, final observations. In other words, it is between 5th – 10th hours of the observation.

The outputs from the second clustering can be seen in the next page. As it can be seen in the table on right, unlike the previous clustering, the algorithm converged only in four iterations in this clustering. This fast converge is also an indicator that the data is distributed more heterogeneously in this part of observation. So, visual distribution is also corrected with this iteration sequence.

Iteration History[a]

Iteration	Change in Cluster Centers	
	1	2
1	.909	1.433
2	.126	.177
3	.059	.081
4	.000	.000

Initial Cluster Centers

	Cluster	
	1	2
mov21	.0790	.8308
mov22	.0790	.7316
mov23	.0386	.8066
mov24	.0128	.9504
mov25	.0128	.9700
mov26	.0128	1.0448
mov27	.0880	1.1682
mov28	.1010	1.2960
mov29	.1010	1.3046
mov30	.1396	1.3020
mov31	.1396	1.2132
mov32	.0826	1.2440
mov33	.0568	1.1318
mov34	.0826	1.0594
mov35	.0848	1.1452
mov36	.0848	1.1098
mov37	.08325	1.08050
mov38	.111000000	1.104666667
mov39	.1020	1.2245
mov40	.000	.735

Number of Cases in each Cluster

Cluster	1	60.000
	2	40.000
Valid		100.000
Missing		.000

Final Cluster Centers

	Cluster	
	1	2
mov21	.2181	.5878
mov22	.2244	.5986
mov23	.2314	.6216
mov24	.2374	.6667
mov25	.2278	.6956
mov26	.2392	.6977
mov27	.2384	.7325
mov28	.2402	.7806
mov29	.2368	.8007
mov30	.2476	.7991
mov31	.2384	.7979
mov32	.2423	.7703
mov33	.2339	.7202
mov34	.2305	.7114
mov35	.2197	.7236
mov36	.2292	.7332
mov37	.22249	.74784
mov38	.228933333	.769225000
mov39	.2395	.7710
mov40	.252	.745

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
mov21	3.280	1	.035	98	94.120	.000
mov22	3.361	1	.038	98	88.100	.000
mov23	3.654	1	.045	98	81.662	.000
mov24	4.424	1	.049	98	89.906	.000
mov25	5.252	1	.054	98	97.594	.000
mov26	5.045	1	.054	98	93.716	.000
mov27	5.859	1	.045	98	131.589	.000
mov28	7.011	1	.038	98	183.114	.000
mov29	7.634	1	.035	98	220.326	.000
mov30	7.298	1	.036	98	203.978	.000
mov31	7.511	1	.034	98	223.454	.000
mov32	6.691	1	.034	98	196.213	.000
mov33	5.674	1	.030	98	187.422	.000
mov34	5.550	1	.036	98	154.517	.000
mov35	6.093	1	.028	98	218.100	.000
mov36	6.096	1	.028	98	214.200	.000
mov37	6.624	1	.032	98	208.030	.000
mov38	7.006	1	.040	98	175.383	.000
mov39	6.779	1	.056	98	120.859	.000
mov40	5.846	1	.094	98	62.119	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

The classification/clustering is more successful in the second part of observation. As it can be seen below, there is only one wrong classification in the result. The success ratio is 97.5%.

Cluster Membership				Cluster Membership			
Number	Layer	Cluster	Distance	Number	Layer	Cluster	Distance
41	FSL-1	1	.617	81	PDL-1	2	1.492
42	FSL-1	1	.384	82	PDL-1	2	1.016
43	FSL-1	1	.623	83	PDL-1	2	1.225
44	FSL-1	1	.381	84	PDL-1	2	1.669
45	FSL-1	1	.615	85	PDL-1	2	1.183
46	FSL-1	1	.611	86	PDL-1	2	.520
47	FSL-1	1	.631	87	PDL-1	2	1.440
48	FSL-1	1	.357	88	PDL-1	2	1.058
49	FSL-1	1	.580	89	PDL-1	2	1.235
50	FSL-1	1	.215	90	PDL-1	2	2.042
51	FSL-2	1	.735	91	PDL-2	1	.328
52	FSL-2	1	.286	92	PDL-2	2	.562
53	FSL-2	1	.387	93	PDL-2	2	1.555
54	FSL-2	1	.250	94	PDL-2	2	1.121
55	FSL-2	1	.516	95	PDL-2	2	1.228
56	FSL-2	1	.439	96	PDL-2	2	1.351
57	FSL-2	1	.566	97	PDL-2	2	1.225
58	FSL-2	1	.656	98	PDL-2	2	.781
59	FSL-2	1	.595	99	PDL-2	2	1.606
60	FSL-2	1	.447	100	PDL-2	2	1.571

4 Insights and Comments

Consequently, although all layers are classified/clustered perfectly this paper provides valuable inferences and insights about cell behaviors on particular layer types. The insights are,

- There are 2 classes in the beginning and maturity phases, since there are actually 5 layer types, other 3 types cannot be classified or clustered statistically.
- Both statistically and visually, cells migrate significantly slower on the layer type FLS in every phase.
- Likewise, cells migrate significantly faster on the layer type RGDS in beginning phase.
- And, cells migrate significantly faster on the layer type PDL in maturity phase both statistically and visually.
- 90% success rate in the beginning phase, and 97.5% success rate in maturity phase have been reached.

Movement velocity is directly related with the growing speed of the cancer cells. Thus, it is possible that these results can help on future cancer cell treatment reasearches.