PNU Industrial Data Science

K-means clustering

Prof. Hyerim Bae

빅데이터로 돈버는 세상

출처: https://www.hyosungfms.com/fms/promote/fms_news_view.do?id_boards=13416

Contents

산업데이터과학은 산업현장에서 수집된 데이터를 분석하는데 필요한 기초 소양을 강의합니다.

Clustering overview

Clustering: The Main Idea

Goal: Form groups (clusters) of **similar** records

Used for **segmenting markets** into groups of similar customers

Example: Segmenting US neighborhoods based on demographics & income: "Furs & station wagons," "Bohemian Mix,""Money & Brains", ...

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

Other Applications

- Classification of species
- Grouping securities in portfolios
- Grouping firms for structural analysis of economy
- Army uniform sizes

Example: Public Utilities

Goal: find clusters of similar utilities

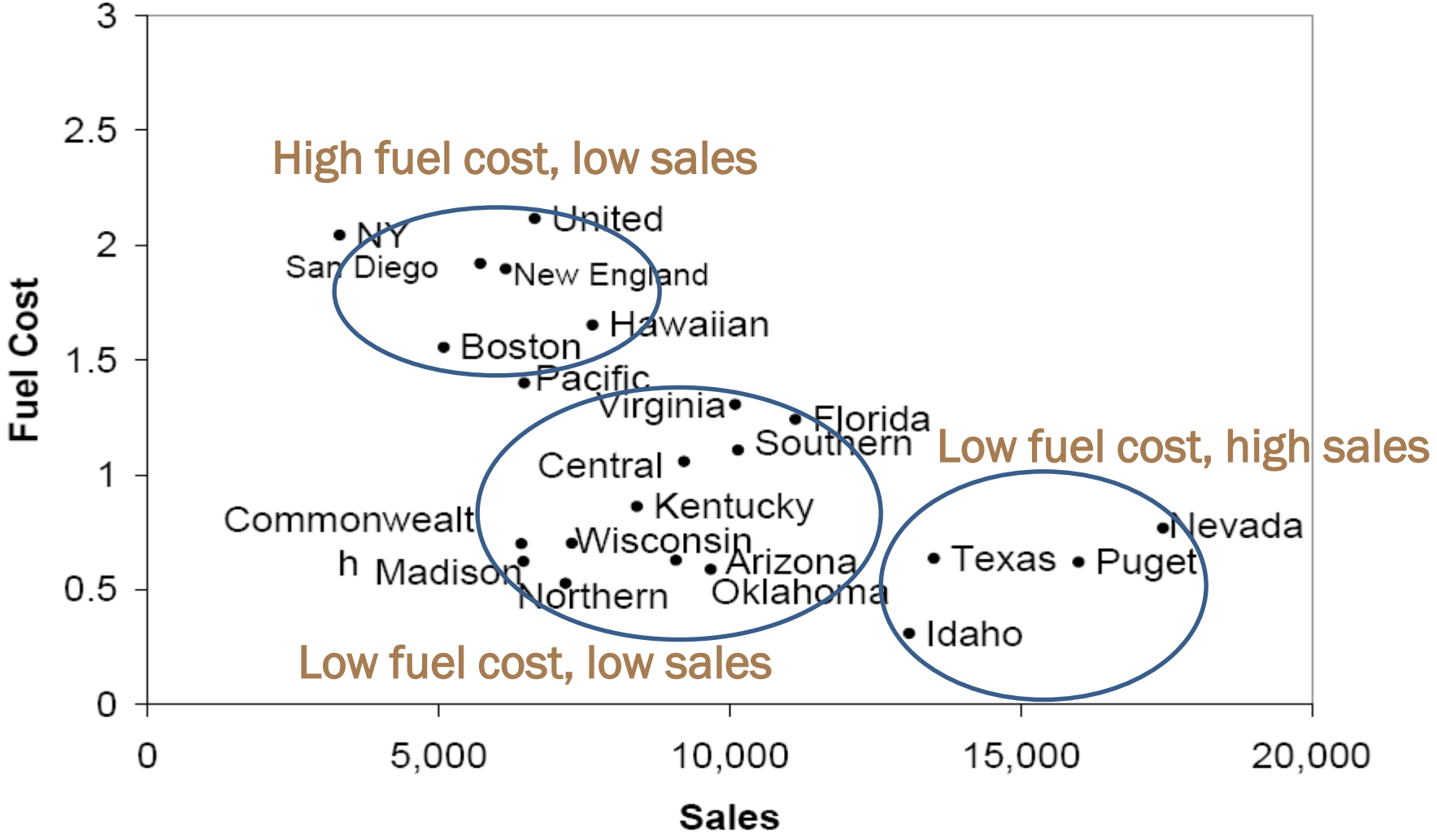
Data: 22 firms, 8 variables

- Fixed-charge covering ratio (고정비 부담율)
- Rate of return on capital (투자수익률)
- Cost per kilowatt capacity (킬로와트당 비용)
- Annual load factor (연간 부하량)
- Growth in peak demand (최대전력 수요량)
- Sales (전력 판매 매출)
- % nuclear (원자력 비율)
- Fuel costs per kwh (연료비)

Dataset

Company	Fixed_charge	RoR	Cost	Load	Δ Demand	Sales	Nuclear	Fuel_Cost
Arizona	1.06	9.2	151	54.4	1.6	9077	0	0.628
Boston	0.89	10.3	202	57.9	2.2	5088	25.3	1.555
Central	1.43	15.4	113	53	3.4	9212	0	1.058
Commonwealth	1.02	11.2	168	56	0.3	6423	34.3	0.7
Con Ed NY	1.49	8.8	192	51.2	1	3300	15.6	2.044
Florida	1.32	13.5	111	60	-2.2	11127	22.5	1.241
Hawaiian	1.22	12.2	175	67.6	2.2	7642	0	1.652
Idaho	1.1	9.2	245	57	3.3	13082	0	0.309
Kentucky	1.34	13	168	60.4	7.2	8406	0	0.862
Madison	1.12	12.4	197	53	2.7	6455	39.2	0.623
Nevada	0.75	7.5	173	51.5	6.5	17441	0	0.768
New England	1.13	10.9	178	62	3.7	6154	0	1.897
Northern	1.15	12.7	199	53.7	6.4	7179	50.2	0.527
Oklahoma	1.09	12	96	49.8	1.4	9673	0	0.588
Pacific	0.96	7.6	164	62.2	-0.1	6468	0.9	1.4
Puget	1.16	9.9	252	56	9.2	15991	0	0.62
San Diego	0.76	6.4	136	61.9	9	5714	8.3	1.92
Southern	1.05	12.6	150	56.7	2.7	10140	0	1.108
Texas	1.16	11.7	104	54	-2.1	13507	0	0.636
Wisconsin	1.2	11.8	148	59.9	3.5	7287	41.1	0.702
United	1.04	8.6	204	61	3.5	6650	0	2.116
Virginia	1.07	9.3	174	54.3	5.9	10093	26.6	1.306

- Sales and Fuel cost



Extension to More Than 2 Dimensions

In prior example, clustering was done by eye

Multiple dimensions require formal algorithm with

- A **distance measure**
- A way to use the distance measure in forming clusters

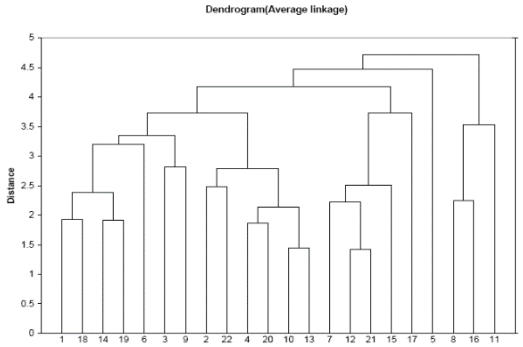
We will consider two algorithms: **hierarchical** and **non-hierarchical**

Types of clustering

- Hierarchical vs. non-hierarchical
- Hierarchical

  Agglomerative Methods
  - Begin with n-clusters (each record its own cluster)
  - Keep joining records into clusters until one cluster is left (the entire data set)
  - Most popular

  Divisive Methods
  - Start with one all-inclusive cluster
  - Repeatedly divide into smaller clusters



- Non-hierarchical

  K-means clustering algorithm
  - K: the number of clusters (Assign the data into a cluster with small distance)
  - Smaller amount of computation than hierarchical methods

  DBSCAN
  - Density based approach
  - Discover clusters of arbitrary shape

Measuring Distance

Between records

Between clusters

Properties of distance measure

1. Nonnegative $d_{ij} \geq 0$
2. Self-Proximity $d_{ii} = 0$
3. Symmetry $d_{ij} = d_{ji}$
4. Triangular Inequality $d_{ij} \leq d_{ik} + d_{kj}$

Distance Between Two Records

Euclidean Distance is most popular:

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

For Categorical Data: Similarity

To measure the distance between records in terms of two 0/1 variables, create table with counts:

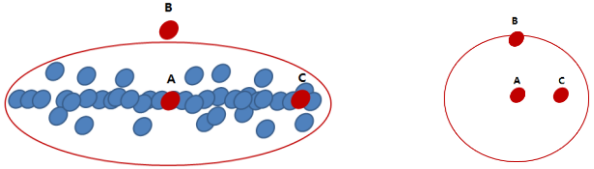
	0	1
0	a	b
1	c	d

Similarity metrics based on this table:

- Matching coef. = (a+d)/(a+b+c+d)
- Jaquard's coef. = d/(b+c+d)
  - Use in cases where a matching "1" is much greater evidence of similarity than matching "0" (e.g. "owns Corvette")
  - 예) 두 사람 모두에게 많은 특징들이 공통적으로 없다고 해서 두 사람이 유사하다고 생각하지 않는다.

Other Distance Measures

- Correlation-based similarity
  - $d_{ij} = 1 - r_{ij}^2$
- Statistical distance (Mahalanobis)
  - Variable with higher correlation has less effect.



$$D(A, B) = \sqrt{(A-B)\sum^{-1}(A-B)^T}$$

- Manhattan distance (absolute differences)
- Maximum coordinate distance
- Gower's similarity (for mixed variable types: continuous & categorical)

Measuring Distance Between Clusters:

- **single linkage**

$$\min(distance(A_i, B_j))$$

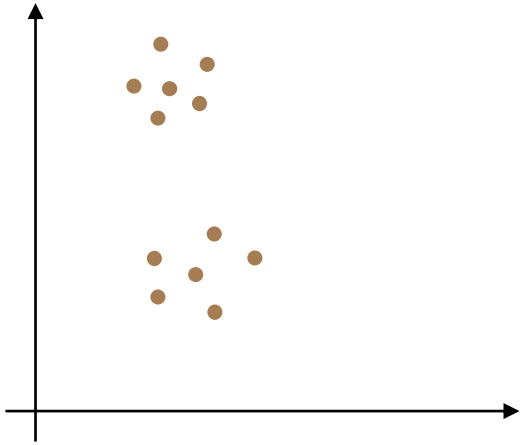
- **complete linkage**

$$\max(distance(A_i, B_j))$$

- **average linkage**

$$\text{Average}(distance(A_i, B_j))$$

- **Centroid distance**
  - Distance between centroids



The Hierarchical Clustering Steps

1. Start with *n* clusters (each record is its own cluster)
2. Merge two closest records into one cluster
3. At each successive step, the two clusters closest to each other are merged

Dendrogram, from bottom up, illustrates the process

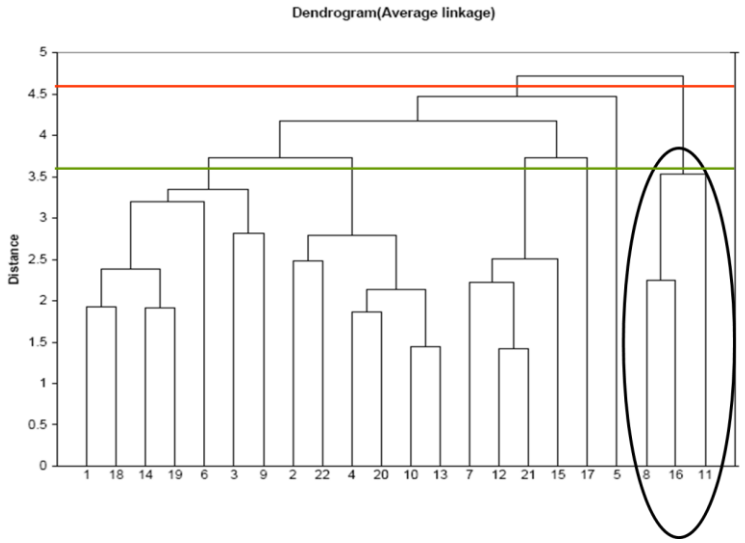
Reading the Dendrogram

See process of clustering: Lines connected lower down are merged earlier

- Records 12 & 21 are closest & form first cluster
- 10 and 13 will be merged next, after 12 & 21

Determining number of clusters: For a given "distance between clusters", a horizontal line intersects the clusters that are that far apart, to create clusters

- E.g., at distance of 4.6 (**red line** in next slide), data can be reduced to 2 clusters -- The smaller of the two is circled
- At distance of 3.6 (**green line**) data can be reduced to 6 clusters, including the circled cluster



Validation: Interpretation

Goal: obtain meaningful and useful clusters

Caveats:

(1) Random chance can often produce apparent clusters

(2) Different cluster methods produce different results

설명가능성(Intepretability):

- Obtain summary statistics
- Also review clusters in terms of variables **not** used in clustering
- Label the cluster (e.g. clustering of financial firms in 2008 might yield label like "midsize, sub-prime loser")

Validation: Desirable Cluster Features

Stability – are clusters and cluster assignments sensitive to slight changes in inputs? Are cluster assignments in partition B similar to partition A?

- Cluster A:분할된 데이터 집합 A에 대해서 군집분석 수행
- Assign to A's cluster using centroid in A:집합A의 중심점을 구하고 이를 이용하여 집합 B에 있는 각 레코드를 군집으로 할당
- Compare with the original result: 위의 군집할당 결과가 모든 데이터에 근거한 군집할당과 비교했을 때 얼마나 일치하는가

Separation – check ratio of between-cluster variation to within-cluster variation (higher is better)

- 군집내 변동에 대한 군집 간 변동의 비를 검토

Disadv. of hierarchical clustering

- Amount of computation (n X n distance matrix)
  - 계산횟수가 많고 계산속도가 느리다.
- Error in the early analysis cannot be corrected:
  - 분석 초기에 잘못된 군집에 들어가면 이를 수정할 수 없다
- Low stability:
  - 데이터를 재정렬하거나 몇 개의 레코드를 제외시킬 경우, 전혀 다른 군집결과가 나타날 수 있다.
- Average distance will seriously affected by distance measure
  - 단일 연결법과 완전연결법은 군집간 거리의 상대적인 순서가 유지되는한 거리측도의 변화에 대해 로버스트하다.
  - 평균연결법은 거리측도에 영향을 많이 받는다.
- Sensitive to anomaly
  - 이상치에 민감하다.

Non-hierarchical: K-Means Clustering Algorithm

1. Choose # of clusters desired, *k*
2. Start with a partition into k clusters
   Often based on random selection of k centroids
3. At each step, move each record to cluster with closest centroid
4. Recompute centroids, repeat step 3
5. Stop when moving records increases within-cluster dispersion

K-means Algorithm: Choosing k and Initial Partitioning

Choose *k* based on the how results will be used

e.g., "How many market segments do we want?"

Also experiment with slightly different *k*'s

Initial partition into clusters can be random, or based on domain knowledge

If random partition, repeat the process with different random partitions

Output: Cluster Centroids

We chose $k = 3$

4 of the 8 variables are shown

Cluster	Fixed_charge	RoR	Cost	Load_factor
Cluster-1	0.89	10.3	202	57.9
Cluster-2	1.43	15.4	113	53
Cluster-3	1.06	9.2	151	54.4

Distance Between Clusters

Clusters 1 and 2 are relatively well-separated from each other, while cluster 3 not as much

Distance between	Cluster-1	Cluster-2	Cluster-3
Cluster-1	0	5.03216253	3.16901457
Cluster-2	5.03216253	0	3.76581196
Cluster-3	3.16901457	3.76581196	0

Within-Cluster Dispersion

Data summary (In Original coordinates)

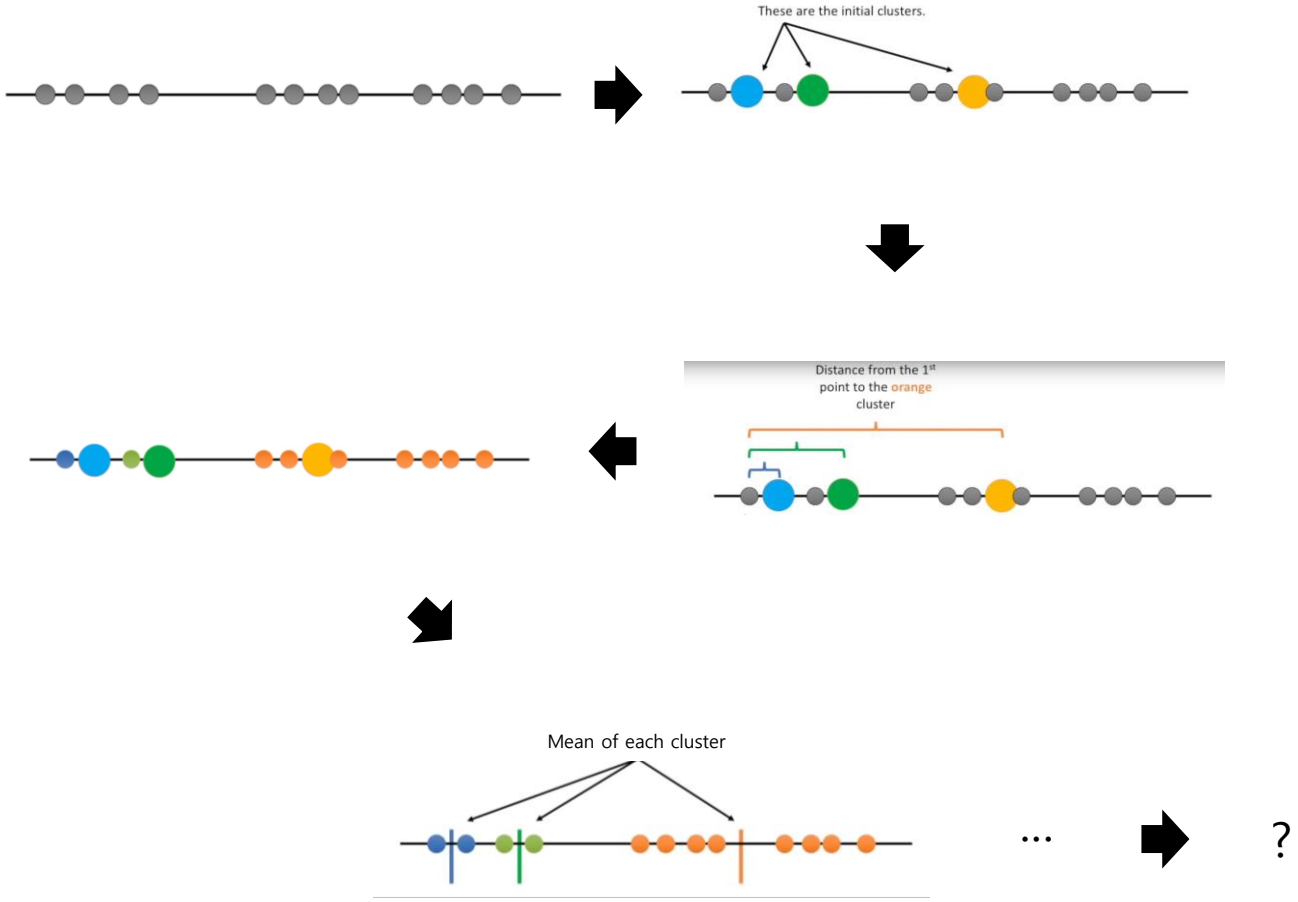
Cluster	#Obs	Average distance in cluster
Cluster-1	12	1748.348058
Cluster-2	3	907.6919822
Cluster-3	7	3625.242085
Overall	22	2230.906692

Clusters 1 and 2 are relatively tight, cluster 3 very loose

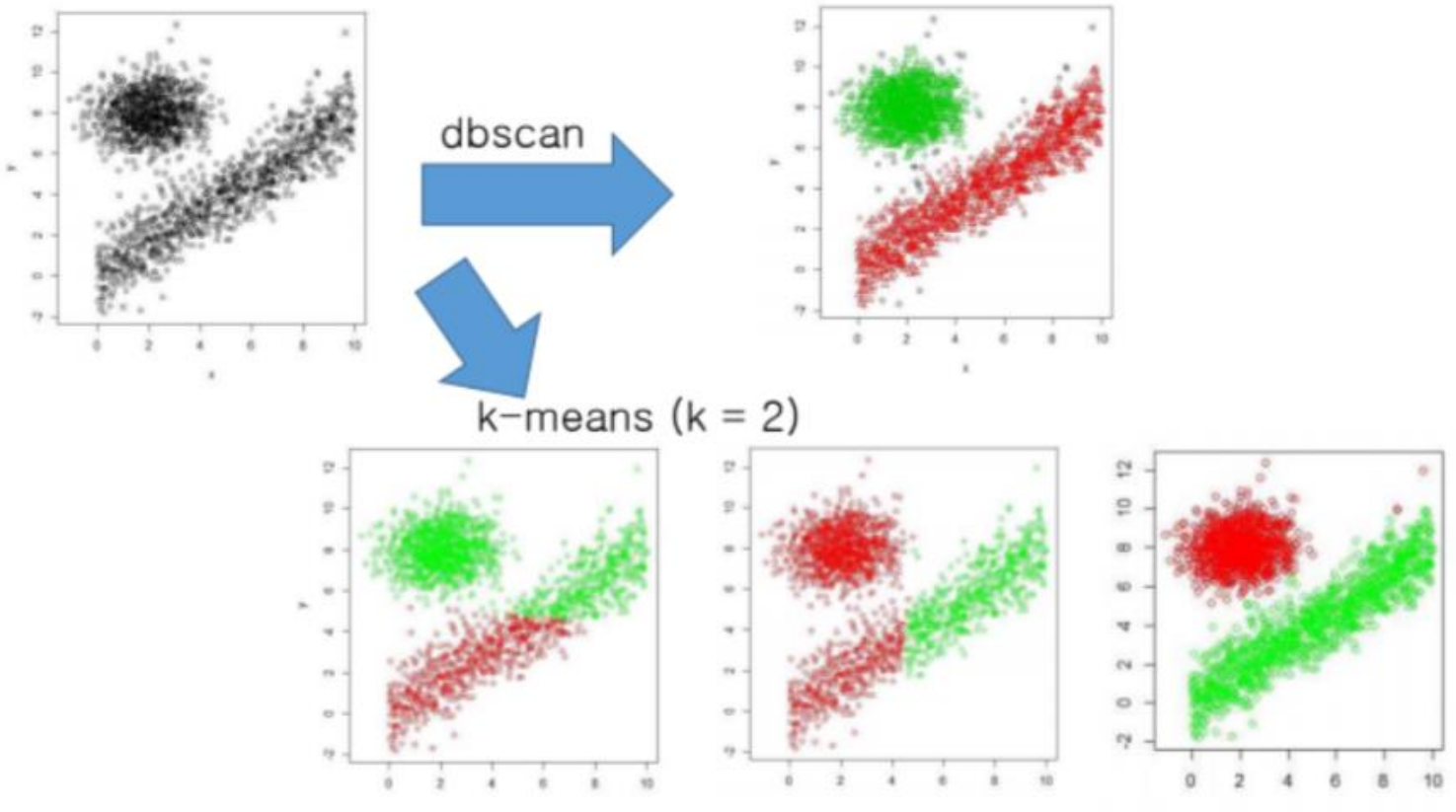
Conclusion: Clusters 1 & 2 well defined, not so for cluster 3

Next step: try again with k=2 or k=4

Drawback of K-means



K-means vs. DBSCAN



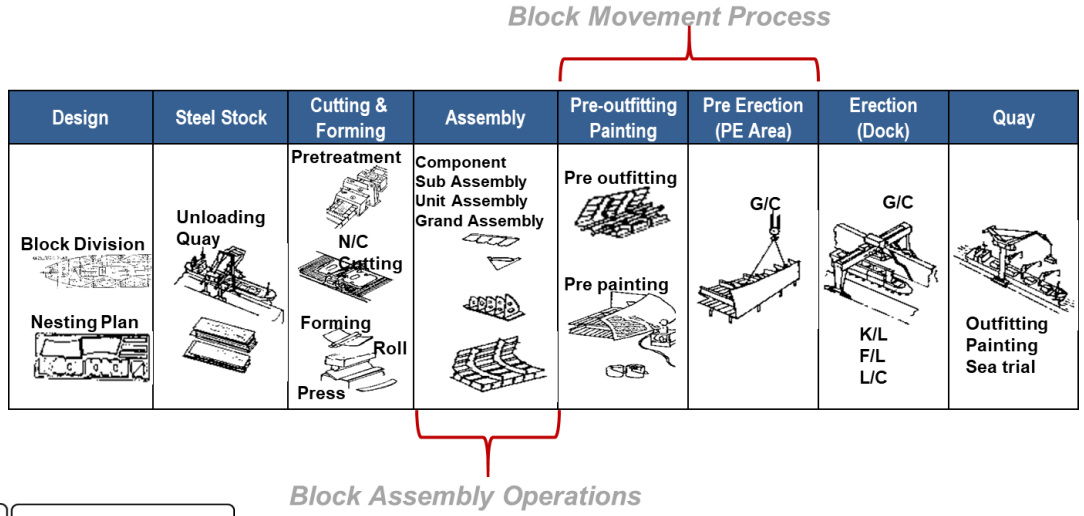
Source: https://ybeaning.tistory.com/27

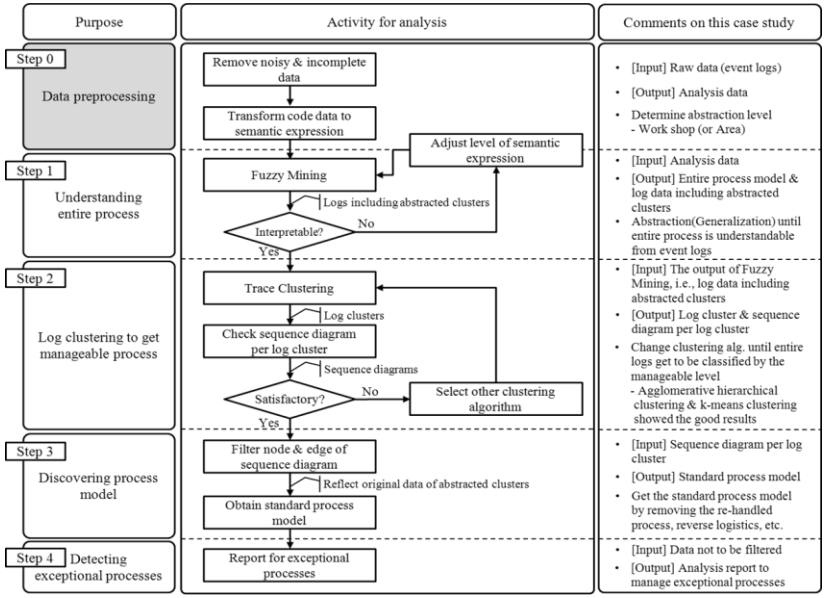
Summary

- Cluster analysis is an exploratory tool. Useful only when it produces **meaningful** clusters
- **Hierarchical** clustering gives visual representation of different levels of clustering
  - On other hand, due to non-iterative nature, it can be unstable, can vary highly depending on settings, and is computationally expensive
- **Non-hierarchical** is computationally cheap and more stable
  - requires user to set *k*
- Can use both methods
- Be wary of chance results; data may not have definitive "real" clusters

Case study: Clustering of parts based on production flow

- Shipyard analysis





Field	Description	Mapping to PM
PROJ	Project (Ship) ID	Case
BLK	Block ID	Case
WORKSTAGE	Work step	Other
SHOP	Work place	Activity
WORKCENTER	Work center	Activity
JL	Amount	
MH	Man hour	
STARTDATE	Start date	
FINISHDATE	End date	Timestamp
SUBCNTRCODE	Organization code	Resource
SUBBANCODE	Sub Org. code	Resource
WS_DESC	Description on ws	Other

Distance measure

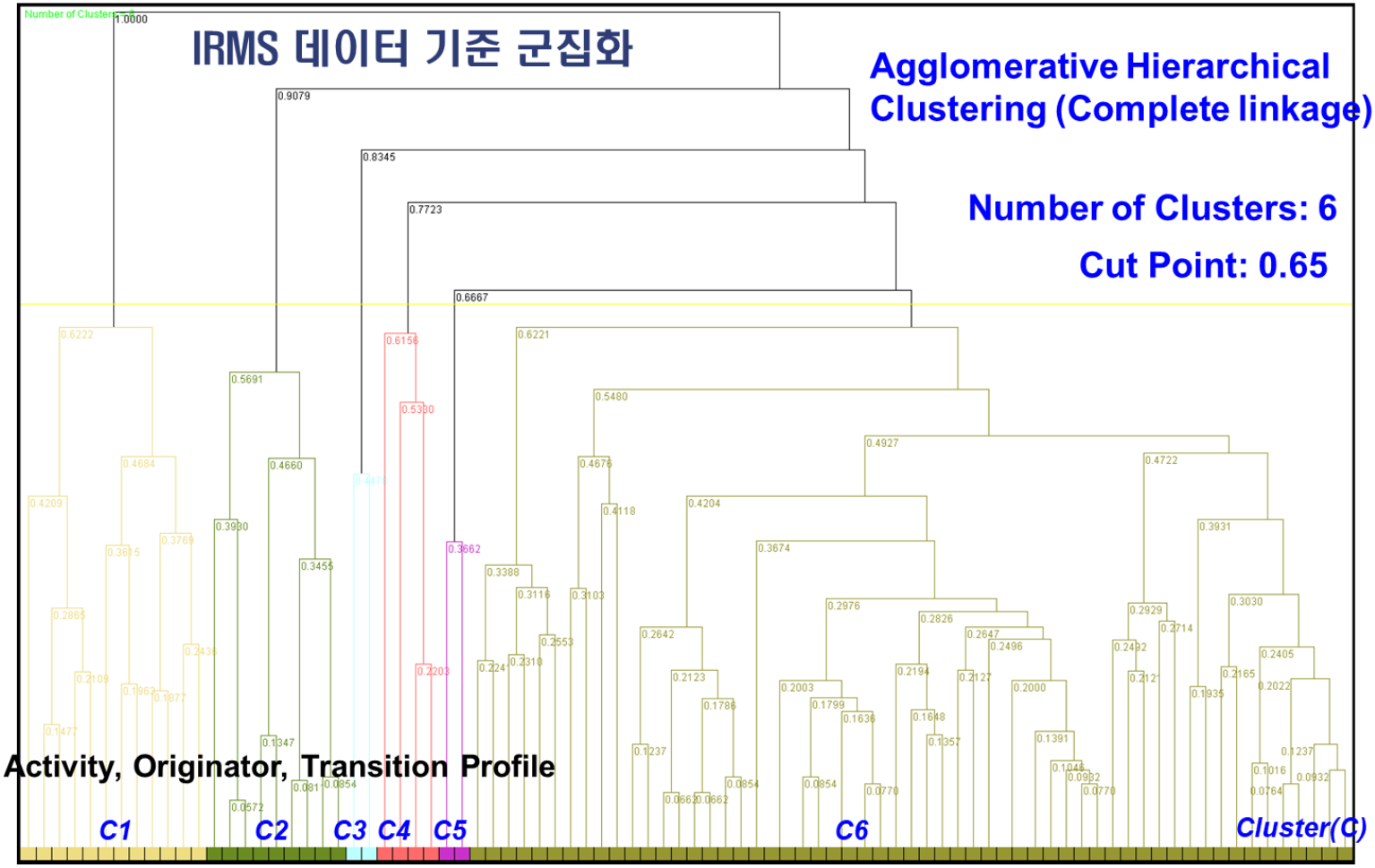
- How to measure the distance between traces?

Project	Block	Workshop	Moving Datetime	Organization	Transporter
A	Block 1	Assembly Shop	Oct-26 09:31	BB	Z
A	Block 1	Outfitting Shop	Nov-24 22:20	BB	Y
A	Block 1	Quay+Stock Area	Nov-26 21:20	CC	X
A	Block 1	Blasting Shop	Dec-08 06:10	DD	Y
A	Block 1	Painting Shop	Dec-09 11:10	DD	Y
A	Block 1	Painting Area (Outdoor)	Dec-13 22:35	DD	Y
A	Block 1	Stock Area	Dec-20 20:14	DD	X
A	Block 1	HEAVY+PE Area	Dec-21 08:46	DD	X

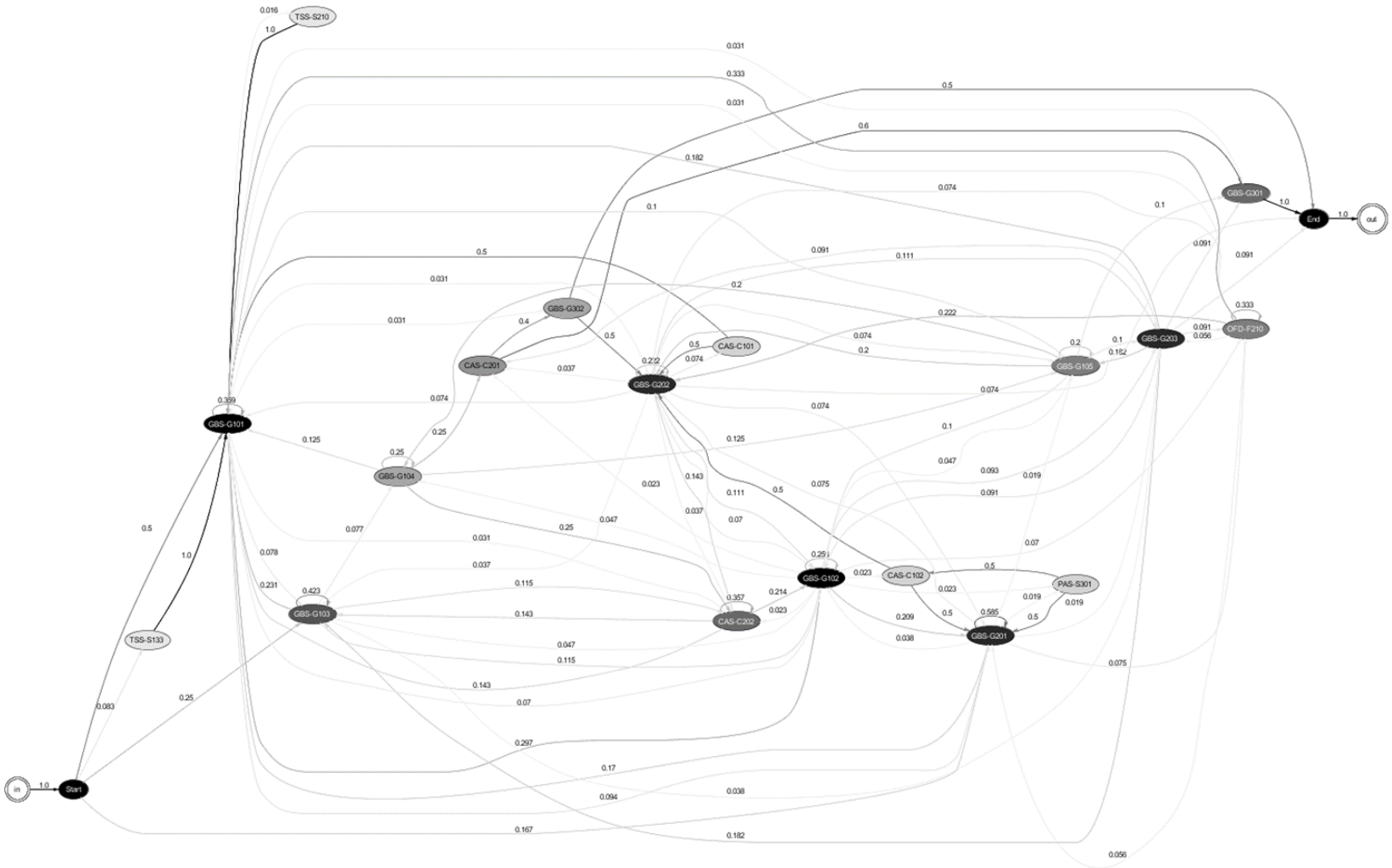
- Activity profile
- Transition profile

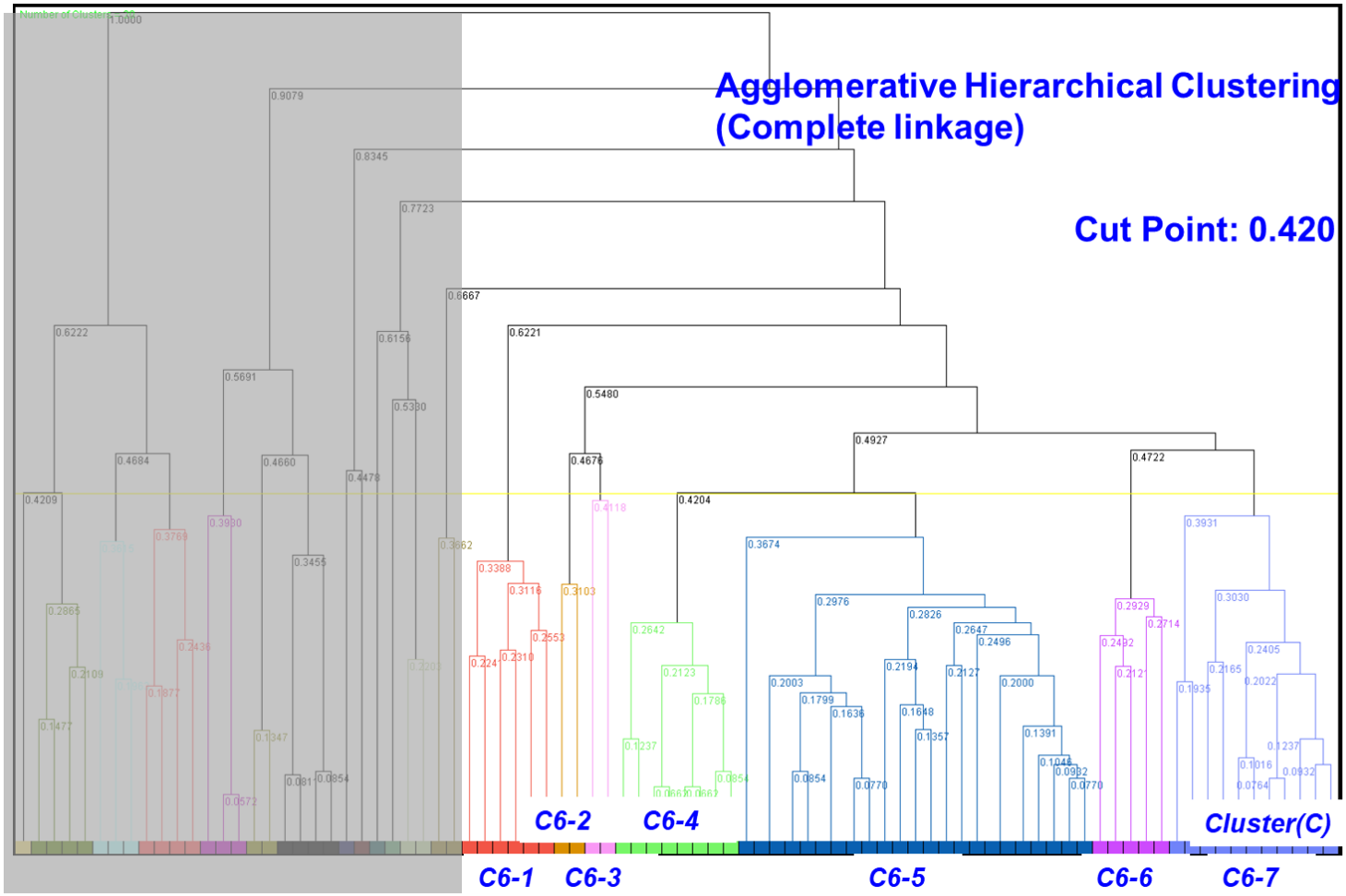
Block	Event log
Block 1	{S, G}
Block 2	{S, U, G}
Block 3	{G}
Block 4	{S, G}
Block 5	{S, U, G}

Block	Activity profile			Transition profile		
	S	U	G	S-G	S-U	U-G
Block 1	1	0	1	1	0	0
Block 2	1	1	1	0	1	1
Block 3	0	0	1	0	0	0
Block 4	1	0	1	1	0	0
Block 5	1	1	1	0	1	1


IRMS 데이터 기준 군집화
Agglomerative Hierarchical Clustering (Complete linkage)
Number of Clusters: 6
Cut Point: 0.65
Activity, Originator, Transition Profile
C1
C2
C3
C4
C5
C6
Cluster(C)

- Cluster 1: Block assembled in GBS




Agglomerative Hierarchical Clustering
(Complete linkage)
Cut Point: 0.420
C6-1
C6-2
C6-3
C6-4
C6-5
C6-6
C6-7
Cluster(C)

번호	군집1	군집2	블록수	군집이름	블록
1	C1		12	조립5공장(GBS)에서 블록 조립	EM0-209, EM0-258, EM0-248, EM0-259, EM0-249, EM0-112, EM0-131, EM0-121, EM0-642,EM0-652, EM0-238, EM0-228
2	C2		9	조립4공장(OFD)에서 소조 작업 후 조립 4공장 또는 조립5공장(GBS)에서 대조 작업	DM9-115, EM0-171, EM0-161, EM0-261, EM0-271, EM0-638, EM0-628, DM9-156, DM9-146
3	C3		2	(정하지 못함)	EM0-803, EM0-213
4	C4		4	CAS, PAS에서 콤프, 판작업 후 조립3공장(NPS)에서 중,대조 작업	DM9-107, EM0-106, DM9-113, EM0-115
5	C5		2	조립2공장(3DS)에서 대조 작업 (CAS에서 콤프 제작)	DM9-105, EM0-105
6	C6	C6-1	6	조립1공장(PBS)에서 대조 작업	DM9-232, DM9-222, DM9-234, DM9-233, DM9-224, DM9-223
7		C6-2	2	조립2공장(3DS)에서 대조 작업 (CAS, TSS, PAS에서 콤프/소조/판 제작)	DM9-121, DM9-131
8		C6-3	2	조립2공장(3DS)에서 대조 작업 (CAS, TSS, PAS, PBS에서 콤프 제작)	
9		C6-4	8	조립1공장(PBS)에서 대조 블록	
10		C6-5	23	(정하지 못함	
11		C6-6	5	3DS 또는 TSS 대조 블록 조립	
12		C6-7	11	3DS/TSS/CAS/PAS에서 소중조 또는 OFD에서 대조 블록 조립	

번호	배량 케이스	블록수	군집특징	블록
1	3D6	19	대조 L/T이 7일이 적용되는 3DS 일반 BLK, 모든 본 송선을 3DS 자체 제작	DM9-121, DM9-803, DM9-801, DM9-105, EM0-195, EM0-193, EM0-164, EM0-185, EM0-183, EM0-804, EM0-803, EM0-802, EM0-801, EM0-174, EM0-213, EM0-212, EM0-211, EM0-105, DM9-131
2	PB4	18	PBS LINE BLK(T/BOX BLK)	DM9-126, EM0-156, DM9-634, DM9-633, DM9-632, DM9-631, EM0-126, DM9-624, DM9-623, DM9-622, DM9-621, EM0-176, EM0-175, EM0-136, DM9-136, EM0-166, EM0-165, EM0-146
3	GB6	17	대조 L/T이 7일이 적용되는 GBS일반 BLK, 일반형상의 Block으로 모든 본 송선 GBS 자체 제작	EM0-261, EM0-259, EM0-258, DM9-114, EM0-112, EM0-249, EM0-248, DM9-104, EM0-131, EM0-238, EM0-121, EM0-228, EM0-642, EM0-652, EM0-104, EM0-271, EM0-209
4	PB1	6	PBS LINE BLK(중조 PBS LINE)	DM9-234, DM9-233, DM9-232, DM9-224, DM9-223 DM9-222
5	3D1	5	대조 L/T이 9일이 적용되는 3DS BLK(E/ROOM D/BOTTOM), 대조 용접 물량이 많은 BLK	DM9-103, DM9-102, DM9-101, EM0-103, EM0-101
6	TS2	5	TSS에서 제작하는 일반적인 BLK	EM0-154, EM0-194, EM0-184, EM0-116, EM0-144
7	OF5	5	일반적인 OFD BLK	EM0-638, EM0-628, DM9-115, EM0-171, EM0-161
8	NP1	4	NPS 고정 Block	DM9-117, DM9-107, EM0-115, EM0-106
9	PA2	2	PBS H2(LINE) 제작 후 OFD 완료 BLK	DM9-156, DM9-146
10	NA2	2	NPS에서 H2(LINE) 제작 후 OFD에서 대조 완료 BLK	DM9-231, DM9-221
11	GB1	1	대조 L/T이 9일이 적용되는 GBS BLK(E/ROOM D/BOTTOM)	EM0-102
12	GB2	1	STERN BOSS BLK(110BLK) 대조 L/T 18일	EM0-110
13	3N1	1	3DS에서 곡중조를 공급해서 NPS에서 대조를 완료하는 BLK	DM9-113

배량 케이스 기준 군집

군집	블록	차이 값		계획		실적	
		Σ Event	Σ Task	Fit	Cross Fit	Fit	Cross Fit
3D6	19	-17	12	0.375	0.118	0.357	0.167
PB4	18	-13	3	0.78	0.141	0.424	0.28
GB6	17	-17	9	0.375	0.16	0.422	0.307
PB1	6	18	5	0.715	0.111	0.353	0.11
TS2	5	5	3	0.883	0.5	0.86	0.6
3D1	5	-1	1	0.823	0.148	0.825	0.28
OF5	5	7	2	0.37	0.15	0.36	0.21
NP1	4	4	2	0.711	0.128	0.6	0.134
PA2	2	-2	0	0.816	0.208	0.833	0.244
NA2	2	4	1	0.875	0.183	0.717	0.196
GB1	1	0	0	1	1	1	1
GB2	1	-1	0	1	0.8	1	0.667
3N1	1	0	0	0.85	0.125	0.875	0.1
정리(전체)	86	89 (+38, -51)	38 (+38,-0)	0.736	0.290	0.664	0.330
정리(상위 30%)	65	70	32	0.626	0.206	0.483	0.293

IRMS 기준 군집

군집	블록	차이 값		계획		실적	
		Σ Event	Σ Task	Fitness	Cross Fit	Fitness	Cross Fit
C6-5	23	-27	10	0.756	0.307	0.706	0.387
C1	12	-17	3	0.305	0.131	0.386	0.237
C6-7	11	4	7	0.634	0.219	0.608	0.131
C2	9	5	-2	0.504	0.125	0.355	0.186
C6-4	8	16	1	0.873	0.128	0.458	0.247
C6-1	6	18	5	0.715	0.111	0.353	0.11
C6-6	5	-3	6	0.6	0.144	0.711	0.215
C4	4	4	-1	0.654	0.131	0.493	0.084
C6-3	2	-5	5	0.726	0.156	0.529	0.21
C3	2	-8	4	0.694	0.117	0.754	0.167
C5	2	0	3	0.78	0.115	0.707	0.101
C6-2	2	0	4	0.696	0.096	0.621	0.068
정리(전체)	86	107 (+47,-60)	51 (+48,-3)	0.661	0.148	0.557	0.179
합계(상위 30%)	63	69	23	0.614	0.182	0.503	0.238