Supplier quality performance evaluation and segmentation using k-means clustering

S Gunawan[1,3], H L Olinger[1,4], H L Chen[1,5], D M Putri[2,6] and Cynthia[2,7]

[1]Department of Industrial & Systems Engineering, Chung Yuan Christian University, Taiwan
[2]Department of International Business, Chung Yuan Christian University, Taiwan

[3]sebastiangunawan25@gmail.com, [4]hanslouisolinger120@gmail.com,
[5]achen@cycu.edu.tw, [6]devanimulya01@gmail.com, [7]Aureliacynthia00@gmail.com

Abstract. In the last few years, Supplier Relationship Management (SRM) becomes an increasingly popular research field in supply chain management. It impacts every aspect of a company, including raw material quality, delivery time, warranties, performance history, price, etc. The competition created by using multi-sourcing strategy will benefit the company by giving him more choices and the most reliable service available. In this study, suppliers are grouped according to their performance evaluation, and recommendations are made based on the material quality history of each supplier to create data-driven model using K-Means Clustering. Based on the "Cashew Trucks Arrival" dataset from Kaggle, we will segment suppliers according to moisture percentage, outturn level, and defect percentage. The result showed there were two groups of suppliers which were "Better Quality" and "Worse Quality" suppliers group. The model must be tested on other critical factors or implemented on real company data.
Keywords: ***Supplier Evaluation, Data Driven, Multi-Sourcing, Supplier Segmentation, K-Means Clustering***

7. Introduction

Supplier Relationship Management (SRM) is recognized as one of the most critical components of business continuity and performance. A successful company will prioritize supplier-buyer relationships because it affects several factors in the supply chain, such as material quality, delivery system, services, price advantages, and even collaboration during product development. To maximize sales, every company has to collaborate with its suppliers; thus, integration between a company's division and its suppliers becomes much more necessary than ever. As an example of supply chain innovation, Vendor Managed Inventory (VMI) is used to increase efficiency in production, inventory, and transportation costs by integrating the same information between buyer and supplier (Marques et al., 2010). Another innovation example is evaluating suppliers to improve supply chain performance.

An important part of a supplier-buyer relationship is supplier evaluation. Imeri (2013) described 23 parameters for assessing suppliers based on importance ratings. Four parameters are most significant for supplier evaluation: quality, delivery, warranties, and performance history. Among the factors for evaluating suppliers, price advantages were added by one researcher. In Ho et al. (2010), various factors can be considered when evaluating and selecting suppliers. As a result, it will be easy to maintain long-term partnerships. Prahinski and Benton's (2004) study explained that supplier evaluations are one of the types of supplier development programs that emphasize current and future business strategies so that all companies involved can contribute to the program's development.

Rather than selecting one supplier, many companies employ multi-sourcing strategies. Hence, companies prefer creating a competitive environment among suppliers to ensure higher quality service. At the same time, having a wide range of suppliers also helps companies minimize the risk of material shortages. Accordingly, supplier segmentation will be a crucial part of a company's strategy for organizations with a large number of suppliers. In this way, they will be able to evaluate their performance and order raw materials from suppliers who perform better. Such segmentation is done by K-Means Clustering, an unsupervised machine-learning approach popular in many applications. For example, Murray et al. (2015) used K-Means to segment customers' behaviour and to forecast consumption behaviour for each customer segment. Sandi et al. (2018) implemented K-Means with an RFM model to cluster customer behaviour during CRM research. Gonçalves et al. (2021) identified a risk profiles in automotive electronics supply chain using descriptive data mining and K-means. While uncovering the inherent cluster structure in data, clusterability is evaluated in various ways. For one-dimensional data reductions, tests on one-dimensional reductions are frequently accurate at identifying clusters in data sets; distance-based methods also perform well in most scenarios (Adolfsson et al., 2019)

Using Cashew Delivery Data from Kaggle, this study seeks to gain an in-depth understanding of this problem. Raw cashew nut consumption and production vary despite a global trend. At the moment, Vietnam and India are two of the world's largest processors and suppliers of cashew nuts. Other emerging developing country suppliers are Indonesia, Brazil, Nigeria, Kenya, Honduras, and others. Many companies establish long-term partnerships with cashew nut suppliers to ensure constant supply and stable prices. This research attempts to create data-driven suppliers segmentation model for the Cashew market, and provide recommendations based on their performance. We will segment each supplier's moisture percentage, outturn level, and defect percentage history. In the following section, we will explain the research methodology briefly, the results, and the conclusion of this research.

8. Research Methodology

A flowchart illustrating the research methodology is shown in **Figure 1**. In the first step, it is imperative to determine the problem through literature research and to identify research objectives. Pre-processing of the data is necessary to begin the research. First, the moisture level, the outturn level, and the defective percentage are presented and compared in this study. Database resources describe some terminologies; for example, moisture level shows raw cashew nut moisture percentage. The Outturn level describes the quality metric used to calculate the ratio of high quality to a total quantity of cashew nuts per 80 kilograms. Last but not least is the defective percentage, which shows the number of defective cashew kernels.

Upon selecting the variables that will be compared in the research, the data will be cleaned to remove "Not Available" data, as shown in **Figure 2**. In order to keep the dataset from becoming smaller, we use the mean for each variable to fill in NA data. The mean is selected because it will not change the pattern of the dataset. It will not affect the overall average itself.

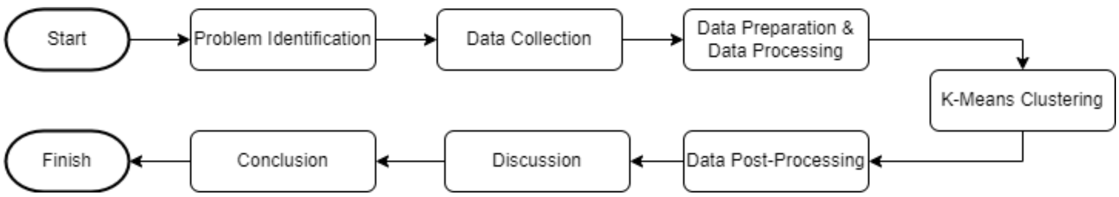

Figure 1. Methodology Implementation

After preprocessing, K-means clustering is applied. This research employs the Elbow and Silhouette methods to determine the optimum number of clusters. With each iteration of the Elbow method, the inertia value will be calculated for a different number of clusters. The inertia value is the sum of the squared distances between variables from their nearest cluster centers according to the

Euclidean distance principle. Clusters with minimum inertia values are the most suitable candidates for choosing the appropriate number of clusters, while more clusters will decrease inertia values. The inertia values will need to be weighed against the number of clusters, so we need to assess the trade-off between them. The number of clusters will be determined if there is a significant decrease in inertia value between the previous number of clusters and the current number of clusters. Silhouette coefficients show how close each cluster is to the centroid. A high Silhouette coefficient indicates a low average distance between the data and the cluster center (Kit & Azmi, 2021).

According to K-Means Clustering, the result from those methods will be several clusters. First, the dataset will be divided into clusters based on the center of the clusters. Then, cluster center coordinates will be compared, and suppliers will be categorized and interpreted. It is possible, through postprocessing, to determine which cluster, out of multiple clusters, will be the most suitable place to locate one supplier. Finally, a summary of supplier performance segmentation will be generated based on data processing and further implications based on the segmentation.

9. Result and Discussion

The data in this study will come from a Kaggle dataset called "Cashew Trucks Arrival," which contains 670 data since March 2015 to May 2017 and 15 variables, including number, date, truck ID, number of bags, net weight, origin, supplier ID, moisture percentage, number of nuts, outturn level, defective percentage, price, year, month, day, and average weight per bag. Out of those 15 variables, 4 variables are being kept. Variable "supplier" shows which supplier has done the delivery activity. While other variables, such as "moisture," "outturn," and "defective," represent cashew nuts quality performance based on the available dataset. **Figure 2** shows available variables in the data set that will be evaluated in the study.

	Date	truckid	nbags	net_weight	origin	supplier	moisture	nut_count	outturn	defective	price	year	month	day	avg_wpb
0	05/03/2015	8553FA02	567	40597	66	119	11.30	191.0	48.40	10.25	NaN	2015	3	5	71.599647
1	09/03/2015	7162EG01	226	20253	82	123	12.60	184.0	48.22	8.90	NaN	2015	3	9	89.615044
2	10/03/2015	4819CJ01	506	38878	51	15	10.10	194.0	45.76	10.90	NaN	2015	3	10	76.833992
3	11/03/2015	8581FK05	422	39354	47	66	10.00	183.0	48.93	9.10	NaN	2015	3	11	93.255924
4	11/03/2015	5066FL01	555	40009	66	119	11.30	193.0	46.29	12.65	NaN	2015	3	11	72.088288
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
666	27/05/2017	1735HH07	460	35240	66	85	16.83	194.0	36.78	NaN	950.0	2017	5	27	76.608696
667	27/05/2017	2344FP01	480	39340	66	85	16.23	182.0	41.88	NaN	950.0	2017	5	27	81.958333
668	27/05/2017	4165HG01	141	12071	43	2	18.56	165.0	41.00	NaN	850.0	2017	5	27	85.609929
669	29/05/2017	1074FL01	442	37878	43	2	17.30	170.0	46.20	NaN	850.0	2017	5	29	85.696833
670	30/05/2017	2027HH07	335	29157	59	67	14.30	170.0	45.67	NaN	950.0	2017	5	30	87.035821

Figure 2. Cashew Truck Arrival Dataset

As shown in **Figure 2,** some data are also "Not Available." Therefore, the data will be filled by mean values from each variable to make them more suitable for K-Means Clustering analysis. Another option is to remove all "Not Available" data, but we did not do that because it would reduce a great deal of data. Given these two considerations, it is safest to fill the data with a mean to not change the existing data pattern and overall average after replacing NA data with the mean ones. The result is shown in **Figure 3** after the dataset has been filled with the mean. For example, the NA on the defective column in **Figure 2,** replaced with 14.532129 in **Figure 3,** shows that the value is the average and **Figure 4** shows that after replacing the NA data, the overall average did not change.

	Date	truckid	nbags	net_weight	origin	supplier	moisture	nut_count	outturn	defective	price	year	month	day	avg_wpb
0	05/03/2015	8553FA02	567	40597	66	119	11.30	191.0	48.40	10.250000	750.390625	2015	3	5	71.599647
1	09/03/2015	7162EG01	226	20253	82	123	12.60	184.0	48.22	8.900000	750.390625	2015	3	9	89.615044
2	10/03/2015	4819CJ01	506	38878	51	15	10.10	194.0	45.76	10.900000	750.390625	2015	3	10	76.833992
3	11/03/2015	8581FK05	422	39354	47	66	10.00	183.0	48.93	9.100000	750.390625	2015	3	11	93.255924
4	11/03/2015	5066FL01	555	40009	66	119	11.30	193.0	46.29	12.650000	750.390625	2015	3	11	72.088288
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
666	27/05/2017	1735HH07	460	35240	66	85	16.83	194.0	36.78	14.532129	950.000000	2017	5	27	76.608696
667	27/05/2017	2344FP01	480	39340	66	85	16.23	182.0	41.88	14.532129	950.000000	2017	5	27	81.958333
668	27/05/2017	4165HG01	141	12071	43	2	18.56	165.0	41.00	14.532129	850.000000	2017	5	27	85.609929
669	29/05/2017	1074FL01	442	37878	43	2	17.30	170.0	46.20	14.532129	850.000000	2017	5	29	85.696833
670	30/05/2017	2027HH07	335	29157	59	67	14.30	170.0	45.67	14.532129	950.000000	2017	5	30	87.035821

Figure 3. Filling "Not Available" Data

```
supplier["defective"].mean()
```

```
14.532129032258066
```

Figure 4. Overall Average After Replacing NA Data

As part of the research, the optimal number of clusters will be determined by the Elbow and Silhouette methods. The result of the Elbow method is shown in **Figure 5**. The most significant reduction is from one cluster to two clusters. As mentioned before, the number of clusters will be determined if there is a significant decrease in inertia value between the previous and current clusters. Meanwhile, the **Silhouette method** shows the same results as the **Elbow method**, where the coefficients show how close each cluster is to the centroid. A high Silhouette coefficient indicates a low average distance between the data and further dispersion between centroids. When both methods yield the same result, only using two clusters will yield the highest Silhouette coefficient.

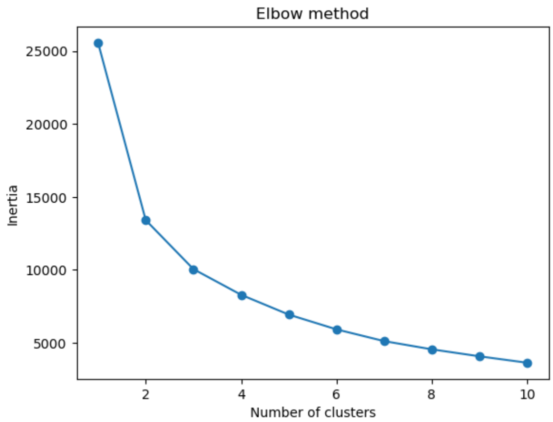

Figure 5. A result of the Elbow method

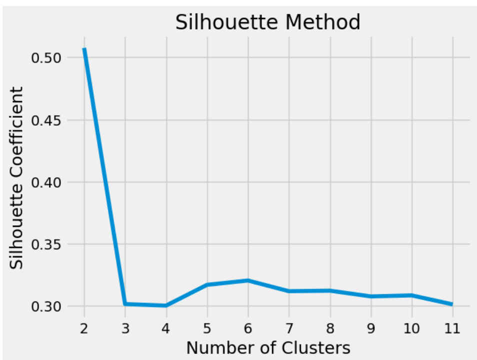

Figure 6. The result of the Silhouette method

Two clusters can be identified through moisture percentages at 12.8954 and 13.7965, output at 46.4313 and 42.6497, and defective percentages at 12.1928 and 22.2358, shown in **Table 1**. Cluster 1 can be categorized as "Higher Quality" supplier groups, while Cluster 2 can be categorized as "Lower Quality" supplier groups. Furthermore, there are differences between those parameters. From moisture percentage, the result from cluster 1 is 12.8954, while the result from cluster 2 is 13.7965. Low moisture affects the quality of cashews (Falade et al., 2003). Therefore, cluster 1, with a higher outturn level but a lower defective percentage than Cluster 2, indicates a better-quality product than Cluster 2. The results range also shows the differences between two clusters with variables perspective for each cluster and variable in **Table 2**. Cluster 1 has the lowest range for Moisture Percentages compared with Cluster 2. Regarding the outturn Level, Cluster 2 generated a slightly lower result than Cluster 1. The most significant difference between variables is the Defective Percentage, in which cluster 1's maximum value is 17.80, but cluster 2's minimum value is 14.53. It explains that there is a significant difference between those two clusters.

Table 1. Result of K-Means Clustering.

Cluster	Moisture Percentage	Outturn Level	Defective Percentage
1	12.8954	46.4313	12.1928
2	13.7965	42.6497	22.2358

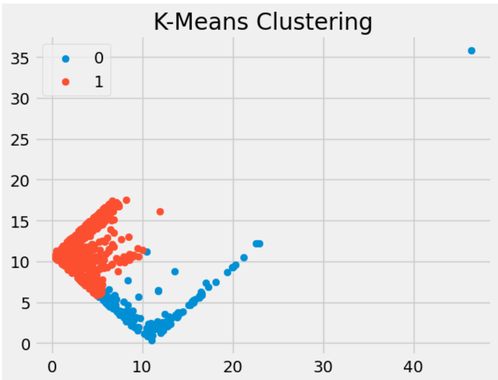

Figure 7. Clustering Result

	supplier	moisture	outturn	defective	Cluster
224	43	11.60	47.69	12.4	1
226	43	12.60	45.84	15.6	1
229	43	18.00	45.40	12.8	1
230	43	14.00	46.11	13.9	1
232	43	13.20	46.28	13.3	1
...	...	...	...	...	...
451	43	13.40	43.82	20.5	0
452	43	17.03	45.32	15.7	1
460	43	11.80	41.71	23.4	0
461	43	12.70	41.18	25.5	0
463	43	16.40	39.00	33.6	0

Figure 8. Supplier 43 Detected on Both Clusters

We found that some suppliers were detected in two clusters when we processed the data. To solve this problem, we can filter on the supplier ID identified in both groups and calculate the number of data for each supplier and cluster. Some data for each supplier should be compared between two clusters. If the data in Cluster 1 is more than in Cluster 2, the supplier will be categorized as Cluster 1. Otherwise, the supplier will be categorized as Cluster 2. as shown in **Appendix 1**. For example, **Figure 8** indicates that supplier 43 was detected in both groups. As we can see on data 224 to 232, it is categorized as Cluster 1, while data 451, 460, 461, and 463 are categorized as Cluster 2. In **Figure 9**, supplier 43's performance in Cluster 1 and Cluster 2 is shown as a percentage.

Table 2. Cluster Variables Range

Cluster	Moisture Percentage	Outturn Level	Defective Percentage
1	1.00 - 21.20	38.80 - 50.68	6.00 - 17.80
2	9.26 - 22.47	36.43 - 47.25	14.53 - 57.50

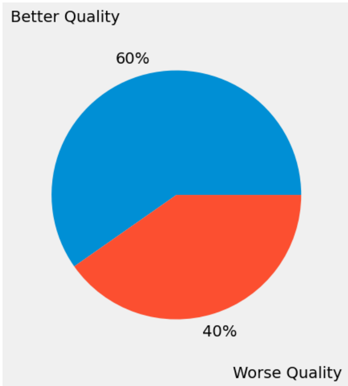

Figure 9. Performance of Supplier 43

In total, 60% of supplier 43's performance history or 46 out of 77 deliveries, is clustered under the "Better Quality" cluster. As a result, supplier 43 will be estimated to have better than worse quality. On the other hand, based on this result, we could consider giving Supplier 43 more evaluations to enhance its product quality.

In the context of management, the company may encourage suppliers whose performance is plotted as cluster 2 to improve their productivity and quality. The company can reward cluster 1 suppliers for their performance by serving high-quality products. After all, we all know that evaluations and rewards affect the relationship between suppliers and buyers.

This research came with some limitations. In the real-industry cases, raw material delivery will occur repeatedly and continuously. This causes delivery data will always increase. Through the use of Machine Learning approach, a data-driven model is created that can generate an adaptive decision to segment supplier based on real-time situation and performances. A dataset called "Cashew Trucks Arrival" were used for this research which had limited amount of data. However, the model has been created and could be implemented to real-based datasets on supplier segmentations.

This dataset also had variables limitations. As known, several factors influence the quality of cashew nuts kernels, such as their size, shape and colour, as well as the percentage of defective nuts. Taste and flavour, however, are subjective and cannot easily be measured based on physical characteristics, according to the industry. As such, the demand and price fluctuations for cashew nuts are highly related to taste and flavour. The datasets contain only 15 variables with 3 quality metrics variables (moisture, outturn, and defective), 6 delivery information variables (date, truckid, origin, supplier, year, month, and day), 2 price variables (price and avg_wpb), and 3 cashew amount variables (nbags, net_weight, and nut_count) as on **Figure 2**. More potential variables are recommended to segment suppliers for the further research.

10. Conclusion

In this study, we have already examined using K-Means clustering to select suppliers based on historical data categories. Two clusters were generated, each with some pros and cons. For example, a company can select the first cluster with better-quality cashew production. However, this research has some limitations. First, there are four parameters for supplier performance evaluation and segmentation. The performance metrics of suppliers should be considered as another parameter. Furthermore, we used secondary datasets that were already available in Internet databases. For future research, this method will be applied to real-based datasets on supplier segmentation based on quality metrics or using another advanced metrics.

References

[1] Adolfsson, A., Ackerman, M., & Brownstein, N. C. (2019). To cluster, or not to cluster: An analysis of clusterability methods. Pattern Recognition, 88, 13-26.

[2] Bai, C., Rezaei, J., & Sarkis, J. (2017). Multicriteria Green Supplier Segmentation. *IEEE Transactions on Engineering Management, 64*(4), 515-528.

[3] Dobos, I., & Vorosmarty, G. (2014). Green Supplier Selection and Evaluation using DEA-Type Composite Indicators. *International Journal of Production Economics, 157*, 273-278.

[4] Falade, K., Akinwale, T., & Adedokun, O. (2003). Effect of Drying Methods on Osmotically Dehydrated Cashew Apples. *European Food Research and Technology, 216*(6), 500-504.

[5] Gonçalves, J. N., Cortez, P., & Carvalho, M. S. (2021). K-means clustering combined with principal component analysis for material profiling in automotive supply chains. European Journal of Industrial Engineering, 15(2), 273-294.

[6] Ho, W., Xu, X., & Dey, P. K. (2010). Multi-Criteria Decision Making Approaches for Supplier Evaluation and Selection: A Literature Review. *European Journal of Operational Research, 202*(1), 16-24.

[7] Imeri, S. (2013). Key Performance Criteria for Vendor Selection - A Literature Review. *Management Research and Practice, 5*(2), 63-75.

[8] Kit, T. C., & Azmi, N. F. (2021). Customer Profiling for Malaysia Online Retail Industry using K-Means Clustering and RM Model. *International Journal of Advanced Computer Science and Applications, 12*(1), 106-113.

[9] Marques, G., Thierry, C., Lamothe, J., & Gourc, D. (2010). A review of Vendor Managed Inventory (VMI): From Concept to Processes. *Production Planning & Control, 21*(6), 574-561.

[10] Murray, P. W., Agard, B., & Barajas, M. A. (2015). Forecasting Supply Chain Demand by Clustering Customers. *IFAC-PapersOnLine, 48*(3), 1834-1839.

[11] Murugi, N. M., & Shalle, N. (2016). Role of Supplier Relationship Management on Procurement Performance in Manufacturing Sector in Kenya: A Case of East African Breweries. *International Academix Journal of Procurement and Supply Chain Management, 3*(4), 1-20.

[12] Prahinski, C., & Benton, W. (2004). Supplier Evaluations: Communication Strategies to Improve Supplier Performance. *Journal of Operations Management, 22*(1), 39-62.

[13] Prasad, S. H., Kamath, G., Barkur, G., & Naik, R. (2016). Does Supplier Evaluation Impact Process Improvement? *Journal of Industrial Engineering and Management, 9*(3), 708-731.

[14] Rezaei, J., & Ortt, R. (2013). Multi-Criteria Supplier Segmentation using a Fuzzy Preference Relations based AHP. *European Journal of Operational Research, 225*(1), 75-84.

[15] Sandi, T. A., Raharjo, M., Putra, J. L., & Ridwan. (2018). Clustering Kesetiaan Pelanggan E-Ritel Dengan Model RFM (Recency, Frequency, Monetary) dan K-Means. *Jurnal Pilar Nusa Mandiri, 14*(2), 239-246.

[16] Walter, A. (2003). Relationship-Spesific Factors Influencing Supplier Involvement in Customer New Product Development. *Journal of Business Research, 56*(9), 721-733.

Appendix 1. Number of Data for Each Cluster and Supplier

Supplier	**Cluster 1**	**Cluster 2**	**Supplier**	**Cluster 1**	**Cluster 2**	**Supplier**	**Cluster 1**	**Cluster 2**	**Supplier**	**Cluster 1**	**Cluster 2**	**Supplier**	**Cluster 1**	**Cluster 2**
0	1	0	27	1	2	54	5	0	81	1	0	108	0	2
1	1	0	28	2	5	55	1	0	82	1	0	109	1	0
2	24	0	29	1	0	56	1	0	83	1	0	110	1	0
3	1	0	30	1	0	57	6	0	84	36	0	111	1	0
4	1	1	31	2	2	58	1	0	85	45	4	112	1	0
5	2	0	32	6	3	59	1	0	86	4	0	113	1	0
6	11	0	33	0	1	60	0	1	87	1	0	114	10	1
7	1	0	34	1	0	61	2	0	88	0	1	115	5	0
8	2	0	35	1	0	62	5	1	89	1	0	116	1	0
9	1	1	36	1	0	63	1	5	90	10	0	117	1	0
10	1	0	37	3	0	64	7	0	91	1	1	119	0	7
11	15	0	38	0	1	65	4	4	92	2	0	119	20	0
12	3	2	39	1	1	66	7	1	93	1	0	120	2	0
13	1	0	40	1	0	67	8	0	94	4	4	121	1	0
14	14	1	41	1	0	68	4	1	95	0	1	122	1	0
15	4	0	42	0	1	69	6	0	96	5	0	123	3	0
16	1	0	43	31	46	70	1	0	97	2	2	124	1	2
17	1	0	44	1	0	71	10	0	98	1	0	125	0	2
18	7	0	45	2	0	72	16	4	99	0	1	126	1	0
19	2	0	46	0	2	73	1	0	100	2	0	127	3	0
20	3	1	47	0	1	74	1	0	101	4	1	128	2	0
21	1	0	48	0	1	75	29	0	102	0	1	129	2	0
22	2	0	49	5	3	76	1	0	103	1	0	130	1	0
23	13	0	50	1	0	77	1	0	104	4	4	131	18	0
24	7	1	51	1	0	78	2	3	105	0	1	132	1	0
25	1	0	52	1	0	79	4	0	106	2	0			
26	1	1	53	7	0	80	3	1	107	6	0			