TEST Engineering & Management

November-December 2019
ISSN: 0193-4120 Page No. 2820 - 2825

Morphometric Analysis of Coleopteran Stored Product Pest using Clustering Techniques

Tun Mohd Firdaus Bin Azis,
Faculty of Applied Sciences, Universiti Teknologi Mara, Cawangan Perlis, 02600 Arau Perlis, Malaysia.
Shamshuritawati Sharif,
School of Quantitative Sciences, UUM College of Arts & Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah MALAYSIA.
Basharoh Abdul-Karim.
Kolej Poly-Tech MARA Alor Setar, Bangunan MARA Mergong, Seberang Jalan Putra, 05150 Alor Setar, Kedah
basharoh@gapps.kptm.edu.my
Haslina Zakaria.
Institute of Engineering Mathematics, Universiti Malaysia Perlis, Kampus Pauh Putra, 02600 Arau, Perlis, MALAYSIA
haslinazakaria@unimap.edu.my

Article Info
Volume 81
Page Number: 2820 - 2825
Publication Issue:
November-December 2019

Article History
Article Received: *5 March 2019*
Revised: *18 May 2019*
Accepted: *24 September 2019*
Publication: *14 December 2019*

Abstract:

Insect pest species identification is a crucial process before starting any pest management program. The similar morphological characteristics among the vast insect species cause identification process difficult. In this paper, the Coleopteran stored product pest is our main concern. It causes severe damage to the stored product and gives a negative impact on its value. Therefore, K-means clustering and Hierarchical Agglomerative Cluster Analysis (HACA) were used in the identification of Coleopteran stored product pest species. Four morphological structure of 38 Coleopteran stored product pest species image was used to generate the measurement that been used as simulation data. In total, there are 100 datasets produce each morphological structure for each image. The result from K-Means Clustering produces 5 clusters while Hierarchical Agglomerative Cluster Analysis (HACA) produces 11. From the Average Silhouette index, HACA performs better in clustering compare to K-means clustering.

Keywords*: Clustering analysis, Coleopteran, K-means clustering, Hierarchical Agglomerative Cluster Analysis.*

I. INTRODUCTION

Coleopteran stored product pest damages grain through their feeding and contamination after they cast their skin and faecal substance. Contamination leads to a drop in value and quality of the stored product in the market. Primary insect pests cause more severe damage because it inhabits and reproduces inside the storage of stored product where the insect density continues to grow from time to time. Consequently, the commodity stored in warehouse tends to reduce. In total, there are 65 Coleopteran stored product pest species. It has been categorized into four-level based on the degree of damages done to commodities [1]–[3].

In implementing any pest management program, the earlier crucial step is an identification of the insect pest species [4]. From the identification of insect species, the detail information their growth, type of damages, level of damages done to the grain will be obtained. [5], [6].The information is vital because different insect pest species cause a different type and level of injuries to the stored product [1]. This information has then been used to choose the best pest management program that uses the least cost and effect no or the least harm to the environment [5], [6]

Morphological key features for the specific species is needed in species identification. This can be obtained from the correct textbook and references [8]. The failure to get the proper references of possible species can lead to misidentification. Consequently, assessments of population status, the distribution could be affected. This can lead to erroneous conservation decisions [8], [9]. To confirm the species also can be done by sending the insect specimens to expert or taxonomist to further verify the species [2]. It demonstrated that in the identification of insect species is difficult.

Moreover, the vast number of insect pest's species make things even more complicated in the identification process, especially using the morphological image. Thus, in this study, we propose a statistical method as a tool to identify the species. In terms of cost, this techniques is cheaper to the aforementioned techniques. Clustering technique is implemented to cluster the Coleopteran stored product pest species.

The remainder of the paper is structured as follows. In the next section, a review on Coleopteran stored product pest is delivered. It is followed by data analysis and statistical analysis in section 3. Later, in the second last and the final section, clustering results and conclusion will be delivered, respectively.

II. REVIEW ON COLEOPTERAN STORED PRODUCT PEST

Stored product insect had a very long history with the human society where it can trace back 5000 to 10000 years ago. This significantly related to the agriculture activities and storage of large quantities of grains that attracted the storage grain pest. Among the insect pest, Coleopteran order becomes a significant pest for seed type storage, especially in the family of Curculionidae and Bruchinae [3].

Existence of insects inside stored product container causes many types of economic losses. The damages done by this insect differ with insect species, type of grain, and transporting and storage in the marketing system. Insects existence will lead to contamination because they cast skins, webbing, and eliminating faecal material that reducing the market value of commodities[1].

The stored product pest will reduce the amount of the commodity by continue to feed the commodities, and it became worse as they continue to have an increment in population growth [1], [3]. For example, larvae and adults of *Sitophilus* sp. can cause severe damage to stored grain. The adult *Sitophilus* sp. will lay an egg inside the grain, and then it hatches into larvae. The larvae then will continue to develop internally inside the grain. The larvae development consume up to 60% of the weight of single grain [10], [11].

This insect infestation also will encourage mould growth, including those fungi that produce mycotoxins [12], [13]. The existence of the insect inside the stored product storage will cause rejection by consumers. Then, there are involving the costs when the application of measures in controlling and preventing infestation. There are also issue in a method in controlling, especially that involving the pesticide and fumigant that is well known to affect other animals, human and environment [3], [14].

III. DATA PROCESSING AND STATISTICAL ANALYSIS

A. Secondary Data

PaDIL online database was used to retrieve the image of all Coleopteran stored product pest. It involves 38 Coleopteran stored product pest species. The list of this 38 Coleopteran stored pest is presented in **Error! Reference source not found.**.

Using the ImageJ 1.52a software morphometric data of thorax width, thorax length, body length, and length between thorax and abdomen were measured from the image of 38 Coleopteran stored product pest. [15]. From the measurement, 100 datasets been generated using the constants ratio increment from the lowest range to the highest range of body length. It follows the assumption that the growth in the other part of the morphological structure will affect the others. This mainly based on the ratio of the body part remains constants at any sized of the insect [7], [16].

A. K-Means

K-mean clustering and Hierarchical Agglomerative Cluster Analysis (HACA) were used in this study to identify the species of Coleopteran stored product pest. K-mean clustering is used in segmenting the 38 Coleopteran stored product pest using the data measured and generated from the image retrieve from PaDIL database. Three steps involved in performing K-mean clustering. The first step involves all item that is morphometric dataset from the four morphological structure of the Coleopteran stored product pest been partition into K initial clusters for the first step. In this step, the level involves scanning through the list of morphometric data. It then assigns each of it to the group whose centroid (mean) the nearest. The second step happens when each time the dataset or item in it reassigned, recalculation the cluster mean or centroid for the cluster receiving that item and the cluster losing that item will be done. Step three by repeating step two over and over again until no more reassignments are made [17].

B. HACA

The other method used in this study is HACA. This method is unsupervised involving identification of pattern and dividing it from a large group of data into the smaller group [18]. Precisely for this study, HACA been used to identify morphometric data of 38 Coleopteran stored product insects and group it data into the members of a significant cluster that shares high similarities [19]. HACA has been employed based on the normal distribution dataset using Ward's method and Euclidean distances, as a measure of the relationship [20].

TABLE 1: Coleopteran stored product pest list

No.	The species	Body length range(mm)
1	*Lasioderma serricorne*	2.0-3.0
2	*Stegobium paniceum*	2.25-3.5
3	*Araecerus fasciculatus*	3.0-5.0
4	*Dinoderus minutus*	2.2-4.5
5	*Prostephanus truncatus*	3.5-5.5
6	*Rhyzopertha dominica*	2.0-3.0
7	*Acanthoscelides obtectu*	3.0-5.0
8	*Zabrotes subfasciatus*	1.8-2.8
9	*Necrobia rufipes*	3.5-7.0
10	*Cryptophagus cellaris*	2.0-3.0
11	*Sitophilus granarius*	3.0-5.0
12	*Sitophilus oryzae*	2.3-3.5
13	*Anthrenus verbasci*	1.7-3.5
14	*Attagenus unicolor*	3.0-5.0
15	*Dermestes maculatus*	5.5-10.0
16	*Dermestes lardarius*	7.0-9.0
17	*Trogoderma glabrum*	2.0-4.0
18	*Trogoderma granarium*	1.4-2.3
19	*Trogoderma inclusum*	2.0-2.5
20	*Trogoderma variabile*	2.6-4.6
21	*Trogoderma versicolor*	2.0-5.0
22	*Lophocateres pusillus*	2.6-3.2
23	*Typhaea stercorea*	2.2-3.0
24	*Carpophilus dimidiatus*	1.6-1.8

November-December 2019
ISSN: 0193-4120 Page No. 2820 - 2825

25	*Carpophilus hemipterus*	1.8-2.1
26	*Ptinus ocellus*	3.5-4.0
27	*Ahasverus advena*	1.82-4.0
28	*Cathartus quadricollis*	2.1-3.5
29	*Oryzaephilus mercator*	2.23.1
30	*Oryzaephilus surinamensis*	2.0-3.0
31	*Alphitobius laevigatus*	5.0-6.5
32	*Gnatocerus cornutus*	3.5-4.5
33	*Latheticus oryzae*	2.5-3.0
34	*Palorus subdepressus*	2.72.8
35	*Tenebrio molitor*	12.0-18.0
36	*Tribolium castaneum*	2.6-4.4
37	*Tribolium confusum*	2.6-4.4
38	*Tenebroides mauritanicus*	5.0-11.0

C. Average Silhouette Indices

In the evaluation for both of the clustering method, Average Silhouette index (Si) was measured. This index used closeness of data of morphometric structure belonging to a cluster. Then it compared with other species dataset according to the nearest group. The value range from -1 to 1. Si value near 1, it indicates that the object is assigned to the proper cluster. Si value near 0 it is unclear which cluster it belongs. If the Si value near to -1 the object is misclassified [21], [22].

IV. CLUSTERING RESULTS

For the K-Means clustering method, it generates five clusters. Compare to HACA it generated 11 cluster. The clustering result has been presented in Table 2 and Table 3.

TABLE 2: K-Means result

Cluster	% of Morphometric dataset in each cluster			
	100-75	75-50	50-25	25-0
Cluster 1	17 species (*A. advena, A. verbasci, C. dimidiatus, C. quadricollis, C. cellaris, L. serricorne, L. oryzae, L. pusillus, O. surinamensis, P. subdepressus, R. dominica, S. paniceum, T. inclusum, T. stercorea, Z. subfasciatus, S. oryzae, T. granarium*)	3 species (*C. hemipterus, T. confusum, T. castaneum*)	6 species (*D. minutus, T. versicolor, T. glabrum, T. variabile, P. ocellus, S. granarius*)	6 species (*A. obtectus, A. fasciculatus, G. cornutus, N. rufipes, P. truncatus*)
Cluster 2	6 species (*A. verbasci, G. cornutus, P. truncates, A. fasciculatus, A. obtectus, N. rufipes*)	6 species (*A. laevigatus, S. granaries, P. ocellus, T. variabile, T. glabrum, T. versicolor, D. minutus*)	4 species (*T. castaneum, T. confusum, C. hemipterus, T. mauritanicus*)	3 species (*D. maculatus, T. granarium, S. oryzae*)
Cluster 3	3 species (*D. lardarius, D. maculatus,*	0	1 species (*A. laevigatus*)	2 species (*N. rufipes, A.*

	T. mauritanicus)			*fasciculatus)*
Cluster 4	1 species *(O. mercator)*	0	0	0
Cluster 5	1 species (*T. molitor*)	0	0	0
Total Species	28	9	11	11

TABLE 3: HACA result

Cluster	% of Morphometric dataset in each cluster			
	100-75	75-50	50-25	25-0
Cluster 1	16 species *(A. verbasci, T. granarium, S. paniceum, P. subdepressus, L. serricorne, L. oryzae, R. dominica, O. surinamensis, Z. subfasciatus, C. quadricollis, C. cellaris, T. inclusum, A. advena, L. pusillus, C. dimidiatus, T. stercorea)*	2 species *(C. hemipterus, S. oryzae)*	6 species *(T. castaneum, T. variabile, T. versicolor, T. confusum, D. minutus, T. glabrum)*	1 species *(S. granarius)*
Cluster 2	10 species *(A. fasciculatus, D. minutus, P. truncates, A. obtectus, N. rufipes, P. ocellus, A. laevigatus, G. cornutus, S. granarius, A. verbasci)*	5 species *(T. variabile, T. versicolor, T. castaneum, T. glabrum, T. confusum)*	4 species *(C. hemipterus, T. mauritanicus, S. oryzae, D. maculatus)*	2 species *(T. granarium, S. paniceum)*
Cluster 3	2 species *(D. lardarius, D. maculatus)*	1 species *(T. mauritanicus)*	0	0
Cluster 4, 5,6,7, 8,9 10	0	0	0	1 species *(O. mercator)*
Cluster 11	1 species *(T. molitor)*	0	0	0
Total Species	27	8	10	4

The result is based on the clustering of 100 morphometric variation dataset from 38 species that give the total of 38,000 datasets. The process involves 38,000 datasets that been measured based on their dissimilarity and been assigned to a cluster. The specific cluster of the dataset that belongs to the specific species the better. The clustering result for the dataset was summarized into four categories that were 100 to 75%, 75 to 50%, 50 to 25% and 25 to 0%. In the example, if 100 to 75% it showed that more than 75% dataset of specific specif belongs to that cluster.

For K-Means, Cluster 1 for 100 to 75% species dataset show 17 species being group in this cluster. This is the highest number of species compared to other clusters. In total 28 species based on the 100 to 75% dataset been group in this five cluster. This can be seen in **Error! Reference source not found.**.

For HACA, Cluster 1 for 100 to 75% species dataset show 16 species being group in this cluster. This is the highest number of species compared to another 10 clusters. In total, 29 species based on the 100 to 75% dataset been group in this 11 cluster. This can be seen in Table 3.

For clustering evaluation, HACA show better in average Si index 0.84 compare to K-Means 0.68. This show HACA assigned the dataset into cluster more proper compared to K-means [21]. This is shown by its HACA average Si Index that higher and closer to 1 compare to K-Means (Figure 1).

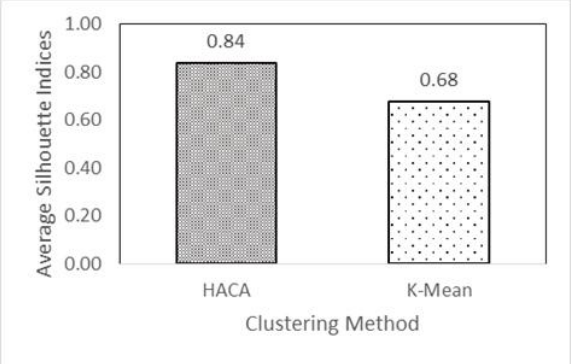

FIGURE 1: Average silhouette indices for HACA and K-Mean

Based on the result from Table 2 and Table 3 many species had been a cluster in the same group. This then leads to difficulty in identification of Coleopteran stored product pest. Thus, the clustering method cannot solely be used in the identification process [23]. The two clustering method also show different clustering result that and cluster evaluation shows HACA perform better in grouping the data in a cluster.

CONCLUDING REMARKS

Clustering technique is beneficial in understanding the relationship between species in the way they can understand their similarity in morphological characteristics. There are two methods implemented which are K-Means clustering and Hierarchical Agglomerative Cluster Analysis (HACA). K-Means clustering is used in segmenting the 38 Coleopteran stored product pest species. Overall, cluster analysis results in other animals the majority of Coleopteran store product pest species been a cluster in a similar group, especially in the first cluster. The segmentation is weak. Therefore, the findings show that it is inadequate to use only the cluster analysis technique in the step of identification. Other multivariate techniques are required to discriminate further and identify the Coleopteran stored product pest using morphometric data such as discriminant analysis.

ACKNOWLEDGEMENT

The authors gratefully acknowledge Government of Malaysia for the sponsorships, Universiti Teknologi MARA and Universiti Utara Malaysia for the facilities. Special thanks go to the anonymous referees for their constructive comments and suggestions.

REFERENCES

1. Hagstrum & Subramanyam. Fundamentals of Stored-Product Entomology David. (AACC Internationla, 2006).
2. Hagstrum, D., Klejdysz, T., Subramanyam, B. & Nawrot, J. Atlas of Stored-Product Insects and Mites. (AACC International, 2017).
3. Rees, D. Insects of Stored Products. (CSIRO, 2004).
4. Alston, D. G. Important Components of a Successful Pest Management Program. Utah Pest Fact Sheet 1–3 (2011).
5. Austen, G. E., Bindemann, M., Griffiths, R. A. & Roberts, D. L. Species identification by experts and non-experts: comparing images from field guides. Sci. Rep. 6, (2016).
6. IAOM Food Protection Committee. IPM Manual Integrated Pest Management. (2016).
7. Daley, Howell, V. Insect morphometrics. Annu. Rev. Entomol. 30, 415–38 (1985).
8. Elphick, C. S. How you count counts: The importance of methods research in applied ecology. J. Appl. Ecol. 45, 1313–1320 (2008).
9. Shea, C. P., Peterson, J. T., Wisniewski, J. M. & Johnson, N. A. Misidentification of freshwater mussel species (Bivalvia:Unionidae): contributing factors, management implications, and potential solutions. J. North Am. Benthol. Soc. 30, 446–458 (2011).
10. Baker, J. E. Properties of glycosidases from the maize weevil, Sitophilus zeamais. Insect Biochem. 21, 615–621 (1991).
11. Campbell, J. F. Influence of Seed Size on Exploitation by the Rice Weevil , Sitophilus oryzae. 15, 429–445 (2002).
12. Germinara, Cristofaro, D. & Rotundo. Behavioral responses of adult Sitophilus granarius to individual cereal volatiles. J. Chem. Ecol. 34, 523–529 (2008).
13. Germinara, G. S., Rotundo, G. & De Cristofaro, A. Repellence and fumigant toxicity of propionic acid against adults of Sitophilus granarius (L.) and S. oryzae (L.). J. Stored Prod. Res. 43, 229–233 (2007).
14. Dent, D. Insect pest management. (CABI Publishing, 2000). doi:10.1079/9780851993409.0000

November-December 2019
ISSN: 0193-4120 Page No. 2732 - 2742

15. Abramoff, M. D., Magalhaes, P. J. & Ram, S. J. Image Processing with ImageJ. Biophotonics Int. 11, 36–42 (2004).
16. Conner, J. K., Cooper, I. A., La Rosa, R. J., Pérez, S. G. & Royer, A. M. Patterns of phenotypic correlations among morphological traits across plants and animals. Philos. Trans. R. Soc. B Biol. Sci. 369, (2014).
17. James Manoharan, J. & Hari Ganesh, S. Initialization of optimized K-means centroids using divide-and-conquer method. ARPN J. Eng. Appl. Sci. 11, 1086–1091 (2016).
18. Almeida, J. A. S., Barbosa, L. M. S., Pais, A. A. C. C. & Formosinho, S. J. Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. Chemom. Intell. Lab. Syst. 87, 208–217 (2007).
19. Bock, H. H. Probabilistic models in cluster analysis. Comput. Stat. Data Anal. 23, 5–28 (1996).
20. Umar, R. et al. Identification Source of Variation on Regional Impact of Air Quality Pattern Using Chemometric. Aerosol Air Qual. Res. 15, 1545–1558 (2015).
21. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65 (1987).
22. Janssens, F., Zhang, L., Moor, B. De & Glänzel, W. Hybrid clustering for validation and improvement of subject-classification schemes. Inf. Process. Manag. 45, 683–702 (2009).
23. Bouket, A. C. Hierarchical cluster analysis of Criconemoides species (Nematoda: Criconematidae), with a proposal for unknown species identification. Arch. Phytopathol. Plant Prot. 47, 90–105 (2014).

AUTHORS PROFILE

Tun Mohd Firdaus Bin Azis currently a lecturer at Universiti Teknologi MARA Cawangan Perlis. Her Master degrees in Entomology were from Universiti Kebangsaan Malaysia in 2013. He is interested in the study of insect, and Biostatistics.

Shamshuritawati Sharif currently a senior lecturer at Universiti Utara Malaysia. She obtained a PhD in Mathematics at Universiti Teknologi Malaysia. Her Master degrees in Decision Science were from Universiti Utara Malaysia in 2003. Her diploma and Bachelor of Science in Statistics were obtained from MARA Institute of Technology (ITM) in 2000. She is interested in multivariate hypothesis testing, industrial statistics, statistical quality control, network analysis and centrality measure.

Basharoh Abdul Karim completed her Bachelor of Science (Honors) in Mathematics at Universiti Putra Malaysia (UPM) on February 9, 2002. She has started her career as a lecturer in Kolej Poly-Tech MARA Alor Setar since June 2002, and her teaching experience had exceeded 17 years. She had obtained her Master of Science (Data Analysis) from Universiti Utara Malaysia in 2019. Her interest in this profession motivates her to work hard in the teaching, research, publications. Her main expertise is in statistics and business mathematics

Haslina Zakaria completed her Decision Science (Hons) first degree in Universiti Utara Malaysia, 2004. She had started her career in Engineering of Mathematics in 2006, and her teaching experience had exceeded 13 years in Universiti Malaysia. Perlis (UniMAP). She had obtained her Master of Science (Data Analysis) from Universiti Utara Malaysia in 2019. Her interest in this profession motivates her to work hard in the teaching, research, publications and consultancy. Her main expertise is in engineering mathematics and statistics.