Monte Carlo Simulation and Clustering for Customer Segmentation in Business Organization

Andry Alamsyah [1], Bellania Nurriz [2]
School of Economics and Business
Telkom University, Bandung, Indonesia
[1] andrya@telkomuniversity.ac.id, [2] bellanian@student.telkomuniversity.ac.id

Abstract— **Utilizing data for segmentation analysis can bring a streamlined way to get potential insight as of decision making support in a business organization. Using appropriate data analytical technique help the organizations in profiling their customer segments accurately. The result brings an effective marketing strategy. However, there are times in doing data analytic, the organization needs another variable of data where the value is unavailable, for example: customer's income data which mostly hard to collect. By using Monte Carlo simulation, the value of customer's income can be generated and then compared with customer spending to construct customer segmentation model. An unsupervised learning for customer segmentation model using K-Means clustering enables us to see the grouping patterns of customer's income towards their spending. Clusters of the dataset might be interpreted as a group of customers that having a similar character.**

This paper shows us how to generate customer's income data and create data cluster to optimizing customer potential by utilizing data. Furthermore, the result brings us insight into which group of the customer might unserved properly considering their average income with their spending behavior.

Keywords— Data Mining; Monte Carlo Simulation; Clustering; K-Means; Customer Segmentation.

I. INTRODUCTION

Perform business analysis has become a liability of business organization to retain their business. Conduct the analysis by utilizing data resources can help a lot to the organization in create knowledge, find positional opportunities, as of support efficiencies in decision making to affect customer satisfaction. The challenge in data abundance, or known as big data, makes human need to adopt technology in processing data into valuable information. By doing large-scale data analytic, the organization has opportunities for modeling, simulating and optimizing based on a whole set of data, rather than depend on the sample [1].

There are times in processing data into information, the organization needs another variable of data that the value is unavailable. An uncertainty value of data can be generated by using function simulation. A Monte Carlo simulation approach, the simulation can be done to generate value on a stochastic event. Stochastic event explains about chance or probability of the event [2]. In addition, generating data by simulation method can provide us to see the worst to the best-case scenario, where we can see the spectrum of the event. Simulation can give us an overview of the actual case become less cost than doing real analytical [3].

Data simulation results can be used in further analysis depend on organization needs. One of eligible utilization is for support customer segmentation case. There are several analytical methods can be used to determine market or customer segments. Unsupervised segmentation research for a metric character commonly solves by grouping the data by clustering methods in data mining. In grouping, clustering is not based on the defining categories beforehand, but by the level of similarity of an actor with the other actors. Centroid-based techniques in K-Means algorithm implemented as a way of partitioning the clustering method [4]. Customer in the same group or cluster will have similar characteristic.

As a case study, we apply these methods in one of the branches of Telkom Indonesia in Makassar city (TWM), in the need to optimize their marketing strategies for home segment customers. Analyze their customer segments in the city based on customer's income information and their payment behavior, makes company can see the purchasing power of a market. But indeed, the company did not have such customer's income data. The nature of customer's income data is fluctuating by time to time. The tens of thousands number of customers also make difficult to collect the data.

In this paper, we show how to apply Monte Carlo simulation to generated customer's income data for TWM. Furthermore, the simulation result data is used to support customer segmentation analysis using the K-Means clustering. From the patterns formed then the company can identify their

customer segments and see the likely potential customers to be served.

II. Theoritical Background

A. Monte Carlo

Monte Carlo is about something that involves deliberateness of use random numbers in the calculation that has the structure of a stochastic process, where the sequence determined by development conditions of random events. As a figure, if there is an expected value denoted by *G*, then by following a game of chance, and then use a set of numerical estimates as *n*, denoted with *Gn*, the fundamental theorem of Monte Carlo guarantees that (*Gn*) = *G* [5]. The main purposes of the development and study Monte Carlo algorithm are; Monte Carlo Simulation, which used to simulate real-life phenomenon by following the corresponding physical, chemical or biological processes under consideration. The other is Monte Carlo Numerical, which used to solve the deterministic problem by modeling random variables [6].

B. Method of Moment Estimation

This method can be used to estimate population parameters such as mean, variance, median, etc. by equating sample moments to unobserved moments population that has a theoretical equation, and then complete the equation for expected amount [7]. The first step is to determine the condition of the moment properly [8].

C. Maximum Likelihood Estimation

This method can be used to make principal parameter inferences of the probability distribution on a dataset, or also can be used as a parameter data distribution with the most likely to occur in an event distributed identically [9].

D. Data Mining

Discovering useful trends and patterns in the large data set can be performed by data mining process. Data mining has six common functions i.e.; description, estimation, prediction, classification, clustering, and association [10].

E. Clustering

Clustering is the basic idea for the process of finding groups of objects (consumer, business, etc.) with an estimated similarity within objects in the same group, and less similar to objects in another group. There are two major types of basic methods are generally used in clustering as known hierarchical model and centroid model. K-Means algorithm is one of the common clustering process based on centroid model. The measure for the case in the cluster is represented by the mean value. K-Means algorithm also works by dividing the dataset into several clusters that requested [11].

F. Market Segmentation

Market segmentation means dividing the market into well-defined groups, so that market consists of customer groups in the same needs and desires. The main variables for the segmentation of the consumer market can be classified as follows: *Geographic region, city or metro size, density, climate, demographic age, family size, family life cycle, gender, income, occupation, etc*. It might also segment based on a combination of several of the categories mentioned [12].

III. Methodology

There are several of data that must be collected to run the customer's income simulation and clustering for this study. The customer's income simulation support by data from Badan Pusat Statistik (BPS) for 14 sub-district in Makassar city, which are GDRP per capita per month, sub-district non-governmental funds, total population, and the number of the household for each sub-district. Afterward, clustering data support by customer's income data and payment data as the spending value of home segment customers. The workflow of segmentation analysis shown in Fig. 1 below:

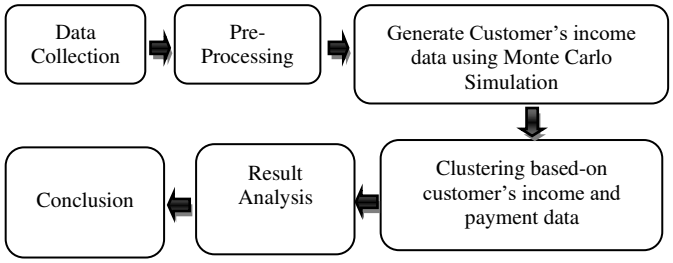


Fig. 1. Research Workflow

The first process is collecting data that use for research. After we get the data, we reduce the noise in the data by conduct pre-processing activity. The third process is generating customer's income data using Monte Caro simulation that consist of several steps which will be explained in the next section. After we have the outcomes of simulation, we can perform clustering process using K-Means algorithm, for customer's income data toward customer's payment. The fifth process is analyzing the insight from simulation and clustering until we find the pattern or model. Finally, we can create the conclusions based on this case study.

IV. Experiment and Result

A. Generate Customer's Income Data

As mention above, customer's income data is generated by Monte Carlo simulation method. The stages of the Monte Carlo simulation in this research is done by below processes;

1) Determining Static and Distribution Input Variable

First, we need a static and distribution model as a variable to performing Monte Carlo simulation [13]. The *mean* $(\bar{x})$ as static variable and standard deviation (σ) as distribution

variable, which set to per capita income for each sub-district in Makassar city are using to generate this simulation. The *mean* $(\bar{x})$ values are produced from BPS data that is processed to obtain an average per capita income value in 14 sub-districts. Standard deviation values are determined in a random manner until we get a number that generates random number which closest to the condition of a minimum per capita income in each district based on the observation. The number of mean and standard deviation of per capita income for each sub-district in Makassar shown in Table I below:

TABLE I. MEAN AND STANDARD DEVIATION OF PER CAPITA INCOME FOR EACH SUB-DISTRICT

Sub-Districts	Average per Capita Income $(\bar{x})$	Standard Deviation (σ)
BIRINGKANAYA	Rp3.308.472	0.7
MANGGALA	Rp6.380.478	1.3
PANAKKUKANG	Rp8.357.546	1.9
BONTOALA	Rp2.119.381	0.4
WAJO	Rp5.599.512	1.3
UJUNG PANDANG	Rp9.601.037	2.2
MAKASSAR	Rp3.248.984	0.7
RAPPOCINI	Rp5.315.408	1.1
TAMALATE	Rp2.331.208	0.5
MAMAJANG	Rp3.060.287	0.6
MARISO	Rp6.450.864	1.5
UJUNG TANAH	Rp5.579.343	1.3
TALLO	Rp5.729.780	1.3
TAMALANREA	Rp9.431.498	2.2

2) Generating the Random Number of Per Capita Income

The second step, we generate the random number of per capita income in each sub-district that refer to the value of mean and standard deviation number. Random number generation can be finished with the normal inverse approach, which calculates the inverse of the cumulative normal distribution function, for a given number of probability, mean and standard deviation of the set. The random number become a specific value of a variable resulting [14]. The function is later described as the following function:

$$X_n = f(p, \bar{x}_{sub-district}, \sigma_{sub-district}) \quad (1)$$

X_n is a specific simulation value of per capita income variable. $f(p, \bar{x}_{sub-district}, \sigma_{sub-district})$ is the normal inverse function of the cumulative normal distribution function for random probability (p), Average per capita income $(\bar{x}_{sub-district})$, and standard deviation $(\sigma_{sub-district})$.

Random numbers generated for 14 sub-districts in Makassar city by using their respective mean and standard deviation, where iterations of random number simulation are done based on the population number for each sub-district. By the random number result, we can see the dispersion, the minimum and the maximum of per capita income for 14 sub-districts. As a representation of overall, we show the distribution model for random numbers results of per capita income in two districts, at the following Fig. 2;

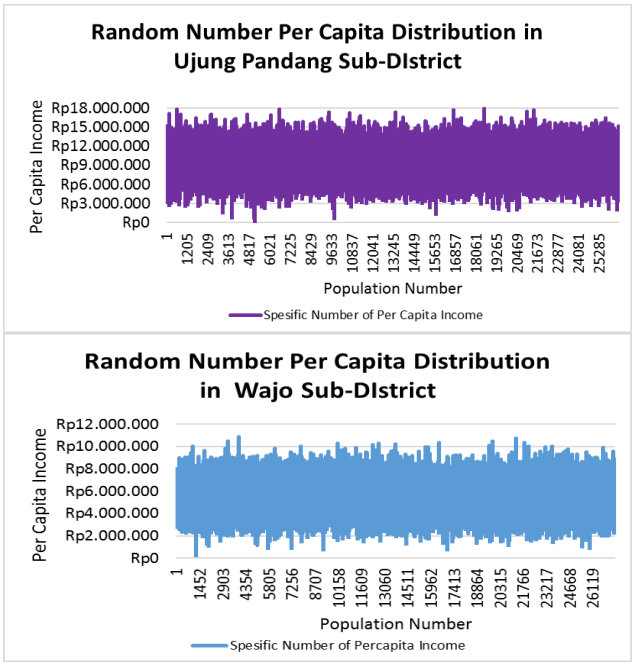


Fig. 2. Distribution of per capita income in Makassar city, (a) Ujung Pandang Sub-district; (b) Wajo Sub-district.

3) Generating Customer's Income Value

The third step is generating customer's income data of home segment customers. Based on currents case study, A home segment customer is a home user, so that the customer income assumed as a household income. To get home segment customer's income we use function of per capita income value (X_n), which are summed up as much as the average population number in a house at each sub-district (q). The function is later described as the following function:

$$X_{customer} = [(X_n)_1 + (X_n)_2 + \ldots + (X_n)_q] \quad (2)$$

Which are $X_{customer}$ is a specific simulation value of per capita income variable, and q is as the average population number in a house at each sub-district. The simulation conducted to every subscriber in each sub-district based on company's payment data. As a representation of overall, we show the distribution model for simulation result of customer's income data in two district, at the following Fig. 3;

(a)

(b)

Fig. 3. Simulation Result of Customer's Income in Makassar city, (a) Ujung Pandang Sub-district; (b) Wajo Sub-district.

The distribution that occurs in simulating for each district above shows the shape of a normal distribution where the result is widely spread nearby the mean value of the data, also be the average customer income according to simulation results. The average result of customer's income data generated for 14 sub-districts are shown in the following Table II:

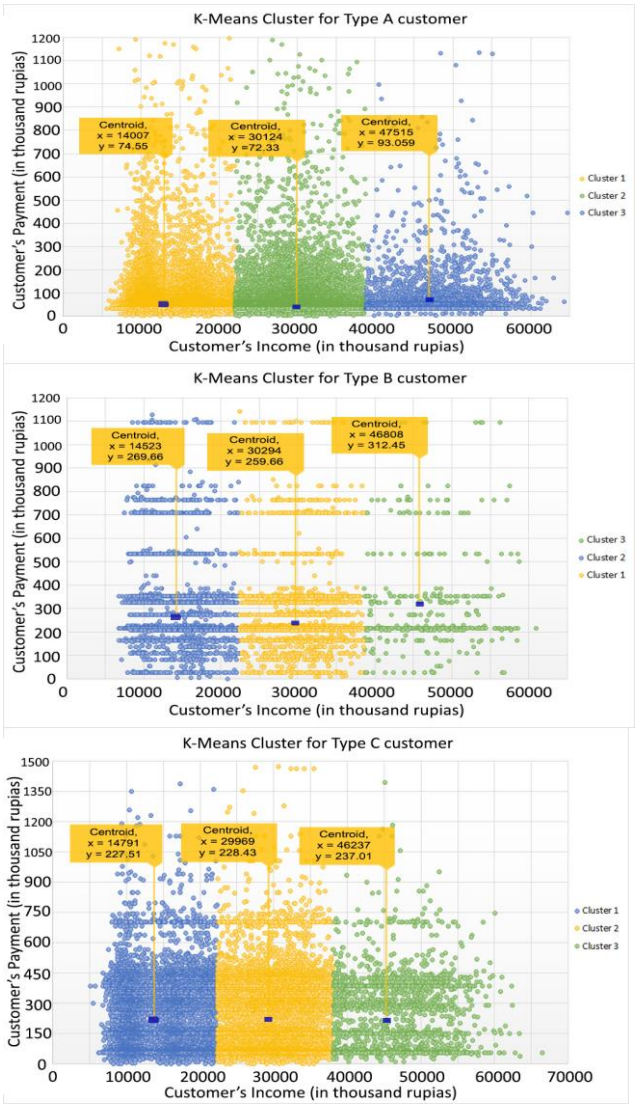
TABLE II. THE AVERAGE RESULT OF CUSTOMER'S INCOME DATA IN 14 SUB-DISTRICT

Sub-districts	Average customer's income
Biringkanaya	Rp16.533.738
Manggala	Rp31.827.076
Panakkukang	Rp33.324.795
Bontoala	Rp10.545.895
Wajo	Rp27.942.371
Ujung Pandang	Rp48.047.476
Makassar	Rp16.250.431
Rappocini	Rp21.242.886
Tamalate	Rp9.278.339
Mamajang	Rp12.223.786
Mariso	Rp32.289.541
Ujung Tanah	Rp11.054.783
Tallo	Rp28.681.484
Tamalanrea	Rp28.316.669

B. *Clustering Customer's Data*

In this research, we conduct K-Means clustering method to analyze customer segmentation. Attributes that used in this clustering are the result of customer's income simulation as variable *x* and customer payments as a variable *y*. The data used are from payment collection data of TWM, has been divided into three main type. Type 'A' with 31127 amount of data, is for the customer that only subscribe to the phone. Type 'B' with 7185 amount of data, is for the customer that subscribe to phone and the internet. Type 'C' with 27652 amount of data, is for the customer that subscribe to the phone, the internet, and interactive television.

The number of optimal cluster for customers type A, B, and C are three clusters, determined from Within-Sum of Square (WSS). Clustering model of payment collection towards customer's income data for each customer types are shown following in Fig. 4:



(a)

(b)

(c)

Fig. 4. K-Means Clustering Plot of (a) Type A Customer, (b) Type B Customer, (c) Type C Customer.

On the Fig. 4, the spread dots are a representation to customers with their income and payment data. Every cluster in a dataset is distinguished into different color. Each cluster also has a centroid cluster point as a mean vector of the cluster [15]. For more detail, the result of clustering of each customer type are shown in Table III below:

TABLE III. K-MEANS CLUSTERING RESULT

Type A			
Cluster	Mean of payment	Mean of Income	size
1	Rp74,55 (thousand)	Rp14,007.41 (thousand)	15415
2	Rp72,34 (thousand)	Rp30,124.10 (thousand)	13317
3	Rp93,06 (thousand)	Rp47,515.47 (thousand)	2395
Type B			
Cluster	Mean of payment	Mean of Income	size
1	Rp269,6882 (thousand)	Rp14.523,88 (thousand)	3579
2	Rp259,6638 (thousand)	Rp30.294,27 (thousand)	3180
3	Rp312,4554 (thousand)	Rp46.808,34 (thousand)	426
Type C			
Cluster	Mean of payment	Mean of Income	size
1	Rp227,5132 (thousand)	Rp14.791,11 (thousand)	12494
2	Rp228,4382 (thousand)	Rp29.969,09 (thousand)	13066
3	Rp237,0105 (thousand)	Rp46.237,09 (thousand)	2092

V. DISCUSSION AND ANALYSIS

By generating customer's income data using Monte Carlo simulation, we get the likely result of customer's income for Organization. We can figure out the average difference income of the customer, such as the low average are in the Bontoala sub-district (Rp10.545.895), also overview of the potential customers who earn quite large, as we can see in table II for Mariso (Rp32.289.541), Panakkukang (Rp33.324.795) and Ujung Pandang (Rp48.047.476). The validity of simulation might very dependent on the accuracy or appropriateness of their input matrix.

A set of customer's income data from simulation result also useful for seeking customer segmentation. By conduct the simulation, the organization has opportunity to get the picture of their customer's income, that was initially unavailable. As we done above for segmentation analysis, we compare the income data towards customer's payment.

Furthermore, we get several patterns by clustering customer data. Commonly, we get two main patterns formed of grouping which are based on the customer's income level and based on customer's spending level. We categorize customer's income level into low income, middle income, and high income. Another, in every cluster, has a crowded area where data congregate towards their payment vector. We categorize customer's spending level into the slight user, moderate user, and heavy user. The frequency of customers about their income level and spending level are shown in table IV and V below:

TABLE IV. INCOME LEVEL FREQUENCY

Category	Income Range	Cluster	frequency
Low income	<Rp23 million	Cluster 1 type A	47.7%
		Cluster 1 type B	
		Cluster 1 type C	
Middle income	Rp23 million – Rp38 million	Cluster 2 type A	44.8%
		Cluster 2 type B	
		Cluster 2 type C	
High income	>RP38 million	Cluster 3 type A	7.5%
		Cluster 3 type B	
		Cluster 3 type C	

TABLE V. SPENDING LEVEL FREQUENCY

Category	Spending range	Cluster	frequency
Slight user	Rp0 – Rp150 thousand	Cluster 1 type A	47.1%
		Cluster 2 type A	
		Cluster 3 type A	
Moderate user	Rp0 – Rp400 thousand	Cluster 1 type B	10.1%
		Cluster 2 type B	
		Cluster 3 type B	
Heavy user	Rp0 – Rp600 thousand	Cluster 1 type C	42.8 %
		Cluster 2 type C	
		Cluster 3 type C	

Likewise, from the two table above, we can see if there are several groups of customers in the cluster, which have middle to high income, but they only consume the product with a small payment. Customer in cluster 2 and cluster 3 for type 'A' only as a slight user, while they are in middle and high-income level of customer. Also for customer in cluster 3 type 'B' that only categorize as moderate user, while they have high income. The information can bring the organization to arrange appropriate marketing strategy and serve their customer more precisely.

The methodology we proposed can evaluate customer segmentation from customer's income simulation and also utilize customer's payment data. The insight can assist the organization in identify customers correctly.

CONCLUSION

In a business organization, we frequently found incomplete data when doing data analytics activities. It is difficult to avoid. However, we can manage to overcome this by simulating value around normal assumption. We get the approximation value, by choosing a method that ensures the value of generated data is not far different from the real condition. Monte Carlo simulation is one of the methods that ensure the conditions of approximation value is fulfilled.

The challenge in performing Monte Carlo simulation is on determining the input value as the variable for simulation, for example the mean and standard deviation value. They must be adjusted to the range value we want to generate, and sometimes we have to construct it by empirically based on our external knowledge.

The other hand, one of the most important for the organization in developing their business is to commit market analysis. The analysis can start by analyzing market segmentation. A method that helps in partitioning customer groups based on similarity can conduct by clustering methodology.

The patterns formed by clustering help organization in identify their customers more specifically. We create clustering model by utilize the data from the Customer's income simulation result and compare the data with the customer's payment. This clustering model beneficial for segmentation analysis, which ultimately bring insight to a potential customer's profile.

For the future research, we suggest using Monte Carlo simulation to generate more complex data such as in the case of association model, classification model or in other scenario unavailable data situation.

REFERENCES

[1] R. Schulte. "Real-Time Analytics: Six Steps for Fast, Precise Decision-Making," www.forbes.com/sites/gartnergroup/2016/06/08/real-time-analytics-six-steps-for-fast-precise-decision-making/#fea05667ae83, June 8, 2016 [Feb. 17, 2017].

[2] Stokastik. (n.d) In *Kamus Besar Bahasa Indonesia Online*. Retrieved from kbbi.web.id/stokastik [Feb. 15, 2017].

[3] F, Lateef. (2010, Oct-Dec.). "Journal of Emergencies, Trauma, and Shock." *Simulation-Based Learning: Just Like the Real Thing*. [Online]. 3(4), pp. 348–352. Available: www.ncbi.nlm.nih.gov/pmc/articles/ [Feb. 17, 2017]

[4] J. Han, M. Kamber, J. Pie. *Data Mining Concepts and Techniques, 3rd ed*.. USA: Morgan Kaufman, 2012, pp. 451.

[5] M. H. Kalos, P. A. Whitlock. *Monte Carlo Methods*. German: WILEY-VCH Verlag GmbH & Co. KGaA, 2004, pp 2 -89.

[6] I. T. Dimov. *Monte Carlo Methods for Applied Scientist*. Singapore: World Scientific Publishing Co. Pte. Ltd, 2008, pp. 4.

[7] A. C. Cohen, B. J. Whitten. (1988). *Parameter Estimation in Reliability and Life Span Models.* [online]. Vol.96. Available: books.google.co.id [february 21, 2017].

[8] D. Harris, L Matyas. "Introduction to the Generalized Method of Moments Estimation" in *Generalized Method of Moments Estimation.* L. Matyas. New York: Cambridge University Press, 1999, pp. 4.

[9] A. M. Law, W. D. Kelton. "Review of Basic Probability" in *Simulation Modeling and Analysis, 2nd ed.* Singapore: McGraw Hill, 1991, pp. 282

[10] D. T. Larose, C. D. Larose. "An Intriduction to Data Mining" in *Discovering Knowledge in Data: An Introduction to Data Mining, 2nd ed.* New Jersey: John Wiley & Sons Inc, 2014, pp. 2-13.

[11] F, Provost. T. Fawcett. Tom. *Data Science for Business.* USA: O'Reilly Media Inc, 2013, pp. 163-170.

[12] P. Kotler, K. L. Keller. *Marketing Management, 15th ed*. USA: Pearson Eductation Inc, 2016, pp 268-269.

[13] S. Rauchaudhuri. "Introduction to Monte Carlo Simulation", in *Proc. Winter Simulation Conference*, 2008, pp. 92-93.

[14] W. L. Winston. *Microsoft Excel 2013: Data Analysis and Business Modeling.* USA: O'Reilly Media Inc, 2014, pp. 709.

[15] B. S. Everitt, S. Landau, M. Leese, D. Stahl. *Cluster Analysis, 5th ed.* John Wiley & Sons Ltd, 2011, pp. 76.