Journal of
Information Systems Engineering
and Business Intelligence

Vol.5, No.2, October 2019
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Clustering of Drug Sampling Data to Determine Drug Distribution Patterns with K-Means Method : Study on Central Kalimantan Province, Indonesia

Wahyuri[1)] [*], Umi Athiyah[2)], Ira Puspitasari[3)], Yunita Nita[4)]

[1)2)4)]*Faculty of Pharmacy, Universitas Airlangga, Indonesia*
Nanizar Zaman Joenoes Building, UNAIR – C Campus, Jl. Mulyorejo, Surabaya, 60115
[1)]wahyuriwahyuri28@gmail.com [2)]umi-a@ff.unair.ac.id,[4)]yunita-n@ff.unair.ac.id

[1)]*Balai Besar Pengawas Obat dan Makanan Di Palangka Raya*
Jl. Cilik Riwut no. 13 Km. 3,5, Palangka Raya, 73112
[1)]wahyuriwahyuri28@gmail.com

[3)]*Information System Study Program, Faculty of Science and Technology, Universitas Airlangga, Indonesia*
UNAIR – C Campus, Jl.Mulyorejo, Surabaya, 60115
[3)]ira-p@fst.unair.ac.id

Article history:

Received 10 August 2019
Revised 16 October 2019
Accepted 16 October 2019
Available online 28 October 2019

Keywords:

Clustering, CRISP-DM, Data Mining, Drug distribution patterns, Drug quality control, Drug sampling

Abstract

Background: Drug sampling and testing in the context of post-marketing control is an important component to ensure drug safety in the supply chains. The results are used by the Indonesian National Agency for Drug and Food Control (NA-FDC) for conducting public warnings, evaluating the Good Manufacturing Practice (GMP) and Good Distribution Practice (GDP) implementation, and enforcing the law against drug violation.
Objective: This study aimed to identify and analyze drug distribution patterns to provide an overview of drug sampling in the public sector.
Methods: The data was collected from Balai Besar Pengawas Obat dan Makanan (BBPOM) Palangka Raya's database. The collected data were the drug sampling data from Integrated Information Reporting Systems (IIRS) application from 2014 to 2018. Next, we employed CRISP-DM methodology to analyze the data and to identify the pattern. K-means clustering model was selected for data modeling.
Results: The dataset contained five attributes, i.e., drug name, therapeutic classes, district/city, sample category, and evaluation of drug surveillance. The drug distribution pattern formed three clusters. First cluster contained 522 drug items in eight therapeutic classes and spread over ten districts, second cluster contained 1542 drug items in five therapeutic classes and spread over five districts, and third cluster contained 503 drug items in eleven therapeutic classes and spread across nine districts.
Conclusion: To conclude, the applied data mining technique has improved the decision on the drug sampling planning. It also provides in-depth information on the improvement of drug post-marketing control performance in Central Kalimantan Province.

I. INTRODUCTION

One of the government's responsibility in providing health services to the community is to ensure drug safety, efficacy and quality in the market. To ensure the quality of drugs in circulation, countries in the world have implemented various regulations/policies for drug control starting from ensuring the quality of drugs through pre-market evaluations to maintaining patient safety for drugs consumed through post-market evaluation. Activities undertaken include: drug registration and appraisal, inspection of drug production and distribution facilities, drug sampling and testing, supervision of drug advertisements and supervision of drug clinical trials [1].

[*] Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print)
doi: http://dx.doi.org/10.20473/jisebi.5.2.208-218

Drug regulatory authority in countries around the world implement drug sampling and testing as part of the Post Marketing Surveillance program. Drug sampling and testing is an important component to ensure drug safety in the supply chains. Many aspects affect its implementation, including the concept of sampling and testing, the geographical characteristics of the surveillance area, the profile of the drug in circulation, and the specific purpose of sampling. In Indonesia, Post Marketing Surveillance is carried out by National Agency of Drug and Food Control (NA-FDC). The results are used by NA-FDC for conducting public warning, evaluating the Good Manufacturing Practice (GMP) and Good Distribution Practice (GDP) implementation and enforcing the law against drug violation [1].

Central Kalimantan is the 3rd largest province in Indonesia after Papua and East Kalimantan with the area of one and a half times of the island of Java, which reaches 153,564 km² and consists of 13 districts and 1 city. With a position in the middle of Kalimantan island, Central Kalimantan Province was created as an interconnection with other areas in the Kalimantan island. With this area, the opportunity to grow new trade locations is increasingly open. This condition has caused the volume of drug in the Central Kalimantan province to increase. On the other hand, substandard drug, counterfeit drug, and drug that contain hazardous substances have the potential to be more easily entered among the people in its area [2].

Implementation of drug sampling in Central Kalimantan Province is carried out by Balai Besar POM (BBPOM) in Palangka Raya. The number of drug samples as many as 500 samples per year are taken from various distribution facilities including Pharmaceutical Wholesalers, District/City Pharmacy Warehouses, government and private hospitals, health centers, pharmacies, and drug stores spreading throughout Central Kalimantan Province. Drug sampling is carried out by following the Sampling Priority Guidelines established by NA-FDC [3]. In Central Kalimantan Province, the distance and reach to the surveillance area is quite difficult. This difficulty is also experienced by officers in carrying out drug sampling in the field. The existence of a mismatch between the sampling implementation plan with the conditions in the field causes the implementation of sampling to be ineffective and inefficient.

Based on the problem above, BBPOM in Palangka Raya is looking for ways to manage and support the decision making process in managing strategies and drug sampling plans to be right on target. Several ways that can be done are knowledge discovery and pattern recognition related to the drug sampling process and its entities in the data owned by BBPOM in Palangka Raya through the Integrated Internal Reporting System (IIRS).

IIRS application is a webbase programming with database using SQL server so that it can be accessed in real time by employees in BBPOM in Palangka Raya. Since being implemented in 2014, IIRS has produced very large amounts of data. Drug sampling data contained fairly complex information about 23 attributes, including sample code, drug name, composition, registration number, batch number, expiration date, manufacturer, therapeutic class, sampling method, sampling criteria, sampling location, number of samples, packaging condition, price and labelling evaluation result. The tipe of data were polynominal (text), nominal (text) and numerik (date_time). Data stored in a database will increasingly accumulate over time but has not been utilized in such a way as to explore the potential of existing information in helping to improve the quality of drug sampling planning and implementation. Knowledge discovery and pattern recognition on drug sampling data contained in the IIRS database will produce a drug distribution pattern that contains information about the location of the therapeutic classes distribution, sampling sites, sampling criteria, and sampling methods refer to drug sampling guidelines according to the Central Kalimantan Province characteristics.

The process of knowledge discovery and pattern recognition can be done using Data Mining method [4][5]. Many studies have been conducted in health sector. Data mining has been able to provide many contributions to the decision-making system for health services [8][13][15][18][19], to find out the pattern of a disease [11], to find the cause of the spread of a disease [14], to predict or early diagnose of a disease [6][12], to be used in finding great information on the costs of care and treatment of patients who have certain diseases [9][16] and can provide alternative recommendations in treatment [7][10][17]. However, study on knowledge discovery in drug surveillance data has never been done.

CRISP-DM methodology is the most widely used in data mining projects, so this methodology is used as a de facto standard in data mining implementation [20]. Some data in health sectors with various categories of variables, clustering techniques with k-means algorithm is one technique that is widely used today [19]. In the context of drug sampling and testing, data mining method uses clustering technique with the k-means algorithm. This technique can be used as a start in understanding and exploring new knowledge in drug sampling data. Clustering functions to group data based on the similarity of data in one group and minimize similarity to other groups [4].

Clustering on drug sampling data needs to be done as an initial understanding in exploring new knowledge about drug as an object in the framework of drug quality control. Clustering identifies the patterns of relationships between attributes, for example between cities/districts with therapeutic classes. This study aimed to determine the drug

distribution pattern in Central Kalimantan Province based on drug sampling data and clustering technique. The drug distribution pattern produced to provide a drug sampling profile in the public sector and according to the characteristics of the Central Kalimantan province. The determination of drug distribution pattern is expected to contribute to the implementation sampling plan and better testing in the future which have an impact on improving the performance in monitoring the quality and safety of post-marketing drug.

II. METHODS

The population of this study was the data of drugs sampled and tested by BBPOM in Palangka Raya which circulated in Central Kalimantan province. The samples were data of drug sampling and testing contained in IIRS database from 2014 to 2018. Process standard of data mining used in this research was Cross Industry Standard Process for Data Mining (CRISP-DM) methodology (Fig. 1).

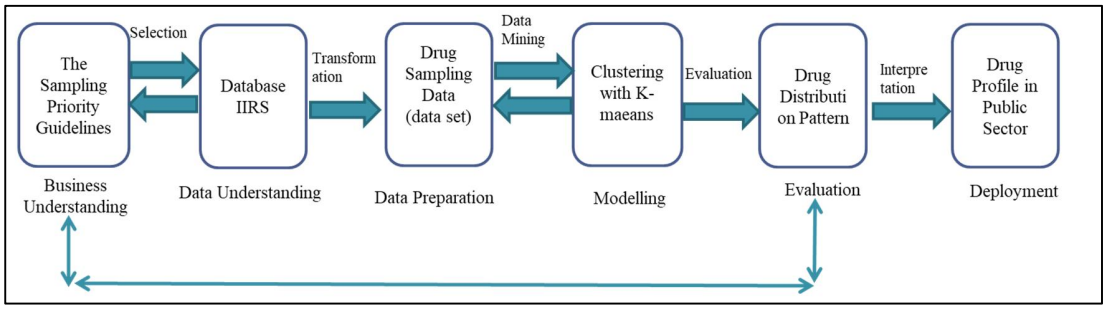

Fig. 1 CRISP-DM Process Schema in Drug Sampling Data

The first step in the CRISP-DM methodology was understanding the substance of data mining which was done as well as identifying needs from the perspective of the business process organization [20]. Business understanding referred to the NA-FDC sampling priority guidelines 2018. At this stage, an understanding of the background, goal, substance, and spirit of the post-marketing drug surveillance program was needed.

Data Understanding was the stage of collecting initial data and studying the data so that it could be known and understood of what could be done on the data [20]. Data understanding referred to drug sampling database contained in IIRS. The database used in this study was drug sampling data in the Central Kalimantan province from 2014 to 2018 that were 2567 records.

Data preparation included all activities to build a dataset that would be included in the modeling tool from the initial raw data to create a new database for data mining set up. Several relevant attributes were selected to formed new attributes for modeling as shown in Table 1.

TABLE 1
ATTRIBUTE FOR MODELLING STAGE

No	Field Name	Data Type	Notes
1.	Sample Name	Polynominal (Text)	Drug sample brand name
2.	Therapeutic Classes	Polynominal (Text)	Classification of drugs based on pharmacological effects following the National Formulary nomenclature
3.	District/City	Polynominal (Text)	Name of district/city in Central Kalimantan province
4.	Sample Category	Polynominal (Text)	Categories of drug samples based on the NA-FDC sampling priority guideline: non-NHI, NHI Upstream, NHI Downstream
5.	Evaluation of Drug Surveillance	Polynominal (Text)	Evaluation of drug samples referring to the list of drug evaluation contained in the NA-FDC sampling priority guideline

The next activity was transforming data on new attribute. Data transformation was the process of changing nominal/polynominal data types (text) on an attribute into the form of integer data number (number/numerical). This was needed to do the clustering model. Data transformation was done by sorting the attribute from the largest to the smallest number of items. The attributes that need to be initialized were the Therapeutic Class, District/City, Sample Category and Evaluation of Drug Surveillance which generated the data set like the example in Table 2.

TABLE 2
SAMPLE DATASET FOR MODELLING

Sample	Therapeutic Class	District/City	Sample Category	Evaluation of Drug Surveillance
1	2	1	1	1
2	2	6	2	2
3	2	8	2	2
4	2	1	1	1
5	2	4	2	1
6	7	11	2	3
7	7	2	1	3

After doing data preparation, then the next step was the process of data mining (modelling phase). Data mining modelling was done using clustering technique with k-means algorithm. The selection of modeling technique and algorithm was done using Rapidminer Version 8.2 software. K-means algorithm was a clustering algorithm that is included in the partition method. K-means was the oldest clustering algorithm and the most widely used in various applications because of its ease of implementation. K-means works by distributing a number of objects into a predetermined cluster, where each cluster center is represented by the average value of the objects that exist in each cluster [23]. The K-means process flow as seen in Fig. 2 below.

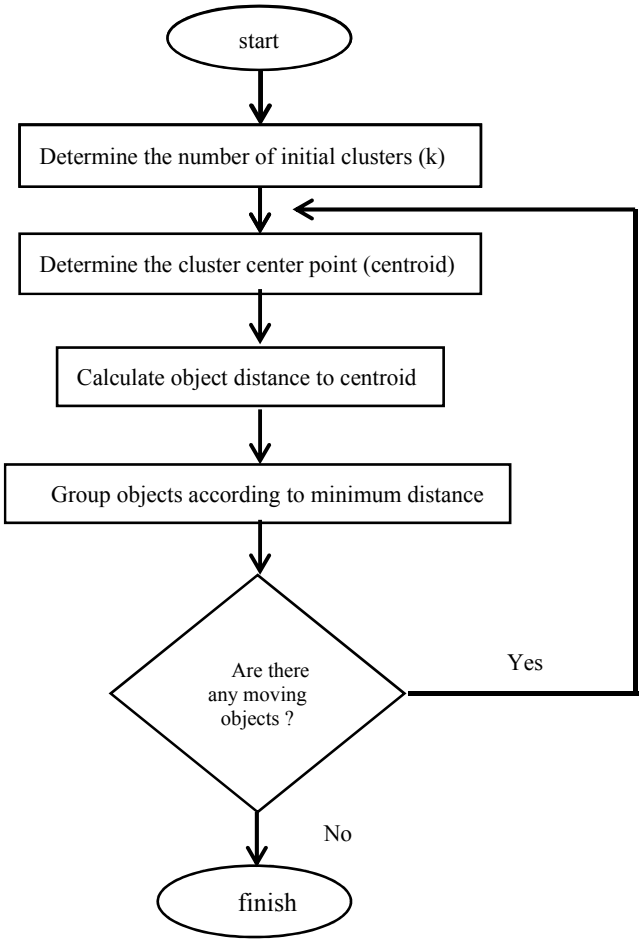

Fig. 2 Process Flow K-means Algorithm

One measurement used to evaluate the quality cluster algorithm is by Davies-Bouldin Index (DB Index) which is introduced by David L. Davies and Donald W. Bouldin in 1979. The DB index is an internal evaluation scheme,

where validation is done to measure how well the grouping has been done using the quantity and features inherent in the dataset. This DB index measures between each cluster that is most similar to a cluster group, the smaller the DB index value of the cluster similarity the better [22].

Clustering visualization was shown through a graph of the relationship between the cluster plot against each parameter. In each parameter, a cluster with the smallest mean was obtained, then a reference was made as the main priority and a cluster with the largest mean was a low priority. Whereas the other clusters were given medium priority. With Rapidminer, clustering visualization results were analyzed using scatter diagrams [21].

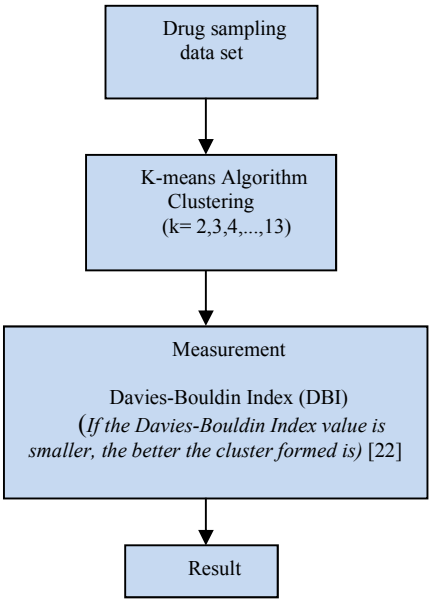

Fig. 3 Computational Framework

III. RESULTS

From the k-means clustering model as shown in Figure 3 above, then it was tested on various k values to obtain the most optimal cluster with Davies Bouldin Index Value and Average Within Centroid Distance parameters, with the result shown in Table 3.

TABLE 3
K-VALUE TO DETERMINE THE BEST CLUSTER NUMBER

k Value	DB index value	Average Within Centroid Distance Value
2	0,982	11,599
3	0,786	6,595
4	0,898	5,376
5	0,923	4,677
6	0,898	3,875
7	1,024	3,584
8	1,031	3,138
9	1,022	2,926
11	1,125	2,750
12	1,190	2,590
13	∞	2,495

The best Davies Bouldin (DB) Index Value was the smallest DB index value obtained in the cluster amount of k=3, then the next model was selected with a k value=3, with the cluster result model as shown in Table 4.

TABLE 4
RESULT OF K-MEANS CLUSTER MODELLING

Cluster Model	
Cluster_0 (red)	522 instances (20,33%)
Cluster_1 (blue)	1542 instances (60,07%)
Cluster_2 (green)	503 instances (19,60%)
Total number of instances	2567 instances

Table 5 showed the calculation of the distance between cluster and centroid (the cluster center point). The cluster distance with centroid (the cluster center point) showed the level of data population proximity with the closest cluster.

TABLE 5
CLUSTER DISTANCE TO CENTROID

Attribute	*Cluster*_0	*Cluster*_1	*Cluster*_2
Therapeutic Class	8.057	2.221	2.658
Sample Category	1.542	1.337	1.656
District/City	2.900	2.090	8.847
Drug Surveillance Evaluation	1.728	1.597	1.698

Three clusters formed had characters based on therapeutic class, the location of the sample which was the district/city from which the sample was obtained, the sample category, and the drug surveillance evaluation based on BPOM drug sampling priority guideline. The patterns formed were illustrated below:

1) Therapeutic Classes Cluster, dividing 12 (twelve) classes of therapy into 3 clusters (cluster_1, cluster_2, and cluster_0). Fig. 4 showed that more upward, the smaller the number of drugs that was sampled.

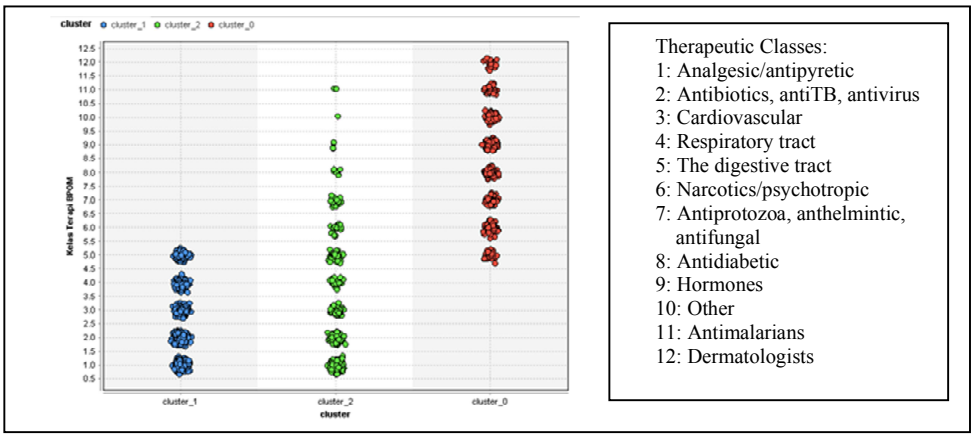

Fig. 4 Result of Therapeutic Classes Clustering

2) District/City Cluster, dividing 14 (fourteen) districts/city into 3 clusters (cluster_1, cluster_2, and cluster_0). Fig. 5 showed that the more upward, the smaller the number of drugs that was sampled.

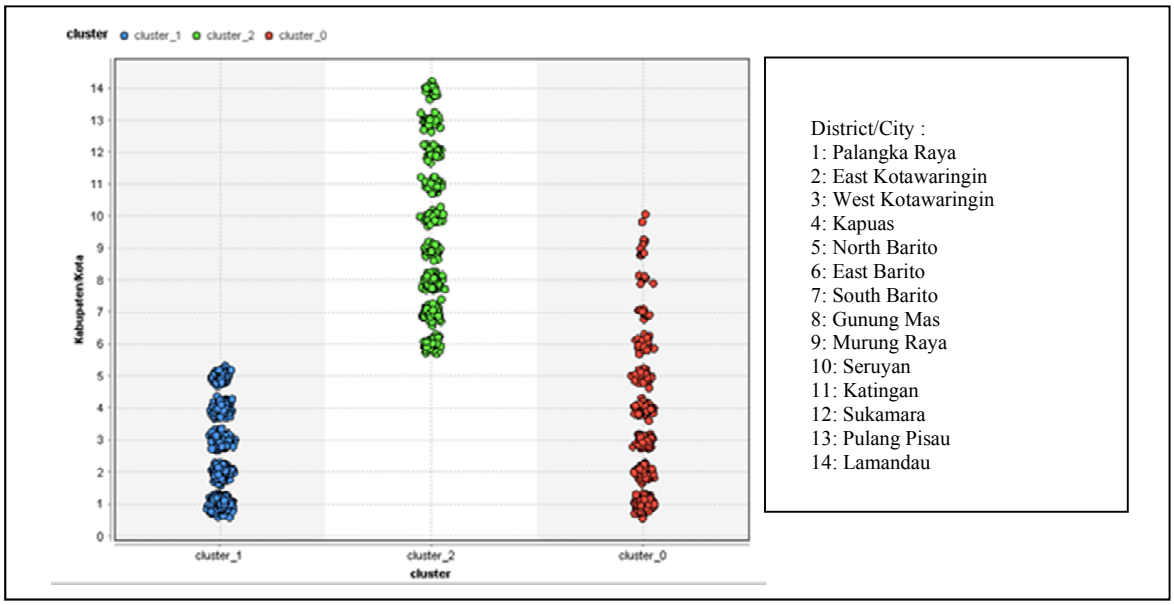

Fig 5. Result of District/City Clustering

3) *Sample Category Cluster*, dividing 3 (three) sample categories into 3 clusters (cluster_1, cluster_2, and cluster_0). Fig. 6 showed which drug sampled on the cluster_1 and the cluster _0 with the number of high Non-National Health Insurance (NHI) drugs, moderate NHI Downstream, and low NHI Upstream. Whereas at the cluster _2, it showed the number of high Non-NHI drugs, moderate NHI Upstream, and low NHI Downstream.

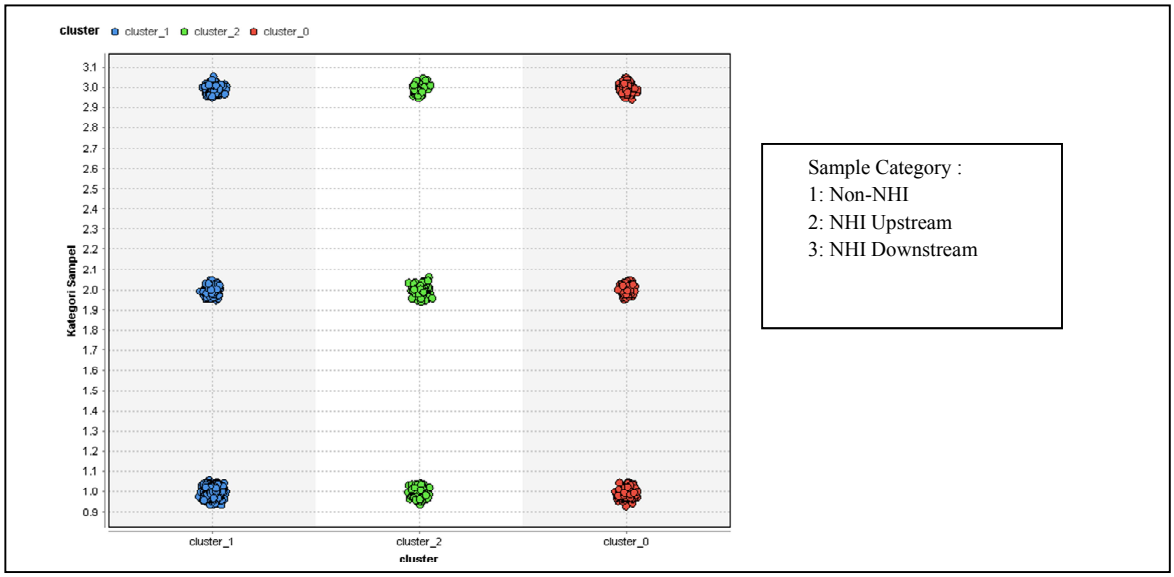

Fig 6. Result of Sample Category Clustering

4) *Evaluation of Drug Surveillance Cluster*, dividing 5 (five) categories into 3 clusters (cluster_1, cluster_2, and cluster_0). Fig. 7 showed that the more upward, the smaller the number of drugs that was sampled and on evaluation of drug surveillance number 5 was not found on cluster_0.

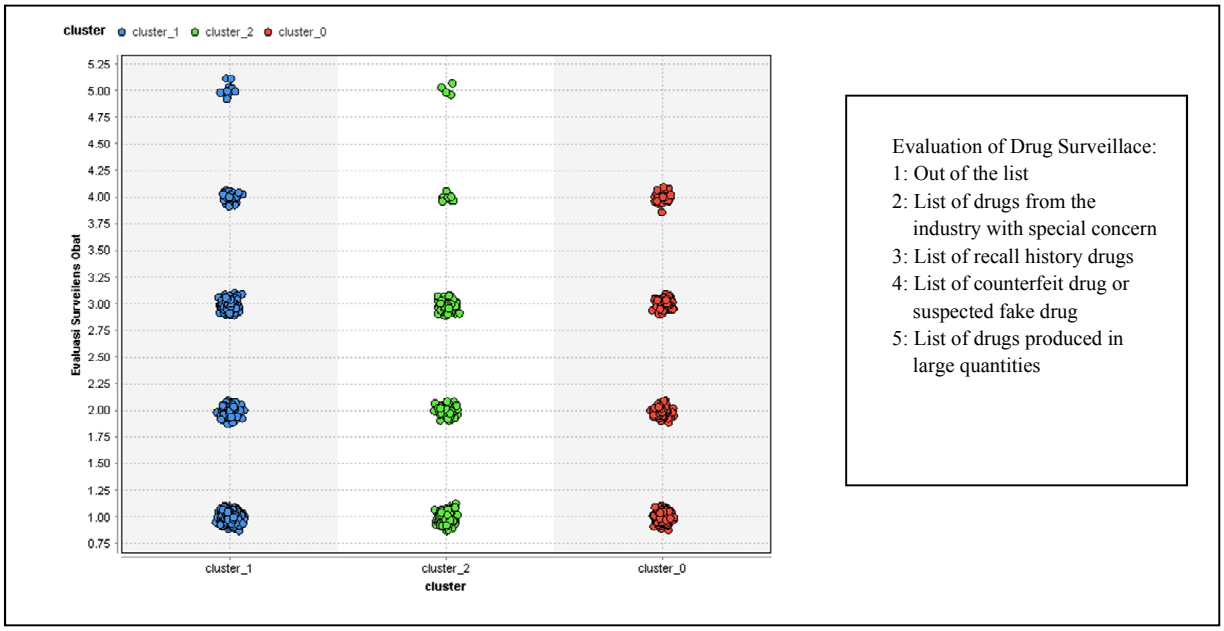

Fig. 7 Result of Evaluation of Drug Surveillace Clustering

IV. DISCUSSION

A. Utilization of Drug Distribution Pattern in Drug Sampling Planning in Central Kalimantan Province

Drug sampling and testing in the post-marketing control activity was one of the drug quality assurance programs.

This was an important part of the regulatory system in all countries and may serve as a source of information about the quality of drugs in circulation. Drug sampling was costly. Limitations on resources may restrict the number of samples collected, parameters tested, techniques to be used for analysis or the number of staff available to conduct the sampling and testing. Therefore it was important to optimize the use of resources by focusing on those drugs and parameters that pose a higher risk to patients and apply risk analysis during sampling planning [24]. Observing the high challenge of drug control related to various types and specific characteristics that circulate in each region including Central Kalimantan, the drug sampling was carried out based on a risk analysis [25].

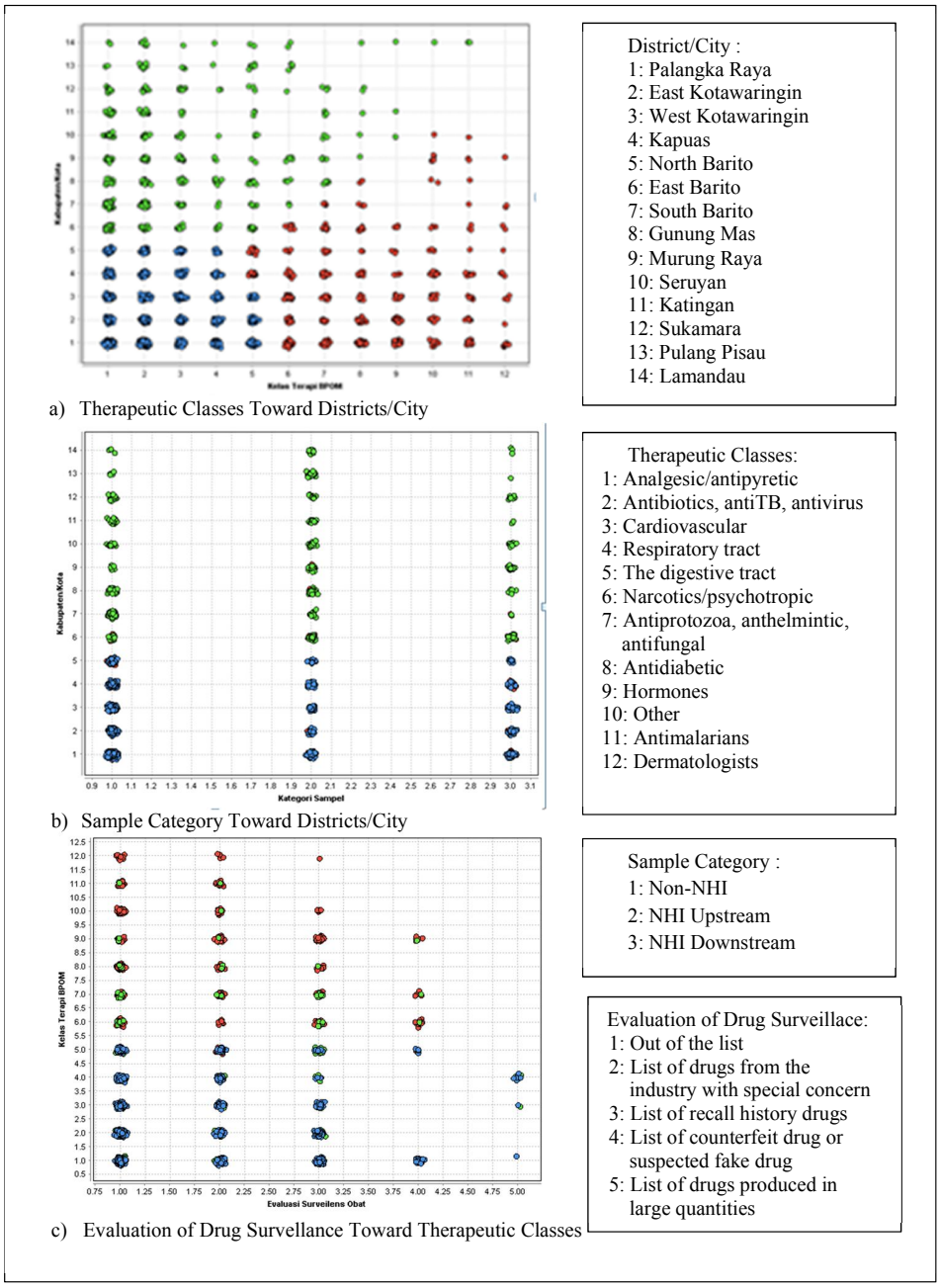

Fig.8 The Drug Distribution Patterns in Central Kalimantan Province

Drug sampling that had been carried out was based on the sampling implementation plan which was made every month. In the practice, the target sample set in the sampling implementation plan often could not be fulfilled because the intended drug sample was not available. This was quite difficult for officers because they had to move to other

sampling sites and did more searching for drugs listed in the sampling plan, so it took a very long time and required a hard effort from the officers.

In districts that were very far away and the number of drug distribution facilities was small (such as Sukamara, Lamandau, Seruyan, and Murung Raya districts), officers experienced difficulties in carrying out drug sampling. Due to the limited number of facilities, officers did not have many choices of drug types or therapeutic classes that were eligible to be taken as samples. Therefore, the determination of the sampling plan at the beginning would greatly affect the success of the sampling implementation in the field.

Data mining results in the form of drug distribution pattern could be used as a reference to make drug sampling target when taking drug samples in the field with due regard to the cluster characteristics formed as in Fig. 8 below. Distribution Pattern of Therapeutic Classes Toward Districts/City showed that analgesic/antipyretic, antibiotic, anti-TB, antivirus, and cardiovascular drugs had a distribution in all districts/city in Central Kalimantan Province; In Central Kalimantan Province there were 6 (six) districts/city that had drug distribution in all therapeutic classes, namely: Palangka Raya, East Kotawaringin, West Kotawaringin, Kapuas, North Barito, and East Barito.

For example, the North Barito District in Fig. 8 was known to be included in cluster_1 and cluster_2. If antidiabetic drugs wanted to be sampled, the characteristics of cluster_2 were needed to be paid attention. Sampling focus can be done at pharmacies using a random sampling method for non-NHI drugs and a targeted sampling method for NHI drugs at the District Pharmacy Warehouse. Also, sampling focus could be directed at antidiabetic drugs produced by industries that were included in the special attention list.

Another example was sampling implementation in the Sukamara district. Sukamara district was the furthest district in Central Kalimantan province, 660 km from the city of Palangka Raya with an 18 hours journey by land [24]. Based on the cluster formed, Sukamara district could be included in the cluster_2. Based on Fig. 8, sampling implementation of this district could be focused on 8 therapeutic classes, namely: antibiotics-antituberculosis-antivirus drugs, cardiovascular drugs, respiratory drugs, analgesic/antipyretic drugs, narcotic/psychotropic drugs, antidiabetic drugs, gastrointestinal drugs, and anti protozoa-anthelmintic-antifungal drugs. The sampling method that could be done was random sampling in pharmacies and drug stores on non-NHI drugs and targeted sampling on District Pharmacy Warehouse. Sampling focus could also be directed at drug surveillance evaluation cluster, that was, drugs produced by industries that were included in the list of specific special attention to respiratory and antidiabetic drugs and list of drugs with recall history to analgesic/antipyretic drugs, gastrointestinal drugs, antiprotozoal-anthelmintic-antifungal drugs.

By paying attention to the drug distribution patterns that were formed, drug sampling planning could be done with more directed and focused on the drugs that were in the distribution in accordance with the patterns that were formed in each district/city.

B. Limitations and Further Development

The limitation of this research was that the clustering technique done was limited to one type of algorithm, which was k-means. Tomar [19] stated that there is no single data mining technique that can provide a consistent result on various types of data in the health sector. Therefore, there are still opportunities for extracting patterns with other algorithms so that performance can be compared between algorithms.

The drug sampling database was also dynamic in nature, so the drug distribution pattern formed in the data mining process was ad hoc. The drug sampling database dynamics would potentially give a different pattern from time to time. This could affect decision making in sampling planning in the future, both in determining the proportion of the number of samples in each therapeutic class, setting the target sampling for NHI drugs, and determining the category of drugs included in the drug surveillance evaluation. Therefore, data mining on drug sampling data should be carried out on an ongoing basis with additional exploration also on other relevant data sources, for example, the latest drug profile circulating, the result of the latest drug control, the result of third-party drug testing, and the report on drug use in the district/city.

For further development, the clustering technique application with several algorithms needs to be done so that the best algorithm performance can be obtained and further exploration of drug control data using other data mining methods, for example, estimation, forecasting, classification, and association.

Then, the drug distribution pattern obtained in this research needs to be applied in real drug sampling planning so that the success of its application in improving the quality of post-marketing drug control planning can be measured. Exploration and implementation of data mining in other types of data are also very possible to do, for example in the database of traditional medicines, cosmetics, and food.

V. CONCLUSIONS

In this study, the CRISP-DM methodology can be used as a framework for conducting the data mining method on drug surveillance data and K-means clustering technique implementation on the drug sampling data can find drug distribution patterns in Central Kalimantan province. The drug distribution pattern that is formed can be used as a reference to set up the drug sampling target when taking drug samples in the field with due regard to the characteristic of the cluster which is formed.

This is very useful for increasing the effectiveness of drug sampling activity in the context of drug post-marketing control. Sampling data characterization result can be used as reference material in making a sampling implementation plan in the district/city, designing the focus of NHI drugs monitoring, and is useful in knowing the distribution of drugs in certain cases by analyzing the cluster formed in the drug surveillance evaluation as an early warning reference.

REFERENCES

[1] BPOM, 2016. 2016 Annual Report, Jakarta, pp. 1-10.
[2] BPOM in Palangka Raya (2014). Strategic Plan for 2014-2019 , Palangka Raya, pp. 1-32.
[3] BPOM, 2018. Guidelines for Priority Sampling in 2018, Jakarta, pp. 1- 180.
[4] Han J, Kamber M (2012). Data Mining: Concepts and Techniques , Urbana-Champaign, Third Edition, University of Illinois, pp. 1-35.
[5] Larose DT, Larose CD, 2014. Discovering Knowledge in Data , Wiley, pp.1-15.
[6] Braga A, Portela F, Santis MF, Belha A, Machado J, Silva A, Rua F, (2016). Data Mining to Predict The Use of Vasopressors in Intensive Medicines Patients, Journal of Technology 78: 6-7.
[7] Chen TJ, Chou LF, Hwang SJ (2003). Application of a Data Mining Technique to Analyze Coprecription Patterns for Antacids in Taiwan, Elsevier, Clinical Therapeutics, Vol. 25, Issue 9: 2453-2463.
[8] Duggirala HJ, Tonning JM, Sith E, Bright RA, Baker JD, Ball R, Bell C, Bright-Ponte SJ, Botsis T, Bouri K, Boyer M, Burkhart K, Condrey GS, Chen JJ, Chirtel S, Filice RW, Francis H, Jiang H, Levine J, Martin D, Oladipo T, O'Neill R, Palmer LAM, Paredes A, Rochester G, Sholtes D, SzarfmanA, Wong HL, Xu Z, Koss-Hout T (2016). Use of Data Mining at the Food and Drug Administration, Journal of the American Medical Informatics Association 23 (2): 428-434.
[9] Erawati S, Mustafa K, Lazuardi L (2016). Pattern of Diabetes Mellitus Inpatient Cost Component Grouping in Hospitals , Journal of Information Systems for Public Health, Volume 1, April 2016: 25-31.
[10] Ibrahim H, Saad A, Abdo A, Eldin S (2016). Mining Association Patterns of Drug-Interactions Using Post Marketing FDA's Spontaneous Reporting Data, Journal of Biomedical Informatics 60 Elsevier: 294-308.
[11] Ilayaraja M, Meyyapan T (2013). Mining Medical Data to Identify Frequent Diseases Using Apriori Algorithm, International Conference in Pattern Recognition Informatics and Mobile Engineering, India: 194-199.
[12] Jen CH, Wang CC, Jiang BC, Chu YH, Chen MS (2012). Application of Classification Techniques on Development of an Early-Warning System for Chronic Illnesses, Expert Systems with Applications 39 (10). Elsevier: 8852–8858.
[13] Koh HC, Tan G (2011). Data Mining Applications in Healthcare, Journal of Healthcare Information Management 19 (2): 65
[14] Moon SS, Kang SY, Jikpitaklert W, Kim SB (2012). Decision Tree Models For Characterizing Smoking Patterns of Older Adults, Elsevier, Expert System With Application, Vol. 39, Issue 1, January 2012: 445-451
[15] Ranjan J (2006). Applications of Mining Data Techniques in Pharmaceutical Industry, Journal of Theoretical and Applied Information Technology : 61-67.
[16] Razali AM, Ali S (2009). Generating Treatment Plan in Medicine: A Data Mining Approach, American Journal of Applied Science, ed. 6: 345-351.
[17] Reddy CK, Aggarwal CC (2015). Healthcare Data Analytics, CRC Press, Wayne State University in Detroit, Michigan, USA, pp. 1-15.
[18] Stuhlinger W, Hogl O, Muller M (2000). Intelligent Data Mining For Medical Quality Management, Workshop on Noter of The 14th European Conference Artificial Intelligence: 1-10.
[19] Tomar D, Agarwal S (2013). A Survey on Data Mining Approaches for Healthcare, International Journal of Bio-Science and Bio Technology, Vol. 5 No. 5: 241-266
[20] Wirth R (2000). CRISP-DM: Towards a Standard Process Model for Data Mining, Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, no.24959: 29-39.
[21] Hofmann M, Klinkenberg R, 2014. Rapid Miner: Data Mining Use Cases and Business Analytics Applications, CRC Press, Boca Raton, FL, pp. 19-29, 162.
[22] Davies DL, Bouldin, DW, 1979. A Cluster Separation Measure, IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2), pp. 224-227.
[23] Suyanto, 2017. Data Mining For Data Classification and Clustering, Penerbit Informatika, pp 247-262.
[24] WHO (2016). Annex 7: Guidelines on The Conduct of Surveys of The Quality of Medicines, WHO Expert Committee on Specifications for Pharmaceutical Preparations, Fiftieth report, WHO Technical Report Series No. 996, Geneva.
[25] BPOM in Palangka Raya (2017). 2017 Performance Report , Palangka Raya, pp. 1-16.
[26] BPOM in Palangka Raya (2016). 2016 Performance Report , Palangka Raya, pp. 1-18.

Supplementary

The new attributes formed from Table 1 above are obtained from:

1) **Therapeutic Classes**, classified based on risk analysis compiled in the NA-FDC sampling priority guideline. The name of the therapeutic classes follow the nomenclature in the National Formulary.

2) **Sample Category** based on the NA-FDC sampling priority guideline for the National Health Insurance (NHI) Program, divided into:
   a. non-NHI: drugs that were sampled from drugstore and pharmacies with random sampling method.
   b. NHI Upstream: drugs that were sampled from Pharmaceutical Wholesalers and District/City Pharmacy Warehouses with purposive targeted sampling method.
   c. NHI Downstream: drugs that were sampled from health centers and government hospitals with purposive targeted sampling method.
3) **Drug Surveillance Evaluation** obtained by evaluating drugs samples by referring to the drug evaluation list contained in the NA-FDC sampling priority guideline consisting of:
   a. List of counterfeit drugs evaluated base on Appendix 4 in the NA-FDC sampling priority guideline.
   b. List of recall history drugs evaluated base on Appendix 5 in the NA-FDC sampling priority guideline.
   c. List of drugs produced by the pharmaceutical industry with special concerns evaluated base on Appendix 6 in the NA-FDC sampling priority guideline.
   d. List of drugs produced in large quantities evaluated base on Appendix 7 in the NA-FDC sampling priority guideline.