Application of Data Mining Using the K-Means Algorithm in Rural and Urban Land and Building Tax (PBB-P2) Receivables Data in Bantul Regency

H D Aprilia¹ and D Agustiani²

¹² Regional Financial and Asset Agency of Bantul Regency, DI Yogyakarta, Indonesia

¹herwinadian@gmail.com, ²dian.agustiani@gmail.com

Abstract. The rural and urban land and building tax (PBB-P2) receivable amount in Bantul is quite large. As of December 31, 2019, there were 3,344,145 PBB-P2 objects, worth for IDR 114,984,991,600. This number tends to increase from year to year, showing that PBB-P2 receivable collection process has not optimal. This study discusses data mining applications for data management using the K-Means clustering method. This paper uses the PBB-P2 existing receivable data, given by the Directorate General of Taxation before PBB-P2 becomes local taxes. The data is between the years of 1994 and 2012. Data is grouped based on the village area and the category of the receivables number. We cluster it into three types, namely the high, medium, and low accounts receivable clusters. Data mining is expected to improve the PBB-P2 receivable data management in the Bantul Regency to make better decision-making. This study's results make it easier to analyze PBB-P2 receivable data pattern based on the grouping of village areas and the receivable number category. The analysis results are expected to provide input for BKAD Bantul to identify certain villages and categories of receivables that need more attention in the PBB-P2 collecting process in the Bantul Regency.

1. Introduction

Rural and Urban Land and Building Tax (PBB-P2) is a tax on land and buildings owned, controlled, and/or utilized by private persons or entities, except for areas used for plantation, forestry, and mining business activities [1]. According to Indonesian law number 28, the year of 2009 (UU 28/2009), on Regional Taxes and Regional Retributions, local governments are given full authority to manage local taxes. One of the significant changes regulated by Law 28/2009 is the PBB-P2 tax management delegation as the local tax and its management become the local Governments authority [2].

As PBB-P2 became one of the local taxes, there were also receivables from the Directorate General of Taxes to the Local Government. Following the accrual-based Government Accounting Standards, PBB-P2 receivables occur when the regional collection rights arise, namely at the time of issuance of SPPT, SKP or STP. If taxpayers have not paid the SPPT/SKP/STP, the PBB receivable will continue to add to the regional government's financial statements.

In 2013, the Central Government gave Bantul Government the PBB-P2 management authority. Since then, Bantul Government has the autonomy to collect and manage it. It is managed by the Bantul Regency Regional Finance and Asset Agency (BKAD Bantul). Not only the authority, but BKAD Bantul also received the transfer of PBB-P2's receivables, amounting to IDR 58,144,699,240 for the 1994 - 2012 tax assessment. The amount of PBB-P2 receivables in the Bantul Regency tends to increase from year to year. As of December 31, 2019, PBB-P2 receivables were recorded at IDR 114,984,991,600.
for 3,344,145 tax objects. This number tends to increase yearly, showing that the collecting management process of PBB-P2 receivables has not been optimal in Bantul Regency.

Managing receivables data is an integral part of the tax collection process. The Directorate General of Taxes divides PBB-P2 receivables management into three sub-activities: administering, collecting, and eliminating the PBB-P2 receivables [2]. The balance grows every year, and the local governments need to do some actions for the increased. Neither to collect nor to eliminate it. Good data analysis is necessary to make the right decision. But, currently, the PBB-P2 Receivables data managed by the BKAD Bantul has not been used for analysis.

Data mining and clustering are some of the methods that can be used to analyze receivables data. We group the data into several groups based on specific criteria. There was research on data mining to process tax data. Jauhari & Mardiani (2013) use data mining to cluster motorized vehicle tax data [3]. This study discusses the application of data mining to cluster PBB-P2 receivable data in the Bantul Regency. We classify data into 3 clusters: high, medium, and low receivables. We use the amount (1) and count (2) of the receivable tax. As a result, there will be some clusters created. There will be villages clustered with various receivables amount and count: high, medium, and low. The local government can use this information as a reference to plan some activities, whether to collect the receivables or eliminate it. The provided information is expected to use as a recommendation for the BKAD Bantul to prioritize villages that need more attention in the receivable managing process.

2. Literature Review

2.1. Data Mining
Data mining is a pattern-finding process of an extensive data set association stored in a database. It uses pattern recognition technology or statistical and mathematical techniques. [4] Larose classifies Data Mining divided into six groups: description, estimation, prediction, classification, clustering, and association [4]

2.2. Clustering
One of data mining methods is data clustering [5]. Garcia-Molina et al. [6] stated that clustering means classifying extensive data into smaller groups according to each essential things in common. Data clustering includes records grouping, observing –or paying attention, and forming objects class’ that have similarities. A cluster is a collection of records similar to one another and not similar to other clusters.

2.3. K-Means Algorithm
There are several steps in K-Means Cluster [7] Analysis algorithm: determine the number of clusters (1), allocate objects into clusters randomly (2), calculate the sample centroid in each cluster (3), allocate each object to the nearest centroid (4). We must go back to step three. if there are objects that moved to another cluster or are still having changes in centroid value or some are above the specified threshold value or the value change in the objective function used is above the threshold value.

3. Research Methods
Data used in this research was obtained from Badan Keuangan dan Aset Daerah of Bantul Regency (BKAD Bantul). The object of this research is Land and Building Tax of Urban and Rural (PBB-P2) Receivables in Bantul Regency. Data sample selected from the year 1994 – 2012 of which the transferred receivable data from Directorate General of Fiscal Balance to the government of Bantul Regency on the PBB-P2 transfer period in 2013.

3.1. K-Means Algorithm processing stages
The method used for clustering PBB-P2 receivable data was K-means algorithm. Data Attributes which were used are village name, receivable amount and number of invoice. The expected output was to
produce 3 clusters namely low receivable cluster (C1), moderate receivable cluster (C2), and high receivable cluster (C3).

Clustering method using K-Means algorithm was carried out in several stages as follows:

1. Generating the smallest value, average value and highest value of each attribute as the centre of the initial cluster (centroid).
2. Calculating the closest distance of the centroid with each data in the cluster using the Euclidean Distance formula:
   \[ d(x, \mu) = \sqrt{(x_1 - \mu_1)^2 + \ldots + (x_i - \mu_i)^2} \] (1)
   With \( d(x, \mu) \) is the distance between the cluster \( x \) and the cluster centre \( \mu \) in the word \( x_i \) is the \( i \)-th weight of the cluster to be searched for distance, \( \mu_i \) the weight of the word to \( i \) at the centre of the cluster.
3. Classifying each data based on its proximity to the centroid (smallest distance).
4. Updating the centroid value. The new centroid value is obtained from the average cluster in question by using the formula:
   \[ C_k = \frac{1}{n} \sum d_i \] (2)
   Where:
   \( n \) : the amount of data in the cluster
   \( d \) : the sum of the value of the incoming distance in each cluster
5. After the new centroid value is obtained, do the second iteration with the same calculation from step 2 to step 5 with the new centroid.
6. Compare each cluster’s members with the previous iteration, if the cluster member composition is still changing, then proceed to the next iteration until the same composition is obtained.

3.2. Implementation of K-means algorithm with Microsoft Excel

Aravind H, C Rajgopal and K P Soman has shown that clustering process using K-Means algorithm can be performed in Microsoft Excel. [8] This paper implements K-means algorithm using Microsoft Excel formulas, with the following steps:

1. Determine the initial centroid using 3 function: MIN function to generate the smallest value, AVERAGE function to generate the average value, and MAX function to generate the highest value
   Centroid1 =MIN(data attribute1, …, data attribute n)
   Centroid2 =AVERAGE(data attribute1, …, data attribute n)
   Centroid3 =MAX (data attribute1, …, data attribute n)
2. Determine the closest distance of the centroid with each data with SQRT function
   Closest Distance (CD) = \[ \sqrt{(x_1 - \mu_1)^2 + \ldots + (x_i - \mu_i)^2} \]
   CD-Cluster1 = SQRT(((data attribute1 - centroid1attribute1)^2) + ((data attribute2 - centroid1attribute2)^2))
   CD-Cluster2 = SQRT(((data attribute1 - centroid2attribute1)^2) + ((data attribute2 - centroid2attribute2)^2))
   CD-Cluster3 = SQRT(((data attribute1 - centroid3attribute1)^2) + ((data attribute2 - centroid3attribute2)^2))
3. Determine the smallest distance of each data to the centroid using IF function
   Smallest Distance (SD) =
   IF(MIN(CD-Cluster1:CD-Cluster3)=CD-Cluster1,"Cluster1", IF(MIN(CD-Cluster1:Cluster3)=CD-Cluster2,"Cluster2","Cluster3"))
4. Calculate the new centroid value using the combination of SUMIF and COUNTIF function
   New Centroid Cluster \( n \)
4. Results and Discussion
The use of sections to divide the text of the paper is optional and left as a decision for the author. Where the author wishes to divide the paper into sections the formatting shown in Table 2 should be used.

4.1. Data Collection
PBB-P2 Receivables Data was collected for 72 villages in Bantul Regency for the year 1994-2012 can be seen in Table 1.

| No | Village Name | Invoice Amount (IDR) | Invoice Number (sheets) | No | Village Name | Invoice Amount (IDR) | Invoice Number (sheets) |
|----|--------------|----------------------|-------------------------|----|--------------|----------------------|-------------------------|
| 1  | Argodadi     | 227,484,070          | 15,938                  | 37 | Potorono     | 1,270,629,509        | 31,795                  |
| 2  | Argomulyo    | 783,994,004          | 32,810                  | 38 | Ringinharjo  | 212,631,595          | 13,217                  |
| 3  | Argorejo     | 603,338,933          | 29,883                  | 39 | Sabdodadi    | 284,878,603          | 13,441                  |
| 4  | Argosari     | 246,835,958          | 16,243                  | 40 | Segoroyoso   | 122,085,652          | 13,240                  |
| 5  | Bangunharjo  | 2,931,075,397        | 37,288                  | 41 | Seloharjo    | 180,615,421          | 24,441                  |
| 6  | Bangunjiwo   | 1,517,889,440        | 87,207                  | 42 | Selopamioro  | 456,176,943          | 63,688                  |
| 7  | Banguntapan  | 5,368,517,564        | 75,448                  | 43 | Sendangssari | 193,111,605          | 17,016                  |
| 8  | Bantul       | 886,421,619          | 38,718                  | 44 | Sidomulyo    | 292,293,695          | 36,662                  |
| 9  | Baturetno    | 1,239,521,266        | 24,461                  | 45 | Singosaren   | 384,106,535          | 6,546                   |
| 10 | Bawuran      | 122,072,967          | 10,874                  | 46 | Sitimulyo    | 1,561,036,318        | 49,187                  |
| 11 | Canden       | 33,346,990           | 2,632                   | 47 | Srigading    | 155,225,250          | 12,542                  |
| 12 | Caturharjo   | 563,292,636          | 50,333                  | 48 | Srihardono   | 162,466,016          | 29,086                  |
| 13 | Donotirto    | 37,538,562           | 975                     | 49 | Sriharjo     | 324,603,009          | 42,455                  |
| 14 | Gadingharjo  | 103,292,353          | 7,718                   | 50 | Srimartani   | 772,210,342          | 50,496                  |
| 15 | Gadingssari  | 36,945,202           | 3,192                   | 51 | Srimulyo     | 527,826,875          | 43,263                  |
| 16 | Gilangharjo  | 513,976,013          | 38,796                  | 52 | Sumberagung  | 275,159,814          | 19,899                  |
| 17 | Girirejo     | 89,756,162           | 7,839                   | 53 | Sumbermulyo  | 399,486,962          | 47,773                  |
| 18 | Guwosari     | 538,719,753          | 35,967                  | 54 | Tamanan      | 895,228,979          | 23,648                  |
| 19 | Imogiri      | 95,840,857           | 3,595                   | 55 | Tamanitiro   | 1,619,524,339        | 53,822                  |
| 20 | Jagalan      | 151,993,018          | 3,060                   | 56 | Temuwuh      | 105,898              | 2                      |
| 21 | Jambidan     | 890,613,282          | 27,924                  | 57 | Terong       | 864                  | 1                      |
| 22 | Karangtalun  | 106,344,605          | 5,251                   | 58 | Timbulharjo  | 1,582,429,770        | 59,962                  |
| 23 | Karangtengah | 58,615,786           | 7,268                   | 59 | Tirtohargo   | 2,166,385            | 104                    |
| 24 | Kebonagung   | 116,472,471          | 11,337                  | 60 | Tirtomulyo   | 8,981,412            | 603                    |
| 25 | Mangunus     | 4,980                | 2                      | 61 | Tirtominolo  | 822,134,307          | 32,069                  |
| 26 | Mulyodadi    | 349,565,997          | 38,221                  | 62 | Tirtosari    | 10,740,717           | 703                    |
| 27 | Murtgading   | 106,596,727          | 8,496                   | 63 | Triharjo     | 714,105,568          | 72,003                  |
| 28 | Ngestiharjo  | 3,443,279,987        | 56,244                  | 64 | Trimulyo     | 553,636,004          | 36,538                  |
| 29 | Palbapang    | 460,652,344          | 32,388                  | 65 | Trimurti     | 400,662,796          | 22,985                  |
| 30 | Panggungharjo| 1,560,264,729        | 42,490                  | 66 | Trienggo     | 1,143,909,937        | 54,471                  |
| 31 | Panjangrejo  | 182,227,384          | 24,606                  | 67 | Triwidiadi   | 26,686,603           | 3,365                   |
| 32 | Parangtitris | 205,741,789          | 7,939                   | 68 | Wijirejo     | 300,153,697          | 24,521                  |
| 33 | Patalan      | 485,109,567          | 38,958                  | 69 | Wirokerten   | 1,745,007,192        | 30,345                  |
| 34 | Pendowoharjo | 929,847,213          | 41,701                  | 70 | Wonokromo    | 477,784,185          | 23,399                  |
| 35 | Pleret       | 412,527,710          | 22,170                  | 71 | Wonolelo     | 9,455,293            | 994                    |
| 36 | Poncosari    | 465,989,666          | 38,718                  | 72 | Wukirsari    | 492,085,508          | 41,927                  |
4.2. Data Normalization

According to Aksoy and Haralick (2001) and Larose (2005) a primary application of geometrical measures (distances) to features having large ranges will implicitly assign greater efforts in the metrics compared to the application with features having smaller ranges. Furthermore, the features need to be dimensionless since the numerical values of the ranges of dimensional features rely upon the units of measurements and, hence, a selection of the units of measurements may significantly alter the outcomes of clustering. Therefore, one should not employ distance measures like the Euclidean distance without having normalization of the data sets [4][9].

Normalizing the data attempts to give all attributes an equal weight. For distance-based methods, normalization helps prevent attributes with initially large ranges from outweighing attributes with initially smaller ranges. [10] Normalization can be done with Min-max normalization formulas [2] as follow:

\[ v'_i = \frac{v_i - \text{min}_A}{\text{max}_A - \text{min}_A} (\text{new}_\text{max}_A - \text{new}_\text{min}_A) + \text{new}_\text{min}_A. \]

\( \text{min}_A \) and \( \text{max}_A \) are the minimum and maximum values of an attribute 
\( v_i \) is the i-th variable 
\( \text{new}_\text{max}_A - \text{new}_\text{min}_A \) are the expected minimum and maximum values of the normalisation 

The result of data normalization can be seen in Table 2

| No | Invoice Amount | Invoice Number | No | Invoice Amount | Invoice Number |
|----|----------------|----------------|----|----------------|----------------|
| 1  | 0.04800        | 0.18277        | 37 | 0.26813        | 0.36461        |
| 2  | 0.16544        | 0.37625        | 38 | 0.04487        | 0.15156        |
| 3  | 0.12732        | 0.34269        | 39 | 0.06012        | 0.15413        |
| 4  | 0.05209        | 0.18626        | 40 | 0.02576        | 0.15183        |
| 5  | 0.61853        | 0.42761        | 41 | 0.03811        | 0.28028        |
| 6  | 0.32031        | 1.00008        | 42 | 0.09626        | 0.73036        |
| 7  | 1.13289        | 0.86523        | 43 | 0.04075        | 0.19513        |
| 8  | 0.18706        | 0.44401        | 44 | 0.06168        | 0.42043        |
| 9  | 0.26157        | 0.28051        | 45 | 0.08106        | 0.07506        |
| 10 | 0.02576        | 0.12469        | 46 | 0.32942        | 0.56407        |
| 11 | 0.00704        | 0.03017        | 47 | 0.03276        | 0.14382        |
| 12 | 0.11887        | 0.57721        | 48 | 0.03428        | 0.33355        |
| 13 | 0.00792        | 0.01117        | 49 | 0.06850        | 0.48686        |
| 14 | 0.02180        | 0.08850        | 50 | 0.16295        | 0.57908        |
| 15 | 0.00780        | 0.03659        | 51 | 0.11138        | 0.49613        |
| 16 | 0.10846        | 0.44490        | 52 | 0.05807        | 0.22819        |
| 17 | 0.01894        | 0.08989        | 53 | 0.08430        | 0.54785        |
| 18 | 0.11368        | 0.41246        | 54 | 0.18891        | 0.27118        |
| 19 | 0.02022        | 0.04122        | 55 | 0.34176        | 0.61722        |
| 20 | 0.03207        | 0.03508        | 56 | 0.00002        | 0.00001        |
| 21 | 0.18794        | 0.32022        | 57 | 0.0018951      | 0.00690        |
| 22 | 0.02244        | 0.06021        | 58 | 0.33393        | 0.68763        |
| 23 | 0.01237        | 0.08334        | 59 | 0.00046        | 0.00118        |
| 24 | 0.02458        | 0.13000        | 60 | 0.00227        | 0.00805        |
| 25 | 0.00000        | 0.00001        | 61 | 0.17349        | 0.36776        |
| 26 | 0.07377        | 0.43831        | 62 | 0.00227        | 0.00805        |
| 27 | 0.02249        | 0.09742        | 63 | 0.15069        | 0.82572        |
| 28 | 0.72662        | 0.64500        | 64 | 0.11683        | 0.41901        |
| 29 | 0.09721        | 0.37141        | 65 | 0.08455        | 0.26358        |
4.3. Implementation of K-Means Algorithm

Determination of the initial centroid value is done by determining the minimum value, the average value, and the maximum value of receivable amount and invoice number. Initial centroid data as can be seen in Table 3.

Table 3. Centroid Initial Data

| Data Cluster                  | Receivable Amount | Invoice Number |
|-------------------------------|-------------------|----------------|
| low receivable cluster (C1)   | 0,000             | 0,000          |
| moderate receivable cluster (C2)| 0,130             | 0,308          |
| high receivable cluster (C3)  | 1,133             | 1,000          |

The centroid then used to calculate the closest distance of each data with the center point of the cluster. The results of calculation in Iteration 1 can be seen in table 4.

After data is in the closest cluster, recalculate the new cluster center based on its average member. Calculations to determine the new cluster center point from the new data is done by summing the values of all cluster members then divided by the total number of cluster members. After the calculation process above, the new centroid will be obtained as can be seen in Table 5. Then do the second iteration with the same calculation as iteration 1. Compare members of each iteration. If the cluster member iteration position is still changing, then proceed to iteration 3. If the cluster member position does not change, then the iteration is stopped.

Based on all data’s grouping results using k-means clustering method, the final grouping results up to the 5th iteration, where the center points no longer changes and there is no data moves between clusters. Cluster member composition for each iteration can be seen in Table 5.

Table 4. New centroid data for Iteration 2 to Iteration 5

| Iteration  | New Centroid Value |
|------------|--------------------|
|            | Receivable Amount  | Invoice Number  |
|            | C1     | C2     | C3     | C1     | C2     | C3     |
| Iteration 2| 0,02   | 0,16   | 0,93   | 0,06   | 0,42   | 0,76   |
| Iteration 3| 0,02   | 0,16   | 0,83   | 0,09   | 0,45   | 0,65   |
| Iteration 4| 0,03   | 0,17   | 0,83   | 0,11   | 0,47   | 0,65   |
| Iteration 5| 0,03   | 0,17   | 0,83   | 0,12   | 0,49   | 0,65   |
Table 5. Cluster member compositions for Iteration 1 to 5

| Iteration | Cluster Member Composition |
|-----------|----------------------------|
|           | C1 | C2 | C3 |
| Iteration 1 | 25 | 45 | 2  |
| Iteration 2 | 30 | 39 | 3  |
| Iteration 3 | 34 | 35 | 3  |
| Iteration 4 | 36 | 33 | 3  |
| Iteration 5 | 36 | 33 | 3  |

5. Conclusion

The clustering process of PBB-P2 receivable data for 72 villages in Bantul Regency over the fiscal year of 1994 to 2012, resulting in three clusters: 36 villages are grouped as the low receivable cluster (C1), 33 villages are grouped into the moderate receivable cluster (C2), and 3 villages are grouped into the high receivable cluster (C3).

From these results, it can be concluded that data mining technique with clustering method can be used as a consideration for the formulation of PBB-P2 receivable management strategies and policies in Bantul Regency.

For accounts receivable whose collection rights have expired, the results of this research can be used as consideration for carrying out receivables write-off. For the process of collecting accounts receivable in the current or future year, the results of this study can provide references for BKAD Bantul to identify villages which have high potential for arrears so that accounts receivable collection policies can be focused to those areas. This research can be further developed by adding more attributes to the data mining process or applying different clustering techniques.

References

[1] Law of The Republic Indonesia Number 28 Year 2009 concerning Local Taxes and Charges (Indonesia)
[2] Ministry of Finance of The Republic Indonesia 2014 Pedoman Umum Pengelolaan Pajak Bumi dan Bangunan Perdesaan dan Perkotaan (Indonesia)
[3] Johari I, Soeyapto D and Mardiani 2015 Penerapan data Mining untuk Data Jumlah Kendaraan Menggunakan Algoritma Expectation Maximization (EM) pada Dispenda Kota Palembang (Indonesia: STMIK MDP)
[4] Larose Daniel T 2005 Discovering Knowledge in Data: An Introduction to Data Mining (New Jersey: John Willey & Sons. Inc)
[5] Gullo F 2015 From patterns in data to knowledge discovery: What data mining can do Phys. Procedia 62 18-22
[6] Garcia-Molina H, Ullman JD and Widom J 2002 Database systems the complete book, International edition (New York: Prentice Hall)
[7] R. P. Kristianto, B. Santoso, and M. W. Sari, “Integration of K-means clustering and naive bayes classification algorithms for smart ac monitoring and control in WSAN,” 2019 4th Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2019, pp. 495–500, 2019.
[8] Aravind H, Raigopal C and Soman KP 2010 A Simple Approach to Clustering in Excel Int. J. of Comp. App. 11 975–8887
[9] Aksoy S and Haralick RM 2001 Feature normalization and likelihood-based similarity measures for image retrieval (Pattern Recogn. Lett., 22: 563-582.)
[10] Han J, Kamber M and Pei J 2012 Data Mining: Concepts and Techniques (Waltham: Morgan Kaufmann)