Fraud Detection of Bus Ticket Sales by Using Spatio Temporal Data Mining

Fajar Delli Wihartiko and Doni Wihartika

Abstract*: *Cheating in the sale of bus tickets is often found in transportation service providers who still use the conductor as a ticket seller on the bus. The high cost of supervision, lack of honesty, unification of the sales function and the ticket control function on the conductor makes this fraudulent practice a problem that companies must handle. By looking at the behavior of ticket sales for each individual through the method of spatio temporal clustering can detect fraudulent behavior that occurs. The bus ticket sales deception process is implemented in Bogor's Bus Rapid Transit (BRT). The results show that there are 3.2% of high-potential officers cheating ticket sales. By knowing the cheating behavior of ticket sales, the company can follow up with the policy so that cheating behavior does not become the company culture.

Index Terms*: *Conservation, Data Mining, C45.

I. INTRODUCTION

Fraud is generally defined by the Association of Certified Fraud Examiners (ACFE) as any attempt to deceive others to benefit. In the ACFE final report of 2012 [1] shows that there are more than 51% of cases of corruption in the survey for the Asian region and 36.1% of cases of corruption found in transportation and workshops. Bogor’s Bus Rapid Transit (BRT) operators were not spared from the various potential fraudulent actions that may be made company employees. One of the biggest potentials of fraud is the ticketing system where Bogor's BRT currently uses the conventional system by using the conductor officer as a ticket seller.

The conventional system implemented by Bogor's BRT means that bus managers must be able to overcome the possible fraud by the conductor. The weakness of this system is due to the unification of sales and ticket control functions charged to the conductor. One solution to the problem is to hire honest employees. However, the value of honesty of employees will continue to decrease if the company is not able to monitor and crack down on employees who are proven to commit fraud. Supervision for all the conductor in charge of each bus fleet and all places of departure will be able to increase the company's operational costs. So the cheating detection of ticket sales conductor is an important issue to be developed by the company before the company is able to implement e-ticketing system.

Fraud detection is an act of identifying fraud as quickly as possible, both systematically and potentially within an organization, through an ever-evolving method of coping with possible frauds [2], [3]. This cheating detection has been applied in various disciplines such as credit card fraud detection, telecommunications and money laundering issues [4]. Research on fraud detection in the field of transportation, especially on bus ticket sales is a rarely discussed theme. This is because bus rapid transit services held in various cities in the world have used smart cards (e-ticketing) to support passenger transaction activities [5] and no longer use conductor to sell tickets.

Bogor's BRT operators currently have no rules / guidelines to detect ticket fraud. The detection activity is only limited to monitor the smallest ticket sales for each month which then the result of detection is used for employee evaluation. This technique cannot provide accurate information about possible employee fraud because it does not take into consideration the employee holiday aspect, the tracks served and the amount of trips the employee receives. Therefore this research is expected to help companies to be able to detect and define cheating ticket sales by taking into account the absenteeism aspects of employees, service lines and the number of passengers per trip. This is in line with Goldstein & Seiichi's [2] statement that the overall anti-cheating strategy includes prevention, detection and fraud investigation.

Fraud detection on financial statements has been done by Efstathios Kirkos, et. Al. through the application of data mining [6]. The recent development of fraud detection research is the use of graph mining to detect fraudulent banking transactions [7]. Various techniques can be used to detect cheating and one of the techniques used is through outlier detection [2]. Outlier detection is the process of finding object data with behavior that is very different from expectations. Such an object is called an outlier [8]. Outlier detection and clustering are two related topics. Clustering finds a majority pattern in a set of data and organizes data to a certain size, while outlier detection attempts to capture extraordinary cases that deviate substantially from the majority pattern.

Bogor's BRT ticket cheating detection process, conducted through a study of ticket sales data. The idea of this research is to analyze ticket sales data that is seen as spatio temporal data (Fig. 1). Spatio temporal data by Kisilevich, et. al [9] is data retrieved and indexed according to the dimension of space and time. Passenger / trip and path sales data are analogous to spatial attributes as well as day sales as temporal attributes. Ticket sales data can also be used to determine the optimal number of Bus trips [10]. In this paper the outlier is viewed as a person or group of people with minimum sales results on the size of the sales difference and the level of certain occurrences.

Revised Manuscript Received on April 25, 2019.

Fajar Delli Wihartiko, Department of Computer Science, Universitas Pakuan, Indonesia.

Doni Wihartika, Department of Management, Universitas Pakuan, Indonesia.

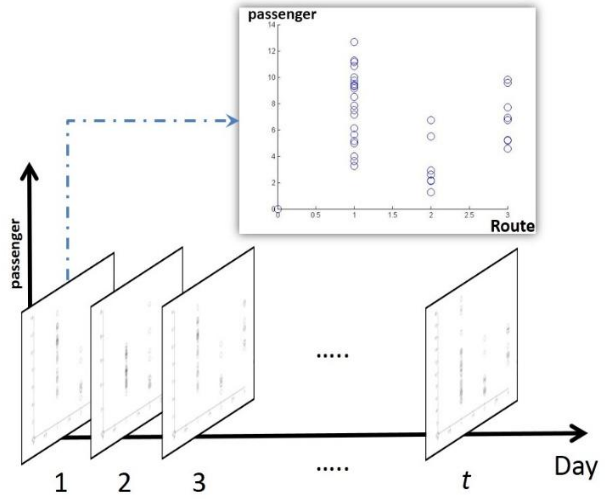


Fig. 1. The main idea of this research

One of the common algorithms used to classify spatial data is DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [11]. The DBSCAN algorithm is an algorithm developed by Esther and has the ability to form clusters well based on density [12]. This algorithm uses the distance measure to calculate the distance between points. The DBSCAN algorithm has the advantage of not requiring initial information about the number of clusters in a data set, finding clusters of various shapes, even finding clusters within other clusters and being aware of noise. The weakness of this algorithm is its performance depending on the function used to measure the distance between data and cannot group data well, if the density of each cluster is much different.

Based on the description, the fraud detection of Bogor's BRT Bus ticket sales is more suitable to be done with DBSCAN Algorithm approach, which is analogized as outlier ticket sales from clustering results. The ticket sales outline is defined as a person or group of employees with the smallest ticket sales in a single day. The search results of ticket sales outliers in a certain period are then categorized according to company needs into the level of possible employee cheating.

II. RESEARCH METHOD

Currently the cheating ticket detection at Bogor's BRT is done by looking at the smallest salespeople for each month. This will result in errors in detecting due to the possibility that the employee with the smallest sale is the result of the employee selling the tickets on a quiet lane, the number of days worked a little or the number of ritase achieved slightly. Alternative detection is done by adding aspects of absence, ritase and paths served. Ticket sales data and added aspects are then analogized into temporal spatio data. Ticket seller outlier search is done after clustering results using DBSCAN from spatio temporal data. The research stages to detect cheating ticket sales of Bogor's BRT buses are done by using the stages as in Fig. 2.

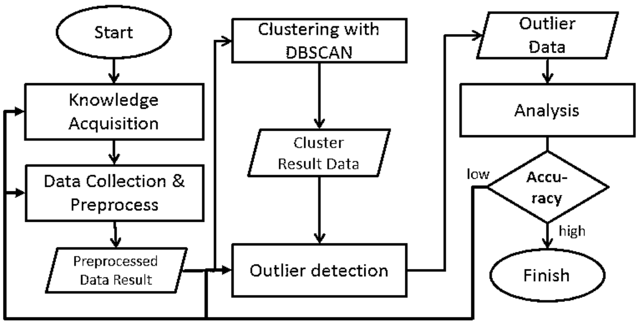


Fig. 2. The Research Methods

A. Knowledge Acquisition

Knowledge acquisition process is done by interview / in-depth interview to the manager of Bogor's BRT. The interview process is determined based on knowledge acquisition techniques [13]. The interview technique was chosen because in this case there is no procedure for the detection of cheating ticket sales and the available knowledge is explicit. The interview was conducted to the Head of Operations Division, Head of Sub Division of Finance and Accounting and Head of Sub Division of Operational Supervision. The results of interviews with bus managers are determination of parameters used and defining and categorizing outlier ticket seller.

B. Preprocess Data

The data used are daily sales reports of Bogor's BRT ticket along with the data about the path taken, the trip achieved and the absenteeism. Preprocess data is done by integrating the available reports into 3-sized cube data u × v × w with the meaning of u is a number Conductor, v is a day in a period of evaluation and there is w is the number of route served. The data cube is then transformed into spatial temporal data by passing passengers per trip and path into spatial data as well as time periods as temporal data. The spatial temporal data structure is shown in the following table:

Table 1. Data Type

No	Attribute	Data Type	Value Range
1.	Employee Name (point)	Nominal	1, 2, ….., 62
2.	Passenger (*z* axis)	Real	[0 – 100]
3.	Route (*x* axis)	Categorical	0, 1, 2, 3
4.	Date (*y* axis)	Ordinal	1, 2, …., 31

C. DBSCAN Clustering for Ticket Sales

The existing groupings on spatio temporal data of Bogor's BRT ticket sales are done by clustering technique by using DBSCAN algorithm. DBSCAN [12] is used to identify and categorize the density of a point. The idea of this algorithm is that for every point of a given radius it must have a certain number of points so that a high density region will be in one cluster whereas a low density

region will be identified as noise. The DBSCAN concept is defined by [14] which is then adjusted to the cheating ticket cheating problem is as follows:
Let $\varepsilon \in \boldsymbol{R}$ and $x \in \boldsymbol{D}$ with $\boldsymbol{D}$ be the set of all ticket sales people.

Definition 1. *Eps-neighborhood* from *x*
Eps-neighborhood of employee *x* denoted by $N_\varepsilon(x) = B_d(x,\varepsilon) = (y|dist(x,y) \le \varepsilon)$

Definition 2. *The Main Point*
For all $x \in \boldsymbol{D}$ defined *x* is *the main point* if at least there are employees of *minpts* in a *Eps-neighborhood*. In other words, *x* is called *the main point* if $|N_\varepsilon(x)|$>*minpts* where *minpts* $\in \mathrm{Z}^+$.

Definition 3. *The boundary point*
The boundary point is defined as an employee who does not meet the *minpts* size but is *the neighborhood* of a main point *z*, for example $x \in N_\varepsilon(z)$.

Definition 4. *Noise Point*
Employees who do not include *the main point* and *the boundary* are referred to as *noise points*.

Definition 5. *Density*
Employee *x* is called *directly density reachable* from another employee *y* if $x \in N_\varepsilon(y)$ and *y*is the main point. A point *x* is called the *reachable density of y* if there is a chain point $x_0, x_1, ..., x_l$ so that $x = x_0$ & $y = x_1$ and the point x_i is called *directly density reachable* from x_{i-1}, $\forall i = 1,2,...,l$. Let defined the point *x* and *y* as *density connected* if there is a main point *z* so the point *x* and *y* is *the reachable density of z. Density-based clusters* are defined as the maximum set of density conected points.

DBSCAN searches for a cluster by checking the neighborhood at each point in a database. If the *Eps-Neighborhood* of a point *x* has more than *MinPts*, a new bunch with a point *x* as a center of bunching is formed, then iteratively combines the reachable density cluster until there is no point that can not be added in clustering. The result of grouping using DBSCAN is used to detect outlier ticket sales.

D. Outlier Detection

The concept of bus ticket sales outlier detection is as follows. Let $j \in J$ be the set of bus routes (in Bogor's BRT has 3 pieces of route so *J* = {1, 2, 3}). Time $t \in T$ represents the period / date of the data used.

Definition 6. Outliers of Bus Ticket Sales
For example, defined Ol_t is the set of ticket seller outliers on day *t*. The ticket seller outliers is an employee or group of employees who sells the smallest ticket for each day on every *j* line with the following conditions:

- An employee $x \in \boldsymbol{D}$ is referred to as a ticket seller *outlier* for line *j* if $x \in noise$ and *x* is the smallest sale on line *j* (min(*psgr*; *j*)).
- All employeesxon a cluster i in route *j* ($C_{i,j}$) is called the ticket seller outlier if $C_{i,j}$ is the cluster with the smallest sales result for route*j* (min($C_{k,j}$); $k \in Z$) and there is one of the conditions in which there is another cluster on route *j* other than $C_{k,j}$ (*k* > 1) or a condition where there is at least *MinNoise* on route *j*.

In other words the ticket sales outlier on day *t* is defined as:

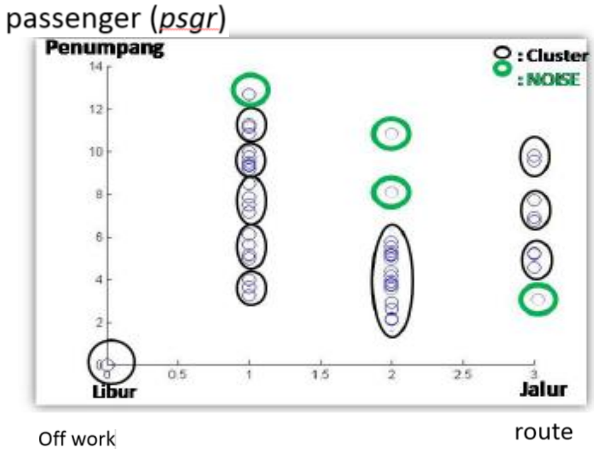
$$Ol_t = \{\{x \in D | x \in noise \wedge x = \min(psgr, j)\} \vee$$
$$\{\forall x \in C_{i,j} | C_{i,j} = \min(C_{k,j}) \wedge$$
$$(k > 1 \wedge noise \ge MinNoise\} : k \in Z, j \in J\}$$

The illustration of the Bus ticket sales outlier is shown in Figure 3.In summary Definition 6 can be written into Algorithm as follows:

Algorithm 1. *Bus Ticket Seller Outlier*

```
For all $x \in \boldsymbol{D}$
    For all $j \in J$
      if $x \in noise$ and $x = \min(psgr,j)$
      then $id_{ol}(x) = 1$
    else if $x \in C_{i,j}$ and
          $C_{i,j} = \min(C_{k,j}) \,\&\, (k > 1 \wedge Noise_j \ge MinNoise)$
    then $id_{ol}(x) = 1$
else if $id_{ol}(x) = 0$
$Ol \leftarrow \{x \in D | id_{ol}(x) = i)$
```

To obtain all bus ticket sales outliers in one period, DBSCAN Algorithm and Algorithm of Teller Sales Outline Determination are used iteratively over the number of days in the period, in which case it is 31 days. The measures used in this study are the results of interviews with management, and are shown in Table 2. To calculate the distance between employee points using Euclidean Distance. Machine learning process is done using MATLAB.



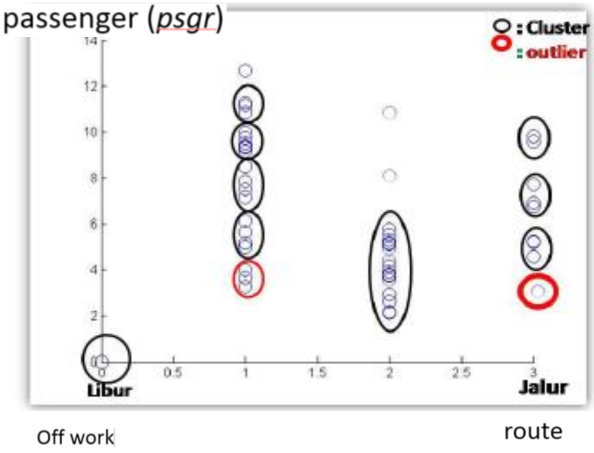


Fig. 3. Illustration of Noise and Outlier Ticket Sellers

Table 2. The Parameters

No	Parameter	Value
1	ε	0.99
2	Minpts	3
3	MinNoise	3

E. Outlier Detection Result Analysis Procedure

The results of the oulier detection are then compared with the employee's attendance to see the percentage of employees who appear to be outlier ticket sellers in a month. The results are then categorized into the level of possible employee fraud obtained based on the results of expert discussions. The level category of possible fraud is shown in Table 3.Statistical analysis is done by looking at the concentration and dissemination of results data. Measurement of accuracy / verification is done by comparing the results obtained with employee's historical data.

Table 3. Possible Fraud Rate Categorization

No	Criterion	Category
1.	$x_{ol} \leq 15\%$	Not potentially
2.	$15\% \leq x_{ol} \leq 30\%$	Low
3.	$30\% \leq x_{ol} \leq 45\%$	Medium
4.	$x_{ol} \geq 45\%$	High

III. RESULT AND ANALYSIS

The final result of the DBSCAN algorithm implementation and the Ticket Sales Outlier Algorithm is an outlier matrix with the dimensions of 62 × 31. This matrix represents information on 62 employees who become outlier ticket sellers for 31 days. The result of recapitulation of outlier appearance in 31 days accompanied by employee attendance data for one month, and percentage of outlier appearance to attendance can be seen in Table 4. Box plot of percentage occurrence of ticket seller outlier in Table 4 can be seen in Figure 4.

Table. 4 Occurrence of Outliers, Attendance and Percentage January 2017

Emp.	O.L.	Prsnc.	%
1	8	22	36
2	9	25	36
3	7	25	28
4	11	25	44
5	5	23	22
6	6	23	26
7	5	22	23
8	4	26	15
9	4	13	31
10	4	25	16
11	5	24	21
12	7	24	29
13	5	23	22
14	9	24	38
15	7	26	27
16	6	21	29
17	13	21	62
18	6	23	26
19	3	25	12
20	4	23	17
21	6	25	24
22	5	25	20
23	5	25	20
24	5	18	28
25	8	25	32
26	4	23	17
27	4	23	17
28	9	22	41
29	6	23	26
30	7	23	30
31	5	19	26
32	5	22	23
33	5	23	22
34	1	24	4
35	6	22	27
36	10	20	50
37	5	22	23
38	7	23	30
39	3	12	25
40	4	19	21
41	7	24	29
42	3	21	14
43	3	22	14
44	8	23	35
45	8	24	33
46	10	22	45
47	6	24	25
48	5	23	22
49	5	25	20
50	10	23	43
51	5	25	20
52	3	21	14
53	7	25	28
54	5	22	23
55	4	22	18
56	3	22	14
57	7	23	30
58	10	22	45
59	9	23	39
60	5	24	21
61	5	25	20
62	3	25	12

Information :

Emp	: Employee to 1, 2, ..., 62
O.L.	: Frequency to Outlier in a month
Prsnc.	: Employee presence in January 2017 (day)
%	: O.L Percentage of Prsnc.

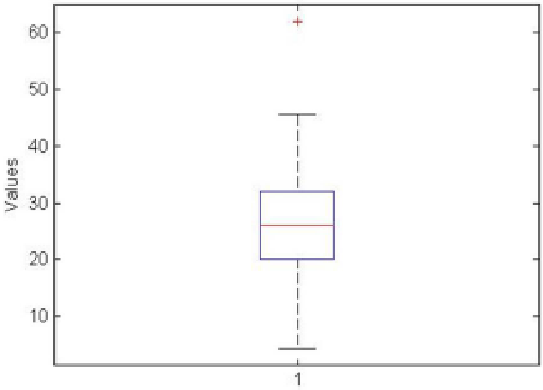


Fig. 4. Box plot of percentage occurrence of ticket seller outlier

Based on Fig. 4 it can be seen that the average percentage of occurrence of ticket sales outlier in January 2017 is 26.5% with data spread in the range of 20% -32%. Based on the distribution of the data, the management needs to increase supervision, because the tendency of cheating ticket sales has been to the middle level. The occurrence of these outliers is grouped according to the likelihood of fraud according to Table 3, so overall employee fraud potential for the January 2017 period can be seen in Fig. 5.

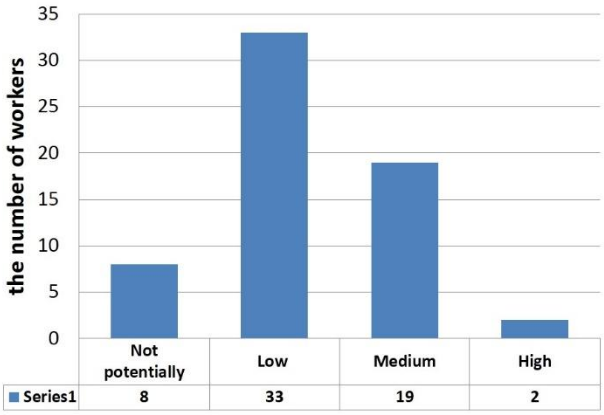


Fig. 5. Potential of Fraudulent Ticket Sales for January 2017

Detection results show that there are 3.2% of high-potential employees with ticket sales fraud, 30.6% are medium potential, 53.2% has low potential and only 12.9% of employees are not potentially cheating tickets. The results of verification with Bogor's BRT internal auditors indicate that the two employees who are detected "high potential" do fraud, do have a "negative" note on the performance of ticket sales. Such records include the negligence of ticket sales administration and delay on the deposit of the proceeds of ticket sales resulting in the small ticket sales. This indicates that the model has been 100% verified specifically for the detection of fraudulent high ticket sales. Bogor's BRT management takes firm action on both employees, by providing warning letters and direct reprimands. This is done to give a deterrent effect as well as coaching for employees to be able to improve its performance. If the fraud happens again in the following month, the management can provide punishment (postponement of employee status increase, decrease of status until Termination of Employment). In this case the model can be used as a preventive media for management in monitoring the cheating rate of ticket sales. This model can be used as a basis in the preparation of coaching programs for employees who are detected to commit fraud. Verification results of "detected" potential employees show that there is only one employee (5.26%) who has been proven to be cheating ticket sales. Management takes the same action of giving warning letters and direct rebuke. In addition to the above, no ticket sales violation was found. The experimental results use different data, have different performance times according to the number of data input. The time performance generated for processing in 3 cases of data according to the study [15] can be seen in Table 5. From the table it appears that relatively long time is needed to work on the problem of big-scale outlier detection.

Table 5. Performance Times Comparison

No	Cases	Size	Result (second)
1	Small	5 × 31 × 1	181
2	Medium	62 × 31 × 3	903
3	Big	277 × 31 × 10	20769

IV. CONCLUSION

The problem of bus ticket bus fraud detection can be solved by using DBSCAN Algorithm and Outlier Ticket Seller Algorithm with algorithm complexity of O (n^6). In the implementation data base shaped report is transformed into spatio temporal data by analysing the attribute of passengers per trip and path into spatial data and period of day into temporal data. The ticket sales outline is defined as a person or group of employees with the smallest ticket sales. The results showed that there were 3.2% of employees with high potential for fraud, and verification results showed 100% accuracy, especially in the case of high potential employees to cheat.

REFERENCES

1. Association of Certified Fraud Examiners (ACFE). 2012. Report to Nation. http://www.acfe.com/uploadedFiles/ACFE_Website/Content/rttn/2012-report-to-nations.pdf.
2. M. Goldstein and S. Uchida. "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data." *PloS one*, vol. 11, no. 4, art. no. e0152173, 2016.
3. E. W. Ngai, Y. Hu, Y. H. Wong, Y. Chen and X. Sun. "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature." *Decision support systems*, vol. 50, no. 3, pp.559-569, 2011.
4. E. Johansson and P. Carey. "Detecting fraud: The role of the anonymous reporting channel." *Journal of business ethics*, vol. 139, no. 2, 391-409, 2016.
5. GTZ. 2007. Bus Rapid Transit Planning Guide
6. E. Kirkos, C. Spathis and Y. Manolopoulos. "Data mining techniques for the detection of fraudulent financial statements." *Expert systems with applications*, vol. 32, no. 4, pp. 995-1003, 2017.
7. G. Sadowski and P. Rathle. *Fraud detection: Discovering connections with graph databases*. White Paper-Neo Technology-Graphs are Everywhere, 2014.
8. P. N. Tan. *Introduction to data mining*. India: Pearson Education India, 2018.
9. S. Kisilevich, F. Mansmann, M. Nanni and S. Rinzivillo. *Spatio-temporal clustering*. *Data mining and knowledge discovery handbook* . Boston: Springer, 2009.
10. F. D. Wihartiko, A. Buono and B. P. Silalahi. "Integer programming model for optimizing bus timetable using genetic algorithm." *IOP Conference Series: Materials Science and Engineering*. vol. 166, no. 1, p. 012016, 2017.
11. H. Bäcklund, A. Hedblom and N. Neijman. "A density-based spatial clustering of application with noise." *Data Mining TNM033*, pp. 11-30, 2011.
12. M. Ester, H. P. Kriegel, J. Sander and X. Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." *Kdd* , vol. 96, no. 34, pp. 226-231, 1996.
13. E. S. Berner, *Clinical decision support systems*. New York: Springer Science+ Business Media, LLC, 2007.
14. M. J. Zaki, Jr. W. Meira and W. Meira. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge: Cambridge University Press, 2014.
15. F. D. Wihartiko, H. Wijayanti and F. Virgantari. "Performance comparison of genetic algorithms and particle swarm optimization for model integer programming bus timetabling problem." *IOP Conference Series: Materials Science and Engineering*. vol. 332, no. 1, p. 012020, 2018.