Int j simul model 22 (2023) 1, 17-28
Original scientific paper

SAMPLING SIMULATION IN PROCESS DISCOVERY

Prasetyo, H. N.[*,**]; Sarno, R.[*]; Wijaya, D. R.[**]; Budiraharjo, R.[*,***] & Waspada, I.[*,****]
[*] Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[**] Department of Diploma of Information System, Telkom University, Bandung, Indonesia
[***] Department of Information Systems, Institut Teknologi Nasional, Bandung, Indonesia
[****] Department of Informatics, Universitas Diponegoro, Semarang, Indonesia
E-Mail: Riyanarto@if.its.ac.id, hanung.207025@mhs.its.ac.id

Abstract
Process model resulting from small event log datasets can be easily done because currently available applications are relatively able to do so. However, when faced with event logs from big data, modelling will force the existing applications to work hard. So far, the steps used are sampling of event logs resulting from the system. The problem arising is that the sampling process must be done several times because it has to check the desired fitness value on the sample taken. If the fitness value has not been got, then the sample size is added and the fitness value at a certain iteration is calculated until the required fitness value is met. There are many steps to do with this mechanism. Thus, this paper proposes an alternative way to reduce the steps by using an appropriate sampling technique. The mechanism used is statistical-based sampling simulation in the event log datasets to determine which sampling method is stable in the process modelling. The simulation results show that the sampling method using Cluster Random Sampling with the error rate or Alpha of 1 % has a relatively stable process model and can represent the process model resulting from the event log population.
(Received in July 2022, accepted in November 2022. This paper was with the authors 3 weeks for 1 revision.)

Key Words: Process Mining, Process Discovery, Big Data, Sampling, Event Log

1. INTRODUCTION

Process modelling is currently becoming an important thing in improving business processes in organizations or companies [1]. In the current process modelling, the approach taken is no longer based on field observations or interviews with the process users, but it is based on an analysis of the system's track record obtained from transactions and operational activities. System's track records are obtained from the event logs resulting through the generated activity database that is stored on the Server [2]. Companies currently have a large volume of transaction activities, so they feel the need for business process analysis [3]. Sometimes some researchers and practitioners refer to the large volume as Big Data [4].

In order for the analysis of the resulting event log to be good, process mining is currently the most qualified concept [5]. Process mining is a set of analytical activities carried out to obtain accurate and precise information related to the discovery of processes that have been running through the mechanism of process modelling, process measurement, and process performance improvement [6]. Process mining algorithms automatically discover process models based on event log data that is captured during the execution of business processes. Several studies have shown that most process discovery algorithms are started by building an internal data abstraction that is based on the entire event log and then applying the applicable filtering steps. Thus, some process discovery algorithms are not able to manage big data, in which the event log data is too large to process. In addition, some process mining tools set a limit on the size of the event log data [7].

Process discovery algorithms tend to use the entire event log data [8]. How much the event log data is required? There are many business processes that have a large volume of transactions. However, the ever-increasing size of data handled by process mining algorithms causes performance issues when implementing the existing process discovery algorithms. At

the same time, the event log data used for analysis usually relates only to certain intervals of process execution, not to the entire history [9]. Thus, when dealing with large event log data, it is no longer possible to use the existing standard hardware. With the increasing availability of event logs associated with business processes, the scalability of the techniques used in discovering the process model of event log is becoming a performance bottleneck. Several studies show that the development of machine learning algorithms to overcome the problem still requires a large amount of money [10]. This indicates that there are many systems that are still too complex to describe analytically or that the computation of solutions requires too many resources and takes a long time [11].

There are not many studies discussing event log sampling in the process discovery. However, there have been several studies in the last decade discussing the use of sampling in process mining or process discovery, for example the research conducted by [12, 13]. His research was related to event log sampling for process modelling. The research was conducted to check the desired fitness value on the sample taken. If the fitness value had not been reached, the sample size was added and then the fitness value was calculated iteratively. The purpose of this mechanism was to add new sections to the sample until the found process model had the required quality and represented the entire event log in a representative way. In 2017 there was another study related to event log sampling. The research was conducted to create an algorithm to perform sampling that was started with an empty sample. The algorithm having been created iteratively added *N* traces (which were randomly selected) to the samples until it stopped at the sample dependency value parameter [14]. Another interesting study came the following year. The study introduced a process discovery framework that relied on pre-processing statistics from event logs. The main thing to do was to reduce the sample size by reducing the runtime and trace in the event log process while setting a guarantee value for sampling errors [9]. The research was then developed in 2019, in which the results of the event log sampling were then combined with the results of estimation to see their suitability [15]. In 2019, researches were developed with the simulation that the approach taken accelerated sophisticated conformance checking algorithms by up to three times, while maintaining analytical accuracy [16].

Another approach was used in the event log sampling technique. In order to facilitate effective event log sampling, researches had been carried out on LogRank$^+$ based sampling. This method ranked the relevance of each trace and then selected the top *N* traces as log samples based on the input sampling rate. In subsequent experiments, each trial log was assigned a set of log samples with varying ratio (ranging from 5 % to 30 % with each 5 % increment). This was carried out by utilizing the LogRank$^+$ based Event Log Sampling plug-in [7]. This research was then developed in 2020, in which LogRank$^+$ sample event logs were used as a reference in the process of improving the discovery for describing business process models. The form of sample event logs were also developed through graphical form in other studies but still using the LogRank$^+$ approach [17].

There was an interesting thing conveyed in research conducted by Sani et al., in which most process discovery algorithms tended to use all event log data. Meanwhile, when dealing with large event data, it was no longer possible to use standard hardware in a limited time. In his research, he stated that a direct approach to overcome this problem was to reduce the size of the data by using random sampling method. This study proposed the concept of selecting multiple process samples from event logs based on variance or traces and applying process discovery algorithms to the selected samples. More metrics were then used to evaluate the quality and complexity of the discovered process models. The research showed that it was possible to speed up discovery techniques using sampling without losing the quality of the resulting process model [18]. This research was then continued by using a sample event log mechanism to monitor business processes that aimed to quickly predict the behaviour of business processes and to reduce the risk of unwanted behaviour in the process. By taking a

sample event log quickly, it could predict the results immediately [10]. Some of the studies above are in line with the research conducted by Knols and van der Werf [19]. The research was conducted to simulate the quality of the sample size. The event log was divided into several sample size groups, such as small, medium, and large sample sizes. The results showed the behavioural quality of the event log sample, in which the larger the sample the better.

Previous research literature shows that no one has ever done a sampling technique simulation in finding a process model. The majority of the research conducted sample event logs and measured the fitness value of the process model. According to some of the studies above, the sampling mechanism from the event logs is good. By taking samples from the event logs, process modelling becomes more effective and faster. In some sampling mechanisms, to measure the quality of the sample event log, it is necessary to measure the fitness of the sample taken. If the fitness is not good or is considered not good, then an event log is added to the sample used to measure the quality of the fitness value until a good fitness value is obtained and close to the population value. This mechanism is good for finding sample sizes of event log that can represent the event log population. However, how many times are needed to sample and add event logs? Doesn't it take a long time and a lot of energy in finding a sample event log that can represent the event log population? What is the optimal and representative sample size for the event log population in process modelling?

Based on the problems mentioned above, this paper shows the experimental results of performing a sampling technique on the event log as a part of the process modelling. The purpose of this study is to see the appropriate sample size of event log, without having to add an event log trail to the sample in order to get sampling results that can represent the event log population. In addition, this study is to see which sampling method is relatively the most stable in representing the event log data population by using all the sampling techniques that can be used in the experiment. This study is also to see the performance of the process model, which is seen by using a fitness trace. In terms of strengthening the sampling results, this study uses a measurement of the similarity of the process model resulting from the sampling technique used with the process model resulting from the event log population. The measurement used is the Jaccard method.

2. MATERIAL AND METHODS

2.1 Process mining and process discovery

Through the recording of event logs of the operating system, process mining is a tool for assessing and monitoring processes within an organization with the aim of identifying areas that need to be improved for process quality [20]. Process discovery, conformance checking, organizational mining, automated simulation model development, model extension, model repair, case prediction, and recommendations are all part of the process of process mining. In more detail, process discovery is a mechanism for searching, mapping, and documenting the existing business process activities automatically and based on data [21]. The automatically collected data is then analysed so that it can be automatically proposed in the process model [22].

2.2 Event log

When analysing a business processes, the current event log is used to track the current event with the specified time when the business process was running [23]. Event logs can be used to understand system activity and to investigate emerging issues in both simple and complex forms. Table I shows the event log form generated by the system. Each activity performed in

the process has a time record. The activity time depends on each case. The activity times recorded from case to case can be sequential or intersecting.

Table I: Event log in CSV format.

Case_ID	Activity	Resource	Start Timestamp	Complete Timestamp	Variant	Lifecycle: transition
1	Register	System	1970-01-02 18:23	1970-01-02 18:23	Variant 2	complete
1	Defect Analysis	Tester3	1970-01-02 18:23	1970-01-02 18:30	Variant 2	complete
1	Repair Test	Tester3	1970-01-02 18:49	1970-01-02 18:55	Variant 2	complete
10	Register	System	1970-01-01 17:09	1970-01-01 17:09	Variant 8	complete
10	Analyse Defect	Tester2	1970-01-01 17:09	1970-01-01 17:15	Variant 8	complete

2.3 Sampling techniques in statistics

Sampling in statistics is an important concept. The sampling technique used is related to the sampling method. The sampling techniques are carried out with several objectives, including [24]:

- to obtain data that is more accurate, but still relevant to the population that is the target of the study,
- to provide information related to the population to be studied, and
- can be used as a guide or reference in making a decision.

Based on the explanation above, it can be concluded that the sampling technique is formed as a way to determine the sample size to be used as a source of research data. This can be done by taking into account the characteristics and distribution of the population so that later a representative sample can be obtained. The sampling technique consists of two forms, namely random and non-random sampling [25]. The distribution of the sampling techniques is shown in Fig. 1.

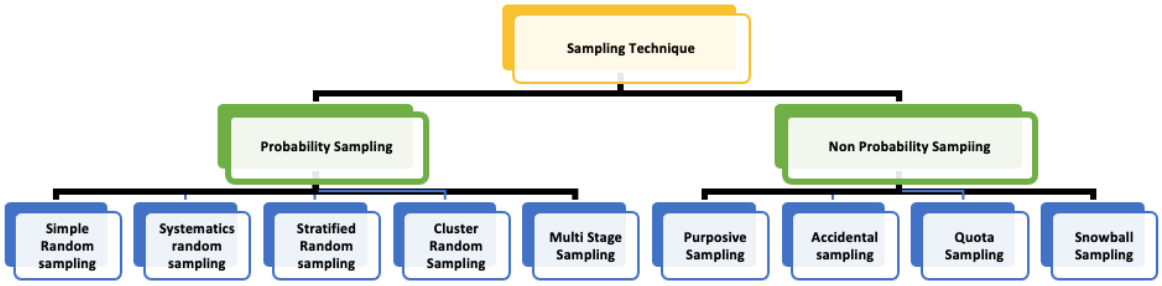

Figure 1: Sampling technique in statistics.

In principle, both random and non-random sampling techniques carry out the same steps. In determining and taking the sample, some steps are required, namely [26]:

- determining the population to be observed,
- determining the sample frame and the set of all possible events,
- determining the appropriate sampling technique or method,
- conduct sampling (data collection), and
- conducting re-examination of the sampling process.

Several machine learning tools already use plug-ins in performing sampling techniques. One of the machine learning tools is ProM. ProM is an independent platform implemented in the Java programming language. ProM has an extensible framework to support various process mining techniques in the form of plug-ins (see Fig. 2). This study used three sampling techniques, namely Random Sampling, Cluster Sampling, and Systematic Random. The basis for selecting the three sampling techniques is because they support the event log data form. Meanwhile, the stratified sampling technique can be represented by cluster random sampling.

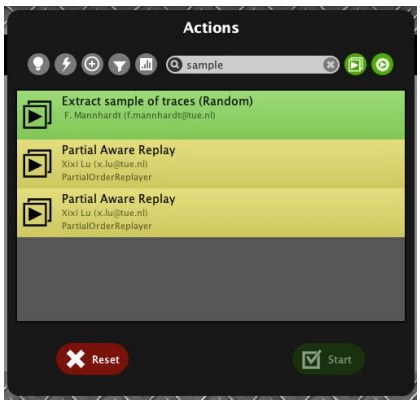

Figure 2: The feature of Random Sampling Trace on ProM.

2.4 Sample size

There are many methods of selecting sample size in probability schemes, such as the Isaac and Michael's method, Krejcie and Morgan's method, and so on. In several studies, the most stable method in determining sample size is the method of Isaac and Michael [27]. This method was developed gradually and continuously by Isaac and Michael (1981) and Smith (1983). The method was developed to determine the number of samples that meet the following requirements: (1) the total population is known; (2) at the error rate (Alpha) of 1 % and 5 %, and (3) this method is specifically used for normally distributed samples. Thus, this method cannot be used for samples that are not normally distributed, such as homogeneous samples [28]. This is what underlies the sampling technique used in this paper. The formula of Isaac and Michael's method is shown in Eq. (1):

$$n = (Z_\alpha^2 \,.\, N.\sigma^2)/(\varepsilon^2.N + Z_\alpha^2 \,.\, \sigma^2) \quad (1)$$

where n is the sample size, N is the population, σ is the standard deviation, Z_α is the desired significance level, and ε is the margin of error.

2.5 Methods

The research methodology carried out in this study is shown in Fig. 3.

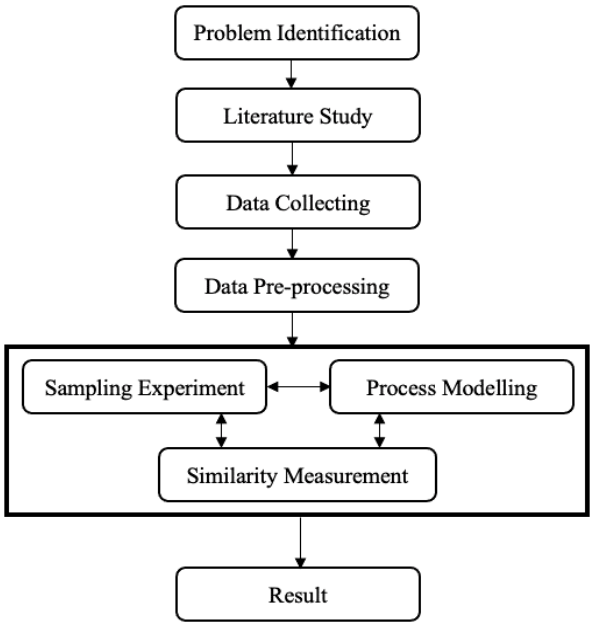

Figure 3: Research methodology.

The approach used is a quantitative approach, in which the flow is as follows:

1. Problem identification

   Problem identification is done to find and understand the main problems encountered. Identification is carried out thoroughly both to identify technical problems and related work structures and phenomena that occur [29]. As previously explained, there are several problems in the process modelling that originate from the event log having a large volume. Thus, we need a sampling technique from the event log. However, in conducting sampling, which method can obtain reliable sampling to produce process modelling that can represent the event log as a whole?

2. Literature study

   This stage is carried out for the data collection method by reading books, articles, notes, and reports related to the problem being solved [30]. The literature studies carried out are related to theories and discussions, such as process mining, process discovery, event log sub-selection, event log sampling, and statistical sampling.

3. Data collection

   Data collection is carried out as a simulation material. The data used is the Artificial-Repair Event Log dataset. The event log data is in the form of a table consisting of Case-ID, Events, Start Timestamps, Complete Timestamps, and Users. The dataset is open and accessible. The dataset used is one of the data taken from the Futurelearn dataset in *https://www.futurelearn.com/info/courses/process-mining/0/steps/15620*.

4. Pre-processing

   The pre-processing stage is carried out to adjust the event log to be used in the simulation.

5. Sampling experiment, process modelling and similarity measurement

   At this stage the sampling technique is used in the event log. After that, the measurements are carried out and the simulation results are presented. The sampling techniques used are Simple Random Sampling, Systematic Random Sampling, and Cluster Random Sampling. Meanwhile, the process modelling uses two approaches, namely Inductive Miner and Heuristic Miner Algorithm [31]. To complete the analysis, measurement of the similarity of the process model resulting from the sampling technique used with the process model resulting from the event log population is carried out. The measurement of the similarity uses the Jaccard Method with the determination of the process model using the Inductive Miner Algorithm and Heuristic Miner Algorithm [32, 33].

6. Result

   This stage is the final stage where the measurement results from the simulation, modelling results, and other results are needed for further analysis.

3. EXPERIMENT

3.1 Description of experimental sampling events

In the experimental section, the sampling technique was used in the event log population which was used as a case study. The search for a large sample size based on the Isaac and Michael's method does not need to test the normality of the event log population so that it can be continued with the sampling technique that will be used in this paper. The next step was to perform an experimental simulation on the event log by using three sampling techniques, namely Simple Random Sampling, Systematic Random Sampling, and Cluster Random Sampling with an alpha (α) at the error rate of 1 % and 5 %. The experimental results are shown in Table II.

Table II: Description of experimental sampling event log.

No.	Component	Event Log (As Population)	Simple Random Sampling		Cluster Random Sampling		Systematic Random Sampling	
			α (0.01)	α (0.05)	α (0.01)	α (0.05)	α (0.01)	α (0.05)
1	Events	7734	3898	1965	7790	4306	7047	6413
2	Cases	1104	552	276	553	311	1006	893
3	Activities	8	8	8	8	8	8	8
4	Resources	13	13	13	13	13	13	13
5	Frequency Mean	594,92	299,85	151,15	599,23	331,23	542,08	493,31
6	Frequency Std. Dev.	911,51	458,91	230,99	916,25	508,15	829,99	747,01

3.2 Fitness value

In each sample event log, the fitness value was measured in each process model resulting from the sampling technique carried out. The Process Modelling used Inductive Miner Algorithm and Heuristics Miner Algorithm. The measurement results can be seen in Table III.

Table III : Fitness value of Process Model with Inductive Miner and Heuristics Miner Algorithm.

Process Model with	Alpha (α)	Inductive Miner Algorithm	Heuristics Miner Algorithm
		Trace fitness	Trace fitness
Simple Random	5 %	0.649	0.851
Simple Random	1 %	0.649	0.857
Cluster Random	5 %	0.623	0.998
Cluster Random	1 %	0.649	0.998
Systematic Random	5 %	0.626	0.994
Systematic Random	1 %	0.649	0.998

3.3 Similarity

The next step was to analyse the similarity of the process model resulting from the sampling technique used with the process model resulting from the event log population. This was done to strengthen the results of previous measurements. The measurement of similarity in this study also used two algorithm approaches, namely the Inductive Miner Algorithm and the Heuristic Miner Algorithm. The use of the two algorithm approaches is to show the comparison of process modelling performance.

3.4 Measurement of similarity

After getting the event log sampling, the process modelling was then carried out on each event log using the Inductive Miner Algorithm. Fig. 4 shows the process model resulting from the event log population which will be used as a comparison. Meanwhile, Figs. 5 to 10 show the process model resulting from the sampling technique.

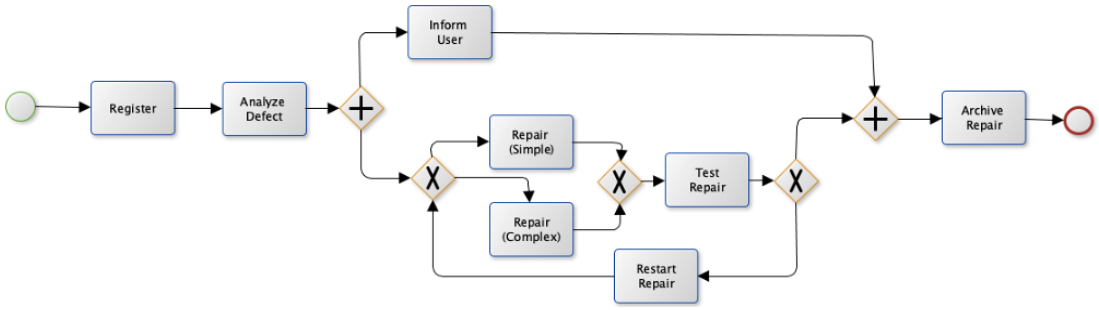

Figure 4: Process Model resulting from Event Log Population.

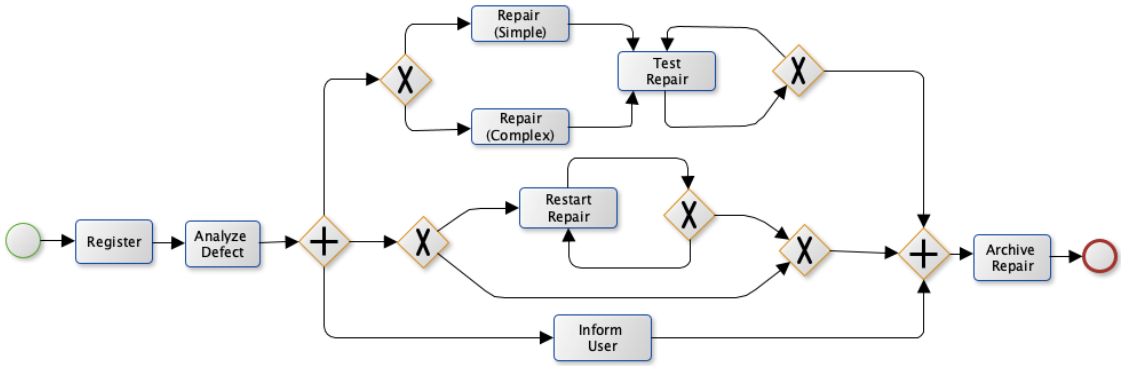

Figure 5: Process Model resulting from Simple Sampling Random with 5 % Event Log.

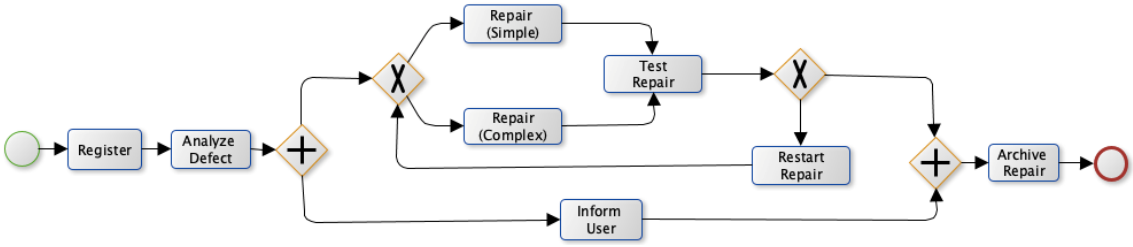

Figure 6: Process Model resulting from Simple Sampling Random with 1 % Event Log.

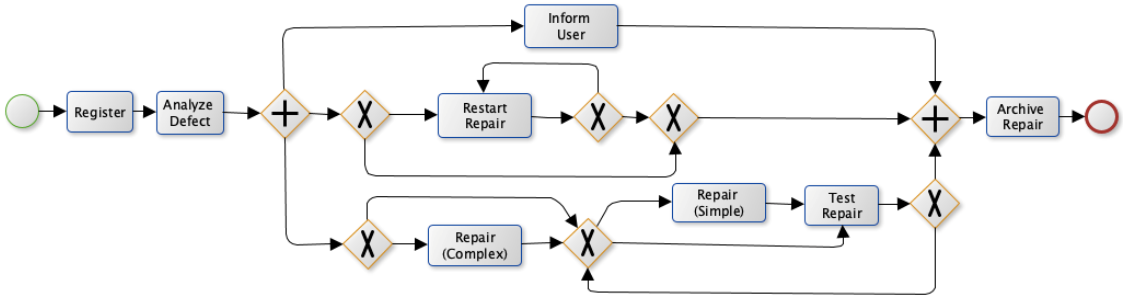

Figure 7: Process Model resulting from Cluster Sampling Random with 5 % Event Log.

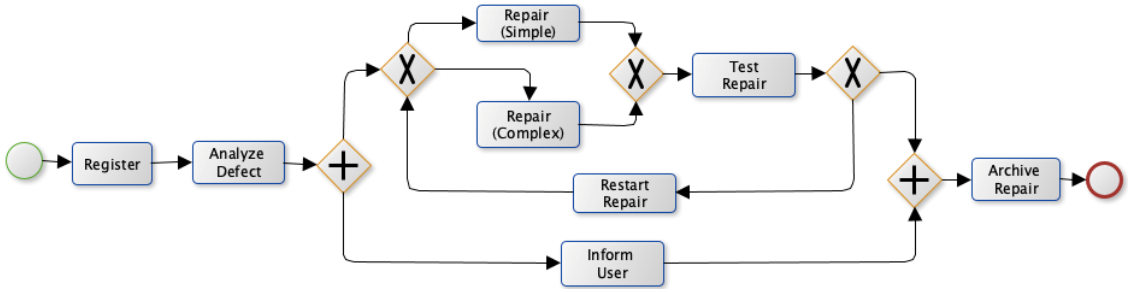

Figure 8: Process Model resulting from Cluster Sampling Random with 1 % Event Log.

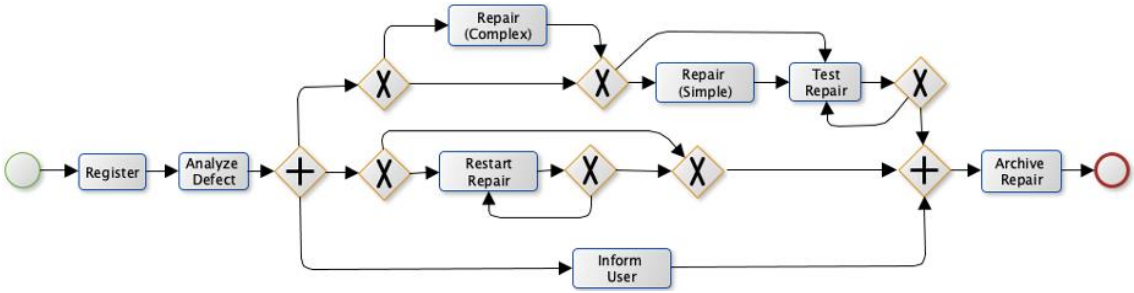

Figure 9: Process Model resulting from Systematic Sampling Random with 5 % Event Log.

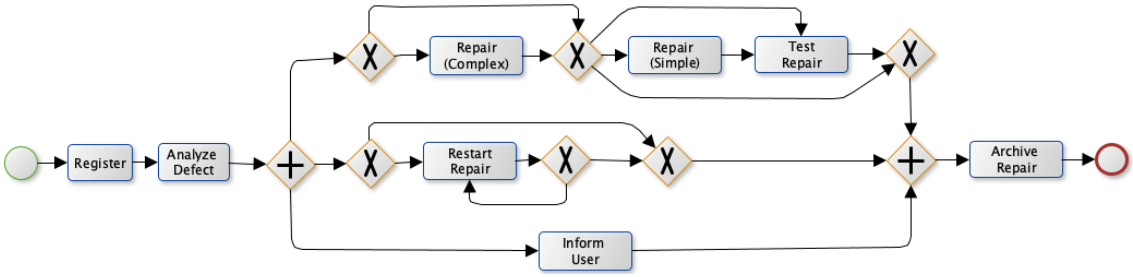

Figure 10: Process Model resulting from Systematic Sampling Random with 1 % Event Log.

The next step was to measure the similarity of the process model resulting from the sampling technique used with the process model resulting from the event log population by using the Jaccard method, which was based on the process model from Figs. 4 to 10. Tables IV, V, VI, and VII show the measurement results.

Table IV : Jaccard Structure Similarity Score.

Process Model (PM)	*ed*	*tk*	*g*	*t*	*e*	Structure Similarity		Score
						| Pop ∩ MP |	| Pop ∪ MP |	
Population of Process Model	2	8	5	17	32			
With Simple Random Sampling (***α***: 5 %)	2	8	7	22	39	32	39	0.820
With Simple Random Sampling (***α***: 1 %)	2	8	4	15	29	29	32	0.906
With Cluster Random Sampling (***α***: 5 %)	2	8	8	24	42	32	42	0.762
With Cluster Random Sampling (***α***: 1 %)	2	8	5	17	32	32	32	1.000
With Systematic Random Sampling (***α***: 5 %)	2	8	8	23	41	32	41	0.780
With Systematic Random Sampling (***α***: 1 %)	2	8	8	24	32	32	32	1.000

Notes: *Structural Similarity (SS): Pop = Event Log Population, MP = Process Model of Sampling Technique, ed = Edges, tk = Task, g = Gateway, t = Transitions, e = Total Number of Elements, | Pop ∩ MP | = Intersection of Pop and Process Model of Sampling Technique, | Pop ∪ MP | = Union of Pop and Process Model of Sampling Technique.*

Table V: Jaccard Behaviour Score.

Process Model (PM)	*e*	Behaviour Similarity		Score
		| Pop ∩ MP |	| Pop ∪ MP |	
Population of Process Model	10			
With Simple Random Sampling (***α***: 5 %)	11	10	11	0.909
With Simple Random Sampling (***α***: 1 %)	10	10	10	1.000
With Cluster Random Sampling (***α***: 5 %)	11	10	11	0.909
With Cluster Random Sampling (***α***: 1 %)	10	10	10	1.000
With Systematic Random Sampling (***α***: 5 %)	12	10	12	0.833
With Systematic Random Sampling (***α***: 1 %)	13	10	13	0.769

Notes: *Behavioural Similarity (BS): e = Total Number of Elements, | Pop ∩ MP | = Intersection of Pop and Process Model of Sampling Technique | Pop ∪ MP | = Union of Pop and Process Model of Sampling Technique*

Table VI: Similarity Score of Process Model with Inductive Miner Algorithm.

Comparison	Structure Similarity	Behaviour Similarity	Final Score
Process Model from Simple Random Sampling (***α***: 5 %)	0.820	0.909	0.864
Process Model from Simple Random Sampling (***α***: 1 %)	0.906	1.000	0.953
Process Model from Cluster Random Sampling (***α***: 5 %)	0.762	0.909	0.836
Process Model from Cluster Random Sampling (***α***: 1 %)	1.000	1.000	1.000
Process Model from Systematic Random Sampling (***α***: 5 %)	0.780	0.833	0.806
Process Model from Systematic Random Sampling (***α***: 1 %)	1.000	0.769	0.884

In the same way, the measurement of similarity was carried out with the Heuristic Miner Algorithm using the Jaccard similarity method. The results are as shown in Table VII.

Table VII: Similarity Score of Process Model with Heuristics Miner Algorithm.

Process Model with	Alpha (α)	Final score
Simple Random	5 %	0.065
Simple Random	1 %	0.065
Cluster Random	5 %	0.197
Cluster Random	1 %	1.000
Systematic Random	5 %	0.197
Systematic Random	1 %	0.301

4. RESULTS AND DISCUSSION

The experimental results of the sampling techniques used (Simple Random Sampling, Systematic Random Sampling, and Cluster Random Sampling) in Table II show descriptive statistics that Cluster Random Sampling has a value that is close to the event log population. The average value of the event log sample frequency using Cluster Random Sampling with the error rate of 1 % is closer to the event log population than samples from other sampling techniques. Likewise, the standard deviation value for the sample from Cluster Random Sampling with the error rate of 1 % also has a value that is close to the event log population, namely 599.23 and 594.92. To strengthen the results of the statistical description, it can be seen that the fitness value of process model resulting from Cluster Random Sampling with the error rate of 1 % and 5 % shows a stable value of 0.649 for the Inductive Miner Algorithm. Meanwhile, the process model resulting from the Heuristics Miner Algorithm shows a stable fitness value at 0.998 in Cluster Random Sampling at the error rate of 1 % and 5 % and Systematic Random Sampling but only at the error rate of 1 %.

To strengthen these results, measurement of the similarity of the process model resulting from the sampling technique used with the process model resulting from the event log population was carried out. Tables VI and VII show that the process model resulting from Cluster Random Sampling with the error rate of 1 % has a stable similarity value of 1 for the two algorithm approaches taken, namely Inductive Miner Algorithm and Heuristics Miner Algorithm. Based on these three measurements, it can be concluded that the process model resulting from the sampling technique using Cluster Random Sampling with the error rate of 1 % shows stable results as a representative process model of event log population. Cluster Random Sampling with the error rate of 1 % can be a recommendation in sampling the event log in the discovery process model. This will certainly be more effective than the sampling mechanism by increasing the sample size and measuring the fitness value repeatedly.

This study has limitations. As previously explained, the proposed method for event log datasets in this study is normally distributed. Further research is needed for event log data that is not normally distributed.

5. CONCLUSION

From the three sampling method experiments, based on descriptive sampling trend values and the measurements of fitness value and similarity using the Jaccard method, it was obtained the results that the sampling method using Cluster Random Sampling with the error rate of 1 % has a relatively stable process model and can represent the process model resulting from event log population, when it is compared to other sampling methods. When being faced with large volume event logs such as Big Data and requiring effective performance, this paper recommends using a sampling technique with Cluster Random Sampling at the error rate of 1 %. This mechanism will make process modelling more effective.

ACKNOWLEDGEMENT

The researchers would like to thank the Smart Information Management Laboratory Community, Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, especially colleagues who have become friends for pleasant discussions. We also thank the PPM Unit of Telkom University which fully supports the course of this research.

REFERENCES

[1] Claes, J.; Vanderfeesten, I.; Pinggera, J.; Reijers, H. A.; Weber, B.; Poels, G. (2015). A visual analysis of the process of process modeling, *Information Systems and E-Business Management*, Vol. 13, 147-190, doi:10.1007/s10257-014-0245-4

[2] Van der Aalst, W. M. P. (2011). Process discovery: an introduction, *Process Mining*, Springer, Berlin, 125-156, doi:10.1007/978-3-642-19345-3_5

[3] Gudelj, M.; Delic, M.; Kuzmanovic, B.; Tesic, Z.; Tasic, N. (2021). Business process management model as an approach to process orientation, *International Journal of Simulation Modelling*, Vol. 20, No. 2, 255-266, doi:10.2507/IJSIMM20-2-554

[4] George, G.; Haas, M. R.; Pentland, A. (2014). Big data and management, *Academy of Management Journal*, Vol. 57, No. 2, 321-326, doi:10.5465/amj.2014.4002

[5] Rautenburger, L.; Liebl, A. (2021). Process mining, Liermann, V.; Stegmann, C. (Eds.), *The Digital Journey of Banking and Insurance*, Palgrave Macmillan, Cham, 259-275, doi:10.1007/978-3-030-78829-2_15

[6] Accorsi, R.; Ullrich, M.; van der Aalst, W. M. P. (2012). Process mining, *Informatik-Spektrum*, Vol. 35, No. 5, 354-359, doi:10.1007/s00287-012-0641-4

[7] Liu, C.; Pei, Y.; Zeng, Q.; Duan, H.; Zhang, F. (2020). LogRank$^+$: a novel approach to support business process event log sampling, Huang, Z.; Beek, W.; Wang, H.; Zhou, R.; Zhang, Y. (Eds.), *Web Information Systems Engineering – WISE 2020*, *Lecture Notes in Computer Science*, Vol. 12343, Springer, Cham, 417-430, doi:10.1007/978-3-030-62008-0_29

[8] Sarno, R.; Sungkono, K. R.; Taufiqulsa'di, M.; Darmawan, H.; Fahmi, A.; Triyana, K. (2021). Improving efficiency for discovering business processes containing invisible tasks in non-free choice, *Journal of Big Data*, Vol. 8, No. 1, Paper 113, 17 pages, doi:10.1186/s40537-021-00487-x

[9] Bauer, M.; Senderovich, A.; Gal, A.; Grunske, L.; Weidlich, M. (2018). How much event data is enough? A statistical framework for process discovery, Krogstie, J.; Reijers, H. (Eds.), *Advanced Information Systems Engineering – CAiSE 2018*, *Lecture Notes in Computer Science*, Vol. 10816, Springer, Cham, 239-256, doi:10.1007/978-3-319-91563-0_15

[10] Sani, M. F.; Vazifehdoostirani, M.; Park, G.; Pegoraro, M.; van Zelst, S. J.; van der Aalst, W. M. P. (2022). Event log sampling for predictive monitoring, Munoz-Gama, J.; Lu, X. (Eds.), *Process Mining Workshops – ICPM 2021*, *Lecture Notes in Business Information Processing*, Vol. 433, Springer, Cham, 154-166, doi:10.1007/978-3-030-98581-3_12

[11] Simon, E.; Oyekan, J.; Hutabarat, W.; Tiwari, A.; Turner, C. J. (2018). Adapting Petri nets to discrete event simulation for the stochastic modelling of manufacturing systems, *International Journal of Simulation Modelling*, Vol. 17, No. 1, 5-17, doi:10.2507/IJSIMM17(1)403

[12] Bratosin, C.; Sidorova, N.; van der Aalst, W. (2010). Discovering process models with genetic algorithms using sampling, Setchi, R.; Jordanov, I.; Howlett, R. J.; Jain, L. C. (Eds.), *Knowledge-Based and Intelligent Information and Engineering Systems – KES 2010*, *Lecture Notes in Computer Science*, Vol. 6276, Springer, Berlin, 41-50, doi:10.1007/978-3-642-15387-7_8

[13] Bratosin, C.; Sidorova, N.; van der Aalst, W. (2011). Distributed genetic process mining using sampling, Malyshkin, V. (Ed.), *Parallel Computing Technologies – PaCT 2011*, *Lecture Notes in Computer Science*, Vol. 6873, Springer, Berlin, 224-237, doi:10.1007/978-3-642-23178-0_20

[14] Berti, A. (2017). Statistical sampling in process mining discovery, *The 9th International Conference on Information, Process, and Knowledge Management*, 41-43

[15] Sani, M. F.; van Zelst, S. J.; van der Aalst, W. M. P. (2019). The impact of event log subset selection on the performance of process discovery algorithms, Welzer, T.; Eder, J.; Podgorelec, V.; Wrembel, R.; Ivanović, M.; Gamper, J.; Morzy, M.; Tzouramanis, T.; Darmont, J.; Kamišalić Latifić, A. (Eds.), *New Trends in Databases and Information Systems – ADBIS 2019*,

Communications in Computer and Information Science, Vol. 1064, Springer, Cham, 391-404, doi:10.1007/978-3-030-30278-8_39

[16] Sani, M. F. (2020). Improving the performance of process discovery algorithms by instance selection, *Computer Science and Information Systems*, Vol. 17, No. 3, 927-958, doi:10.2298/CSIS200127028S

[17] Liu, C.; Pei, Y.; Cheng, L.; Zeng, Q.; Duan, H. (2021). Sampling business process event logs using graph-based ranking model, *Concurrency and Computation: Practice and Experience*, Vol. 33, No. 5, Paper e5974, 14 pages, doi:10.1002/cpe.5974

[18] Sani, M. F.; van Zelst, S. J.; van der Aalst, W. M. P. (2021). The impact of biased sampling of event logs on the performance of process discovery, *Computing*, Vol. 103, No. 6, 1085-1104, doi:10.1007/s00607-021-00910-4

[19] Knols, B.; van der Werf, J. M. E. M. (2019). Measuring the behavioral quality of log sampling, *Proceedings of the 2019 International Conference on Process Mining*, 97-104, doi:10.1109/ICPM.2019.00024

[20] Van der Aalst, W. (2012). Process mining, *Communications of the ACM*, Vol. 55, No. 8, 76-83, doi:10.1145/2240236.2240257

[21] Sungkono, K. R.; Ahmadiyah, A. S.; Sarno, R.; Haykal, M. F.; Hakim, M. R.; Priambodo, B. J.; Fauzan, M. A.; Farhan, M. K. (2021). Graph-based process discovery containing invisible non-prime task in procurement of animal-based ingredient of halal restaurants, *Proceedings of the 2021 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, 134-140, doi:10.1109/APWiMob51111.2021.9435261

[22] Van der Aalst, W. M. P. (2018). Process discovery from event data: relating models and logs through abstractions, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 8, No. 3, Paper e1244, 21 pages, doi:10.1002/widm.1244

[23] Van der Aalst, W. M. P.; van Dongen, B. F. (2013). Discovering Petri nets from event logs, Jensen, K.; van der Aalst, W.M.P.; Balbo, G.; Koutny, M.; Wolf, K. (Eds.), *Transactions on Petri Nets and Other Models of Concurrency VII, Lecture Notes in Computer Science*, Vol. 7480, Springer, Berlin, 372-422, doi:10.1007/978-3-642-38143-0_10

[24] Mitchell, J.; McDaniel, W. (1969). Adaptive sampling technique, *IEEE Transactions on Automatic Control*, Vol. 14, No. 2, 200-201, doi:10.1109/TAC.1969.1099144

[25] Djauhari, M. A. (2020). *Sample Size: Generic Formulas for Social Practitioners (Ukuran Sampel: Formula Generik Bagi Praktisi Sains Sosial)*, 1st ed., ITB Press, Bandung (in Indonesian)

[26] Narayanan, S.; Lin, F. C. (1985). Sampling technique, Wong, S. H. Y. (Ed.), *Therapeutic Drug Monitoring and Toxicology by Liquid Chromatography*, CRC Press, New York, 79-88

[27] Isaac, S.; Michael, W. B. (1995). *Handbook in Research and Evaluation: A Collection of Principles, Methods, and Strategies Useful in the Planning, Design, and Evaluation of Studies in Education and the Behavioral Sciences*, 3rd ed., Edits Publishers, San Diego

[28] Johanson, G. A.; Brooks, G. P. (2010). Initial scale development: sample size for pilot studies, *Educational and Psychological Measurement*, Vol. 70, No. 3, 394-400, doi:10.1177/0013164409355692

[29] Kumar, R. (2019). *Research Methodology: A Step-by-Step Guide for Beginners*, 5th ed., Sage Publications Ltd, Los Angeles

[30] Snyder, H. (2019). Literature review as a research methodology: an overview and guidelines, *Journal of Business Research*, Vol. 104, 333-339, doi:10.1016/j.jbusres.2019.07.039

[31] Saint, J.; Fan, Y.; Singh, S.; Gasevic, D.; Pardo, A. (2021). Using process mining to analyse self-regulated learning: a systematic analysis of four algorithms, *LAK21: 11th International Learning Analytics and Knowledge Conference*, 333-343, doi:10.1145/3448139.3448171

[32] Coupette, C.; Vreeken, J. (2021). Graph similarity description: how are these graphs similar?, *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 185-195, doi:10.1145/3447548.3467257

[33] Dijkman, R.; Dumas, M.; van Dongen, B.; Käärik, R.; Mendling, J. (2011). Similarity of business process models: metrics and evaluation, *Information Systems*, Vol. 36, No. 2, 498-516, doi:10.1016/j.is.2010.09.006