The Effect of Duration Heteroscedasticity to the Bottleneck in Business Process Discovered by Inductive Miner Algorithm

Hanung Nindito Prasetyo[1,2]
[1]Department of Informatics
Institut Teknologi Sepuluh November, Surabaya
[2]Department of Information System Diploma
Telkom University
Indonesia
[1]hanung.207025@mhs.its.ac.id

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh November
Surabaya
Indonesia
riyanarto@if.its.ac.id

Raden Budiraharjo[1,2]
[1]Departement of Informatics
Institut Teknologi Sepuluh November, Surabaya
[2]Department of Information System
Institut Teknologi Nasional, Bandung
Indonesia
[1]budiraharjo.207025@mhs.its.ac.id

Kelly Rossa Sungkono
Department of Informatics
Institut Teknologi Sepuluh November
Surabaya
Indonesia
kelly@its.ac.id

2021 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob) | 978-1-7281-9475-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/APWIMOB51111.2021.9435199

***Abstract*— One way to do business process modelling is to use the process mining. Process mining links the gap between traditional model-based process analysis such as business process management simulation and data-centric analysis techniques such as machine learning and data mining. In process modelling, bottleneck conditions are often found. Bottlenecks conditions can be found in the process models generated using Process Mining applications such as ProM and Disco based on event log data. There is another alternative to find the bottleneck condition of the event log using a statistical approach. The alternative is to view the event log as an asset that can be explored without using a normative process model. This paper proposes a statistical test of heteroscedasticity in event log data. Then the heteroscedasticity test results from the event log are compared with the results of normative process modelling with the Inductive Miner algorithm using the Process Mining application. The comparison results show that the detected event log data having heteroscedasticity problems will ensure a bottleneck condition in the process model. The approach taken can be an alternative in evaluating the process model based on its event log.**

Keywords*— *Process Discovery, Heteroscedasticity, Bottleneck, Inductive Miner Algorithm.

I. INTRODUCTIONS

The quality of information technology-based service processes is currently an important part of the organization in providing and accommodating stakeholders. The quality of the service process is highly dependent on the business processes that have been built and established by the organization. Almost all aspects of work in an organization are based on predefined business processes. Today's organizations have realized how important business process development is. The business process is a reference for all users, both internal and external to the organization[1]. To answer the needs of the times, business processes are always improved every time according to the needs of the organization.

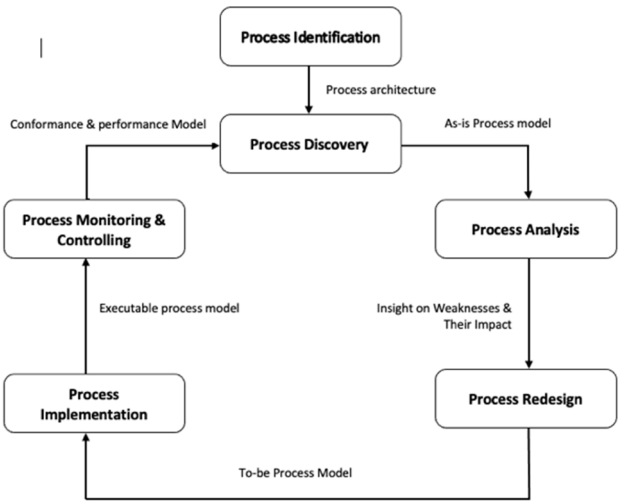

Fig. 1 Business Process management life Cycle[2]

In the business process management life cycle as shown in Figure 1, business process improvement is based on an analysis of ongoing business processes. The results of the process analysis form the basis for process improvement or redesign of the process model. In principle, process analysis is to do Most organizations perform process analysis based on various weaknesses found directly in the ongoing business process.

Process mining is currently an important requirement for organizations to improve the quality of their business processes[3]. Various benefits can be obtained from the mining process, such as to find out how the process actually happened. Find out whether the running process is in

accordance with the previously designed model. Currently, to improve the quality of business processes is to use process discovery mechanisms as part of process mining[4]. In the current context, the discovery process is no longer based on manual field reports but can be explored using the event log generated from the current system. Through the event log, all business activities that occur in the system can be seen so that the process model can be automatically predicted through this mechanism[5].

One of the important problems in the resulting process model is looking for bottleneck conditions in the process. A bottleneck condition is a condition that shows the occurrence of a process queue which causes the process to look ineffective[4]. In principle, a bottleneck is a condition of events in a transaction that has a longer waiting time than other transactions in a business process[6]. On the normative discovery process, this can be seen by "firing an event log" on the mining application process using an algorithm according to user needs. Through a mechanism like this one can evaluate the resulting process model. However, apart from using a process mining application, there is an alternative way to determine a bottleneck condition by exploring the resulting event log. the event log is not only a tool to do the discovery process but also an important asset as part of the analysis process. The approach taken in exploring the event log is to use the linear regression method to identify any heteroscedasticity problems from the event log data held. Based on calculations and analysis carried out on the resulting event log, it is found that if the regression model has a heteroscedasticity problem, the process model generated through a particular modelling algorithm has a bottleneck problem in the process. The modelling algorithm used to verify the bottleneck condition in this study uses the inductive miner algorithm.

II. RELATED WORK

Other research searches are conducted to see the contribution of the research that is being carried out. Other research related to the exploration of Log events as an asset has not been done much. Through searches on scientific databases such as ACM, IEEE Xplore, SpringerLink, ScienceDirect, and Taylor & Francis online, 16 papers were found. After reviewing 16 papers, only 6 were actually related to the event log exploration. This can be seen in the research conducted by Lu et al[7]. The research being conducted is to take a new approach to detect deviations at the event level by identifying frequent and unusual behaviors among examples of processes being carried out without finding a normative model. What it does is check for traces of abnormal behaviors or traces of irregular patterns in the event log. Another study conducted by Bauer et al. [8] who presented a framework for process discovery that relies on statistical pre-processing of the event log and significantly reduces its size by means of sampling. it reduces the runtime and memory footprint of the process discovery algorithm, while providing assurance against introduced sampling errors. In another study, Sani [9] in his paper evaluated various subset selection methods and evaluated their performance on real event data. The proposed method has been implemented in the ProM and RapidProM platforms. Experimental results suggest that it is possible to significantly accelerate discovery using a ranking-based strategy. Furthermore, the results show that biased selection of process samples compared to random selection will result in a higher quality process model. while Ceravolo [10] in his research proposed a method based on statistical inference for event log pre-processing. What it does is measure the distance between different segments of the event log, calculating the probability distribution of the observed activity at a given position. This study demonstrates the approach taken by developing case studies with logs of real life events and shows that the proposed method offers interesting properties in terms of computational complexity.

One of the highlights is the research conducted by Aydemir, et al. [11] where the aim of the paper is to summarize the development of a hybrid analytic approach that utilizes a back-end platform based on SQL and NO-SQL in harmony to run process mining for participating banks in Turkey. In this research, he create a process mining framework that visualizes process performance indicators and proposes workflow design changes and performs statistical tests to identify performance fluctuations by primarily using a parallel in-memory processing framework, named Apache Spark. Unfortunately, the description of statistical performance measurement is not shown how in this paper. statistical testing is not submitted and only stated to have been carried out. Based on the search until the writing of this thesis, it can be ascertained that there has been no research that specifically explores the Event Log as an asset with a statistical approach, especially the problem of heteroscedasticity which is part of linear regression analysis. This research will pave the way for event log exploration as part of the discovery process in the mining process.

III. MATERIAL & METHODS

A. *Process mining*

Process mining is an activity or mechanism used to explore and analysed event data to gain knowledge and insights generated during process execution [2]. as for the details of process mining activities are discovery, modelling, monitoring, and optimizing the underlying processes.

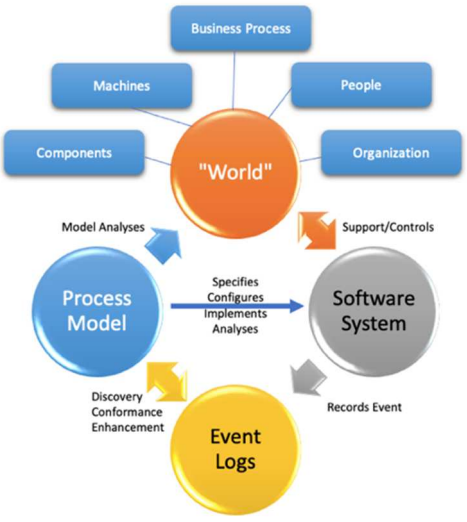

Fig. 2 Process Mining[8]

The following are the advantages that can be obtained when carrying out process mining [8]

- Process discovery is the activity undertaken to convert an event log into a process model for analysis.
- Conformity checking is the activity of gathering information about the differences between models and what happens in real life. With this step, the company will be able to find various things that are needed to improve the ongoing process.
- Throughput / bottleneck analysis is an activity to calculate the intensity of the implementation of the event to determine the potential bottlenecks in the process. This kind of analysis can be used to improve Key Performance Indicator (KPI) in time related processes to minimize throughput / overhead time.

In principle, current process mining links the gap between traditional model-based process analysis such as business process management simulation and data-centric analysis techniques such as machine learning and data mining [9].

B. Process Discovery

Process Discovery is an automated, data-based mechanism to find, map, and document existing business process activities. Then an analysis is carried out on the automatically obtained data so that it can be recommended automatically in modelling processes, or workflows [4]. Digital footprint left in the system and a detailed representation of business processes is generated automatically. Organizational business process discovery and analysis can be used in identifying key problem areas, not only at the start of a digital transformation initiative, but also when improving the performance of existing processes. Along with business process modelling, Process Discovery is an important part in improving the quality of business process management [10].

C. Event log

In business process analysis, event log is a collection of event records with timestamps generated by running business processes [5]. The event log records every event that occurs in system execution to provide an audit trail that can be used to understand system activity and diagnose problems that occur in both simple and complex forms.

The definition of Event Log *(Event, trace, Event Log). Let A be a set of activities. $\sigma \in A$ is trace, which is the sequence of events. $L \in B(A)$ is an event log, referred to as a multi-set(Bags)[6].*

The event log form is usually in the CSV extension. As shown in table 1.

TABLE 1 FORM OF EVENT LOG

Case ID	Activity	Start Time	End Time
Case 1	Claims submission	02/03/20 03.15	02/03/20 03.30
Case 1	Claims Acceptance	02/03/20 03.35	02/03/20 03.39
Case 1	Claims Survey	03/03/20 08.20	03/03/20 12.15
Case 1	Claims Rejected	05/03/20 13.15	05/03/20 13.20
Case 2	Claims submission	02/03/20 08.15	02/03/20 08.25
Case 2	Claims Acceptance	02/03/20 13.35	02/03/20 13.39

To be able to do process modelling from the event log, at least Case ID, Task name/Activity, and Time Stamp are needed[8].

D. Heteroscedasticity

Heteroscedasticity is the opposite of homoscedasticity, which is a condition where there is an inequality of variants of the error for all observations of each independent variable in the regression model. On the other hand, the notion of homoscedasticity is a condition in which there is a similarity of variants of the error for all observations of each independent variable in the regression model. If the variance of the residual value from one observation to another is constant, it is called Homoscedasticity[11].

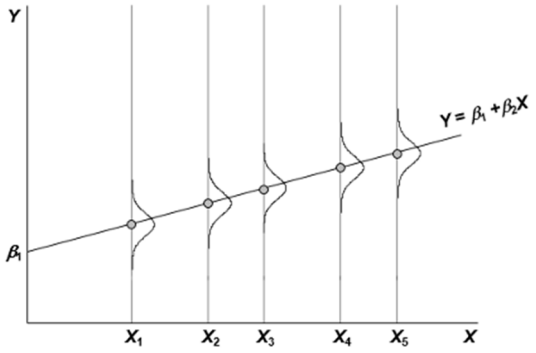

Fig. 3 Homoscedasticity concept[11]

Heteroscedasticity is the opposite of homoscedasticity, which is a state (an indicator) where the variance of the error is inequality for all observations of each independent variable in the regression model[12]. Why is it necessary to detect heteroscedasticity? The answer is to find out if there is a deviation from the classic assumption requirements in linear regression, where the regression model must be met with the conditions for the absence of heteroscedasticity. Figure 4 is a pattern of variance of heteroscedasticity.

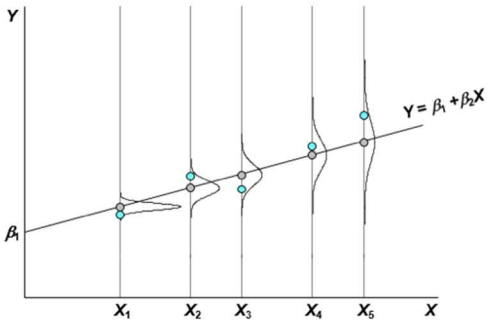

Fig. 4 Heteroskedasticity concept[11]

Heteroscedasticity is part of the classical assumptions Linier Regression in statistics. This study uses the Glejser approach to look for heteroscedasticity. The Glejser test is used because it has more stringent requirements than other tests in detecting heteroscedasticity [13]. The steps of the Glejser test are as follows[14]:

(Step 1) Perform an initial Linear analysis and use the sample regression equation

$$\hat{Y} = b_0 + b_1 X_{1i} \quad (1)$$

To obtain the residual

$$e_i = Y_i - \hat{Y}_I \quad (2)$$

For each of the n observations.

(Step 2) Then perform a second regression analysis using the absolute value of the remainder of $|e_I|$ as the dependent variable and the initial predictor variable as the independent variable. The secondary regression equation is

$$|\hat{e}_i| = b_0^* + b_1^* X_{1i} \quad (3)$$

(Step 3) Use the secondary regression analysis to obtain the test statistic tGMS = b_1* /Sb_1*, the ratio of the slope to its standard error, which enables a t-test to the population slope B_1* from the secondary analysis. Under the Null hypothesis, the Glejser /Mendenhall-Sincich test Statistic tGMS follows at distribution with n – 2 degrees of freedom.

(Step 4) Based on the distribution with n – 2 degrees of freedom in the GMF test statistic, it can be concluded that heteroscedasticity occurs or does not occur

E. Inductive Miner Algorithm

The Inductive Miner algorithm is an algorithm built to improve the performance of Alpha Miner and Heuristics Miner. The inductive miner algorithm guarantees the process model has a good fitness value [15]. The basic concept of this algorithm is to find the separations that occur in the event log such as sequential, parallel, concurrent, and loop. After finding the splits, the algorithm is repeated on the sub-log (found by applying the splits) until a fundamental model is found in the cases under investigation. The use of an inductive mining algorithm is also considered an important factor capable of analysing the overall activity.

F. Methods

The methods proposed in this study are as shown in Figure 5.

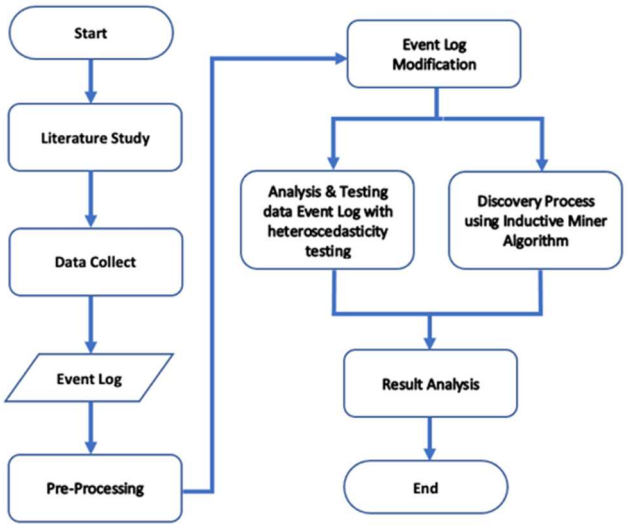

Fig. 5 Research Diagram

Figure 5 shows the method used which focuses more on the pre-processing process and event log analysis with heteroscedasticity testing. The process discovery using inductive Miner Algorithm is a comparison to the results of heteroscedasticity testing.

IV. RESULT AND DISCUSSION

A. Data Collect

In this study, the event log data collection process was carried out through some experimental data that was deliberately made based on real life data. The purpose of doing this is to demonstrate the concept of the relationship between event log analysis and the process model. There are 2 types of event logs, Type I is an event log that has a trace with normal activity times and Type II is an event log that has an abnormal activity time. The two event logs are based on the business process in the Figure 6.

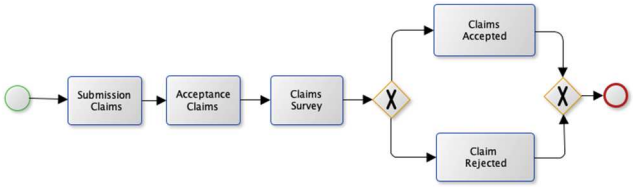

Fig. 6 Insurance Claim Process

B. Event Log Pre-Processing & Modification

After obtaining the event log as simulation material, the event log modification is carried out through the Disco application to obtain the variables needed in the calculation process. The Disco application performs event log extraction as shown in Figure 7.

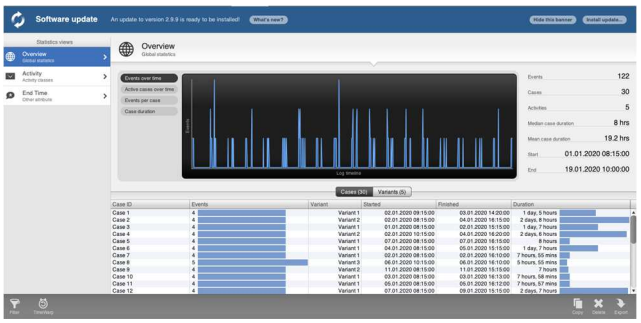

Fig. 7 get data & Extract From Disco Application

Then perform the extraction to get the variable Evens and Time Duration as testing variables. Obtained the following data in Table 2.

TABLE 2 THE RESULTS OF EXTRACTING DATA FROM THE DISCO APPLICATION

Case ID	Events	Duration	Duration (seconds)
Case 1	4	1 day, 5 hours	104700
Case 2	4	2 days, 8 hours	201600
Case 3	5	1 day, 7 hours	111600
Case 4	4	2 days, 6 hours	194700
Case 5	3	8 hours	28800

Therefore, it is necessary to modify the event log data extracted from the Disco application. From the results of data extraction, it can be determined that the independent variable

is Events (X_i) and the dependent variable is Duration (Y_i). Both of these variables are needed in heteroscedasticity testing.

C. *Analysis & Testing data Event Log with heteroscedasticity testing*

The Glejser test is performed by regressing between the independent variables and their absolute residual values (ABS_RES). If the significance value (p-value) between the independent variables and the absolute residual is more than 0.05, there is no heteroscedasticity problem. The testing process is carried out on both types of event logs. This research uses the help of the SPSS version 25 application, here are the results of the calculation of the heteroscedasticity test for event log Type I that has traces of normal activity. The output results can be seen in Table 3.

TABLE 3 EVENT LOG TYPE I TEST RESULTS (COEFFICIENTS[A])

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	30576.923	8664.876		3.529	.001
	events	-1450.385	2089.287	-.130	-.694	.493

a. Dependent Variable: duration

The result of the p-value from Table 3 shows a value of 0.493 > 0.05, which means that the $Y = a + bX + \epsilon$ model does not have a heteroscedasticity problem. Then do the same thing with the event log Type II which has traces with abnormal activity times. The output results can be seen in the Table 4.

TABLE 4 EVENT LOG TYPE II TEST RESULTS(COEFFICIENTS[A])

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	277340.816	98061.797		2.828	.009
	Events	-55078.163	24068.321	-.397	-2.288	.030

a. Dependent Variable: duration

The result of the p-value shows a value of 0.03 < 0.05, which means that the $Y = a + bX + \epsilon$ model does have a heteroscedasticity problem. To see the testing trend, it is necessary to carry out repeated testing of the both event logs. Repeated testing was carried out with different confidence levels, namely 90% and 99%. The results of the Heteroscedasticity Trend can be seen in the Table 5.

TABLE 5 TREND OF HETEROSCEDASTICITY TESTING WITH INTERVAL CONFIDENCE 90%, 95% AND 99%.

No	Interval Confidence	p-Value Indicators
Type I(without Heteroscedasticity)		
1	90%	$0.510 > \propto$
2	95%	$0.493 > \propto$
3	99%	$0.493 > \propto$
Type II(with Heteroscedasticity)		
1	90%	$0.03 < \propto$
2	95%	$0.03 < \propto$
3	99%	$0.03 < \propto$

The trend results show that the Type I model with heteroscedasticity problems has a consistent p-value, while the Type II model which does not have heteroscedasticity has a change in value but it is not too significant.

D. *Process Discovery using Inductive Miner Algorithm*

The next step is to carry out a discovery process using the ProM application based on the Type I event log using the Inductive Miner Algorithm. The resulting process modelling uses Petri-Net. The next stage is to test the performance of the resulting Petri-Net process model. The performance test results can be seen in the Figure 8.

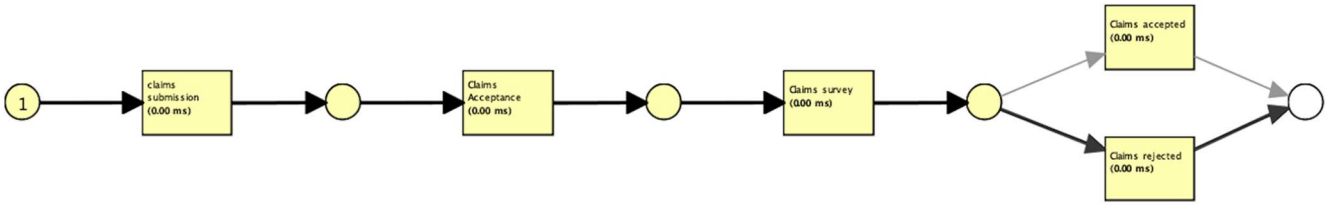

Fig. 8 Process model using Inductive Miner from event log without heteroscedasticity

In the same way, the discovery process is carried out on the Type II event log using the same algorithm. Performance test results can be seen in the Figure 9.

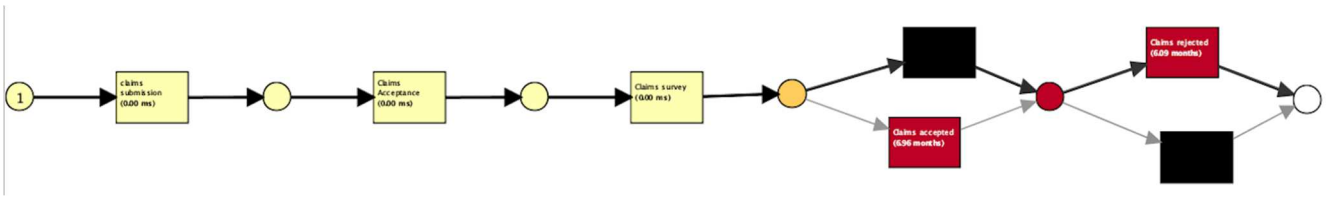

Fig. 9 Process model using Inductive Miner from event log with heteroscedasticity

In the process model generated from the type II event log which has heteroscedasticity problems, there are activities that have red colour on Accepted Claims and Rejected Claims Activities. The red colour indicates a bottleneck condition in the model process. Meanwhile, the process model generated from the Type I event log does not have a bottleneck condition.

E. Result Analysis

The distribution of duration data from Type I event logs that do not have heteroscedasticity problem shows more even variance and tends to be uniform. This can be shown in Figure 10.

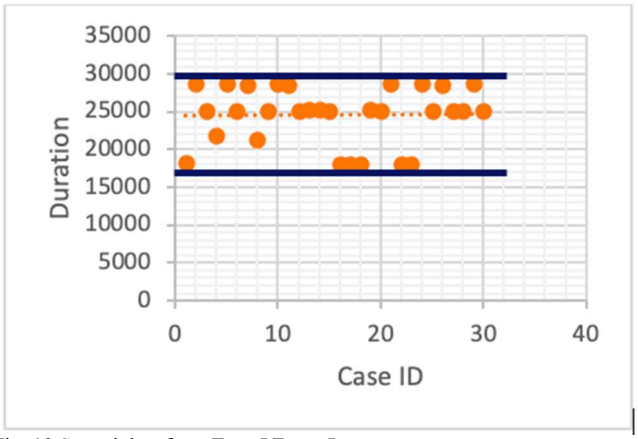

Fig. 10 Spread data from Type I Event Log

Meanwhile, the distribution of data from the Type II event log which has heteroscedasticity problems has variance that tends to be non-uniform. This can be shown in Figure 11.

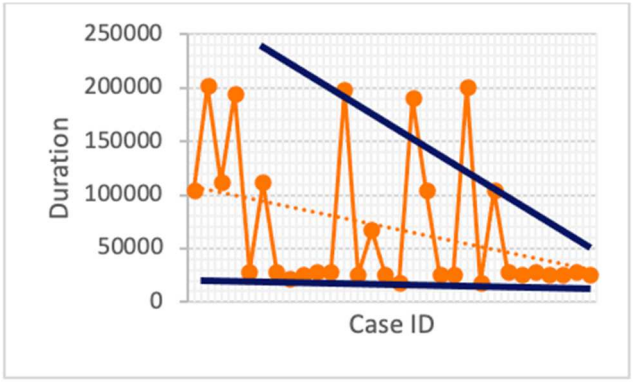

Fig. 11 Spread data from Type II Event Logs

V. CONCLUSIONS

Based on the results of the study, it can be concluded that the Type I Event Log after being tested with the Glejser approach has a p-value of 0.493, which means with a p-value greater than the error rate $\propto$=0.05, that mean the event log type I model does not have a heteroscedasticity problem. Meanwhile, Event log type II has a p-value of 0.03 where the significance value is smaller than the error rate $\propto$=0.05 which means that the event log type 2 model has a heteroscedasticity problem. The results of the heteroscedasticity testing trend show that the event log type II which has heteroscedasticity problems has a consistent p-value when tested with an error rate $\propto$ of 10%, 5% and 1% compared to the type I event log.

The Log Type II event model which has a heteroscedasticity problem has a process model that contains a bottleneck in the insurance claims discovery process using the inductive miner algorithm. Meanwhile, the Type I event log model which does not have heteroscedasticity problems by modeling the process using the inductive miner algorithm does not have a bottleneck condition. Thus, analysis of the event log data using a statistical approach can complement the evaluation of process modeling based on the process discovery.

REFERENCES

[1] S. Sadiq, P. Soffer, and H. Völzer, "Business Process Management" *(LNCS 8659)*, vol. 8659. 2014.

[2] M. Dumas, M. La Rosa, J. Mendling, and H. A. Reijers, "Fundamentals of Business Process Management". 2013.

[3] W. Van Der Aalst *et al.*, "Process mining manifesto," *Lect. Notes Bus. Inf. Process.*, vol. 99 LNBIP, no. Part 1, pp. 169–194, 2012, DOI: 10.1007/978-3-642-28108-2_19.

[4] W. M. P. van der Aalst, "Process discovery from event data: Relating models and logs through abstractions", *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 3, pp. 1–21, 2018, DOI:10.1002/widm.1244.

[5] N. Tax, N. Sidorova, and W. M. P. van der Aalst, "Discovering more precise process models from event logs by filtering out chaotic activities", *arXiv*, pp. 107–139, 2017.

[6] W. M. P. Van Der Aalst and B. F. Van Dongen, "Discovering Petri nets from event logs," *Lect. Notes Comput. Sci. (including Subser . Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7480 LNCS, pp. 372–422, 2013, DOI: 10.1007/978-3-642-38143-0_10.

[7] X. Lu, D. Fahland, F. J. H. M. van den Biggelaar, and W. M. P. van der Aalst, "Detecting deviating behaviors without models," *Lect. Notes Bus. Inf. Process.*, vol. 256, pp. 126–139, 2016, DOI: 10.1007/978-3-319-42887-1_11.

[8] I. H. Kwon, "Book Review: Process Mining: Discovery, Conformance and Enhancement of Business Processes", vol. 20, no. 2. 2014.

[9] W. Van Der Aalst, "Using process mining to bridge the gap between BI and BPM," *Computer (Long. Beach. Calif).*, vol. 44, no. 12, pp. 77–80, 2011, DOI: 10.1109/MC.2011.384.

[10] W. M. P. van der Aalst, "Business Process Management: A Comprehensive Survey," *ISRN Softw. Eng.*, vol. 2013, pp. 1–37, 2013, DOI: 10.1155/2013/507984.

[11] O. L. O. Astivia and B. D. Zumbo, "Heteroskedasticity in multiple regression analysis: What it is, how to detect it and how to solve it with applications in R and SPSS, Pract. Assessment, Res. Eval.", vol. 24, no. 1, pp. 1–16, 2019.

[12] P. J. Rosopa, M. M. Schaffer, and A. N. Schroeder, "Managing heteroscedasticity in general linear models," *Psychol. Methods*, vol. 18, no. 3, pp. 335–351, 2013, DOI: 10.1037/a0032553.

[13] Machado, José AF, and JMC Santos Silva. "Glejser's test revisited." Journal of Econometrics 97.1 (2000): 189-202.

[14] H. Glejser, "A New Test for Heteroskedasticity Author (s): H . Glejser Source : Journal of the American Statistical Association , Vol . 64 , No 325 (Mar ., 1969), pp . 316

[15] A. Bogarín, R. Cerezo, and C. Romero, "Discovering learning processes using inductive miner: A case study with learning management systems (LMSs)," Psicothema, vol. 30, no. 3, pp.322–329, 2018, DOI: 10.7334/psicothema2018.116.