ICIC Express Letters
Part B: Applications
Volume **7**, Number **11**, November **2016**

ICIC International ©2016 ISSN 2185-2766
pp. 2323–2329

STATISTICAL VERIFICATION OF PROCESS MODEL FITNESS IN PROCESS MINING

Sung-Hyun Sim, Yulim Choi and Hyerim Bae*

Department of Industrial Engineering
Pusan National University
2, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Korea
{ ssh; flamethrower }@pusan.ac.kr; *Corresponding author: hrbae@pusan.ac.kr

Received April 2016; accepted July 2016

Abstract. *The primary purpose of process mining is to explore a process model from an event log and analyze it in order to suggest enhancements to the process. Evaluation of the conformance of process models is of great importance in this regard. However, due to their large data size and complex structure, this is not easy. Previous studies on conformance checking have applied fitness measuring methods that use token replay and node-arc relations based on Petri net. Fitness thus far has not considered statistical significance, but just offers a numeric ratio. We herein propose a statistical fitness test based on the Kolmogorov-Smirnov test to formulate statistical process model fitness guidelines and conformance parameters for model selection. We also propose a new concept of 'maximum confidence dependency' to solve the problem of the trade-off between model abstraction and process conformance.*

Keywords: Process mining, Conformance checking, Kolmogorov-Smirnov test, Process model selection, Maximum confidence dependency

1. **Introduction.** These days, multiple suppliers and customers make process models complex. In order to cope with such complexity, information system managers want to make their process models be simplified. However, simplification reduces process conformance, which means that the resultant model explains only a small portion of real process executions [1]. In the process mining [2] field, this decision making process is closely related with process model quality, and is typically expressed as conformance checking [3]. When we check conformance of large scale data, it is regarded hard to handle because this data usually generates a spaghetti process model [7]. The easiest approach to the analysis of such data is to simplify the process model for effective representation by controlling the size of the activity or path. For this purpose, process mining researchers have adopted the concept of fitness measure [3], which replays tokens on process model to check conformance. In this paper, we develop a statistical equality test that provides insight into the process model's abstraction level and its reflection of the original dataset. We chose to apply the Kolmogorov-Smirnov test [4], to evaluate a process model's conformance with the original dataset. For the conformance testing of a process model, this proceeds in four stages. First: discover the process model. Second: play-out the log from the process model. Third: perform an equality test between the original log and the comparative log (extracted from the model). Fourth: analyze the statistical significance. By this procedure, we can develop a better conformance-checking method according to changes of dataset size and/or complexity. The paper is configured with five chapters. The first and the second explain basis. The third is statistical proof and explanation about the algorithm. The fourth is experiment to compare the original one and the proposed one. The fifth concludes the paper.

2. **Conformance Checking.** In process mining, certain index values are used to evaluate the discovered model. It is important to judge whether the process model derived from an event log is a proper model having sufficient conformance. The existing indicators of process model quality are fitness, precision, and generalization [6]. The most commonly utilized indicator of conformance checking is fitness. The fitness calculation equation [6] is

$$fitness(L, M) = 1 - \frac{fcost(L, M)}{move_L(L) + |L| * move_m(M)} \tag{1}$$

All the notations on Equation (1) are defined at Buijs et al. [6]. We used the fitness for comparison with our methodology. Previous research on the evaluation of process model conformance to the event log has entailed checking the node-arc relation or using log replay through token play.

3. Statistical Method for Goodness-of-Fit of Heuristic Process Model.

3.1. Kolmogorov-Smirnov test procedure.

The Kolmogorov-Smirnov test [4] evaluates goodness-of-fit based on the theorem that if two continuous observations' cumulative distribution functions are equal, the observation's probability density function also is equal. According to [5], let $F(x)$ be the population distribution function and $F_0(x)$ be the specific distribution function. Then, the hypothesis test is

$$\begin{aligned} H_0 &: F(x) = F_0(x) \text{ for every } x \\ H_1 &: F(x) \neq F_0(x) \text{ for some } x \end{aligned} \tag{2}$$

The following is the Kolmogorov-Smirnov test procedure [4,8].

Step 1. *Let probability sample $X_1, \ldots, X_n$ be an empirical distribution function $F(x)$.*

Step 2. *Test statistics, $D = sup_x\{|F_0(x) - F(x)|\}$.*

Step 3. *If $D > d\left(\frac{\alpha}{2}, n\right)$, the null hypothesis is rejected. In $d\left(\frac{\alpha}{2}, n\right)$, α represents the upper-bound 100α percentile, and n is the sample size.*

3.2. Statistic for goodness-of-fit test of heuristic process model.

The following is a four-step statistical method for evaluation of goodness-of-fit. We define pre-processed data as original log data and played-out data as comparative log data.

Step 1. *Prepare a process model from original* log(L_0) *by discovering it using heuristic miner (play-in)* [9].

Step 2. *Extract activity occurrence probability vector from original* log *and prepare comparative* log(L_1) *by artificially executing discovered model (play-out).*

Step 3. *Extract empirical cumulate distribution function for original* log *and comparative* log.

Step 4. *Use Kolmogorov-Smirnov test to compare original* log *with comparative* log *for conformance.*

Suppose that m activities occur in L_0 and n activities are found in L_1. Let $Pr(a_i)$ be the probability of the ith activity $(1 \leq i \leq m)$ in L_0, and let $Pr(a'_j)$ be the probability of the jth activity $(1 \leq j \leq n)$ in L_1.

$$\begin{aligned} E &= \{E_1 = Pr(a_1), \ldots, E_m = Pr(a_m)\}^T \\ H &= \{H_1 = Pr(a'_1), \ldots, H_n = Pr(a'_n)\}^T \end{aligned} \tag{3}$$

Let the activity occurrence probability vector for the original log and comparative log be $(E_1, \ldots, E_m)$, $(H_1, \ldots, H_n)$ and let its order statistics be $(E_{(1)}, \ldots, E_{(m)})$, $(H_{(1)}, \ldots, H_{(n)})$ respectively. We define $L_0(x)$ as the activity occurrence probability for the empirical distribution of the original log data and $L_1(x)$ as the activity occurrence probability for the empirical distribution function of the comparative log, where x is a variable of activity occurrence probability

$$\begin{aligned} &E_{(1)}, \ldots, E_{(m)} \sim L_0(x) \quad L_0(x) = \sum_{i=1}^{m} \frac{I_{(0,x)}(E_{(i)})}{m} \\ &H_{(1)}, \ldots, H_{(n)} \sim L_1(x) \quad L_1(x) = \sum_{j=1}^{n} \frac{I_{(0,x)}(H_{(j)})}{n} \end{aligned} \tag{4}$$

$$I_{(0,x)}(t) = \begin{cases} 1, & 0 \le t \le x \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

In Equation (5), $I_{(0,x)}(t)$ denotes the number of observations satisfying $0 \le t \le x$ when a specific t is selected in the domain of definition. Then, the hypothesis test for process model goodness-of-fit is

$$\begin{aligned} &H_0\colon\ L_0(x) = L_1(x) \text{ for every } x \\ &H_1\colon\ L_0(x) \neq L_1(x) \text{ for some } x \end{aligned} \tag{6}$$

We define our test statistics as

$$D_{th} = sup_x\{|L_0(x) - L_1(x)|\} \tag{7}$$

After the process model is discovered, m and n are determined. A parameter α also can be determined, according to user preference. After that, we test D_{th} according to the Kolmogorov-Smirnov test statistics in the Kolmogorov-Smirnov table. Using these statistics, D_{th} is checked as to whether it follows the inequality $D_{th} > d\left(\frac{\alpha}{2}, m, n\right)$. Then based on the result, we will decide whether to accept hypothesis H_1. Algorithm 3.1 is a pseudo code that calculates conformance using the Kolmogorov-Smirnov test.

Algorithm 3.1. KS-Conformance

```
Algorithm KS-Con (Matrix AOPV[][])
  Input: Matrix AOPV[m][2]
           //Activity Occurrence Probability Vector original and comparative log
  Output: Boolean rht //Result of Hypothesis Test
           function Kss (float α/2, int i, int j)
             //Kolmogorov-Smirnov Statistics from a given table
             int m//number of activity in original log
             int n//number of activity in comparative log
             float loS //Level of significance
               Matrix SUM[m][2]
  for ($i \leftarrow 2$ to $m; i++$) {
             SUM[1][1] ← AOPV[1][1]
             SUM[i][1] ← AOPV[i][1] + SUM [i – 1][1]
             SUM[1][2] ← AOPV[1][2]
             SUM[i][2] ← AOPV[i][2] + SUM [i – 1][2]
             Matrix KSD[i][1] ←SUM [i][1] – SUM [i][2]}
             int comp ← KSS(loS,m,n)
  if (max(KSD[]) ≦ comp) rht ← 1
  else rht ← 0
  return rht
```

3.3. **Maximum confidence dependency.** One of our purposes in this paper is to suggest a threshold value called maximum confidence dependency (MCD). The procedure for

MCD determination is related to analysis of the maximum abstraction level with statistical confidence. When the Kolmogorov-Smirnov statistics satisfy Equation (8), we say that the process model satisfies goodness-of-fit with the original log data

$$D_{th} \leq d\left(\frac{\alpha}{2}, m, n\right) \tag{8}$$

Among the values of D_{th} satisfying the above equation, the maximum value is defined as the MCD.

Algorithm 3.2. Maximum Confidence Dependency

```
Algorithm MCD //Maximum Confidence Dependency
  Input Matrix AOPV [m][n]
              //Calculate Activity Occurrence Probability
  Output int mcd //Maximum Confidence Dependency

  Matrix DTT[101]
          do (i <- 100 to 0; i--){
                  DTT[i] <-KS-Con(AOPV)
          }while (DTT[i] = 1)
  return i + 1
```

3.4. **Algorithm performance.** To verify the performance of our algorithm, an experiment was carried out to measure the time required for a different number of cases. We repeated each case 1,000 times using artificial experiment data, as shown in Figure 1. We were able to observe a bell-shaped curve with a low standard deviation value, which confirmed the adequate reliability of our experiment.

Furthermore, this result shows that our algorithm does not incur a proportional increase of processing time with increasing case number. Table 1 summarizes the result.

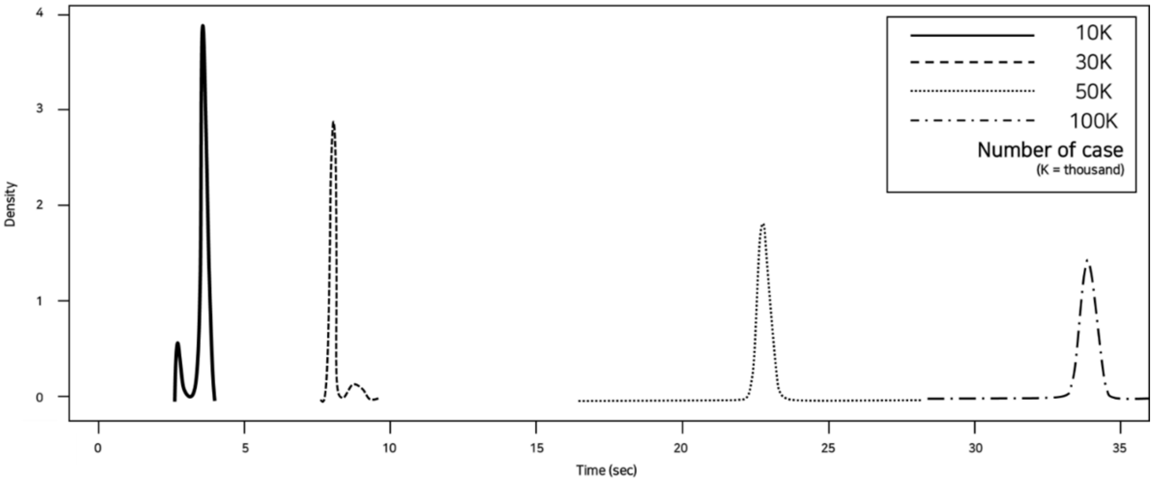

Figure 1. Performance of the algorithm with the number of cases

Table 1. Result of time analysis for algorithm performance

System time (sec)		
Case size	*Mean*	*Standard deviation*
10,000	2.61	0.287
30,000	7.28	0.461
50,000	23.23	1.172
100,000	34.11	0.818

4. **Experiment.** In order to validate our approach, we carried out experiment using real data from steel manufacturing company. We evaluate methodology by using p-test and CDF (cumulate distribution functions). Data contains 10313 cases, 89226 events and 11 activities, which were generated in the course of the execution of steel manufacturing processes within a Korean company. We used heuristic miner [9] to obtain a process model. To simplify the process model, we applied a high dependency threshold to generating abstract process model. In this experiment, we adjusted the abstraction level by selection of an appropriate dependency threshold.

Table 2 shows the result of a Kolmogorov-Smirnov verification test to verify the model with two dependency thresholds: 0.97 and 0.98. With the dependency threshold of 0.97, the p-value is exactly 1, which means that the model almost perfectly fits the original log data. However, with the dependency threshold of 0.98, the p-value is 2.2^{-16}, which means that H_0 cannot be supported statistically. In this case, we cannot say that the process model follows its original dataset. Notwithstanding the similarity of the fitness values for the two cases, there is a serious difference in statistical significance.

TABLE 2. Steel manufacturing process data conformance checking result

Dependency	*Test statistic*	*P-value*	*Fitness*
0.97	*0.0019*	*1*	*0.8723*
0.98	*0.0687*	*2.2e-16*	*0.8635*

Figure 2 and Figure 3 are cumulate distribution functions (CDFs) of the original and comparative datasets, respectively. They show the difference between the two cases. When a process model conforms to an event log with a statistical significance of 0.99, as shown in Figure 2, the CDF curves of model and event log fit very well, whereas Figure 3 shows some gaps between them.

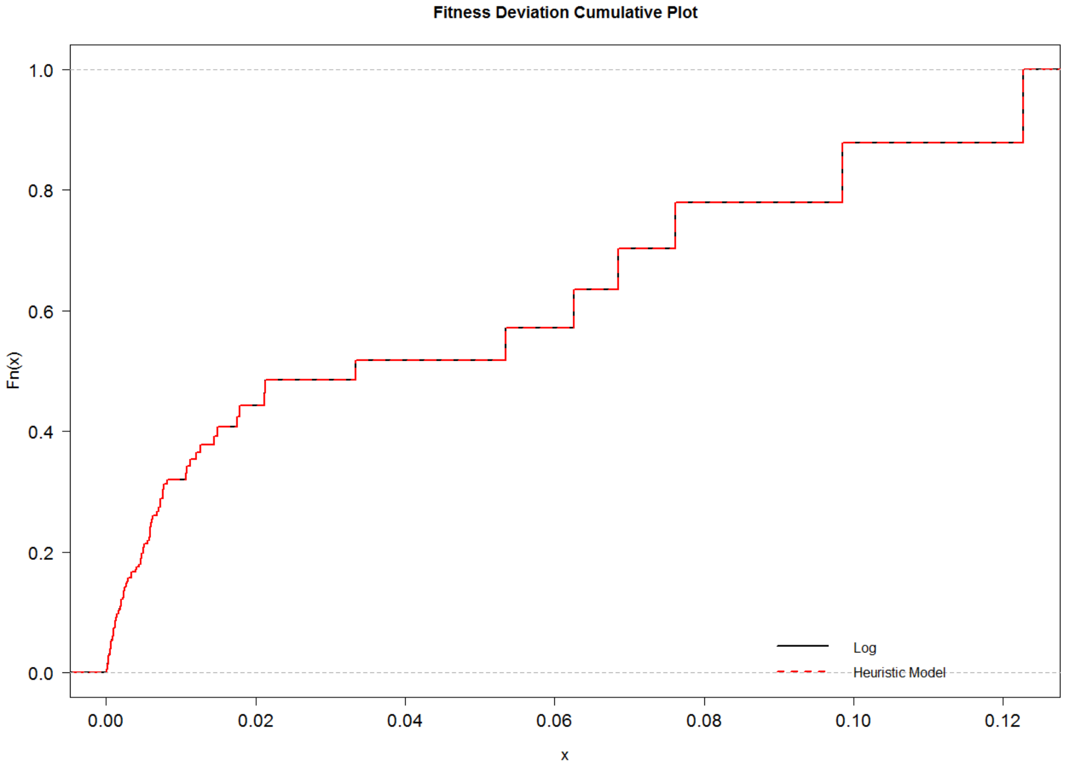

FIGURE 2. CDF at dependency threshold 0.97

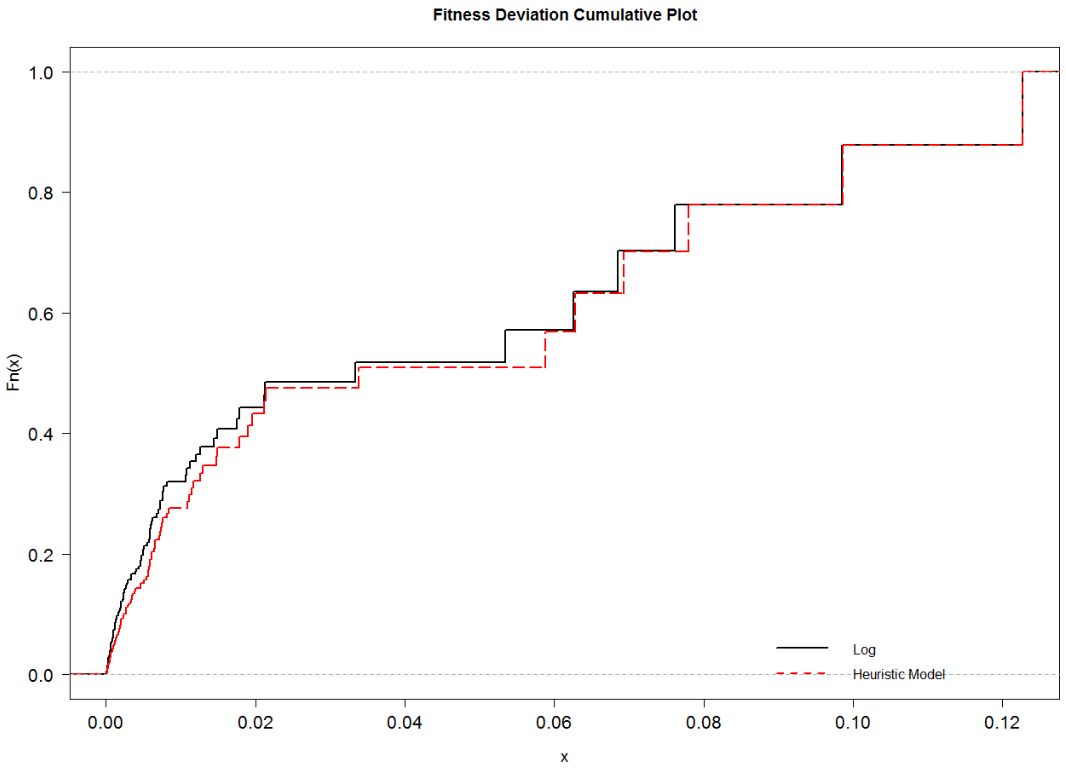

Figure 3. CDF at threshold dependency threshold 0.98

We define maximum confident dependency (MCD) as the largest dependency value which makes a process model conform event log to a given statistical significance. In the above case, MCD is 0.97 at 99% significance level.

5. Conclusion.

In this paper, we develop a method for checking the conformance of a model to the event log by comparison of the statistical variation of two datasets, namely the original log dataset and the log data set generated by playing out the process model. Additionally, we propose a method for checking process model quality by means of the concept of MCD, which is a threshold value for satisfaction of the equality condition between two log data sets according to a certain level of statistical significance. We expect that our approach will prove to be easily applicable for conformance checking of process models generated by process mining techniques. In this paper, we handle data which play-out from heuristic miner. However, in our further study for this methodology, we handle various process mining techniques to generalize our methodology and suggest optimal model.

Acknowledgment. This research was partially supported by the National Research Foundation of Korea (NRF) (No. NRF-2015R1D1A1A09061331) and the IT R&D Program of MOTIE/KEIT (Ministry of Trade, Industry and Energy/Korea Evaluation Institute of Industrial Technology; No. 10054514), with grant funded by the Korean government (MEST).

REFERENCES

[1] A. Rozinat, M. Veloso and W. M. P. van der Aalst, Evaluating the quality of discovered process models, *The 2nd Intl. Workshop on the Induction of Process Models*, Antwerp, Belgium, 2008.

[2] W. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, Springer Science and Business Media, 2011.

[3] A. Rozinat and W. M. P. van der Aalst, Conformance checking of processes based on monitoring real behavior, *Information Systems*, vol.33, no.1, pp.64-95, 2008.

[4] P. H. Kvam and B. Vidakovic, *Nonparametric Statistics with Applications to Science and Engineering*, John Wiley and Sons, 2007.

[5] P. Sprent and N. C. Smeeton, *Applied Nonparametric Statistical Methods*, CRC Press, 2007.

[6] J. C. A. M. Buijs, B. F. van Dongen and W. M. P. van der Aalst, On the role of fitness, precision, generalization and simplicity in process discovery, in *On the Move to Meaningful Internet Systems: OTM 2012*, 2012.

[7] W. M. P. van der Aalst and C. W. Günth, Finding structure in unstructured processes: The case for process mining, *The 7th International Conference on Application of Concurrency to System Design*, 2007.

[8] S. Facchinetti, A procedure to find exact critical values of Kolmogorov-Smirnov test, *Statistica Applicata*, vol.21, pp.337-359, 2009.

[9] A. J. M. M. Weijters, W. M. P. van der Aalst and A. A. de Medeiros, Process mining with the heuristics miner-algorithm, *Tech. Rep. WP 166*, Technische Universiteit Eindhoven, 2006.