Stochastic Process Model-Log Quality Dimensions: An Experimental Study

Adam Burke (✉)
Queensland University of Technology
at.burke@qut.edu.au

Sander J.J. Leemans
RWTH Aachen
s.leemans@bpm.rwth-aachen.de

Moe T. Wynn
Queensland University of Technology
m.wynn@qut.edu.au

Wil M.P van der Aalst
RWTH Aachen
wvdaalst@pads.rwth-aachen.de

Arthur H.M. ter Hofstede
Queensland University of Technology
a.terhofstede@qut.edu.au

***Abstract*—Stochastic process models are a type of model that explicitly include elements of probability in describing an organization, facilitating different modes of analysis and simulation. Having obtained models of an organizational process, say through process mining, using them well depends on understanding their quality, and being able to compare different models. There may not be a single optimal stochastic model for a process, but trade-offs between models, decided by their intended use. Reasoning about trade-offs in a precise way requires quantitative measures, and an understanding of how these measures relate, including whether they capture independent underlying properties.**

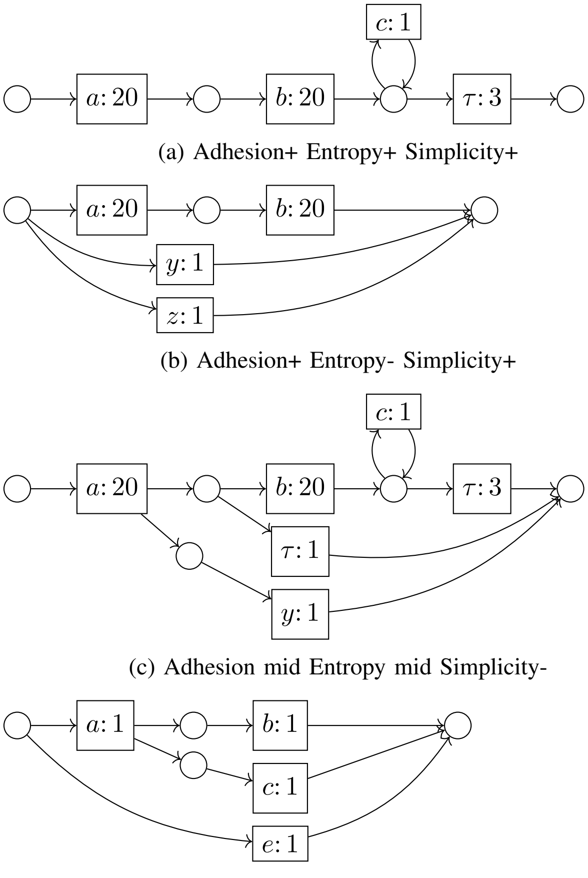
This paper is an empirical investigation of measures for stochastic process models built from real-life logs. The experimental design assembles a large collection of models built both randomly and by discovery techniques. A wide spectrum of candidate measures, drawn from and inspired by the process mining literature, are applied using these models. Based on this analysis, three stochastic quality dimensions are proposed: adhesion, entropy and simplicity.

***Index Terms*—stochastic process mining, process conformance, Stochastic Petri Nets, adhesion, entropy, simplicity**

I. Introduction

The everyday behaviour of an organization - and hence much of its impact in the world - depends on people and systems in the organization. When the actions of such human and machine "street-level bureaucrats" [1] are recorded in information systems, they can be transformed into models of the organization through process mining [2]. The notion of whether something is routine or unusual is captured explicitly with probability weightings in stochastic process models. Formalized models, such as stochastic Petri nets [3], can be used for process analysis and improvement, to investigate performance, or as a component of business process simulation. This automated discovery and analysis is the focus of stochastic process mining.

Process model quality measures, such as proportion of replayable traces, or number of model components, have been defined to evaluate and guide model construction, as part of the study of process conformance [2]. In (non-stochastic) control-flow mining, the many quality measures that exist are organized according to four quality dimensions: fitness, precision, simplicity and generalization [2, p188]. This scheme allows statements like "this discovery technique trades off precision for simplicity" to be backed up with empirical evidence from measures for those dimensions.



(a) Adhesion+ Entropy+ Simplicity+

(b) Adhesion+ Entropy- Simplicity+

(c) Adhesion mid Entropy mid Simplicity-

(d) Adhesion- Entropy+ Simplicity mid

Fig. 1: Models exemplifying adhesion and entropy variations relative to log $L_E = [\langle a,b\rangle^{20}, \langle a,b,c\rangle^{2}, \langle a,b,c,c\rangle^{1}, \langle e,f\rangle^{1}]$.

Using established control-flow process mining dimensions as a starting point, in this paper, we investigate the question of *what dimensions may describe the quality of stochastic process models*. The *mathematical* space being explored is not purely analytical, as the underlying event logs to which models are compared are real-life empirical data on the social behaviour of organizations. This suggests using an exploratory quantitative analysis. To conduct it *empirically*, we collected and evaluated stochastic process models for real-life processes. Models were tested against a set of computationally cheap measures, based mainly on log-model comparisons. Further testing was performed with two stochastic process conformance measures [4], [5]. The process model collection consisted of Stochastic Labelled Petri Nets (SLPNs) [4] and was constructed by both random generation and discovery techniques. Existing stochastic process discovery techniques [6], [7], [8] were leveraged, and supplemented by the use of a new genetic miner, Stochastic Evolutionary Tree Miner (SETM).

This empirical data was analysed for correlation and principal components (PCA), discovering three components. We name these as stochastic process model quality dimensions: *adhesion*, *entropy* and *simplicity*. Adhesion represents how little change is required to modify one process into another. Figure 1 previews these dimensions with multiple models for an example event log; more detailed discussion is in Section V.

This paper contributes 1) An experimental design for investigating stochastic process quality measure relationships 2) measures supporting that investigation 3) proposed model quality dimensions. After formal preliminaries in Section II, experimental design is explicated in Section III, including measures and model generation. Section IV is a quantitative analysis, with dimensions named in Section V. Section VI discusses related work, and Section VII concludes.

II. Preliminaries

Petri net and process mining definitions are used throughout (see [3], [2, p80], [4]). Sequences are shown as $\langle a_1, \ldots, a_n \rangle$ and concatenation as $+$. The set of multisets (bags) over type C is $\mathcal{B}(C)$, and real-valued multisets are $\mathcal{B}^+(C)$. Multiset union and intersection are $\sqcup$ and $\sqcap$ respectively.

Definition II.1 (Activities and Event Logs)**.** Let A be a set of activities in a process, and A^* the possible sequences of those activities. A *trace* $\sigma \in A^*$ is a sequence of activities. *Event logs* are multisets of traces $\mathcal{B}(A^*)$.

$\mathcal{L}$ is the set of all logs. $|L|$ is the number of traces in a log $L \in \mathcal{L}$, and $||L||$ the number of events. The number of cases matching trace σ in log $L \in \mathcal{L}$ is denoted $L[\sigma]$.

Definition II.2 (Stochastic Language)**.** A *stochastic language* Θ for traces over activity set A is a function $\Theta : A^* \to [0, 1]$ which denotes a probability for each trace, and sums to unity.

Definition II.3 (Petri nets)**.** A *Petri net* is a tuple (P, T, F, M_0) of places P, transitions T, flow relation $F \subseteq (P \times T) \cup (T \times P)$ and initial marking M_0. Markings are multisets of places $M \in \mathcal{B}(P)$ indicating a state of the Petri net.

The first node in F represents an incoming node and the second an outgoing. Transitions are *enabled* when every incoming place contains a token. Transitions *fire*, changing the state of the net by consuming incoming tokens and producing tokens for outgoing places.

Petri nets can be extended to model probabilities. To avoid ambiguity, we refer to nets without probability constructs as *place-transition nets*, following [3].

Definition II.4 (Stochastic Labelled Petri Net)**.** An *SLPN* [4] is a tuple $(P, T, F, M_0, W, \lambda)$ such that (P, T, F, M_0) is a place-transition net. A weight function $W \colon T \to \mathbb{R}^+$ assigns each transition a weight. Labelling function $\lambda \colon T \to A \cup \{\tau\}$ then provides a mapping from transitions to a symbol library of activities A. τ is a silent label where $\tau \notin A$.

When transitions $T_e \subseteq T$ are enabled in a particular marking, a transition $t \in T_e$ fires according to the probability given by $\frac{W(t)}{\sum_{t' \in T_e} W(t')}$. The sequences of labels generated by a series of transitions through the model forms a trace, and the collection of such traces and their probabilities is the SLPN's stochastic language. We assume traces end in deadlock.

Figure 1 shows example SLPNs. We define the set of all SLPNs as $\mathcal{G}$. SLPNs are a labeled variant of *Stochastic Petri Nets* (SPNs) [3] and Generalized SPNs (GSPNs) [3].

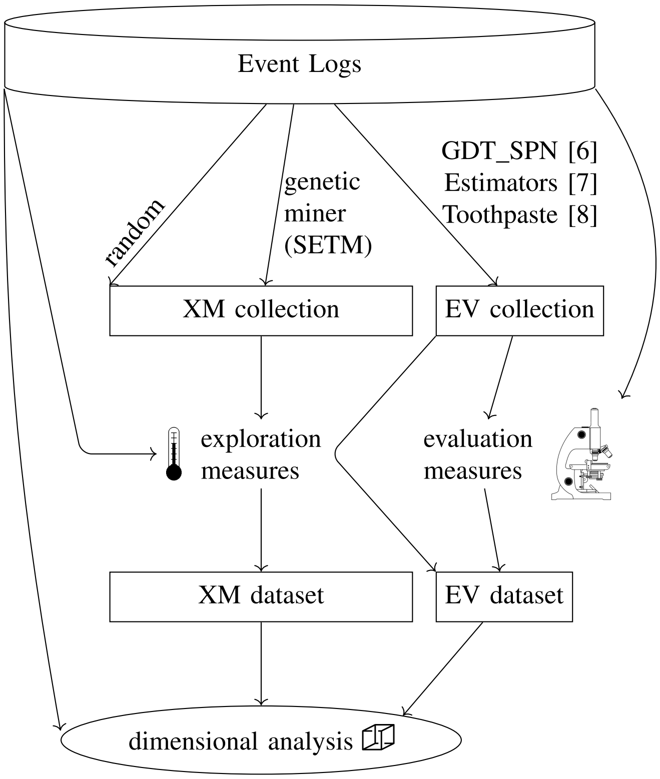


Fig. 2: Experiment design for investigating stochastic quality dimensions, from logs to models, measures and analysis.

III. Experiment Design

Figure 2 illustrates the experiment design. A large number of models were generated from event logs, then a variety of measures calculated. The resulting collection was then analyzed for possible model and measure relations. Models were collected using random generation, a genetic miner (SETM), and existing stochastic process discovery techniques. This diversity of models allows for variation in model quality across different measures. Measures were partitioned into experimental and computationally cheap *exploration measures*, and already existing, more expensive, *evaluation measures*. The use of cheap measures made testing a large set of models practical. The full set of models, called the *XM collection*, were measured using exploration measures. The discovery technique subset, called the *EV collection*, was also measured using evaluation measures. The measure results for these models are termed the XM and EV *datasets* respectively.

Firstly, this section reviews existing evaluation measures, to both explain their application here, and familiarize readers with measures specific to stochastic process models. Secondly, we introduce a technique for generating a play-out log for an SLPN model. Thirdly, play-out logs and other techniques are used to define a number of exploration measures, including new measures. Lastly, the techniques used for model generation are outlined, as well as the data collection procedure.

A. Evaluation Measures

Conformance measures for stochastic process models have been a focus of recent research. The three published techniques known to the authors are used here as evaluation measures.

The truncated Earth-movers' distance [5] (*tEMSC*) represents the shared probability mass between logs or stochastic models. This measure is calculated using both the string edit (Levenshtein) distance and the Earth-movers' distance across a reallocation matrix of paths in the stochastic languages of model and log. This experiment used a probability mass threshold of 0.8, shown as *tEMSC*0.8.

Entropy precision (H_P) and recall (H_F) [4] calculate the respective log-facing and model-facing components of a Stochastic Deterministic Finite Automata (SDFA). Both measures have a stronger theoretical basis for representing partial traces than the exploration measures in Table I. Uses have shown both success differentiating high-quality models and limitations on application to lower quality models [7]. Unsound, non-SDFA, or lower quality models may result in extended runtimes or non-termination.

B. Stochastic Play-out Logs

A number of exploration measures make use of play-out logs [2, p41], an established process mining technique for generating event log traces based on process models. For place-transition Petri nets, a standard way of generating play-out logs is by "playing the token game": noting the traces generated when the model advances from the initial marking through subsequent states. The token game with place-transition nets requires making arbitrary choices or introducing a probability function for choosing between enabled transitions. Stochastic models, such as SLPNs, already include explicit probability functions which define behaviour when multiple transitions are enabled. This can productively constrain the generation of possible play-out logs to a far smaller set of outputs. The play-out log can then substitute for the model when comparing other logs or models, allowing measurement of models which otherwise could not be practically included in the experiment. By using a finite representation to approximate the possibly infinite stochastic language of the model, a stochastic play-out log eliminates or greatly reduces the need for multiple samples to represent possible traces.

Definition III.1 (Play-out log)**.** A play-out log $L_p \in \mathcal{B}^+(A^*)$ is a positive real-valued multiset of traces.

In real-valued multisets, the count of a member is in $\mathbb{R}^+$, e.g., $[\langle a \rangle^{3.4}, \langle b, c \rangle^{2.0}]$. The set of all play-out logs is $\mathcal{L}^+ \supset \mathcal{L}$. The $\cdot$ operator scales all occurrence values by a factor.

Definition III.2 (Stochastic play-out log generator)**.** A log *generator* is a function $f \colon \mathcal{G} \times \mathbb{N} \rightarrow \mathcal{L}^+$, taking an SLPN and a target size, and giving a play-out log, such that $|f(g,k)| = k$.

Detail of implementation function spg is in Appendix A.

C. Exploration Measures

Eighteen exploration measures were used. To allow the analysis of a large set of models, computational cheapness was prioritized. Measures, summarised in Table I and detailed in Appendix B, are either existing measures in the literature, or simplifications of existing measures or dimensions. We found a play-out log target generation size of 1000 traces allowed for good granularity within reasonable execution times.

Definition III.3 (Measures and Play-out Measures)**.** A *measure* is a function comparing models and logs, $\mu \colon \mathcal{G} \times \mathcal{L}^+ \rightarrow [0,1]$. A *play-out measure* is a function comparing play-out logs and event logs, $\pi \colon \mathcal{L}^+ \times \mathcal{L}^+ \rightarrow [0,1]$.

Given a play-out measure π, a model-log measure μ_π is calculated from a model $g \in \mathcal{G}$, a play-out log size k, and an event log $L \in \mathcal{L}$ using $\mu_\pi(g,k,L) = \pi(\frac{|L|}{k} \cdot spg(g,k), L)$.

The experimental design concept for each measure, and categorization by existing control-flow dimensions, are summarized in Table I. The log-only measures trace count (LTC) and event count (LEC) were also captured.

D. Model Generation

A variety of models were included in the experiment to allow for the observation of more general relationships. Generation techniques were chosen to include different anticipated quality profiles, including low, moderate and higher quality models, using random generation and discovery techniques.

Each technique was applied to six public event logs, summarized in Table II. For XM (exploration), 18 exploration measures were calculated across each model, with sample size $n = 9301$. For EV (evaluation), a smaller set of discovery and estimated higher quality models had three additional evaluation measures calculated.

Random models were generated using *Probabilistic Process Trees (PPTs)* [8]. PPTs are a form of weighted process tree which correspond to a subset of SLPNs. To limit complexity, models larger than the arbitrary cutoffs of 1000 transitions or

TABLE I: Measures and their design rationale

Abbrv.	Measure Name	Design Concept
Exploration measures		
EMT	Earth Movers With Play-out Trace	Earth Movers
TMO	Trace Probability Mass Overlap	Probability Mass
TOR	Trace Overlap Ratio	Probability Mass
ARG	Activity Ratio Gower	Fitness
TRG2	Trace Ratio Gower length 2	Fitness
TRG3	Trace Ratio Gower length 3	Fitness
TRG4	Trace Ratio Gower length 4	Fitness
HIFT	Play-out Entropy Intersection Fitness	Fitness
HIPT	Play-out Entropy Intersection Precision	Precision
HJFT	Play-out Entropy Projection Fitness	Fitness
HJPT	Play-out Entropy Projection Precision	Precision
SSENC	Structural Simplicity by entity count	Simplicity
SSEDC	Structural Simplicity by edge count	Simplicity
SSS	Structural Simplicity incl. stochastic	Simplicity
TGF1	Generalization by Trace Floor count 1	Generalization
TGF5	Generalization by Trace Floor count 5	Generalization
TGF10	Generalization by Trace Floor count 10	Generalization
TGDU	Generalization by trace uniqueness	Generalization
Evaluation measures		
$tEMSC0.8$	Earth Movers truncated	Earth Movers
H_P	Entropy Precision	Precision
H_F	Entropy Recall	Fitness
Log measures		
LTC	Log Trace Count	Log
LEC	Log Event Count	Log

TABLE II: Event logs

Log	Traces	Variants	$\lvert A\rvert$	Domain
BPIC 2013 closed	1487	183	4	Issue tracking
BPIC 2013 incidents	7554	1511	3	Incident tracking
BPIC 2018 control	43808	59	7	EU Agriculture policy
BPIC 2018 reference	43802	515	6	EU Agriculture policy
Road Traffic Fines	150370	231	11	Italian policing
Sepsis	1054	846	16	Hospital diagnosis

a tree depth of 30 were excluded. Up to 1000 random models were generated for each log. Randomly generated models were anticipated to be of lower quality.

A novel genetic miner for discovering stochastic process models, the Stochastic Evolutionary Tree Miner (SETM), was implemented for this experiment. It is based on the Evolutionary Tree Miner [9]. The SETM generates random PPTs for the initial generation of models. Next, four possible mutations are applied: to add a node (including silent transitions), mutate a single node, remove a subtree, or remove useless nodes (specifically to apply Preserving Compression rules [8]). These mutations preserve valid and consistent tree weights. Models were exported as SLPNs.

Where used, the SETM was run across 1000 generations with a fitness function incorporating all eighteen exploration measures with equal weight. The fittest model in each generation was added to the XM (exploration) collection, generating a spectrum of models of moderate quality. The genetic miner yielded results for four of the logs in this experiment; the two logs with most activities did not yield results.

Models generated by existing discovery techniques were also included in the XM collection, and were anticipated to be of higher quality. Public implementations of stochastic process discovery techniques for GSPNs [6], [7], [8] created a further 103 models relating to the selected event logs. The EV collection comprised only these discovered models. The state of the art in stochastic discovery techniques still yield low-quality models in a number of cases, so there were broad ranges of values available for analysis of the EV dataset.

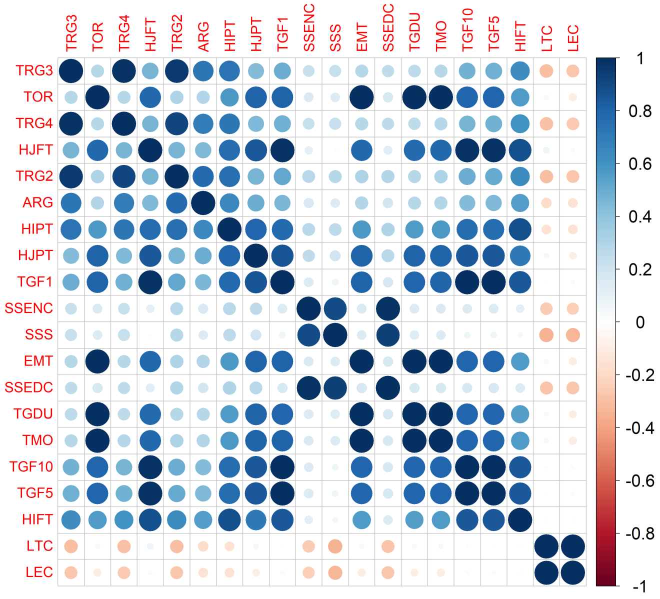


Fig. 3: Correlations between exploration measures and log measures on the XM and EV combined dataset, ordered to show groups of correlated measures.

Exploration measures and SETM were implemented in Java using the ProM framework[1]. Experiments were run on a Linux clustered data centre using 50 Gb of RAM.

IV. Quantitative Analysis

We performed an exploratory quantitative analysis on the dataset of model measures, discovering three orthogonal components, which were examined for robustness.

A. Analysis

Analyses of correlations and principal components [10] were performed to determine commonality, and orthogonality between measures, that indicated potential quality dimensions.

To weight the three sources of models equally across random, SETM and discovery sources, sources with less than 1000 models had data points repeated as if resampled. Sample sizes quoted throughout exclude resampling.

A correlation matrix for the exploration measures was generated for the XM and EV datasets. As seen in Figure 3, some measures are identical or very highly correlated (> 0.99), with observable groupings of correlation and orthogonality. For example, TGDU and TOR are very highly correlated. The number of dimensions was estimated using a scree plot, suggesting three dimensions covering 89.3% of the dataset

[1] All source code is accessible at https://github.com/adamburkegh/spm_dim

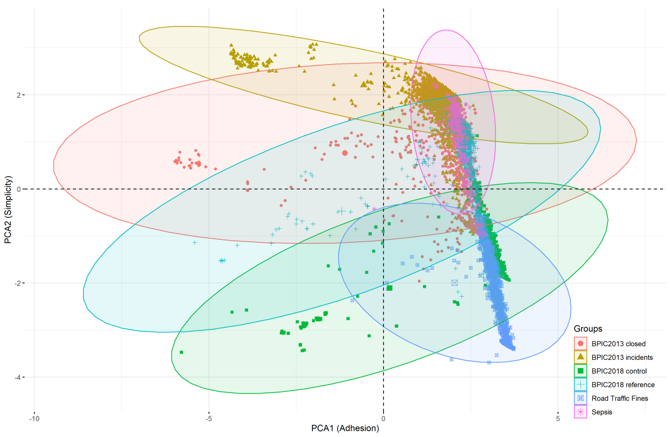


Fig. 4: Variation by log across the first two principal components on the XM and EV combined dataset.

variance for exploration measures. An Anderson-Darling test for normality showed no variables fit a normal distribution ($p < 0.001$), ruling out techniques such as factor analysis.

Next, a scaled Principal Component Analysis (PCA) was performed to analyse the contribution of measures and identify potential dimensions. PCA uses an input data set to calculate orthogonal components that successively maximize variance. Different logs showed different profiles on the principal components, as seen in Figure 4, but the components do not simply correspond to the source logs. For example, though the Road Traffic Fines models are disjoint with the BPIC2013 incidents models, models from other logs intersect in measure space. A check against model generation type correspondence was similar. The first three principal components covered 54.5%, 19.2%, and 15.5% of the variance.

The principal component analysis was then repeated for the EV dataset, including evaluation measures. A scree plot again suggested three dimensions. Including only models for which all measures were available, sample size $n = 72$, the three components cover 52.9%, 17.8% and 11.0% of the variance.

B. Robustness Tests

We performed robustness tests on result subsets to evaluate generalizability. These are summarized below, with additional figures excluded due to space limitations.

The EV dataset already excludes random and genetically-mined (SETM) models. When SETM models are excluded from the XM dataset, the three dimensions remain identifiable. The second and third components are reversed across the XM dataset and the EV dataset, but are associated with many of the same measures. Entropy Trace Fitness (HIFT) remains associated with the first component, but Entropy Trace Precision (HIPT) is more closely correlated with HIFT, and the second component is strongly associated with ARG. Similar, but not identical, components are identified when including evaluation measures, with HIFT more correlated with the second component rather than the first.

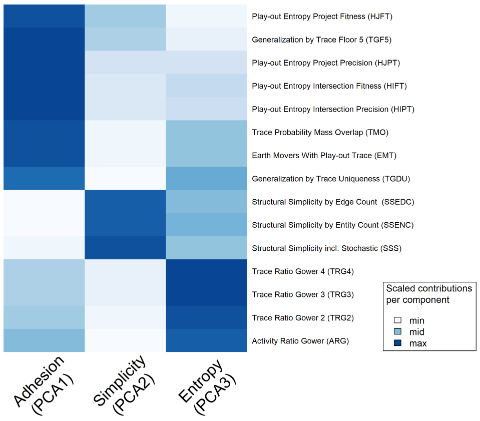


Fig. 5: Measure contribution heatmap by component, on exploration measures for the combined XM and EV dataset.

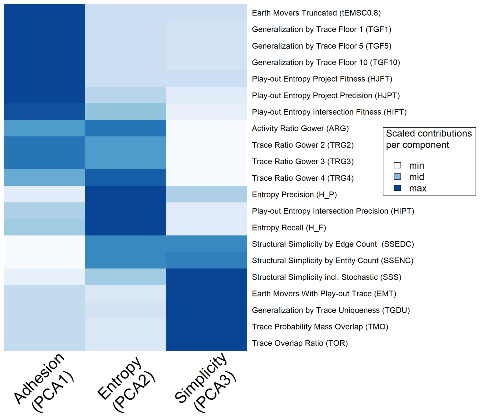


Fig. 6: Measure contribution heatmap by component, for exploration and evaluation measures on the EV dataset.

Analysis was also performed on subsets selected by log, which was largely consistent with the identified dimensions. On some subsets closely correlated measures became collinear and an arbitrary measure had to be excluded to proceed with analysis. Measures showing this behaviour were the groups (TOR,TMO) and the TGF measures.

V. Quality Dimensions

We propose the names *adhesion*, *entropy* and *simplicity* for the three empirical dimensions. Figures 5 and 6 show the contribution of the principal component for a measure on the exploration and evaluation datasets respectively. Measures are ordered by dimension contribution. In the EV dataset, Entropy becomes more explanatory than Simplicity.

Adhesion. To represent how little effort is required to transform one stochastic language into another, we introduce the term *adhesion*. Such a transformation can involve both modifying which traces the process accepts, and the probability of those traces. An informal interpretation is how few changes a manager needs to make for their team to adhere to a different way of working. In this experiment the PCA1 component showed correlation with the evaluation measure Earth Movers Distance (*tEMSC*0.8), and exploration measures Entropy Trace Fitness (HIFT) and Trace Generalization Floor (TGF5/10). It also showed moderate correlation with Earth Movers With Play-out Trace (EMT).

Entropy measures the amount of information in a system: in this case the combination of log and model. The measures of entropy precision H_P and recall H_F [4] both contributed to this dimension, along with entropy precision tracewise HIPT. Activity and subtrace measures ARG and TRG2-4 approximated this dimension on the EV dataset and proxied for more granular measures on the XM dataset. The identification of entropy is supported by related stochastic formalizations [4], and its well-established status in information theory.

Simplicity in models is well-recognized as a virtue in science generally [11], and a desirable dimension in process models specifically [2, p118]. The three simplicity measures used, SSEDC, SSENC, and SSS, were highly correlated and largely orthogonal to other components, with SSS best aligned with the PCA dimension.

The SSEDC and SSENC measures showed some correlation with entropy on the EV dataset, and are sensitive to only control-flow variation in models. Most divergence between simplicity measures was caused by SSS having zero values for 22 discovery models, including 1 from the GDT_SPN technique [6] and 21 from the estimator technique [7]. This is a limitation of the measure, but for this analysis, it provides evidence for the sensitivity of the experiment design, and a common simplicity dimension.

A. Discussion

The proposed dimensions have a number of explanatory features, within the constraints of the experiment.

The adhesion dimension is distinct from control-flow quality dimensions in the following way. Fitness represents whether the model can represent the entirety of the language of the log. Adhesion instead represents how little change is required to alter the stochastic language to match the target. *tEMSC*0.8 measures this dimension by capturing both small trace-level alterations (for instance., adding a missing activity) and re-allocated probability mass. Precision characterizes a type of modelling efficiency: how much of the model is needed to represent shared elements between model and log. Adhesion relates to stochastic language probability weightings, so the TGF measures, which examine traces with more probability mass, approximate it. It also encompasses the idea of edit cost to change paths, so substring-based measures such as TRG2-4 have some correlation in sets of higher quality models.

Figure 1 shows four models for the simple example log L_E. This log has a frequently used path $\langle a, b \rangle$, with a number of much rarer variations. The examples use *tEMSC*0.8 to measure adhesion and the sum of H_P and H_F to measure entropy.

Model 1a has high adhesion and entropy by modeling only the most frequent path and its suffixes. 1b models the most frequent trace simply, but includes paths not represented in the log such as activities y and z. The model covers most of the probability mass, and little needs to change to represent missing traces, but much of the information in the log is not represented. Model 1c includes much information not in the log, but still describes some of the probability mass of the paths. Lastly, model 1d represents many of the paths in the log, but is not representative of their probability, so has high entropy and low adhesion. Simplicity also varies across these examples, with 1a being simpler than 1c.

The entropy dimension discovered combines fitness and precision measures. This suggests considering these as two aspects of a single dimension, and the future design of new entropy-based measures to reflect the dimension directly. Subtrace sensitivity is an approximation to entropy in both datasets, particularly for lower quality models where trace-based measures are zero. The HIPT also shows some correlation with this dimension on both datasets, despite only considering full traces. On the other hand, the examples show that like their control-flow counterparts, entropy fitness and precision can be made to vary independently of each other in some cases. Perhaps further study may establish that they are not only useful descriptive tools for stochastic process models, but can also act as orthogonal dimensions under more specific conditions than those observed in this experiment.

Exploration measures correlated with the adhesion and entropy dimensions might be employed as estimators for adhesion in contexts where *tEMSC*0.8 or H_P cannot be calculated, or where execution cost is crucial.

These results have limitations. A wide range of models and logs were used, but other datasets may reveal other elements. Larger logs of over 200,000 traces or 16 activities were not used, and SETM use was limited by larger numbers of log activities. The stochastic models used were limited to SLPNs, though some of the discovery models were derived from discovery algorithms with BPMN output, and a mix of discovery algorithms were used. The use of PPTs for random generation and for the seed generation in the SETM limits the possible models generated, though it also constrains them to sound models with consistency constraints on stochastic weights. The SETM evolutionary fitness function, which includes all the exploration measures, may tend to a correlative effect between measures. The recurrence of the same correlations when excluding SETM models suggests this effect, if it exists, was not large.

As an experimentally derived theory, further experimentation will be the ultimate test of generality for all of the proposed dimensions. Currently, these orthogonal dimensions comprise a descriptive tool, not a predictive quantitative model, and future experiments may yield further insights. It is promising that features corresponding to the dimensions can be qualitatively observed in specific models above.

VI. Related Work

In both the natural sciences [11] and statistical learning [12], the aim is usually to choose a single best model based on some universal calculation, with the exact form of that calculation a matter of robust debate [11]. In process mining, choice of model is seen as a context-driven judgement call, where desirable criteria are necessarily traded-off against one another. That approach is supported by the concept of quality dimensions. Initial investigations explored dimensions such as structural and behavioural appropriateness [13] before converging on fitness, precision, simplicity and generalization [14], [2].

The current study builds directly on the analysis of genetically-mined control-flow models [14], both in study design, and direct extension of the Evolutionary Tree Miner code [9]. That work conducted a qualitative study on classes of models generated with different genetic miner constraints. In this work, the dimensions derived through quantitative analysis in Section IV are applied to a small qualitative analysis in Section V-A.

At least one quantitative study comparing process measures and dimensions has been conducted for control-flow dimensions [15]. Factor analysis was used to identify common components, for which clear correspondence was found for existing fitness and precision measures. Simplicity was excluded from the analysis. Like this study, it does not provide evidence for a generalization dimension. Also like this study, it abstracts from the log in ways that could prevent such a dimension being detected. Correspondence among groups of measures was clearer than for the current study, which may be due to concepts not translating to stochastic models, or the computationally cheaper exploration measures not being as generally useful as proven control-flow measures.

Stochastic conformance measures are reviewed in Section III-A. A framework for calculating behavioural simplicity [16] can also accommodate stochastic models. In contrast to the structural simplicity used in this paper, behavioural simplicity represents the underlying problem complexity rather than the specific model representation. The relationship between behavioural simplicity and other quality measures remains of research interest.

As well as the GSPN discovery techniques used in this experiment [6], [7], [8], other recent research has shown techniques for discovering probability-annotated BPMN models [17], probabilistic declarative models [18], non-classical probability Bayesian networks [19] and Bayesian models for place-transition Petri nets [20]. The Bayesian technique [20] also has potential applications for model comparison and new conformance measures.

VII. Conclusion

Stochastic process models provide a description of the "everydayness" of work in organizations. To help in the use and evaluation of these models, we conducted an empirical study of stochastic process model quality measures and relationships. Models were generated from six real-life logs and collected using both random model generation and process discovery. Analyzing a variety of computationally cheap measures across thousands of models, three quality dimensions were observed using principal component analysis. Two dimensions were identified with model simplicity and log-model system entropy. An adhesion dimension is also proposed, which represents how little change is needed for a model to match the stochastic language of a log. Future work may further explore the theoretical basis of these dimensions, construct correlating measures, and challenge the theory with further empirical tests.

Appendix

A. Stochastic Play-out Log Generator

The stochastic play-out log generator implemented for these experiments is represented as function spg. Function $eb\colon \mathcal{G} \times \mathcal{B}(P) \to \mathbb{P}(T)$ returns all enabled transitions for a net and a marking. Function $tg\colon \mathcal{G} \times \mathcal{B}(P) \times T \to \mathcal{B}(P)$ returns the new marking after a transition fires.

$$\begin{aligned}
&\text{Let } g = (P, T, F, M_0, W, \lambda) \text{ below}\\
&tr\colon T \times \mathcal{G} \to A^*\\
&tr(t, g) = \langle\rangle \text{ if } \lambda(t) = \tau \text{ else } \langle \lambda(t) \rangle\\
&sdlg\colon \mathcal{G} \times \mathcal{B}(P) \times \mathbb{N} \to \mathcal{L}\\
sdlg(g, m, n) = &\biguplus_{t \in eb(g,m)} [\sigma^f \mid \sigma = tr(t, g) + \sigma_{tl}\\
&\qquad \wedge d = \text{floor}(\frac{nW(t)}{W_s}) + surplus(t, g, m, n)\\
&\qquad \wedge \sigma_{tl} \in sdlg(g, tg(g, m, t), d)]\\
&\text{where } W_s = \sum_{t' \in eb(g,m)} W(t')\\
spg(g, n) = &sdlg(g, M_0, n)
\end{aligned}$$

The function takes a target size as a trace "budget", then recursively splits the budget according to each possible state in a token game, and the relative weights of enabled transitions.

Rounding, represented by the $surplus$ function, is done by lexical order of the transition labels, then to the transition with the least allocation, then arbitrarily. The implementation also imposes a maximum trace length of 5000 to limit the impact of models generating very long traces (from certain loops).

B. Detailed Exploration Measures

This section details the measures summarized in Table I. For the measure definitions below, let event log $L \in \mathcal{L}$, model $g \in \mathcal{G}$. To obtain the playout log $M \in \mathcal{L}^+$, the model g is played out to k traces, then occurrences are scaled to match the original log: $M = \frac{|L|}{k} \cdot spg(g, k)$.

The first measure is a simplification of the stochastic Earth Movers' distance [5].

EMT Earth Movers with play-out trace weighting.

$$EMT(M, L) = 1 - \frac{1}{|L|} \sum_{\sigma \in L} \max(L[\sigma] - M[\sigma], 0)$$

Two measures address how much of the probability mass of the log is in shared traces.

TMO Trace Probability mass overlap.

$$TMO(M, L) = \sum_{\sigma \in L \sqcap M} \frac{(L \sqcap M)[\sigma]}{|L|}$$

TOR Trace overlap ratio.

$$TOR(M,L) = \frac{|L \sqcap M|}{|L|}$$

Analysis of which subtraces occur in both log and model (represented by the play-out log) approximate fitness.

ARG The Gower's similarity [21] between activity count ratio vectors. This measure is designed to be deliberately sensitive to variation between poor quality models, when other measures may be zero. Given log L, take $ST_n(L)$ to be the subtraces of length n, $\sigma_s\#L$ the subtrace frequency of σ_s, with each occurrence in a trace counted, and $||L||_n$ to be the total subtraces of length n. ARG is a special case: ARG=TRG1.

TRGn Subtrace ratios, activity ratios generalized to sub-traces of length n. **TRG2**, **TRG3** and **TRG4** are all measured.

$$TRGn(M,L) = \sum_{\sigma \in ST_n(L \sqcup M)} 1 - y_\sigma$$

$$\text{where } y_\sigma = \frac{1}{\max(\sigma\#L, \sigma\#M)} \left| \frac{\sigma\#L}{||L||_n} - \frac{\sigma\#M}{||M||_n} \right|$$

Two simplified variants of evaluation measure entropy [4], based on play-out logs, are used to define fitness and precision measures. The first uses bag intersection.

HIFT Play-out entropy intersection fitness.

$$HIFT(M,L) = \min(1, \frac{H(L \sqcap M)}{H(L)})$$

HIPT Play-out entropy intersection precision.

$$HIPT(M,L) = \min(1, \frac{H(L \sqcap M)}{H(M)})$$

The second entropy variant uses SDFA projection [4] function $\mathcal{P} \colon \mathcal{L}^+ \times \mathcal{L}^+ \rightarrow \mathcal{L}^+$, where traces are used as SDFA tokens.

$$\mathcal{P}(L_1, L_2) = L_P \sqcup [\langle\rangle^{|L_1| - |L_P|}]$$

$$\text{where } L_P = [\sigma^i \in L_1 \mid \exists_{j>0}\ \sigma^j \in L_2]$$

HJFT Play-out entropy projection fitness.

$$HJFT(M,L) = \frac{H(\mathcal{P}(L,M))}{H(L)}$$

HJPT Play-out entropy projection precision.

$$HJPT(M,L) = \frac{H(\mathcal{P}(M,L))}{H(M)}$$

Three simplicity measures are scaled by log size to impose a valid upper bound of 1.

SSENC Structural simplicity by entity count [22].

$$SSENC(g,L) = \max(1 - \frac{|P| + |T|}{|L|}, 0)$$

SSEDC Structural simplicity by edge count [22].

$$SSEDC(g,L) = \max(1 - \frac{|F|}{|L|}, 0)$$

SSS Structural simplicity by all structural components in SLPNs. This accounts for stochastic features not found in existing structural simplicity measures.

$$SSS(g,L) = \max(1 - \frac{1}{|L|}(|P| + |T| + |F| + |\bigcup_{t \in T} W(t)|), 0)$$

The following generalization measures are at a trace level, and are taken from example measures in [23].

TGF1 Generalization by trace floor, gen_{L2M_q} [23]. We also use **TGF5** and **TGF10** as measures for trace floors of 5 and 10 respectively.

$$TGF1(M,L) = \frac{|[\sigma \in l | \sigma \in M \wedge L[\sigma] \geqslant q]|}{|L|} \text{ with } q \geqslant 1$$

TGDU Generalization by trace uniqueness difference, $gen_{L2M_{HB}}$ [23].

$$TGDU(M,L) = \frac{|[\sigma \in L | \sigma \in M]| - |L \sqcap M|}{|L|}$$

Acknowledgements. The eResearch Office at QUT provided computational resources for this work.

References

[1] A. Ammitzbøll Flügge, T. Hildebrandt, and N. H. Møller, "Street-Level Algorithms and AI in Bureaucratic Decision-Making: A Caseworker Perspective," *Proc. ACM Hum. Comput. Interact*, pp. 40:1–40:23, 2021.

[2] W. M. P. van der Aalst, *Process Mining: Data Science in Action*, 2nd ed. Springer-Verlag, 2016.

[3] F. Bause and P. Kritzinger, *Stochastic Petri Nets: An Introduction to the Theory*. Vieweg+Teubner Verlag, 2002.

[4] S. J. J. Leemans and A. Polyvyanyy, "Stochastic-Aware Conformance Checking: An Entropy-Based Approach," in *CAiSE*, 2020, pp. 217–233.

[5] S. J. J. Leemans, W. M. P. van der Aalst, T. Brockhoff, and A. Polyvyanyy, "Stochastic process mining: Earth movers' stochastic conformance," *IS*, p. 101724, 2021.

[6] A. Rogge-Solti, W. M. P. van der Aalst, and M. Weske, "Discovering Stochastic Petri Nets with Arbitrary Delay Distributions from Event Logs," in *BPM Workshops*, 2014, pp. 15–27.

[7] A. Burke, S. J. J. Leemans, and M. T. Wynn, "Stochastic Process Discovery by Weight Estimation," in *Process Mining Workshops*, 2021.

[8] ——, "Discovering Stochastic Process Models By Reduction and Abstraction," in *Petri Nets*, 2021, pp. 312–336.

[9] J. C. A. M. Buijs, B. F. van Dongen, and W. M. P. van der Aalst, "A genetic algorithm for discovering process trees," in *2012 IEEE Congress on Evolutionary Computation*, 2012, pp. 1–8.

[10] I. Jolliffe, "Principal Component Analysis," in *International Encyclopedia of Statistical Science*, M. Lovric, Ed., 2011, pp. 1094–1096.

[11] E. Sober, *Ockham's Razors: A User's Manual*. CUP, 2015.

[12] U. von Luxburg and B. Schölkopf, "Statistical Learning Theory: Models, Concepts, and Results," in *Handbook of the History of Logic*, Gabbay *et al.*, Eds., 2011, pp. 651–706.

[13] A. Rozinat and W. M. P. van der Aalst, "Conformance checking of processes based on monitoring real behavior," *IS*, pp. 64–95, 2008.

[14] J. C. A. M. Buijs, B. F. van Dongen, and W. M. P. van der Aalst, "Quality Dimensions in Process Discovery: The Importance of Fitness, Precision, Generalization and Simplicity," *International Journal of Cooperative Information Systems*, p. 1440001, 2014.

[15] G. Janssenswillen, N. Donders, T. Jouck, and B. Depaire, "A comparative study of existing quality measures for process discovery," *IS*, 2017.

[16] A. Kalenkova, A. Polyvyanyy, and M. La Rosa, "A Framework for Estimating Simplicity of Automatically Discovered Process Models Based on Structural and Behavioral Characteristics," in *BPM*, 2020.

[17] M. Camargo, M. Dumas, and O. González-Rojas, "Automated discovery of business process simulation models from event logs," *DSS*, 2020.

[18] F. M. Maggi, M. Montali, and R. Peñaloza, "Probabilistic Conformance Checking Based on Declarative Process Models," in *CAiSE*, 2020.

[19] C. Moreira, E. Haven, S. Sozzo, and A. Wichert, "Process mining with real world financial loan applications: Improving inference on incomplete event logs," *PLOS ONE*, vol. 13, p. e0207806, 2018.

[20] G. Janssenswillen, B. Depaire, and C. Faes, "Enhancing Discovered Process Models using Bayesian Inference and MCMC," in *Proceedings of the 2020 BPI Workshop*, 2020.

[21] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, pp. 857–871, 1971.

[22] J. Mendling, G. Neumann, and W. M. P. van der Aalst, "Understanding the Occurrence of Errors in Process Models Based on Metrics," in *OTM*, 2007, pp. 113–130.

[23] W. M. P. van der Aalst, "Relating Process Models and Event Logs-21 Conformance Propositions." in *ATAED@ Petri Nets/ACSD*, 2018.