# Identifying and Explaining Undesirable Traces in Business Processes Using Descriptive and Predictive Analysis

Ali Norouzifar(ali.norouzifar@pads.rwth-aachen.de, RWTH Aachen University

## Abstract

This study proposes data analytic approaches to identify possibly undesirable traces in complex business processes. Two types of analysis are employed: descriptive and predictive. In the descriptive analysis, cohorts of traces with different control flows are identified based on a case attribute or process indicator. A hierarchical clustering approach is used for this purpose by employing the Earth Mover's Distance to measure differences between the cohorts. In the predictive analysis, machine learning models are trained to predict the cohort for unseen cases, and SHAP values are used as an explainable AI technique to extract the most critical features. The approach is demonstrated through experiments that show its practical and effective use in identifying possibly undesirable traces and providing insights to improve the process or discover better process models.

## Introduction

What is an event log?


| Case id | Application type | Loan goal | Requested amount | Throughput time |
|---|---|---|---|---|
| Application_1000311556 | New credit | Car | 45000 | 734 |
| Application_1000339879 | New credit | Existing loan takeover | 37500 | 308 |
| Application_1000034150 | New credit | Existing loan takeover | 5000 | 267 |
| Application_1000557783 | New credit | Home improvement | 10000 | 306 |

| Case id | Timestamp | Activity name | resource |
|---|---|---|---|
| Application_1000339879 | 3/17/2016 12:57 | A_Create Application | User_1 |
| Application_1000339879 | 3/17/2016 12:57 | A_Submitted | User_1 |
| Application_1000339879 | 3/17/2016 12:58 | A_Concept | User_1 |
| Application_1000339879 | 3/18/2016 8:25 | A_Accepted | User_28 |
| Application_1000339879 | 3/18/2016 8:30 | O_Create Offer | User_28 |
| Application_1000339879 | 3/18/2016 8:30 | O_Created | User_28 |
| Application_1000339879 | 3/18/2016 8:31 | O_Sent (mail and online) | User_28 |
| Application_1000339879 | 3/18/2016 8:31 | A_Complete | User_28 |
| Application_1000557783 | 3/18/2016 9:23 | A_Validating | User_119 |
| Application_1000557783 | 3/18/2016 9:23 | O_Returned | User_119 |
| Application_1000557783 | 3/22/2016 13:09 | O_Accepted | User_30 |
| Application_1000557783 | 3/22/2016 13:09 | A_Pending | User_30 |
| Application_1000339879 | 3/23/2016 8:16 | A_Validating | User_119 |

Earth Mover's distance [1]: Let $f$ and $g$ be finite stochastic languages and let $\delta$ be a trace distance function. The earth movers' distance between $f$ and $g$ is:

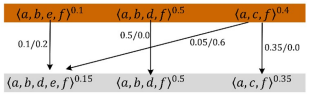
$$EMD(f,g) = min \sum_{t \in S_f} \sum_{u \in S_g} x_{tu} \cdot \delta(t,u)$$

$$\forall t \in S_f : \sum_{u \in S_g} x_{tu} = f(t)$$

$$\forall u \in S_g : \sum_{t \in S_f} x_{tu} = g(u)$$

$$\forall t \in S_f \ \forall u \in S_g : x_{tu} \geq 0$$

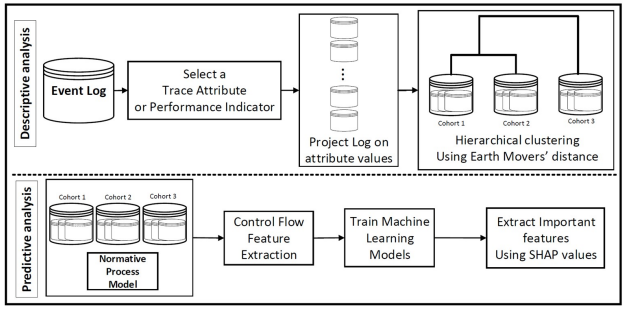
Example:




## Cohort analysis

Descriptive analysis: we focus on identifying cohorts, which are groups of traces with different control flows [2]. Our strategy is to partition the event log into smaller segments based on a feature of interest and then use hierarchical trace clustering to merge the partitions based on the earth movers' distance. A user-defined distance threshold is used to stop the hierarchical clustering.

Predictive analysis: Next to finding interesting cohorts, we may train some machine learning models to predict the cohort based on some designed control flow features. After training and evaluating a machine learning model using well-known machine learning evaluation metrics, SHAP (SHapley Additive exPlanations) values are used as an explainable AI technique to extract the most important features contributing in the prediction.



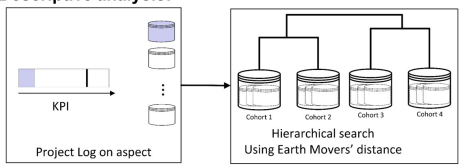

## Experiments

**Event log:** The BPIC 2017 event log comprises 11000 loan applications (randomly sampled) and provides a realistic publicly available dataset for evaluating our framework.

**KPI:** throughput time of traces

**Parameters:** Number of bins = 50, Minimum stop distance = 0.25

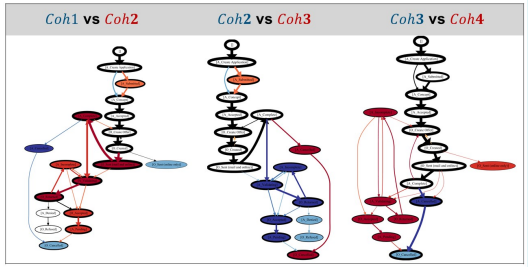
**Descriptive analysis:**




$\boldsymbol{Coh1}$: $0 \leq throughput\ time < 4\ days$ (size: 306)
$\boldsymbol{Coh2}$: $4\ days \leq throughput\ time < 30\ days$, (size: 7064)
$\boldsymbol{Coh3}$: $30\ days \leq throughput\ time < 36\ days$, (size: 2761)
$\boldsymbol{Coh4}$: $36\ days \leq throughput\ time < 170\ days$, (size: 869)

$$EMD(coh1, coh2) = 0.29$$
$$EMD(coh2, coh3) = 0.33$$
$$EMD(coh3, coh4) = 0.3$$

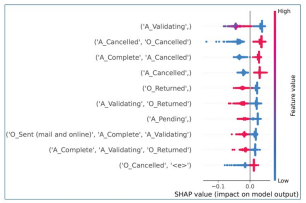
Comparing the cohorts using statistical tests [3]:




**Predictive analysis:**

**Features:** only encoded control flow using n-gram encoding with $n \in \{1,2,3\}$

The SHAP values are extracted for the trained machine learning model using Random Forest algorithm.




## Conclusion

The experimental results show that this approach offers valuable insights into identifying potentially undesirable traces. It provides a practical and effective way to analyze and understand complex processes, which can be used for process improvement and better process discovery. Although the initial results are promising, further justification is required for the design choices, and the framework could be enhanced to make it applicable to real-life use cases.

Some interesting directions to continue this research:

- Collaborating with our industrial partner, the UWV agency, to validate the findings using our framework.
- Exploring alternative KPIs to gain additional insights from the framework.
- The binning strategy employed in this framework may face difficulties if the behavior is changing on the bin boundaries.
- Considering several KPIs for the cohort analysis.

## Acknowledgment

This research was supported by the research training group Dataninja (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia.

## References

[1] Leemans, S.J., Syring, A.F. and van der Aalst, W.M.P., 2019. Earth movers' stochastic conformance checking. In *Business Process Management Forum: BPM Forum 2019, Vienna, Austria, September 1–6, 2019, Proceedings 17* (pp. 127-143). Springer International Publishing.

[2] Leemans, S.J., Shabaninejad, S., Goel, K., Khosravi, H., Sadiq, S. and Wynn, M.T., 2020. Identifying cohorts: Recommending drill-downs based on differences in behaviour for process mining. In *Conceptual Modeling: 39th International Conference, ER 2020, Vienna, Austria, November 3–6, 2020, Proceedings 39* (pp. 92-102). Springer International Publishing.

[1] Bolt, A., de Leoni, M. and van der Aalst, W.M.P., 2018. Process variant comparison: using event logs to detect differences in behavior and business rules. *Information Systems*, *74*, pp.53-66.