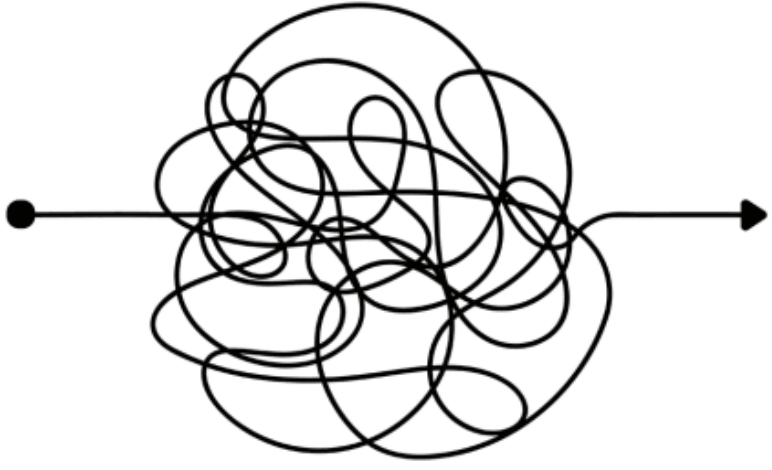
"Simplicity is the ultimate sophistication." - Leonardo da Vinci



Replication project:

The Impact of Process Complexity on Process Performance: A Study Using Event Log Data Vidgof/Wurm/Mendling (2023)

Advanced Process Mining – FSS 2024

Group 4:
Leonie Starke, Progha Labani Das, Max Frigolis, Biel Alcaraz, Frederik Röckle

Agenda

UNIVERSITY OF MANNHEIM

Status quo

Authors approach

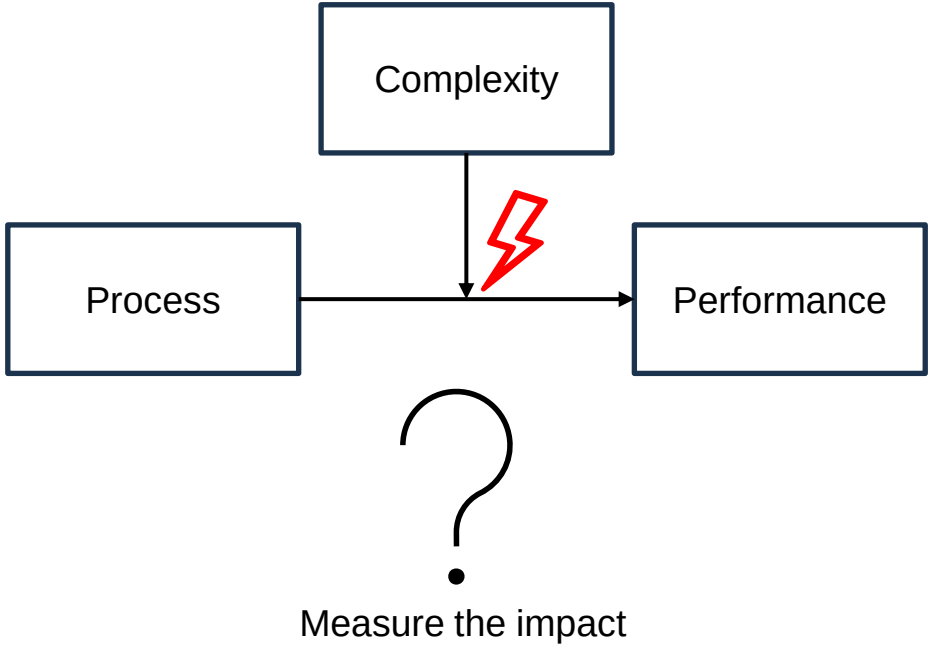
Replication Experience

Results

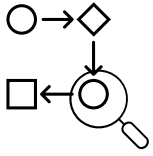
Additional Insights

Conclusion

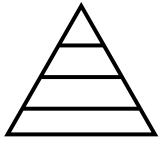
Alleged negative impact of complexity on performance



Status quo measured only by behavioral studies



Only point-spot observations of the process



Results of studies vary across hierarchy

Very subjective perception hence not reliable

Authors approach quantifies correlation by regression

UNIVERSITY OF MANNHEIM

Research goal: Quantify the impact of process complexity on process performance by analyzing the relation based on event log data

Two-fold process:

1. Obtain datasets and compute process complexity metrics and throughput time

2. Conduct a statistic regression analysis to investigate the correlation between complexity and performance (throughput time)

Linear model: X (Complexity metrics) -> y (median throughput time)

Authors results explain up to 0.96 variance and high corr.

UNIVERSITY OF MANNHEIM

1. Obtain data and computation of complexity metrics:
   - Use event logs from BPIC datasets from 2011, 2015 2017, 2018, 2019, 2020

2. Regression Analysis
   - Built 6 regression models with forward, backward and both selection method
   - Encoded industry domain of event log as a dummy variable

Results

R1. Best performing model reached an R-squared value of 0.969

R2. Industry dummy variable alone accounts for 80% of variance in the dependent variable

38 complexity metrics to fit throughput time

UNIVERSITY OF MANNHEIM

Size:
- No. Events
- No. Sequences
- Avg Sequence length

Variation:
- No. Acyclic paths
- No. Unique sequences

Distance:
- Avg Affinity
- Deviation from random

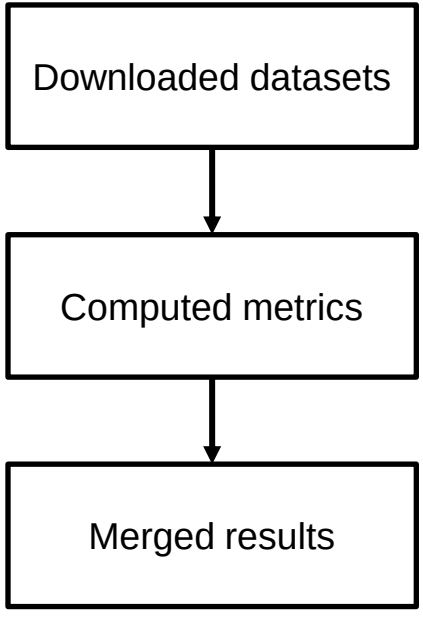
Simple Entropy:
- Sequence entropy
- Variant entropy

Enriched Entropy:
- Enriched sequence entropy
- Enriched variant entropy

only examples

Our Replication Experience – Obtaining the datasets

UNIVERSITY OF MANNHEIM



Downloaded all datasets from the TU Eindhoven website

Computed the complexity metrics and throughput time on a Azure VM 32gb ram instance with script provided by authors

Merged all metrics in one file

Our Replication Experience – the regression analysis

Built regression models

Built OLS regression models on Google Colab with statsmodels. Authors didn‘t provide any code for regression

1. With our replicated dataset

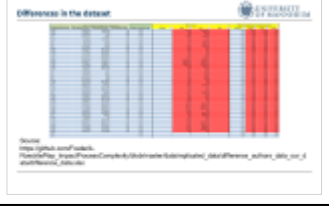
$R^2 > 0.98$

2. With provided dataset

$R^2 > 0.98$

Authors $R^2 > 0.96$

Investigated differences between datasets



Apply authors approach on new data – external validity

UNIVERSITY OF MANNHEIM

1. We built a new dataset based on (BPIC 2012, Hospital Billing and Sespsi Cases)
2. Applied the same pipeline
3. Reached a similiar good R^2-value

$R^2 > 0.98$

Explanation of total variance (R2) by industry

R2. *„Industry dummy variable alone accounts for 80% of variance in the dependent variable„*

X (Industry variable) -> y (median throughput time)

```
model = sm.OLS(y,X).fit()
```

```
model.rsquared
```

```
0.26560601353812285
```

Source: https://github.com/Frederik-Roeckle/Rep_ImpactProcessComplexity/blob/master/models/Only_industry_variable_as_predictor.ipynb

Additional Insights during our replication

UNIVERSITY OF MANNHEIM

methodology

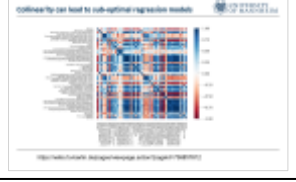
- Category groups of complexity metrics are not always aligned
- Criterias for outlier removement were not clearly specified
- Mapping of variables between paper and authors script was unclear in the beginning

dataset

- Difference between our obtained data and the one from the author in the dataset
- The authors used a dataset (finance_log) where they didn't state the origin

regression analysis

- Authors didn't provide code for regression analysis
- High correlation between some variables
- Authors didn't share their best performing model (neither variables or weights)



Conclusion

UNIVERSITY OF MANNHEIM

Paper quantifies the relation of process complexity and process performance through a Linear model: X (Complexity metrics) -> y (median throughput time)

We partly replicated and extended the results of the author's work:

✓ R1.:Achieved similar R^2-scores for same data sets
=> test external validity on new data sets

✗ R2.: Only explain 0.26 instead of 0.8 variance in throughput time by the industry

Lookout from the authors: “apply dimensionality reduction techniques to find best describing variables and even use regression models for prediction”

Repo

Link to our GitHub-Repo:

https://github.com/Frederik-Roeckle/Rep_ImpactProcessComplexity

Results – computed data

UNIVERSITY OF MANNHEIM

COMPUTED DATA			
Models without dummy	Remaining Variables	R-squared	
Full forward	47	0,988081	
Full backward	40	0,98815	
Full stepwise	41	0,98815	
Significant forward	18	0,988149	
Significant backward	27	0,983590	
Significant stepwise	15	0,988345	
Minimal forward	5	0,988188	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper
Minimal backward	5	0,931998	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper
Minimal stepwise	5	0,931998	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper
COMPUTED DATA			
Models with dummy	Remaining Variables	R-squared	
Full forward	48	0,988177	
Full backward	41	0,988359	
Full stepwise	45	0,979055	
Significant forward	18	0,988149	
Significant backward	24	0,983590	
Significant stepwise	13	0,983475	
Minimal forward	5	0,988399	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper
Minimal backward	5	0,931998	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper
Minimal stepwise	5	0,931998	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper

dataset used for that model: https://github.com/Frederik-Roeckle/Rep_ImpactProcessComplexity/blob/master/data/replicated_data/generated_merged.csv

Results – authors provided data

UNIVERSITY OF MANNHEIM

PROVIDED DATA			
Models without dummy	Remaining Variables	R-squared	
Full forward	47	0,988188	
Full backward	44	0,988166	
Full stepwise	46	0,988185	
Significant forward	23	0,988188	
Significant backward	28	0,988188	
Significant stepwise	51	0,985422	
Minimal forward	5	0,988188	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper
Minimal backward	5	0,931999	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper
Minimal stepwise	5	0,931999	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper

PROVIDED DATA			
Models with dummy	Remaining Variables	R-squared	
Full forward	48	0,9884	
Full backward	44	0,988392	
Full stepwise	42	0,988359	
Significant forward	19	0,9884	
Significant backward	26	0,988392	
Significant stepwise	10	0,974088	
Minimal forward	5	0,9884	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper
Minimal backward	5	0,931999	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper
Minimal stepwise	5	0,931999	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper

dataset used for that model: https://github.com/MaxVidgof/complexity-data/blob/main/merged.csv

Results – our own test data

UNIVERSITY OF MANNHEIM

EVALUATION DATA			
Models without dummy	Remaining Variables	R-squared	
Full forward	47	0,988149	
Full backward	40	0,98815	
Full stepwise	45	0,988184	
Significant forward	18	0,988149	
Significant backward	27	0,983590	
Significant stepwise	15	0,983475	
Minimal forward	5	0,988188	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper
Minimal backward	5	0,931998	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper
Minimal stepwise	5	0,931998	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper

EVALUATION DATA			
Models with dummy	Remaining Variables	R-squared	
Full forward	48	0,988304	
Full backward	41	0,988359	
Full stepwise	45	0,988359	
Significant forward	18	0,988304	
Significant backward	24	0,988359	
Significant stepwise	13	0,971739	
Minimal forward	5	0,988399	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper
Minimal backward	5	0,931998	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper
Minimal stepwise	5	0,931998	here we just have 1 remainining variable for each of the 5 categories, because this is what we understood in the paper

dataset used for that model: https://github.com/Frederik-Roeckle/Rep_ImpactProcessComplexity/blob/master/data/test_data/generated_merged.csv

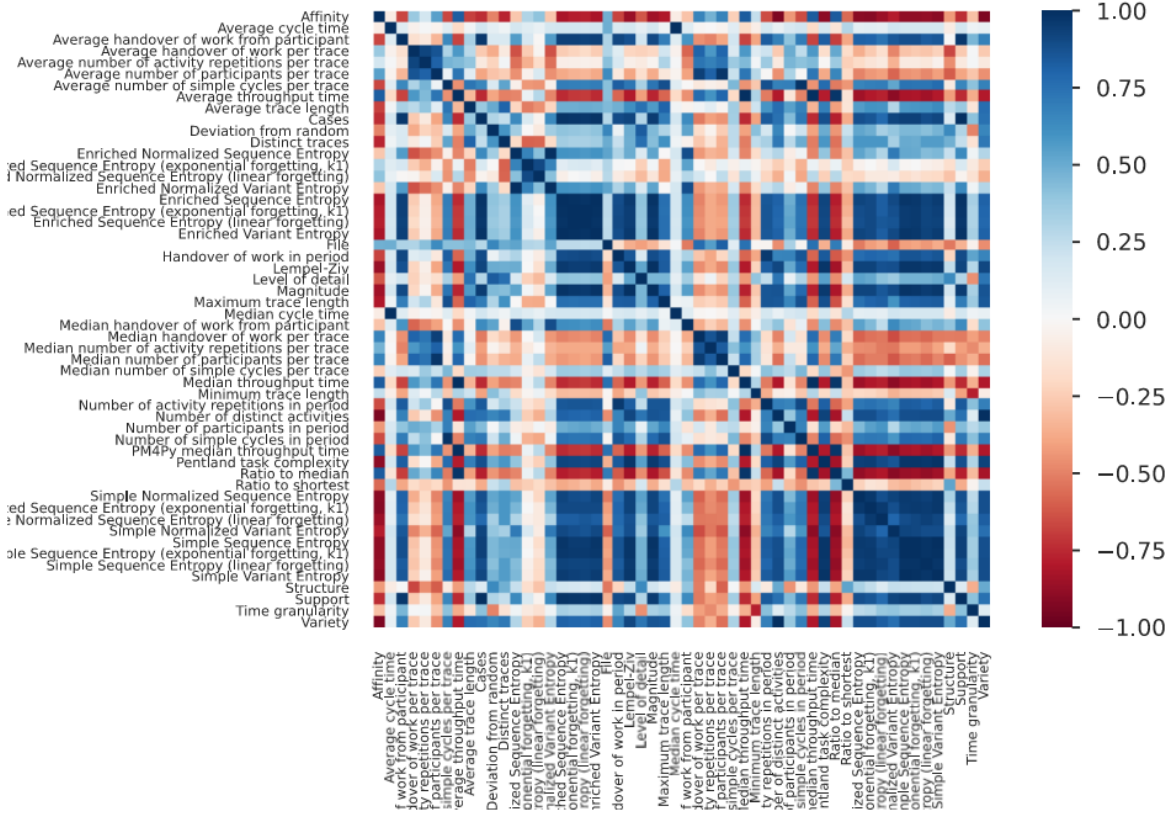
Differences in the dataset

UNIVERSITY OF MANNHEIM

Generated Merged (Created during reproduction)					Difference Total				Difference Relative (based on original merged.csv)			
Datatype_generated_merg	Mean_generated_merg	Sum_generated_merge	Count_generated_merg	Missing_in_generated_merg	Datatype	Mean	Sum	Count	Datatype2	Mean3	Sum4	Count5
float64	11988,43445	4195952,058	350	FALSE	0	6173,565	2269719,891	6	0,00%	33,99%	35,10%	1,69%
float64	43,60835184	64017,0605	1468	FALSE	0	5,802	8813,506	6	0,00%	11,74%	12,10%	0,41%
float64	0,342302452	502,5	1468	FALSE	0	0,009	16,000	6	0,00%	2,69%	3,09%	0,41%
int64	3839,917575	5636999	1468	FALSE	0	39,867	81804,000	6	0,00%	1,03%	1,43%	0,41%
float64	0,24216328	355,4956947	1468	FALSE	0	0,002	4,245	6	0,00%	0,78%	1,18%	0,41%
int64	1049,76158	1541050	1468	FALSE	0	7,175	16874,000	6	0,00%	0,68%	1,08%	0,41%
int64	1049,76158	1541050	1468	FALSE	0	7,175	16874,000	6	0,00%	0,68%	1,08%	0,41%
int64	3127,540872	4591230	1468	FALSE	0	21,361	50251,000	6	0,00%	0,68%	1,08%	0,41%
float64	0,199122216	292,3114125	1468	FALSE	0	0,001	2,801	6	0,00%	0,54%	0,95%	0,41%
float64	0,137087137	201,2439165	1468	FALSE	0	0,001	1,813	6	0,00%	0,49%	0,89%	0,41%
float64	0,115794664	169,9865675	1468	FALSE	0	0,001	1,496	6	0,00%	0,47%	0,87%	0,41%
float64	2,514175703	3690,809931	1468	FALSE	0	0,011	31,086	6	0,00%	0,43%	0,84%	0,41%
float64	23143,07795	33974038,43	1468	FALSE	0	-96,387	-3216,217	6	0,00%	0,42%	-0,01%	0,41%
float64	940817998,9	1,38112E+12	1468	FALSE	0	-3826748,614	4280536,476	6	0,00%	0,41%	0,00%	0,41%
float64	941186177,1	1,38166E+12	1468	FALSE	0	-3828230,950	4304641,713	6	0,00%	0,41%	0,00%	0,41%
float64	1010818685	1,48388E+12	1468	FALSE	0	-4109901,562	6917205,313	6	0,00%	0,41%	0,00%	0,41%
float64	2652,425545	3893760,701	1468	FALSE	0	10,860	31922,076	6	0,00%	0,41%	0,81%	0,41%
float64	169,834471	249317,0034	1468	FALSE	0	-0,650	60,352	6	0,00%	0,38%	0,02%	0,41%
int64	9,857629428	14471	1468	FALSE	0	-0,027	19,000	6	0,00%	0,28%	0,13%	0,41%
int64	248,876703	365351	1468	FALSE	0	-0,651	533,000	6	0,00%	0,26%	0,15%	0,41%
int64	63,31675749	92949	1468	FALSE	0	-0,164	138,000	6	0,00%	0,26%	0,15%	0,41%
int64	63,31675749	92949	1468	FALSE	0	-0,164	138,000	6	0,00%	0,26%	0,15%	0,41%
float64	0,437153158	641,7408355	1468	FALSE	0	0,001	4,269	6	0,00%	0,25%	0,66%	0,41%
float64	0,437153158	641,7408355	1468	FALSE	0	0,001	4,269	6	0,00%	0,25%	0,66%	0,41%
int64	20896,47207	30676021	1468	FALSE	0	48,902	197460,000	6	0,00%	0,23%	0,64%	0,41%
float64	0,604953758	848,1451684	1402	FALSE	0	-0,001	1,701	6	0,00%	0,23%	0,20%	0,43%
float64	12,28625586	18036,22361	1468	FALSE	0	-0,024	39,045	6	0,00%	0,19%	0,22%	0,41%
int64	29603,54428	43458003	1468	FALSE	0	55,257	259070,000	6	0,00%	0,19%	0,59%	0,41%
float64	29,49208731	43294,38417	1468	FALSE	0	-0,054	96,709	6	0,00%	0,18%	0,22%	0,41%
float64	17,20583145	25258,16057	1468	FALSE	0	-0,031	57,664	6	0,00%	0,18%	0,23%	0,41%
float64	0,331149333	486,1272206	1468	FALSE	0	0,000	2,687	6	0,00%	0,14%	0,55%	0,41%
float64	0,289884327	425,5501927	1468	FALSE	0	0,0003	2,160	6	0,00%	0,10%	0,51%	0,41%
float64	120028,4514	176201766,6	1468	FALSE	0	-110,552	557217,198	6	0,00%	0,09%	0,32%	0,41%
float64	271763,2656	398948473,9	1468	FALSE	0	243,012	1988779,544	6	0,00%	0,09%	0,50%	0,41%
float64	271763,2656	398948473,9	1468	FALSE	0	243,012	1988779,544	6	0,00%	0,09%	0,50%	0,41%
float64	0,469348714	689,003912	1468	FALSE	0	0,0002	3,135	6	0,00%	0,05%	0,45%	0,41%
float64	129053,1048	189449957,8	1468	FALSE	0	-52,632	696739,193	6	0,00%	0,04%	0,37%	0,41%
int64	19002,67643	27895929	1468	FALSE	0	-7,689	102683,000	6	0,00%	0,04%	0,37%	0,41%
float64	3,65599455	5367	1468	FALSE	0	-0,001	20,000	6	0,00%	0,04%	0,37%	0,41%
float64	0,748687783	1099,073665	1468	FALSE	0	0,0002	4,795	6	0,00%	0,03%	0,43%	0,41%
float64	196609,7917	288623174,2	1468	FALSE	0	51,629	1255760,254	6	0,00%	0,03%	0,43%	0,41%
float64	204032,2981	299519413,5	1468	FALSE	0	49,103	1296571,321	6	0,00%	0,02%	0,43%	0,41%
float64	145326,8635	213339835,6	1468	FALSE	0	-22,384	838966,846	6	0,00%	0,02%	0,39%	0,41%
float64	175734,0809	257977630,7	1468	FALSE	0	-24,215	1018711,533	6	0,00%	0,01%	0,39%	0,41%
float64	0,852795576	1251,903906	1468	FALSE	0	0,000	4,964	6	0,00%	0,01%	0,39%	0,41%
object				FALSE	0	0	0	0	0,00%	0,00%	0,00%	0,00%
object				FALSE	0	0	0	0	0,00%	0,00%	0,00%	0,00%

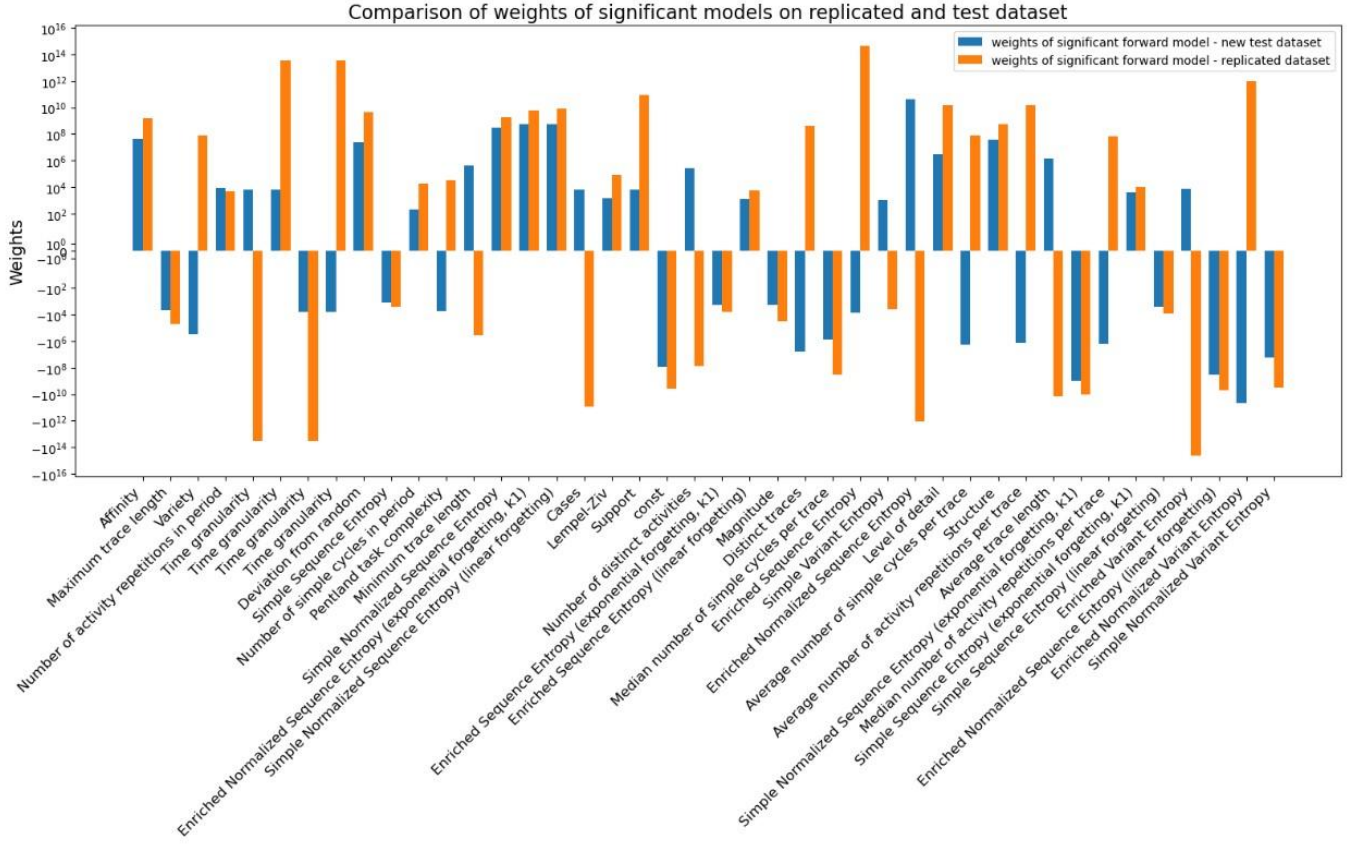
Source:
https://github.com/Frederik-Roeckle/Rep_ImpactProcessComplexity/blob/master/data/replicated_data/difference_authors_data_our_data/difference_data.xlsx

Collinearity can lead to sub-optimal regression models



https://wikis.fu-berlin.de/pages/viewpage.action?pageId=736857812

Differences in model weights



Authors linear regression models

UNIVERSITY OF MANNHEIM

APPROACH

Building statistical models to explain throughput time based on process complexity. To do so, authors performed a regression analysis with the following methodology:

Two Sets:
Complexity metrics **with** and **without** *Industry*

Automated Model Selection - Based on Akaike Information Criterion with directions:
Forward - Backward - Both

Significant Variables in Model Selection - Remove variables with p-value > 0.001

Create **Minimal Models** - At most one independent variable from categories:
Size - Variation - Distance - Entropy - Generic

Authors Evaluation Process

UNIVERSITY OF MANNHEIM

APPROACH

Before dealing with the main experiment, authors **preprocessed** and **prepared** the data:

	Dataset Used: 14 Publicly available real-life event logs from the BPIC

- **Splitting** into Time Periods
- Computed 38 **Complexity Measurements** / Time Period
- Computed **Median Throughput Time** / Time Period
- **Combine** all Logs into a single Dataset
- **Removal** of Outlier Periods