Statistical & Causal Analysis of the Chimbuko Provenance Database

Author(s) :Serges Love Teutu Talla, Margaret O Ajuwon, Isabelle Kemajou-Brown, Christopher Kelly, Kerstin Kleese Van Dam

BACKGROUND

Performance data are collected to establish how well Exascale applications are doing with executing their workflow efficiently. Chimbuko collects performance anomalies that are being detected and saved them into its Provenance Database, together with as much contextual information as needed and will be used for our analysis.

GOAL AND OBJECTIVES

Develop a set of algorithms to query data, perform analysis and visualization using Python to determine if the information collected for each anomaly is sufficient to conduct causal analysis.

METHODOLOGY

Performed correlation analysis using Theil's U correlation method, applied machine learning by regression for prediction and K-Prototypes for clustering, and ran Random Forest Model and Decision Tree for feature selection

.

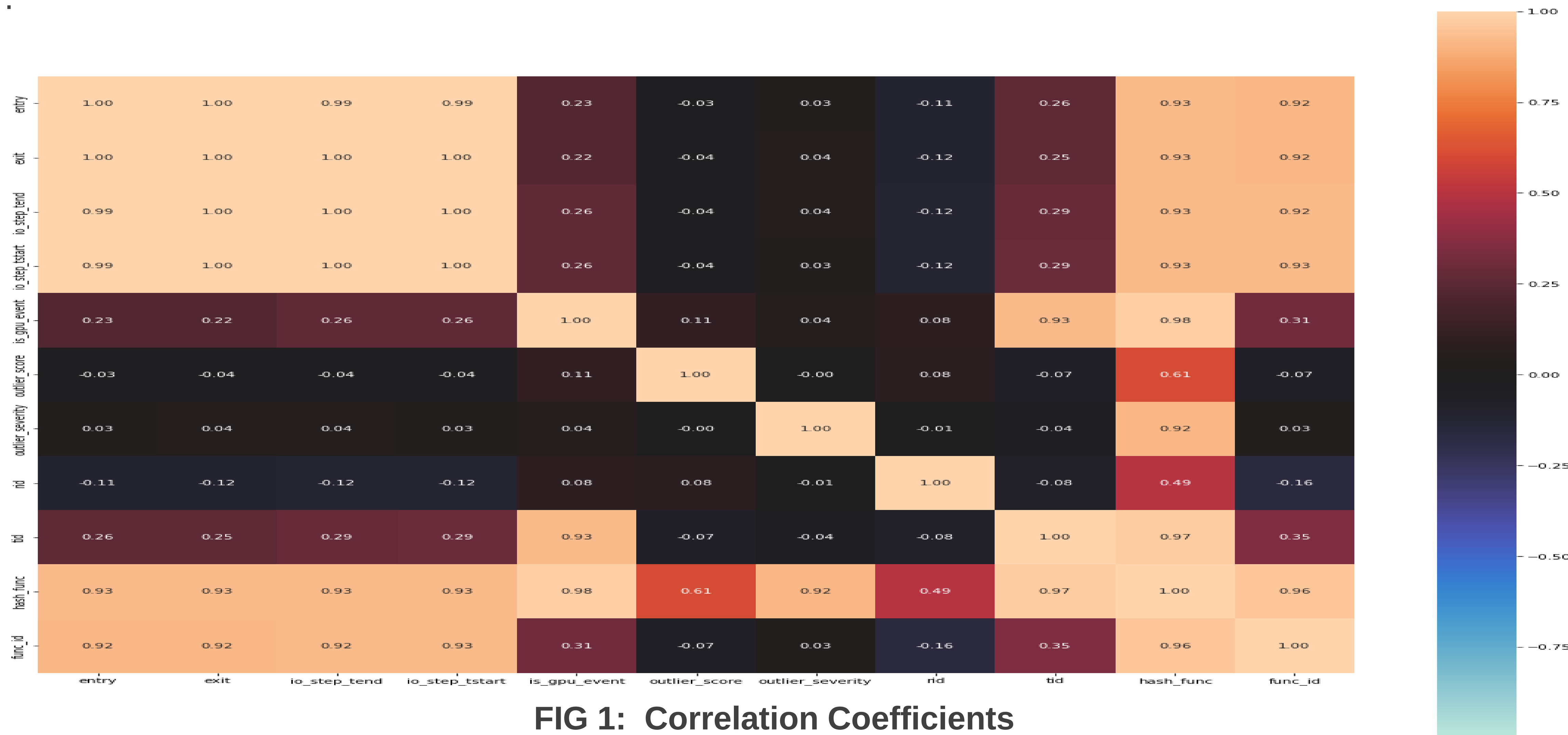

FIG 1: Correlation Coefficients

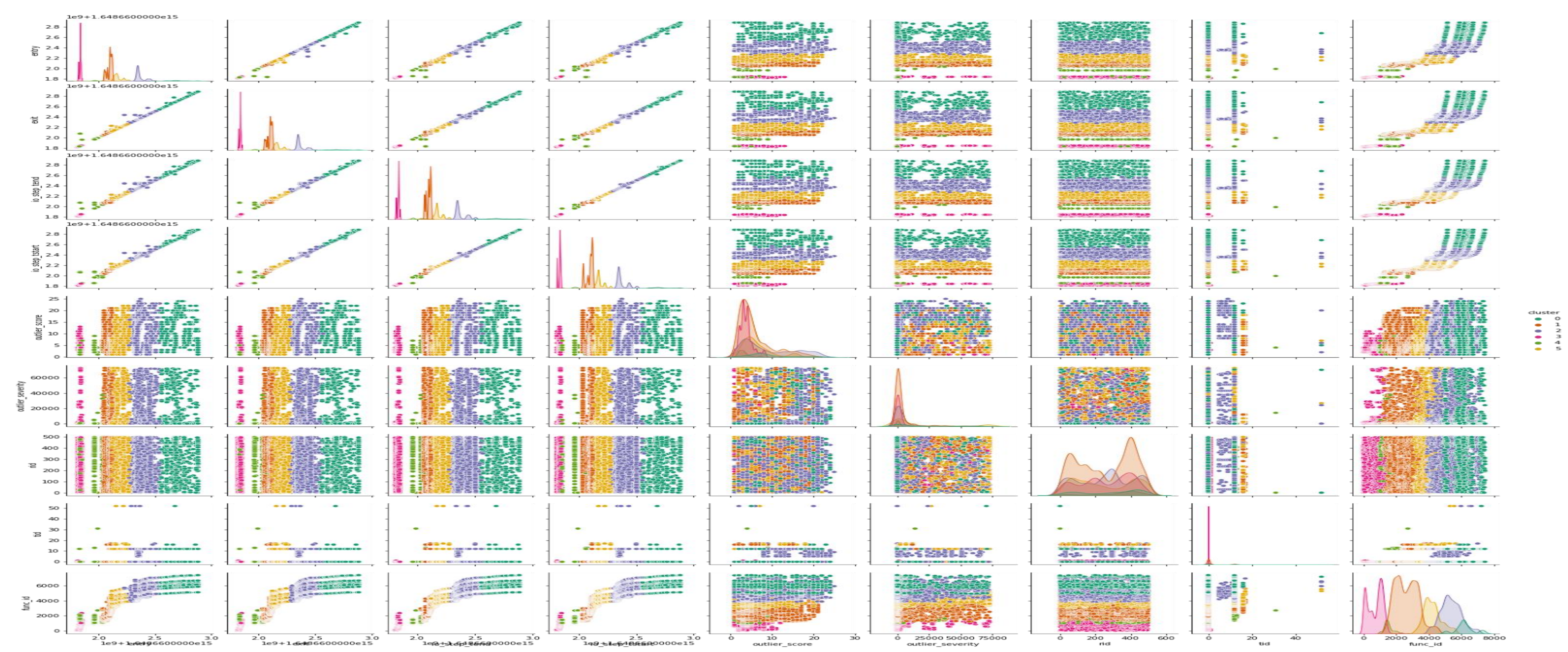

FIG 2: k-Prototypes Clustering

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                  score   No. Observations:                10944
Model:                          Logit   Df Residuals:                    10937
Method:                           MLE   Df Model:                            6
Date:                Wed, 27 Jul 2022   Pseudo R-squ.:                 0.04225
Time:                        10:26:58   Log-Likelihood:                -7248.4
converged:                      False   LL-Null:                       -7568.1
Covariance Type:            nonrobust   LLR p-value:                 7.071e-135
====================================================================================
                       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------
const             5921.4988   1.59e+05      0.037      0.970   -3.05e+05    3.17e+05
entry            -3.592e-12   9.63e-11     -0.037      0.970   -1.92e-10    1.85e-10
is_gpu_event        -3.1446      0.543     -5.787      0.000      -4.210      -2.079
outlier_severity  1.569e-05   1.29e-06     12.197      0.000    1.32e-05    1.82e-05
rid              -3.922e-05      0.000     -0.305      0.761      -0.000       0.000
runtime_total    -3.766e-09   1.71e-09     -2.208      0.027   -7.11e-09   -4.23e-10
tid                  0.3316      0.045      7.367      0.000       0.243       0.420
====================================================================================
```

FIG 3: Binary Logistic Regression (BLR) Results

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                  score   No. Observations:                10944
Model:                          Logit   Df Residuals:                    10941
Method:                           MLE   Df Model:                            2
Date:                Wed, 27 Jul 2022   Pseudo R-squ.:                 0.01678
Time:                        12:53:36   Log-Likelihood:                -7441.1
converged:                      False   LL-Null:                       -7568.1
Covariance Type:            nonrobust   LLR p-value:                  7.002e-56
====================================================================================
                       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------
const             6943.0211   1.48e+05      0.047      0.963   -2.84e+05    2.98e+05
entry            -4.211e-12   9.01e-11     -0.047      0.963   -1.81e-10    1.72e-10
outlier_severity  1.879e-05   1.24e-06     15.164      0.000    1.64e-05    2.12e-05
====================================================================================
```

FIG 4: BLR With Random Forest Model & Decision Tree

RESULTS AND CONCLUSION

- The correlation coefficient of call_stack is above 0.93 with all features except for rid (rank index), and outlier_score with the values of 0.49 and 0.61 respectively. This implies that call_stack is a very important feature of the dataset and could highlight interesting trends.
- From the scatter plot, we can conclude that rid are spread out on the entire dataset. This result suggests there is no apparent inhomogeneity between ranks in the job.
- Basic grouping performed by sum of outlier_severity shows that 'MPI_Allreduce()' has the highest sum of outlier_severity and the top five call_stack with highest sum of outlier_severity were all ending with 'OpenMP_Implicit_Task' . The outlier_sevrerity reflects the importance of an anomaly.
- Regression analysis performed before feature selection (fig 3) indicates that there is a strong evidence of association between the outcome (outlier_score) and predictors: is_gpu_event, tid, runtime_total and outlier_severity.
- Feature selection was performed on selected variables (entry, is_gpu_event, outlier_severity, rid, runtime_total, and tid) using Random Forest and Decision Tree and, showed that outlier_severity and entry are the only important features. The regression analysis results (fig 4) indicate that there is a strong evidence of association between the outlier_score and the outlier_severity.

ACKNOWLEDGEMENT

The authors thank DOE, ECP, Sustainable Horizons Institute, and Brookhaven National Lab for their support

REFERENCE

https://chimbuko-performance-analysis.readthedocs.io/en/ckelly_develop/
https://github.com/margaretajuwon/Statistical-and-Causal-Analysis-of-Chimboku-Provenace-Database

U.S. DEPARTMENT OF ENERGY

ECP EXASCALE COMPUTING PROJECT

www.bnl.gov

SUSTAINABLE HORIZONS INSTITUTE

Brookhaven National Laboratory