Metrics and Domains From the Qualitas.class corpus

A Dataset For *Metrics As Scores*

Sebastian Hönel

1 Description

This dataset was created by extracting software metrics data from the *Qualitas.class corpus* (Terra et al. 2013; Tempero et al. 2010). Therefore, the principal quantity type is `Metric`, and the context is given by a system's `Domain` (e.g., "Game", "Middleware", etc.). Some metrics were obtained on program-level, while others are package- or method-level metrics. Most of the metrics in the corpus are of discrete/integral nature.

The corpus holds 23 types of pre-computed software metrics for a total of 111 systems which are spread across eleven different domains.

This dataset has the following *discrete* **Quantity Types** (*Metrics*):

- *CA*: Afferent Coupling
- *CE*: Efferent Coupling
- *DIT*: Depth of Inheritance Tree
- *MLOC*: Method Lines of Code
- *NBD*: Nested Block Depth
- *NOC*: Number of Classes
- *NOF*: Number of Attributes
- *NOI*: Number of Interfaces
- *NOM*: Number of Methods
- *NOP*: Number of Packages
- *NORM*: Number of Overridden Methods
- *NSC*: Number of Children
- *NSF*: Number of Static Attributes
- *NSM*: Number of Static Methods
- *PAR*: Number of Parameters
- *TLOC*: Total Lines of Code
- *VG*: McCabe Cyclomatic Complexity
- *WMC*: Weighted Methods per Class

The following Metrics are continuous:

- *LCOM*: Lack of Cohesion in Methods
- *RMA*: Abstractness
- *RMD*: Normalized Distance
- *RMI*: Instability
- *SIX*: Specialization Index

It has a total of 11 **Contexts** (*Domains*): *3D; Graphics; Media*, *Databases*, *Diagrams; Visualiz.*, *Games*, *IDE*, *Middleware*, *Parsers; Generators*, *Progr. Language*, *SDK*, *Testing*, and *Tool*.

2 Analysis

In this section, results for the analysis of variance (ANOVA) and Tukey's Honest Significance Test (TukeyHSD) are shown. These tests will give a first indication as to how different the quantity types are across contexts. These two tests were used in the original paper (Hönel et al. 2022) that *Metrics As Scores* was initially conceived for.

These tests are conducted to help answering related questions, such as:

- Are there significant statistical differences for each type of quantity across all contexts?
- Is each context in its entirety (i.e., considering all types of quantities) distinguishable from the other contexts?
- What are good/bad or common/extreme scores for each context of the given dataset?

2.1 ANOVA

This test analyzes the differences among means (Chambers and Hastie 2017). For each type of quantity, this test analyzes if means of its samples are significantly different across contexts. The null hypothesis of this test is that there are *no* significant differences. This test yields a p-value and an F-statistic. The latter is the mean square of each independent variable divided by the mean square of the residuals. Large F-statistics indicate that the variation among contexts is likely. The p-value then indicates how likely it is for the F-statistic to have occurred, given the null hypothesis is true.

2.2 KS2

The Two-sample Kolmogorov–Smirnov Test (KS2) is a non-parametric and tests whether two samples stem from the same probability distribution (Stephens 1974). KS2 does not check for a certain type of probability distribution since it uses the samples' empirical CDFs. Its test statistic is the maximum vertical distance between the two CDFs. For two samples $\mathbf{x}, \mathbf{y}$, the statistic is calculated as $D_{\mathbf{x},\mathbf{y}} = \sup_t \, | \, F_{\mathcal{X}}(t) - F_{\mathcal{Y}}(t) \, |$. The null hypothesis is that the samples' CDFs are identical, that is, $F_{\mathcal{X}} = F_{\mathcal{Y}}$. This test is used to compare one type of quantity between two contexts.

2.3 TukeyHSD

This test is used to gain insights into the results of an ANOVA test. While the former only allows obtaining the amount of corroboration for the null hypothesis, TukeyHSD performs all pairwise comparisons (Tukey 1949). For example, by choosing a certain type of quantity and context, we obtain a list of other contexts that are significantly statistically different. The null hypothesis of this test is the same as for the ANOVA test.

3 Results

Here we present some insights from conducting the ANOVA- KS2-, and TukeyHSD tests.

3.1 ANOVA

Table 1 show the results of the ANOVA analysis. For each type of quantity, it indicates whether the quantity types' means vary significantly across contexts. For this test, we also add a virtual contexts, in which we simply merge the values of all contexts and effectively disregard the context. Therefore, the ANOVA test also indicates whether quantity types' values are different in a specific context when compared to all recorded values.

Table 1 shows, for each quantity type, if there was corroboration for the null hypothesis, which here is whether the samples of the same quantity type have the same mean across contexts. So, for example, samples of the first quantity type **CA** were compared across contexts *3D/graphics/media*, *IDE*, *SDK*, *database*, *diagram generator/data visualization*, *games*, *middleware*, *parsers/generators/make*, *programming language*, *testing*, *tool*, and ***ALL***. We cannot accept the null hypothesis for a significance level of $\alpha = 0.05$, meaning that the the means of samples of the quantity type CA have statistically significantly different means across contexts.

Table 1: Results of the ANOVA analysis.

Quantity Type	p-Value	F-Statistic
CA	5.68366e-15	8.40893
CE	8.45762e-33	16.4429
DIT	0	418.961
LCOM	3.2494e-137	61.2257
MLOC	0	146.497
NBD	0	1146.74
NOC	4.35712e-49	23.6001
NOF	4.7448e-49	23.5012
NOI	3.559e-50	24.0754
NOM	1.22564e-160	71.1691
NOP	9.7519e-10	6.03588
NORM	3.61094e-154	68.4245
NSC	8.35946e-09	5.47685
NSF	9.58398e-30	15.0548
NSM	2.91267e-26	13.5088
PAR	0	1171.11
RMA	1.65443e-65	30.7523
RMD	2.5677e-35	17.5561
RMI	8.35271e-77	35.6537
SIX	8.56776e-91	41.4512
TLOC	1.13942e-08	5.49735
VG	0	340.904
WMC	5.28363e-109	49.2226

3.2 KS2

The two-sample Kolmogorov–Smirnov test checks if two samples stem from an identical distribution (which is the null hypothesis). The KS2 test requires pairwise comparisons. For each type of quantity, we compare it within all the contexts. Therefore, all pair-wise combinations are computed. This dataset has 11 contexts plus one virtual (in which we effectively disregard the context, therefore, $n = 12$). The number of pair-wise comparisons per type of quantity, therefore, is $n \times (n-1) \div 2 = 66$. In other words, each type of quantity can significantly stick out anywhere between zero and 66 times. Figure 1 shows the results of the KS2 test using the significance threshold $\alpha = 0.05$.

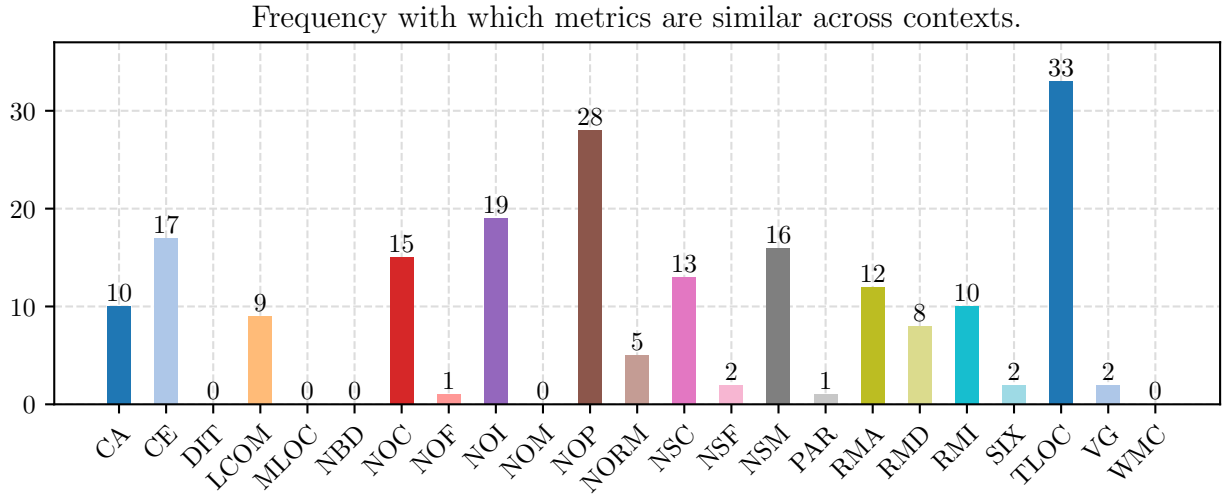

Figure 1: The frequency with which quantity types were considered to come from the same distribution. Numbers close to the max indicate that the type of quantity is not significantly different across contexts.

3.3 TukeyHSD

Per context, this test allows us to report how many types of quantities are different, compared to all other contexts' types of quantities. Consider the set of contexts C and the set of quantity types Q. Given a context c_i and a quantity type q_j, this test reports all other contexts $C^{\setminus c_i}$ that have a statistically significantly different distribution for q_j. However, for c_i, we aggregate these counts across all quantity types. We have a total of $n = 11 + 1$ contexts (including the virtual context). Therefore, the amount of quantity types ($m = 23$) that are different in different contexts can maximally be $(n-1) \times m = 253$.

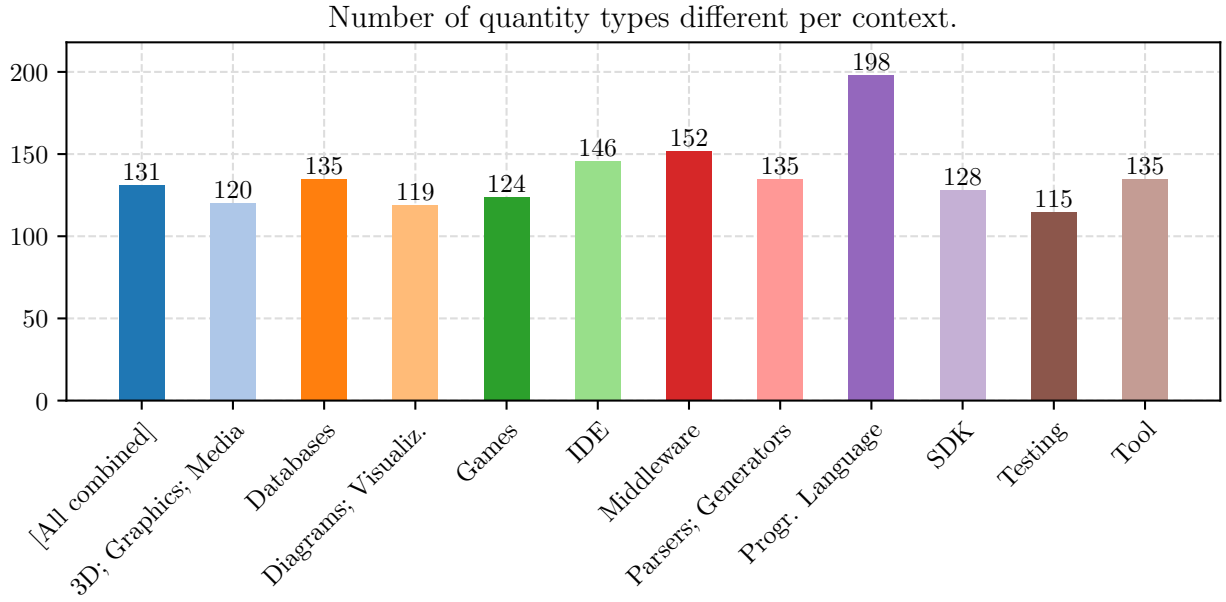

Figure 2: Number of quantity types that are different per context, by comparing each context to all other contexts.

3.4 Common/extreme Values

In this section, we demonstrate some common or extreme values by showing what (maximum) distance is required to achieve a score as shown in header of Table 2. The example shown here is that of the first type of quantity (**TLOC**), which has been transformed using each context's expection, $\mathbb{E}[X]$. Since it is practically unlikely to have uniformly distributed quantities, a linear increase in score is usually associated with a non-linear decrease in distance. Note that the values in Table 2 are approximate.

Table 2: The required maximum distance from the context's expectation $\mathbb{E}[X]$ to achieve a score less than or equal to x.

Domain	0.001	0.1	0.2	0.33	0.5	0.66	0.8	0.9	0.999	$\mathbb{E}[X]$
[All combined]	$2.4e^6$	$1.0e^5$	$5.4e^4$	$4.2e^4$	$3.0e^4$	$2.1e^4$	$1.3e^4$	6650	70	$2.3e^4$
3D; Graphics; Media	$1.9e^5$	$8.7e^4$	$6.2e^4$	$5.2e^4$	$3.7e^4$	$2.7e^4$	$1.8e^4$	$1.1e^4$	135	$3.8e^4$
IDE	$2.5e^6$	$6.4e^4$	$5.0e^4$	$3.8e^4$	$2.7e^4$	$1.8e^4$	$1.1e^4$	5270	54	$1.1e^4$
SDK	$2.1e^5$	$2.5e^4$	$2.5e^4$	$1.4e^4$	$1.1e^4$	7340	4600	2470	27	$1.0e^4$
Databases	$7.0e^5$	$3.8e^5$	$2.3e^5$	$1.7e^5$	$1.2e^5$	$9.1e^4$	$6.0e^4$	$3.5e^4$	448	$1.1e^5$
Diagrams; Visualiz.	$2.0e^5$	$1.2e^5$	$8.8e^4$	$6.7e^4$	$5.0e^4$	$3.5e^4$	$2.3e^4$	$1.3e^4$	158	$6.7e^4$
Games	$2.5e^5$	$1.6e^5$	$1.4e^5$	$1.2e^5$	$9.6e^4$	$7.2e^4$	$4.7e^4$	$2.5e^4$	286	$1.2e^5$
Middleware	$3.4e^5$	$7.8e^4$	$5.4e^4$	$4.2e^4$	$3.1e^4$	$2.2e^4$	$1.4e^4$	7910	92	$3.3e^4$
Parsers; Generators	$1.9e^5$	$9.8e^4$	$5.8e^4$	$4.2e^4$	$3.1e^4$	$2.3e^4$	$1.5e^4$	8560	106	$2.8e^4$
Progr. Language	$1.1e^6$	$6.5e^5$	$4.4e^5$	$3.2e^5$	$2.4e^5$	$1.7e^5$	$1.1e^5$	$6.0e^4$	711	$2.4e^5$
Testing	$1.0e^5$	$6.8e^4$	$5.0e^4$	$3.8e^4$	$3.0e^4$	$2.3e^4$	$1.7e^4$	$1.1e^4$	165	$3.3e^4$
Tool	$4.2e^5$	$1.3e^5$	$9.0e^4$	$7.1e^4$	$5.4e^4$	$3.9e^4$	$2.6e^4$	$1.4e^4$	176	$5.9e^4$

References

Chambers, John M, and Trevor J Hastie. 2017. "Statistical Models." In *Statistical Models in S*, 13–44. Routledge. https://doi.org/10.1201/9780203738535.

Hönel, Sebastian, Morgan Ericsson, Welf Löwe, and Anna Wingkvist. 2022. "Contextual Operationalization of Metrics as Scores: Is My Metric Value Good?" In *22nd IEEE International Conference on Software Quality, Reliability and Security, QRS 2022, Guangzhou, China, December 5-9, 2022*, 333–43. IEEE. https://doi.org/10.1109/QRS57517.2022.00042.

Stephens, M. A. 1974. "EDF Statistics for Goodness of Fit and Some Comparisons." *Journal of the American Statistical Association* 69 (347): 730–37. https://doi.org/10.1080/01621459.1974.10480196.

Tempero, Ewan D., Craig Anslow, Jens Dietrich, Ted Han, Jing Li, Markus Lumpe, Hayden Melton, and James Noble. 2010. "The Qualitas Corpus: A Curated Collection of Java Code for Empirical Studies." In *17th Asia Pacific Software Engineering Conference, APSEC 2010, Sydney, Australia, November 30 - December 3, 2010*, edited by Jun Han and Tran Dan Thu, 336–45. IEEE Computer Society. https://doi.org/10.1109/APSEC.2010.46.

Terra, Ricardo, Luis Fernando Miranda, Marco Túlio Valente, and Roberto da Silva Bigonha. 2013. "Qualitas.class Corpus: A Compiled Version of the Qualitas Corpus." *ACM SIGSOFT Softw. Eng. Notes* 38 (5): 1–4. https://doi.org/10.1145/2507288.2507314.

Tukey, John W. 1949. "Comparing Individual Means in the Analysis of Variance." *Biometrics* 5 (2): 99–114. http://www.jstor.org/stable/3001913.