When and Why Your Code Starts to Smell Bad —Additional Analyses—

Michele Tufano*, Fabio Palomba†, Gabriele Bavota‡, Rocco Oliveto§, Massimiliano Di Penta‡, Andrea De Lucia†,
Denys Poshyvanyk* *The College of William and Mary, Williamsburg, VA, USA
†University of Salerno, Fisciano (SA), Italy ‡University of Sannio, Benevento, Italy
§University of Molise, Pesche (IS), Italy

As explained in our paper, the main *construct validity* threat that could affect our results is related to the use of DECOR rules to detect smells. Indeed, our results can be affected by the presence of false positives and false negatives. Concerning false positives, Moha *et al.* reported a precision above 60% and a recall of 100% on Xerces 2.7.0. Other than relying on their assessment, we have manually validated a statistically signifiant subset of the detected smells. Such a manual validation has been performed by two of the authors independently, and cases of disagreement were discussed. In total, 1,107 smells were validated, including 241 Blob instances, 317 CDSBP, 166 Complex Class, 65 Spaghetti Code, and 318 Functional Decomposition. The results of the manual validation indicated a mean precision of 73%, and specifically of 79% for Blob, 62% for CDSBP, 74% for Complex Class, 82% for Spaghetti Code, and 70% for Functional Decomposition. In this document we replicated the analysis performed in our paper to answer the two formulated research questions ($\mathbf{RQ}_1$ and $\mathbf{RQ}_2$) by just considering the smell-introducing commits (2,555) involving smell instances that have been manually validated as true positives (and thus, only those commits that we are sure introduced a smell in the systems). This analysis was needed to verify if the main findings of our paper are confirmed when just considering commits introducing smells that have been manually verified.

In particular:

- Table I reports the descriptive statistics of the number of commits needed by each smell type to affect code files—$\mathbf{RQ}_1$.
- Table II presents the descriptive statistics (mean and median) of the slope of the regression line computed, for each metric, for both smelly and clean files. Also, Table II reports the results of the Mann-Whitney test and Cliff's d effect size (**L**arge, **M**edium, or **S**mall) obtained when analyzing the difference between the slope of regression lines for clean and smelly artifacts. Finally, column *cmp* in Table II shows a $\uparrow$ ($\downarrow$) if for the metric m there is a statistically significant difference in the m's slope between the two groups of artifacts with the smelly ones exhibiting a higher (lower) slope; otherwise, a $-$ is present in the *cmp* column—$\mathbf{RQ}_1$.
- Tables III, IV, and V report the commit-goal, project status, and developer status tags, respectively, assigned to the 2,555 investigated smell-inducing commits.

From the obtained results, all findings reported in our paper are confirmed on the set of manually validated smell-inducing commits.

Results for our $\mathbf{RQ}_1$.

TABLE I
$\mathbf{RQ}_1$: DESCRIPTIVE STATISTICS OF THE NUMBER OF COMMITS NEEDED BY EACH SMELL TYPE TO AFFECT FILES

Ecosystem	Bad Smell	avg	median	st. dev.
Android	Blob Class	17.91	0	21.15
	Class Data Should Be Private	3.75	0	11.31
	Complex Class	0.78	0	4.24
	Functional Decomposition	0.49	0	2.65
	Spaghetti Code	0.78	0	1.14

TABLE II
$\mathbf{RQ}_1$: SLOPE AFFECTED *vs* SLOPE NOT AFFECTED - MANN-WHITNEY TEST (ADJ. P-VALUE) AND CLIFF'S DELTA (d)

Smell	Affected	LOC			LCOM			WMC			RFC			CBO			NOM			NOA		
		mean	med	cmp	mean	med	cmp	mean	med	cmp	mean	med	cmp	mean	med	cmp	mean	med	cmp	mean	med	cmp
Blob	NO	0.18	0		0.14	0		0.05	0		0.07	0		0.03	0		0.06	0		0.02	0	
	YES	57.32	26.21	↑	468.40	7.84	↑	16.75	2.57	↑	18.76	2.63	↑	0.94	0.24	↑	5.37	0.94	↑	1.96	0.09	↑
	p-value	<0.01			<0.01			<0.01			<0.01			<0.01			<0.01			<0.01		
	Cliff's d	0.78 (L)			0.58 (L)			0.68 (L)			0.59 (L)			0.25 (S)			0.82 (L)			0.41 (M)		
CDSP	NO	0.23	0		0.37	0		0.07	0		0.08	0		0.1	0		0.02	0		0.03	0	
	YES	7.26	2.04	↑	5.78	0	—	0.67	0	↑	0.25	0	↑	0.71	0.23	↑	0.34	0	↑	3.18	1.00	↑
	p-value	<0.01			0.23			0.03			0.02			<0.01			0.01			<0.01		
	Cliff's d	0.65 (L)			0.02 (N)			0.05 (N)			0.07 (N)			0.19 (S)			0.02 (N)			0.79 (L)		
CC	NO	0.22	0		0.34	0		0.06	0		0.04	0		0.10	0		0.02	0		0.03	0	
	YES	46.93	15.12	↑	3033.00	115.20	↑	18.51	3.64	↑	7.78	0.78	↑	0.75	0.16	↑	10.64	1.46	↑	3.85	0.01	↑
	p-value	<0.01			<0.01			<0.01			<0.01			<0.01			<0.01			<0.01		
	Cliff's d	0.81 (L)			0.75 (L)			0.88 (L)			0.59 (L)			0.35 (S)			0.86 (L)			0.41 (M)		
FD	NO	0.31	0		0.77	0		0.04	0		0.05	0		0.15	0		0.02	0		0.06	0	
	YES	-18.14	-4.17	↓	-6.59	-0.34	↓	-3.97	-1.51	↓	-7.54	-0.61	↓	0.74	0.58	—	-1.75	-0.30	↓	-0.10	0	↓
	p-value	<0.01			<0.01			<0.01			<0.01			0.83			<0.01			<0.01		
	Cliff's d	-0.65 (L)			-0.54 (L)			-0.71 (L)			-0.46 (L)			0.02 (N)			-0.74 (L)			-0.17 (S)		
SC	NO	0.14	0		0.96	0		0.05	0		-0.01	0		0.09	0		0.02	0		0.04	0	
	YES	128.10	49.21	↑	822.30	221.4	↑	20.32	6.34	↑	11.80	0.07	↑	1.81	0	—	7.79	3.60	↑	9.64	1.06	↑
	p-value	<0.01			<0.01			<0.01			<0.01			0.43			<0.01			<0.01		
	Cliff's d	0.88 (L)			0.61 (L)			0.63 (L)			0.52 (L)			0.02 (N)			0.62 (L)			0.71 (L)		

Results for our $\mathbf{RQ}_2$.

TABLE III
$\mathbf{RQ}_2$: COMMIT-GOAL TAGS ASSIGNED TO SMELL-INDUCING COMMITS

Ecosystem	Smell	Bug Fixing	Enhancement	New Feature	Refactoring
Overall	Blob	6	53	37	4
	CDSP	4	56	30	10
	CC	19	54	23	4
	FD	13	51	27	9
	SC	4	63	22	11

TABLE IV
$\mathbf{RQ}_2$: PROJECT-STATUS TAGS TO SMELL-INTRODUCING COMMITS

Ecosystem	Smell	Working on Release				Project Startup			
		One Day	One Week	One Month	More	One Week	One Month	One Year	More
Overall	Blob	0	44	45	11	0	4	41	55
	CDSP	1	11	78	10	1	10	38	51
	CC	0	4	87	9	4	0	51	45
	FD	1	13	75	11	2	3	59	36
	SC	3	7	79	11	0	4	41	55

TABLE V
$\mathbf{RQ}_2$: DEVELOPER-STATUS TAGS TO SMELL-INTRODUCING COMMITS

Ecosystem	Smell	Workload			Ownership		New-Comer	
		High	Medium	Low	True	False	True	False
Overall	Blob	72	26	2	94	6	5	95
	CDSP	68	29	3	93	7	2	98
	CC	53	40	7	91	9	11	89
	FD	68	29	3	92	8	6	94
	SC	54	40	6	71	29	13	87