SCIENCE ▪ PASSION ▪ TECHNOLOGY

SFL for Spreadsheets: Influence of Correct Output Cells on the Fault Localization Quality

Birgit Hofer, Graz University of Technology

5th IEEE Int. Workshop on Program Debugging (IWPD), November 2014, Naples, Italy

Why spreadsheets?

- Used in nearly every company
- Basis for decisions
- Error prone
  - 3-5 % chance to make a fault in a formula
  - 88 % of spreadsheets contain faults
- Hard to debug
  - Size of spreadsheets
  - Structure hidden

Spectrum-based Fault Localization (SFL) can help to narrow down the search space.

Birgit Hofer: „SFL for Spreadsheets: Influence of Correct Output Cells on the Fault Localization Quality“
5th IEEE International Workshop on Program Debugging (IWPD 2014), Novermber 2014, Naples, Italy

Outline

1. SFL for Spreadsheets
2. Initial Situation
3. Research questions
   - **RQ1** Do spreadsheets contain correct output cells that positively or negatively influence the ranking of the faulty cells?
   - **RQ2** If yes, is it possible to a-priori determine which correct output cells would positively influence the ranking?
   - **RQ3** Is it possible to avoid a decreasing fault localization quality when adding more correct output cells?
4. Future work

Test Cases for Spreadsheets

- **Input cells:** cells that do not reference other cells

  I = {B2=23, C2=31, E2=15, B3=35, C3=34, E3=17}

- **Output cells:** any formula cell, determined by user

  O = {B4=58, C4=65, D4=123, F2=810, F3=1173}

	A	B	C	D	E	F
1		**week 1**	**week 2**	**Total**	**$/h**	**Gross Pay**
2	Green	23	31	23	15	$345,00
3	Jones	35	34	69	17	$1.173,00
4	**Total**	58	65	92		

Birgit Hofer: „SFL for Spreadsheets: Influence of Correct Output Cells on the Fault Localization Quality“
5th IEEE International Workshop on Program Debugging (IWPD 2014), Novermber 2014, Naples, Italy

SFL for Spreadsheets (1)

Program debugging: execution traces, slices
Spreadsheets: cones (borrowed from hardware debugging)

$$\mathrm{CONE}(c) = c \cup \bigcup_{c' \in \rho(c)} \mathrm{CONE}(c')$$

The function *ρ(c)* returns all cells referenced in *c*.

	A	B	C	D	E	F
1		**week 1**	**week 2**	**Total**	**$/h**	**Gross Pay**
2	Green	23	31	=SUM(B2)	15	=D2*E2
3	Jones	35	34	=SUM(B3:C3)	17	=D3*E3
4	**Total**	=SUM(B2:B3)	=SUM(C2:C3)	=SUM(D2:D3)		

CONE(F2) = {B2, D2, E2, F2}

Birgit Hofer: „SFL for Spreadsheets: Influence of Correct Output Cells on the Fault Localization Quality“
5th IEEE International Workshop on Program Debugging (IWPD 2014), Novermber 2014, Naples, Italy

SFL for Spreadsheets (2)

Spectra: Cones of erroneous and correct output cells

	A	B	C	D	E	F
1		**week 1**	**week 2**	**Total**	**$/h**	**Gross Pay**
2	Green	23	31	=SUM(B2)	15	=D2*E2
3	Jones	35	34	=SUM(B3:C3)	17	=D3*E3
4	**Total**	=SUM(B2:B3)	=SUM(C2:C3)	=SUM(D2:D3)		

CONE(F2) = {B2,D2,E2,F2}

CONE(D4) = {B2,D2,B3,C3,D3,D4}

CONE(B4) = {B2,B3,B4}

CONE(C4) = {C2,C3,C4}

CONE(F3) = {B3,C3,D3,E3,F3}

SFL for Spreadsheets (3)

CONE(F2) = {B2,D2,E2,F2}
CONE(D4) = {B2,D2,B3,C3,D3,D4}
CONE(B4) = {B2,B3,B4}
CONE(C4) = {C2,C3,C4}
CONE(F3) = {B3,C3,D3,E3,F3}

$$a_{11}(c) = |\{c' | c \in CONE(c') \wedge c' is\ erroneous\}|$$
$$a_{10}(c) = |\{c' | c \in CONE(c') \wedge c' is\ correct\}|$$
$$a_{01}(c) = |\{c' | c \notin CONE(c') \wedge c' is\ erroneous\}|$$

$$Ochiai(c) = \frac{a_{11}(c)}{\sqrt{(a_{11}(c)+a_{01}(c)) \times (a_{11}(c)+a_{10}(c))}}$$

	F2	D4	B4	C4	F3	SC	Rank.
B2	●	●	●			0.82	2
B3		●	●		●	0.41	7
B4			●			-	
C2				●		-	
C3		●		●	●	0.41	7
C4				●		-	
D2	●	●				1.00	1
D3		●			●	0.50	6
D4		●				0.71	3
E2	●					0.71	3
E3					●	-	
F2	●					0.71	3
F3					●	-	
Error	●	●					

Birgit Hofer: „SFL for Spreadsheets: Influence of Correct Output Cells on the Fault Localization Quality“
5th IEEE International Workshop on Program Debugging (IWPD 2014), Novermber 2014, Naples, Italy

Evaluation Methods

- Best Case

$$Rank_{Best} = |\{c \mid Ochiai(c) > f\}| + 1$$

- Average Case

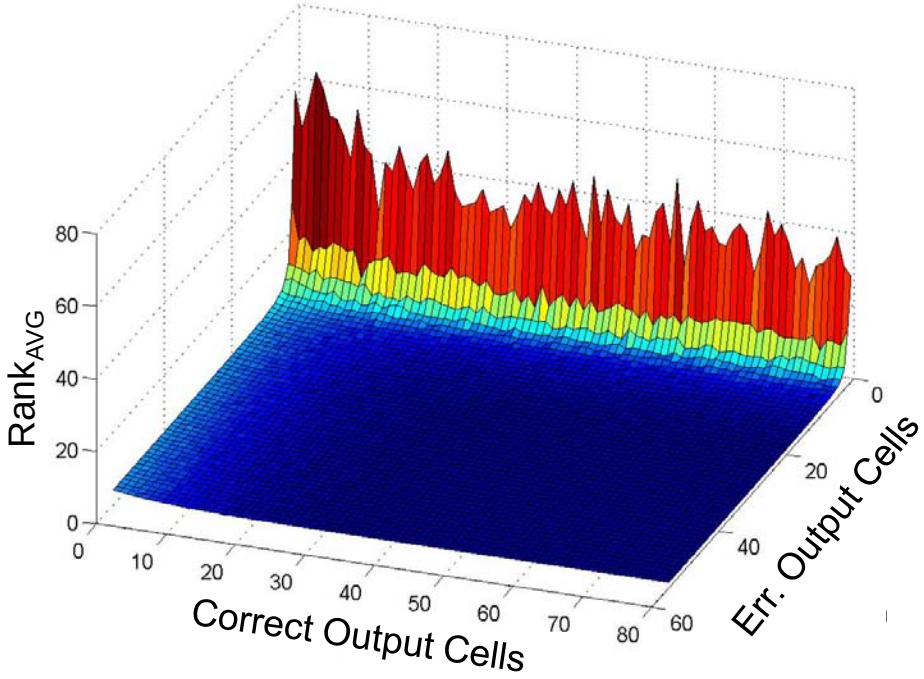
$$Rank_{AVG} = |\{c \mid Ochiai(c) > f\}| + \frac{|\{c \mid Ochiai(c) = f\}|}{2} + 0.5$$

- Worst Case

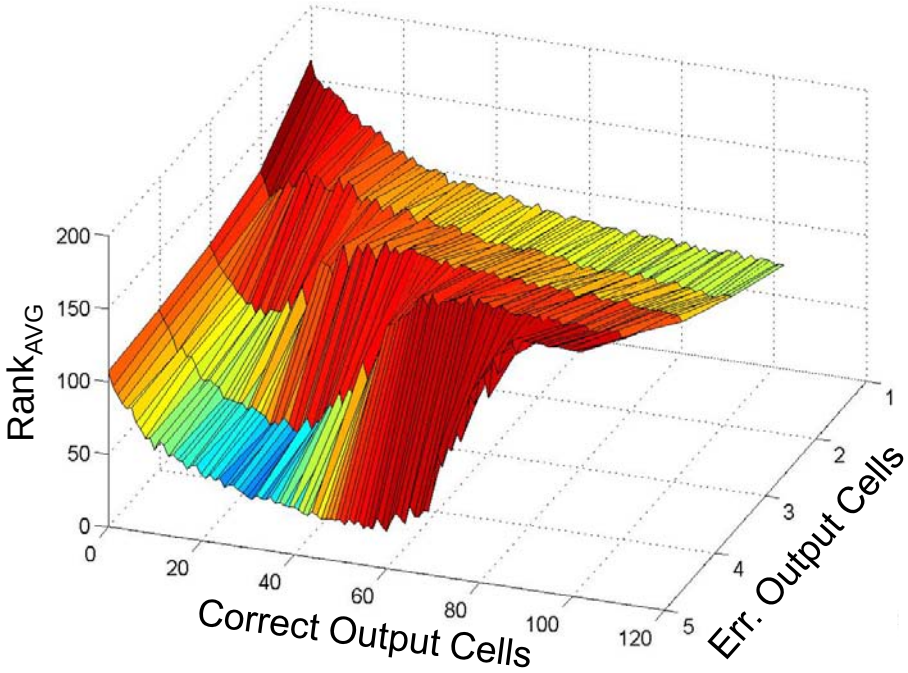
$$Rank_{WORST} = |\{c \mid Ochiai(c) \geq f\}|$$

Initial situation – Average SFL Ranking

EUSES
my_financial_model_1FAULTS_V5



ISCAS85
c7552_BOOL_tc1_96_1Fault



No user wants to indicate for so many output cells if they are correct.

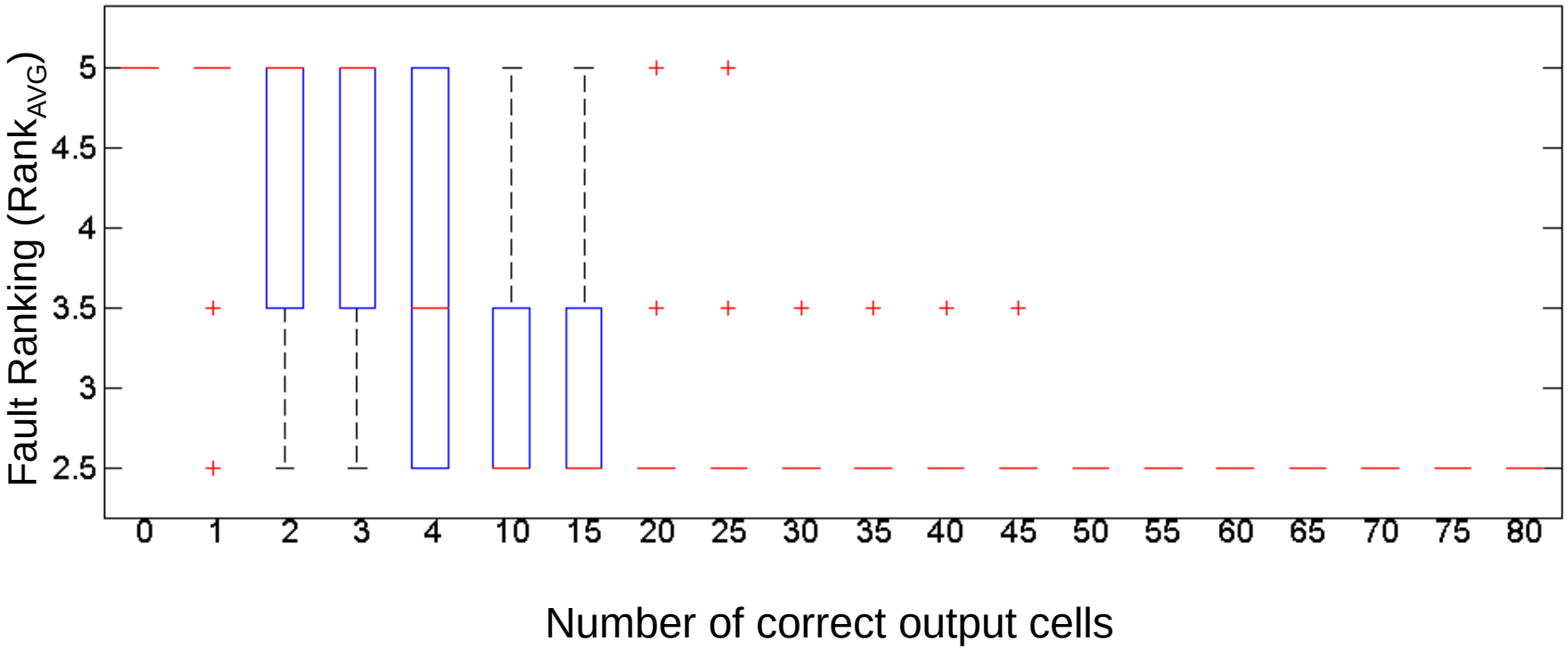
Source: Hofer, Perez, Abreu, and Wotawa: "*On the empirical evaluation of similarity coefficients for spreadsheets fault localization*", Automated Software Engineering, 2014.

Outline

1. SFL for Spreadsheets
2. Initial Situation
3. Research questions
   - **RQ1** Do spreadsheets contain correct output cells that positively or negatively influence the ranking of the faulty cells?
   - **RQ2** If yes, is it possible to a-priori determine which correct output cells would positively influence the ranking?
   - **RQ3** Is it possible to avoid a decreasing fault localization quality when adding more correct output cells?
4. Future work

RQ1: Do spreadsheets contain correct output cells that positively or negatively influence the ranking of the faulty cells?

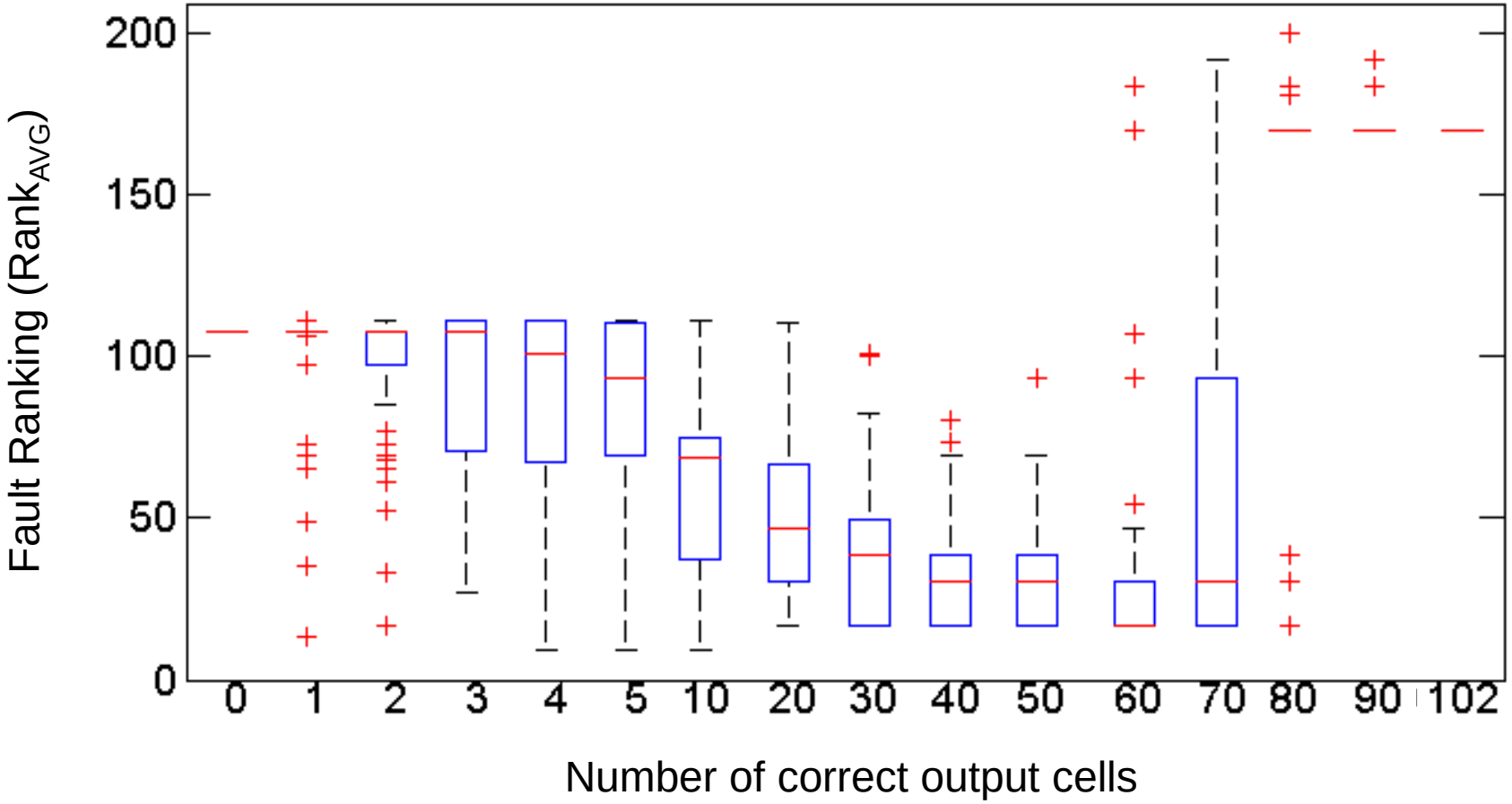
EUSES my_financial_model



Birgit Hofer: „SFL for Spreadsheets: Influence of Correct Output Cells on the Fault Localization Quality“
5th IEEE International Workshop on Program Debugging (IWPD 2014), Novermber 2014, Naples, Italy

RQ1: Do spreadsheets contain correct output cells that positively or negatively influence the ranking of the faulty cells?

ISCAS85 c7552



Birgit Hofer: „SFL for Spreadsheets: Influence of Correct Output Cells on the Fault Localization Quality“
5th IEEE International Workshop on Program Debugging (IWPD 2014), Novermber 2014, Naples, Italy

RQ2: If yes, is it possible to a-priori determine which correct output cells would positively influence the ranking?

- Avoid coincidental correct output cells
  - A-priori definition not possible
  - Too many potential coincidental correct output cells

- Take output cells with largest cones first

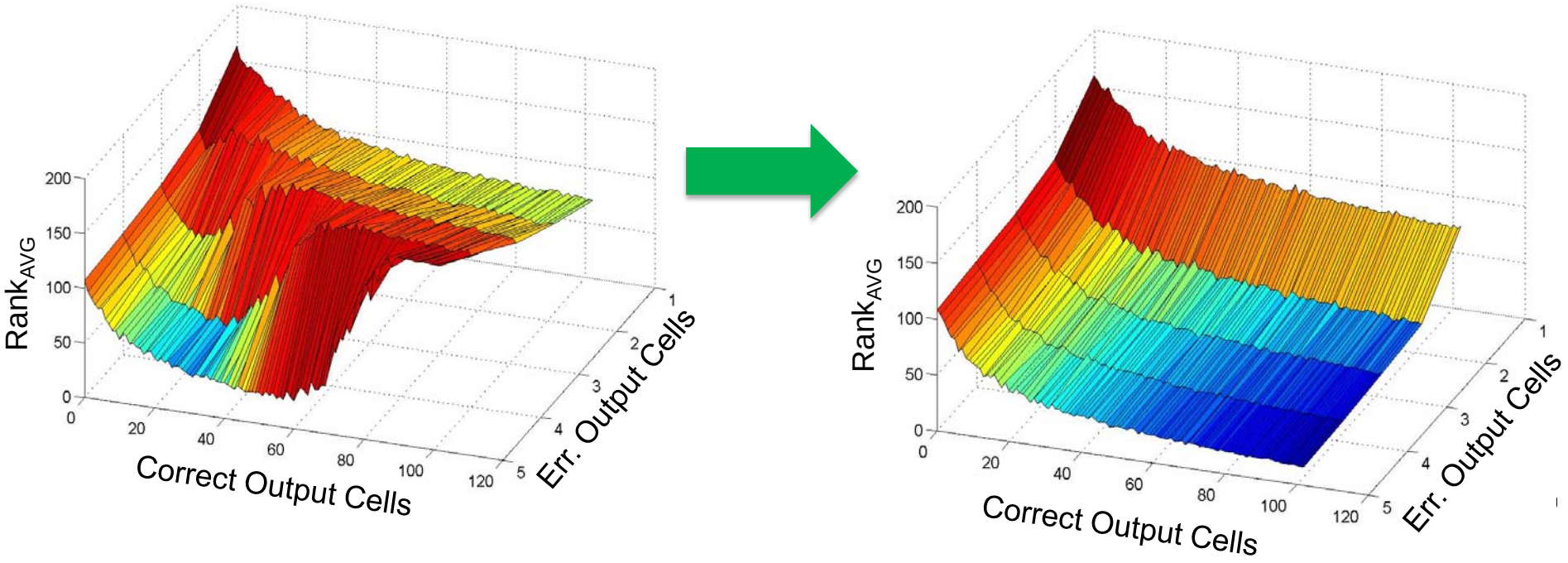
$Rank_{AVG}$ for one correct output cell	Random selection	Largest cone
EUSES my_financial	5.8	2.5
ISCAS85 c7552	100.6	69.5

Birgit Hofer: „SFL for Spreadsheets: Influence of Correct Output Cells on the Fault Localization Quality“
5th IEEE International Workshop on Program Debugging (IWPD 2014), Novermber 2014, Naples, Italy

RQ3: Is it possible to avoid a decreasing fault localization quality when adding more correct output cells?

→Balance ratio of correct and erroneous output cells
→Duplicate cones of erroneous output cells

ISCAS85 c7552



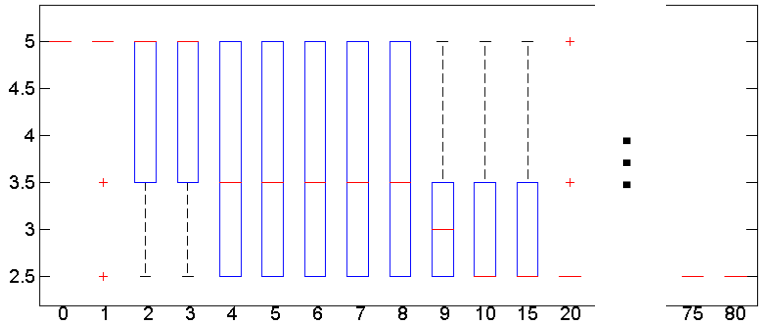
Future work

- Extend evaluation
  - All EUSES and ISCAS85 spreadsheets
- Weighting instead of duplication
  - Wong's Heuristic III
- Several faults
  - Influence on ranking
  - Clustering

Summary

RQ1 Do spreadsheets contain correct output cells that positively or negatively influence the ranking of the faulty cells?

YES



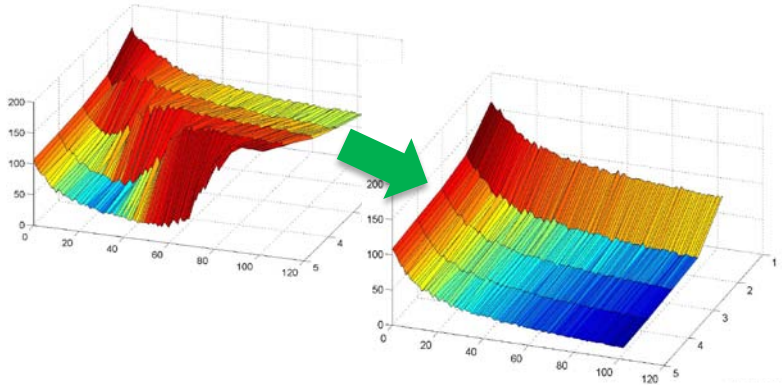
RQ2 If yes, is it possible to a-priori determine which correct output cells would positively influence the ranking?

YES (use largest cones first)

$Rank_{AVG}$ (1 correct output cell)	Random	Largest cone
EUSES my_financial	5.8	2.5
ISCAS85 c7552	100.6	69.5

RQ3 Is it possible to avoid a decreasing fault localization quality when adding more correct output cells?

YES (duplicate cones of erroneous output cells)



Birgit Hofer: „SFL for Spreadsheets: Influence of Correct Output Cells on the Fault Localization Quality“
5th IEEE International Workshop on Program Debugging (IWPD 2014), Novermber 2014, Naples, Italy