Complexity Metrics for Spreadsheet Models

Andrej Bregar

University of Maribor

Faculty of Electrical Engineering and Computer Science

Introduction

- According to conducted research studies, up to 60% of spreadsheets produce wrong outputs.
  - Most spreadsheets are large and complex.
  - Formal development strategies are rarely applied.
- In order to reduce risks, complexity of models or particular formulae should be measured.
  - Complex spreadsheets make error finding difficult.
  - Errors come in relation with cells that have a high potential for faults.
- Several complexity metrics are proposed.
  - Product and process metrics are extensively used in traditional software engineering.

Complexity metrics

Formula size •Number of operators •Number of operands	*Cell range* •Range width (height)
Formula structure •Nesting level of a token •Average nesting level •Depth of nesting •Decision count	*Cell cascade* •Cell fan-in (precedents) •Cell fan-out (dependents) •Reachability of a cell •Average reachability •Average path length •Maximal path length •Total number of paths
Cell references of a formula •Dispersion of references •Column (row) span •Column (row) reference delta •Maximal positive (negative) delta	*Modular structure* •Data binding triples •Unreferenced data cells

Formula structure

- Decision count.
  - Number of simple conditions within a formula.
- Depth of nesting.
  - Nesting occurs because each function operand can be a result of another function.
- Average nesting level.

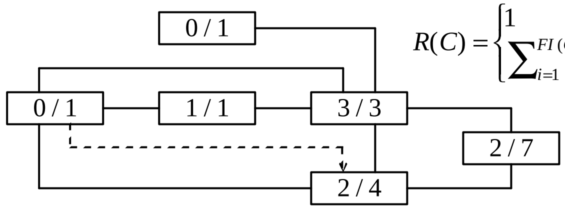
$$NL_{Avg} = \frac{\sum_{i=1}^{N_1+N_2} NL_i}{N_1 + N_2}$$

N_1 - number of operators
N_2 - number of operands
NL_i - nestig level of i-th token

- Research question: “Does reducing nesting improve modifiability and auditability?”

Cell cascade

- Cell fan-in (precedents).
  - Number of direct links that lead in a cell.
  - Count of references to another cells.
- Cell fan-out (dependents).
  - Number of direct links that lead out of a cell.
- Reachability of a cell.
  - Each path is a sequence of arcs from one of the start cells to the relevant terminal cell.



$$R(C) = \begin{cases} 1 & , FI(C) = 0 \\ \sum_{i=1}^{FI(C)} R(C_i) & , \text{otherwise} \end{cases}$$

Cell references of a formula (1/2)

- Dispersion of references.

$$DR = 1 - \exp(-a \cdot \Delta)$$

$$\Delta = \sum_{i=1}^{N} |DX_i \cdot DY_i|$$

DX_i and DY_i are i-th reference deltas

  - Error rates rise when equations contain references to cells that are in both different columns and rows than a cell containing the formula.
  - Readability is not influenced linearly by distances.
  - Constant *a* sets the slope of the dispersion curve.
    - It has an order of magnitude of 10^{-2}.
    - It determines which distance sums are reasonable in terms of quality characteristics.

Cell references of a formula (2/2)

- Additional equations to estimate the degree of dispersion will be defined.
  - Manhattan and Euclidean distances will be applied in place of the $|DX_i \cdot DY_i|$ product.
  - Since references can be balanced or unbalanced, angles between 2-D vectors of references will be calculated.
- Column and row spans.
  - These measures supplement the dispersion.
  - A matrix of only about twenty times twenty cells is visible to the spreadsheet developer or user.

Cell ranges and modules

- Width and height of cell ranges.
  - Each cell in a range is accessed and processed separately.
  - Ranges tend to be more auditable than a group of cells with different formulae, but they exhibit a risk potential because of possible invalid references.
- Data binding triples.
  - The sharing of data among modules.
- Percentage of unreferenced data cells.
  - Every data cell or range has to be referenced, because all input values must be analysed.

Measuring error rates (1/2)

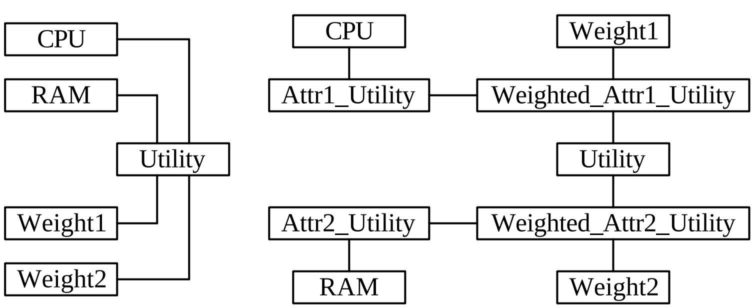
- Cell error rate.
  - Percentage of non-label cells containing errors.
  - It is estimated to be between 1% and 2%.
- Bottom-line error rate.

$$E = 1 - (1 - e)^N$$

  - Error rates multiply along cascades of subtasks.
    - Bottom-line values are computed through cascades of formula cells.
    - Any cell error leads to an incorrect result.

Measuring error rates (2/2)

- Complex formulae are considerably more liable to errors than simple formulae or cells containing data.
- Reliability of cascades can be reasonably estimated only if cell error rate is adjusted accordantly with the complexity of each individual cell!



$e = 0.02$, $N = 5 \rightarrow E = 0.0961$, $N = 9 \rightarrow E = 0.1663$

Conditional constructs (1/2)

- If conditions are nested on many levels, formulae become overly complex.
  - Reduced auditability, modifiability and reliability.
- Approaches to enabling efficient branching.
  - Condition block.
    - It is a slight modification of LOOKUP.
    - It returns the value of a cascade belonging to the positively assessed condition.
    - Bottom-line operations of cell paths are declared.
  - Reference branching condition cell.
    - It declares two “forward” references.
    - Paths that are not executed should be shaded in a predefined way.

Conditional constructs (2/2)

- Complexity of logical structure.

$$O(S) = \left(\sum_{i=1}^{M} O(S_i) + N \right)^{1+b}$$

M - nested conditionals
N - conditionless paths
$O(S)$, $O(S_i)$ - complexities

- If $b = 0$, number of logically disjunctive branches leading to the same bottom-line cell is returned.
- If $b \approx 10^{-1}$, two factors are considered.
  - Perceived complexity increases, if conditionals are nested on many levels of computational cascades.
  - Different degrees of risk should be associated with various types of conditional constructs.

Directions for further work

- The proposed metrics will be:
  - applied to actual spreadsheets and validated,
  - substituted with more appropriate metrics,
  - supplemented with additional complexity factors,
  - correlated to quantitative process measurements,
  - correlated to cell error rates,
  - used in an automated analysis tool.