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

<table>
<thead>
<tr>
<th>Formula size</th>
<th>Cell range</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Number of operators</td>
<td>• Range width (height)</td>
</tr>
<tr>
<td>• Number of operands</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Formula structure</th>
<th>Cell cascade</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Nesting level of a token</td>
<td>• Cell fan-in (precedents)</td>
</tr>
<tr>
<td>• Average nesting level</td>
<td>• Cell fan-out (dependents)</td>
</tr>
<tr>
<td>• Depth of nesting</td>
<td>• Reachability of a cell</td>
</tr>
<tr>
<td>• Decision count</td>
<td>• Average reachability</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cell references of a formula</th>
<th>Modular structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Dispersion of references</td>
<td>• Data binding triples</td>
</tr>
<tr>
<td>• Column (row) span</td>
<td>• Unreferenced data cells</td>
</tr>
<tr>
<td>• Column (row) reference delta</td>
<td></td>
</tr>
<tr>
<td>• Maximal positive (negative) delta</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cell cascade</th>
<th>Modular structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Cell fan-in (precedents)</td>
<td>• Data binding triples</td>
</tr>
<tr>
<td>• Cell fan-out (dependents)</td>
<td>• Unreferenced data cells</td>
</tr>
<tr>
<td>• Reachability of a cell</td>
<td>• Average reachability</td>
</tr>
<tr>
<td>• Average reachability</td>
<td></td>
</tr>
<tr>
<td>• Average path length</td>
<td></td>
</tr>
<tr>
<td>• Maximal path length</td>
<td></td>
</tr>
<tr>
<td>• Total number of paths</td>
<td></td>
</tr>
</tbody>
</table>

### 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
  - \(NL_i\) - nesting 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.

\[
R(C) = \begin{cases} 
1 & F(C) = 0 \\
\sum_{i=1}^{F(C)} R(C_i) & \text{otherwise}
\end{cases}
\]

---

**Cell references of a formula (1/2)**

- Dispersion of references.

\[
DR = 1 - \exp(-\alpha \cdot \Delta)
\]

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

- 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.
  \[ O(S) = \left( \sum_{i=1}^{M} O(S_i) + N \right)^{1+\beta} \]
  \[ M - \text{nested conditionals} \]
  \[ N - \text{conditionless paths} \]
  \[ O(S), O(S_i) - \text{complexities} \]
- If \( \beta = 0 \), number of logically disjunctive branches leading to the same bottom-line cell is returned.
- If \( \beta = 10^{-1} \), 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.