1 Supplementary Material for *Creating a Historical Migration Dataset from Finnish Church Records, 1800–1920*

This supplementary material provides further explanation, examples, and analysis related to the table structure detection process described in the main article. Although the method works reliably across a variety of historical sources, Finnish church records from the 19th and early 20th centuries present certain challenges. These include inconsistent layouts, mixed formatting styles, and varying document quality.

The examples included here highlight some of the more difficult cases and common errors that occur when detecting tables, rows, and columns in both printed and hand-drawn formats. This material helps clarify where current methods may fall short and where improvements can be made for historical document collections.

1.1 Table Detection

The method for detecting tables generally works well, but some recurring errors were observed:

- **Split Two-Page Tables**: A frequent issue arises when a single table spans two facing pages. The model sometimes detects each page as a separate table, missing the fact that they form a continuous whole. This leads to two detections for what is actually one table (Figures 1 and 2).
- **Overlapping Tables on Sparse Pages**: When only part of a table contains visible content—such as a few entries or marks—the model may detect both the partially filled area and the full table layout as separate tables. This leads to overlapping detections on the same page (Figure 3).

These errors suggest that the model has difficulty identifying structural continuity across pages and may respond to low-density content in ways that produce extra or conflicting detections.

1.2 Row Detection

Row detection works in most cases, but some error patterns were identified for both printed and hand-drawn tables:

- **False Detection of Empty Rows**: On pages with many empty cells, the model sometimes detects rows that do not exist. This may happen when horizontal lines or gaps are interpreted as row boundaries. The result is an inflated row count and incorrect data structure (Figures 4 and 5).

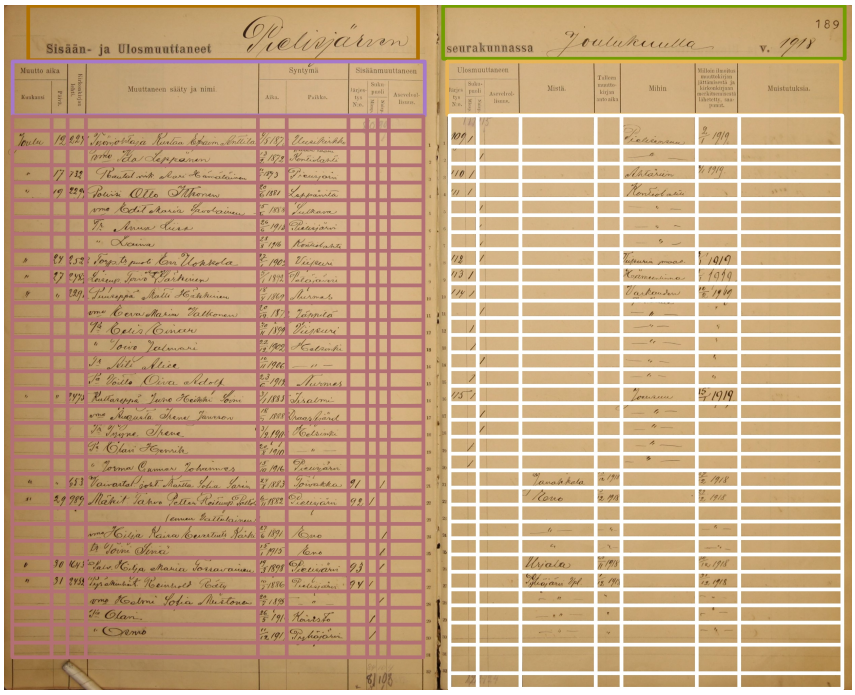

Figure 1: Two-page table detected as two separate tables (printed).

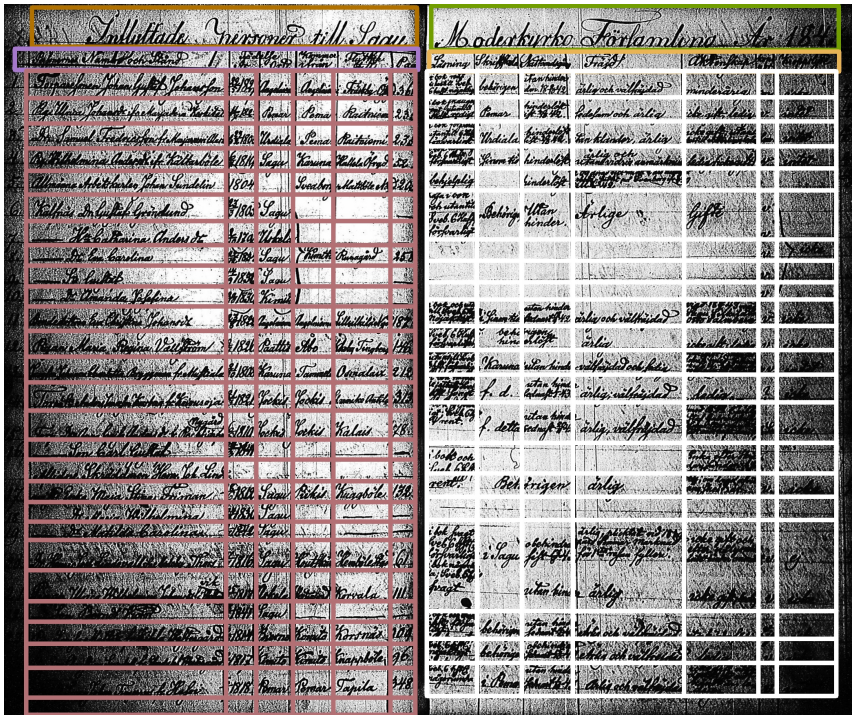

Figure 2: Two-page table detected as two separate tables (hand-drawn).

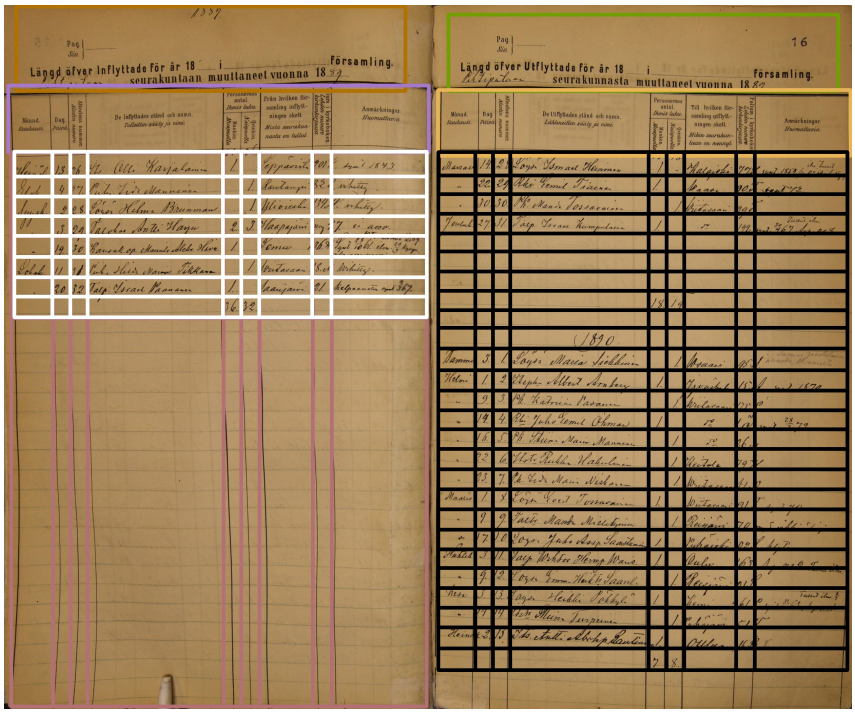

Figure 3: Overlapping table detections on a sparse page (printed).

- **Missing Final Row**: The last row on a page is sometimes missed, especially if it is close to the page margin or if the visual cues marking its boundary are weak. This affects the completeness of the detected data (Figures 6 and 7).
- **Misplaced Row Boundaries with Multi-line Text**: When a cell contains more than one line of text, spacing inside the cell may be mistaken for row boundaries. This can lead to rows being split or misaligned. Examples of this issue appear in both printed and hand-drawn formats (Figures 8 and 7).

These observations show that interpreting structure on sparse or visually ambiguous pages remains difficult for current models.

1.3 Column Detection

Column detection is generally accurate, but some common problems were noted:

- **Missing Edge Columns**: The model may fail to detect the leftmost or rightmost column. This is more likely when the column border is faint, broken, or very close to the edge of the page. An example is shown in Figure 6.
- **Single Column Split into Two**: A wide column with varied content may be mistakenly split into two separate columns. This leads to extra divisions that do not reflect the actual table structure (Figures 9 and 10).

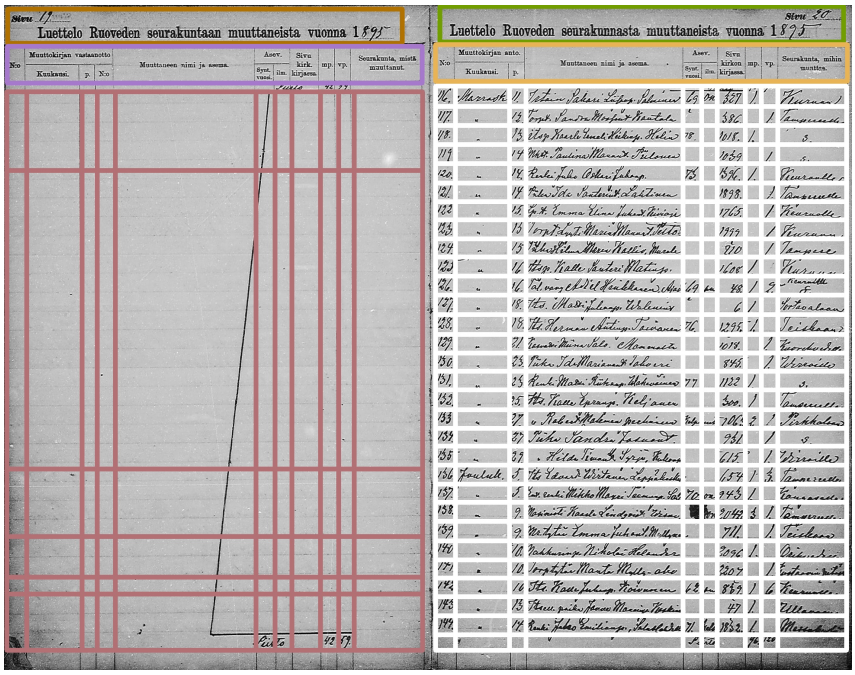

Figure 4: False detection of empty rows (printed).

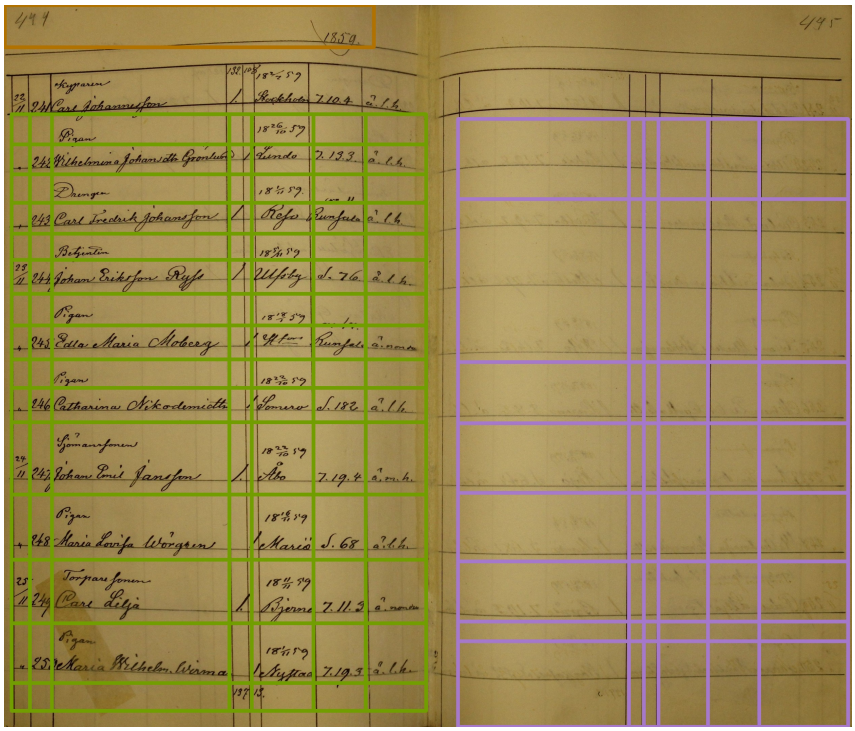

Figure 5: False detection of empty rows (hand-drawn).

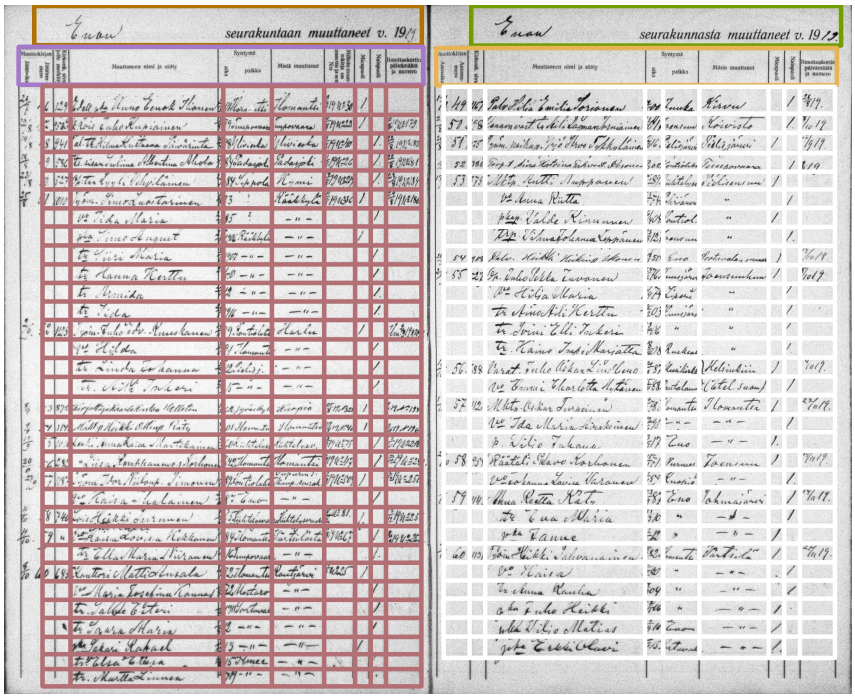

Figure 6: Missing final row (printed).

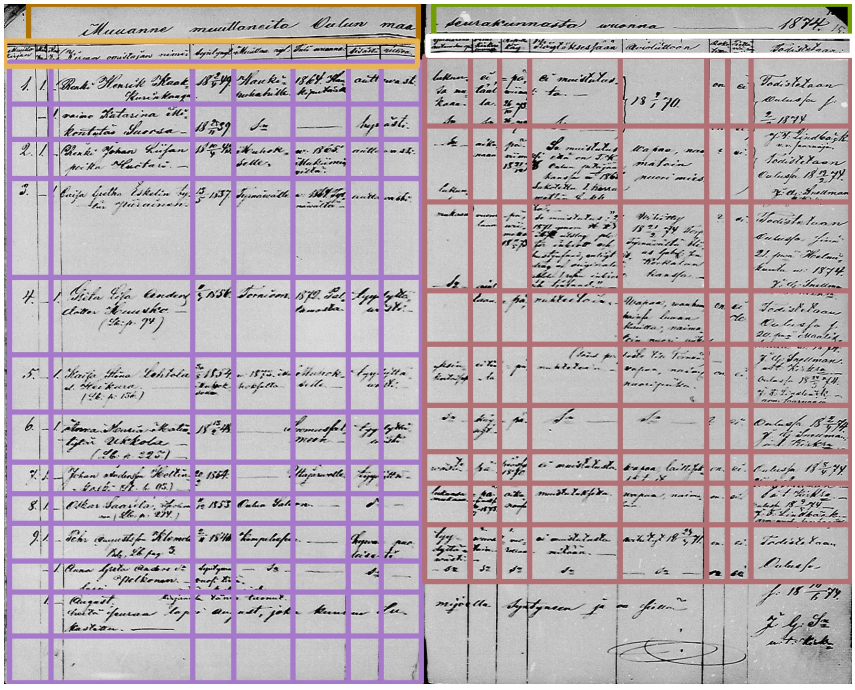

Figure 7: Missing final row and misplaced row boundaries with multi-line text (hand-drawn).

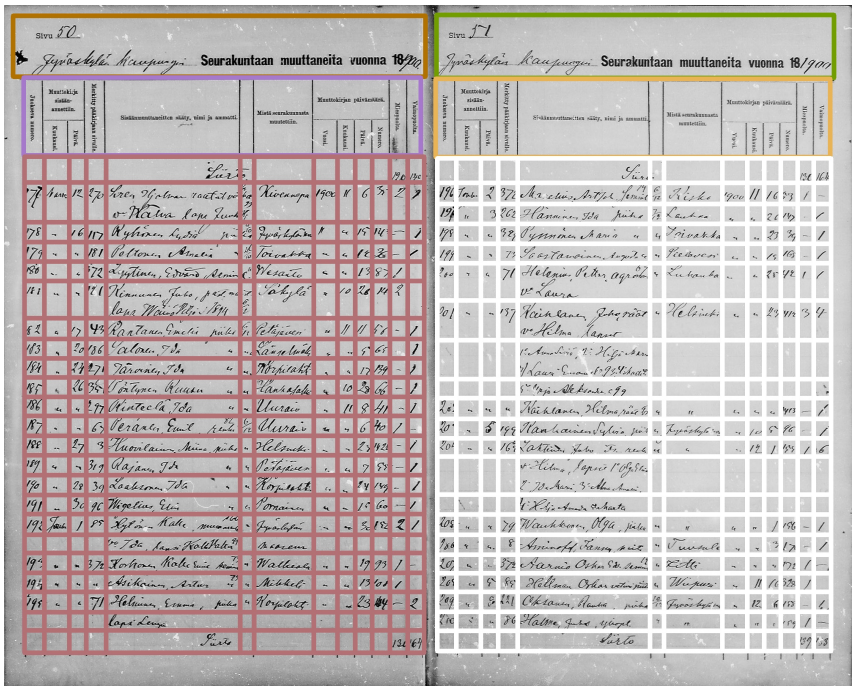

Figure 8: Misplaced row boundaries due to multi-line text (printed).

- **Two Columns Merged into One**: When two narrow columns are placed closely together, and the line separating them is unclear or missing, the model may treat them as one column. This leads to grouping of content that should be kept separate (Figures 5 and 6).

These issues point to difficulties in detecting vertical structure when borders are weak or inconsistent.

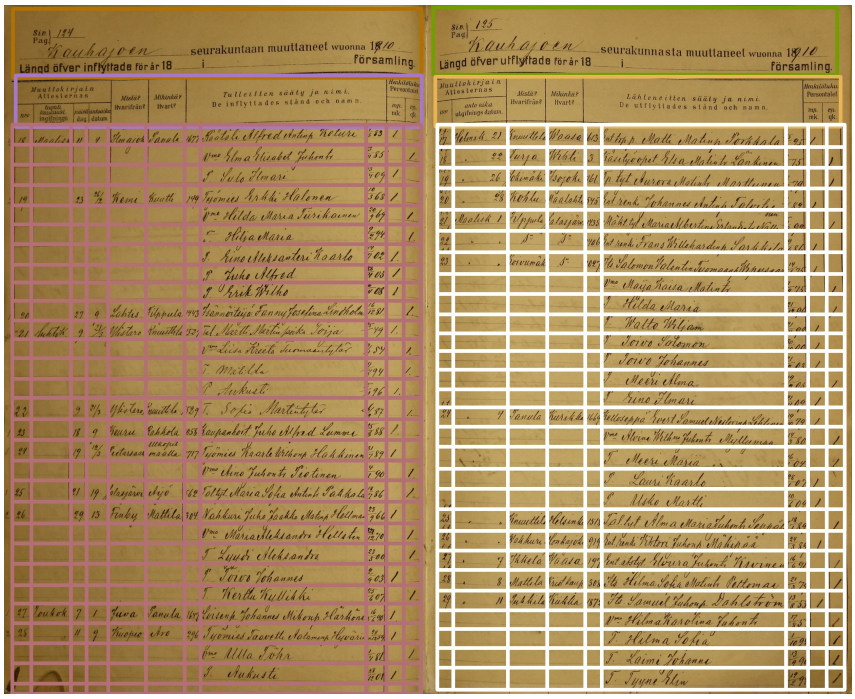

Figure 9: Single column split into two columns (printed).

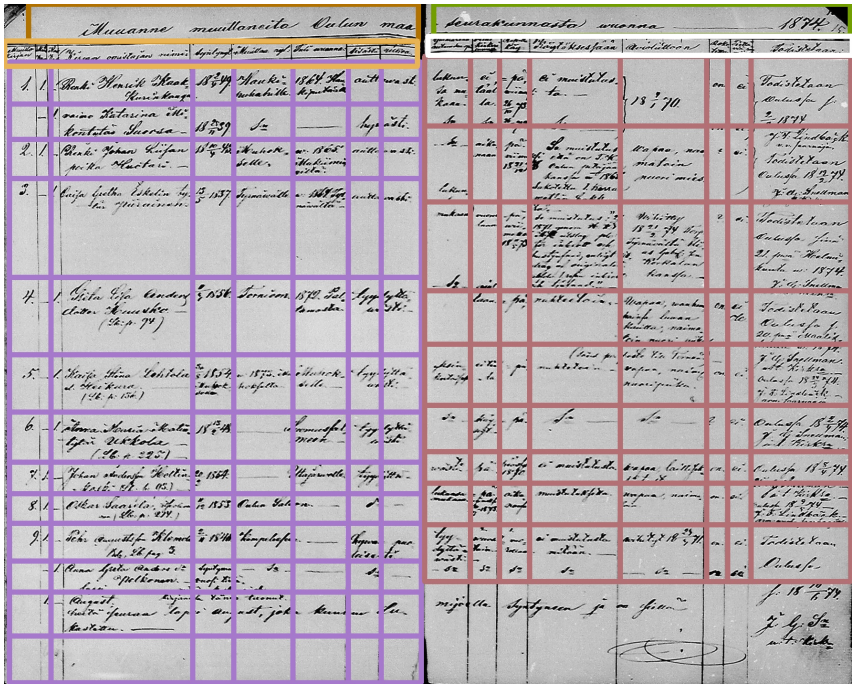

Figure 10: Single column split into two columns (hand-drawn).