CombiTab

Version 1.0

Archaeological Seriation & Combination Table Tool

Comprehensive User Documentation

A Streamlit-based application for creating, editing, and analyzing archaeological seriation diagrams (combination tables)

Table of Contents

1. Introduction

1.1 What is CombiTab?

CombiTab is a specialized software application designed for archaeologists to create, edit, and analyze seriation diagrams, commonly known as combination tables or Kombinationstabellen. These visual tools are fundamental to archaeological chronology, allowing researchers to establish relative time sequences based on the co-occurrence patterns of artifact types across different archaeological contexts.

The application is built using Streamlit, a modern Python web framework, and integrates powerful data analysis libraries including Pandas for data manipulation, NumPy and SciPy for mathematical computations, and Plotly for interactive visualizations. This combination provides a user-friendly interface while maintaining the computational rigor required for professional archaeological analysis.

1.2 Historical Background of Seriation

Seriation as an archaeological method dates back to the late 19th century. Sir Flinders Petrie pioneered the technique in 1899 when analyzing Egyptian predynastic graves. By arranging graves according to the overlapping distributions of pottery types, Petrie created a relative chronological sequence without relying on absolute dating methods.

The method rests on two fundamental assumptions. First, artifact types have limited lifespans: they are introduced at some point in time, reach peak popularity, and eventually fall out of use. Second, contexts (such as graves or settlement layers) that share similar artifact assemblages are likely to be roughly contemporary. When contexts are properly ordered, artifact types should display a characteristic pattern where each type occurs in a continuous block without gaps, forming the distinctive diagonal pattern visible in well-seriated combination tables.

1.3 Core Concepts

Understanding CombiTab requires familiarity with several key concepts:

The combination matrix is the fundamental data structure. Rows represent archaeological contexts (such as graves, pits, or stratigraphic layers), while columns represent artifact types (such as fibula types, pottery forms, or weapon categories). Each cell indicates whether a particular type is present in a particular context. The matrix can use binary presence/absence values (0 or 1) or frequency counts.

Seriation is the process of rearranging both rows and columns to achieve the optimal ordering. The goal is to create a pattern where filled cells concentrate along the diagonal, and each type forms a continuous block without gaps. This ordering should ideally reflect the chronological sequence of the contexts.

Index types (German: Leittypen) are artifact types with particularly distinctive chronological distributions. These types are especially useful for establishing relative chronology because they appear only during specific time periods and thus serve as chronological markers.

2. Getting Started

2.1 System Requirements

CombiTab runs as a web application in any modern browser. For local installation, the following Python packages are required: Streamlit (web framework), Pandas (data handling), NumPy and SciPy (numerical computation), Matplotlib and Plotly (visualization), and scikit-learn (for MDS analysis). The application is optimized for matrices containing up to several hundred rows and columns.

2.2 Launching the Application

To start CombiTab, run the following command in your terminal:

```
streamlit run combitab.py
```

The application will open in your default web browser. The interface is divided into two main areas: a sidebar for project management and data import, and a main content area with multiple tabs for different functions.

2.3 User Interface Overview

The sidebar provides access to project naming, data import (CSV or Excel files), and project file management (save/load). The main area contains six tabs:

- Matrix View: The primary working area showing the combination table visualization with sorting controls and manual ordering tools.
- Analysis: Statistical analysis tools including Correspondence Analysis and quality metrics.
- Cell Annotations: Interface for adding certainty, fragmentation status, and notes to individual cells.
- Column Labels: Metadata management for artifact types including material group assignment.
- Row Labels: Metadata management for contexts including context type classification.
- Export: Options for exporting the matrix as images or data files.

3. Data Import and Project Management

3.1 Data Format Requirements

CombiTab accepts data in CSV or Excel format. The expected structure places context names in the first column (which becomes the row index) and artifact type names as column headers. Cell values can be either binary (0/1) for presence/absence data or integers for frequency counts. The application automatically detects the data type upon import.

Example data structure:

Context	Fibula_A	Fibula_B	Pot_Type1	Beads
Grave_001	1	0	1	0
Grave_002	1	1	1	1

3.2 Project Files

CombiTab uses JSON-based project files that store the complete project state. This includes the data matrix, row and column orderings, all metadata (material groups, context types), cell annotations, and visualization settings. Project files enable you to save your work and continue later without losing any information.

To export a project, click the Export Project button in the sidebar. The resulting JSON file can be shared with colleagues or archived for future reference. To import a previously saved project, use the Import Project file uploader in the sidebar.

4. The Matrix View

4.1 Visualization Styles

CombiTab offers three distinct visualization styles for rendering the combination matrix:

Classic Squares: The traditional representation where filled black squares indicate presence and white squares indicate absence. For frequency data, cell darkness scales with count. This style is most commonly used in archaeological publications.

Battleship Bars: Horizontal bars within each cell represent values, with bar height proportional to frequency. This style is particularly effective for visualizing quantity variations across the matrix.

Sized Dots: Circular markers scaled by frequency. This style provides clear visual distinction between different count levels while maintaining a clean appearance.

4.2 Interactive Features

The interactive Plotly-based view provides powerful navigation tools for exploring large matrices. Scroll to zoom in and out, click and drag to pan across the matrix, and double-click to reset the view to its original state. The toolbar provides additional controls including a home button to reset the view, zoom buttons for precise control, and a camera icon to download the current view as an image.

Hovering over any cell displays detailed information including the context name, artifact type, value, material group, and any annotations that have been added to that cell.

4.3 Material Group Colors

The header row displays color-coded bars indicating the material group of each artifact type. Default material groups include Ceramic (tan), Metal (steel blue), Glass (teal), Bone/Antler (burlywood), Stone (gray), and Organic (brown). You can customize these colors and add new material groups in the Column Labels tab.

Index types (Leittypen) are marked with a star symbol in the header, making them easily identifiable in the visualization.

4.4 Manual Ordering

The right sidebar provides controls for manually adjusting the order of rows and columns. Select any row or column from the dropdown menus, then use the Up/Down or Left/Right buttons to move it within the sequence. This allows fine-tuning of the seriation beyond what automatic algorithms achieve.

Rows and columns can be fixed in place using the lock checkbox. Fixed items maintain their position during automatic sorting operations, which is useful when stratigraphic constraints or other external information must be respected.

5. Seriation Algorithms and Mathematical Foundations

5.1 The Centroid Method

The centroid method is the simplest and most intuitive seriation algorithm. For each row and column, it calculates a weighted average position (centroid) based on where values occur.

For a row with values v_1, v_2, ..., v_n at positions 1, 2, ..., n, the row centroid is calculated as:

$$\text{Centroid} = \Sigma(\text{position} \times \text{value}) / \Sigma(\text{value})$$

Rows are then sorted by their centroid values in ascending order. The same calculation is performed for columns using the column vectors. The process places rows and columns with earlier distributions toward the beginning and those with later distributions toward the end, naturally creating the diagonal pattern characteristic of a well-seriated matrix.

Consider a simple example: if a row has value 1 at positions 2, 3, and 4 out of 10 columns, its centroid would be (2+3+4)/3 = 3.0. A row with values at positions 7, 8, and 9 would have centroid 8.0 and thus appear later in the sequence.

5.2 Correspondence Analysis (CA) Sorting

Correspondence Analysis is a multivariate statistical technique that reveals the underlying structure in contingency tables. When applied to seriation, CA identifies the main axis of variation in the data, which in chronological contexts typically corresponds to time.

The mathematical foundation of CA involves the following steps:

1. Calculate the correspondence matrix P by dividing each cell by the grand total: $P = X / n$
2. Compute row masses r (row sums of P) and column masses c (column sums of P)
3. Calculate the matrix of standardized residuals: $S = Dr^{-1/2} \times (P - rc') \times Dc^{-1/2}$
4. Perform Singular Value Decomposition (SVD) on S: $S = U \times \Sigma \times V'$
5. Extract row and column coordinates from the left and right singular vectors

The first dimension scores typically capture the chronological gradient. Sorting rows and columns by their first dimension scores produces an ordering that maximizes the explained variation in the data.

The eigenvalues from the SVD indicate how much variation each dimension explains. The first dimension usually explains the largest proportion of variation, and subsequent dimensions capture progressively less. In archaeological data with a clear chronological signal, the first two dimensions often explain 50-80% of the total variation.

5.3 Iterative Seriation

Iterative seriation repeatedly applies either the centroid or CA method until the ordering converges to a stable state. The algorithm works as follows:

1. Start with the current row and column ordering
2. Apply the chosen sorting method to reorder rows and columns
3. Calculate the quality score of the new ordering
4. If the ordering changed, return to step 2; otherwise, stop

Iteration continues until either the ordering remains unchanged between iterations (convergence) or a maximum number of iterations (default: 10) is reached. This approach often produces better results than a single sorting pass because each iteration refines the ordering based on the updated positions.

5.4 Fixed Elements

Both seriation methods respect fixed elements. When a row or column is marked as fixed, it maintains its position during automatic sorting. The algorithm sorts only the unfixed elements, then reconstructs the complete ordering by inserting the fixed elements at their original positions.

This feature is essential when external constraints apply, such as stratigraphic relationships that determine the relative position of certain contexts, or when index types must be placed at specific positions in the sequence.

6. Quality Metrics

CombiTab calculates several metrics to evaluate the quality of a seriation. These metrics help identify whether the ordering successfully captures the underlying chronological pattern and highlight areas that may need attention.

6.1 Concentration Index

The Concentration Index measures how close non-zero values cluster around the matrix diagonal. It is calculated by computing the weighted average distance of all values from the diagonal:

For each cell at row i and column j with value v:

- Normalize positions: row_pos = i / (n_rows - 1), col_pos = j / (n_cols - 1)
- Calculate diagonal distance: diag_dist = |row_pos - col_pos|
- Weight by value: weighted_distance = v × diag_dist

The Concentration Index equals 1 minus the average weighted distance. Values close to 1 indicate strong diagonal concentration (good seriation), while lower values suggest values are scattered throughout the matrix.

6.2 Anti-Robinson Index (Continuity)

The Robinson property states that in a properly ordered similarity matrix, values should decrease monotonically as you move away from the diagonal. In seriation terms, this translates to the expectation that each artifact type should form a continuous block without gaps.

The Anti-Robinson Index measures how well the seriation satisfies this property by counting gap cells. For each column, the algorithm identifies the first and last rows containing the type, then counts any empty cells between them. The index equals 1 minus the proportion of gap cells:

```
Anti-Robinson = 1 - (gap_cells / total_span_cells)
```

A perfect score of 1.0 indicates no gaps in any type distribution. Lower scores indicate more gaps, which may suggest problems with the data or the need for further refinement of the ordering.

6.3 Type Continuity

For each artifact type, Type Continuity measures how densely the type occurs within its span:

```
Continuity = actual_occurrences / span_length
```

A type present in 5 contexts spanning positions 3 through 7 (span = 5) would have perfect continuity of 1.0. If the same type were present in only 3 of those 5 positions, its continuity would be 0.6. The Average Type Continuity aggregates this metric across all types.

6.4 Overall Quality Score

The Overall Quality Score combines the individual metrics into a single value using weighted averaging:

```
Quality = 0.4 × Concentration + 0.4 × Anti-Robinson + 0.2 ×
Avg_Continuity
```

This weighting emphasizes diagonal concentration and the absence of gaps (each weighted at 40%) while giving secondary importance to type continuity (20%). Scores above 0.8 generally indicate good seriations, while scores below 0.6 suggest significant room for improvement.

7. Correspondence Analysis

7.1 Understanding the CA Biplot

The CA Biplot provides a two-dimensional visualization of the relationships between contexts (rows) and artifact types (columns). In this display, contexts appear as points and artifact types appear as arrows emanating from the origin.

The biplot uses a joint display where both rows and columns are represented in the same space. Points close together represent similar assemblages or similar distributions. The direction and length of type arrows indicate their contribution to the ordination: types that point in the same direction co-occur frequently, while types pointing in opposite directions rarely co-occur.

7.2 Interpreting Dimensions

Each dimension represents a different axis of variation in the data. In archaeological seriation, the first dimension typically corresponds to time, with earlier contexts and types at one end and later ones at the opposite end. The second dimension often captures secondary variation such as functional differences, spatial patterns, or social distinctions.

The explained variance percentages shown on the axis labels indicate how much of the total variation each dimension captures. A first dimension explaining 45% of variance and a second explaining 20% together account for 65% of the data structure. The remaining variation is distributed across additional dimensions.

7.3 The Scree Plot

The scree plot displays eigenvalues (or explained variance percentages) for successive dimensions. In the plot, dimensions are shown on the x-axis and variance on the y-axis. A characteristic elbow in the curve indicates where meaningful dimensions end and noise begins.

For seriation purposes, attention typically focuses on the first one or two dimensions. If the first dimension dominates strongly (explaining more than 40-50% of variance), the data likely contains a clear chronological signal. If multiple dimensions explain similar amounts of variance, the data may have more complex structure beyond simple chronological ordering.

7.4 Material Group Visualization

In the biplot, type arrows are colored according to their material group assignment. This coloring can reveal patterns such as functional associations (weapons clustering together, jewelry forming a separate group) or chronological trends within material categories. The legend shows which colors correspond to which material groups.

8. Metadata and Cell Annotations

8.1 Column Metadata (Artifact Types)

Each artifact type column can be enriched with metadata:

- Material Group: Categorizes the type by material (Ceramic, Metal, Glass, etc.), affecting visualization colors.
- Index Type (Leittyp): Flag indicating whether this type serves as a chronological marker.
- Fixed Position: Prevents the column from moving during automatic sorting.
- Notes: Free-text field for recording observations about the type.

Material group assignment can be done individually through the Column Labels tab or in bulk using the Type Statistics table in the Matrix View, which allows filtering and sorting types while editing their material assignments.

8.2 Row Metadata (Contexts)

Context rows support the following metadata:

- Context Type: Classifies the context (Grave, Pit, Ditch, Layer, Posthole, Structure).
- Area: Records the excavation area or spatial location.
- Fixed Position: Prevents the row from moving during sorting.
- Notes: Free-text observations about the context.

Context type information appears in abbreviated form next to context names in the matrix visualization, aiding in the identification of context categories at a glance.

8.3 Cell Annotations

Individual cells (where context meets type) can receive detailed annotations:

Certainty: Indicates confidence in the type identification.

- Certain: The identification is confident and unambiguous.
- Uncertain: Some doubt exists about the correct type assignment.
- Questionable: The identification is tentative and may require revision.

Fragmentation: Records the preservation state of the artifact.

- Complete: The artifact is whole or nearly complete.
- Fragmentary: Only a fragment is present.
- Unknown: Preservation status is not recorded.

Additional fields include Count Range (for uncertain frequency values), Inventory Numbers (links to museum catalogs), and free-text Notes.

In the matrix visualization, certainty is rendered using transparency (uncertain items appear faded) and hatching patterns. This provides immediate visual feedback about data quality across the entire matrix.

9. Filtering and Focus Mode

9.1 Material Group Filtering

For large matrices with many artifact types, filtering by material group can simplify the display. The material filter allows selecting one or more groups to display while hiding others. This is particularly useful when focusing on specific artifact categories, such as analyzing only metal objects or comparing ceramic types in isolation.

The filter operates on the visualization only; it does not affect the underlying data or the sorting algorithms. When filters are active, an information banner indicates how many rows and columns are currently visible.

9.2 Focus Mode

Focus Mode provides range-based filtering for both rows and columns. Using slider controls, you can select a subset of the matrix to display. This is invaluable for examining specific portions of large matrices in detail.

Row Focus: Displays only contexts within the selected position range. For example, focusing on rows 20-40 shows only those 21 contexts.

Column Focus: Displays only types within the selected position range, useful for examining specific chronological phases.

9.3 Hide Empty Rows/Columns

The Hide Empty options remove rows or columns that contain no data within the currently visible filtered view. When material filtering is active, hiding empty rows removes contexts with no finds in the selected material groups. This creates cleaner visualizations focused on relevant data.

9.4 Performance Considerations

CombiTab provides an Edit Mode toggle that hides the matrix visualization while editing metadata. For matrices larger than approximately 50×50 cells, enabling Edit Mode can significantly speed up the interface when making multiple changes to material assignments or annotations. The matrix display can then be refreshed with a single click once editing is complete.

The application also implements figure caching, automatically reusing previously rendered visualizations when no relevant parameters have changed. This optimization maintains responsiveness even with large datasets.

10. Export Options

10.1 Image Export

The matrix visualization can be exported in three formats:

- PNG: Raster format suitable for presentations and web use. Export DPI can be adjusted for higher resolution output.
- SVG: Vector format ideal for publication and further editing in graphics software.
- PDF: Vector format suitable for direct inclusion in documents.

The legend showing material group colors and annotation symbols can be exported separately, allowing flexible layout in publications.

10.2 Data Export

Data can be exported in multiple formats:

CSV (Sorted Matrix): Exports the matrix data with rows and columns in the current sorted order. This is useful for further analysis in other software or for archival purposes.

Excel with Metadata: Creates a multi-sheet workbook containing the sorted matrix, column metadata (types and material groups), row metadata (contexts and types), and type statistics.

Excel with Annotations: Includes all metadata plus a sheet listing all cell annotations with their associated contexts, types, and values.

10.3 CA Biplot Export

The Correspondence Analysis biplot can be exported as an SVG file at high resolution. Export options allow customization of label sizes, point sizes, and arrow widths to optimize the output for different publication requirements.

11. Keyboard Shortcuts and Performance

11.1 Keyboard Shortcuts

CombiTab supports the following keyboard shortcuts for efficient operation:

Shortcut	Action
Ctrl + Z	Undo the last ordering change
Ctrl + Y	Redo a previously undone change
Ctrl + Shift + Z	Alternative redo shortcut
Ctrl + S	Trigger project export

11.2 Undo/Redo System

CombiTab maintains an undo history of up to 20 steps. Each sorting operation and manual row/column movement is recorded, allowing you to experiment with different orderings and easily return to previous states. The redo stack allows you to restore changes that were undone.

The current history status is displayed below the undo/redo buttons, showing how many undo and redo steps are available.

11.3 Large Dataset Handling

The application implements several strategies for handling large datasets efficiently:

- Figure Caching: Rendered matrix images are cached and reused when visualization parameters haven't changed.
- Edit Mode: Disabling the matrix display during bulk metadata editing prevents unnecessary redraws.
- Batched Operations: Metadata changes are collected and applied together rather than triggering individual updates.
- Adaptive Settings: The application recommends appropriate visualization settings based on matrix size.

12. Glossary of Terms

Anti-Robinson Property: The characteristic of a properly ordered matrix where similarity values decrease monotonically away from the diagonal. In seriation, this manifests as continuous type distributions without gaps.

Centroid: The weighted average position of occurrences. For a row, the centroid indicates the center of mass of its type distribution across columns.

Combination Table (Kombinationstabelle): A matrix showing the co-occurrence of artifact types (columns) across archaeological contexts (rows). The standard visualization tool for displaying seriation results.

Concentration Index: A metric measuring how closely non-zero values cluster around the matrix diagonal. Higher values indicate better seriation quality.

Context: An archaeological unit such as a grave, pit, layer, or feature that forms a row in the combination table.

Correspondence Analysis (CA): A multivariate statistical technique for analyzing contingency tables. In archaeology, CA extracts the main axes of variation, typically corresponding to chronological and functional dimensions.

Eigenvalue: In CA, eigenvalues indicate the amount of variance explained by each dimension. Larger eigenvalues correspond to more important dimensions.

Index Type (Leittyp): An artifact type with a narrow temporal distribution that serves as a chronological marker. Index types are particularly useful for establishing relative chronology.

Inertia: In Correspondence Analysis, the total amount of variation in the data. Individual dimension inertias indicate how much each dimension contributes to explaining this total.

Material Group: A classification of artifact types by their primary material (ceramic, metal, glass, bone, stone, organic).

Presence/Absence Data: Binary data where cells contain either 0 (type absent) or 1 (type present), regardless of quantity.

Frequency Data: Count data where cells contain the number of items of each type found in each context.

Seriation: The archaeological method of establishing relative chronology by ordering contexts and artifact types based on their co-occurrence patterns.

Span: The range of positions over which an artifact type occurs, measured from its first to its last occurrence in the ordered sequence.

Type Continuity: A metric indicating how completely a type fills its span. Perfect continuity (1.0) means no gaps exist within the span.

End of Documentation