Piccard: A Graph-Base Approach for Census Data Analysis



KAUST: King Abdullah University of Science and Technology

AI Summer School

Project Mentor: Fernando Figueroa

Student Name: Abdulmohseen Ali AlAli

Department of Human Geography, Faculty of Arts & Science University of Toronto Scarborough

21 August 2025

Abstract

Urban researchers rely on census data to identify and analyze demographic trends over time. Understanding these trends is essential for planning new infrastructure, supporting immigrant communities, and providing local services, among several other goals. However, due to the same demographic trends that census data is meant to illuminate, population changes can lead to new census boundaries being drawn. This makes the seemingly simple task of analyzing changes in a region difficult because a region defined by a specific boundary in any given year may not exist in other years.

We present piccard2, a Python package that combines a novel solution to this problem with data clustering algorithms to streamline the data analysis process. piccard2's solution significantly improves on the traditional one, geographical harmonization, which involves defining a common set of regions across all years and fitting data to these regions. This method always introduces some amount of error, and harmonization methods are not readily available for some types of data, which makes it difficult to analyze and visualize that data.

By contrast, piccard2 represents census regions as nodes in a graph network. Each census region in a specific year represents a node, and two nodes are connected if they represent consecutive census years and share at least a specific percentage of geographical overlap. When identifying trends in a specific region over time, every path through the graph containing that region is analyzed. piccard2's solution does not rely on harmonization because it does not assume that every geographical region stays constant over time; instead, it considers every change in the boundaries of every region as a separate path through time.

piccard2 integrates network creation, data clustering, and visualization into one tool, an approach that makes useful analysis possible for data that cannot easily be harmonized. Also, piccard2 is able to efficiently create networks by utilizing parallel processing for large datasets and incorporating flexibility over different coordinate systems. These advantages, combined with piccard2's innovative approach to the problem of inconsistent geographic boundaries over time, will make piccard2 an important tool for urban studies researchers, policymakers, and any professional working with census data.

1. Project objectives and goals:

The primary objective of piccard is to provide a robust framework for analyzing census data across time, overcoming spatial boundary inconsistencies without resorting to simplified harmonization. To achieve this, the project pursues several interconnected goals:

1- **Graph-Based Geospatial Modelling:** Developing a new approach to overcome the current issues with changing boundaries without forcing a fixed one over the years (harmonization), in order to minimize errors and preserve historical data.

2- **Temporal Clustering of Geographic Trajectories:** Implement and integrate multivariate time-series clustering algorithms to identify patterns of demographic change across evolving spatial regions.

3- **Insightful and Customizable Visualizations:** Provide a suite of interactive, customizable visual tools, to help researchers explore temporal and multidimensional patterns clearly and intuitively.

4- **Metadata-Driven Column Linking:** Build a hierarchical structure that traces how variable names changed over the years. Use a combination of text similarity, semantic embedding (e.g., SBERT) to correctly link columns across time.

5- **Inference of Missing Parents Columns:** develop a method for inferring missing parent variables by leveraging metadata links and the additive relationship where child variables determine their parent values.

6- **Software Packaging and Reproducibility:** Deliver piccard as a maintainable, well-documented Python library, complete with installation instructions, examples, and published on PyPI to maximize accessibility for researchers and practitioners.

2. Current progress and milestones achieved:

Public Release and Deployment

Successfully deployed piccard to PyPI, making the tool publicly available to researchers worldwide. The package underwent comprehensive refactoring applying software design principles, resulting in a unified framework integrating network creation, data clustering, and visualization.

Clustering Implementation

Adapted TSCluster algorithms (GreedyTSCluster and OptTSCluster) into the piccard framework, achieving high-performance clustering capable of processing several thousand

geographic regions within seconds. Implemented parallel processing support for scalability with larger datasets.

Visualization Framework

Developed a comprehensive visualization suite featuring three specialized approaches which are temporal evolution visualization, cluster comparison, and multidimensional radar charts. All functions are highly customizable with color-blind accessibility features.

VariableLinker Development

Created the VariableLinker framework to address census data inconsistencies across years. Implemented multiple matching algorithms. The system generates hierarchical tree visualizations showing parent-child relationships between census variables allowing for deeper analysis of the categories and variables of the census data.

VariableLinker Validation

Evaluating the results from VariableLinker using Statistics Canada's official matching tables as ground truth. Successfully validated three matching approaches through confusion matrix analysis, demonstrating positive results for the linking methodology effectiveness.

Parent Prediction

Developed advanced inference functionality that leverages the additive property of census data to predict missing parent variables. The system analyzes tree structures and infers parent values through child aggregation when all children are available in a given year.

Documentation

Produced comprehensive documentation for the project explaining each part of it with details showing the functions implemented and their objective

3. Methodology and approach:

Clustering Analysis

We implemented time-series clustering using the GreedyTSCluster and OptTSCluster algorithms to identify coherent groups of geographic regions exhibiting similar temporal trajectories. This clustering approach was specifically selected for its ability to handle multivariate time-series data while maintaining computational efficiency across large spatial datasets. The algorithms operate by iteratively grouping regions based on their temporal similarity across the selected variables,

thereby revealing underlying patterns of neighborhood evolution and transformation over the study period.

Visualization Framework

To facilitate comprehensive analysis and interpretation of the clustering results, we developed a suite of specialized visualization tools designed to support multi-dimensional exploration of the temporal patterns. The visualization framework encompasses three primary components:

1- **Temporal Evolution Visualization:** This component focuses on demonstrating the changes in the total number of regions within each cluster over time, utilizing two complementary graphical methods. The first employs parallel categories (parallelcats) visualization, which represents the total number of each cluster in boxes with connecting lines between years, illustrating both the magnitude of change and the specific transformation pathways. The second utilizes count plots to display the overall changes of each cluster while abstracting the detailed transformation paths, providing a clearer overview of cluster evolution patterns.

2- **Cluster Comparison Across Time and Within Years:** This approach employs standard analytical plots, including bar plots and line plots, which focus on the underlying data and present them in formats that facilitate easy comparison across different years and clusters. These visualizations enable researchers to examine cluster characteristics at specific time points and identify temporal variations in cluster composition.

3- **Multidimensional Variable Exploration:** To represent multidimensional relationships that capture multiple variables simultaneously within a single visualization, we utilize radar charts due to their capacity to display multiple columns and represent their features comprehensively. This approach provides more in-depth analytical perspectives and may reveal new insights that would not be apparent through conventional univariate or bivariate plotting methods.

VariableLinker Development

Created the VariableLinker framework to address census data inconsistencies by constructing hierarchical tree structures that visualize parent-child relationships among census variables. The methodology involves several key functions:

1- **Tree Construction Process**: Selected the most recent census year as the tree root, then mapped columns from earlier years by comparing their descriptions to those in the base year. This hierarchical approach enables comprehensive visualization of how different column vectors relate and evolve over time.

2- **Text Similarity Matching Approaches**:

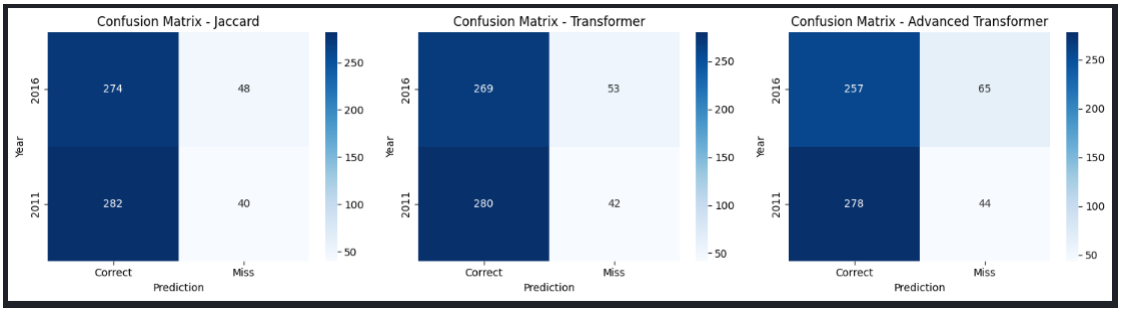
a) **Jaccard Similarity Method**: Implemented token-based comparison using cleaned descriptions with special characters and stop words removed, calculating Jaccard similarity between token sets. Provides good performance for exact and near-exact matches but may miss semantic similarities and is sensitive to phrasing variations.

b) **Multithreaded Processing:** Implemented parallel processing for similarity computation to accelerate performance on large datasets. Improved processing speed while maintaining accuracy of the underlying Jaccard similarity method

c) **Sentence Transformer Method:** Utilized LLM encoders to generate vector representations of descriptions, significantly improving speed through faster semantic similarity checks. Captures semantic meaning and handles paraphrased descriptions effectively while being more robust to word variations. However, has limited ability to process numeric values and ranges within text descriptions

d) **Advanced Transformer with Tie-Breaking**: Enhanced transformer approach using detailed metadata sections to resolve cases where multiple columns share identical descriptions. While semantically powerful and aimed at improving disambiguation, evaluation showed higher error rates than basic transformers, adding complexity without tangible benefits.

3- **Evaluation Framework**: Validating the results of VariableLinker using Statistics Canada's official matching tables as ground truth, constructing confusion matrices to assess performance by quantifying correct matches, mismatches, and overall accuracy of each approach.



Parent Prediction Algorithm

Developed inference functionality based on the additive property principle where parent census variables represent the sum of their children. The algorithm scans tree structures to identify cases where all children exist in a given year, but the parent is not recorded. By aggregating children's values and verifying they satisfy parent requirements, the system generates predicted parent values, improving data completeness for historical analyses.

4. Preliminary results or findings

1. Clustering Analysis

We successfully developed an efficient clustering framework capable of processing several thousand rows within seconds. Parallel processing support further enables scaling to larger datasets without performance loss. This approach revealed meaningful temporal groupings of regions, providing a strong foundation for understanding neighborhood evolution patterns.

2. Visualization Framework

The specialized visualization tools significantly improved interpretability. Temporal evolution plots highlighted how clusters changed over time, while comparison charts (bar plots and line plots) facilitated year-to-year and cross-cluster comparisons. Radar charts allowed multidimensional exploration, uncovering hidden relationships that would be difficult to detect using traditional visualization methods.

3. VariableLinker

The tree construction methodology successfully revealed structural relationships within census data, providing deeper understanding of how data attributes evolved over time. The hierarchical representation identified consistent relationships between variables and added valuable interpretive context to raw metadata, making complex census structures more accessible for analysis.

4. Parent Prediction

The parent prediction functionality added an advanced layer of inference to the analysis. By leveraging the tree structure, this feature identified potential parent attributes that were not explicitly recorded in certain years but could be inferred from the presence and aggregation of their children. This improved data completeness and enhanced the reliability of historical analyses.

5. Timeline and schedule

Task	Start Date	End Date	Status
Domain Knowledge Acquisition	06/06/2025	15/06/2025	Completed
Understanding the structure of the data, and learning how to use the provided tool to download the datasets	11/06/2025	05/07/2025	Completed
Clustering Algorithm Integration	18/06/2025	24/06/2025	Completed
Comprehensive Visualization	25/06/2025	09/7/2025	Completed

Framework Development			
VariableLinker Development for Census Data Harmonization	10/7/2025	25/7/2025	Completed
VariableLinker Documentation and Testing	26/7/2025	1/8/2025	Completed
Real-World Case Study Implementation and Validation	27/7/2025	3/8/2025	Completed
Package Refactoring and Deployment Preparation	2/8/2025	5/8/2025	Completed
PyPI Package Release and Distribution	2/8/2025	8/8/2025	Completed
SUDS Presentation and Poster Development	9/8/2025	15/8/2025	Complete

6. Challenges and obstacles encountered

- I am currently facing challenges in addressing inconsistencies within the dataset, as it contains multiple issues and some of which appear difficult or potentially impossible to resolve completely. However, I am actively exploring several approaches that may help mitigate these problems and improve the clarity and reliability of the data, even if only partially.
- Currently, I am facing some efficient challenges with the tree construction process, as it takes a considerable amount of time to complete. To address this, I am exploring alternative approaches for similarity checking. I am considering the integration of multithreading to enhance performance. Also, I am planning to use large language models, which can improve the accuracy and performance of similarity calculations by understanding the semantic meaning of the text, rather than relying solely on token-level comparisons.

7. Next Steps and Future Work

- Submitting our work in “Environment and Planning B” to be published under the Urban Data/Code series
- Expanding the package by including more advanced variable linking methodology
- Further testing for the package and use it in real research

8. Resources and tools used

- **Piccard Source Code and Package**
  - **GitHub:** https://github.com/ecorbin567/piccard2.git
  - **PyPI:** https://pypi.org/project/piccard/
- Data Processing and Analysis Libraries
  - Pandas: Data manipulation and analysis for census datasets
  - NumPy: Numerical computing and array operations
  - TSCluster: Time-series clustering algorithms (GreedyTSCluster and OptTSCluster)
  - NLTK/Stopwords: Natural language processing for text similarity
- Machine Learning and Similarity Computation
  - SBERT (Sentence-BERT): Sentence transformer models for semantic similarity
    - Model: all-mpnet-base-v2 for high accuracy semantic matching
    - Available models: https://sbert.net/docs/sentence_transformer/pretrained_models.html
- Data Sources and References
  - Statistics Canada Census Data:
    - 2021 Census Metadata: https://www12.statcan.gc.ca/census-recensement/2021/ref/dict/app/index-eng.cfm?ID=a2_2
    - 2016 Census Metadata: https://www12.statcan.gc.ca/census-recensement/2021/ref/dict/app/index-eng.cfm?ID=A2_5

Research and Academic Resources

- Springer Article: Graph-based clustering methodology reference
  - Link: https://link.springer.com/article/10.1007/s10618-024-01054-7
- Original document for piccard 0.0.1
  - Link: https://htmlpreview.github.io/?https://github.com/lodi-m/piccard/blob/main/docs/documentation.html