Smart HVAC: Inferring TU–RTU connections through Clustering and Correlation

DS5110 with Professor Ryan Bockmon
April 24th, 2025

Bini Chandra
Northeastern University
chandra.bi@northeastern.edu

Brian West
Northeastern University
west.bri@northeastern.edu

Cody Snow
Northeastern University
snow.cod@northeastern.edu

Introduction

This project focused on developing a data-driven mapping method to determine which terminal units (TUs) are fed by which rooftop units (RTUs) in a building's HVAC system. Without this detailed mapping, building management systems may run inefficiently and waste significant energy. Our client, Jeff Kimmel, co-founded Elipsa, an AI company that helps building owners save energy. He provided our team with real client sensor data from multiple terminal units and rooftop units to analyze and discover component relationships.

Background

Buildings account for approximately 40% of total energy consumption worldwide, with 40-60% of that energy dedicated to HVAC systems—much of which is wasted due to inefficiencies. Energy Usage Intensity (EUI), measured in kWh per square foot, is a key metric for assessing building efficiency. Elipsa's primary focus is on reducing this waste through artificial intelligence applications.

The complexity of building management systems (BMS) presents significant challenges. These systems collect data from thousands of different data points to manage equipment, but their configuration is highly specific and unique to each building. Despite the existence of standards, they are rarely followed consistently across the industry. This lack of standardization is compounded by commercial pressures: contractors often omit data point tagging during installation to remain competitive in bidding processes, even though proper tagging would enable more advanced BMS functionality.

A more fundamental issue exists beyond tagging—understanding the physical connections between HVAC components. In many buildings, especially older ones, documentation about how terminal units connect to rooftop units is often incomplete or lost entirely. Inferring these connections requires analyzing system behavior over time. For example, if a fan is off in one unit while air is flowing in another terminal unit, this suggests they are not connected. The challenge lies in using sensor data from different equipment to deduce these physical relationships.

This disconnect between system designers and data analysts is particularly problematic. Modern building management systems increasingly use the MQTT protocol due to its low overhead, and Elipsa leverages this by implementing a self-hosted MQTT broker on the building's network to capture and analyze this data. Through this analysis, they aim to infer connections between equipment without requiring complete system documentation—addressing a critical gap in building energy management.

Methods/Analysis

In this project, we followed two different analysis approaches to find which terminal units (TUs) are connected to which rooftop units (RTUs). Below is a breakdown of the dataset, cleaning process, and algorithms applied.

1. **Data Demographics**

Our dataset consists of:

- 47 Terminal Unit (TU) files, each containing approximately 4,800 rows of time-series sensor data.
- 3 Rooftop Unit (RTU) files, each also containing around 4,870 rows.
- Each row represents a single timestamp, and the data spans a 51-day period.

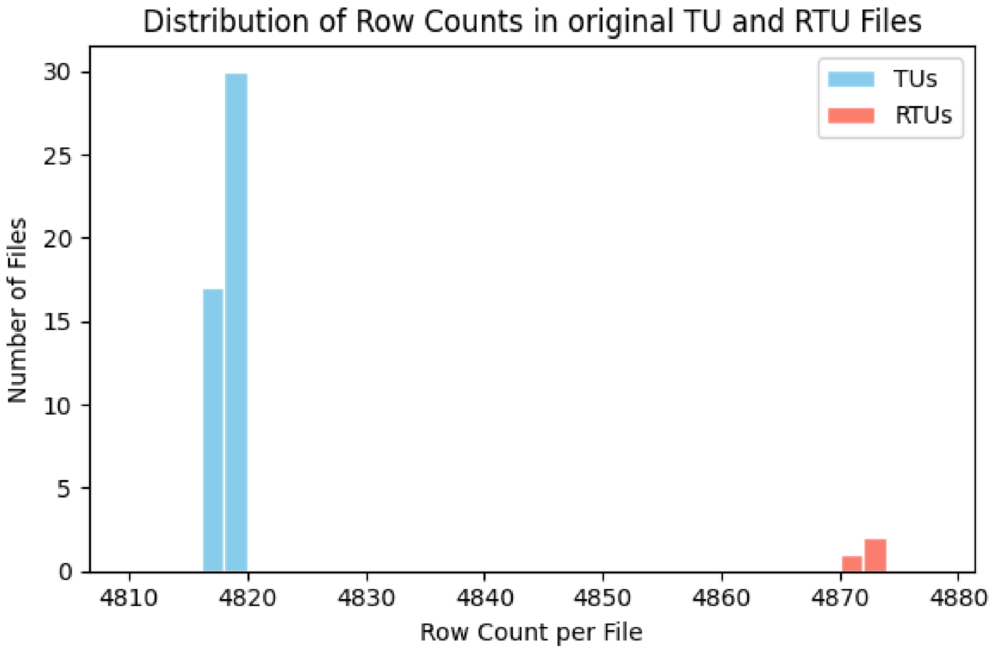


Figure: Histogram showing the distribution of row counts per file in the raw dataset.

The variables in our dataset include:

- Airflow rate
- Discharge air temperature
- Damper position
- Heating and cooling setpoints
- Room temperature
- Various operational commands and states

We analyzed the presence of variables across all TU and RTU files. The chart below shows how frequently each feature appeared across files:

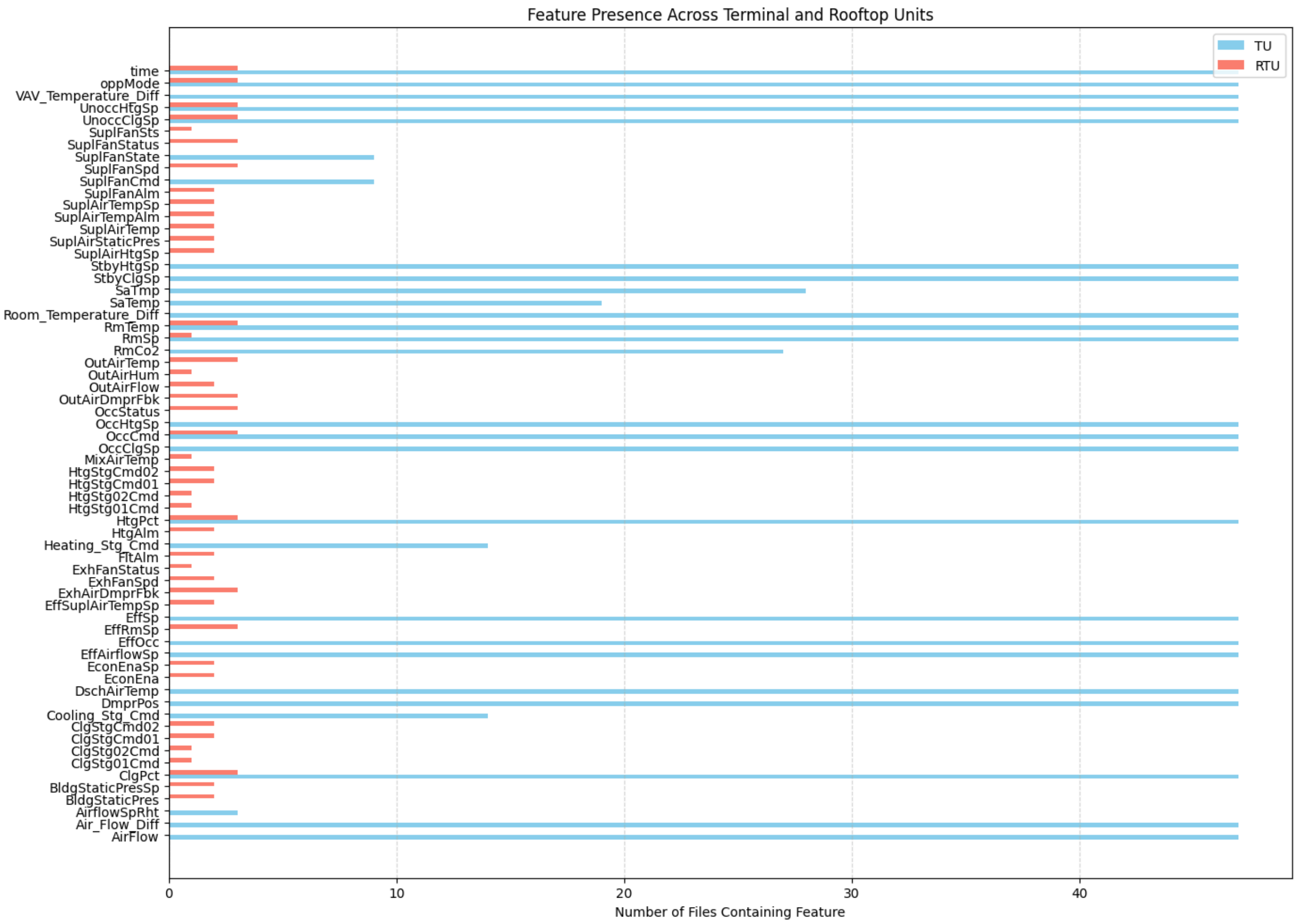


Figure: Chart showing how frequently each feature appeared across files.

2. Data Formatting and Cleaning

To prepare the data for analysis, we followed a multistep cleaning and formatting process that ensured consistency and quality across all terminal unit (TU) and rooftop unit (RTU) files:

- **Forward-Filling Setpoint Values:**
  Many of the “SetPoint” columns (e.g., RmSp, StbyClgSp) had missing values at random timestamps. These were forward-filled so that the last known value carried forward, which reflects how building control systems usually behave in real-world conditions.

- **Dropping Low-Value Columns:**
  Columns with zero standard deviation were dropped, as they contribute no useful information for clustering or correlation. Additionally, we also removed several variables based on the client’s recommendation, such as Air_Flow_Diff and

oppMode.

- **Dropping Incomplete Rows:**
  To improve data quality, we also removed rows that had too many missing values.

- **Timestamp Synchronization (for Approach 2):**
  We synchronized TU and RTU files based on common timestamps across all files to ensure accurate time-aligned correlation analysis.

Variables Dropped:

Unit Type	**Reason**	**Dropped Features**
Terminal Units	Client Advice/ 0 Standard Deviation	Air_Flow_Diff, Room_Temperature_Diff, VAV_Temperature_Diff, SaTemp, SaTmp, oppMode EffOcc, Heating_Stg_Cmd, OccCmd, SuplFanCmd, SuplFanState, Cooling_Stg_Cmd, AirflowSpRht, RmCo2
Rooftop Units (Approach 1)	0 Standard Deviation	'ClgStgCmd02', 'EconEna', 'EconEnaSp', 'EffRmSp', 'ExhAirDmprFbk', 'FltAlm', 'HtgAlm', 'HtgPct', 'HtgStgCmd01', 'HtgStgCmd02', 'OccCmd', 'OccStatus', 'SuplAirHtgSp', 'SuplAirTempAlm', 'SuplAirTempSp', 'SuplFanAlm', 'UnoccHtgSp', 'oppMode'
Rooftop Units (Approach 2)	0 Standard Deviation	'EconEna', 'EconEnaSp', 'ClgStgCmd02', 'HtgAlm', 'HtgStgCmd01', 'FltAlm', 'HtgStgCmd02', 'SuplAirHtgSp', 'SuplAirTempAlm', 'SuplAirTempSp', 'SuplFanAlm', 'UnoccClgSp', 'UnoccHtgSp', 'ExhAirDmprFbk', 'SuplFanStatus',

		'ClgStg01Cmd', 'ClgStgCmd01', 'SuplFanSts', 'ExhFanStatus', 'oppMode'

All other variables were retained as they described the system in ways that could potentially help associate TUs with RTUs.

3. Excluding RTU_3

We observed that RTU_3 lacks several important variables that are available in RTU_1 and RTU_2, such as SuplAirTemp, SuplAirStaticPres, and EffSuplAirTempSp. These missing features limit its comparability with terminal units. Based on this and its low activity in key HVAC signals, we treat RTU_3 as an unconnected or inactive unit and exclude it from the mapping analysis.

4. Techniques Used

We applied a combination of the following techniques across both approaches:

1. **Principal Component Analysis (PCA):**
    - Used to reduce the dimensions of data to focus on the variables that most effectively described the system
    - Helped reduce "noise" in the data and simplified further analysis

2. **K-Means Clustering:**
    - Process involved:
        - Choosing the number of clusters
        - Randomly placing cluster centers
        - Calculating distances between centers and data points
        - Assigning data points to the closest cluster center
        - Moving cluster centers to the average distance between assigned data points
        - Repeating the process multiple times
    - Goal: Cluster terminal units that likely belong to the same rooftop unit

3. **Correlation Analysis:**
    - Used to determine the relationship between pairs of variables
    - Goal: Identify terminal unit variables with strong correlations to rooftop unit variables to help determine which rooftop unit a terminal unit belongs to

Approaches 1, 2 and Results

Two of the three approaches we took utilized Principal Component Analysis, K-Means Clustering, and Correlation analysis to attempt grouping the TUs and RTUs together. These approaches were less successful than the third approach, detailed after this summary.

Principal Component Analysis

Our PCA identified the following variables as contributing most significantly to describing the data:

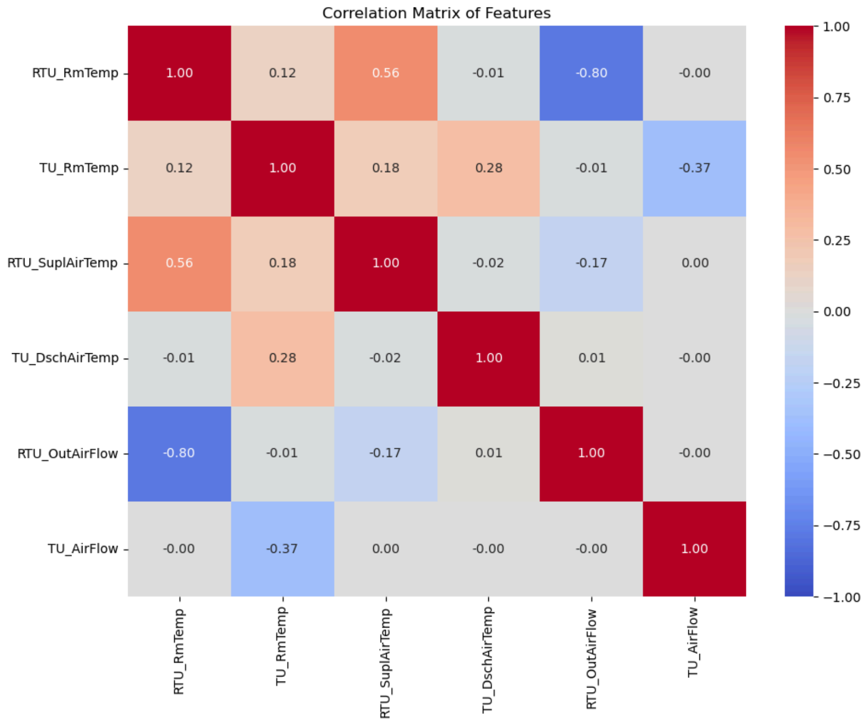
- RTU_RmTemp (Rooftop Unit Room Temperature)
- TU_RmTemp (Terminal Unit Room Temperature)
- RTU_SuplAirTemp (Rooftop Unit Supply Air Temperature)
- TU_DschAirTemp (Terminal Unit Discharge Air Temperature)
- RTU_OutAirFlow (Rooftop Unit Out Air Flow)
- TU_AirFlow (Terminal Unit Air Flow)

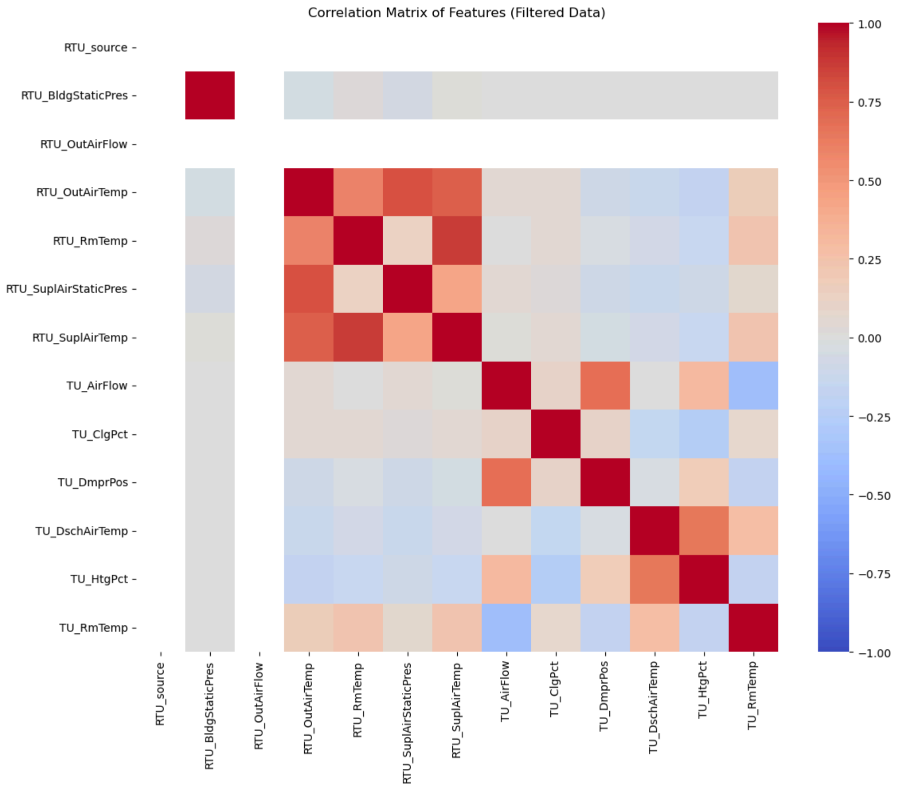
K-Means Clustering

Using two clusters of TUs only, we achieved 65% accuracy in associating terminal units that are likely fed by the same rooftop unit. Merging the TU and RTU data together and performing PCA and K-Means Clustering on the merged dataset resulted in a lower accuracy score, yielding only 57%. While this provides a baseline for mapping, we suspected that additional approaches might improve accuracy.

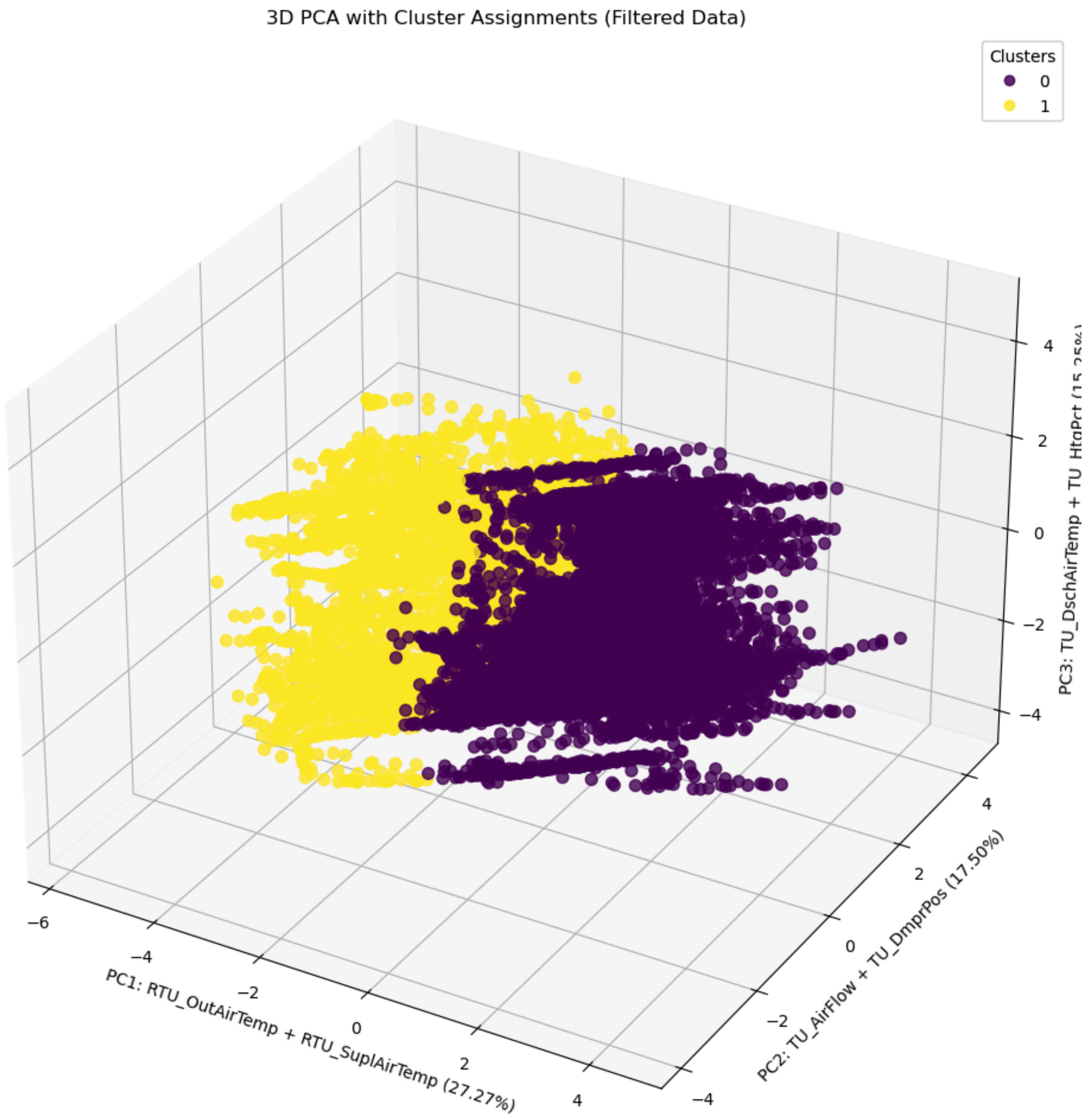
Correlation Analysis

Our correlation analysis between terminal unit variables and rooftop unit variables yielded mostly weak relationships. In cases where stronger correlations were observed, they were approximately equally strong with both rooftop units, making it difficult to establish clear mappings. This suggests that more sophisticated approaches may be needed to determine the relationships definitively.





These correlation matrices show correlation of fields in the merged RTU and TU dataset, which enables us to create a three-dimensional visualization of the clusters.



The merged dataset yielded two closely grouped clusters that did not meaningfully differentiate between the two RTUs (this was the 56% accuracy approach). Axes are labeled based upon which features contribute most to each principal component.

Approach 3 and Results:

Analysis Methods

As a third analytical strategy, we focused on clustering terminal units based on their overall behavioral summaries instead of directly merging them with RTU data. This method attempted to uncover natural TU groupings without relying on any predefined

TU–RTU pairings.

Step 1: Feature Summarization (Median + IQR)
Since each TU file contains time-series sensor readings over 51 days, we summarized every numeric feature using its median and interquartile range (IQR). This gave us a stable snapshot of how each unit behaves across time, helping reduce the effect of outliers and short-term fluctuations.

Step 2: Dimensionality Reduction (PCA)
We applied Principal Component Analysis (PCA) on the summarized features to reduce dimensionality and highlight the most important behavioral patterns across units. This made it easier to visualize and cluster the TUs.

Step 3: Clustering
KMeans clustering (k = 2) was used on the PCA-reduced features to divide the 45 terminal units (we skipped BB.csv and D.csv, due to excessive missing values) into two distinct clusters.These clusters were assumed to correspond to RTU_1 and RTU_2.

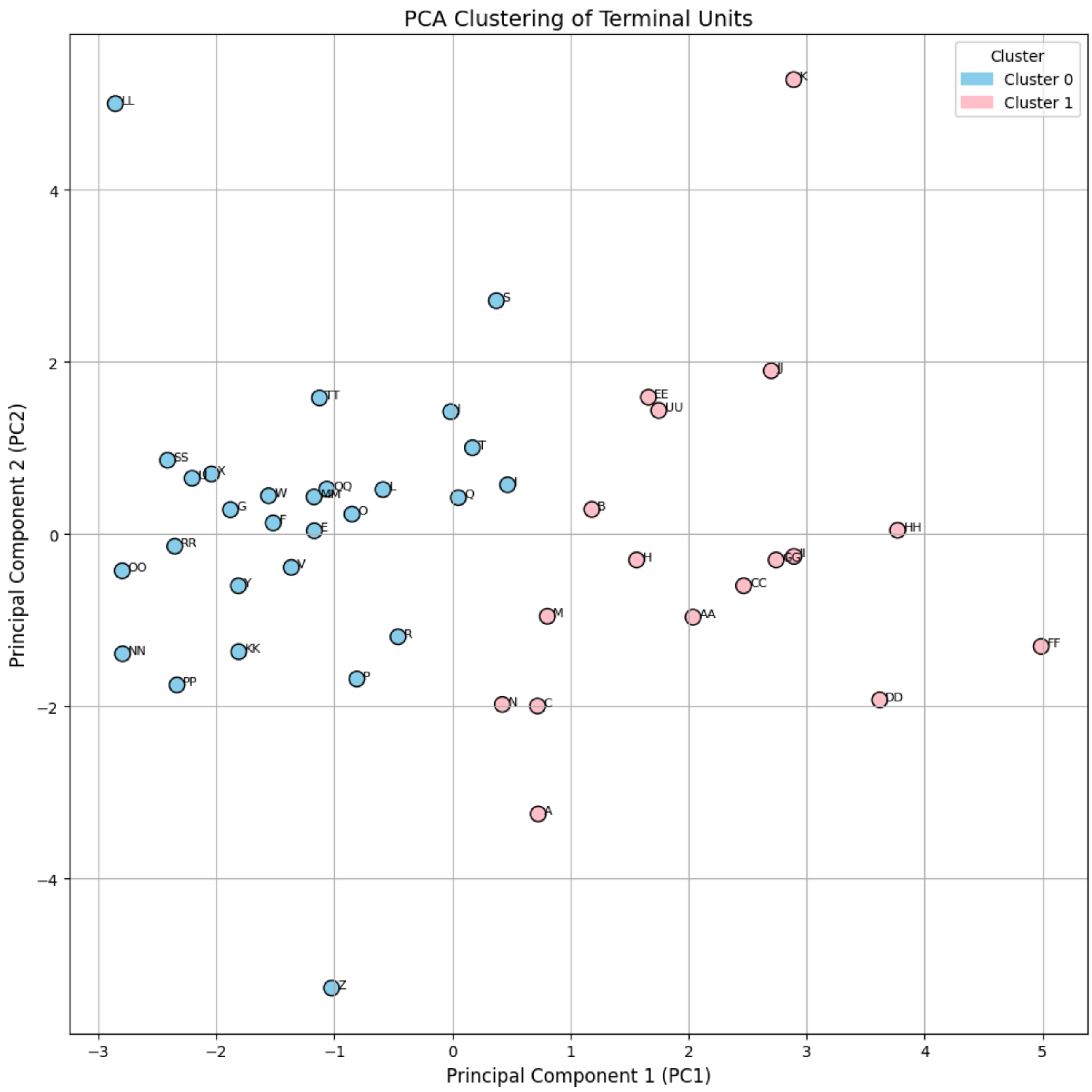
Step 4: RTU Assignment by Correlation
Instead of directly comparing each TU to RTUs one at a time, we averaged the correlation of each TU's raw time series with each RTU's raw time series — using only the columns common to both groups. The average correlation across these shared columns was used to identify which RTU each cluster aligned with most closely.

Step 5: Mapping Evaluation
Finally, we compared our TU-to-RTU assignments with the client's ground truth groupings. Since the cluster numbers could be flipped, we tested both possible mappings and picked the one that gave the best match (77% accuracy).

Results:

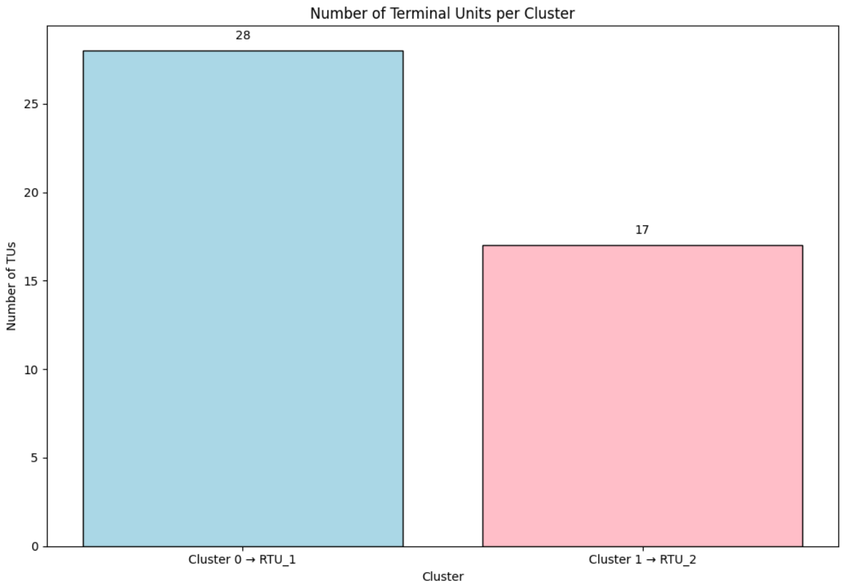
Common Columns Used:
['ClgPct', 'HtgPct', 'RmTemp']

PCA Clustering of Terminal Units
We applied Principal Component Analysis (PCA) to reduce the high-dimensional TU feature set (median + IQR) into two principal components. The scatter plot below shows the natural grouping of TUs into two distinct clusters:



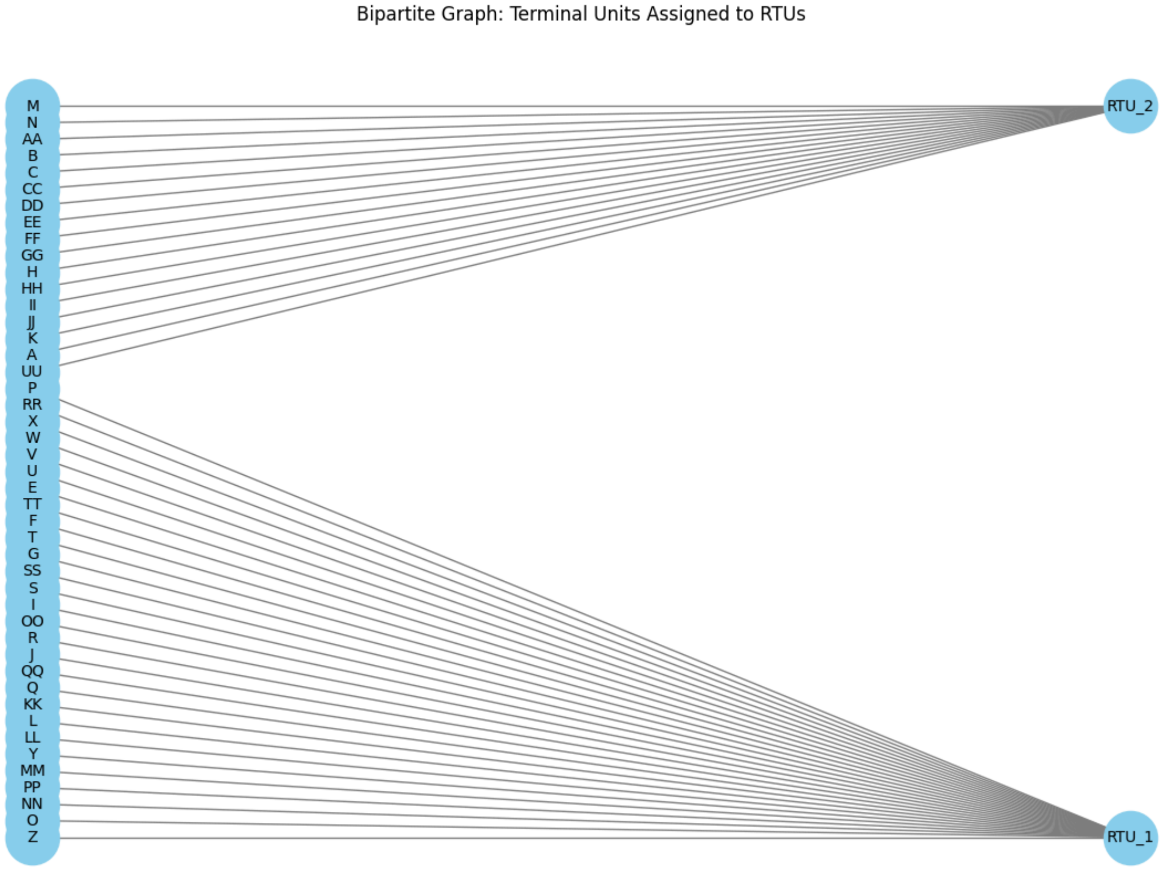
Number of Terminal Units per Cluster

The bar chart below summarizes how many terminal units were assigned to each cluster after KMeans:



Bipartite Graph – Final TU-to-RTU Mapping

We visualized the final assignments using a bipartite graph, where each TU is linked to its assigned RTU:



Final Mapping Accuracy:
77%

Mismatches Identified:
Some terminal units (e.g., CC, DD, E, F, G, I, J, L, N, UU) were difficult to classify confidently and appeared as mismatches.

Discussion

Our analysis of terminal unit (TU) and rooftop unit (RTU) connections through clustering and correlation techniques yielded several valuable insights into HVAC system mapping. The most successful approach (Approach 3) achieved 77% accuracy in correctly associating TUs with their corresponding RTUs, demonstrating that data-driven methods can effectively infer physical connections within building systems even without complete documentation.

Interpretation of Results

The varying success rates across our three approaches highlight important characteristics of HVAC system data. The relatively modest performance of our initial clustering and correlation attempts (65% and 57% accuracy) suggests that simple statistical relationships between TU and RTU variables are insufficient for accurate mapping. This is likely due to the complex, non-linear relationships between components in HVAC systems, where multiple factors influence the behavior of connected units.

Our most successful approach (77% accuracy) didn’t rely on direct feature comparisons at every timestamp. Instead, it summarized each terminal unit’s behavior over the full dataset using statistical metrics like the median and interquartile range (IQR). This made the clustering process more robust by focusing on overall behavioral trends rather than short-term fluctuations. As a result, the TU-to-RTU matching became more accurate, based on general patterns rather than noisy data. The superior performance of this method validates our hypothesis that components connected to the same RTU exhibit similar behavioral patterns over time.

The persistent misclassification of certain terminal units (e.g., CC, DD, E, F, G, I, J, L, N, UU) across our approaches suggests these units may have unique operational characteristics that make them difficult to classify. These units may:

- Be located in areas with atypical thermal loads
- Have non-standard configurations or settings
- Be influenced by adjacent systems or environmental factors
- Require different variables or longer time periods for accurate classification

Practical Applications

These results provide several practical benefits for building management and energy efficiency efforts:

1. **Improved System Documentation**: Our method offers building managers a way to reconstruct missing system documentation through data analysis, addressing a critical gap in legacy building management.
2. **Energy Optimization Potential**: With more accurate TU-RTU mapping, building management systems can implement more efficient control strategies, potentially reducing the 40-60% of building energy consumed by HVAC systems.
3. **Fault Detection Enhancement**: Understanding component connections enables more sophisticated fault detection algorithms that can differentiate between issues in individual TUs versus problems at the RTU level.
4. **Retrofit Planning**: When planning system upgrades, accurate mapping information helps engineers make more informed decisions about which components to replace or modify.
5. **Digital Twin Development**: These mappings form a critical input for creating digital twins of building systems, allowing for simulation and predictive optimization.

Comparison with Existing Methods

Traditional methods for determining HVAC connections rely on manual inspection of physical systems or as-built documentation, both of which can be time-consuming, expensive, or impossible in older buildings. Our data-driven approach offers several advantages:

- **Non-invasive**: Our method requires no physical inspection or system modifications
- **Cost-effective**: Analysis can be performed using existing sensor data without additional hardware
- **Scalable**: The approach can be applied to multiple buildings with minimal customization
- **Continuous improvement**: The method can be refined over time as more data becomes available

When compared to existing computational methods in the literature, our approach emphasizes practical applicability by using readily available BMS data and achieves reasonable accuracy without requiring specialized sensors or system modifications.

Limitations

While our results are promising, several limitations should be acknowledged:

1. **Data Quality Constraints**: Missing values and inconsistent variable availability across files limited the completeness of our analysis. In particular, the exclusion of RTU_3 due to missing variables highlights how data quality issues can impact mapping efforts.
2. **Temporal Resolution**: The 51-day period in our dataset may not capture all seasonal operational patterns that could improve classification accuracy.
3. **Ground Truth Validation**: Our accuracy assessment relies on client-provided ground truth, which itself may contain errors or ambiguities in complex systems.
4. **Domain Knowledge Integration**: Our purely data-driven approach doesn't fully leverage HVAC domain knowledge that could improve classifications, such as known physical constraints on airflow or temperature relationships.
5. **Variable Selection Uncertainty**: Without comprehensive domain expertise, we may have excluded variables that contain important signals for mapping relationships.
6. **Binary Classification Limitation**: Our approach assumes each TU connects to exactly one RTU, which may not capture more complex configurations where TUs receive conditioning from multiple sources.
7. **Time Constraints:** Due to limited time, we couldn’t explore all ideas in depth—like seasonal patterns, supervised learning models, or advanced time-series methods. With more time, we could have tested additional techniques that might improve accuracy and reliability.

Future Work

Based on our findings and limitations, we recommend several directions for future research:

1. **Supervised Learning Approaches**: With the ground truth mappings now available, supervised machine learning models could be trained to recognize patterns indicative of specific TU-RTU connections.
2. **Feature Engineering**: Developing specialized features that capture the physical relationships between HVAC components could improve classification accuracy.
3. **Multi-building Validation**: Testing these methods across different building types, sizes, and HVAC configurations would assess generalizability.
4. **Controlled Experiments**: Deliberately creating operational changes (e.g., adjusting setpoints or temporarily disabling components) could generate stronger correlation signals for mapping.
5. **Integration with Building Information Models (BIM)**: Combining our data-driven approach with spatial information from BIM could provide additional context for resolving ambiguous classifications.
6. **Time-series Analysis Enhancement**: More sophisticated time-series analysis techniques, such as dynamic time warping or recurrent neural networks, might better capture the temporal relationships between components.
7. **Hybrid Physics-Data Models**: Incorporating basic thermodynamic principles into the analysis could improve accuracy by enforcing physical constraints on possible connections. Using concepts like Bernoulli's Principle to analyze pressure readings across the system could reveal interesting relationships.

The 77% accuracy achieved represents a significant step toward automated HVAC system mapping using existing sensors, but further refinement is needed before such methods can be deployed at scale in commercial applications. By addressing the limitations identified and incorporating the suggested enhancements, future iterations of this approach could potentially achieve the accuracy levels required for practical implementation in building management systems.