High-dimensional Reduction on the Context of Urban Planning

Gabriela Correia Zhipei Wang

Abstract

Urban land cover data is essential for sustainable urban planning, aiding in resource management, environmental resilience, and equitable infrastructure development. However, the high dimensionality of such data introduces challenges related to noise, redundancy, and computational complexity, complicating classification efforts. This study examines the performance of Hierarchical Clustering, both with and without dimensionality reduction via Principal Component Analysis (PCA), on a high-resolution urban land cover dataset from Deerfield Beach, Florida. While PCA improved the Silhouette coefficient from 0.16 to 0.19 and enhanced cluster separability for certain classes such as shadow, significant misclassification persisted among spectrally similar classes like tree vs. grass and soil vs. concrete. These results underscore PCA's potential to complement clustering by reducing dimensionality and noise, but highlight its limitations in resolving intrinsic spectral ambiguities. Future research should explore alternative dimensionality reduction techniques, such as UMAP or t-SNE, and multi-modal datasets, including LiDAR and hyperspectral imagery, to improve class separability and support evidence-based urban planning.

1 Introduction

Urban planning plays a crucial role in climate and environmental decision-making, as it directly influences how cities grow, how resources are allocated, and how sustainable these environments can be over time. One critical aspect of this process is understanding the distribution and dynamics of urban land cover , which directly impacts infrastructure planning, resource management, and climate adaptation strategies. High-resolution land cover data offers valuable insights into urban growth patterns, theoptimization of green spaces, and the design of resilient transportation networks. For example, identifying impervious surfaces and vegetation cover can help mitigate urban heat islands, while mapping flood-prone areas can enhance emergency response strategies. Initiatives such as the Copernicus Urban Atlas (Agency (2022)) underscore the importance of detailed land cover data in providing actionable guidance for urban planning decisions.

Urban land cover classification also facilitates the identification of vulnerable areas that may face disproportionate climate risks, such as extreme heat or flooding. Such data enables planners to create equitable and resilient urban environments, ensuring access to critical resources and enhancing overall livability (Johnson (2012)). Moreover, the ability to track changes in vegetation, water bodies, and built environments over time

supports biodiversity conservation efforts and informs strategies to reduce pollution and promote green infrastructure (Durduran (2015)). Given the increasing complexity of urban environments, accurate and scalable methods for classifying land cover are essential for enabling data-driven decisions that support sustainable urban development.

In this study, we utilize the Urban Land Cover (ULC) dataset from the University of California, Irvine (UCI) Machine Learning Repository to explore methods for enhancing urban land cover classification. This dataset consists of high-resolution atmospheric images of Deerfield Beach, Florida, USA, and classifies the region into nine distinct land cover types: trees, grass, soil, concrete, asphalt, buildings, vehicles, pools, and shadows (See Figure 1). It contains 675 observations and 147 features, representing various spatial image characteristics such as:

- Pixel's Spectral Magnitude: mean, standard deviation of spectral bands;
- Textural: Mean, variance and correlation derived from the Grey Level Co-occurance Matrix (GLCM) as proposed by Haralik (Shanmugam and Dinstein (1973)), one of the most common methods to obtain texture measures;
- Formal Characteristics: region's shape description.

Figure 1: Example map of Deerfield Beach, Florida classification into seven land cover types. Extracted from Johnson (2012)

Given the dataset's high dimensionality and rich feature space, traditional clustering methods like Hierarchical Clustering face challenges related to noise, feature redundancy, and computational inefficiency. This complexity is particularly problematic when attempting to separate spectrally similar classes. To improve clustering performance, we propose incorporating Principal Component Analysis (PCA) as a dimensionality reduction step prior to applying Hierarchical Clustering. PCA transforms the original high-dimensional data into a smaller set of uncorrelated components that retain most of the variance in the dataset. By reducing noise and redundancy, PCA has the potential to sharpen cluster boundaries and enhance the interpretability of clustering outcomes.

The central question guiding this analysis is: *Does the incorporation of dimensionality reduction techniques like PCA improve the modeling and classification of urban land cover data compared to traditional Hierarchical Clustering?* By evaluating the performance of

these methods using the ULC dataset, we aim to provide insights into the utility of PCA for clustering high-dimensional land cover data and its implications for urban planning.

2 Methods

2.1 Hierarchical Clustering

Hierarchical clustering is chosen for its flexibility, particularly in unsupervised contexts where the number of clusters is not known beforehand. In the context of urban land cover, it makes sense to allow the data to define clusters based on similarity, because the exact number of distinct land cover categories may vary or overlap in different contexts. This method also helps us focus on the relationships between data points, which can be crucial when distinguishing between, for example, semi-overlapping categories like shadows and buildings.

The primary dissimilarity measure will be the Euclidean distance with normalized features. Normalization ensures that all variables contribute equally to the clustering process, avoiding bias introduced by differing units or scales, enhancing consistency.

Two linkage methods will be considered. First, Complete linkage will be the primary choice due to its tendency to produce compact and well-separated clusters, which aligns with the clear distinctions expected between categories such as trees, asphalt, and water. The Average linkage will serve as a secondary option, as it balances compactness and separation, making it suitable for scenarios with moderate overlap between clusters (e.g., shadows and buildings).

2.2 Dimension reduction

Principal Component Analysis (PCA) is employed to handle high-dimensional data in an unsupervised manner—an approach that is well-established in clustering research (see, for example, Shanmugam and Dinstein (1973); Johnson (2012)).

By identifying those components that capture the majority of data variance, PCA projects the high-dimensional dataset onto a lower-dimensional subspace. This subspace retains most of the meaningful variability, improving cluster separability and reducing computational complexity.

Since our ultimate goal is to recover classes via clustering—comparing the results against actual labels only for validation—we do not introduce class information during training. This means PCA and Hierarchical Clustering work together, since they are both unsupervised methods.

2.3 Validation of results

To assess how well our clustering aligns with true class labels, we will assess cluster assignments in comparison to the known land cover categories. For each cluster obtained

from the model, we will map each cluster to one of the real classes by assigning it to the class that appears most frequently in that cluster. We will then compare the inferred results to the real classes through the use of tables and visualizations.

To validate the quality of the clusters, we will evaluate Bootstrapped stability, the Within-Cluster Sum of Squares (Cohesion) and Between-Cluster Distance (Separation). Bootstrapped stability measures how robust the cluster boundaries are when repeatedly sampling from the dataset, indicating whether minor variations in the data lead to significant changes in cluster membership. Cohesion measures the compactness of clusters by calculating the sum of squared distances between each data point and the centroid of its cluster, whereas Separation measures the distinctiveness of clusters by computing the distance between the centroids of different clusters. They are defined as follows:

$$\text{Cohesion} = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \mu_k||^2 \tag{1}$$

$$\text{Separation} = \sum_{k=1}^{K} ||\mu_k - \mu||^2 \tag{2}$$

Where:

- K: Number of clusters.
- C_k: The set of points in cluster k.
- μ_k: Centroid of cluster k.
- μ: Overall mean of the dataset.

To evaluate both of this metrics, we will calculate the Silhouette Coefficient, defined as

$$s_i = \frac{(b_i - a_i)}{max(a_i, b_i)}, \tag{3}$$

where a_i is the cohesion and b_i is the separation. This measure ranges from -1 to 1, and will be used to quantify how well each data point is matched to its own cluster compared to other clusters, where a higher silhouette coefficient indicates better-defined clusters.

3 Results

3.1 Hierarchical Clustering

In Figure 2, we present the dendrogram produced by applying Hierarchical Clustering with complete linkage to the scaled dataset. Although the dendrogram could, in principle, guide decisions on where to "cut" the tree, here we impose a cut at nine clusters based on

prior knowledge that nine distinct land-cover classes exist in the data. Visually, this means we partition the dendrogram at the level where the vertical branches divide observations into nine separate groups. Notably, some clusters merge at relatively high distances, indicating potential overlap between classes, whereas others coalesce at lower distances, suggesting more cohesive subgroups. By enforcing nine clusters to match the known categories, we can then compare these clusters directly to the true labels and assess how faithfully hierarchical clustering recovers the real-world classes under this constraint

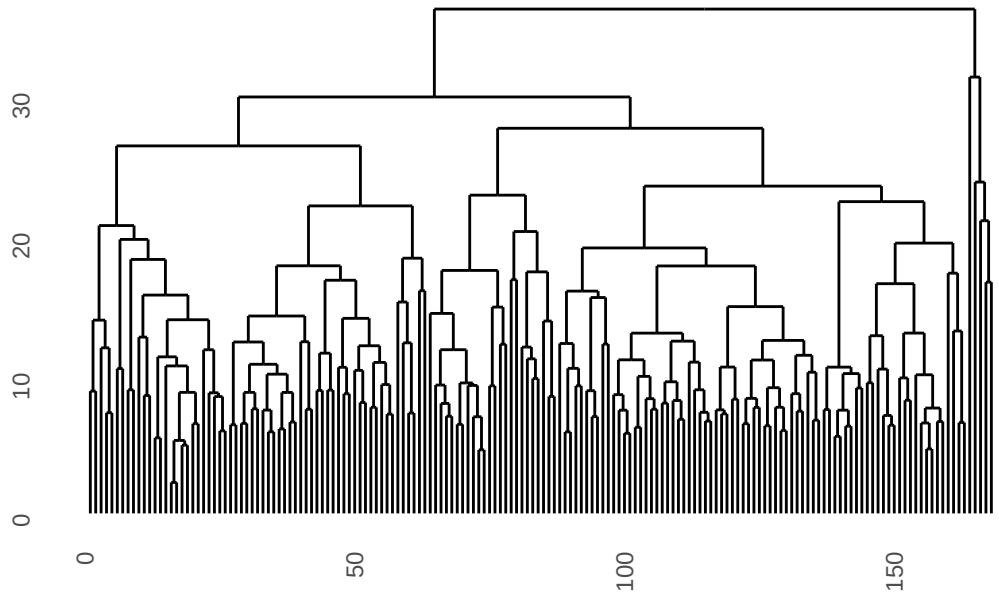

Figure 2: Dendogram for the Hierarchical Cluster

Table 1 shows the bootstrap stability of the clusters. Values close to one indicate stable clusters, so the first 3, fifth and seventh clusters have relatively higher stability. We next perform external validation by mapping each of the nine clusters to the most frequent true class label within that cluster and examining the confusion matrix, displayed under Table 2. While certain clusters largely matched their corresponding classes, notable confusion arose between classes such as “tree” vs. “grass,” and “concrete” vs. “soil.” This overlap highlights the spectral and textural similarities that complicate unsupervised separation. The Silhouette coefficient for this solution (0.16) further indicates relatively weak cluster separation, consistent with the observed misclassifications.

Table 1: Confusion Matrix for Hierarchical Clustering

	asphalt	building	car	concrete	grass	pool	tree	shadow	soil
asphalt	13	0	0	0	1	0	0	0	0
building	1	18	0	1	5	0	0	0	0
car	0	4	9	0	0	2	0	0	0
concrete	0	2	0	21	0	0	0	0	0
grass	2	1	0	5	21	0	0	0	0
pool	0	1	0	0	1	13	0	0	0
tree	0	0	0	1	16	0	0	0	0
shadow	11	0	0	0	5	0	0	0	0
soil	0	0	0	14	0	0	0	0	0

Table 2: Cluster Bootstrap Stability for Hierarchical Clustering

Cluster	Stability
1	0.628
2	0.618
3	0.675
4	0.585
5	0.652
6	0.519
7	0.796
8	0.512
9	0.437

3.2 PCA-Reduced Hierarchical Clustering

To address the high dimensionality of the original dataset, we first performed Principal Component Analysis (PCA) before repeating the hierarchical clustering procedure. Figure 3 shows the scree plot (top), which indicates that nine principal components explain a substantial portion of the total variance, and the dendrogram (bottom) derived from these nine components. By reducing noise and redundancy in the data, this PCA transformation aimed to sharpen cluster boundaries and potentially enhance the interpretability of the dendrogram.

Despite the dimensionality reduction, the bootstrap stability measures (Table 3) reveal somewhat lower values for most clusters than in the full-dimensional scenario, suggesting a heightened sensitivity to sampling variation within the PCA-transformed space. Moreover, the external validation (Table 4) indicates a slight decline in overall accuracy compared to standard hierarchical clustering, although the Silhouette coefficient (0.19) is marginally higher.

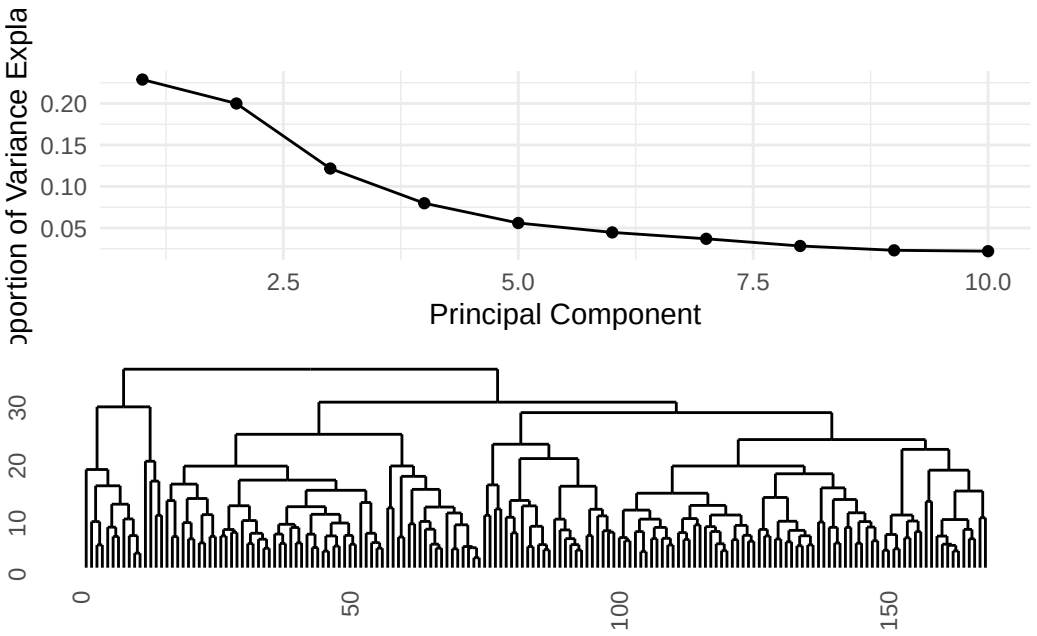

Figure 3: Screeplot and Dendogram for the PCA + Hierarchical Cluster

Table 3: Cluster Bootstrap Stability for PCA + Hierarchical Clustering

Cluster	Stability
1	0.428
2	0.326
3	0.495
4	0.721
5	0.586
6	0.511
7	0.437
8	0.395
9	0.534

Table 4: Confusion Matrix for PCA + Hierarchical Clustering

	asphalt	building	car	concrete	grass	pool	shadow	soil	tree
asphalt	10	0	0	0	4	0	0	0	0
building	0	14	0	8	2	0	1	0	0
car	0	2	4	1	0	8	0	0	0
concrete	0	0	0	23	0	0	0	0	0
grass	0	1	0	9	19	0	0	0	0
pool	0	1	0	0	1	13	0	0	0
shadow	2	0	0	0	7	0	7	0	0
soil	0	0	0	14	0	0	0	0	0
tree	0	0	0	1	16	0	0	0	0

This finding implies that certain clusters—particularly those separating "shadow" from visually similar classes—benefit from PCA, whereas classes with overlapping spectral profiles (for example, "tree" vs. "grass" or "soil" vs. "concrete") remain problematic. In Figure 5 below, we compare clustering assignments with the true class labels. While incorporating PCA helps distinguish certain categories—such as "shadow" from visually similar classes—some clusters still exhibit overlap. In particular, Figure 5 highlights how pairs of classes with highly similar profiles (e.g., "tree" vs. "grass" and "soil" vs. "concrete") remain problematic, resulting in substantial misclassification. This underscores the inherent challenge of separating land-cover types in urban environments, where vegetation (e.g., trees and lawns) can exhibit similar reflectance properties and materials like soil and concrete can appear alike in certain spectral bands. Consequently, although PCA helps reduce dimensionality and enhance some cluster boundaries, further strategies or additional domain-specific features may be needed to fully disentangle these closely related classes.

4 Conclusion

In this report, we have investigated the effectiveness of dimensionality reduction via Principal Component Analysis (PCA) when combined with Hierarchical Clustering (HC) for classifying high-dimensional urban land cover data. The dataset, comprising nine distinct land cover classes, was analyzed to evaluate the performance of clustering algorithms in recovering meaningful groupings of spectrally and texturally similar features. By comparing the outcomes of clustering with and without PCA, we explored whether PCA could improve the separability and robustness of clusters while addressing challenges associated with high dimensionality.

For this dataset, we found that incorporating PCA slightly improved clustering performance in certain aspects but did not entirely resolve the challenges posed by overlapping properties of some classes. PCA improved the Silhouette coefficient from 0.16 to 0.19, reflecting marginally better-defined clusters. Additionally, it enhanced cluster separability for specific categories, such as shadow, which benefited from the reduced dimensionality. However, substantial misclassification persisted between classes with similar profiles,

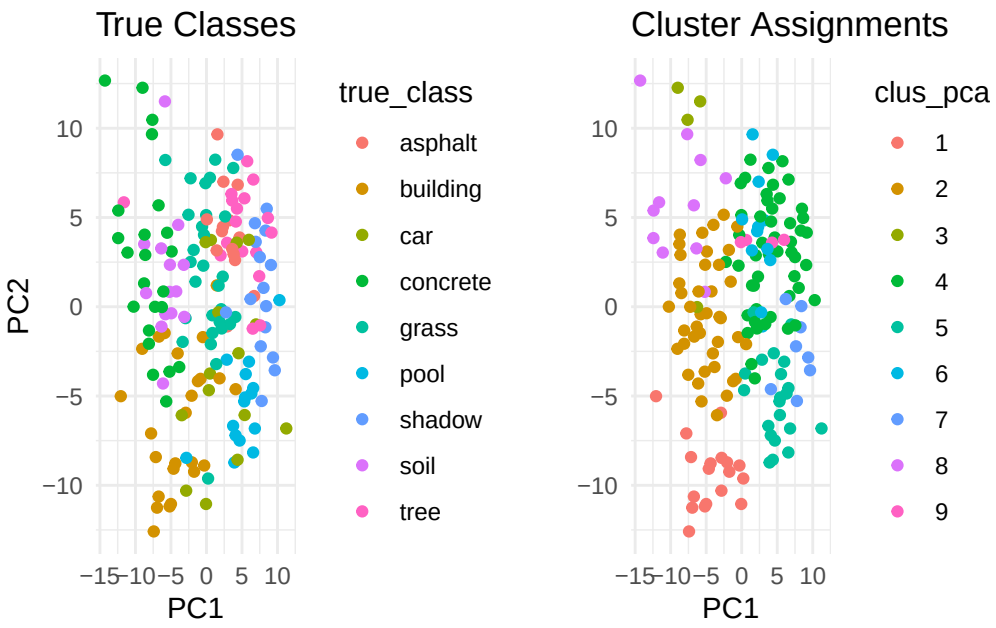

Figure 4: Comparison of PCA-Based Clusters vs. True Classes

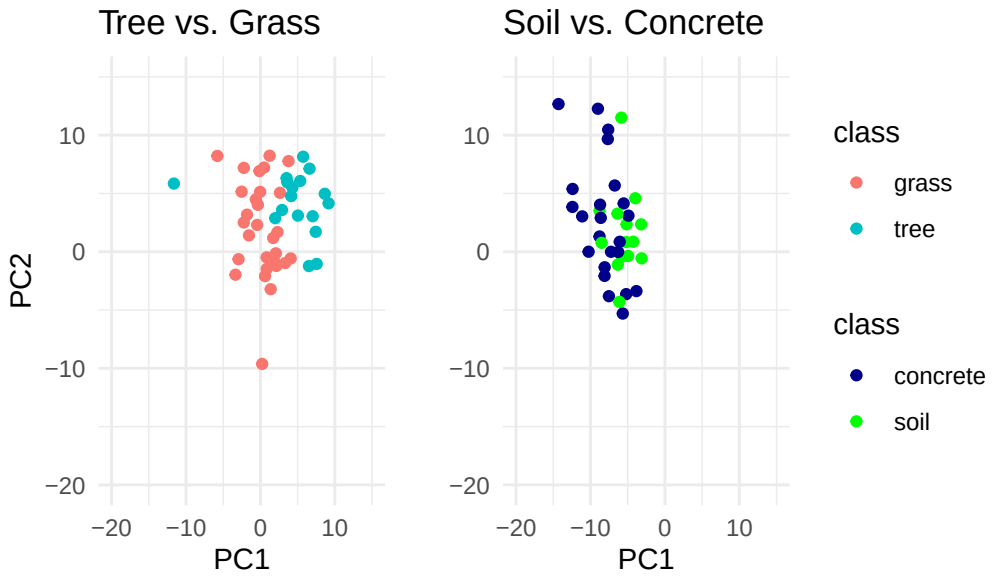

Figure 5: Class overlaps in PCA space. The left panel compares 'tree' vs. 'grass' distributions, and the right panel, 'soil' vs. 'concrete.'

such as tree vs. grass and soil vs. concrete. These results suggest that while PCA can complement HC by reducing noise and redundancy, its utility in this context is limited by the inherent similarities among certain land cover types.

This conclusion aligns with findings in the broader literature on urban land cover classification, where distinguishing between spectrally similar classes has been a persistent challenge. Vegetation types like trees and grass often exhibit overlapping reflectance in multispectral imagery, while urban materials like soil and concrete may be difficult to differentiate based on their spectral characteristics alone. These challenges highlight the need for more robust features, such as temporal information, hyperspectral data, or ancillary spatial information, to improve classification accuracy. For instance, previous studies have shown that incorporating temporal data from satellite imagery or vegetation indices can better separate vegetation types by capturing seasonal changes in reflectance patterns (Lu and Weng (2004)).

Several limitations in this analysis provide avenues for future research. First, the dataset was limited to a single geographic region, Deerfield Beach, Florida, which may limit the generalizability of the findings to other urban contexts with different environmental or climatic conditions. Expanding the analysis to larger, more diverse datasets could provide more robust insights into the effectiveness of PCA and HC. Additionally, while PCA focuses on maximizing variance, alternative dimensionality reduction techniques such as t-SNE or UMAP, which prioritize local neighborhood structure, might yield better separability for overlapping classes (McInnes, Healy, and Melville (2018)). Different sparse clustering methods or ensemble techniques could also be explored to improve performance further.

Finally, future studies could benefit from incorporating multi-modal datasets, such as combining multispectral imagery with LiDAR or social data, to enrich feature space and improve class separability. Previous research demonstrates that integrating these complementary data sources can significantly enhance land-cover classification accuracy (Zhu et al. (2017)). Such approaches could address not only spectral ambiguities but also contextualize land cover data within broader urban dynamics, offering more actionable insights for planners. Overall, this study demonstrates the potential of dimensionality reduction techniques to enhance clustering performance but underscores the continued need for innovation and integration in urban land cover analysis.

5 References

Agency, European Environment. 2022. "Urban Atlas." https://land.copernicus.eu/en/products/urban-atlas.

Durduran, Süleyman Savaş. 2015. "Automatic Classification of High Resolution Land Cover Using a New Data Weighting Procedure: The Combination of k-Means Clustering Algorithm and Central Tendency Measures (KMC–CTM)." *Applied Soft Computing* 35: 136–50. https://doi.org/https://doi.org/10.1016/j.asoc.2015.06.025.

Johnson, Brian. 2012. "Mapping Urban Land Cover Using Multi-Scale and Spatial Autocorrelation Information in High Resolution Imagery." PhD thesis.

Lu, Dengsheng, and Qihao Weng. 2004. "Urban Land-Cover Classification Using Airborne Hyperspectral Imagery: A Comparison of Spectral and Spatial-Based Techniques." *Remote Sensing of Environment* 87 (4): 456–70.

McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv Preprint arXiv:1802.03426*.

Shanmugam, Kalaivani, and Ih Dinstein. 1973. "Textural Features for Image Classification." *IEEE Trans Syst Man Cybern* SMC-3 (January): 610–21.

Zhu, Xiao Xiang, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. 2017. "Fusing Multispectral and LiDAR Data for Urban Land-Cover Classification with Random Forests and Texture Analysis." *International Journal of Applied Earth Observation and Geoinformation* 63: 141–52.