CS7641: Unsupervised Learning Report

[1st] Nha Van Thanh Do
College of Computing
Georgia Institute of Technology
Atlanta, Georgia
ndo46@gatech.edu

***Abstract*—This report continues the Covertype dataset from previous Supervised Learning and Optimization reports but using unsupervised methods to reason about the geometry of the data. K-Means and EM/GMM are used for clustering techniques. PCA, ICA, and Randomized Projections are used to create lower dimensional representations. The OL best model (Nesterov MLP) will be reused while the label column in this dataset is used as external evidence for evaluation purpose only.**

I. Introduction

I reuse the identical working sample from SL and OL reports with a stratified of 20,000 instance subset of Forest Cover Type (581,012 rows) drawn with `StratifiedShuffleSplit` at seed 85. The feature matrix has 54 columns, 10 of those are continuous variables (elevation, aspect, slope, 4 distance measures, 3 hillshade indices) and 44 binary indicators (4 wilderness areas, 40 soil types). Covertype is a seven class label that is heavily imbalanced: Lodgepole Pine (48.8%) and Spruce/Fir (36.5%) dominate, while Cottonwood/Willow is only 0.47%. The label is used only after every unsupervised model is fit, and only for external evaluation and interpretation.

As in SL/OL, only the 10 continuous features are standardized with `StandardScaler` while leaving the 44 binary indicators as 0/1. Standardizing is required because raw ranges differ by orders of magnitude (elevation 1876–3846 m versus 0/1 soil flags), so any Euclidean method would otherwise be dominated by elevation and roadway distance. Leaving binaries unscaled is deliberate. These columns are 0 for most rows, scaling them to unit variance would inflate a near constant column into a high magnitude axis and let a handful of samples dominate distances.

For the exploratory clustering and dimensionality reduction, I fit on the whole 20,000 working sample because labels are consumed only after that. For the downstream Neural Network, it is fit on the training split only and then applied to validation and test. The 80/20 split is used so this part numbers compare back to SL/OL on an identical test set. All randomized components use seeds 85,86,87 (RP additionally 88,89) and I report medians.

II. Hypothesis

For clustering, I believe K-Means may align weakly on this dataset because Euclidean geometry on standardized continuous features captures elevation, distance, and wilderness structure rather than ecological species labels. Looking at Fig 1, there are overlapping features on 4 of the continuous attributes that could impact to this algorithm heavily and may result in grouping different types together. The dominant geometric split should be far coarser than seven classes. I also expect EM/GMM to align better than K-Means, because soft elliptical densities can model the imbalanced, overlapping class distributions that K-Means cannot.

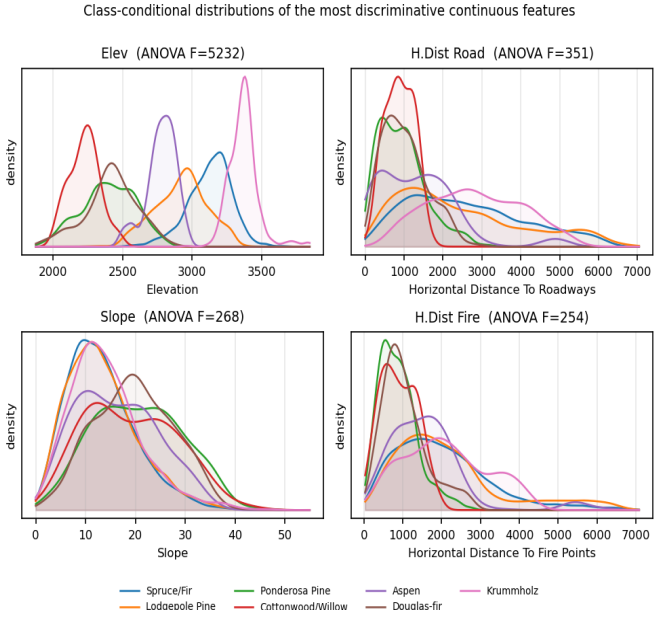


Fig. 1: Class Conditional Distribution

I also expect PCA to preserve downstream neural network performance better than ICA or RP. Fig 2 shows the correlation between continuous features and that is what Jolliffe and Cadima mentioned that PCA is effective when variables are correlated and when most variance can be represented by a few principal components [1]. Whereas RP preserves pairwise distances only approximately [2] and ICA optimizes statistical independence rather than class separation [3].

III. Clustering on Original Features

A. Setup

K-Means uses Euclidean distance on the standardized feature matrix with k-means++ initialization, `n_init`=10, `max_iter`=300, and random seed 85. Although Covertype contains seven classes, $k = 7$ is treated only as a reference point rather than the expected optimum. Accordingly,

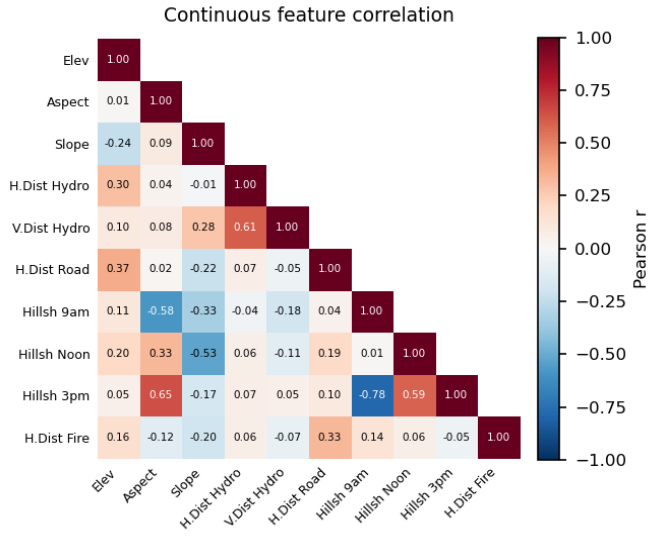


Fig. 2: Pearson Correlation Matrix

I evaluate $k \in \{2, \ldots, 20\}$ using silhouette, Davies-Bouldin (DB), Calinski-Harabasz (CH), and inertia to determine the supported cluster structure. For EM/GMM, I use diagonal covariance with `n_init`=5 and `reg_covar`=10^{-4}, selecting the number of components by BIC. Because covariance type is itself an important modeling choice, I first compare spherical, diagonal, tied, and full covariance at $k = 7$. The resulting trade-offs between density fit, label alignment, and computational cost lead to the adoption of diagonal model for the remaining experiments.

B. K-Means

The K-Means internal metrics in Fig 3 suggest that the original Covertype feature space does not naturally separate into 7 compact clusters. Silhouette score evaluates how well each point fits within its assigned cluster relative to the nearest neighboring cluster, so higher values indicate clusters that are both compact and well separated. Here, the silhouette score is highest at $k = 2$ with score of 0.191 and decreases steadily as more clusters are added. At $k = 7$, the score drops to just 0.133 indicating that it favors a small number of clusters rather than 7 distinct groups. Davies-Bouldin index (DB) measures the ratio of each cluster to its most similar cluster, so lower DB means the clusters are tight internally and far apart from each other. DB decreases from $k = 2$ to around $k = 7$ meaning that splitting the 2 large groups into smaller clusters does improve cluster compactness. However, it is weak and small. Likewise, inertia measures the total of squared distances of samples to their closest cluster center, decreases smoothly as k increases because each additional centroid reduces within-cluster variation. Since inertia almost always decreases with larger k, it is interpreted using the elbow method rather than its absolute value. The inertia curve shows no elbow near $k = 7$ suggesting there is no natural 7 cluster solution. Calinski-Harabasz (CH) also favors relatively small values of k, reinforcing the conclusion that the dominant structure is much coarser than the seven cover types.

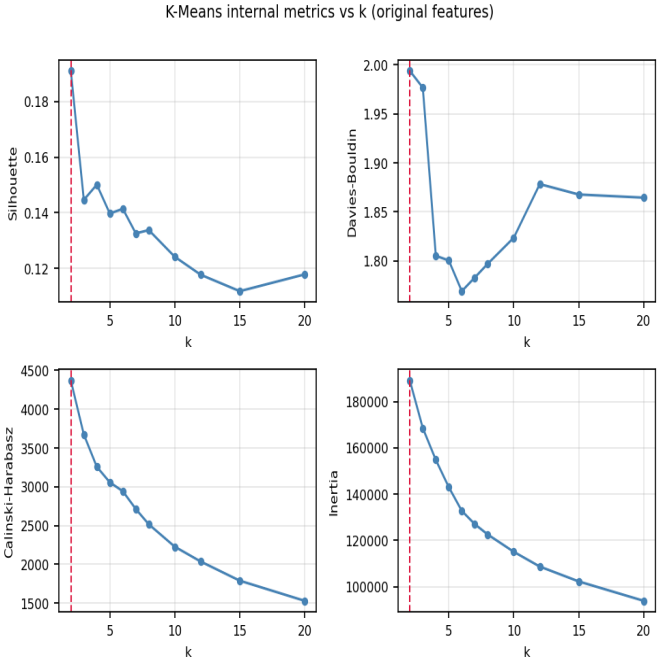


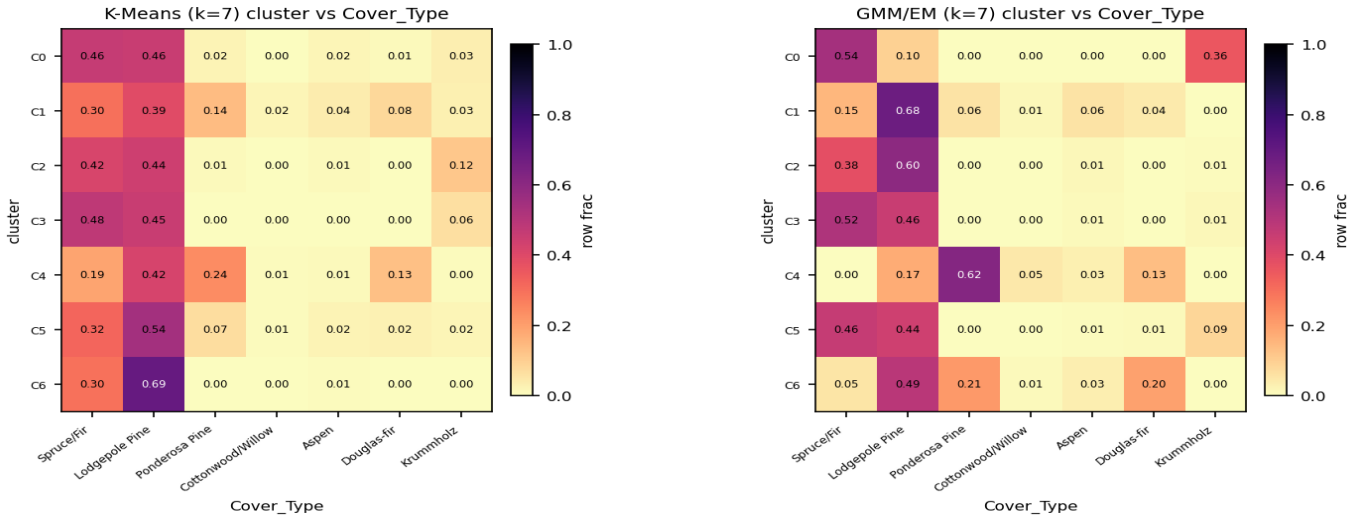
Fig. 3: K-Means Internal Metrics

Taken together, these internal metrics tell a consistent story. The data contain meaningful geometric structure, but that structure does not naturally partition into seven compact Euclidean clusters. Instead, K-Means primarily discovers two broad groups and then progressively subdivides them as k increases, without revealing a clear intrinsic seven cluster organization.

TABLE I: Original-space clustering

Algo	k	Sil.	DB	ARI	NMI	V-meas
K-Means	2 (sil)	**0.191**	1.994	0.004	0.002	0.002
K-Means	7 (ref)	0.133	1.783	0.021	0.064	0.064
GMM/EM	7 (ref)	0.020	3.548	**0.115**	**0.209**	**0.209**
GMM/EM	20 (BIC)	−0.029	3.208	0.074	0.186	0.186

Table I shows that although k = 2 achieves the best silhouette score, its label alignment is essentially zero (ARI = 0.004, NMI = 0.002). Even at the reference point k = 7, alignment remains weak (ARI = 0.021, NMI = 0.064). It's because every cluster contains substantial mixtures of the two dominant classes, Spruce/Fir and Lodgepole Pine, rather than isolating individual cover types. Heatmap in Fig 4a shows that every cluster is dominated by the 2 majority classes (class 1 and 2). This indicates that Euclidean distance is capturing continuous geographic variation such as elevation, slope, and distance-based features rather than the ecological boundaries that define the target labels.



(a) K-Means Heatmap

(b) GMM Heatmap

Fig. 4: Contingency Heatmaps for K-Means and GMM

C. EM/GMM

For GMM, I first compared the 4 covariance models at $k = 7$ as shown in Table II. Although full covariance achieves the lowest BIC, it's about 8 times slower than diagonal covariance and produces lower ARI and NMI. Therefore, I use diagonal covariance which provides the best balance between model quality, computational cost and interpretability.

TABLE II: GMM covariance-type check at $k = 7$

Covariance	BIC	ARI	NMI	Fit (s)
spherical	+789,883	0.021	0.065	1.2
diag	−4,634,306	**0.115**	**0.209**	**1.4**
tied	−2,065,918	0.051	0.115	5.4
full	−5,084,384	0.091	0.194	10.4

With the setup using diagonal covariance, Fig 5 shows that BIC and AIC decrease steadily across the entire search range and reach their minimum at the edge of the grid ($k = 20$). Since BIC measures goodness of density fit while penalizing model complexity, this trend indicates that additional Gaussian components continue improving the density model. Although GMM has much weaker internal clustering quality than K-Means (silhouette = 0.020 at k=7), it aligns substantially better with the Cover Type labels. In table I, at the same k=7, GMM achieves ARI = 0.115 and NMI = 0.209, compared with 0.021 and 0.064 for K-Means. Looking at Fig 4b for GMM heatmap, we can easily observe that it begins to recover meaningful ecological structure. Cluster C4 is predominantly Ponderosa Pine (62%), while C0 captures a high elevation Spruce/Fir–Krummholz group (54% and 36%, respectively). Other components remain mixtures, particularly those dominated by Spruce/Fir and Lodgepole Pine, showing that substantial overlap still exists between cover types.

For this experiment, K-Means wins every internal metric (positive silhouette, low DB) but it barely tracks the labels.

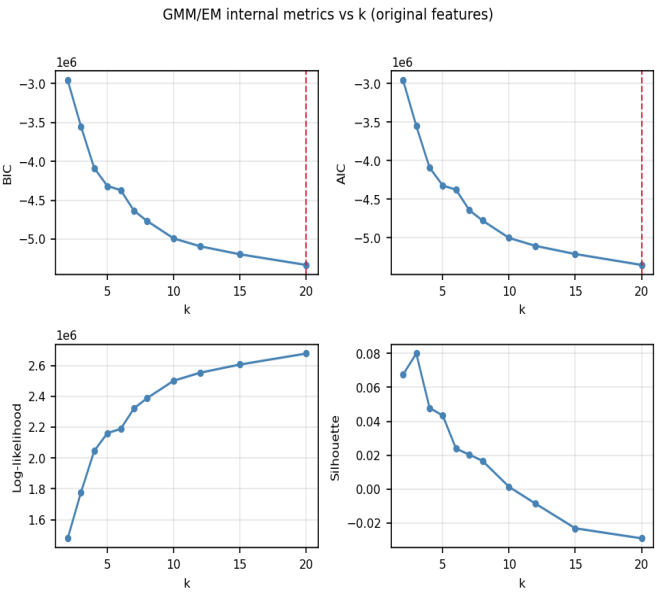


Fig. 5: GMM Internal Metrics

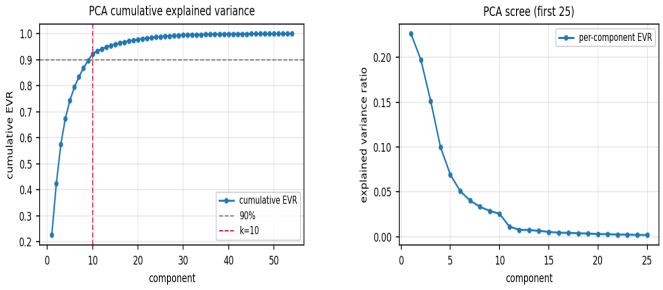
GMM has a poor silhouette (0.020) yet captures three times the label information. However, it does not mean that GMM finds the species but it's a more flexible probabilistic model that preserves substantially more ecological structure than hard Euclidean clustering.

IV. Dimensionality Reduction

A. PCA

PCA is evaluated using the explained variance ratio (EVR), which measures how much of the total variance is retained by each principal component. Fig 6b shows the first three components are 22.6%, 19.7%, and 15.1% of the variance

which contribute more than half of total variance. On the other hand, 10 components in fig 6a retain 92.2% and 14 retain 95%. The covariance spectrum also has an effective rank of only 7.4, indicating that much of the information is concentrated in a small number of correlated directions despite the dataset containing 54 features.



(a) PCA Cumulative Variance (b) PCA per-component

Fig. 6: PCA Diagnostics

These results indicate substantial redundancy in the original feature space, which is expected because many continuous terrain variables are correlated and the wilderness-area variables are one-hot encoded. Therefore, I select 10 principal components, the smallest number that preserves 90% of the variance. Beyond this point, each additional component contributes little additional information.

The first two principal components also provide a useful visualization of the clustering results from Part 1 as shown in Fig 7. Instead of seven well-separated groups, the data forms one broad continuous manifold. Coloring the points by GMM cluster (left) or Cover Type (right) shows considerable overlap, confirming that the dominant directions of variation correspond to terrain geometry rather than species boundaries.

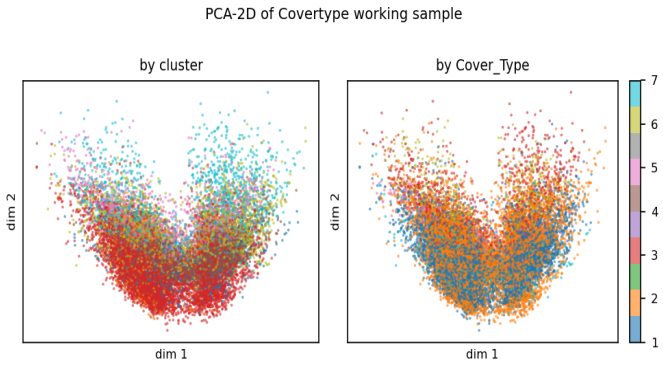


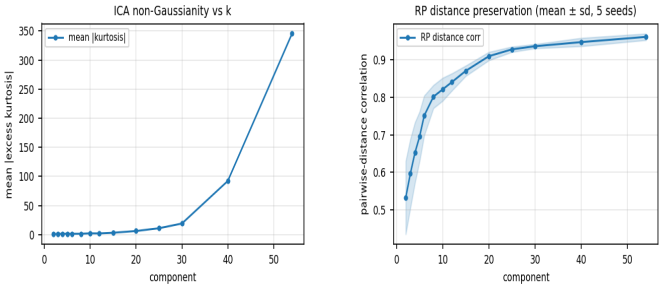
Fig. 7: Working Sample Projected onto PCA's Components

B. ICA

ICA is evaluated using absolute excess kurtosis, the non-Gaussianity measure that FastICA maximizes when estimating statistically independent components. Fig 8a shows that the mean absolute kurtosis increases steadily with the number of components instead of reaching a stable plateau, rising from approximately 2.2 at 10 components to over 340 at 54 components. That's because Kurtosis-based non Gaussianity is known to be highly sensitive to outliers [3], and a column that is 0 for 99% of rows behaves exactly like an outlier, dominated signal, so the most non-Gaussian directions in this space are simply the individual rare soil-type binaries. Unrestricted selection would pick $k = 54$ but to make a fair downstream comparison, I continue to use the same 10 component budget that is selected by PCA. Seed sensitivity is also low. As shown in table III, for $k = 10$ across 3 seeds of 85, 86, 87, the sorted kurtosis profile barely moves, each position varies by ≈ 0.001, so the degeneracy is reproducible, not seed noise. However, at $k = 54$, it becomes sensitive as different seeds latch onto different rare flags. The effective rank is 7.4, so beyond this threshold, ICA is forced to manufacture independent axes in a subspace with almost no independent structure left. It's the result of overlearning at high k [4].

TABLE III: Seed Sensitivity

Method	k	Diagnostic	Seed metric
ICA	10	mean|kurt| 2.24, max 6.5	CV 0.001
ICA	15	mean|kurt| 3.28, max 7.3	CV 0.012
ICA	54	mean|kurt| 345, max 2850	CV 0.151
RP	2	dist. corr 0.53	sd 0.097
RP	10	dist. corr 0.82	sd 0.031
RP	20	dist. corr 0.91	sd 0.011
RP	54	dist. corr 0.96	sd 0.009



(a) ICA Kurtosis Diverges (b) RP Distance Correlation

Fig. 8: ICA and RP DR Diagnostics

C. RP

Random Projection is evaluated using pairwise distance correlation, which measures how well distances between observations are preserved after projection. Because RP is stochastic, I also measure variability across 5 random seeds 85,86,87,88,89 with different k as shown in table III.

Fig 8b shows that mean Pearson distance correlation increases with dimensionality, reaching approximately 0.90 at 20 components, while the variation across seeds decreases substantially. Therefore, I select 20 components, the smallest dimension that preserves pairwise distances with correlation above 0.90 and low seed variability. Compared with PCA,

RP requires roughly twice as many dimensions just to exceed a 0.90 distance correlation. This is expected because RP preserves distances approximately through random projections rather than learning directions from the data.

D. Cross-method Comparison

TABLE IV: Cross-method Comparison

Method	k	Dist. corr	Diagnostic	Runtime
PCA	10	0.997	92.2% cum. var	0.4 s
ICA	10	0.853	mean |kurt|=2.24	21.6 s
RP	20	0.909	seed sd=0.011	2.7 s

Looking at the summary table IV. PCA preserves the dominant variance structure, retaining 90% of the variance with only 10 components while almost perfectly preserving pairwise distances (0.997). ICA seeks statistically independent directions rather than preserving variance or geometry. It also reveals that the dataset contains strongly non-Gaussian sparse features. RP approximately preserves pairwise distances but requires 20 components to achieve a distance correlation of 0.909 and introduces small, measurable variability across random seeds. ICA is also by far the slowest (21.6 s vs. 0.4 s for PCA) because of its iterative fixed-point optimization.

V. Clustering After Dimensionality Reduction

I reran K-Means and GMM in each reduced feature space, then re-select the clustering settings instead of reusing those from the original data. For K-Means, the number of clusters was selected by the silhouette score in each transformed space. For GMM, BIC was evaluated again after each dimensionality reduction method. However, as in the original feature space, BIC decreased monotonically to the largest value in the search grid ($k = 20$), providing no meaningful interior optimum. Therefore, while the selected GMM model is reported for completeness, I compare label alignment at the common $k = 7$ reference so that all representations can be evaluated fairly. Table V summarizes the quantitative results, and Fig 9 illustrates how the cluster assignments change after dimensionality reduction.

TABLE V: Clustering before and after DR

Space	Algo	sel. k	Sil.	DB	ARI_7	NMI_7
Original	K-Means	2	0.191	1.994	0.021	0.064
Original	GMM/EM	20	−0.029	3.208	**0.115**	**0.209**
PCA	K-Means	2	0.210	1.881	0.019	0.060
PCA	GMM/EM	20	0.080	1.962	0.011	0.036
ICA	K-Means	7	0.145	1.865	**0.036**	**0.128**
ICA	GMM/EM	20	0.045	2.424	0.004	0.047
RP	K-Means	2	**0.259**	**1.562**	0.019	0.051
RP	GMM/EM	20	0.100	1.889	0.022	0.060

Nearly every DR and algorithm pair has a higher silhouette and lower DB than its original space (the exception is ICA + K-Means, whose silhouette falls to 0.145). RP and K-Means gives the best silhouette of the whole study (0.259, DB 1.562), and GMM's silhouette jumps from -0.029 in full 54 features

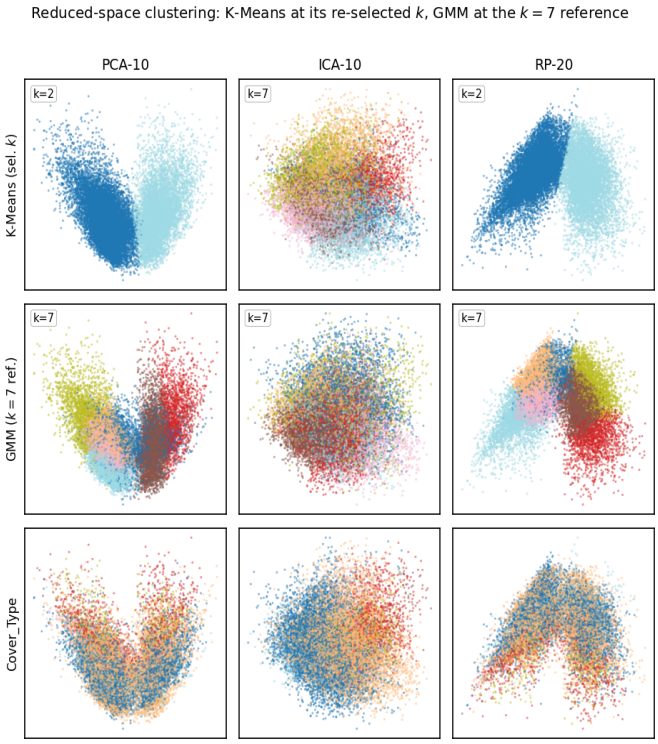


Fig. 9: Clustering after Dimensionality Reduction

to 0.080–0.100 after PCA/RP. K-Means still prefers k=2 in most reduced spaces, confirming that the dominant geometric split survives compression.

However, better internal clustering does not necessarily mean better recovery of the true Cover Type classes. The original space GMM remains the most label aligned model, achieving ARI 0.115 and NMI 0.209. After applying PCA, ICA, or RP, GMM's label alignment drops substantially, suggesting that the sparse soil and wilderness variables contain class discriminative information that is compressed or mixed during dimensionality reduction. The only notable exception is ICA combined with K-Means. ICA is the only representation that changes the preferred cluster count to $k = 7$, and this nearly doubles K-Means' label alignment (NMI 0.064 to 0.128 and ARI 0.021 to 0.036).

Overall, dimensionality reduction improves the geometric quality of the clusters but does not reveal substantially better Cover Type structure. The main effect of dimensionality reduction is therefore to simplify the geometry of the feature space rather than uncover previously hidden ecological structure.

VI. Neural Network with Reduced Data

For this part, I keep the neural network recipe fixed across all experiments. The best $64 \rightarrow 64$ ReLU architecture from the supervised learning study and the Nesterov SGD optimizer identified in the optimization study. Only the input dimension changes after dimensionality reduction, and every transformation is fit on the training split before being applied to the validation and test sets.

TABLE VI: Neural network after DR

Rep.	dim	Acc	Macro-F1	Bal. Acc	F1 sd	Fit (s)	Pred (ms)
Original	54	**0.799**	**0.668**	**0.655**	0.016	0.67	0.5
PCA	10	0.776	0.642	0.606	**0.006**	0.82	0.2
ICA	10	0.773	0.639	0.627	0.010	0.77	0.2
RP	20	0.785	0.607	0.569	0.018	0.78	0.3

Table VI is showing the result of neural network after DR with median over 3 seeds of 85, 86, 87. The original 54 feature representation remains the strongest overall, achieving the highest Accuracy (0.799), Macro-F1 (0.668), and balanced accuracy (0.655). None of the reduced representations improves predictive performance, indicating that the original feature space already provides the network with the most discriminative information. Among the reduced representations, PCA performs best (Macro-F1 0.642), followed closely by ICA (0.639), while RP degrades the most (0.607).

Although dimensionality reduction does not improve accuracy, it changes the training behavior. PCA produces the smallest variation across random seeds (Macro-F1 standard deviation 0.006 versus 0.016 for the original features), suggesting that removing low-variance directions creates a more stable optimization landscape. ICA shows similar overall performance but slightly higher balanced accuracy than PCA (0.627 vs. 0.606), indicating a more even distribution of recall across classes. In contrast, RP exhibits both the lowest Macro-F1 and the highest seed-to-seed variation, reflecting the instability introduced by random projections. Overall, dimensionality reduction slightly improves training stability but not generalization performance.

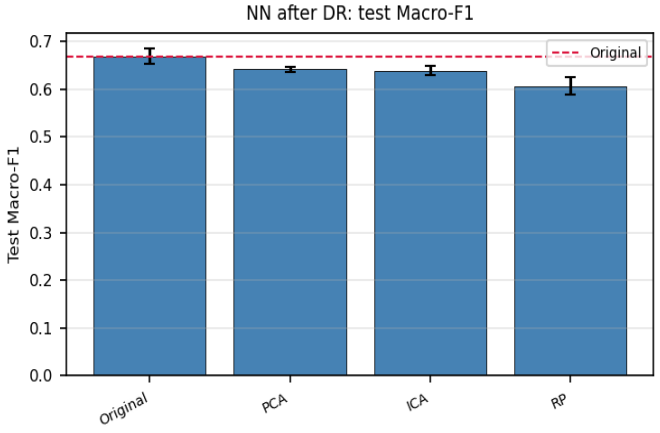


Fig. 10: Macro-F1 Across Algorithms

Reducing the input dimension provides almost no runtime benefit. Training remains between approximately 0.7 and 0.8 seconds and prediction requires only 0.2–0.5 ms across all representations. This is expected because only the first linear layer depends on the input dimension, while the computational cost is dominated by the fixed hidden layers and the identical 3,000-gradient-evaluation training budget. Since the neural network architecture that I'm using is relatively small, dimensionality reduction offers representation compression rather than meaningful computational savings.

On class level behavior, the summary is shown in table VII and heatmap in Fig 11. The effect of dimensionality reduction is concentrated in the minority classes rather than the majority classes. Spruce/Fir and Lodgepole Pine maintain nearly identical F1 scores under every representation, showing that their decision boundaries are robust. In contrast, the rare classes respond differently to each transformation. PCA improves Douglas-fir from 0.43 to 0.50, Krummholz from 0.78 to 0.80, and slightly improves Aspen, suggesting that removing noisy dimensions benefits these classes. However, PCA substantially reduces Cottonwood/Willow performance from 0.62 to 0.40, implying that its discriminative information lies in directions discarded during variance-based compression. RP largely preserves Cottonwood/Willow (0.59) but performs worse on Aspen, while ICA lands between PCA and RP on some minority classes but is weakest on others. These results indicate that class-discriminative information is distributed unevenly across the feature space rather than concentrated in the dominant variance directions.

TABLE VII: Per-class test F1

Class	Orig	PCA	ICA	RP
1 Spruce/Fir	0.797	0.777	0.772	0.783
2 Lodgepole Pine	0.834	0.823	0.812	0.831
3 Ponderosa Pine	0.760	0.724	0.721	0.757
4 Cottonwood/Willow	**0.615**	0.400	0.480	0.593
5 Aspen	0.463	**0.476**	0.376	0.388
6 Douglas-fir	0.428	**0.496**	0.432	0.408
7 Krummholz	0.781	**0.796**	0.743	0.755

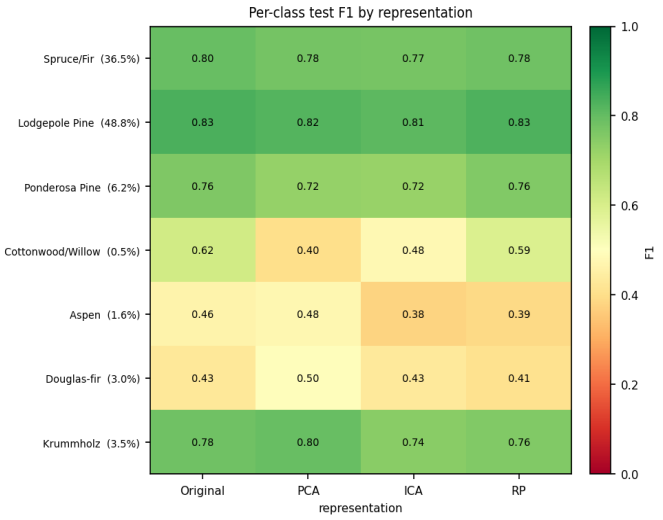


Fig. 11: Per-class F1 by Representation

With current architecture, dimensionality reduction does not improve downstream neural network performance on the Covertype dataset. Instead, it illustrates the trade off between

denoising and information loss. PCA removes enough redundant variation to improve training stability while preserving most predictive performance despite reducing the feature dimension from 54 to 10. ICA behaves similarly but offers no consistent advantage. RP preserves global geometry but loses discriminative information needed for minority classes. Overall, the original feature representation remains the best choice for supervised learning, suggesting that class relevant information is distributed across many dimensions and that compressing the feature space inevitably sacrifices part of that signal.

VII. Extra Credit - Nonlinear Manifold Learning

To explore whether the data contain nonlinear structure that linear methods cannot capture, I evaluated Isomap as a nonlinear manifold learning technique. Unlike t-SNE, Isomap provides an out-of-sample transform, making it suitable for a leakage free train/validation/test pipeline. I used 10 neighbors to construct the neighborhood graph and fit the model only on the training data, then transformed the validation and test sets with the learned embedding. For visualization, I compared a 2-D Isomap embedding against a 2-D PCA embedding on the same 4,000 point subsample.

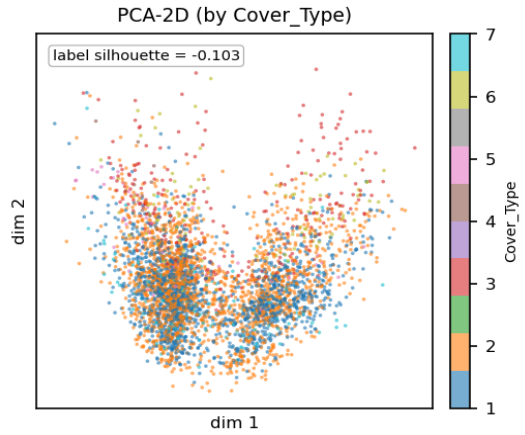


Fig. 12: PCA-2D

Fig 12 and 13 are showing heavily overlapping Cover Type classes and indicating that the non linear embedding does not reveal additional class structure beyond the linear methods. PCA 2D achieves a silhouette of -0.103, while Isomap 2D performs even worse at -0.133. Thus, even after accounting for nonlinear neighborhood relationships, samples from the same class remain closer to samples from other classes than to their own.

TABLE VIII: Isomap-10 vs. Linear Reductions

Rep.	Macro-F1	Bal. Acc	F1 sd
PCA-10	0.642	0.606	0.006
ICA-10	0.639	0.627	0.010
RP-20	0.607	0.569	0.018
Isomap-10	0.522	0.493	0.029

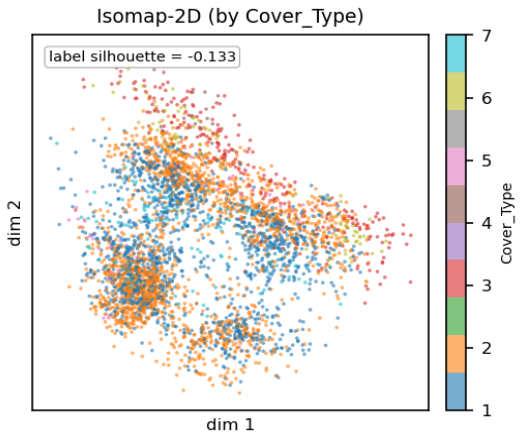


Fig. 13: Isomap-2D

As shown in table VIII, the Isomap representation achieves a Macro-F1 of 0.522, substantially below PCA (0.642), ICA (0.639), RP (0.607), and the original feature space (0.668 table VI). It also produces the lowest balanced accuracy (0.493) and the highest variability across random seeds (0.029). Rather than uncovering hidden nonlinear structure, Isomap appears to distort the mixed continuous and sparse binary feature space, causing useful class information to be lost.

Overall, Isomap provides no advantage over the linear dimensionality reduction methods on the Covertype dataset. Both the visualization and quantitative results suggest that the dataset does not lie on a low dimensional nonlinear manifold that is more informative than the representations learned by PCA, ICA, or RP. Among all dimensionality reduction methods evaluated, PCA remains the strongest choice, preserving the best balance between dimensionality reduction, predictive performance, and training stability.

VIII. Discussion

This report shows that Covertype contains strong geometric structure, but that structure is only weakly related to the ecological Covertype labels. Across every experiment, clustering algorithms readily discovered compact groups in the feature space, but those groups rarely corresponded to the seven classes. K-Means consistently preferred a coarse 2 cluster partition with relatively high silhouette scores, indicating that the dominant organization of the data is a broad geometric split driven by elevation, distances, and wilderness characteristics rather than tree species. GMM uncovered a different type of structure by modeling overlapping elliptical densities, producing substantially better label alignment than K-Means (best NMI 0.209 versus 0.064), although this agreement remains weak in absolute terms. Together, these results suggest that the dataset contains meaningful unsupervised structure, but that the ecological labels are not the primary organizing principle of the original feature space.

For dimensionality reduction algorithms, PCA, ICA, and RP all improved internal clustering quality by producing

more compact and better separated clusters, but none of them improved Covertype labels. Instead, reducing the feature space consistently widened the gap between geometric quality and semantic relevance. RP produced the strongest internal clustering, while ICA was the only method that changed K-Means' preferred solution from two clusters to seven, modestly improving label alignment. Nevertheless, the original-feature GMM remained the best representation of the class labels. The evidence therefore demonstrates that better geometric clusters do not necessarily represent more meaningful classes.

On the downstream neural network experiments, none of the reduced representations surpassed the original feature space, indicating that discriminative information is distributed across many dimensions rather than concentrated in a small subspace. Thus, on this dataset, dimensionality reduction is valuable primarily as a tool for compression, denoising, and improved optimization stability, not for increasing predictive accuracy.

On the other hand, K-Means and RP maximize silhouette while minimizing label agreement. GMM shows the opposite behavior, producing lower internal clustering quality but substantially better alignment with the true classes. Dimensionality reduction generally widens this disconnect. It improves silhouette while reducing NMI in five of the six DR clustering combinations.

The effects of dimensionality reduction are concentrated in the minority classes rather than the majority classes. PCA helped Douglas-fir, Aspen, and Krummholz but halved Cottonwood, whose signal sits in low variance directions. RP protected Cottonwood but hurt Aspen. These results demonstrate that no dimensionality reduction method consistently benefits all minority classes. Instead, its impact depends on how each class's discriminative information is distributed across the feature space, highlighting the trade-off between denoising and information preservation.

Furthermore, there are several limitations. First, the sample is small with only 20,000 observations and is also highly imbalanced that causes some classes like Cottonwood/Willow to have only 94 samples, making minority class metrics and any associated cluster estimates inherently noisy. Second, silhouette and DB depend on the continuous/binary scaling choice. A different weighting would shift the internal metrics. Third, ICA's most non-Gaussian axes are dominated by sparse binary flags and RP carries seed variability, so both are only conditionally stable, and Isomap was fit on a 3,000 point landmark subsample for tractability, making its embedding approximate. Finally, GMM's BIC never reached an interior optimum, so the component count leans on the k = 7 reference rather than a decisive elbow.

In my first hypothesis, I expect K-Means aligns weakly on Covertype dataset and in fact, it preferred a 2 cluster partition than the correct 7 clusters. GMM aligned about 3 times better in term of NMI because its soft elliptical densities fit the imbalanced, overlapping classes. So this hypothesis holds true.

For the second hypothesis with dimensionality reduction algorithms, PCA did preserve downstream performance best on Macro-F1 and accuracy, and was the most stable, matching the correlated, low-rank continuous features reasoning. But ICA tied it on Macro-F1 and beat it on balanced accuracy. RP kept the best raw accuracy while being worst for minority classes. So PCA is best expectation is true for majority metrics and stability, but not universally. So this hypothesis is partially correct.

IX. Conclusion

This study demonstrates that unsupervised learning reveals the intrinsic geometry of the Covertype feature space rather than the ecological classes themselves. K-Means and GMM expose complementary aspects of that geometry, while PCA, ICA, and Random Projection simplify the feature space into cleaner and more interpretable representations. However, improved cluster geometry rarely translates into better agreement with the Covertype labels or stronger downstream neural network performance, indicating that the natural structure of the data is driven primarily by environmental gradients rather than species boundaries. Among the dimensionality reduction methods, PCA provides the best overall trade-off by preserving nearly all neural network performance with a much smaller and more stable representation, whereas ICA offers comparable performance with slightly better class balance and RP favors majority classes at the expense of minority ones. Overall, clustering should be viewed as an exploratory tool for understanding data geometry, not as a proxy for semantic labels, and dimensionality reduction is most valuable for representation learning, denoising, visualization, and model simplification rather than improving classification accuracy on this dataset.

AI USE STATEMENT

I used ChatGPT, Claude and Windsurf IDE to brainstorm the project plan, generate/ debug/ refactor code and OverLeaf AI suggestions for Latex syntax. I also used ChatGPT to learn more about the concepts and Python libraries. The final design, workflow and analysis are my own. I reviewed, verified, and understand all assisted content.

References

[1] I. T. Jolliffe and J. Cadima, "Principal Component Analysis: A Review and Recent Developments," Philosophical Transactions 2016.

[2] E. Bingham and H. Mannila, "Random Projection In Dimensionality Reduction: Applications to Image and Text Data," In Proc. 7th ACM SIGKDD, 2001

[3] A. Hyvarinen and E. Oja, "Independent Component Analysis: Algorithms and Applications," Neural Networks, vol. 13, 2000.

[4] J. Sarela and R. Vigario, "Overlearning in marginal distribution-based ICA: analysis and solutions," J. Mach. Learn. Res., vol. 4, 2003.