**Geospatial Classification of Lincoln County, Oregon Using K-Prototypes Clustering**

**Abstract**

This project applies unsupervised machine learning (k-prototypes clustering) to geospatial datasets in Lincoln County, Oregon. By combining terrain features such as slope, land cover, and soil/hydrology attributes, the study identifies regions with similar environmental characteristics. The approach demonstrates how clustering can uncover spatial patterns in heterogeneous landscapes and highlights the potential of geospatial machine learning for exploratory land classification.

**Introduction**

Geospatial clustering provides a systematic way to analyze large, multi-layered environmental datasets without relying on pre-labeled outcomes. Unlike supervised learning approaches, clustering groups regions based on inherent similarities in terrain and land attributes, enabling the discovery of natural patterns and hidden structures within the landscape.

**Why Lincoln County?**  
Lincoln County, located on the Oregon Coast, presents a unique and diverse terrain that includes coastal lowlands, forested uplands, and river valleys. The county’s combination of varied land cover, hydrological conditions, and drainage characteristics makes it a representative testbed for exploring unsupervised clustering methods. The geographic diversity also allows meaningful interpretation of clusters in relation to environmental and land use patterns.

Why K-Prototypes?  
K-prototypes (Huang, 1997) allows clustering of mixed-type data by combining numeric and categorical features. Numeric variables use Euclidean distance, categorical features use simple matching dissimilarity, and a weighting parameter balances their influence. This makes it ideal for datasets with terrain metrics (e.g., slope) and land attributes (e.g., land cover, hydrology, drainage class).

**Applications**  
Potential applications of such an analysis include land management planning, environmental monitoring, and the preliminary assessment of areas with distinct ecological or terrain-based characteristics. While this project does not directly aim to predict hazards, it provides an analytical foundation that could later support research in hazard zoning, resource allocation, and sustainable development strategies.

**Objective**

To apply k-prototypes clustering on geospatial features (slope, land cover, and hydrological attributes) in order to classify the landscape into distinct environmental zones.

**Data and Preprocessing**

Datasets used:

Slope (derived from DEM) – continuous

Land cover (NLCD / classification raster) – categorical

Hydrological group code (hydgrpdcd) – categorical

Drainage class (drclassdcd) – categorical

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Pixel size** | **Values** |
| Slope (derived from DEM) | Continuous | 98.42519685000000607,  -98.42519685000000607 | 0.03–26.5 degrees |
| Landcover | Categorical | 98.42519685000000607,  -98.42519685000000607 | 11–95 |
| Hyd group code | Categorical | 98.42519685000000607,  -98.42519685000000607 | A–D and mixed groups A/D, C/D |
| Drainage class | Categorical | 98.42519685000000607,  -98.42519685000000607 | Drainage properties |

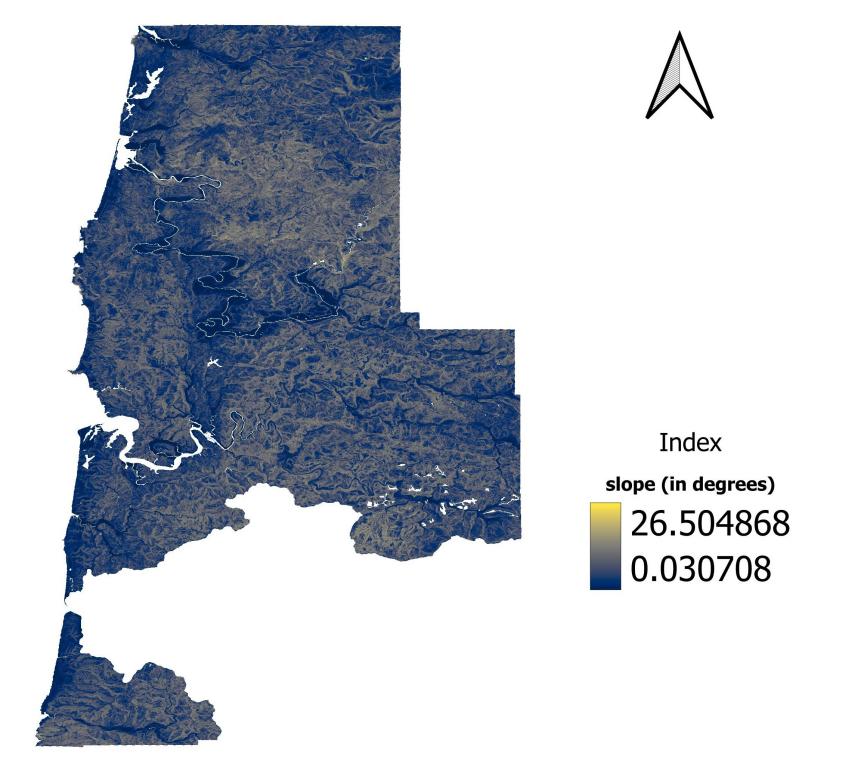
****

Figure: Slope expressed in degrees

Later slope values were normalized to the [0, 1] range to ensure comparability with categorical features for clustering.

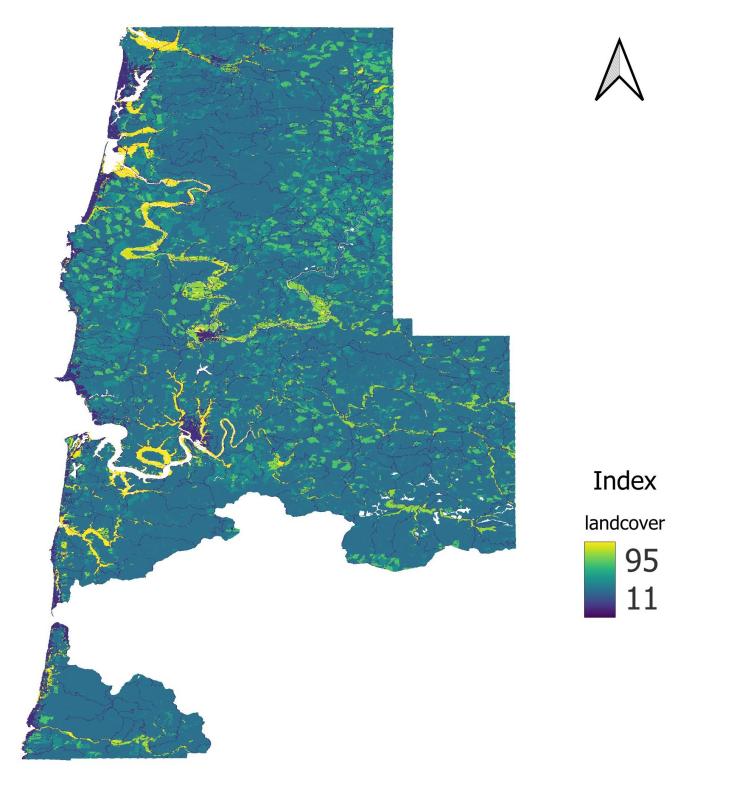
****

Figure: Landcover

NLCD Land Cover Classification Codes (2019 standard)

Water

11 – Open Water

12 – Perennial Ice/Snow

Developed

21 – Developed, Open Space

22 – Developed, Low Intensity

23 – Developed, Medium Intensity

24 – Developed, High Intensity

Barren

31 – Barren Land (Rock/Sand/Clay)

Forest

41 – Deciduous Forest

42 – Evergreen Forest

43 – Mixed Forest

Shrubland

51 – Dwarf Scrub

52 – Shrub/Scrub

Herbaceous

71 – Grassland/Herbaceous

72 – Sedge/Herbaceous

73 – Lichens

74 – Moss

Planted/Cultivated

81 – Pasture/Hay

82 – Cultivated Crops

Wetlands

90 – Woody Wetlands

95 – Emergent Herbaceous Wetlands

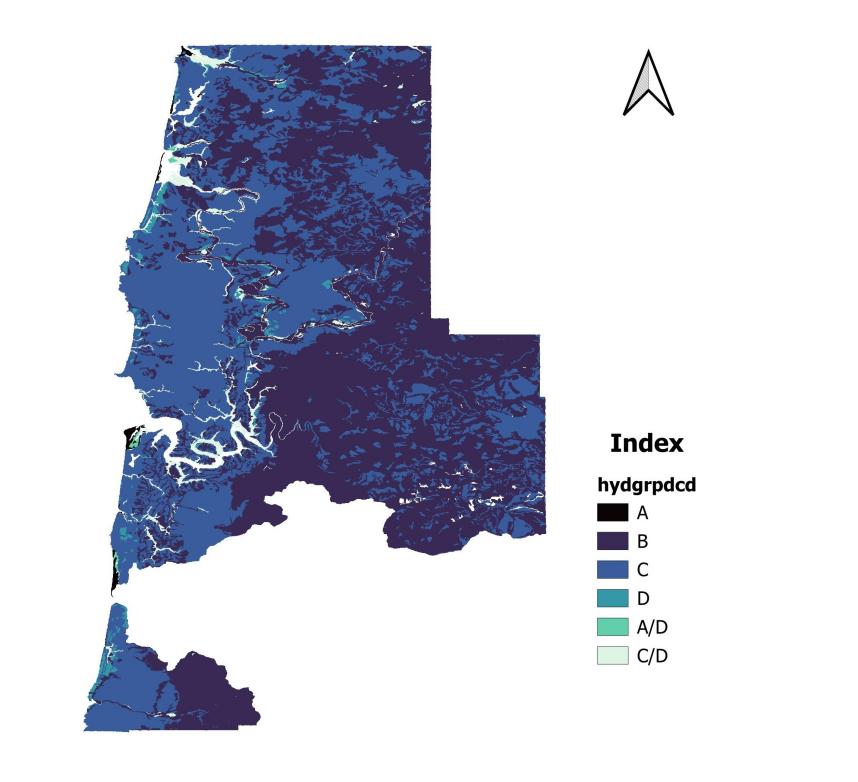
****

Figure: Hydgrpdcd

Hydgrpdcd stands for Hydrologic Soil Group Code. It’s a standard classification used in hydrology and soil science to categorize soils based on their runoff potential and infiltration characteristics.

A - High infiltration capacity, very low runoff potential (sand, loamy sand, gravelly soils).

B - Moderate infiltration capacity, moderate runoff potential (silt loam).

C - Low infiltration capacity, moderately high runoff potential (sandy clay loam).

D - Very low infiltration capacity, very high runoff potential (clay soils, shallow soils, or water table near surface).

For dual groups (A/D, B/D, C/D):

The first letter applies if the soil is drained (artificial drainage or naturally well-drained).

The second letter (D) applies if the soil is undrained (poorly drained, high water table).

Each categorical feature was encoded with numerical values ranging from 1 to 6, enabling seamless integration into the machine learning workflow.

A - 1

B - 2

C - 3

D - 4

A/D - 5

C/D - 6

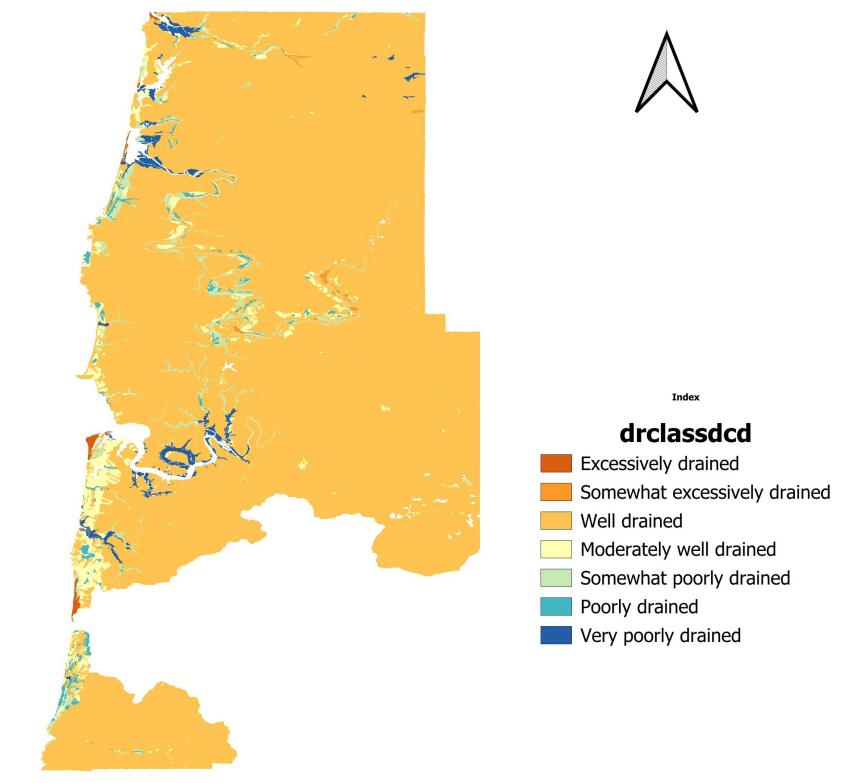
****

Figure: Drclassdcd

drclassdcd usually refers to “Drainage Class”, a categorical attribute in soil or terrain datasets. It indicates how quickly water moves through or drains from the soil in a given area.

Each categorical feature was encoded with numerical values ranging from 1 to 7, enabling seamless integration into the machine learning workflow.

Excessively drained - 1

Somewhat excessively drained - 2

Well drained - 3

Moderately well drained - 4

Somewhat poorly drained - 5

Poorly drained - 6

Very poorly drained - 7

Data Preprocessing

These datasets in raster format was loaded in QGIS for proper vizualization. The geospatial datasets were preprocessed to ensure consistency and compatibility with the k-prototypes clustering algorithm, which accommodates both numeric and categorical variables. To achieve uniformity, all layers were projected to the same coordinate reference system and aligned spatially. Since the slope raster initially had a finer pixel resolution, it was resampled to match the resolution of the landcover data.

The final features used in the analysis were:

Slope (numeric, Float32, range 0.03–26.5)

Landcover (categorical, Float32, values 11–95)

Hydrological group (hydgrpdcd) (categorical, classes A–D and mixed groups A/D, C/D)

Drainage class (drclassdcd) (categorical)

Hydrological group and drainage class were derived from the soil dataset and clipped to the study area for consistency.

Preprocessing steps included:

Handling missing values: All NoData or invalid values were masked and replaced with a sentinel value (-9999) to avoid interfering with clustering.

Normalization: The numeric slope values were normalized to a 0–1 scale, ensuring comparability with categorical features in the clustering process.

Categorical encoding:

hydgrpdcd categories were mapped to numeric codes (1–6) for ML processing.

landcover values, although originally numeric (11–95), were treated as categorical classes.

drclassdcd values were used as integer codes.

Clipping and alignment: All rasters were clipped to the study area and aligned to a common CRS (EPSG:2992) and pixel resolution. Raster pixels common to all the raster files and having a valid data were only selected for analysis.

Extraction for clustering: Raster values were extracted to arrays corresponding to valid pixels, excluding sentinel NoData values, to form the input matrix for k-prototypes.

This preprocessing ensured that the numeric slope and categorical environmental features were properly scaled, encoded, and aligned for unsupervised clustering.

**Methodology**

The project applied unsupervised clustering using K-Prototypes to classify environmental patterns in Lincoln County, Oregon. The workflow consisted of the following steps:

Feature Selection

Four key features were chosen based on their relevance to terrain and land characteristics as well as data availability:

* Slope (numeric and continuous) : Represents terrain steepness, an important factor in land stability.
* Land Cover (categorical): Indicates surface characteristics such as forest, agricultural land, or built-up areas.
* Hydrological Soil Group (categorical): Classifies soils based on their runoff potential.
* Drainage Class (categorical) – describes the natural drainage condition of soils.

These features were selected because complete datasets were available, ensuring consistency and minimizing missing data issues.

Raster datasets representing slope, land cover, hydrological group (hydgrpdcd), and drainage class (drnclassdcd) were read using rasterio.

Preprocessing

Using QGIS these resulting rasters were clipped with the datasets of the study area boundary.

Invalid or missing data were filtered out, including values outside expected ranges, NaN, or -9999.

Slope, a numeric feature, was normalized between 0 and 1 to standardize its scale relative to categorical features.

Categorical features (land cover, hydgrpdcd, drnclassdcd) were converted to integer codes. K-Prototypes requires numeric features for Euclidean distance and integer-coded categorical features for simple matching dissimilarity. Normalization prevents numeric dominance in clustering.

Clustering Approach

Numeric and categorical columns were specified to allow the algorithm to treat different data types appropriately. K-Prototypes allows clustering on mixed-type data (numeric + categorical) by combining Euclidean distance for numeric features and dissimilarity for categorical features.

Python programming was performed using VS Code. To train the K-Prototypes model, a 5% random sample of valid pixels was selected. Sampling was necessary to speed up computation while still preserving representative patterns from the large raster datasets. When a 50% sample was tested, the run time extended for several hours without completing due to limitations in RAM and processing power. By reducing the sample size to 5%, the clustering process successfully completed in about 10 minutes.

Applied the Elbow Method to determine an appropriate number of clusters by analyzing the cost function trend.

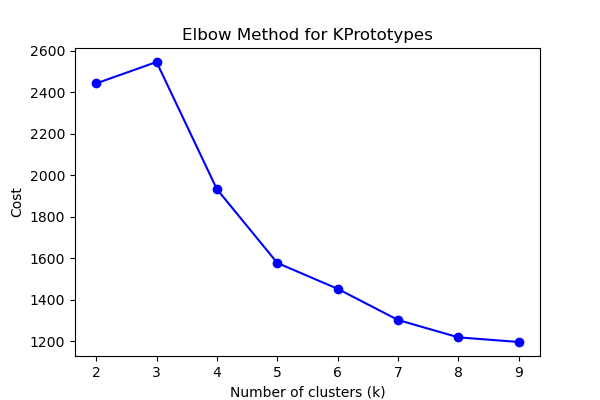


Figure: Plot of Cost vs number of clusters (k)

K-Prototypes was applied with k = 5 clusters, using the Huang initialization method and multiple runs (n\_init=5) to ensure stability.

The full Python scripts used for data preprocessing, feature extraction, and K-Prototypes clustering are provided in the scripts/ folder of the project repository. These scripts include:

Raster cleaning and nodata handling

Normalization of numeric features and encoding of categorical features

Clustering using K-Prototypes and mapping of cluster results to raster format

Visualization of clustered maps

Cluster labels from sampled data were assigned to the full dataset using the trained centroids. A full raster of clusters was reconstructed and saved in the results folder.

Evaluation

Assessed the clustering performance using the cost function (minimization of distance-based dissimilarity).

Interpreted clusters in terms of their relative combinations of slope, land cover, hydrological group, and drainage class.

**Results**

Below shows the output generated after running the code.

Init: initializing centroids

Init: initializing clusters

Starting iterations...

Run: 1, iteration: 1/100, moves: 23076, ncost: 4345.480609876242

Run: 1, iteration: 2/100, moves: 2153, ncost: 4342.014461126517

Run: 1, iteration: 3/100, moves: 309, ncost: 4341.841649113543

Run: 1, iteration: 4/100, moves: 78, ncost: 4341.82382305947

Run: 1, iteration: 5/100, moves: 26, ncost: 4341.820917435647

Run: 1, iteration: 6/100, moves: 7, ncost: 4341.820821662217

Run: 1, iteration: 7/100, moves: 1, ncost: 4341.820821429787

Run: 1, iteration: 8/100, moves: 0, ncost: 4341.820821429787

Init: initializing centroids

Init: initializing clusters

Starting iterations...

Run: 2, iteration: 1/100, moves: 17207, ncost: 4903.425997992716

Run: 2, iteration: 2/100, moves: 8381, ncost: 4784.0297284735125

Run: 2, iteration: 3/100, moves: 7133, ncost: 4748.690987712513

Run: 2, iteration: 4/100, moves: 3433, ncost: 4720.110594196892

Run: 2, iteration: 5/100, moves: 11307, ncost: 4646.483832237952

Run: 2, iteration: 6/100, moves: 2503, ncost: 4644.561762278467

Run: 2, iteration: 7/100, moves: 242, ncost: 4644.536855396017

Run: 2, iteration: 8/100, moves: 44, ncost: 4644.535908906638

Run: 2, iteration: 9/100, moves: 7, ncost: 4644.535877096176

Run: 2, iteration: 10/100, moves: 0, ncost: 4644.535877096176

Init: initializing centroids

Init: initializing clusters

Starting iterations...

Run: 3, iteration: 1/100, moves: 26111, ncost: 4667.070876654311

Run: 3, iteration: 2/100, moves: 8176, ncost: 4646.859176295018

Run: 3, iteration: 3/100, moves: 7570, ncost: 4599.122713393841

Run: 3, iteration: 4/100, moves: 3157, ncost: 4594.291840041943

Run: 3, iteration: 5/100, moves: 188, ncost: 4594.2734961076

Run: 3, iteration: 6/100, moves: 13, ncost: 4594.273381815071

Run: 3, iteration: 7/100, moves: 0, ncost: 4594.273381815071

Init: initializing centroids

Init: initializing clusters

Starting iterations...

Run: 4, iteration: 1/100, moves: 3342, ncost: 4805.267065862894

Run: 4, iteration: 2/100, moves: 268, ncost: 4805.232212882113

Run: 4, iteration: 3/100, moves: 28, ncost: 4805.231918901234

Run: 4, iteration: 4/100, moves: 0, ncost: 4805.231918901234

Init: initializing centroids

Init: initializing clusters

Starting iterations...

Run: 5, iteration: 1/100, moves: 8410, ncost: 4812.178991077252

Starting iterations...

Run: 4, iteration: 1/100, moves: 3342, ncost: 4805.267065862894

Run: 4, iteration: 2/100, moves: 268, ncost: 4805.232212882113

Run: 4, iteration: 3/100, moves: 28, ncost: 4805.231918901234

Run: 4, iteration: 4/100, moves: 0, ncost: 4805.231918901234

Init: initializing centroids

Init: initializing clusters

Starting iterations...

Run: 5, iteration: 1/100, moves: 8410, ncost: 4812.178991077252

Run: 4, iteration: 1/100, moves: 3342, ncost: 4805.267065862894

Run: 4, iteration: 2/100, moves: 268, ncost: 4805.232212882113

Run: 4, iteration: 3/100, moves: 28, ncost: 4805.231918901234

Run: 4, iteration: 4/100, moves: 0, ncost: 4805.231918901234

Init: initializing centroids

Init: initializing clusters

Starting iterations...

Run: 5, iteration: 1/100, moves: 8410, ncost: 4812.178991077252

Run: 4, iteration: 2/100, moves: 268, ncost: 4805.232212882113

Run: 4, iteration: 3/100, moves: 28, ncost: 4805.231918901234

Run: 4, iteration: 4/100, moves: 0, ncost: 4805.231918901234

Init: initializing centroids

Init: initializing clusters

Starting iterations...

Run: 5, iteration: 1/100, moves: 8410, ncost: 4812.178991077252

Run: 4, iteration: 4/100, moves: 0, ncost: 4805.231918901234

Init: initializing centroids

Init: initializing clusters

Starting iterations...

Run: 5, iteration: 1/100, moves: 8410, ncost: 4812.178991077252

Run: 5, iteration: 2/100, moves: 8681, ncost: 4737.058789836662

Run: 5, iteration: 3/100, moves: 3781, ncost: 4716.4919011857255

Init: initializing centroids

Init: initializing clusters

Starting iterations...

Run: 5, iteration: 1/100, moves: 8410, ncost: 4812.178991077252

Run: 5, iteration: 2/100, moves: 8681, ncost: 4737.058789836662

Run: 5, iteration: 3/100, moves: 3781, ncost: 4716.4919011857255

Run: 5, iteration: 1/100, moves: 8410, ncost: 4812.178991077252

Run: 5, iteration: 2/100, moves: 8681, ncost: 4737.058789836662

Run: 5, iteration: 3/100, moves: 3781, ncost: 4716.4919011857255

Run: 5, iteration: 2/100, moves: 8681, ncost: 4737.058789836662

Run: 5, iteration: 3/100, moves: 3781, ncost: 4716.4919011857255

Run: 5, iteration: 4/100, moves: 7830, ncost: 4665.03143748339

Run: 5, iteration: 4/100, moves: 7830, ncost: 4665.03143748339

Run: 5, iteration: 5/100, moves: 2964, ncost: 4659.447481862506

Run: 5, iteration: 5/100, moves: 2964, ncost: 4659.447481862506

Run: 5, iteration: 6/100, moves: 454, ncost: 4659.3287103596185

Run: 5, iteration: 7/100, moves: 60, ncost: 4659.327656025226

Run: 5, iteration: 8/100, moves: 3, ncost: 4659.327649145338

Run: 5, iteration: 9/100, moves: 0, ncost: 4659.327649145338

Best run was number 1

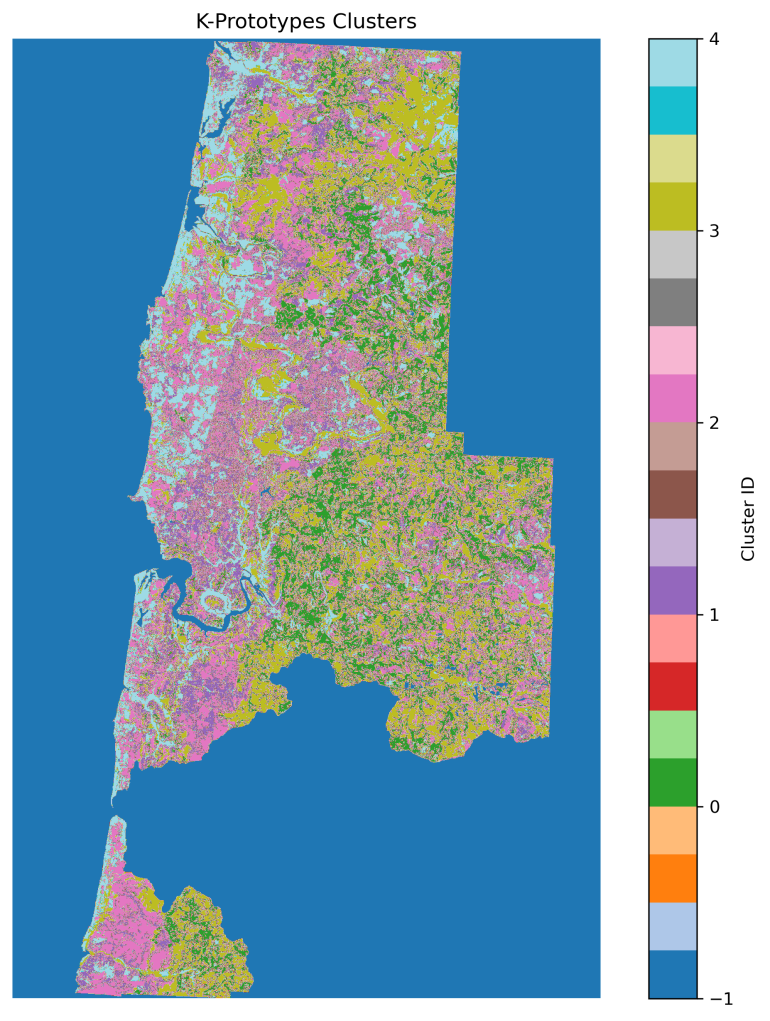
****

Figure: It represents different clusters formed after the kprotypes clustering. Here, -1 represents no data

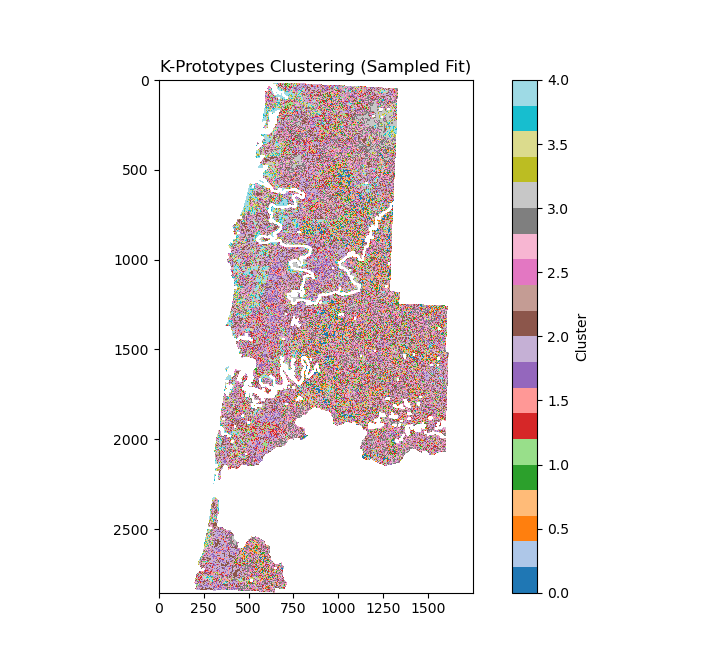
****

Figure: Clustered output with noise removed and mapped back to the original raster map from code after completion

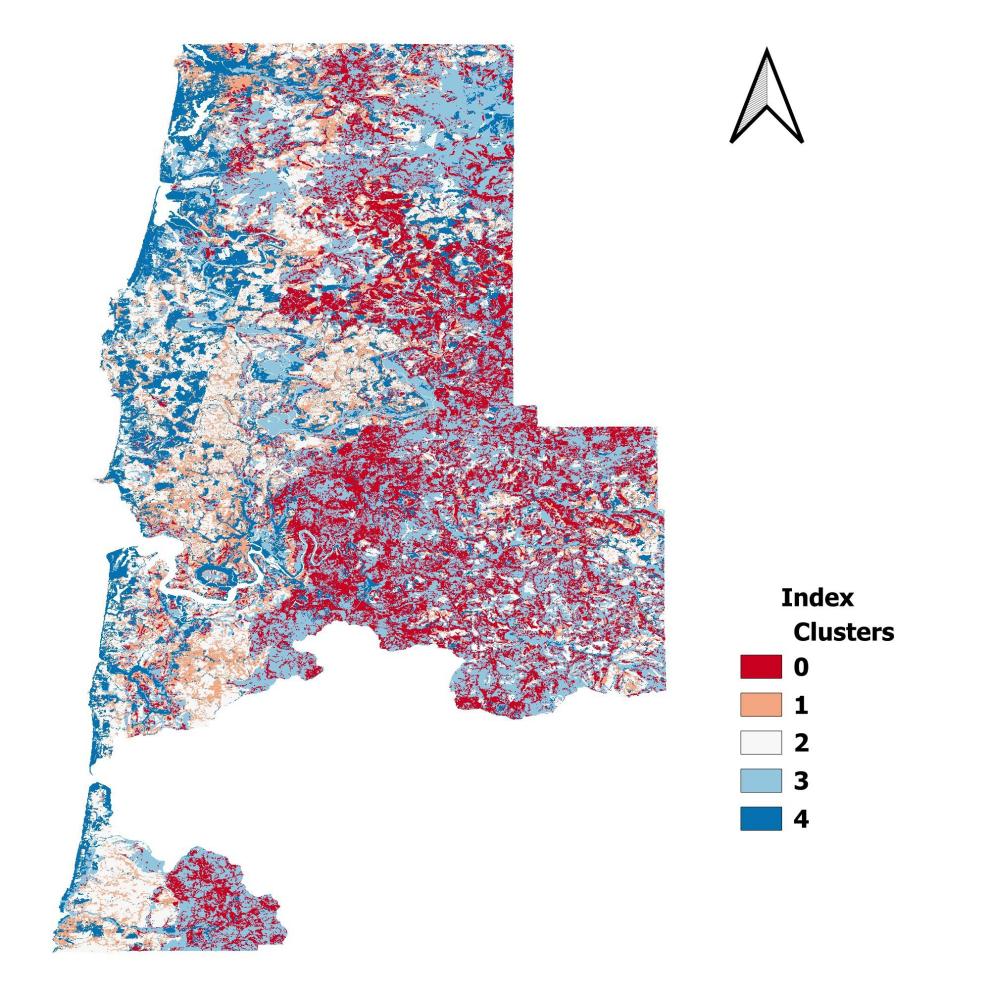
****

Figure: Final clean clustered map derived from QGIS

Cluster Summary (for k=5):

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster** | **Dominant Hydrological Group** | **Dominant Drainage Class** | **Dominant Landcover** | **Slope (in degrees)** | | | |
| mean | median | min | max |
| 0 | B (99.9%) | Well drained (99.8%) | Mixed Forest (66.8%) | 8.684886 | 9.029634 | 0.103834 | 26.50487 |
| 1 | C (97.9%) | Well drained (94.5%) | Mixed Forest (82.0%) | 7.426737 | 7.424306 | 0.12319 | 25.95542 |
| 2 | C (99.8%) | Well drained (94.0%) | Evergreen Forest (94.8%) | 6.388248 | 6.062947 | 0.091171 | 26.44507 |
| 3 | B (98.4%) | Well drained (98.2%) | Evergreen Forest (73.6%) | 6.334602 | 6.040799 | 0.035669 | 17.54043 |
| 4 | C (67.6%) | Well drained (70.4%) | Shrub/Scrub (55.9%) | 3.964393 | 3.471002 | 0.030708 | 12.56138 |

The raster file and images representing the clustered map and used in this report can be found in the results folder.

**Discussion**

The K-Prototypes clustering algorithm was run multiple times to ensure stability. Each run converged after a few iterations, with the cost function (ncost) indicating the internal consistency of clusters. The best run (lowest ncost) was selected for reconstructing the cluster map. This demonstrates that the clustering process reached a stable solution and the resulting clusters reliably reflect underlying spatial patterns.

Cluster Results

Cluster -1:

Consists of nodata regions.

Cluster 0:

Dominated by Hydrological Group B (99.9%) and Well-drained soils (99.8%), indicating relatively stable drainage conditions.

Landcover: Mostly Mixed Forest (66.8%), suggesting a heterogeneous forest type.

Slope: Mean slope is ~8.68°, with values ranging from 0.10° to 26.50°, indicating a mix of gentle to moderately steep areas.

Interpretation: This cluster represents moderately sloped forested areas with well-drained soils and a mix of tree types, likely less prone to waterlogging but with some variability in terrain.

Cluster 1:

Dominated by Hydrological Group C (97.9%) and Well-drained soils (94.5%).

Landcover: Mixed Forest (82%), slightly more homogeneous than Cluster 0.

Slope: Mean ~7.43°, range 0.12°–25.96°.

Interpretation: This cluster highlights slightly gentler slopes than Cluster 0 with well-drained soils and mixed forests. These areas may have moderate susceptibility to slope-related processes due to terrain variability.

Cluster 2:

Dominated by Hydrological Group C (99.8%) and Well-drained soils (94%).

Landcover: Evergreen Forest (94.8%), indicating highly homogeneous forest type.

Slope: Mean ~6.39°, range 0.09°–26.44°.

Interpretation: Mostly flat to gently sloping evergreen forests with well-drained soils, representing stable forested areas.

Cluster 3:

Dominated by Hydrological Group B (98.4%) and Well-drained soils (98.2%).

Landcover: Evergreen Forest (73.6%), slightly more mixed than Cluster 2.

Slope: Mean ~6.33°, range 0.04°–17.54°.

Interpretation: Similar to Cluster 2 but with slightly lower slopes and minor variations in forest type. These areas may have low-to-moderate hazard potential given their slope and drainage characteristics.

Cluster 4:

Dominated by Hydrological Group C (67.6%) and Well-drained soils (70.4%).

Landcover: Shrub/Scrub (55.9%), indicating more open, less forested areas.

Slope: Mean ~3.96°, range 0.03°–12.56°, the gentlest slopes among all clusters.

Interpretation: These areas are mostly gentle, shrub-dominated landscapes with well-drained soils. They likely experience lower slope-related risks but may be more exposed to land cover changes or human disturbances.

Hydrology and Drainage: Almost all clusters are dominated by well-drained soils, reflecting stable subsurface conditions in the study area.

Landcover Variation: Evergreen forests dominate clusters with lower slopes, while mixed forests occupy slightly steeper areas. Shrub/Scrub dominates the gentlest slopes, indicating a gradient in vegetation types corresponding to slope and possibly land management.

Slope Trends: Clusters with higher slopes tend to have mixed forests, whereas lower slopes correspond to evergreen forests or shrublands. This suggests terrain is a key factor influencing vegetation distribution.

Applications:

Provides exploratory zoning of the landscape.

Could inform preliminary land planning and environmental assessment.

Limitations:

Results sensitive to categorical encoding balance.

Limited diversity in hydrological/soil attributes (drainage nearly constant).

No ground truth validation at this stage.

**Conclusion**

This project successfully applied k-prototypes clustering for mixed-type geospatial classification in Lincoln County. The analysis demonstrates that unsupervised clustering can reveal broad environmental zones, primarily structured by slope and land cover.

Future Work:

The methodology developed in this study can be extended to other regions, including areas like Nepal, where high-quality geospatial datasets are often limited. With improved data availability, similar unsupervised clustering can reveal environmental patterns and support preliminary hazard assessments. Future work could include:

* Applying the workflow to regions with available or newly collected datasets, adapting to local data constraints.
* Incorporating higher-resolution soil, hydrology, and land cover data to improve cluster differentiation.
* Validating clusters against historical hazard or environmental records to assess the reliability of the classification.
* Comparing K-Prototypes results with other clustering approaches, such as hierarchical and density-based methods, to evaluate performance and interpretability.

**References**

Lincoln County. (n.d.). Lincoln County GIS Maps [Data set]. Retrieved from [https://maps.co.lincoln.or.us](https://maps.co.lincoln.or.us" \t "_new)

Oregon Department of Forestry. (n.d.). Oregon Digital Elevation Model (DEM) [Data set]. State of Oregon. Retrieved from https://www.oregon.gov

U.S. Geological Survey. (2023, November 2). National Land Cover Database (NLCD) – Oregon. Oregon GIS Framework. U.S. Geological Survey. Retrieved from <https://www.mrlc.gov/>

U.S. Department of Agriculture, Natural Resources Conservation Service. (2024, August 29). Soil Survey Geographic (SSURGO) database for Lincoln County Area, Oregon (or638). Fort Worth, TX: U.S. Department of Agriculture, Natural Resources Conservation Service. Retrieved from [https://websoilsurvey.sc.egov.usda.gov/](https://websoilsurvey.sc.egov.usda.gov/" \t "_new)

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining.