Graph Based Approach to Unsupervised Clustering of USA Land Cover

Benjaimin Elfner *CEAS*
University of Cincinnati
Cincinnati, USA
elfnerbm@mail.uc.edu

Abstract—We present a method to create a clustering of land cover using planar graphs. Doing so transforms the problem into a vertex clustering one. This new representation is created using adjacent segments of land cover classes. Network representation techniques are then used to create component vectors for each node of the graph. These node representations are then clustered using standard clustering algorithms such as k-means or hierarchical to create regions with similar land patterns.

Index Terms—clustering, NLCD, land cover, pattern analysis, graph representation

1 Introduction

Land cover analysis is crucial to forming decisions for tasks such as climate change or land management [1]. Regions can be formed by finding areas of land that exhibit similar spatial patterns of land cover also known as land pattern types (LPTs). These regions are important to many fields since they allow generalizations to be made of the land contained which can speed up analysis [1]. For example, measurements in one part of a region could apply to all related area of land which reduces the work required to collect data.

2 Underling Data Mining Problem

For a given area of classified land cover (represented as a image), form regions of similar land pattern types.

3 Data Used

The data that will be used is the NLCD 2019 Land Cover (CONUS) dataset. This data, collected in 2019, consists of a classification for each 30x30 meter square in the contiguous 48 US states. There are 20 land cover classes that describe contents of the land cover for each square and ar grouped into 8 super classes. [2]

4 Related Works

Previously this problem was approached used the formation of motifels, small tiling subsections of the land cover data [1]. A co-occurrence histogram was created for each motifel where pixel adjacency are categorized and counted. These histograms are combined into contiguous segments using the similarity of motifels' histogram then those segments were clustered using hierarchical clustering. The benefit of this approach is reduced computational complexity since the data can be split into sections larger than a single pixel. The downside to this approach is the measure of similarity only considers the connections between two pixel and ignores larger patterns found in the data.

5 Our Approach

We will approach this problem by viewing land cover data as a planar graph. Each contiguous segment of land cover that shares a common class will represent the vertices and the edges are between adjacent segments. A possible modification of this method is to give a weight to each edge corresponding to the length of the shared perimeter.

With this representation new approaches can be used to compare vertices of the graph. In this paper we will be comparing three methods. These methods were selected due to their varying methodologies, optimization methods, and experimental performance. Basic details are given in TABLE 1.

Each method produces vector embedding for each vertex of the graph. These vectors will be clustered using hierarchical clustering with ward linkage.

5.1 Linear Graph Autoencoders [3]

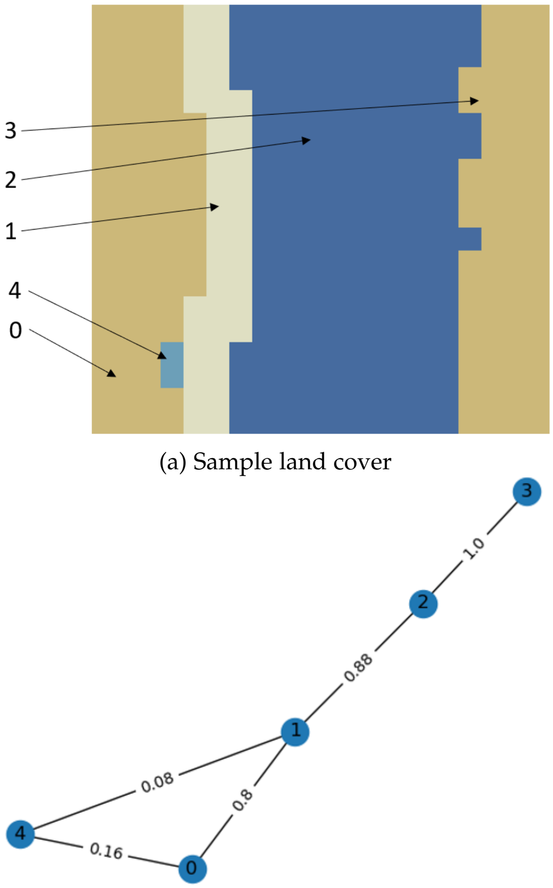
An autoencoder is a model that is able to take data and convert it to a different representation of the data. Most of the time this new representation has reduced dimensionality or is able to represent the original data in a new format that is more conducive to other data mining tasks. The strength of an autoencoder is judged on its ability to recreate the input from the reduced form. A graph autoencoder is able to convert a vertex to a vector representation.

5.2 GraRep [4]

GraRep creates a vector of step representations for each vertex on a graph. The kth value in the vector encodes information about graph k steps from the vertex.

Method	Complexity	Structure Captured	Training Method
LINE	$O(d\|E\|)$	Local	Stochastic Gradient Descent
LGAE	$O(d\|V\|)$		
GraRep	$O(\|V\|\|E\| + d\|V\|2)$	Global	Eigen Decomposition

TABLE 1: Comparison of basic attributes of the selected network representation methods



(a) Sample land cover

(b) Resulting Graph

Fig. 1: This figure demonstrates how a section of land cover is converted into a graph. Each of the labeled sections of the image correspond to the node with the same label. The weights on the graph are the length of the shared perimeter relative to the length of the largest perimeter observed in the image which in this case is the one between 2 and 3.

5.3 LINE [5]

LINE creates vector representations for each vertex but unlike GraRep, the vector only represents a single k steps from the vertex.

6 Proposed Research

Further research needed to complete the project How to handle very large and small regions? Should node label data be introduced?

6.1 Plan

- Finish implementation of GraRep - 09 March 2022
  - Understand remaining parts mathematically of algorithm
  - Create toy model to understand GraRep in a more overall sense
- Perform runs with land cover classes combined into super classes 10 March 2022
- Figure out how to handle very small regions - 12 March 2022
  - What is the threshold for a small region?
  - Create method to determine what surrounding region to merge it into
- Figure out how to handle very large regions - 14 March 2022
  - What is the threshold for a large region?
  - What to do about disjoint graphs?
  - Should they be left out of the final clustering?
- Figure out if node labels should be used - 16 March 2022
  - What would benefits look like?
  - How should they bee introduced into the pipeline?
  - How should the ”strength” of the node labels be in relation to the components found if labels are introduced after the components been found?
- Finish implementation of Linear Graph Autoencoders - 09 March 2022
  - Understand the parts of the algorithm mathematically
  - Create toy model to understand it in a more overall sense
- Finish implementation of LINE - 09 March 2022
  - Understand the parts of the algorithm mathematically
  - Create toy model to understand it in a more overall sense
- Begin running tests on entire pipeline - 09 March 2022

7 Evaluation

The resulting trees for each embedding algorithm can be split such that a desired number of regions are formed. Due to this fact as well as the methods being unsupervised means the resulting regions do not have any human defined meaning nor can the clusters be quantitative compared a ground truth. Therefore, the assessment will be based solely on the usefulness of the clusters. This will be done by examining the patterns found in each cluster qualitatively. The resulting clusters structure will also be compared to regions created by human conducted surveys.

7.1 Results

Example Result of Pipeline

- Input image: 30x30 km region near New Orleans, LA
- Number of segments: 8096
- Network Representation Method used: GraRep
- GraRep Arguments:
  - k=5
  - d=5
- Clustering Method: Hierarchical

Results can be viewed in Fig. 2

8 Conclusion

References

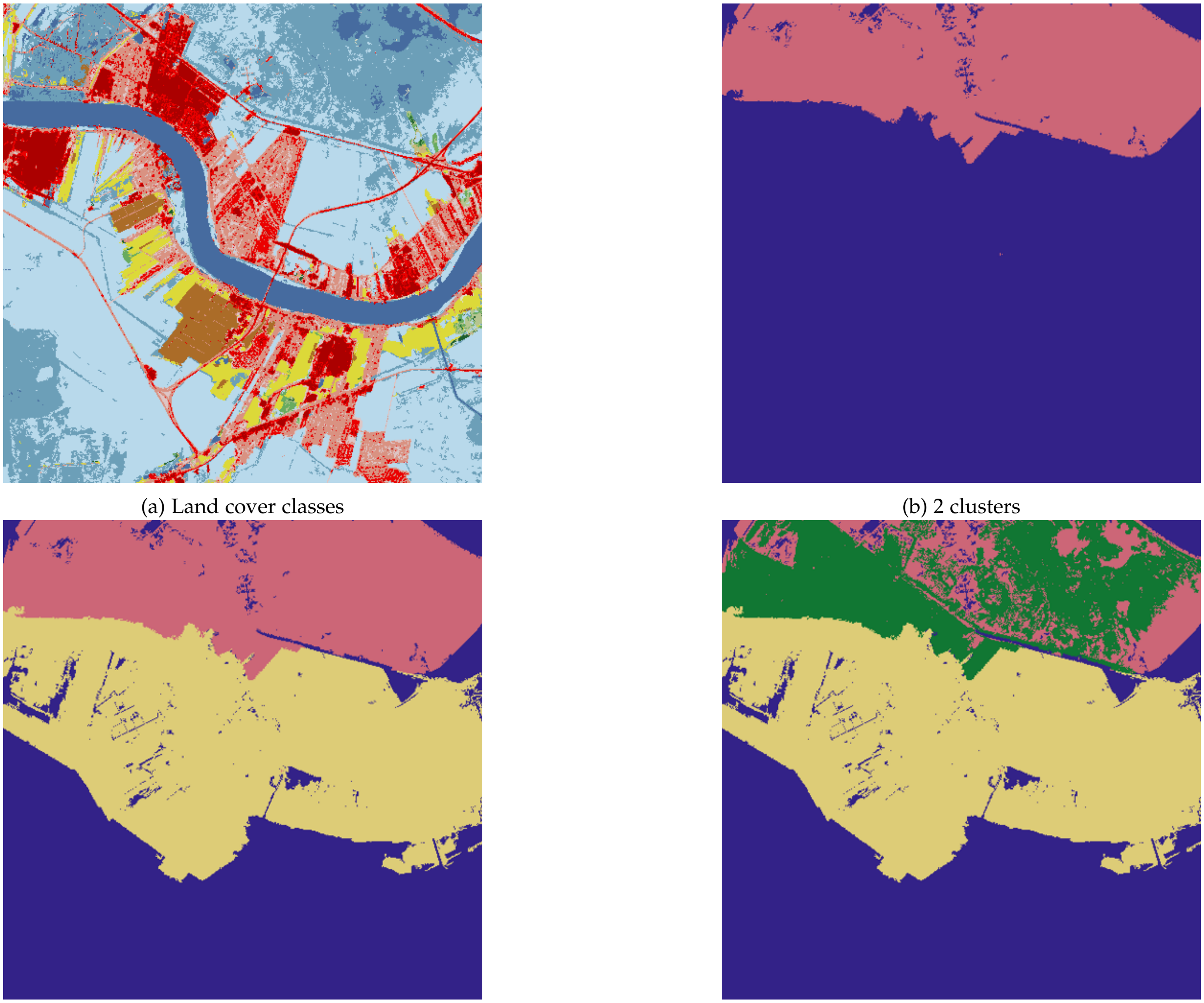
[1] J. Niesterowicz, T. Stepinski, and J. Jasiewicz, "Unsupervised regionalization of the united states into landscape pattern types," *International Journal of Geographical Information Science*, vol. 30, no. 7, pp. 1450–1468, 2016. [Online]. Available: https://doi.org/10.1080/13658816.2015.1134796

[2] (2019) Multi-Resolution Land Characteristics (MRLC) Consortium nlcd 2019 land cover (conus). [Online]. Available: https://www.mrlc.gov/data/nlcd-2019-land-cover-conus

[3] G. Salha-Galvan, R. Hennequin, and M. Vazirgiannis, "Keep it simple: Graph autoencoders without graph convolutional networks," *ArXiv*, vol. abs/1910.00942, 2019.

[4] S. Cao, W. Lu, and Q. Xu, "Grarep: Learning graph representations with global structural information," *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015.

[5] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding," *Proceedings of the 24th International Conference on World Wide Web*, 2015.



(a) Land cover classes

(b) 2 clusters

(c) 3 clusters

(d) 4 clusters

Fig. 2: Pipeline results