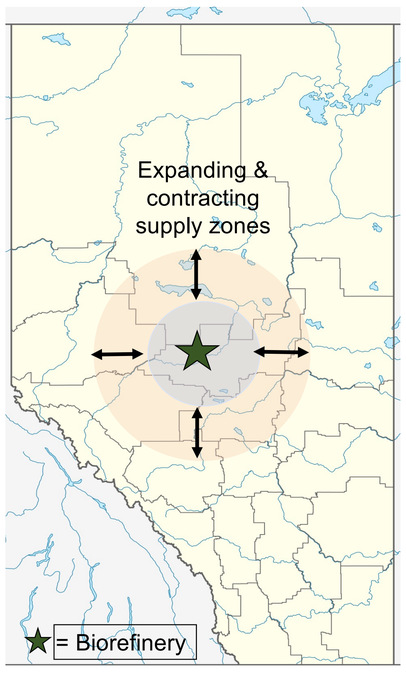
Modeling Variability in Biomass Feedstock Supplies with Limited High-Dimensional Data: An Application of Hierarchical Data Clustering

Amy Xu, Grant Hauer, Marty Luckert[1], Feng Qiu[2]

Large scale production of bioenergy is impacted by the variability of biomass supply

- Concerns about the environmental impact of fossil fuels have led to increased interest in bioenergy production using agricultural crops and residues. But using agricultural crops and residues is economically risky for investors.



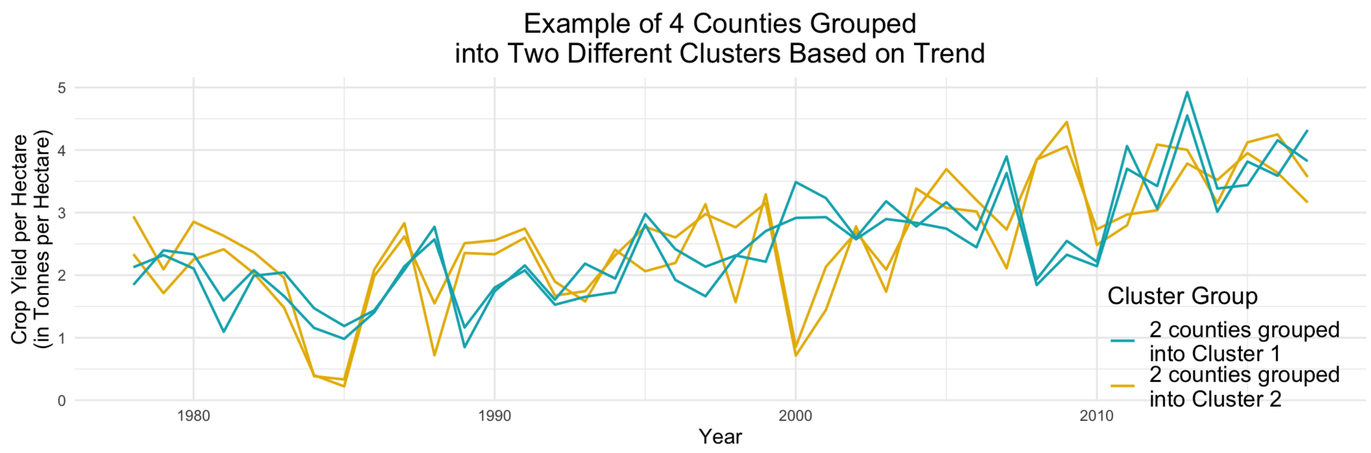
- Since, crop yields vary from location to location and from year to year it is important to accurately model crop yield variability.
- E.g., supplying a biorefinery with a fixed capacity requires an expanding and contracting supply zone over time.

Accurate modeling of variability is impacted by limited data

- We analyze crop yield variability across the 69 counties in Alberta over 40 years.
- To model variability over location and time across counties, the typical approach is to estimate the sample correlation using traditional statistical approaches.
- However, having only 40 years of data for 69 counties leads to a dimensionality problem that makes the sample correlation invalid when using traditional statistical approaches.
- To overcome the problem of limited data, we use data clustering algorithms.

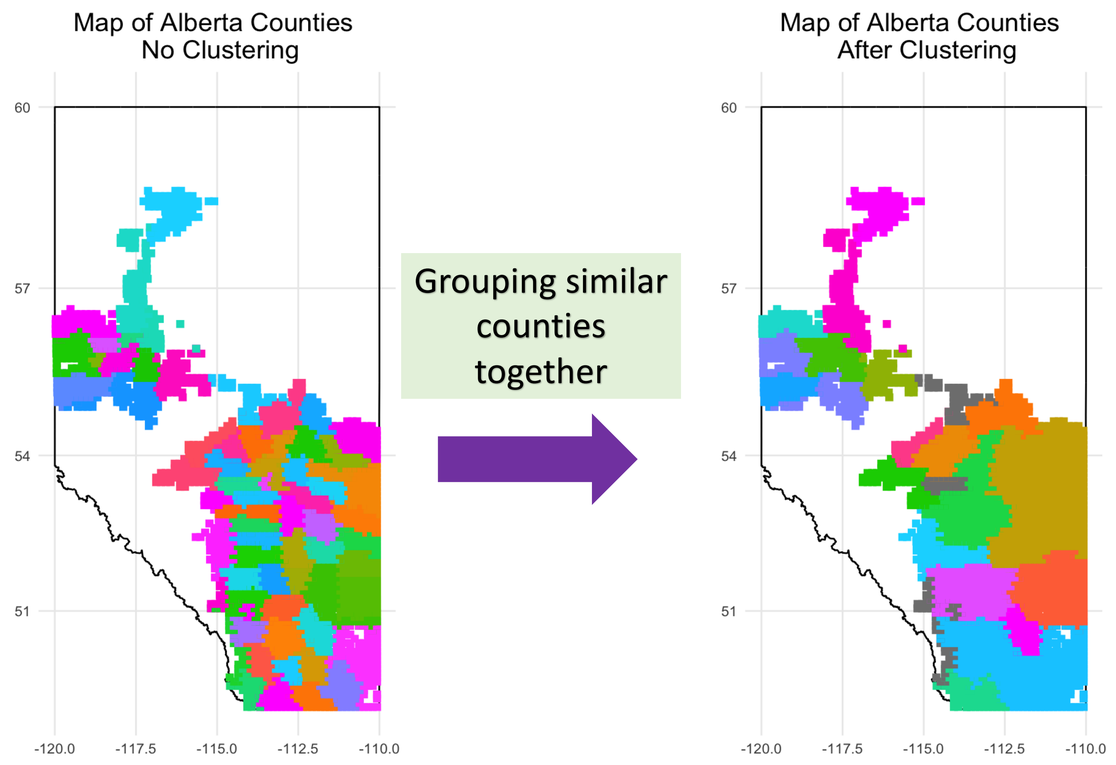
Data clustering groups similar counties

- We group together counties that have similar trends in crop yields over time as measured by hierarchical clustering approaches (see example below).



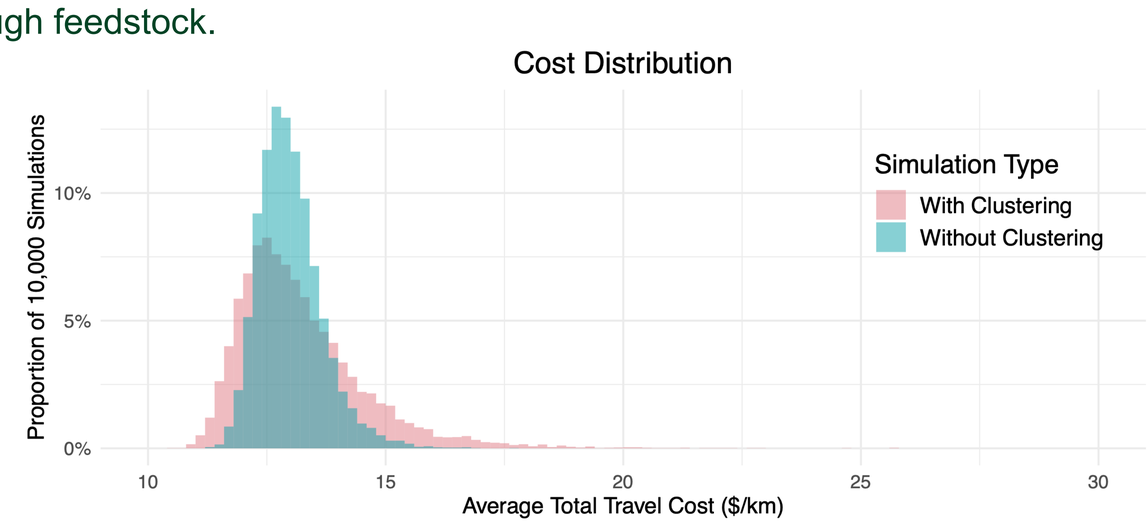
Clustering allows us to reduce the number of variables and retain differences across dissimilar counties

- By clustering similar counties, we can reduce the number of variables by creating a smaller number of aggregated counties.
- The map shows that counties that are grouped together are often close in proximity. This makes sense intuitively as counties that are near one another would share similar weather and soil conditions.
- We used two distance measurements for the hierarchical clustering:
  1. Average linkage (merges based on the average distance)
  2. Complete linkage (merges based on the maximum distance)
- The map on the far right represents the clustered counties obtained after hierarchical clustering with average linkage.
- After clustering, we re-estimated the crop yield variability.



There is more cost variability when counties are clustered together

- Once we estimate the crop yield variability using clustering, we can apply it to improve analysis of economic factors like average total travel cost.
- An example is shown (see Figure) where we assume we want to supply a biorefinery with a fixed capacity of 1 million tonnes per year, travelling farther and farther from the plant location until we obtain enough feedstock.
- We see from the distribution curve of the simulations that the range for average total travel cost is much wider.
- KEY POINT: Cost analysis without clustering can result in underestimation of the variability in average total travel costs for supplying a biorefinery.
- IMPLICATION: Clustering based on the increase in cost variation could lead to different evaluations of economic risk. Those assessing the investment and risk reducing strategies related to bioenergy production will need to take this into account.



[1] Principal Investigator. Email: mluckert@ualberta.ca

[2] Department of Resource Economics and Environmental Sociology, 515 General Services Building, University of Alberta, Edmonton, AB, T6G 2H1