Deciphering Economic Clusters in Real-Time: Applying Machine Learning to Registre des Entreprises du Québec Data

Lucien Chaffa [1]

Thierry Warin [2]

Société Canadienne de Sciences Économiques

May 17, 2024

[1] Chercheur Post-Doc, CIRANO, Presenter

[2] HEC Montréal, CIRANO, and Digital Data and Design (D^3) Institute, Harvard Business School

Introduction

Motivation

- Economic cluster is a way to investigate industry dynamics: their interdependencies and impacts on the economic landscape
- New data science techniques are now available to collect and analyze data

Question

- Can data science techniques help revisit economic cluster analysis?

Goal

- Define clusters and analyze their changing dynamics in near real-time

Literature review

- Agglomeration economies (Marshall (1890), Robinson (1956), etc.)
  - This work: Internal economies and external economies of agglomeration economies
- Economic cluster
  - Comparable cluster definitions using multi-regions and many industries (Porter (2003), Delgado et al. (2016), etc.)
  - Region-specific cluster definitions: qualitative, case-studies (Feser et al.(2009), (Porter and Ramirez-Vallejo (2013), etc.)
  - This work: New measure of inter-industry linkages applicable in both approaches

Contributions

- We propose a new quantitative cluster definition: measuring inter-industry linkages using **Growth Trajectory**
  - Applicable both for region-specific and multi-region cluster definitions
  - Data-driven cluster definition
- Near real-time insight into changing inter-industry dynamics within a cluster

Findings

- Clustering using growth trajectory groups diverse industries within the same cluster
  - Hidden inter-industry linkages are captured
  - Beyond labor market pooling, specialized suppliers, knowledge spillovers
- Identifies the direction of effects on industries within specific clusters
  - Business cycles
  - Economic policies

Methodology Approach: Step 1

- Step 1: Application of unsupervised machine learning algorithms (Hierarchical, K-means, K-medoids)
  - Inputs: S data points with p features ($S \in \mathbb{R}^p$); Goal: $S = \bigcup_{k=1}^{K} C_k$ where ($K \in \mathbb{R}$)

Methodology Approach: Step 1

- Step 1: Application of unsupervised machine learning algorithms (Hierarchical, K-means, K-medoids)
  - Inputs: S data points with p features ($S \in \mathbb{R}^p$); Goal: $S = \bigcup_{k=1}^{K} C_k$ where ($K \in \mathbb{R}$)
  - After all $\sum_{k=1}^{K} W(C_k)$ is the minimum, where $W(C_k) = \frac{1}{card(C_k)} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$, using *squared Euclidean distance*, is the within-cluster variation

Methodology Approach: Step 1

- Step 1: Application of unsupervised machine learning algorithms (Hierarchical, K-means, K-medoids)
  - Inputs: S data points with p features ($S \in \mathbb{R}^p$); Goal: $S = \bigcup_{k=1}^{K} C_k$ where ($K \in \mathbb{R}$)
  - After all $\sum_{k=1}^{K} W(C_k)$ is the minimum, where $W(C_k) = \frac{1}{card(C_k)} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$, using *squared Euclidean distance*, is the within-cluster variation
    - K-means: Choose K, start by random assignment of observations, then iterate assignation to the closest centroid (average point of C_k) till convergence

Methodology Approach: Step 1

- Step 1: Application of unsupervised machine learning algorithms (Hierarchical, K-means, K-medoids)
  - Inputs: S data points with p features ($S \in \mathbb{R}^p$); Goal: $S = \bigcup_{k=1}^{K} C_k$ where ($K \in \mathbb{R}$)
  - After all $\sum_{k=1}^{K} W(C_k)$ is the minimum, where $W(C_k) = \frac{1}{card(C_k)} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$, using *squared Euclidean distance*, is the within-cluster variation
    - K-means: Choose K, start by random assignment of observations, then iterate assignation to the closest centroid (average point of C_k) till convergence
    - K-medoids: Choose K, start by random assignment of observations, then iterate assignation to the closest centroid (median point of C_j) till convergence

Methodology Approach: Step 1

- Step 1: Application of unsupervised machine learning algorithms (Hierarchical, K-means, K-medoids)
  - Inputs: S data points with p features ($S \in \mathbb{R}^p$); Goal: $S = \bigcup_{k=1}^{K} C_k$ where ($K \in \mathbb{R}$)
  - After all $\sum_{k=1}^{K} W(C_k)$ is the minimum, where $W(C_k) = \frac{1}{card(C_k)} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$, using *squared Euclidean distance*, is the within-cluster variation
    - K-means: Choose K, start by random assignment of observations, then iterate assignation to the closest centroid (average point of C_k) till convergence
    - K-medoids: Choose K, start by random assignment of observations, then iterate assignation to the closest centroid (median point of C_j) till convergence
    - Hierarchical: Start with S groups and merge groups till one group.

Methodology Approach: Step 2-4

- Step 2: Identification of sub-clusters
  - Use the same cluster algorithms or network analysis clustering on the clusters identified in Step 1
- Step 3: Geo-spatial mapping of clusters
  - Set a threshold of employment distribution across MRCs
  - Identify clusters distribution
- Step 4: Analysis of temporal dynamics of clusters
  - Insights into cluster life cycles
  - Insights into cluster composition

Data and Summary Statistics I

- REQ data: *(Warin, T. (2021). "Req: Client for accessing Quebec company registrar. v0.1.0")*
  - The dataset is updated every two weeks
  - Dataset description

Number of observations	Number of variables	Number of industries
2.36 millions	61	1045

  - Variables selected: Registration date, cessation date, industry code (4 digits CAE), number of employees (range), Latitude, Longitude
  - Dataset transformation into time series
  - Population over sample

Data and Summary Statistics II

- Evolution of the number of firms in selected industries from 1990 to 2022

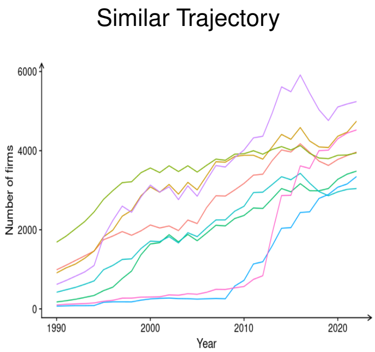

Figure 1: Evolution of Industry Size in Quebec

Data and Summary Statistics III

- Correlation between industries' growth rate

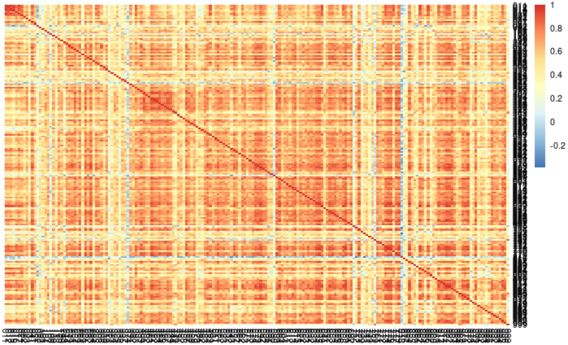

Figure 2: Heatmap of the correlation between industries

Preliminary Results: Industry selection

- Selection of industries with correlation coefficient ≥ 0.95

Preliminary Results: Industry selection

- Selection of industries with correlation coefficient ≥ 0.95
- Network graph

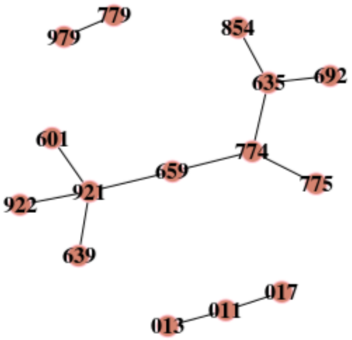

Figure 3: Network graph of highly correlated industries

Preliminary Results: Industry selection

- Selection of industries with correlation coefficient ≥ 0.95

- Network graph

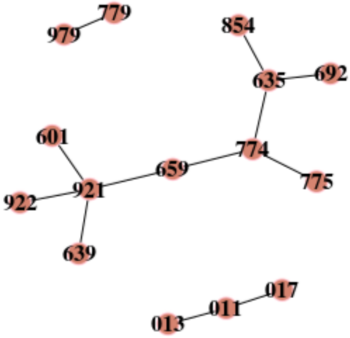

Figure 3: Network graph of highly correlated industries

- Minimum spanning tree representation

- Industry Description:

CAE	Description
011	Livestock and livestock farming poultry
013	Field crops
017	Farms Field crops and horticultural production
601	Food Stores
639	Other Types of Motor Vehicle Retail
921	Restoration
922	Taverns, bars and nightclubs
635	Workshops Motor Vehicle Repair
692	Direct Selling Companies
854	Teaching Personal & Popular Training
659	Other types of trade detail
774	Advertising Services
775	Offices architects and engineers and other services
779	Other business services
979	Other Personal Services & Domestic

Preliminary Results: Industry selection

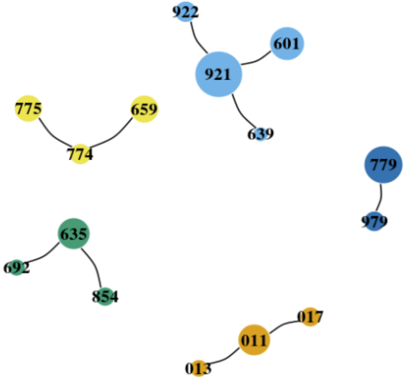

Figure 4: Network graph of sub-clusters of highly correlated industries

Preliminary Results: Industry selection

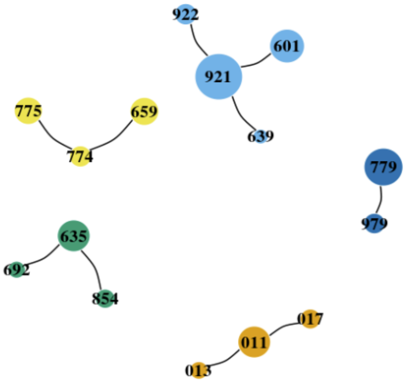

Figure 4: Network graph of sub-clusters of highly correlated industries

- Minimum spanning tree representation
- Clustering using hierarchical approach and modularity measure
- Each node size is proportional to the average of the industry's size

Preliminary Results: Geo-spatial mapping

- Geographic distribution of the clusters

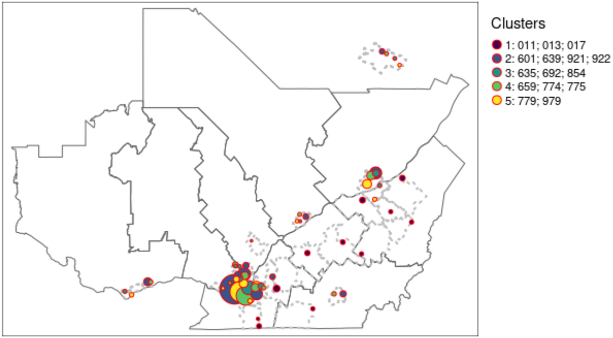

Figure 5: Geographic mapping of the clusters

Preliminary Results: Geo-spatial mapping

- Geographic distribution of the clusters

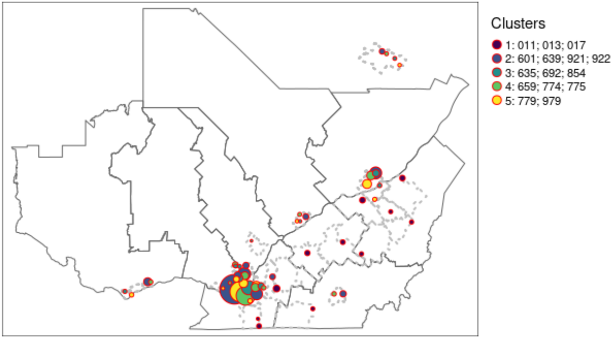

Figure 5: Geographic mapping of the clusters

- Criterion of mapping: Employment distribution across MRC (90th percentile)
- Each circle is proportional to the average size of the cluster
- Agriculture clusters are isolated
- Most clusters are concentrated in and around Montreal

Preliminary Results: Clusters summary statistics

- The size of a cluster is the accumulated number of f its industries

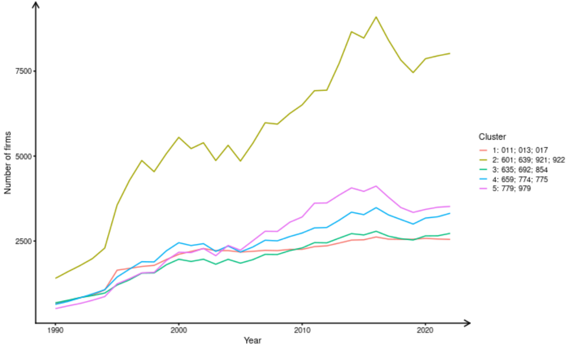

Figure 6: Evolution of the size of clusters

- The evolution of the cluster of related industries shows their perimeter needs to be redefined

Take away

- Inductive approach that considers the changing dynamics of industries
- New inter-industry linkages measure that captures wider agglomeration economies and extends the perimeter of the cluster definition
- Economic cluster definitions based on our measure of inter-industry linkages are applicable in region-specific and multi-region contexts

THANK YOU !

Learn more about our work at CIRANO's Pole on

Data Science for Trade and Intermodal Transportation