A Pre-launch Startup Guide for Cluster Mapping

An appendix to
Cluster Mapping as a Tool for Development

Richard Bryden
Director of Information Products
Institute for Strategy and Competitiveness
Harvard Business School

June 16, 2017

A Pre-launch Startup Guide for Cluster Mapping

The objective of this startup guide is to concisely outline the (1) team resources required and the (2) key steps for proceeding on a cluster mapping project. A sketch for a (3) project timeline is also presented.

(1) Assembling the cluster mapping team

1.1. Core project team

A cluster mapping project is an exercise in data collection, analysis and communication. It requires a team combining both technical skills as well as a capacity for understanding and interpreting industrial statistics and overall economic activity in a country. These technical and general economic skills may be available in the same person but are more likely to be assembled across a team.

The core project team

- Project Manager
- Data analyst
- Economist / consultant
- Web publishing pro (optional)

The skills of a modern **data analyst** include the ability to collect, manipulate and analyze data in a consistent and reproducable manner. In practice this requires some skills in a statistical software package such as R, Stata, SAS, or similar. While the ability to work with data in a spreadsheet program is assumed for the data analyst, a more programatic approach to working with data will be essential to extracting and manipulating large datasets in a consistent and reproduceable manner.

The label **"economist"** is used here to stand in for the team member(s) with the experience to identify and interpret detailed industrial and general economic statistics. An economist or consultant with domain knowledge in insustrial statistics, and economic development topics more broadly, will be the natural choice for a team lead on the project, directing and reviewing the work of the technical staff. Alternatively, or in addition, a **project manager** will direct and coordinate work on the project.

In many cases making the cluster mapping data set and related analyses publicly available via a web portal, not just as a report, will be within the project scope. In these cases, some additional skills in **web publishing** may be added to the team. The terminolgy here is left very generallly as "web publishing"

because there are an increasing range of tools available for moving data onto the web that may or may not require the skills of a full web development team. Some further discsussion follows below in the discussion of “Platform for Access.”

1.2. Steering group

Convening a steering group of key project stakeholders will benefit the project in numerous ways:

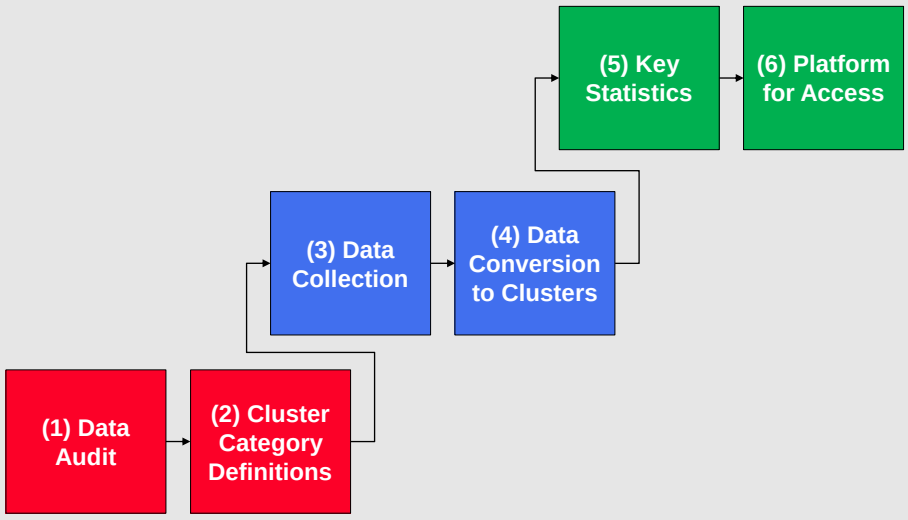
- Providing input to overall project objectives, ensuring maximum relevance to stakeholders.
- Identifying resources, financial and personnel, for the project.
- Gaining visbility for the project outputs in key constituencies.
- Helping to ensure the longevity of the project as an ongoing and updated resource rather than a one-time effort, ideally embeding the project in the portfolio of a governmental statistical agency, in a researh institute, or perhaps a new cluster- or competitivenss-focused entity.

The steering group should draw members from **official statistical offices**, **government agencies**, **NGOs**, **the business community**, **academic institutions**, or **think tanks**. Ideally, the the chair(s) of the steering group will have links to top government leaders to enhance the chances that the project will have impact in policymaking.

(2) The key steps in a cluster mapping project

A cluster mapping project will proceed through three broad phases:

- **Data audit and cluster model adaptation:** (1) Identify and audit data suitable for cluster mapping; (2) Develop cluster category definitions aligned to the locally available data.
- **Preparation of the core cluster mapping database:** (3) Extract, Transform and Load the source data sets; (4) Merge the data to the cluster category definitions and make data available in a format to support exploration and analysis.
- **Presentation and publication of key statistics and analyses:** (5) Design templates for key analyses for clusters and regions; (6) Provide broad access to data and analyses.



(1) Data Audit

Key actions required

- Investigate sources of region-industry data for the core cluster mapping exercise. Document key characteristics and any limitations of the data.
- As a complement to the cluster data, identify data sources for indicators that are region-, industry-, or economy-wide.

Deliverables

- A descriptive inventory of all potential data sources with an emphasis on the characteristics of the region-industry data.
- A complete reference for the local industrial classification system that corresponds to the exact level of detail found in the reported region-industry data, suitable as a starting point for the development of the localized cluster category definitions.

2.1. Data Audit

The objective of the data audit is to identify sources of data that will allow measurement of the scale of industrial activity in the sub-regions of a country. The availability of the data should be benchmarked against these dimensions:

- **Comprehensiveness of industries and activities covered**
    - ***Are all industries covered?*** Or, are some industries systematically or sporadically excluded? Most commonly the data required for cluster mapping will have been collected by national or regional governments, often as a by-product of building an exhaustive registry of businesses for purposes of taxation. Some limitations may derive from the idiosyncrasies of the collection system. For example, some core industry statistics in the U.S. are associated with the collection of social security payroll taxes and will exclude government employees, some public school system employees, railroad

workers, and domestic household workers not subject to the standard system will be omitted from the data.

- ***Are all employment types covered?*** Or, are there potentially multiple sources of data corresponding to different categories of workers? Again drawing on an example familiar to the authors, the main industrial statistics in the U.S. are drawn from data associated with payroll taxes for employees of firms. Separate statistics at some lesser degree of detail exist for self-employed workers and another for agricultural field workers. In some applications the main statistics are sufficient for the objectives; in others the three sources are melded for the most comprehensive coverage.
- ***Is there a large informal economy beyond the scope of any of the available statistics?*** If so, are there measures available to benchmark the extent of non-coverage?
- ***Is the data collected by comprehensive census or by survey?*** If the data is collected by survey, identify the methodology and requirements for extrapolating the sample to the population.

- **Industry classification system and granularity of classification system in use**
  - Identify the industry classification system in use and collect descriptive documentation.
  - Are concordances available to other known industry classifications systems?
  - What is the actual level of detail from the classification system found in the reported statistics? The industry classification system may include thousands of potential codes. An important benchmark here is a count of the actual number of codes in use in the reported data at the maximum level of detail, i.e. do not double-count codes at a more rolled-up detail from the most specific. In our experience, a granularity of 600 to 1,000 or more codes at this level of maximum detail is an indicator of excellent data for cluster mapping. But, projects can still be of high value at granularity below this range.
  - Does the industrial classification system change over the time covered by the statistics?
- **Regional granularity**
  - In which sub-national region types are the data available?
  - Do the characteristics of the data available change for different region types?
- **Data elements available**
  - Make an inventory of the data elements available in the source data. Employment is the most commonly available data element available as an indicator of cluster presence.

Additional elements that may be available include value-added production, revenues, payroll, number of establishments, etc.

  - Can the data be matched to company names?

- **Data suppression issues**
  - Laws governing the collection of data from private firms may include prohibition on releasing data that might be used to individually identify characteristics of individual firms. It is important to understand the impact of this issue. Are data simply omitted? Are range estimates provided in lieu of the suppressed data?
- **Frequency of publication and history available**
  - Are data published annually or at another frequency?
  - What is the lag between data collection and publication?
  - How many years of comparable historical data exist?
- **Data acquisition costs and permissions for redistribution**
  - What fees are associated with gaining access to the data?
  - Do the data come with restrictions on re-publication?

(2) Cluster Category Definitions

Key actions required

- Align the target industry classification to the benchmark cluster definition, creating an industry concordance through the NAICS if one does not already exist.
- Resolve exceptional cases through an understanding of the local economy and an observation in the data of the distribution and co-location of industries.

Deliverables

- The "cluster definition": a mapping of all relevant industry classification detail to the cluster categories and subcluster categories (as data availability allows.)
- Documentation of methodology.

2.2. Cluster Category Definitions

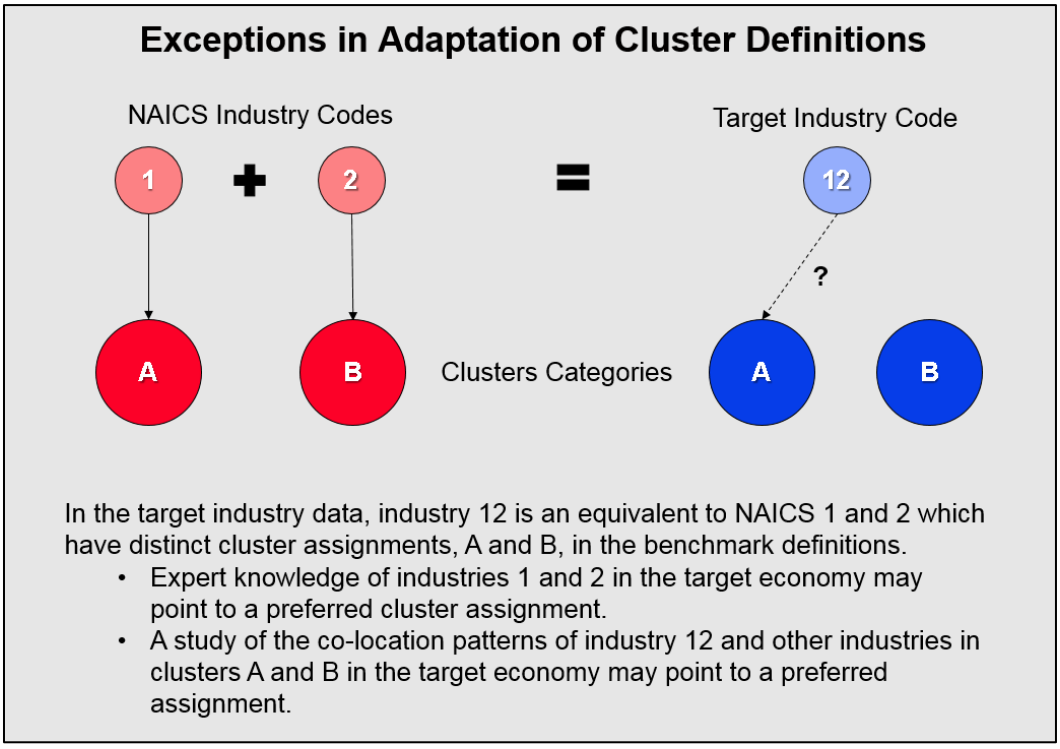
The recommended approach for developing the cluster category definitions proceeds as an adaption of the U.S. benchmark definitions (US Cluster Definitions.xlsx) to the target industrial classification system. For the original research underlying these benchmark cluster definitions see Delgado, Porter, and Stern, "Defining Clusters of Related Industries" (2016). It will be helpful to have a general understanding of the approach in this research on the question of traded vs. local industries and on the broad objectives of the clustering methodology. A detailed reading of the novel clustering methodology itself is less important given an approach based upon adaptation of the existing cluster model.

If the recommended approach is taken, the benchmark definitions need to be translated from their existing form using 5-digit NAICS industry codes into the industrial classification system found in the target data. The starting point to this work will be a listing of the industry classification system to the level implemented in the source data – a document that is expected as an output of the Data Audit step above.

In some cases a concordance may exist between the target industrial classification system and the NAICS. In others, the development of a concordance will be an early component of the project. In the vast majority of cases, industries will map easily from one classification system to another.

Exceptions will occur where industries not covered in the US benchmark definitions are found in the target system. For example, an exception may occur when an industry is uniquely significant to a local economy, for example tortilla manufacturing in Mexico. In this case an observation of the distribution of its employment across regions shows that tortilla manufacturing behaves very similarly to bakeries. Both industries exhibit the characteristics of a local industry, unique among manufactured food products which are otherwise observed in traded clusters. The industry is assigned to "Local Food and Beverage Processing and Distribution" alongside bakeries.

Exceptions will also occur when the local data reports industries at a more aggregated level than in the benchmark definitions *and* the disaggregated industries are found across multiple clusters in the benchmark definitions. Some simple resolutions occur when knowledge of the local economy reveals that one "side" of the aggregated code is essentially not present or minimally present in the local economy and the code can be treated only for the relevant portion. In other cases a data analysis exercise to observe the patterns of colocation of the aggregated industry versus the industries in relevant clusters will indicate the best fit in the spirit of the cluster model.



Adaptation of the benchmark cluster model to target data classification benefits from a collaboration between the data analyst and other project staff with an understanding of the regional economy and industries, ideally with further input or review from researchers familiar with the application of the cluster model in other regions of the world. One to two man-weeks of effort should be anticipated here for a team of two or three people.

The final output of this work should look much like the "NAICS" tab in the US Cluster Definitions spreadsheet linked above - with an assignment of all relevant industry codes to clusters (and subclusters as industry detail allows.) This target country definition should also take into account changes in the industry classifications over time as seen in the benchmark definition file.

(3) Data Collection

Key actions required

- Extract, transform and load all project data as identified in the audit to a consistently formatted and accessible project repository.

Deliverables

- An accessible project repository for all source data and metadata.
- Code and documentation to facilitate replication of the data repository and collection of future updates to the data.

2.3. Data collection

In the most common scenario good region-industry statistics are easily and publicly available from a national statistical office. But, collection and use of this data may be complicated by any of several scenarios that have been observed in the course of recent cluster mapping projects:

- Underlying payroll data or tax data may be collected and curated by the statistical office but not translated into a publicly accessible dataset. Project team members may need special

permissions to access a protected research data center and process the underlying data into a useable format also meeting standards of the research center for removal of data from the premises. Significantly greater time must be allotted to data collection under this scenario.

- If the best available data are derived from surveys rather than from a more complete census or from administrative registries, additional care must be taken in understanding the sampling methodology and best practice for extrapolating the data to the whole population.
- The data may come with restrictions on redistribution at some level of granularity. This is most likely to occur if the data are sourced from a private vendor or if the statistical office subcontracts some role in the collection and/or distribution of its data to a private vendor. But, it may also occur where a statistical office treats data dissemination as a revenue-raising activity. These restrictions are most likely to impact the step (6) "Platform for Access" phase of the project. The aggregation of data into clusters may help in this regard. While republication of the detailed industry data may be prohibited, sharing of data aggregated to clusters may be negotiated and allowed. Construction of a separate database for aggregated and otherwise derived data will be needed to support the Platform for Access or other data distribution.

Good general project methodology will include the careful collection and archiving of

- all source data in its original format,
- source metadata such as coding schemas and other documentation necessary to interpret the source data,
- local cluster definitions with documentation of methodology and exception handling,
- well-documented code and additional records of any other transformations made on the data not captured in code, and
- all intermediate and final datasets that would be needed to replicate the project.

Careful collection and archiving of data will continue throughout the life of the project.

(4) Data Conversion to Clusters

Key actions required

- Merge the region-industry data source(s) to the cluster category definition.

Deliverables

- A database of cluster data allowing easy access to data by region, cluster, year and target data element.
- Code documenting the database creation and allowing for replication and ease of future updates.

2.4. Data Conversion to Clusters

Once the data is acquired and organized and the cluster model adapted to local circumstances, the construction of a cluster database becomes a mechanical exercise easily accomplished given adequate data analysis skills. Several days may be allocated to this step to include careful quality assurance and review.

To support the production of key statistics and analyses, the organization of the cluster database should at a minimum allow for the easy selection of data via indices for:

- Level of regional detail: national and individual or multiple sub-national regions
- Leve of cluster detail: total regional, total traded or local, individual or multiple clusters, individual or multiple subclusters, individual or multiple industries
- Time period
- Data element, e.g. employment, value-added production, payroll, etc.

It may also be convenient and time-saving to calculate and store in the database secondary, derived data elements such as ranks on cluster employment across regions or share of national cluster employment within a region. A well-organized cluster database will ease the calculation of key statistics and the population of various analyses.

(5) Key Statistics

Key actions required

- Query project databases to create a set of key statistics and analyses for regions and clusters.
- Iteratively support requests from the broader economic research project team.

Deliverables

- A report of basic regional and cluster statistics.
- Interactive tools allowing economic research project team members to explore regional and cluster data through a set of standard templates.
- Ad hoc analyses supporting requests from the economic research project team.

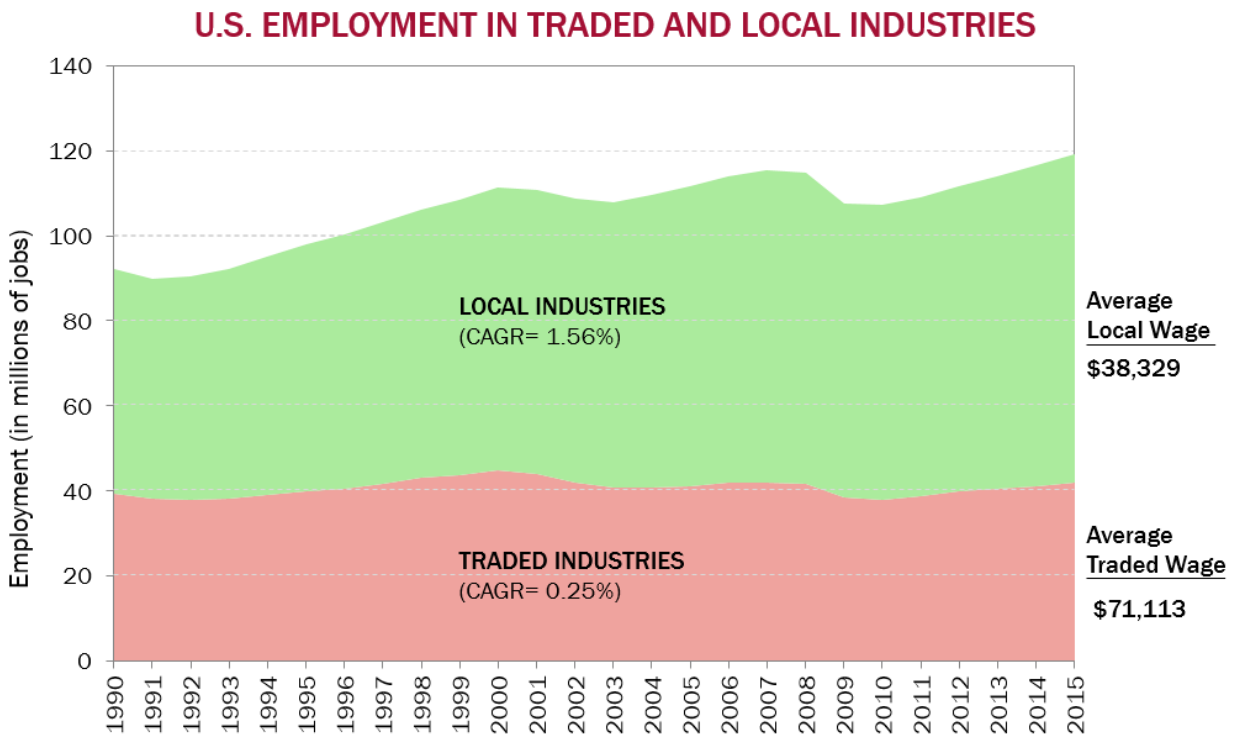
2.5. Key Statistics and Analyses

Often a cluster mapping project may be initiated as part of a broader initiative gathering country competitive diagnostics, or it may be a component of a regional competitiveness project or a cluster-focused research effort. These contexts will have implications for the selection and focus of the statistics, analyses and data visualizations drawn from the cluster data. But, for the purposes of illustration, three typical analyses are described here to demonstrate ranges of economic questions which may easily be addressed with data from the cluster mapping database.

2.5.1.Example: National Benchmarks

Question: What are the long-term trends of traded vs. local industries in the overall economy? Employment and payroll data in these categories are easily pulled from the cluster project database and summed over time to produce the following analysis.

Traded industries employment in the U.S. has been mostly stagnant over the period 1990-2015, hovering around 40 million employees and trending only slightly upward. Growth in overall employment has come primarily from local industries more sheltered from trade with other regions. Unfortunately, this decline in the share of employment in traded industries is associated with a shift to lower wages. Average traded industry wages are $71k versus $38k in local industries. These trends coincide with observations in the general economic data of declining real household income and increasing income inequality in the U.S.

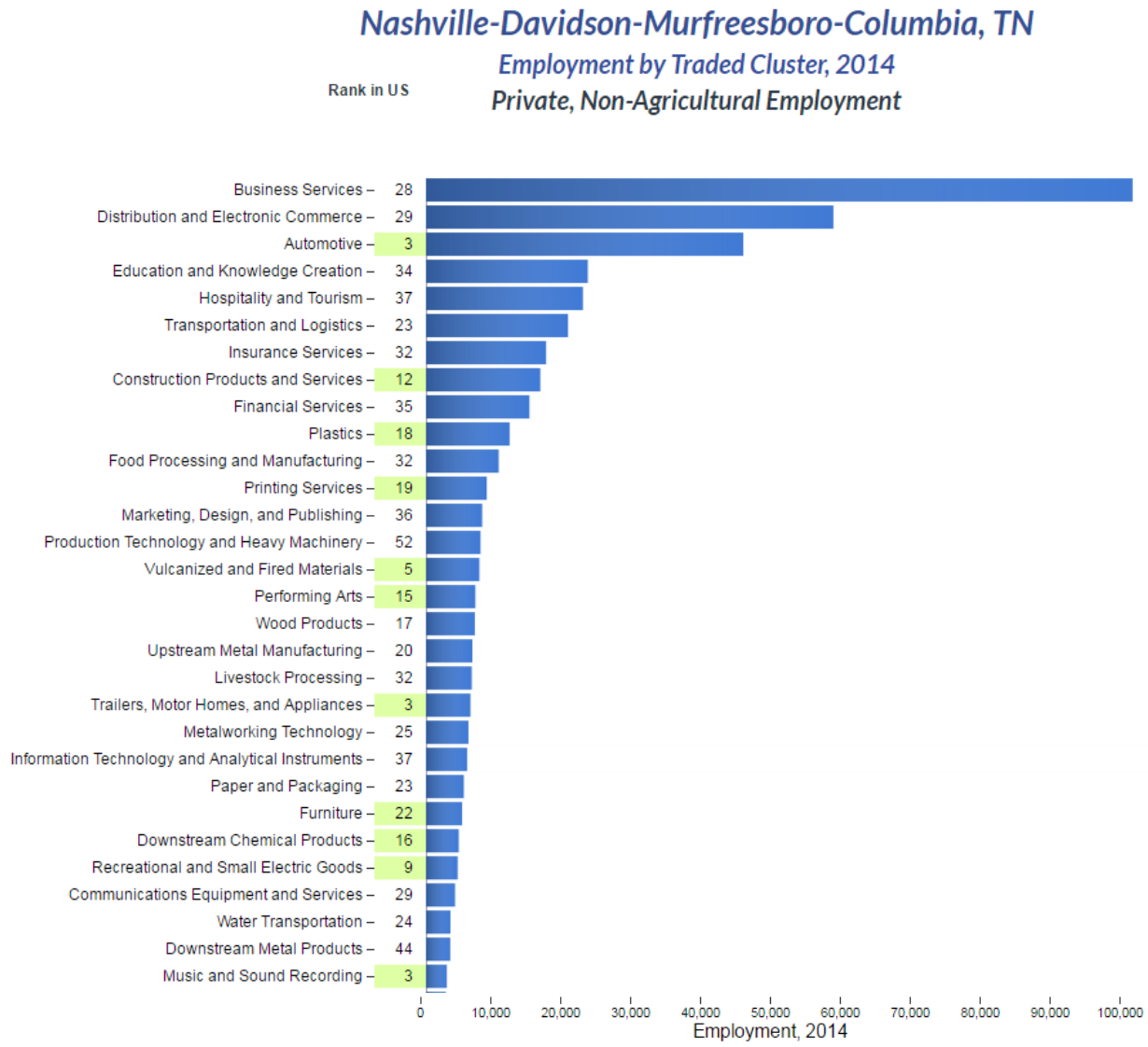


2.5.2.Example: Regional perspective

Question: What are the "strongest" clusters in a particular region? How have they performed recently relative to peers?

In Delgado, Porter, and Stern "Clusters, convergence, and economic performance" (2014), a measure for strong cluster presence in regions is found to have strong correlation with positive general economic outcomes in a region. The strong cluster measure sets a minimum threshold for agglomeration of employment and number of establishments in a region and then looks for clusters in the highest quartile of employment specialization across regions active in the cluster category. Both the cluster agglomeration and specialization measures are based upon distributions across all regions in the national dataset and are easily replicated from a complete cluster mapping dataset.

Here is a (partial) view of the Nashville Economic Area's employment in traded clusters with the green highlighting flagging strong clusters. The measure of a strong cluster is distinct from either simple employment levels or even the cross-regional rankings of cluster employment.

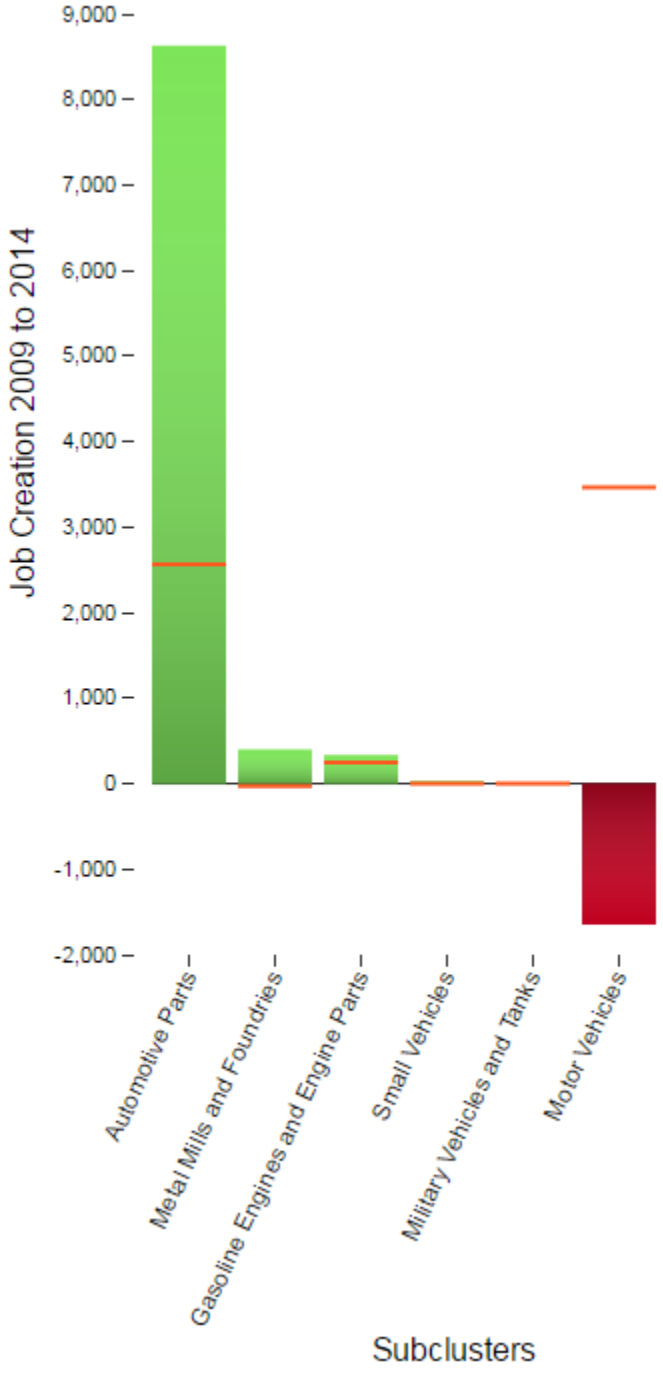


Here is a look at job growth in the post-recession period in one of Nashville's strong clusters. The Nashville Automotive Cluster has outperformed national benchmarks for job growth in the upstream subclusters for Automotive Parts and also Metal Mills and Foundries. Job growth in downstream subclusters for Gasoline Engines and Engine Parts and final Motor Vehicles assembly has been at or significantly below national benchmark rates.

Nashville-Davidson-Murfreesboro-Columbia, TN

Automotive Cluster

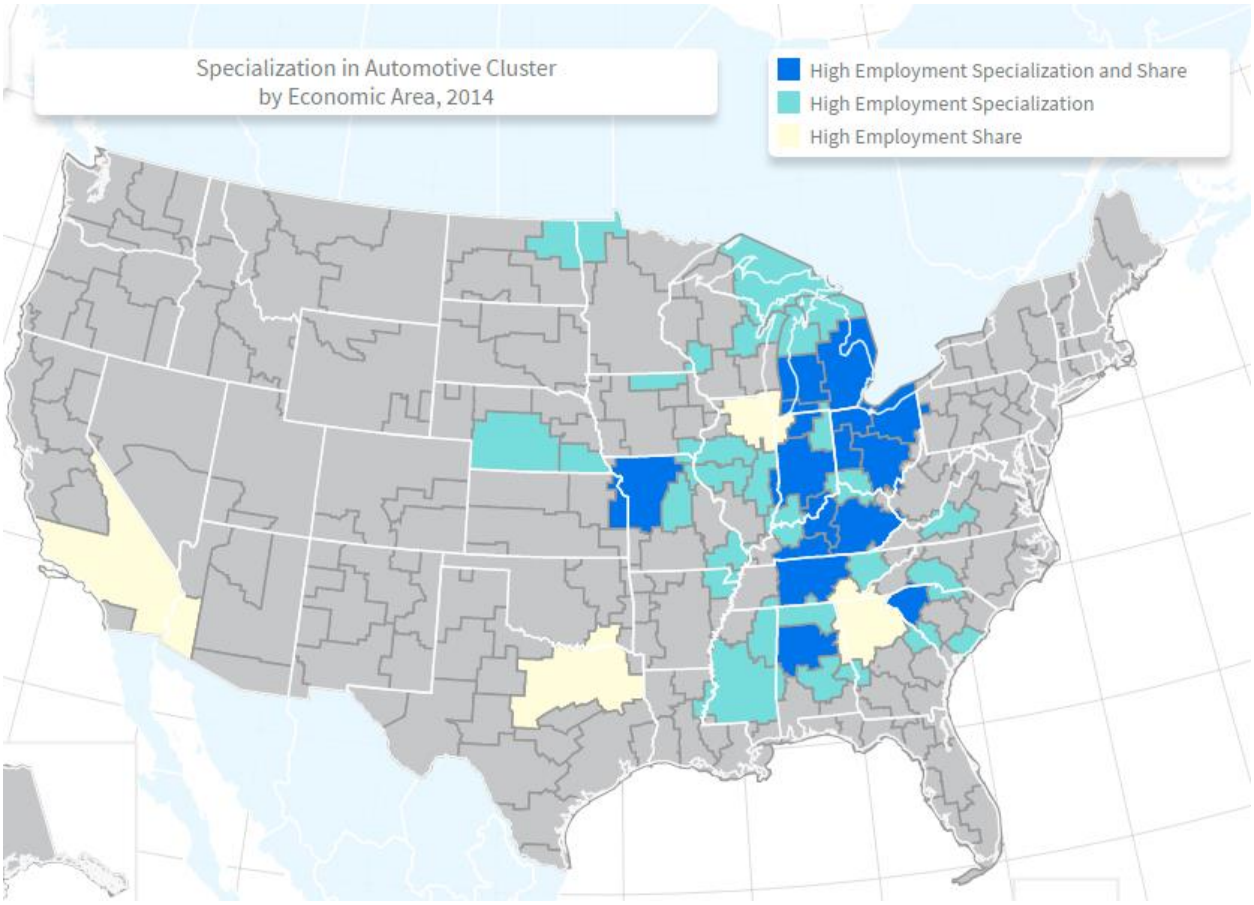
Job Creation by Subcluster, 2009-2014



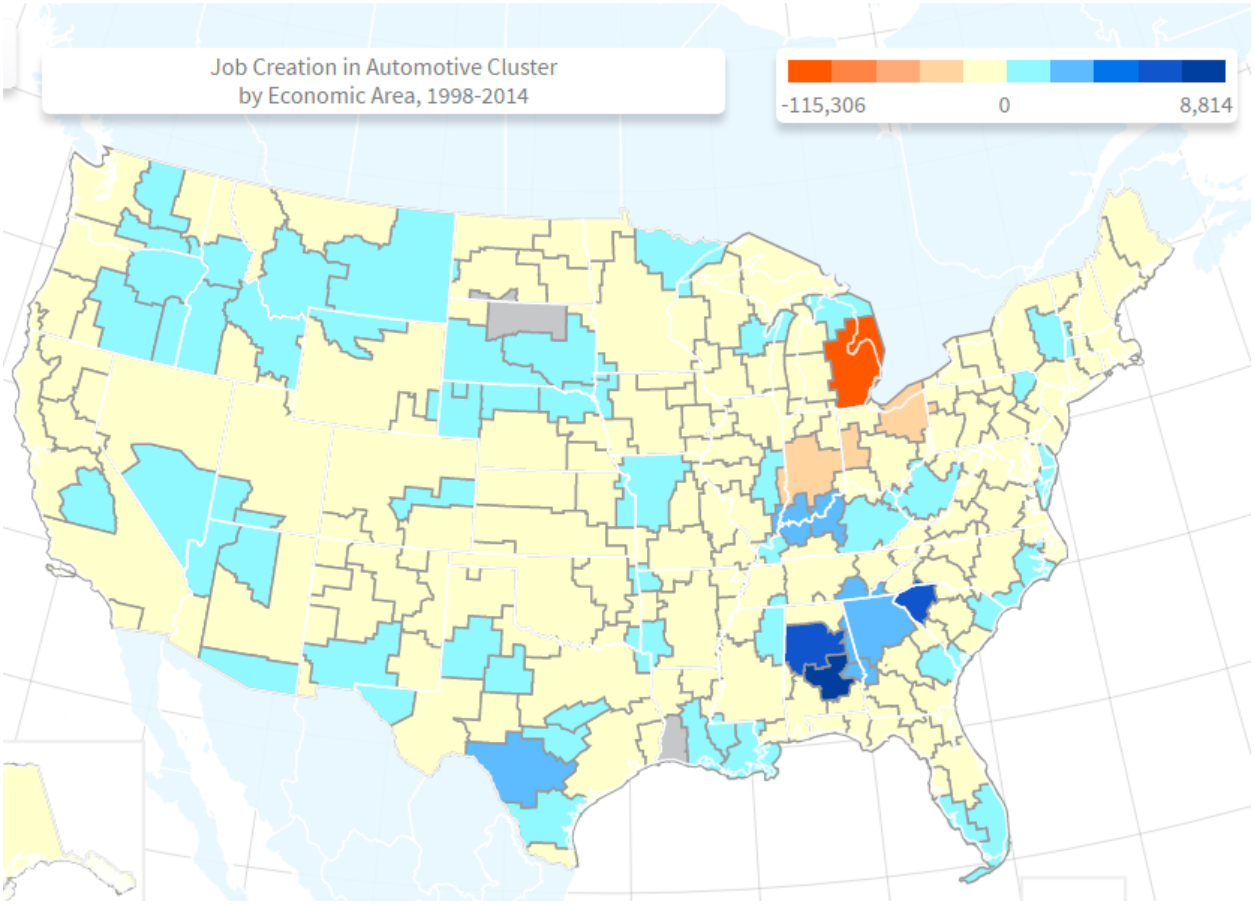
2.5.3.Example: Cluster perspective

Question: What is the geographic footprint of the Automotive cluster in the U.S. and how is it changing over time?

We define regions with high employment share based simply on high absolute employment levels in Automotive clusters. We define regions with high employment specialization as those having a high proportion of regional employment in Automotive relative to national averages. The dark blue here indicates both high share and high specialization. Light blue indicates high specialization only. And a few larger regions in yellow capture significant employment without meeting our cutoffs for high specialization.



Over the recent two decades, employment has declined in traditional Automotive clusters in Detroit and the Upper Midwest region of the U.S. Employment gains have been concentrated just south of these areas in Kentucky, Alabama, Georgia, South Caroline and in Texas to some extent.



These examples are intended to illustrate the key statistics such as employment, employment share, specialization, and wage levels that should be easily retrievable or calculable from the database, and the manner in which they may be assembled in more comprehensive analysis to address specific research questions. The companion document "Cluster Mapping as a Tool for Development" lists other typical analyses for regions and for clusters. The US Cluster Mapping portal, http://clustermapping.us/region/state/massachusetts/cluster-portfolio, also hosts a larger set of standard analyses that can serve as examples.

(6) Platform for Access

Optional

Key actions required

- Investigate and document requirements for public access to data and analyses.
- Develop or direct development of public access tools.

Deliverables

- Deliverables will vary depending upon determined requirements.
    - A more minimal set of deliverables might include only a set of Tableau Public dashboards for standard analyses that could be framed into an existing website.
    - Developing of a full website with custom data visualizations and complementary content would be a much large project.

2.6. Platform for Access (optional)

2.6.1. Scope

In many cases it is useful to make the cluster mapping data set itself publicly available, not just a report on key descriptive statistics or overall findings. Such 'open data' portals have been launched in a number of countries, both to provide critical information to policy makers at the regionals level and also to companies that are considering the attractiveness of different locations.

Such data portals can be created with different levels of ambition and functionalities. At the minimum, a website can be created to download the entire data set as well as key reports. With existing software packages it is now also fairly easy to create websites that include geographic maps to allow accessing data for specific regions, or to create other functions that allow users to choose specific clusters or indicators. Some existing portals have also added further functionality, for example a connected data set on regional economic performance and competitiveness, a registry of cluster-based organizations, and further tools for web-based collaboration across clusters.

The choice about which level of sophistication to aspire for with a cluster mapping portal depends on the policy objectives as well as the available resources. Making data available for the policy community and expert users can often be easily done. Designing a website that engages a broader community of practitioners, including many who are not familiar with the concept of clusters, is a more ambitious undertaking that requires different ways of communicating the content and a more developed set of tools to select and analyze the data.

2.6.2. Web publication tools

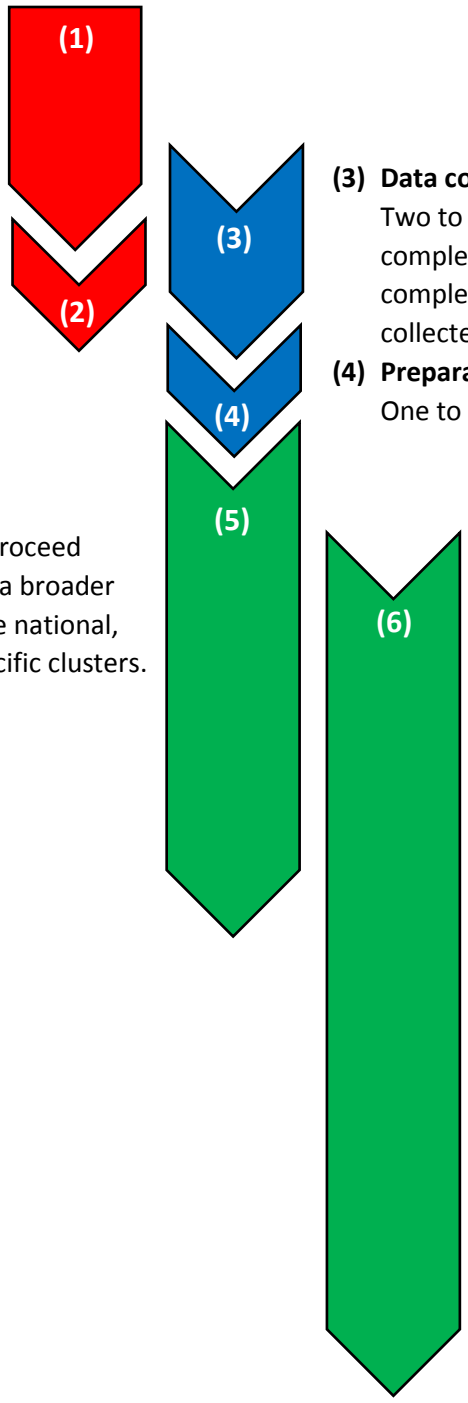
Advances in tools for web publication and data visualization, both open source and proprietary, present new opportunities for inexpensively yet robustly publishing richly interactive, data-driven websites. These advances lower the barriers to creating websites that might include a core set of functionality for charting clusters and regional portfolios and geographically visualizing cluster presence. These data-driven components may also be complemented by topical news, directories of cluster organizations, or links to published reports on clusters and economic development.

It is beyond the scope of the discussion here to fully describe options for creating websites with interactive data visualizations. But, to point to very different approaches available, please see the Github repository for the U.S. Cluster Mapping Project (https://github.com/clustermapping/cmp) for the full open-source code for this custom developed website. And, for a look at a general approach leveraging a proprietary but inexpensive platform, see Tableau Public at https://public.tableau.com. This tool allows a capable data analyst to create open-data web visualizations without the need of a web development professional's expertise.

(3) A project timeline typically run six to nine months.

(1) Data audit
Two to five weeks, dependent primarily on ease of access to region-industry source data and complexity of the source data.

(2) Cluster category definitions
One to two weeks.



(3) Data collection
Two to four weeks, dependent on accessibility and complexity of source data and the extent of the complementary datasets that will be processed and collected.

(4) Preparation of cluster mapping database
One to two weeks.

(5) Key statistics and analyses
Six to twelve weeks, likely to proceed iteratively in conjunction with a broader project to write a review of the national, regional economy, and/or specific clusters.

(6) Platform for access (*optional*)
Six to twenty-four weeks, quite variable depending upon design scope and technologies.