Journal of Economic Geography Advance Access published June 3, 2015

Journal of Economic Geography (2015) pp. 1–38 doi:10.1093/jeg/lbv017

Defining clusters of related industries

Mercedes Delgado[*,†], *Michael E. Porter*[**] *and Scott Stern*[***]

[*]Temple University and Institute for Strategy and Competitiveness
[**]Harvard University
[***]MIT Sloan and NBER
[†]Corresponding author: Mercedes Delgado, Temple University and Institute for Strategy and Competitiveness, 1801 Liacouras Walk, 542 Alter Hall, Philadelphia, PA 19122-6083.
email <mdelgado@temple.edu>

Abstract

Clusters are geographic concentrations of industries related by knowledge, skills, inputs, demand and/or other linkages. There is an increasing need for cluster-based data to support research, facilitate comparisons of clusters across regions and support policymakers in defining regional strategies. This article develops a novel clustering algorithm that systematically generates and assesses sets of cluster definitions (i.e., groups of closely related industries). We implement the algorithm using 2009 data for U.S. industries (six-digit NAICS), and propose a new set of benchmark cluster definitions that incorporates measures of inter-industry linkages based on co-location patterns, input–output links, and similarities in labor occupations. We also illustrate the algorithm's ability to compare alternative sets of cluster definitions by evaluating our new set against existing sets in the literature. We find that our proposed set outperforms other methods in capturing a wide range of inter-industry linkages, including the grouping of industries within the same three-digit NAICS.

Keywords: Cluster analysis, inter-industry linkages, industry clusters
JEL classifications: C38, L6, R1
Date submitted: 21 November 2013 **Date accepted:** 17 April 2015

1. Introduction

There is an increasing need for useful data tools to measure the cluster composition of regions and support regional policy development as well as business strategy. This article addresses this need by providing a methodology for generating and assessing sets of cluster definitions—groups of industries closely related by skill, technology, supply, demand and/or other linkages—that are regionally comparable (i.e., the industries that constitute a cluster are the same for all regions). We develop a novel approach for defining clusters in a way that accounts for multiple types of inter-industry linkages, and also allows us to compare alternative sets of cluster definitions through the calculation of scores for each set. We implement this clustering algorithm to create a new set of U.S. Benchmark Cluster Definitions (BCD), capturing a broad range of inter-industry linkages.

The agglomeration of related economic activity is a central feature of economic geography (Marshall, 1920; Porter, 1990; Krugman, 1991; Ellison and Glaeser, 1997). Marshall (1920) highlighted three distinct drivers of agglomeration: input–output (IO) linkages, labor market pooling and knowledge spillovers, which are associated with cost

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

© The Author (2015). Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

or productivity advantages to firms. Over time, an extensive literature has broadened the set of agglomeration drivers, including local demand conditions, specialized institutions, the organizational structure of regional business and social networks (Porter, 1990, 1998; Saxenian, 1994; Storper, 1995; Markusen, 1996; Sorenson and Audia, 2000; among others). Thus, clusters contain a mix of industries related by various linkages (knowledge, skills, inputs, demand and others).

The bulk of the cluster literature has been based on detailed case studies (Marshall, 1920; Porter, 1990, 1998; Swann, 1992; Saxenian, 1994). Over time, an emerging literature has sought to analyze clusters through larger-scale quantitative studies that span across many regions and industries (Porter, 2003; Feser, 2005). Using particular cluster definitions, studies have shown that the presence of related economic activity matters for regional and industry performance, including increased job creation, patenting, and new business formation (see among others, Feldman and Audretsch, 1999; Porter, 2003; Feser et al., 2008; Glaeser and Kerr, 2009; Delgado et al., 2010, 2014b; Neffke et al., 2011).[1] This evidence has informed key questions of both research and policy interest: the size of cluster effects, which mechanisms are most important in driving agglomeration and how clusters diversify and grow in a region.

Based on different definitions of clusters, covering different portions of the economy (ranging from high technological intensity industries to manufacturing to all industries defined in the industrial classification system), existing research has generated a range of results on these issues. However, the lack of a comprehensive methodology to define and compare alternative sets of cluster definitions makes it difficult to reconcile these findings. This article addresses this issue by developing a novel clustering algorithm to generate, assess and compare alternative sets of cluster definitions.

Our cluster definitions are groups of industries related by skill, technology, supply, demand and/or other linkages. Inter-industry linkages are identified through the co-location patterns of industries across regions, and with national measures of IO and labor occupational relatedness among industries. These linkages are used to group industries into a set of defined clusters. This procedure yields regionally comparable cluster definitions (i.e., the industries that constitute a cluster (e.g., Biopharmaceuticals) are the same for all regions, and so we can compare the specialization of any two regions in the cluster). Thus, these cluster definitions are not explicitly regional (in the sense of measuring region-specific linkages like in Hill and Brennan (2000) and Kerr and Kominers, 2010), but instead group industries in ways that capture geographically bounded inter-industry linkages.

To generate a set of cluster definitions, we use clustering analysis—numerical methods for the classification of similar objects into groups (Everitt et al., 2011; Grimmer and King, 2011). Our algorithm generates many different cluster configurations, *C*s, through a clustering function that utilizes a particular measure of the relatedness between any two industries and well-specified parameter choices (e.g., the number of groups). Each configuration is composed of mutually exclusive groups of related industries (i.e., clusters). The algorithm then provides scores that assess the ability of each *C* to capture meaningful inter-industry linkages within clusters. This allows us to identify the configuration, *C**, that best captures certain types of

[1] See Rosenthal and Strange (2004) and Cortright (2006) for a review of economies of agglomeration studies.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

inter-industry links. Because an algorithm cannot perfectly substitute for expert judgment, the methodology concludes with an expert assessment and adjustment of individual clusters in *C** to determine a final set of cluster definitions, *C***.

Our article contributes to the literature on clusters and economies of agglomeration in several ways. First, the clustering algorithm allows us to compare alternative sets of clusters using a common approach that generates consistent scores (i.e., most clustering methods do not provide scores to help compare across groupings (Everitt et al., 2011; Grimmer and King, 2011)). We can assess cluster configurations that are generated using different inter-industry linkage measures. For example, we can evaluate *C*s generated using pairwise industry co-location patterns to those based on IO or other measures. We can also compare existing sets of cluster definitions. The ability to score sets of clusters can help identify the appropriate sets for addressing particular research and policy questions.

Our scoring approach utilizes a basic clustering principle: creating groupings so that industries within a cluster are more related to each other than to industries in other clusters based on various measures of inter-industry linkages. The score for a given *C* depends on how well it captures various types of industry linkages. However, what constitutes a useful set of cluster definitions may change depending on the research context. Some studies may be interested in a particular type of industry link (e.g., labor occupational links), making sets of cluster definitions that perform better in that link more useful.

Second, our algorithm allows for experimentation with multiple inter-industry linkage measures used in the economies of agglomeration literature, including IO linkages, occupational linkages, the co-location patterns of industries, and combinations among them. We can then examine the cluster configurations generated based on these different measures. The methodology can incorporate additional inter-industry linkage measures as they become available (e.g., a measure that specifically captures knowledge linkages), and score their resulting cluster sets.

Third, although generating cluster definitions will require expert judgment for some individual clusters, our algorithm is transparent. In the last stage of the algorithm, there is room for expert judgment to correct for inevitable anomalies that arise due to data imperfections. For example, in a given *C**, there could be industry outliers that do not seem to belong to their assigned cluster. These can be reallocated to their 'next best' cluster using a score that assesses the relatedness of the particular industry with another cluster, using a transparent process. Users can assess how adjustments (re-allocations of industries, combining or dividing clusters) impact cluster scores.

Another important contribution of this article is that we implement our method to generate a new set of BCD for the USA (BCD or *C***), which captures a wide array of inter-industry linkages. The U.S.-based empirical analysis focuses on grouping 778 6-digit NAICS industries in manufacturing and services in 2009, and uses County Business Patterns (CBP), the Benchmark IO Account of the United States, and Occupational Employment Statistics (OES) datasets to define multiple industry relatedness measures.

The proposed BCD consists of 51 clusters generated by using inter-industry measures of co-location patterns of employment and the number of establishments, IO links and labor occupation links (OCC). We examine the relative performance of the BCD and three existing sets of (mutually exclusive) cluster definitions: industries within the same three-digit NAICS, Porter (2003), and Feser (2005). Grouping industries within the

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

same three-digit NAICS scores poorly in capturing multiple inter-industry linkages, which is perhaps not surprising since the industrial classification system groups industries based on the similarities of their products/services, not on their inter-industry complementarities. Moreover, manufacturing and service industries are assigned to different parts of the NAICS and are thus, by definition, unrelated. Numerous empirical studies on economies of agglomeration have relied solely on industry codes to define inter-industry relatedness, potentially limiting their ability to capture a broad array of relevant industry linkages.

Our benchmark set, the BCD, scores higher in capturing a broader range of inter-industry linkages than the three other prominent sets of cluster definitions. The BCD also scores better or the same as the other sets in IO linkages and shared labor occupations.

While the analysis is based on U.S. data, our clustering methodology can be implemented in other large and integrated economies with sufficient data availability. At present, however, U.S. data offer several advantages. Comparably large and diversified economic areas (EAs) like the EU have heterogeneous data accessibility across countries, and current and past barriers to trade across locations can limit economies of agglomeration. Smaller economies (like the Nordics) have access to rich data, but their relative small size leads to specialization in a narrow range of industries.

Our BCD offer a useful tool for research and policy on regional economic development. They allow the comparison of clusters across locations and over time by mapping the defined clusters into regional units and measuring the specialization patterns of regions in the clusters. Using the BCD, the U.S. Cluster Mapping Project has created a detailed regional cluster dataset that facilitates comparisons across regions and across clusters on numerous dimensions.[2] For example, the Boston, MA, and San Diego, CA, EAs have high employment specialization in Biopharmaceuticals, but the size and breadth of the cluster is lower in San Diego, which lacks specialization in biological products (with the exception of in-vitro diagnostic substances). The Project also includes data on the business environment and cluster institutions to inform research and policy. This data tool can be used in combination with cluster methods that focus on examining region-specific links among firms, individuals, and supportive institutions in clusters to shed light on the mechanisms at play in particular regional clusters.

The rest of the article is organized as follows: Section 2 reviews the literature on industry cluster definitions. Section 3 describes our clustering algorithm. In Section 4, we discuss our main findings regarding the generation and assessment of cluster configurations, and Section 5 proposes a set of BCD. In Section 6 we compare the BCD to existing sets of cluster definitions in the literature, and in Section 7 we discuss some research and policy applications. Section 8 concludes.

2. Defining industry clusters

There are various types of economies of agglomeration identified in the literature, including IO linkages, labor market pooling, knowledge spillovers, sophisticated local

[2] The U.S. Cluster Mapping Project (http://clustermapping.us/) has been supported by the U.S. Economic Development Administration, U.S. Department of Commerce.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

demand, specialized institutions (e.g., trade organizations) and the organizational structure of business and social networks (Marshall, 1920; Porter, 1990, 1998; Swann, 1992; Saxenian, 1994; Storper, 1995; Markusen, 1996; among others). These economies of agglomeration manifest themselves in geographic concentrations of related economic activity. Within regional clusters (a subset of Porter's (1990, 1998) clusters), firms may operate more efficiently and innovate faster due to sharing common technologies, infrastructure, pools of knowledge and skills, inputs and responding to demanding local customers.

To implement cluster research and policy, however, we need a more operational definition of clusters that measure their boundaries (e.g., what set of related economic activities constitutes a cluster?) Two main approaches to defining clusters have developed over the last 20 years: clusters based on inter-industry linkages inferred from multi-region analysis (which we refer to as *comparable cluster definitions*) and cluster definitions based on observed linkages among industries or firms in a single region (which we refer to as *region-specific cluster definitions*). Many empirically derived cluster definitions have been generated by researchers and practitioners over the years based on both approaches (see Cortright (2006) and Feser et al. (2009) for a review). The goal of this article is to develop a novel methodology to generate and assess sets of comparable cluster definitions. We next explain both approaches to define clusters and the contribution of our clustering method to the literature.

2.1. Comparable cluster definitions

A set of comparable cluster definitions allocates individual industries to specific clusters. By defining clusters as a fixed set of industries, it is possible to compare different regions in terms of that cluster definition, as well as the overall cluster composition of regions. According to this approach, we can define a group of related industries to compose a cluster (e.g., Aerospace Vehicles and Defense cluster), and then evaluate how different regions compare in terms of their specialization in that cluster (e.g., the Seattle-Tacoma-Olympia WA region might be more specialized in the Aerospace Vehicle and Defense cluster than the Minneapolis-St. Paul-St. Cloud, MN-WI region). Regional cluster strength reflects specialization in an array of related industries, not specialization in a narrowly defined single industry (Porter, 1998, 2003; Feldman and Audretsch, 1999; Delgado et al., 2014b). Thus, regional cluster strength is conceptually similar to the notion of 'related variety' introduced by Frenken et al. (2007).

There are two types of inter-industry relatedness measures that have been developed in the literature (see Section 3.1 for a detailed explanation). Some studies use national-level data to capture particular inter-industry linkages, including knowledge links based on co-patenting (see e.g., Scherer, 1982; Koo, 2005a; Glaeser and Kerr, 2009); input and output links (see e.g., Feser and Bergman, 2000; Feser, 2005); skill links (see e.g., Koo, 2005b; Glaeser and Kerr, 2009; Neffke and Henning, 2013) and product similarity as defined by the industry classification system (e.g., same three-digit NAICS). Still other studies define measures based on the co-location patterns of industries across many regions to capture various types of linkages (Ellison and Glaeser, 1997; Porter, 2003; Ellison et al., 2010). Only a few studies use the inter-industry relatedness measures to then define clusters of related industries. We next discuss the main existing sets of comparable cluster definitions.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

2.1.1. Knowledge clusters

Studies of knowledge clusters focus on a selected set of U.S. manufacturing industries with high technological intensity. For example, Feldman and Audretsch (1999) group industries that have a common science and technological base, using the Yale Survey of R&D Managers. This survey assesses the relevance of key academic disciplines for a product category. Industries with similar rankings of the importance of different academic disciplines are grouped into six mutually exclusive clusters. Alternatively, Koo (2005b) groups manufacturing industries into seven mutually exclusive knowledge-based clusters using principal component factor analysis on an inter-industry patent-citation flow matrix.

2.1.2. IO clusters

Feser and Bergman (2000) define a set of U.S. manufacturing clusters using IO links based on the Benchmark IO Accounts of the United States. They group IO classification codes (IO codes) into 23 clusters using principal component factor analysis on an inter-industry IO link matrix. The factor analysis method tends to create highly uneven clusters, with a large number of IO codes grouped into a few clusters. To address this issue, Feser (2005) develops a new methodology based on hierarchical clustering on an IO link matrix for manufacturing and service activities. This transparent method creates a set of 45 mutually exclusive clusters. Overlapping clusters are then created in a second stage by identifying secondary IO codes highly related to the primary codes within a cluster. For each cluster, the method provides scores of the fit of each IO code within its cluster. The IO codes are then matched to 2002 NAICS codes to create a final set of clusters of related industries.

2.1.3. Co-location-based clusters

Porter (2003) examines the co-location patterns of narrowly defined service and manufacturing industries to define clusters, following the principle that co-location reveals the presence of linkages across industries. The methodology first distinguishes traded and local industries. Local industries are those that serve primarily the local markets (e.g., retail), whose employment is evenly distributed across regions in proportion to regional population. Traded industries are those that are more geographically concentrated and produce goods and services that are sold across regions and countries. The set of traded industries excludes natural-resource-based industries, whose location is tied to local resource availability (e.g., mining).

To measure the relatedness between a pair of traded industries (four-digit SIC), Porter (2003) computes the pairwise correlation of industry employment across states using 1996 data. This measure of co-location patterns is referred to as the 'locational correlation' (LC) of employment (*LC-Employment*) and captures various types of inter-industry linkages. Porter (2003) uses an iterative approach to define clusters rather than a clustering function approach. A set of 41 narrow (mutually exclusive) clusters are created using an iterative process to identify pairs and then groups of industries highly linked based on statistically significant LCs. In a second stage, a set of broad (overlapping) clusters is created by including other industries that have a high LC with the core industries within the narrow cluster. While the cluster definitions are mainly based on the empirical patterns of employment co-location among industries, the

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Benchmark IO Account of the United States and industry definitions are used to correct the placement of industries with high co-location but low economic relatedness. These cluster definitions have proven very useful in the empirical analysis of the role of clusters on regional performance (see e.g., Porter, 2003; Delgado et al., 2010, 2014b).

2.2. Region-specific cluster definitions

Comparable cluster definitions can capture most economic activities and are necessary for studies that aim to examine clusters across regions. However, one limitation of any multi-region cluster approach is that it overlooks specific inter-industry linkages that may exist in particular regional clusters. These idiosyncratic regional linkages are the focus of the region-specific cluster definitions. This approach focuses on a single region to measure industry and/or firm interdependencies and define the region's clusters. Such studies vary in their industry coverage, types of economic units (industry, technology classes or firms), types of regional units (administrative or non-administrative) and methods.

A small set of papers defines region-specific clusters for a large set of economic activities. Some of these studies identify specific 'driver' industries in which a region has a competitive advantage. Then they use region-specific inter-industry linkages, such as regional IO models, to define the clusters around the driver industries (Hill and Brennan, 2000). Other studies focus on identifying the (non-administrative) geographic boundaries of a given cluster. To do so, they examine the spatial density of businesses for particular industries (Duranton and Overman, 2005) or the spatial density of patents for particular technology classes (e.g., Kerr and Kominers, 2010; Alcacer and Zhao, 2013). The goal is to identify locations with a high density of economic activity in a particular field that will facilitate inter-firm connections and externalities.

The bulk of region-specific cluster definitions are qualitative and based on case studies that tend to focus on particular clusters (see e.g., Bresnahan and Gambardella, 2004; Cortright, 2010; Porter and Ramirez-Vallejo, 2013). These studies rely on existing cluster organizations, industry directories, and other primary data collection to identify clusters. They offer rich details on the firms and institutions within particular defined clusters, but may be less appropriate for comparing clusters across regions.

The conceptual limitation to region-specific approaches to define clusters is that such definitions are based on observed linkages among existing economic activities in a region (Bathelt et al., 2004; Maskell and Malmberg, 2007; Feser et al., 2009; Bathelt and Li, 2013). Activities that are not present in a region (e.g., industries, technology classes and labor occupations that are not present) are classified as unrelated to the other activities in the region. However, such non-present activities could be related to the activities in the region, but historical factors, market imperfections or other factors may have prevented their development. Region-specific cluster definitions could thus be too narrow (or myopic) in terms of the linkages captured because they abstract from the linkages that may be present in other locations. Thus, region-specific cluster definitions could be complemented by comparable cluster definitions derived from patterns across multiple regions.

This article creates a new cluster methodology that systematically generates comparable cluster definitions based on multiple types of inter-industry linkages. This approach provides scores that assess the ability of each set of definitions to capture

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

high inter-industry linkages within individual clusters. For example, we can compare the Feser (2005) and Porter (2003) sets of cluster definitions as well as other sets.

We implement the algorithm and propose a new set of U.S. BCD that captures a wide variety of inter-industry linkages. This set can be updated over time as new data (e.g., new industry definitions) become available.

3. The clustering algorithm

In order to derive clusters of industries, we use cluster analysis, or numerical methods to classify similar objects (cities, people, genes, industries, etc.) into groups (Everitt et al., 2011). In contrast to network analysis, where each object is related to any other object,[3] cluster analysis creates groups (termed clusters) in such a way that objects in the same cluster are more similar among themselves than to those in other clusters.

Defining clusters of related industries involves a number of key choices that can be parameterized in a clustering algorithm. The algorithm includes criteria for scoring alternative cluster configurations. Once the most promising configuration is identified, the algorithm addresses outlier industries to develop a final set of cluster definitions.

The clustering algorithm is designed to define mutually exclusive clusters, where each industry is uniquely assigned to one cluster. The methodology also allows the measurement of relatedness between any pair of (mutually exclusive) clusters and the creation of overlapping clusters (with individual industries shared by multiple clusters).

Drawing on Porter (2003), our method first distinguishes between traded industries (geographically concentrated) and local industries (geographically dispersed). We use the traded categorization defined in Delgado et al. (2014a). There are 1088 six-digit NAICS-2007 industries in the 2009 CBP data (excluding farming and some government activity). They identify 778 traded industries using the specialization and concentration patterns of each industry across U.S. regions. In 2009, the traded industries account for 36% of total U.S. employment, 50.5% of payroll, and more than 90% of patenting activity.

The analysis focuses on grouping the 778 traded industries in service and manufacturing into non-overlapping groups. We refer to each cluster configuration as C and its individual clusters as c. There are five inter-related steps to create and assess each configuration C: (1) define a similarity matrix M_{ij} that captures the relatedness between any two industries; (2) make broad parameter choices β; (3) use a clustering function to create a configuration C based on the similarity matrix and parameter choices ($C = F(M_{ij}, \beta)$); (4) calculate performance scores for each C and identify the most promising configuration C^* and (5) assess and correct the individual clusters in C^* to determine the finalized set of cluster definitions C^{**}. Each of these clustering algorithm steps is explained in detail below.[4]

[3] For example, some papers focus on defining the 'product space'—the network of relatedness between products. Hidalgo et al. (2007) define the product space for exported goods. Other studies focus on specific dimensions of the product space, such as the technology, knowledge or market space (Jaffe et al., 1993; Neffke et al., 2011; Bloom et al., 2012).

[4] All the steps of the algorithm were implemented using STATA software. The online Appendix A: Clustering Analysis (Stata Files) contains the codes to implement the algorithm and replicate the tables in this article.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

3.1. Step 1: similarity matrix

The first step to group related industries into clusters is to define the degree to which each pair of industries is related. A similarity matrix M_{ij} provides the relatedness between any pair of industries i and j. The matrix is based on the choice of indicator and the similarity measure. Indicators used in the literature include employment, number of establishments, measures of buyer-supplier linkages and measures of shared labor requirements. The choice of a similarity measure allows the user to decide how the distance between two industries i and j should be measured (e.g., correlation coefficient, Euclidean, Jaccard index or user-defined measures).

There are many alternative similarity matrices. Our analysis focuses on the inter-industry relatedness measures most frequently used in the field of regional studies to capture economies of agglomeration. The similarity matrices can be divided into three types. First, there are M_{ij} that exploit co-location patterns across many regions to capture various types of inter-industry linkages. This group includes the LC of employment developed by Porter (2003) and the Ellison and Glaeser (1997) co-agglomeration index (COI). Second, there are M_{ij} that focus on national-level inter-industry linkages, including measures based on national input and output tables (see Feser and Bergman, 2000; Feser, 2005; Ellison et al., 2010) and on OCCs (Glaeser and Kerr, 2009). Third, we create multidimensional matrices that use a combination of these matrices. In what follows, we explain each of these similarity matrices as well as additional industry linkages we do not directly measure.

3.1.1. Pairwise industry co-location patterns

Before we explain the co-location similarity matrices used in our analysis, we need to clarify the regional unit used for these measures and the source of the underlying data. There are two spatial approaches to measuring the co-location patterns of industries across regions: using discrete spatial units like states (Ellison and Glaeser, 1997; Porter, 2003; Ellison et al., 2010) and using continuous spatial units that are based on the density of businesses (Duranton and Overman, 2005). Discrete spatial units that capture relevant regional markets offer a reasonable starting point for understanding co-location patterns. The differences between discrete and continuous co-location measures in their ability to capture inter-industry externalities can be tested. For example, Ellison et al. (2010) show that their COI based on states and an approximation of the continuous co-agglomeration metric developed by Duranton and Overman (2005) both capture similar inter-industry Marshallian effects (IO, skill and knowledge links). Using continuous spatial measures is beyond the scope of this article due to data limitations.

We use meaningful administrative regional units: EAs as defined by the Bureau of Economic Analysis (BEA). EAs represent 179 relevant regional markets that cover the entirety of the continental USA (Johnson and Kort, 2004). The underlying employment and count of establishments of an EA-industry is sourced from the CBP 2009 data.[5]

[5] The CBP data are made available at the county, state and U.S. level. EA data are built up from the county file. CBP data use cell suppression in certain geography-industries with a small presence of firms. When employment data are suppressed, a range is reported. We utilize the midpoint in the range in our data.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

3.1.1.1. Locational correlation. Porter (2003) examines the employment co-location patterns of pairs of industries to capture inter-industry linkages of various types (e.g., technology, skills, supply or demand links). He defines the LC of employment (*LC-Employment*) of a pair of industries as the correlation coefficient between employment in industry i and employment in industry j in a region r:

$$\mathrm{LC}-\mathrm{Employment}_{ij}=\mathrm{Correlation}(\mathrm{Employment}_{ir},\mathrm{Employment}_{jr}). \tag{1}$$

Similarly, we also define an alternative LC based on the count of establishments in a region-industry:

$$\mathrm{LC}-\mathrm{Establishments}_{ij}=\mathrm{Correlation}(\mathrm{Establishments}_{ir},\mathrm{Establishments}_{jr}). \tag{2}$$

Economies of agglomeration channels include firms as well as employees. The presence of numerous establishments can facilitate inter-firm interactions that result in spillovers (Glaeser and Kerr, 2009). Thus, the co-location patterns of count of establishments could help capture inter-industry linkages that are facilitated by the number of businesses.

The correlation coefficient is a well-known distance measure for continuous data used in clustering analysis (Everitt et al., 2011). The LC measures can be implemented for very granular industry definitions, and its scale is easy to interpret, with values between −1 and 1. Positive and large values suggest that there are relevant economic interdependencies between a pair of industries. For example, if the location of employment (count of establishments) in electronic computers and software is highly correlated, it would suggest that both industries are linked. While the LC measures tend to capture relevant linkages, it is possible that in some cases, industries with high co-location may have little economic relatedness but instead, capture shared natural resources. As we discuss further below, this does not limit the usefulness of co-location measures.

LC also can be sensitive to the size of the regions (Porter, 2003). For example, for pairs of industries with employment concentrated in large regions (i.e., with many pairs of zero activity across regions), the LC can be biased. We limit this problem by using EAs versus using smaller regional units (like counties) that do not fully capture the regional market. We also implement several sensitivities to EA size (i.e., dropping the largest and smallest EAs) that suggest this problem is limited. In our data, the average *LC-Employment* and *LC-Establishments* of a pair of industries are 0.30 and 0.52, respectively (Table 1).

3.1.1.2. The co-agglomeration index. This index developed by Ellison and Glaeser (1997) captures whether two industries are more co-located than expected if their employment were to be distributed randomly. We use the revised version of the COI in Ellison et al. (2010):

$$COI_{ij}=\sum\nolimits_r(s_{ri}-x_r)\left(s_{rj}-x_r\right)/\left(1-\sum\nolimits_r x_r^2\right), \tag{3}$$

where s_{ri} is the share of industry i's employment in region r; and x_r measures the aggregate size of region r, which they model as the mean employment share in the region across industries. A value of zero or negative for COI would suggest no externalities-driven co-agglomeration. The higher the positive value of the COI, the greater is the potential for externalities between two industries, but it is not easy to assess whether particular positive values are large or small.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Table 1. Descriptive statistics for similarity matrices (778 industries (six-digit NAICS-2007 codes), 2009 data; N = 604,506)

Similarity matrices M_{ij}	Mean	Standard Deviation	Min	Max	Median	Pctile90
LC-Employment (LC-Emp_{ij})	0.296	0.232	−0.176	0.993	0.263	0.631
LC-Establishments (LC-Est_{ij})	0.519	0.259	−0.174	0.998	0.555	0.840
IO (IO_{ij})	0.017	0.046	0	1	0.001	0.064
Labor occupation (Occ_{ij})	0.183	0.202	−0.013	1	0.113	0.450
COI (COI_{ij})	−0.000	0.010	−0.051	0.372	−0.000	0.007
LC-IO-Occ_{ij}*	−0.002	0.645	−1.437	6.644	−0.037	0.801

Notes: An observation is any pair of industries (ij, $i \neq j$).
[a]LC-IO-Occ_{ij} is an average of the (standardized) LC-Employment, LC-Establishments, IO and Occ.

Ellison et al. (2010) compute the COI for pairs of manufacturing industries (three-digit SIC codes) and use states as the main regional unit. They find that each of the three Marshallian effects (IO, skill and knowledge links) matter for the co-agglomeration of a pair of industries. Shared natural advantages (e.g., coastal access) also matter for the co-agglomeration, but this effect is less important than the cumulative effect of the Marshallian factors. Their findings suggest that co-location captures not only meaningful economic interdependencies and externalities between industries, but also some natural advantages.

We extend the Ellison et al. (2010) analysis, and compute the COI for six-digit NAICS manufacturing and service industries, using EAs as the regional unit (see Table 1 for the descriptive statistics of this measure).

3.1.2. National-level inter-industry links

We explain the next two similarity matrices that are based on national-level data: IO and OCCs. Because these measures do not consider location patterns, they may capture industry interdependencies that are not geographically bounded.

3.1.2.1. IO links. Measures based on the Benchmark IO Accounts of the United States are widely used to capture supplier and buyer flows between industries (see Feser, 2005; Ellison et al., 2010; Alcacer and Chung, 2014). Following Ellison et al. (2010), we construct a symmetric IO link between any pair of industries i,j based on the maximum of all unidirectional input and output links:

$$\text{IO}_{ij} = \text{Max}\{\text{input}_{i \to j}, \text{input}_{i \leftarrow j}, \text{output}_{i \to j}, \text{output}_{i \leftarrow j}\}. \tag{4}$$

The $\text{input}_{i \to j}$ link is the share of industry i's total value of inputs that comes from industry j, and the $\text{output}_{i \to j}$ link is the share of industry i's total value of outputs that goes to industry j.[6] The IO_{ij} link takes a minimum value of zero if the two industries do

[6] To properly capture the strength of the IO links between two industries, we compute these percentages excluding final consumption and value-added commodity codes.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

not buy from or sell to each other, and a maximum value of 1 if any of the two industries buy or sell exclusively from or to the other.

To compute this variable, we use the 2002 Benchmark IO Account of the United States developed by the BEA. Most pairwise industrial combinations have a small IO link (also documented at Ellison et al., 2010), making the distribution over all pairwise combinations skewed to the right (Table 1). Overall, IO tables are more detailed for manufacturing than service industries, and so may better capture links among manufacturing industries.

In the sensitivity analysis, we also compute a more conservative IO link score that takes the average (versus maximum) of the unidirectional input and output links, correcting downward the score for pairs of industries with large asymmetries in their links. The average and maximum pairwise IO links are highly correlated, and our findings are robust to using these alternative measures.[7]

3.1.2.2. Labor occupation links. Labor occupations have been used to measure the extent to which industries share similar skills (Koo, 2005a; Glaeser and Kerr, 2009). We use the OES Survey of the Bureau of Labor Statistics (2009 data). The OES data provide 792 non-governmental occupations and information on the prevalence of these occupations for each industry (i.e., for each occupation (e.g., computer programmers); it provides the percentage of that occupation in the total occupational employment of the industry). Using these data and following Glaeser and Kerr (2009), we compute the pairwise correlation between the occupation composition of any two industries:

$$\mathrm{Occ}_{ij} = \mathrm{Correlation}(\mathrm{Occupation}_i, \mathrm{Occupation}_j), \tag{5}$$

where $\mathrm{Occupation}_i$ is a vector with the percentage of each of the 792 occupations in the total occupational employment of industry *i*. A limitation of this measure is that occupation data are aggregated at the four-digit NAICS level (i.e., industries with the same four-digit NAICS will have the maximum occupational link by construction).[8] The average labor occupation correlation in our sample is 0.18.[9]

3.1.3. Multidimensional similarity matrices

We also create multidimensional similarity matrices that average the unidimensional matrices described above (see Table A1 in the Appendix for the definitions of all M_{ij} used in the analysis). Creating multidimensional similarity matrices begins with understanding the relationship between the unidimensional matrices. Looking at the correlations in Table 2, *LC-Employment* is highly correlated with *LC-Establishments* (correlation of 0.77) and with the COI (correlation of 0.36). These high correlations are robust to the size of the industry and to manufacturing and service industries. IO links have a modest positive correlation with the other measures. We also explore a matrix that captures product similarity as defined by the industry code (NAICS-3). This matrix

[7] Other papers use measures of indirect IO links that capture the extent to which a pair of industries have meaningful suppliers and buyers in common (see Feser, 2005).

[8] We are using seven-digit Standard Occupational Classification and four-digit NAICS data because of better coverage. The data can be accessed at http://www.bls.gov/oes/oes_dl.htm.

[9] Another way to measure skill links between industries is to examine the actual flow of employment using matched employer–employee data for the workforce of a country. See Neffke and Henning (2013) inter-industry skill-relatedness analysis for the Swedish economy.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Table 2. Correlation between similarity matrices (778 industries, 2009 data; $N = 604{,}506$)

	LC-Emp	LC-Est	IO	Occ	COI	NAICS-3
LC-Employment	1.00					
LC-Establishments	0.77	1.00				
IO	0.16	0.13	1.00			
Labor occupation	0.03	0.10	0.12	1.00		
COI	0.36	0.17	0.07	0.14	1.00	
NAICS-3[a]	0.05	0.05	0.09	0.45	0.09	1.00
LC-IO-Occ$_{ij}$	0.76	0.78	0.55	0.49	0.29	0.25

Notes: An observation is any pair of industries (ij, $i \neq$j). All coefficients are significant at 1% level. All variables are based on 2009 data except for IO, which is based on 2002 data.
[a]NAICS-3 matrix is equal to 1 for a pair of industries with the same three-digit NAICS code (and 0 otherwise).

is equal to 1 for pairs of industries with the same three-digit NAICS code (and 0 otherwise), and relates very poorly with all similarity matrices except with occupational linkages (correlation of 0.45).

There are two potential benefits from using a multidimensional similarity matrix. First, as mentioned earlier, each unidimensional matrix has strengths and weaknesses (e.g., IO is overly aggregated for some industries, while LC likely overweighs the relationship among industries that are co-located due to unobserved common inputs such as shared natural resources). By combining multiple measures, we reinforce those areas of similarity that are present across multiple distinct measures while placing less weight on industry pairings that are strong according to one measure but not others.[10] Second, given that our measures of industry relatedness are noisy (i.e., are observed with error), averaging across multiple similarity matrices would be helpful even if we were capturing multiple indices of the same agglomerative force: averaging increases the signal-to-noise ratio.

For example, we compute an M_{ij} that we call LC-IO-Occ$_{ij}$, which is the average of the four (standardized) individual matrices: LC-Employment$_{ij}$, LC-Establishments$_{ij}$, IO$_{ij}$ and Occ$_{ij}$. The multidimensional LC-IO-Occ has a high and statistically significant correlation with each of the individual matrices (Table 2). This suggests that a pair of highly linked industries based on one particular measure (e.g., IO) will also tend to be meaningfully related based on the multidimensional similarity matrix. Thus, LC-IO-Occ seems to better measure various inter-industry links.

Through our algorithm, we can compare the performance of cluster configurations derived from different similarity matrices using the validation scores (VSs) developed in

[10] For example, the IO data are aggregated above the **six**-digit NAICS classification for some industries (e.g., the NAICS-488000 industry is aggregated at the three-digit level). The disaggregation of the LC measures at the six-digit NAICS level allows us to assign industries that are within a three-digit NAICS to different clusters based on these more granular measures (e.g., the *Port and Harbor Operations* industry (NAICS-488310) to a Water Transportation cluster and the *Air Traffic Control* industry (NAICS-488111) to a more general Transportation and Logistics cluster). At the same time, the information from the IO matrix is nonetheless informative in determining the overall choice of most appropriate cluster fit for those industries. Thus, a combined matrix can help capture a broader set of inter-industry linkages.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Step 4. We can then assess which matrices seem to result in cluster configurations that capture the broadest range of inter-industry links (see Section 4).

3.1.4. Similarity matrices and alternative agglomeration mechanisms

While we explore a particular set of relevant inter-industry measures, there are specific agglomeration mechanisms that we do not measure explicitly, such as knowledge linkages and social linkages.

Prior studies that focus on aggregated industries in manufacturing examine inter-industry knowledge linkages using patent citation patterns (e.g., Koo, 2005b; Ellison et al., 2010). We cannot create inter-industry patenting linkages due to data limitations. However, knowledge linkages may be partly captured by our industry linkage measures. For example, co-location patterns of industries could capture some knowledge links as shown by Ellison et al. (2010). Two industries may co-locate across regions because they share knowledge, and proximity facilitates the flow of knowledge. Similarly, if two industries share labor occupations, knowledge linkages could flow more easily.

We also do not measure social linkages of firms and individuals, which are important to define regional clusters. The inter-industry economic links captured by our measures can facilitate opportunities for inter-firm interactions. For example, firms operating in industries that share labor requirements or other inputs are more likely to interact and develop socioeconomic links.

If measures of inter-industry knowledge linkages or social linkages become available, they can be incorporated into our clustering algorithm. We can compare them against other similarity matrices, and assess how cluster configurations that are generated using the new matrices perform in the validation scores defined in Step 4.

More broadly, the nature and intensity of knowledge and socioeconomic linkages can vary significantly across regional clusters. Studies of the network among firms, individuals and associated institutions will be especially informative as to the mechanisms at work in specific clusters (Sorenson and Audia, 2000; Rosenthal and Strange, 2003; Bathelt et al., 2004; Storper and Venables, 2004; Feldman et al., 2005; Lorenzen and Mudambi, 2013).

3.2. Step 2: Broad parameter choices

The parameter choices (β) required as inputs to the clustering functions include setting the initial number of clusters (i.e., number of groups), determining how the underlying data should be normalized and determining the starting values for the clustering function.

An important parameter choice in clustering analysis is the initial number of clusters (*numc*). There are 41 clusters in Porter (2003) and 45 IO-based clusters in Feser (2005). Current methods to identify the 'optimal' number of clusters in clustering analysis are very inconclusive (Everitt et al., 2011). Therefore, we explore values for the number of clusters between 30 and 60. Overall, too few or too many groups could result in less useful cluster definitions. Too few clusters could result in large clusters that include industries that are not very related; and too many clusters could result in clusters that are not meaningfully different from each other. Using Step 4 in the cluster algorithm (described below), it is possible to compare the scores of different configurations based on differences in the initial number of clusters.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

The other two parameter choices refer to the starting values and the type of normalization of the underlying data for the clustering functions. The starting values were chosen at random. The underlying data were either untransformed (raw) or row-standardized (rst). These two parameter choices are relevant only for partition-clustering functions: *kmeans* and *kmedians*. The normalization of the data can be important for these two clustering functions since it could result in a better centroid for each individual cluster.[11]

3.3. Step 3: Clustering function

Clustering functions are designed to find the greatest relatedness among industries within each cluster. There are several clustering functions $F(\bullet)$ for grouping industries into clusters (see Everitt et al., 2011; Grimmer and King, 2011). Each function creates a new grouping C based on the similarity matrix and parameter choices: $C = F(M_{ij}, \beta)$. Our analysis uses the main cluster functions for continuous data: the *hierarchical* function (with Ward's linkage) and centroid-based clustering functions (*kmean* and *kmedian*).

Only hierarchical functions allow the user to import a particular similarity matrix. In contrast, *kmean* and *kmedian* functions require the underlying raw data to directly compute the similarity matrix (and centroids). Thus, for similarity matrices that require additional manipulation of the underlying data (e.g., IO or COI), we can only use the hierarchical function.

3.3.1. Example of steps 1–3 of the clustering algorithm

To illustrate how numerical clustering analysis works, we implement Steps 1–3 of the algorithm to replicate a set of cluster definitions that we already know, namely the three-digit NAICS groupings. We define the similarity matrix $NAICS\text{-}3_{ij}$ as a symmetric binary matrix where pairs of industries within the same three-digit NAICS code are assigned a value of 1 (and a value of 0 otherwise). Then, we set the broad parameters (β) so that there are 66 clusters just as there are 66 different three-digit NAICS codes for our 778 industries. Finally, we run the hierarchical clustering function using the *NAICS-3* matrix and 66 clusters. As expected, we find that the resulting grouping C is equal to the NAICS-3 groupings.

3.4. Step 4: Performance scores for each C

Given the number of possible similarity matrices, parameters and clustering functions that could be chosen, the number of alternative cluster configurations is quite large. By combining the choices described above in Steps 1–3 in different ways, we have generated 713 different Cs. For example, choosing the *LC-Employment* similarity matrix, 40 clusters, raw underlying data and the *kmean* clustering function will result in one configuration C_1 (see Table 3 for examples of Cs).

Without some way to evaluate these Cs, it is very hard to find the most useful sets of definitions that incorporate a broad range of inter-industry linkages. The cluster

[11] The centroid of a cluster is the mean industry employment (for *kmean*) and the median industry employment (for *kmedian*). These centroids could be biased toward larger regional industries. To limit this problem, we allow for row-standardization of the region-industry employment/establishment data.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Table 3. Examples of cluster configurations generated

Similarity matrix M_{ij}	Parameter choices β		Clustering function $C = F(M_{ij}, \beta)$	Number of Cs
	Number of clusters (numc)	Data		
LC-Employment (LC-Emp$_{ij}$)	30–60	Raw	Hierarchical-Ward's	31
		Raw/Rst	Kmean	62
		Raw/Rst	Kmedian	62
LC-Establishments (LC-Est$_{ij}$)	30–60	Raw	Hierarchical-Ward's	31
		Raw/Rst	Kmean	62
		Raw/Rst	Kmedian	62
Labor occupation (Occ$_{ij}$)	30–60	Raw	Hierarchical-Ward's	31
		Raw	Kmean	31
		Raw	Kmedian	31
IO (IO$_{ij}$)	30–60	Raw	Hierarchical-Ward's	31
COI	30–60	Raw	Hierarchical-Ward's	31
LC_IO_Occ$_{ij}$[a]	30–60	Raw	Hierarchical-Ward's	31

[a]This M_{ij} is an average of the (standardized) LC-Employment, LC-Establishments, IO and Occ matrices. See Table A1 for a list of all the similarity matrices used. The Hierarchical function uses Ward's linkages.

analysis literature often lacks satisfactory methods for evaluating different categorization schemes (Grimmer and King, 2011). In contrast, our approach provides a score for each C. In order to generate these scores, we must first address the question, *What makes a good set of cluster definitions?*

In our analysis, the primary criterion for a good set of cluster definitions is that industries within a particular cluster (e.g., the Automotive cluster) should be more closely related among themselves than to industries in other clusters. In other words, individual clusters should be meaningfully different from each other, and individual industries should fit well within their own cluster. Our score approach assesses this by using alternative measures of inter-industry linkages to generate validation sub-scores (e.g., sub-scores based on IO links). Our view is that a useful set of clusters will capture various types of industry linkages, including demand, supply, skills and others (Marshall, 1920; Porter, 1998). Thus, we develop an overall validation score (VS) for each C that combines sub-scores based on alternative industry measures.

A secondary criterion is that the configurations should be robust. We would prefer cluster definitions that are similar to other well-performing cluster definitions generated by the algorithm, since this would suggest that they are more robust. We develop Overlap Scores (OSs) to capture the overlap of each C to other configurations.

Those Cs with the higher ranked VSs are then subject to the robustness criteria to select the better configurations. The configuration that does relatively well in all criteria is the C^* selected to undergo further assessment in Step 5. In the remainder of this section, we explain the validation and OSs.

3.4.1. Validation scores

We develop VSs that capture the extent to which individual clusters and industries have high within cluster relatedness (WCR) relative to between cluster relatedness (BCR)

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

with other clusters. The VSs assess a cluster configuration C along two dimensions. The first score, *VS-Cluster*, captures whether *individual clusters* in C are meaningfully different from each other. The second score, *VS-Industry*, assesses the fit of *individual industries* within their own cluster.

At the cluster level, we define $\mathrm{WCR_c}$ as the average relatedness between pairs of industries within a cluster, while $\mathrm{BCR_c}$ is the average relatedness between industries in cluster c and those in another cluster. For example, consider two clusters in C: cluster c_1 with industries a_1, a_2 and cluster c_2 with industries b_1, b_2 and a similarity matrix M_{ij} (e.g., *LC-Employment*) that may be different from the one used to generate C. Then, the WCR of focal cluster c_1 is $\mathrm{WCR}_{c_1} = M_{a_1 a_2}$, and the BCR of c_1 and c_2 is $\mathrm{BCR}_{c_1,c_2} = \mathrm{Avg}(M_{a_1 b_1}, \mathrm{M}_{a_1 b_2}, \mathrm{M}_{a_2 b_1}, \mathrm{M}_{a_2 b_2})$.

For each focal cluster c in C, we compute its BCR with every other cluster and examine the resulting distribution to compute two cut-off values—the average and the 95th percentile values ($\mathrm{AvgBCR_c}$ and $\mathrm{Pctile95BCR_c}$). For example, if there are 51 clusters in C, for each focal cluster c we then have 50 different $\mathrm{BCR_c}$ values, and we compute the average and the 95th percentile of the $\mathrm{BCR_c}$ values. We can then assess whether a cluster's WCR_c is higher than these two threshold values. Specifically, for each cluster in C, we compute a VS that captures the percent of *clusters* with high $\mathrm{WCR_c}$ (*VS-Cluster*). This score is made up of two broad sub-scores. The first calculates the percent of clusters in C with $\mathrm{WCR_c}$ higher than $\mathrm{AvgBCR_c}$ (*VS-Cluster Avg*) based on a particular similarity matrix M_{ij}:

$$\mathrm{VS-Cluster\ \ Avg}\ {}^{\mathrm{M}}_{\mathrm{C}}= \left({}^{100}\!/_{N_c}\right)*\sum_{\mathrm{c}} \mathrm{I}[\mathrm{WCR_c}(M_{ij}) > \mathrm{AvgBCR_c}(M_{ij})], \qquad (6)$$

where N_c is the number of clusters in C and I is an indicator function equal to 1 if for a given cluster c the $\mathrm{WCR_c} > \mathrm{AvgBCR_c}$. The second sub-score is similar but more restrictive; it calculates the percent of clusters in C with $\mathrm{WCR_c}$ higher than $\mathrm{Pctile95BCR_c}$ (*VS-Cluster Pctile95*). We then average these two sub-scores to compute *VS-Cluster*$^{\mathrm{M}}$.

For each cluster configuration C, we compute these validation sub-scores based on four distinct unidimensional M_{ij} (LC-Employment, LC-Establishments, IO and Occ). It is useful to note that these similarity matrices are not dependent on the similarity matrix used to create C, and so we can calculate sub-scores that can be consistently compared regardless of the underlying measures used to generate C. As shown in Table 4, the resulting sub-scores are then averaged in computing an overall cluster VS (*VS-Cluster*), which takes a maximum value of 100 (i.e., all the individual clusters in C contain industries that are highly related based on multiple linkages). Table 5 then illustrates how to use this score approach by examining how a cluster configuration C* generated using the multidimensional LC-IO-Occ matrix performs against each of the four unidimensional matrices.

So far, we have computed a VS that examines individual clusters. We then compute a VS based on the fit of individual industries within their own cluster (*VS-Industry*). For a given industry i, we want it to be more related to the industries within its own cluster than to industries outside its cluster. Similar to our calculation of *VS-Cluster*, we measure the percent of industries (out of 778) with WCR_{ic} higher than their average BCR_i (*VS-Industry Avg*) and higher than the 95th percentile of

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Table 4. Descriptive statistics: Validation scores for cluster configurations (Number of *C*s = 713)

Validation scores	Description	Mean	Standard Deviation	Min	Max	Pctile90
VS-Cluster	% of clusters with high WCR_c (average of *VS-Cluster* sub-scores)	73.9	4.8	63.8	83.8	81.0
VS-Industry	% of industries with high WCR_{ic} (average of *VS-Industry* sub-scores)	66.2	3.1	56.9	73.1	70.1
VS	Average *VS-Cluster* and *VS-Industry*	70.1	3.6	63.4	78.0	75.1
Sub-scores based on individual similarity matrix (LC-Emp, LC-Est, IO or Occ)						
$\text{VS-Cluster}^{\text{LC-Emp}}$	% of clusters with high $\text{WCR}^{\text{LC-Emp}}$	76.8	14.8	50.0	100.0	96.9
$\text{VS-Cluster}^{\text{LC-Est}}$	% of clusters with high $\text{WCR}^{\text{LC-Est}}$	72.0	10.8	50.0	96.7	89.5
$\text{VS-Cluster}^{\text{IO}}$	% of clusters with high WCR^{IO}	67.5	15.9	41.7	98.3	94.2
$\text{VS-Cluster}^{\text{Occ}}$	% of clusters with high WCR^{Occ}	79.3	18.1	47.3	100.00	100.0
$\text{VS-Industry}^{\text{LC-Emp}}$	% of industries with high $\text{WCR}^{\text{LC-Emp}}$	70.7	15.9	45.9	97.8	92.9
$\text{VS-Industry}^{\text{LC-Est}}$	% of industries with high $\text{WCR}^{\text{LC-Est}}$	68.8	13.4	45.6	95.2	88.7
$\text{VS-Industry}^{\text{IO}}$	% of industries with high WCR^{IO}	54.6	12.0	40.1	79.4	75.1
$\text{VS-Industry}^{\text{Occ}}$	% of industries with high WCR^{Occ}	70.8	19.1	46.5	99.8	98.3

Notes: For each *C*, *VS-Cluster* is the average of the sub-scores ($\text{VS-Cluster}^{\text{LC-Emp}}$, VS $\text{Cluster}^{\text{LC-Est}}$, $\text{VS-Cluster}^{\text{IO}}$, $\text{VS-Cluster}^{\text{Occ}}$); and similarly *VS-Industry* is the average of the VS-Industry sub-scores. See Table 5 for an illustration of the scores and sub-scores using a particular *C*.

Table 5. Validation scores (sub-scores) for selected cluster configuration *C** (778 industries)

*C** = Hierarchical (LC-IO-Occ, 51 clusters)

Validation sub-scores	VS-Cluster % Clusters with high WCR_c	VS-Industry % Industries with high WCR_{ic}	VS (Avg VS-Cluster, VS-Industry)
$\text{VS}^{\text{LC-Emp}}$ (Avg)	69.6	58.7	64.1
WCR > AvgBCR	92.2	89.6	
WCR > Pctile95BCR	47.1	27.8	
$\text{VS}^{\text{LC-Est}}$ (Avg)	69.6	61.2	65.4
WCR > AvgBCR	90.2	91.4	
WCR > Pctile95BCR	49	31	
VS^{IO} (Avg)	95.1	79.4	87.2
WCR > AvgBCR	100	88.4	
WCR > Pctile95BCR	90.2	70.3	
VS^{Occ} (Avg)	98.0	92.0	95.0
WCR > AvgBCR	100	98.3	
WCR > Pctile95BCR	96.1	85.7	
Validation scores	83.1	72.8	78.0
Rank (1 = best)	2	5	1

Notes: We compute to what extent individual clusters and industries in *C* have WCR greater than a BCR cut-off value based on particular similarity matrices M_{ij}. We then average these sub-scores into the VSs: VS = Avg($\text{VS}^{\text{LC-Emp}}$, $\text{VS}^{\text{LC-Est}}$, VS^{IO}, VS^{Occ}). The VS Rankings are computed across 713 sets of cluster definitions.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

BCR_i (*VS-Industry Pctile95*) based on each of the four unidimensional similarity matrices.[12]

The overall VS of a cluster configuration is computed as the average of the *VS-Cluster* and *VS-Industry* scores. Those *C*s with highly ranked scores for both *VS-Cluster* and *VS-Industry* are the most promising configurations (the 'candidates' *C**s). The final candidate *C** is the configuration with the maximum VS score. For example, Table 5 illustrates the VSs and sub-scores for the candidate configuration *C**.

3.4.2. Overlap scores

The candidate *C**s are subject to the robustness criteria. We develop scores that capture the robustness of a particular *C* by comparing the industry overlap between the clusters in *C* and the clusters in other candidate configurations. To compare a configuration C_1 to another C_2, for each individual cluster *c* in C_1, we find a matching cluster *b* in C_2 (i.e., the cluster *b* that has the highest industry overlap with *c*). Specifically, we compute the overlap between a pair of clusters *c*, *b* using the geometric mean of the industry overlap in each direction:

$$\mathrm{overlap}_{c,b}=100\left(\text{Shared Industries}_{c,b}\Big/\sqrt{\text{Industries}_c*\text{Industries}_b}\right),$$

where Shared Industries is the number of industries in common in *b* and *c*; and Industries are the number of industries in each cluster. The maximum overlap of a pair of matched clusters is 100. Then we define the OS of C_1 to C_2 as the average industry overlap across C_1's clusters: Overlap $\mathrm{Score}_{\mathrm{C_1-C_2}}=\frac{1}{N_\mathrm{c}}\sum_{\mathrm{c}\in\mathrm{C_1}}\mathrm{overlap}_{\mathrm{c,b}}$. Similarly, we compute the average overlap of C_1 with all other candidate configurations ($\mathrm{OS}_{\text{C-Candidates}}$). For example, on average the proposed candidate *C** has an industry overlap of 86% with other relevant candidates, indicating that these alternative *C*s tend to provide, on average, similar groupings of industries (Table 7).

The configuration that does relatively well in the VS (and OS) is the *C** selected to undergo further assessment in Step 5. Generally, the higher the VSs, the better the *C**. However, there could be anomalies within individual clusters that would require some assessment and reallocation of individual industries to obtain the finalized set of cluster definitions *C***.

3.5. Step 5: Assessing individual clusters of candidate *C**

Because clustering analysis cannot perfectly substitute for expert judgment, the methodology concludes with a transparent correction of anomalies and characterization of the individual clusters in *C**, resulting in a finalized set of cluster definitions. Although Steps 1–4 systematically assign industries to clusters, the resulting *C*s can be improved. Limitations in the underlying data may create spurious industry relatedness that will place some industries into clusters where they are not the best fit. Some clusters may contain conceptually distinct groups that have not been separated because of the choice of initial number of groups (*numc* parameter); and other clusters may be better

[12] The industry WCR_{ic} score is the average pairwise relatedness between industry *i* and the other industries within the cluster, whereas BCR_i is the average relatedness between industry *i* and industries in a different cluster.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

off combined. Step 5 allows us to examine the clusters in C^* to assess whether there are industry outliers that are better placed into different clusters and whether to combine or break individual clusters to improve the coherence of the clusters. We can use our score approach to inform these expert-driven choices. Users can assess how certain changes impact the WCR scores of individual industries and clusters, and the VSs of the cluster configuration relative to the initial values.

We define two types of possible outlier industries: *systematic* and *marginal* outliers. Systematic outlier industries are those with a low overall WCR_{ic} score (based on the average of standardized sub-scores for $\text{WCR}^{\text{LC-Emp}}$, $\text{WCR}^{\text{LC-Est}}$, WCR^{IO} and WCR^{Occ}).[13] Systematic outliers are identified and corrected with a simple sub-process. They are identified based on two criteria: the industry WCR is low relative to other clusters (i.e., WCR_{ic} is below the 75th percentile value of BCR_{i}); or WCR is low relative to other industries in the same cluster (i.e., WCR_{ic} is two standard deviations below the average WCR_{ic}). Then, these systematic outliers are reassigned to the cluster where their WCR is highest. This sub-process is iterated several times until there are no systematic outliers.

Marginal outlier industries are those industries that, even with a high WCR_{ic}, could be conceptually better in another cluster. These outliers are often the result of limitations in the underlying data.[14] For example, *Men's and Boy's Clothing Manufacturing* industries (NAICS 315221-315228) are in the Printing Services cluster for C^*, but they likely best belong to an Apparel cluster. Identifying these marginal outliers requires examining each cluster and analyzing the main product/service lines of the industries based on the detailed definitions offered by the NAICS system. The outliers are reallocated to their 'next best' cluster using the WCR_i scores. Reallocated marginal outliers can be easily tracked and documented so that the process is transparent.

We also examine whether some individual clusters should be combined or partitioned. If two individual clusters have very high BCR and they do not seem conceptually different, they can be combined. In contrast, some individual clusters can be partitioned if we find clear conceptual and relatedness differences among certain sub-groups of industries in a cluster. Because of these corrections, the initial number of clusters (*numc*) and the number of clusters in the finalized set of cluster definitions may differ.

After all five steps in the cluster algorithm are complete, we are able to recommend a final set of BCD C^{**} (the BCD). We explain the main findings and the proposed new cluster definitions in the next sections.

[13] The WCR score is based on these four sub-scores to have a more robust score that captures multiple inter-industry linkages within the cluster.

[14] In some cases, the IO link between two industries may be overestimated due to the level of aggregation of underlying data and/or due to our symmetric measure of IO links. For example, R&D industries (NAICS 541700) appear very highly linked to Water Transportation (NAICS 483000) industries because Water Transportation supplies a large percentage of its output to R&D industries. This induces the R&D industries to be grouped with Water Transportation if IO links are considered in the similarity matrix. For industries where the underlying IO data are highly aggregated and for industries with very high IO links in the cluster, we check that these industries also fit well in their assigned cluster based on the other measures (LC and Occ).

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

4. Generating and assessing cluster configurations

We apply the clustering algorithm to generate 713 different cluster configurations that group 778 six-digit NAICS industries using 2009 U.S. data. These configurations are based on 13 different similarity matrices (M_{ij}) and the parameter and clustering function choices discussed in the prior section ($C = F(M_{ij}, \beta)$). As illustrated in Table 1, some Cs are generated using unidimensional matrices (e.g., IO) and others using multidimensional matrices (e.g., LC-IO-Occ). We then generate the VSs for each configuration to compare Cs generated using different similarity matrices. We then explain the properties of the proposed candidate C^* that will be subject to assessment and adjustments of individual clusters in the last step of the algorithm to obtain the BCD.

4.1. Validation scores by choice of similarity matrix

Through our algorithm, we can compare how well the configurations derived from different similarity matrices perform in the VSs. We can then assess which similarity matrices seem to result in cluster configurations that capture the broadest range of inter-industry linkages and potential externalities (e.g., demand, supply, skills, knowledge and others).

We use our score method to compare Cs generated by either a unidimensional similarity matrix (LC-Emp, LC-Est, IO, Occ and COI) or the multidimensional LC-IO-Occ matrix. Each of these matrices can be used to create a number of different Cs by changing the type of clustering function and/or the parameter choices. For example, there are 31 Cs generated using a hierarchical function, $\mathrm{M}_{ij} = \text{LC-IO-Occ}$, and 31 different values for the number of clusters (*numc*). We then compute the average VSs of Cs generated with the same similarity matrix. The analysis is shown in Table 6, which reports the mean VS and sub-scores ($\mathrm{VS}^{\text{LC-Emp}}$, $\mathrm{VS}^{\text{LC-Est}}$, VS^{IO} and VS^{Occ}) by similarity matrix.

If we want to capture various inter-industry linkages within clusters, then Cs with higher overall VSs will be preferred. We find that Cs generated with the LC-IO-Occ matrix have, on average, statistically significant higher VS scores than other Cs generated based on unidimensional matrices. For example, the mean VS of Cs(LC-IO-Occ) is 76, and the next best mean VS is 71 for Cs(LC-Est) (Table 6). This difference in the means is statistically significant at 1%. The matrix LC-IO-Occ_{ij} seems to generate meaningful sets of cluster definitions that capture a broad set of industry interdependencies. One of the Cs generated with this matrix is the proposed candidate C^*, which is the basis for our final set of cluster definitions.

4.2. Choosing the candidate configuration C*

The choice of C^* depends primarily on the VSs. Those configurations with highly ranked scores for both *VS-Cluster* and *VS-Industry* (i.e., top-50 rankings in both scores across the 713 Cs) are the candidates, C^*s. There are 19 candidates (see Table 7). Then the configuration with the maximum overall VS is the proposed candidate C^*. The C^* is generated using a hierarchical clustering function with 51 clusters and the multidimensional similarity matrix LC-IO-Occ. This configuration has the highest VS across all the Cs, with a score of 78% (see Table 7).

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Table 6. Mean of validation scores (and sub-scores) by selected similarity matrices (M_{ij})

M_{ij}	Number of $Cs(M_{ij})$	Validation score	Validation sub-scores			
		VS (Avg sub-scores)	$VS^{LC\text{-}Emp}$	$VS^{LC\text{-}Est}$	VS^{IO}	VS^{Occ}
LC-IO-Occ	31	76	62	65	84	94
LC-Emp	155	67	94	70	49	54
LC-Est	155	71	79	89	53	63
IO	31	71	56	59	85	82
Occ	93	67	55	55	57	99
COI	31	65	89	70	49	53

Notes: VSs and sub-scores take value [0,100]. The differences in the means of the VSs and sub-scores of Cs(LC-IO-Occ) versus $Cs(M_{ij})$ are all statistically significant at 1% level, except for the sub-score VS^{IO} of *C*s(LC-IO-Occ) and *C*s(IO) (84 and 85).

Table 7. Candidate cluster configurations *C**s (top-50 rankings in both VS-Cluster and VS-Industry)

	Model choices			Validation scores						Overlap score
				VS-Cluster		VS-Industry		VS		$OS_{C\text{-}Candidates}$
*C**s	M_{ij}	Numc	Function	Rank	Score	Rank	Score	Rank	Score	Score
C*	**LC-IO-Occ**	**51**	**Hiw**	**2**	**83.1**	**5**	**72.8**	**1**	**78.0**	86.0
C_2	LC-IO-Occ	43	Hiw	11	82.6	3	73.1	2	77.8	88.5
C_3	LC-IO-Occ	44	Hiw	4	83	15	72.6	3	77.8	88.6
C_4	LC-IO-Occ	45	Hiw	3	83.1	17	72.5	4	77.8	88.6
C_5	LC-IO-Occ	42	Hiw	13	82.4	1	73.1	5	77.8	88.5
C_6	LC-IO-Occ	50	Hiw	7	82.8	8	72.8	6	77.8	86.4
C_7	LC-IO-Occ	53	Hiw	6	82.8	12	72.7	7	77.7	84.4
C_8	LC-IO-Occ	54	Hiw	10	82.6	6	72.8	8	77.7	83.7
C_9	LC-IO-Occ	52	Hiw	9	82.7	10	72.7	9	77.7	85.2
C_{10}	LC-IO-Occ	47	Hiw	8	82.7	16	72.6	10	77.6	87.9
C_{11}	LC-IO-Occ	49	Hiw	15	82.4	6	72.8	11	77.6	87.0
C_{12}	LC-IO-Occ	55	Hiw	17	82.3	4	72.8	12	77.6	82.8
C_{13}	LC-IO-Occ	46	Hiw	5	82.9	21	72.2	13	77.6	88.1
C_{14}	LC-IO-Occ	41	Hiw	21	82	1	73.1	14	77.6	88.1
C_{15}	LC-IO-Occ	48	Hiw	20	82	9	72.7	15	77.4	87.5
C_{16}	COI-IO-Occ	41	Hiw	45	81.4	14	72.6	16	77.0	78.5
C_{17}	LC-Est	43	Hiw	46	81.4	18	72.5	17	76.9	32.9
C_{18}	LC	42	Hiw	27	81.8	39	71.4	19	76.6	33.1
C_{19}	COI-IO-Occ	45	Hiw	48	81.4	35	71.5	23	76.5	76.4

Notes: Rank is across the 713 *C*s. $OS_{C\text{-}Candidates}$ is the average cluster overlap between the focal *C* and the 18 other sets of definitions. Hiw refers to the Hierarchical-Ward's clustering function.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

For sensitivity, we assess the robustness of the candidate C^* by comparing its overlap to the other promising configurations. Table 7 shows that on average C^* has a high overlap with these other Cs, indicating that they tend to provide very similar groupings of industries. The configuration C^* will be subject to assessment and improvement of individual clusters in Step 5 to derive C^{**}.

5. Proposed set of BCD C^{**}

Our methodology concludes with an assessment and correction of the individual clusters in C^* to derive the finalized set of U.S. BCD. We explain this process here, present a summary overview of the BCD and illustrate a few selected clusters. A detailed overview of the cluster definitions, with a description of each cluster, associated industry NAICS codes, summary calculations of the fit of each industry within its cluster and a full explanation of the process to get to these clusters, is available in the supplementary online Appendix B: Set of BCD C^{**}.

The proposed set of cluster definitions C^{**} has 51 clusters (see Table 8).[15] In this set, 136 industries are re-allocated into other clusters: 9 industries are systematic outliers and 127 industries are marginal outliers.[16] We partitioned and combined some clusters in C* to improve the coherence and usefulness of the cluster definitions. These modifications had a trivial effect on the VS score (see Tables 5 and 10).

There are six cases where the algorithm divided industries into two clusters, but we created a single combined cluster because they had high BCR and seemed conceptually similar. Specifically, we combined the following pairs of clusters into an individual cluster: two textile clusters, two financial services clusters, two food clusters, two upstream metal manufacturing clusters, two oil and gas clusters and two transportation and logistics clusters.

We also partitioned the original clusters in five cases: four individual clusters were each partitioned into two clusters, and one cluster was partitioned into three. These partitions are supported by the underlying data and expert opinion, and create clusters that focus on different markets and have better properties (WCR_c) than their original larger cluster. We separated the Agricultural Inputs and Services cluster from the original Forestry cluster, Downstream Metal Products from the original Production Technology cluster and Coal Mining from Electric Power Generation and Transmission. One partition happened systematically when we slightly increased the parameter value for the number of initial clusters (all else being equal): the Performing Arts cluster was separated from the original Hospitality and Tourism cluster. Finally, the original Lighting and Electrical Equipment cluster was partitioned into three clusters: the focal Lighting and Electrical Equipment, Medical Devices and

[15] The BCD is composed of mutually exclusive clusters. Accordingly, some industries are at the boundary of different clusters. For example, *Men's and Boy's Clothing and Furnishings Merchant Wholesalers* (NAICS-424320) is placed within the Distribution and Electronic Commerce cluster based on the clustering algorithm as well as substantive linkages between this industry and others within that cluster, but there are also meaningful linkages between that industry and the Apparel cluster. In ongoing work, we are creating a grouping of 'Related Clusters' and 'Related Industries' for each cluster to allow researchers to accommodate cluster overlaps.

[16] Most industries are moved to clusters with higher WCR_i scores relative to their initial cluster. More than 40% of these industries are re-allocated into a new cluster where they have a WCR_{ic} rank of 1 (best fit out of 51 clusters). A list of the marginal industries can be accessed at the online Appendix B.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Table 8. Overview of proposed set of BCD *C***

Cluster name	Number of industries	Traded employ	WCR		WCR LC-Emp [−1, 1]	WCR LC-Est [−1, 1]	WCR IO [0,1]	WCR Occ [−1, 1]
			Rank	Score				
Aerospace Vehicles and Defense	7	1.3%	1	2.21	0.20	0.63	0.19	0.87
Agricultural Inputs and Services	9	0.2%	2	0.83	0.35	0.53	0.06	0.46
Apparel	21	0.4%	1	2.28	0.45	0.74	0.11	1.00
Automotive	26	1.9%	1	2.26	0.31	0.62	0.22	0.65
Biopharmaceuticals	4	0.6%	1	3.33	0.59	0.76	0.23	1.00
Business Services	33	24.2%	1	1.18	0.66	0.83	0.04	0.25
Coal Mining	4	0.2%	2	2.28	0.44	0.53	0.22	0.62
Communications Equipment and Services	8	1.2%	1	2.36	0.47	0.79	0.23	0.41
Construction Products and Services	20	1.8%	1	1.79	0.39	0.61	0.21	0.29
Distribution and Electronic Commerce	62	13.0%	1	2.18	0.67	0.82	0.12	0.63
Downstream Chemical Products	13	0.6%	1	1.29	0.39	0.69	0.04	0.71
Downstream Metal Products	16	1.0%	1	1.02	0.28	0.58	0.02	0.82
Education and Knowledge Creation	15	6.8%	1	1.33	0.70	0.85	0.03	0.38
Electric Power Generation and Transmission	5	0.3%	2	0.90	0.30	0.31	0.00	1.00
Environmental Services	7	0.2%	1	2.80	0.57	0.78	0.22	0.67
Financial Services	26	4.9%	1	2.03	0.54	0.71	0.16	0.51
Fishing and0 Fishing Products	5	0.1%	1	3.38	0.48	0.63	0.28	1.00
Food Processing and Manufacturing	47	2.2%	1	0.81	0.26	0.46	0.03	0.72
Footwear	6	0.0%	1	5.17	0.09	0.31	0.72	0.75
Forestry	4	0.2%	1	3.52	0.59	0.66	0.38	0.51
Furniture	12	0.9%	1	1.37	0.34	0.76	0.03	0.83
Hospitality and Tourism	31	7.0%	5	0.44	0.50	0.60	0.01	0.21
Information Technology and Analytical Instruments	27	2.5%	1	1.30	0.43	0.78	0.03	0.69
Insurance Services	8	3.8%	1	4.32	0.59	0.82	0.39	0.91
Jewelry and Precious Metals	4	0.1%	1	5.46	0.53	0.79	0.55	1.00
Leather and Related Products	6	0.1%	2	1.32	0.35	0.75	0.01	0.86
Lighting and Electrical Equipment	15	0.7%	1	1.49	0.36	0.75	0.02	0.94
Livestock Processing	5	1.2%	1	1.18	0.34	0.56	0.07	0.65
Marketing, Design, and Publishing	22	2.9%	1	1.68	0.76	0.90	0.04	0.48
Medical Devices	5	0.7%	1	2.12	0.52	0.89	0.07	0.87
Metal Mining	8	0.1%	1	0.62	0.09	0.23	0.05	0.81
Metalworking Technology	17	1.2%	1	1.48	0.62	0.82	0.03	0.59
Music and Sound Recording	5	0.1%	1	6.16	0.85	0.94	0.57	1.00
Nonmetal Mining	13	0.2%	1	0.73	0.13	0.19	0.05	0.89
Oil and Gas Production and Transportation	12	1.3%	1	1.47	0.61	0.64	0.10	0.36
Paper and Packaging	20	0.9%	1	1.62	0.28	0.52	0.08	0.97
Performing Arts	8	0.7%	1	1.64	0.65	0.86	0.07	0.42
Plastics	15	1.6%	1	2.03	0.49	0.76	0.10	0.82
Printing Services	13	1.3%	1	2.53	0.54	0.87	0.15	0.78
Production Technology and Heavy Machinery	41	2.3%	1	1.08	0.29	0.59	0.01	0.89
Recreational and Small Electric Goods	15	0.5%	1	1.30	0.32	0.74	0.01	0.90
Textile Manufacturing	23	0.5%	1	1.19	0.35	0.52	0.10	0.49
Tobacco	3	0.0%	1	7.47	0.09	0.35	1.00	1.00
Trailers, Motor Homes, and Appliances	9	0.2%	1	0.52	0.08	0.26	0.01	0.90
Transportation and Logistics	17	3.8%	1	1.13	0.42	0.76	0.10	0.22
Upstream Chemical Products	12	0.4%	1	1.23	0.12	0.32	0.09	0.95
Upstream Metal Manufacturing	26	0.9%	1	0.99	0.25	0.50	0.04	0.76
Video Production and Distribution	6	0.5%	1	3.37	0.69	0.83	0.27	0.69
Vulcanized and Fired Materials	17	0.6%	1	0.90	0.26	0.53	0.02	0.78
Water Transportation	12	0.7%	1	1.71	0.37	0.69	0.16	0.45
Wood Products	13	0.9%	1	1.70	0.36	0.54	0.10	0.87
Average				2.05	0.42	0.65	0.15	0.71

Notes: WCR is the average of the (standardized) WCR$^{\text{LC-Emp}}$, WCR$^{\text{LC-Est}}$, WCR$^{\text{IO}}$ and WCR$^{\text{Occ}}$.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Recreational and Small Electronic clusters. While these clusters share skills, they are distinct groups. We explain these three clusters below.

After this process of assessment and correction of individual clusters, we finalize the characterization of clusters in C^{**} with the creation of conceptual subcategories (termed 'subclusters') to help describe the content of each cluster. These subclusters are based mainly on industry definitions.

Table 8 offers a summary overview of the clusters in C^{**} (the BCD) with information by cluster on the number of industries, the average WCR score, and sub-scores. All the individual clusters have a high WCR score, but there is variation across clusters with Tobacco, Music and Sound Recording, and Jewelry and Precious Metals having the highest scores. Most clusters have an average WCR score greater than the maximum BCR score (i.e., WCR Rank of one).

5.1. Cluster examples

The clusters in the BCD include a diverse array of industries, often from different three-digit NAICS but in some cases from the same three-digit NAICS. Some clusters include a mix of service and manufacturing industries, while others are service or manufacturing focused. Clusters also differ in the types of corrections undertaken in Step 5, from clusters with no reallocation of industries from the original algorithmic placement, to clusters that are combined or partitioned from the original cluster in C^*. We provide here detailed descriptions of six clusters that illustrate these cases.

5.1.1. Aerospace vehicles and defense cluster

This cluster includes establishments that manufacture aircraft, space vehicles, guided missiles and related parts (Table 9). It was systematically generated by Steps 1–3 of our algorithm, and contains seven industries in two different three-digit NAICS (336 and 334). These industries are categorized into three subclusters: Aircraft, Missiles and Space Vehicles, and Search and Navigation Equipment. All the industries fit best in this cluster when compared with being placed in any of the other 50 clusters (i.e., the rank for each industry based on WCR_{ic} score equals 1). The industry with the highest WCR score is Aircraft Manufacturing (NAICS 315999), suggesting that this is a focal activity with relevant links to the other industries within the cluster. The seven industries that constitute the Aerospace Vehicles and Defense cluster are geographically mapped to measure the specialization of U.S. regions in the cluster. For example, Figure 1 shows the EA that are highly specialized in this cluster, including, among many others, Seattle-Tacoma-Olympia, WA; Wichita-Winfield, KS and Dallas-Fort Worth, TX.

5.1.2. Oil and gas production and transportation cluster

This cluster includes firms involved in locating, extracting, refining and transporting oil and gas (Table 9). It contains 12 industries that were originally divided into two clusters in C^*, but we combined them into a single cluster because of their high BCR score and conceptual similarity. The mix of manufacturing and service industries in this cluster (and in ten other clusters) contrasts with other papers that classify industries in service and manufacturing as unrelated (e.g., Frenken et al., 2007).

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Table 9. Cluster description

NAICS	NAICS name	Sub-cluster name	WCR$_{ic}$	
			Rank (1 = best)	Score
Aerospace vehicles and defense cluster[a]				
336411	Aircraft Mfg	Aircraft	1	3.51
336412	Aircraft Engine & Engine Parts Mfg	Aircraft	1	1.65
336413	Other Aircraft Parts & Auxiliary Equipment Mfg	Aircraft	1	2.21
336414	Guided Missile & Space Vehicle Mfg	Missiles & Space Vehicles	1	2.29
336415	Guided Missile & Space Vehicle Propulsion Unit & Unit Parts Mfg	Missiles & Space Vehicles	1	2.04
336419	Other Guided Missile & Space Vehicle Parts & Auxiliary Equipment Mfg	Missiles & Space Vehicles	1	1.98
334511	Search, Detection, Navigation, Guidance, Aeronautical, & Nautical System & Instrument Mfg	Search & Navigation Equipment	1	1.76
Oil and gas production and transportation cluster[b]				
324110[c]	Petroleum Refineries	Petroleum Processing	1	2.57
324199[c]	All Other Petroleum & Coal Products Mfg	Petroleum Processing	1	0.82
213112	Support Activities for Oil & Gas Operations	Support Activities for Oil & Gas Operations	1	2.15
541360	Geophysical Surveying & Mapping Services	Support Activities for Oil & Gas Operations	1	0.79
213111	Drilling Oil & Gas Wells	Drilling Wells	1	0.94
211111	Crude Petroleum & Natural Gas Extraction	Oil & Gas Extraction	1	2.45
211112	Natural Gas Liquid Extraction	Oil & Gas Extraction	1	2.16
333132	Oil & Gas Field Machinery & Equipment Mfg	Oil & Gas Machinery	1	1.14
486110[c]	Pipeline Transportation of Crude Oil	Pipeline Transportation	1	1.28
486210	Pipeline Transportation of Natural Gas	Pipeline Transportation	1	1.09
486910[c]	Pipeline Transportation of Refined Petroleum Products	Pipeline Transportation	2	1.10
486990[c]	All Other Pipeline Transportation	Pipeline Transportation	1	1.09
Insurance services cluster[d]				
524291	Claims Adjusting	Insurance Related Services	1	6.13
524298	All Other Insurance Related Activities	Insurance Related Services	1	6.16
524113	Direct Life Insurance Carriers	Insurance Carriers	1	3.91
524114	Direct Health & Medical Insurance Carriers	Insurance Carriers	1	3.90
524126	Direct Property & Casualty Insurance Carriers	Insurance Carriers	1	3.90

(continued)

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Table 9. Continued

NAICS	NAICS name	Sub-cluster name	WCR_{ic}	
			Rank (1 = best)	Score
524127	Direct Title Insurance Carriers	Insurance Carriers	1	3.53
524128	Other Direct Insurance Carriers	Insurance Carriers	1	3.30
524130	Reinsurance Carriers	Reinsurance Carriers	1	3.75
Medical devices cluster[e]				
333314[f]	Optical Instrument & Lens Mfg	Optical Instruments & Ophthalmic Goods	1	1.97
339115	Ophthalmic Goods Mfg	Optical Instruments & Ophthalmic Goods	1	2.48
339112	Surgical & Medical Instrument Mfg	Surgical & Dental Instruments & Supplies	1	2.18
339113	Surgical Appliance & Supplies Mfg	Surgical & Dental Instruments & Supplies	1	2.34
339114	Dental Equipment & Supplies Mfg	Surgical & Dental Instruments & Supplies	1	1.61
Lighting and electrical equipment cluster[g]				
335110	Electric Lamp Bulb & Part Mfg	Lighting Fixtures & Parts	1	1.33
335121	Residential Electric Lighting Fixture Mfg	Lighting Fixtures & Parts	1	1.76
335122	Commercial, Industrial, & Institutional Electric Lighting Fixture Mfg	Lighting Fixtures & Parts	1	1.77
335129	Other Lighting Equipment Mfg	Lighting Fixtures & Parts	1	1.68
335311	Power, Distribution, & Specialty Transformer Mfg	Electrical Equipment	1	1.62
335312	Motor & Generator Mfg	Electrical Equipment	1	1.30
335313	Switchgear & Switchboard Apparatus Mfg	Electrical Equipment	1	1.48
335314	Relay & Industrial Control Mfg	Electrical Equipment	1	1.64
335921	Fiber Optic Cable Mfg	Electrical Components	1	1.39
335929	Other Communication & Energy Wire Mfg	Electrical Components	1	1.61
335931	Current-Carrying Wiring Device Mfg	Electrical Components	1	1.67
335932	Noncurrent-Carrying Wiring Device Mfg	Electrical Components	1	1.23
335991	Carbon & Graphite Product Mfg	Electrical Components	1	1.15
335999[h]	All Other Miscellaneous Electrical Equipment & Component Mfg	Electrical Components	1	1.65
335911	Storage Battery Mfg	Storage Batteries	1	1.00

(continued)

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Table 9. Continued

NAICS	NAICS name	Sub-cluster name	WCR_{ic}	
			Rank (1 = best)	Score
Recreational and small electric goods cluster[i]				
337920	Blind & Shade Mfg	Recreational & Decorative Goods	3	1.37
339992	Musical Instrument Mfg	Recreational & Decorative Goods	3	1.60
339993	Fastener, Button, Needle, & Pin Mfg	Recreational & Decorative Goods	3	1.38
339999	All Other Miscellaneous Mfg	Recreational & Decorative Goods	3	1.60
339931	Doll & Stuffed Toy Mfg	Games, Toys, & Children's Vehicles	2	1.17
339932	Game, Toy, & Children's Vehicle Mfg	Games, Toys, & Children's Vehicles	1	1.44
336991[j]	Motorcycle, Bicycle, & Parts Mfg	Motorcycles & Bicycles	1	0.94
339920	Sporting & Athletic Goods Mfg	Sporting & Athletic Goods	3	1.25
333313[j]	Office Machinery Mfg	Office Supplies	4	0.99
333315[j]	Photographic & Photocopying Equipment Mfg	Office Supplies	3	1.34
339941	Pen & Mechanical Pencil Mfg	Office Supplies	2	1.43
339942	Lead Pencil & Art Good Mfg	Office Supplies	1	1.50
339943	Marking Device Mfg	Office Supplies	1	1.61
339944	Carbon Paper & Inked Ribbon Mfg	Office Supplies	1	1.11
335211[j]	Electric Housewares & Household Fan Mfg	Electric Housewares	1	0.78

[a] Establishments in this cluster manufacture aircraft, space vehicles, guided missiles and related parts. It also contains firms that manufacture the necessary search and navigation equipment used by these products.

[b] This cluster includes firms involved in locating, extracting, refining and transporting oil and gas. This includes companies that manufacture the equipment necessary to extract oil and gas, as well as companies that provide support services for oil and gas operations and pipeline transport.

[c] Industry that was originally part of a cluster that was combined with another one to produce this cluster.

[d] This cluster consists of firms providing a range of insurance types, as well as support services such as reinsurance and claims adjustment.

[e] Establishments in this cluster primarily manufacture surgical, dental and optical instruments and supplies.

[f] Marginal industry outliers reallocated from other cluster (see online Appendix B).

[g] This cluster contains firms involved in the manufacture of electrical equipment and electronic components. The companies in this cluster manufacture wire for communications, wiring devices, fiber-optic cables, switchboards, lighting fixtures, motors, transformers and related products.

[h] Marginal industry outliers reallocated from other clusters (see online Appendix B).

[i] This cluster contains establishments that manufacture end-use products for recreational and decorative purposes. These products include games, toys, bicycles, motorcycles, musical instruments, sporting goods, art supplies, office supplies, shades and home accessories. This cluster also incorporates firms that produce small, simple electric goods like hairdryers, fans and office machinery.

[j] Marginal industry outliers reallocated from other clusters (see online Appendix B).

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

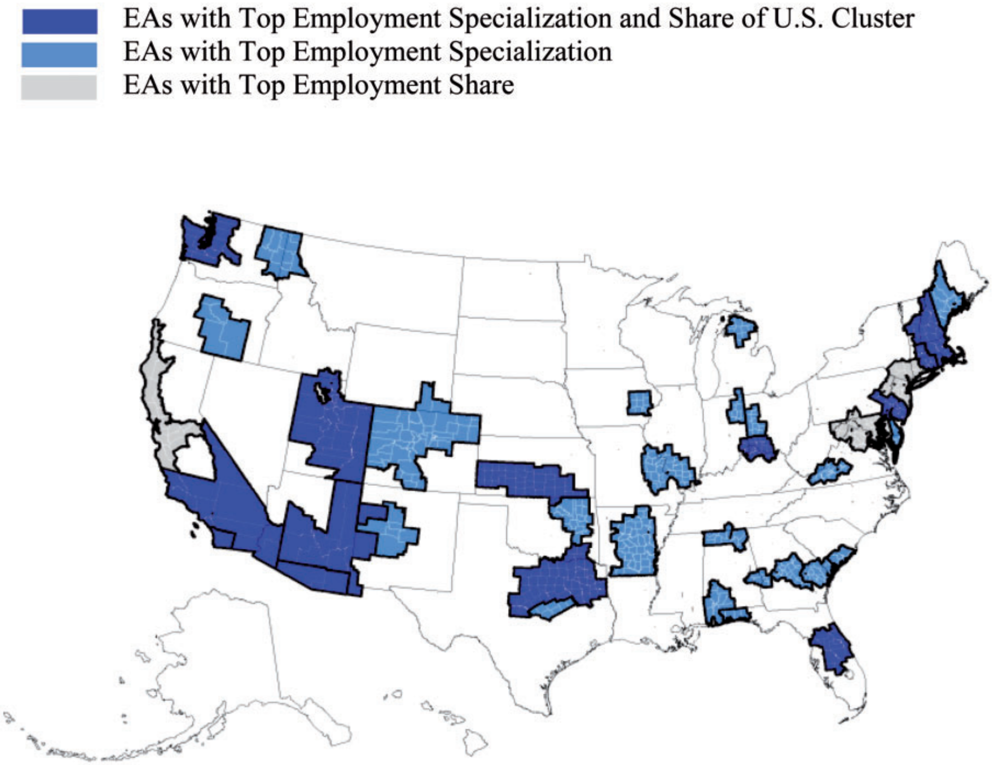
EAs with Top Employment Specialization and Share of U.S. Cluster
EAs with Top Employment Specialization
EAs with Top Employment Share

Figure 1. Top regional aerospace vehicles and defense clusters in 2010 (EAs; *C***).

Notes: EAs with top employment specialization in a cluster meet these criteria: Location quotient (LQ) of cluster employment must be greater than 75th percentile when measured across all EAs with non-zero employment in the cluster. Secondary criteria to differentiate marginal cases: LQ of cluster employment greater than 1.0, share of national cluster employment and share of national cluster establishments greater than 25th percentile. EAs with top employment share in a cluster meet this criterion: share of national cluster employment must be greater than 90th percentile when measured across all EAs with non-zero employment in the cluster. EAs with top employment specialization and share meet all of the above criteria.

5.1.3. Insurance services cluster

The firms in this cluster provide a range of insurance products as well as insurance support services (Table 9). It contains eight service industries and is one of three clusters that is exactly equivalent to a three-digit NAICS group (524). Insurance-related services (e.g., Claims Adjusting, NAICS 524291; All Other Insurance Related Activities, NAICS 524298) is a focal part of the cluster, supporting a broad range of insurance types.

5.1.4. Medical devices; lighting and electrical equipment; and recreational and small electric goods clusters

The algorithm originally grouped all three clusters into one large cluster. Using expert judgment, we separated the overall cluster into three clusters based primarily on the

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

industry definitions (Table 9). We checked that the $\mathrm{WCR_c}$ of each of the new clusters improved or changed minimally relative to the score of the initial larger cluster, and that they each had a $\mathrm{WCR_c}$ rank of 1 (Table 8). We also examined the geographic concentration of the three cluster categories across EA regions as an additional criterion, and concluded that they have different geographic concentration patterns.

The three clusters share skills but they are conceptually different. The Medical Devices cluster (Table 9) consists of firms that manufacture different medical devices and supplies. It has five industries representing two 3-digit NAICS codes (333 and 339). Firms in the Lighting and Electrical Equipment cluster (Table 9) are not involved in manufacturing medical devices, but do manufacture other electrical equipment and electronic components. It is a larger cluster, consisting of 15 industries representing one 3-digit NAICS code (335). Finally, firms in the Recreational and Small Electric Goods cluster (Table 9) are related to those in the Lighting and Electrical Equipment cluster, but the focus is different. These establishments manufacture end-use products for recreational, decorative and office purposes. There are 15 industries that represent five different 3-digit NAICS codes.

6. Comparison of our *C*** and existing sets of cluster definitions

Using the clustering algorithm, we can assess the relative performance of our proposed set of BCD *C*** and other existing sets of cluster definitions: NAICS-3, Porter (2003), and Feser (2005). We use the mutually exclusive sets of cluster definitions offered by Porter (2003) and Feser (2005) since the differences are more readily apparent when each industry is assigned to one cluster, and their overlapping clusters rely on having well-defined mutually exclusive clusters. Before providing the details of the comparisons, it is important to clarify that the number of industries and the number of groups are different across the four sets. Our *C*** includes 778 industries and 51 clusters, NAICS-3 grouping includes 778 industries and 66 clusters, Porter's (2003) set includes 685 industries and 41 traded clusters, and Feser's (2005) set includes 910 industries and 44 clusters (excluding farming and a few other industries due to data limitations).[17]

The VSs for the selected sets of cluster definitions are presented in Table 10. Our analysis shows that a definition based on grouping industries with the same three-digit NAICS code performs relatively poorly when trying to capture a broad set of industry interdependencies. The overall VS of NAICS-3 is the lowest (58) not only among the existing sets, but also among all the cluster configurations (*C*s) generated by the algorithm. The NAICS-3 groupings do poorly in most sub-scores except for occupational linkages. The low VSs are perhaps not surprising, since the industry code system groups industries based on similarities in products or services, not based on broader inter-industry complementarities. This suggests that many studies that classify industries from different parts of the industry code as unrelated may fail to capture relevant inter-industry linkages.

[17] One of the clusters in Feser's (2005) set is Farming, which is excluded from the CBP data; therefore, we focus the analysis on the other 44 clusters and 910 industries (out of 969 industries) that we can bridge into 2007 NAICS codes and CBP data. The larger number of industries in Feser (2005) is in part due to the inclusion of local health services.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Table 10. VSs (sub-scores) for selected sets of cluster definitions

	BCD (*C***) 51 clusters, 778 industries			Three-digit NAICS 66 clusters, 778 industries			Porter (2003) 41 clusters, 685 industries			Feser (2005) 44 clusters, 910 industries		
	VS Cluster	VS Industry	VS[a]	VS Cluster	VS Industry	VS	VS Cluster	VS Industry	VS	VS Cluster	VS Industry	VS
VS Scores	82	73	78	56	59	58	76	68	72	70	67	69
VS Rank	26	1	12	717	709	717	264	177	229	533	250	430
VS sub-scores												
$VS^{LC\text{-}Emp}$	70	63	66	48	46	47	68	66	67	48	59	53
$VS^{LC\text{-}Est}$	73	63	68	51	49	50	67	62	65	61	58	59
VS^{IO}	87	74	81	39	44	41	76	66	71	70	61	66
VS^{Occ}	98	93	95	88	97	92	94	80	87	99	92	95

[a]VS is the average of VS-Cluster and VS-Industry. The VS rankings are computed across the 713 *C*s and the four sets included in the table (1 = best, 717 = worst). Each set of clusters contains mutually exclusive groups.

In Table 10, we find that *C*** scores better in capturing a broad set of industry interdependencies than any of the other sets of clusters. *C*** receives the highest VS (and highest *VS-Cluster* and *VS-Industry* scores) with a VS value of 78 when compared with 72 for Porter (2003), 69 for Feser (2005) and 58 for NAICS-3. *C*** also seems to perform well in particular inter-industry measures. Specifically, *C*** scores better or the same as the three other sets in the validation sub-scores for pairwise industry co-location of the number of establishments ($VS^{LC\text{-}Est}$), IO links (VS^{IO}) and shared labor occupations (VS^{Occ}). The Porter (2003) set scores highest in the validation sub-score for pairwise industry co-location of employment ($VS^{LC\text{-}Emp}$), but holds only a marginal difference from the *C*** score (67 versus 66). The Feser (2005) set scores similarly well than *C*** in the Occ sub-score and relatively worse than *C*** in the LC and IO sub-scores. The lower validation sub-score for IO links is likely due to the fact that Feser's (2005) set of clusters are generated using a particular *indirect IO* link measure that is different from the one used in this article (i.e., his measure captures the percent of meaningful suppliers and buyers in common for a pair of industries rather than the direct selling and buying to and from each other). Overall, the findings suggest that the clusters in our proposed set of BCD contain industries that are meaningfully related (as measured by our LC, IO, and Occ matrices), and may facilitate externalities of various types.

We can also analyze the industry overlap between the clusters in *C*** and the clusters in the other three sets of cluster definitions to evaluate whether different clustering methods tend to generate similar clusters. In each case, we examine the overlap for the sub-set of industries in common with *C***. For example, *C***and NAICS-3 definitions have 778 industries in common, which belong to 51 clusters in *C*** and 66 clusters in NAICS-3. We then assess if the industries are grouped in similar clusters using our OS (see Section 3.4). We compute the OS in each direction (i.e., *C*** to NAICS-3 and NAICS-3 to *C***) and then take the average in both directions. The findings are presented in Table 11. The average industry overlap is 61% between *C*** and NAICS-3, which means that on average a pair of clusters shares 61% of their industries. Only three clusters are identical in both sets (100% overlap): Environmental Services

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Table 11. Overlap between C^{**} and other sets of cluster definitions

Overlap score	C^{**} and three-digit NAICS	C^{**} and Porter (2003)	C^{**} and Feser (2005)
$C^{**} \rightarrow C$	66%	55%	54%
$C^{**} \leftarrow C$	56%	58%	54%
$C^{**} \leftrightarrow C$ (Average)	61%	57%	54%
Industries in C^{**} and C	778	671	734
No. clusters in C^{**}, C	51, 66	47, 41	49, 44

Note: We compute the OS for the sub-set of industries in common with C^{**}.

(NAICS-562); Insurance Services (NAICS-524; Table 9) and Paper and Packaging (NAICS-322).[18] Similarly, the average cluster overlap is 57% between C^{**} and Porter's (2003) set, and 54% between $C^{**}$and Feser's (2005) set. While there is overlap between sets, the scores indicate that the clusters in the BCD are meaningfully different from those in the other existing sets.

7. Research and policy applications for the BCD

What constitutes a good set of cluster definitions depends on the particular research or policy question. We implement our clustering algorithm to offer a set of BCD that captures numerous inter-industry linkages and could facilitate externalities of various types (e.g., skills, supply, demand and others).

For research questions that focus on a particular type of inter-industry link (e.g., occupational links and potential labor market pooling benefits), groupings that better capture that specific link may be preferred. For example, if the policy goal is to promote training and skills that could be shared by industries with similar labor needs, cluster configurations that best capture occupational links will be useful.

However, if the goal is to promote multiple complementarities across industries, we believe our BCD may be more useful. Industries within our cluster categories are highly related based on a mix of various links. Some industries in a cluster may have strong skill links, while other industries may be closely related by IO, technology and/or other types of links. Policies that focus on improving a specific link for a sub-set of industries within the cluster (e.g., training and skills shared by some industries) can facilitate complementarities of many types among all industries in the cluster.

We believe our proposed BCD is particularly important for studying the economic development of regions. We map the finalized set of cluster definitions into different regional units (Metropolitan Statistical Areas (MSAs), EAs and States) over time, creating a regional cluster dataset that allows for a systematic comparison of the cluster composition of regions using different metrics (e.g., employment, specialization, wages and number of establishments). For example, Figure 1 shows the top regional Aerospace Vehicles and Defense clusters across EAs in 2010. This database is publicly available at the U.S. Cluster Mapping Project website (and the BCD can be accessed at the online Appendix B).

[18] The number of different three-digit NAICS in a cluster in C^{**} is on average 2.7.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

For a particular cluster category, we can assess its presence in a region and examine what industries are under-represented (or non-existing) in the region when compared with the national cluster or when compared with similar clusters in other regions. These comparisons can then be the focus of key research and policy actions. To design policies that help a regional cluster, we need to examine why certain industries are under-represented or under-performing. Several explanations are plausible: the lack of skills, inputs, technology, sophisticated demand and/or institutions for collaboration. Other methods that focus on examining region-specific links among firms and individuals in clusters could complement our BCD and offer important insights into the mechanism at play in a particular regional cluster.

The ability to compare a cluster across regions can also facilitate the evaluation of regional cluster policies (Feldman et al. 2012). For example, we could identify a few regional clusters that look very similar in a base year in terms of various cluster attributes (size, specialization, number of firms, industry composition, etc.), but where one cluster is the target of a relevant investment (say a private-public grant) and the others are not. We can then assess if over time we observe differences in the composition and performance of the treated regional cluster relative to the others. In addition, researchers can use our clustering method and the BCD-based database to examine relevant questions on the role of clusters in the performance of regions and firms.

8. Conclusion

In order to compete more effectively, regions need to understand their cluster strengths when compared with those of other regions. To make this comparison, a set of regionally comparable cluster definitions that marks the industry boundaries of each cluster is necessary. This article responds to this need by providing a clustering methodology to generate and assess sets of comparable cluster definitions (i.e., the industries that constitute a cluster are the same for all regions). In our algorithm, each cluster configuration is generated by a clustering function that uses as inputs a particular inter-industry similarity matrix and well-specified parameter choices. The clustering algorithm provides scores that allows us to identify the candidate configuration that best captures multiple types of inter-industry links. The methodology concludes with a correction of anomalies of the individual clusters in the most promising configuration to determine our finalized set of cluster definitions.

Using U.S. data, we implement the clustering algorithm to generate a transparent set of BCD that captures many inter-industry interdependencies. The proposed definitions use measures of inter-industry linkages based on the co-location patterns of employment and establishments, IO linkages and shared labor occupations. The BCD contains 51 clusters that can be mapped consistently into U.S. regions to create a regional cluster database.

With an updateable algorithm for defining and assessing alternative cluster definitions, a number of extensions are possible. First, we can add additional inter-industry similarity matrices as they become available (e.g., specific measures of knowledge linkages or labor flows among industries) to generate improved cluster definitions.

Second, while the analysis here focuses on mutually exclusive clusters, the methodology also provides scores of the relatedness between any pair of clusters and

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

between any industry and any cluster. These scores are based on various inter-industry linkage measures. Thus, we can assess which mutually exclusive clusters are meaningfully related (e.g., industries in the Financial Services cluster and in the Insurance Services cluster are highly related). We can also develop overlapping cluster definitions by adding secondary industries that are highly related to the industries that constitute each (mutually exclusive) cluster (Porter, 2003; Feser, 2005). Defining measures of the relatedness among clusters is important since economies of agglomeration arise across related clusters as well as within individual clusters (Delgado et al., 2014b).

Third, our clustering method can be applied to other countries using their specific data. Defining clusters is best undertaken using data from large and diverse economies with numerous highly integrated regions. Since the USA is a large and diverse economy, the U.S. BCD are a good starting point, especially for economies that lack the data needed to implement the clustering methodology. However, there are some limitations on using the BCD in other economies that are weighted toward economic activities that are less prevalent in the USA (e.g., ship building) or that are not well captured by U.S. data (e.g., farming). They may also be less useful in countries with a lower level of technological development, but here, our definitions offer important insights into how clusters could form with reduced internal barriers to trade and technological improvements. Finally, the BCD will be especially useful for countries with an industry code schema similar in detail to the one in the USA (e.g., Mexico and Canada). It can also be applied with higher aggregation to a large set of countries through matching the U.S. NAICS code to the U.N. International Standard Industrial Classification (ISIC). Definitions based on ISIC would facilitate an examination of the trade and foreign direct investment links of clusters across countries (e.g., Bathelt and Li, 2013; Delgado et al., 2013).

A fourth extension is the further examination of local industries (e.g., retail industries, hospitals) and their linkages with traded regional clusters. Our clustering analysis excludes local industries because they are geographically dispersed across regions, and focus on serving a region's population. However, within a region certain local industries can geographically concentrate (e.g., some retail industries can be co-located nearby other local and traded industries), which has implications for policy.

Fifth, the algorithm can be used to track the evolution of the industry boundaries of clusters over time. For example, while IT and analytical instruments industries are highly related today, they may have been less complementary a decade or two ago. The emergence and evolution of clusters have not been widely studied due to lack of data (Swann, 1998; Porter, 1998; Bathelt and Boggs, 2003; Klepper, 2010). However, an understanding of cluster emergence and relatedness could have wide-ranging implications for forward-looking regional strategy.

Another area for future research is the development of methods to adapt cluster definitions to specific regions (i.e., the industry boundaries of a cluster can sometimes vary by region). For example, regional IO tables could be used to measure region-specific buyer–supplier linkages in clusters.

Finally, our BCD can be mapped into continuous spatial units as well as administrative units (e.g., MSAs, EAs and States). We could then analyze the micro-geography of clusters within (and across) jurisdictions (e.g., Duranton and Overman, 2005; Kerr and Kominers, 2010; Alcacer and Zhao, 2013). For example, in a particular EA (e.g., Los Angeles, CA), we could assess whether a cluster is geographically separated in distant parts within the region or whether the whole cluster is closely co-located.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Understanding the micro-geography of clusters can help inform policies to facilitate the connectivity of firms and supportive institutions within clusters.

The BCD, combined with other available data sources, can be used to greatly inform economic development. For example, using the BCD and other data sources, the U.S. Cluster Mapping Project has created a regional cluster dataset together with multiple regional performance and business environment indicators (e.g., employment, specialization, wages and number of establishments). The Project provides a powerful tool for researchers and policymakers, and offers a new interactive tool for practitioners and firms looking to identify opportunities in regions and design cluster-based regional economic development policies. The tool also maps the cluster composition of regions to encourage connections previously not identified.

Supplementary material

Supplementary data for this paper are available at *Journal of Economic Geography* online.

Acknowledgments

This article is a significantly revised version of the National Bureau of Economic Research (NBER) Working Paper 20375. This article is part of the U.S. Cluster Mapping Project (USCMP), an initiative supported by the Economic Development Administration of the U.S. Department of Commerce, and hosted at the Harvard Business School's Institute of Strategy and Competitiveness. The cluster definitions, underlying data, and additional resources can be accessed at the USCMP (www.clustermapping.us). The authors acknowledge the excellent comments of the editor and two anonymous reviewers. They thank Bill Simpson, Xiang Ao, Rich Bryden, Jorge Guzman, Sam Zyontz and Jane Wu for their valuable assistance with the analysis. They also thank Ed Feser, Frank Neffke, Juan Alcacer, Bill Kerr, Fiona Murray, Christian Ketels, Maryann Feldman, Ram Mudambi, James Delaney, Sergiy Protsiv, Brandon Stewart, Muhammed Yildirim, Sarah Jane Maxted, and the participants in the Industry Studies Association Conference, NBER Productivity Seminar, Temple University Seminar, and the Symposium on the Use of Innovative Datasets for Regional Economic Research at George Washington University.

References

Alcacer, J., Chung, W. (2014) Location strategies for agglomeration economies. *Strategic Management Journal* 35: 1749–1761.

Alcacer, J., Zhao, M. (2013) Zooming. In *A Practical Manual for Identifying Geographic Clusters*. Harvard Business School Working Paper, No. 14-042, November 2013

Bathelt, H., Boggs, J. S. (2003) Toward a reconceptualization of regional development Paths: Is Leipzig's media cluster a continuation of or a rupture with the past? *Economic Geography* 79: 265–293.

Bathelt, H., Li, P. F. (2013) Global cluster networks—foreign direct investment flows from Canada to China. *Journal of Economic Geography* 13: 3–14.

Bathelt, H., Malmberg, A., Maskell, P. (2004) Clusters and knowledge: Local buzz, global pipelines, and the process of knowledge creation. *Progress in Human Geography* 28: 31–56.

Bloom, N., Schankerman, M., Van Reenen, J. (2012) Identifying technology spillovers and product market rivalry. *Econometrica* 81: 1347–1393.

Bresnahan, T., Gambardella, A. (eds) (2004) *Building High-Tech Clusters. Silicon Valley and Beyond*. New York: Cambridge University Press.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Cortright, J. (2006) Making sense of clusters: Regional competitiveness and economic development. The Brookings Institution Metropolitan Policy Program. Available online at: http://www.brookings.edu/reports/2006/03cities_cortright.aspx [Accessed May 2015].

Cortright, J. (2010) *The Athletic and Outdoor Industry Cluster: A White Paper*. Portland: Impresa Economics.

Delgado, M., Bryden, R., Zyontz, S. (2014a) Categorization of traded and local industries in the U.S. economy. Mimeo. Available online at: http://www.clustermapping.us/ [Accessed May 2015].

Delgado, M., Kyle, M., McGahan, A. M. (2013) Intellectual property protection and the geography of trade. *Journal of Industrial Economics* 61: 733–762.

Delgado, M., Porter, M. E., Stern, S. (2010) Clusters and entrepreneurship. *Journal of Economic Geography* 10: 495–518.

Delgado, M., Porter, M. E., Stern, S. (2014b) Clusters, convergence, and economic performance. *Research Policy* 43: 1785–1799.

Duranton, G., Overman, H. G. (2005) Testing for localization using micro-geographic data. *Review of Economic Studies* 72: 1077–1106.

Ellison, G., Glaeser, E. (1997) Geographic concentration in U.S. manufacturing industries: A Dartboard approach. *Journal of Political Economy*, 105: 889–927.

Ellison, G., Glaeser, E., Kerr, W. (2010) What causes industry agglomeration? Evidence from coagglomeration patterns. *The American Economic Review* 100: 1195–1213.

Everitt, B. S., Landau, S., Leese, M., Stahl, D. (2011) *Cluster Analysis*. 5th edn. Chichester, UK: John Wiley & Sons, Ltd.

Feldman, M. P., Audretsch, D. (1999) Innovation in cities: Science-based diversity, specialization, and localized competition. *European Economic Review* 43: 409–429.

Feldman, M. P., Francis, J., Bercovitz, J. (2005) Creating a cluster while building a firm: Entrepreneurs and the formation of industrial clusters. *Regional Studies* 39: 129–141.

Feldman, M. P., Reed, A. G., Lanahan, L., McLaurin, G., Nelson, K., Reamer, A. (2012) Innovative data sources for economic analysis. e-book

Feser, E. J. (2005) Benchmark value chain industry clusters for applied regional research. Regional Economics Applications Laboratory, University of Illinois at Urbana-Champaign

Feser, E. J., Bergman, E. M. (2000) National industry cluster templates: A framework for applied regional cluster analysis. *Regional Studies* 34: 1–19.

Feser, E. J., Renski, H., Goldstein, H. (2008) Clusters and economic development outcomes. *Economic Development Quarterly* 22: 324–344.

Feser, E. J., Renski, H., Koo, J. (2009) Regional cluster analysis with interindustry benchmarks. In Goetz S. J., Deller S. C., and Harris T. R. (eds) *Targeting Regional Economic Development*, pp. 213–238. London: Routledge.

Frenken, K., Van Oort, F. G., Verburg, T. (2007) Related variety, unrelated variety, and regional economic growth. *Regional Studies* 41: 685–697.

Glaeser, E. L., Kerr, W. R. (2009) Local industrial conditions and entrepreneurship: How much of the spatial distribution can we explain? *Journal of Economics and Management Strategy* 18: 623–663.

Grimmer, J., King, G. (2011) General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences* 108: 2643–2650.

Hidalgo, C. A., Klinger, B., Barabasi, A. L., Hausmann, R. (2007) The product space conditions the development of nations. *Science* 317: 482–487.

Hill, E. W., Brennan, J. F. (2000) A methodology for identifying the drivers of industrial clusters: The foundation of regional competitive advantage. *Economic Development Quarterly* 14: 65–96.

Jaffe, A., Trajtenberg, M., Henderson, R. (1993) Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 108: 577–598.

Johnson, K. P., Kort, J. R. (2004) *2004 redefinition of the BEA economic areas*. Available online at: http://www.bea.gov/scb/pdf/2004/11November/1104Econ-Areas.pdf [Accessed May 2015].

Klepper, S. (2010) The origin and growth of industry clusters: The making of silicon valley and detroit. *Journal of Urban Economics* 67: 15–32.

Kerr, W., Kominers, S. (2010) Agglomerative forces and cluster shapes. NBER Working Paper 16639, NBER.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Koo, J. (2005a) How to analyze the regional economy with occupation data. *Economic Development Quarterly* 19: 356–372.

Koo, J. (2005b) Knowledge-based industry clusters: Evidenced by geographical patterns of patents in manufacturing. *Urban Studies* 42: 1487–1505.

Krugman, P. (1991) Increasing returns and economic geography. *Journal of Political Economy* 99: 483–499.

Lorenzen, M., Mudambi, R. (2013) Clusters, connectivity, and catch-up: Bollywood and Bangalore in the global economy. *Journal of Economic Geography* 13: 501–534.

Marshall, A. (1920) *Principles of Economics*. London: MacMillan.

Markusen, A. (1996) Sticky places in slippery space: A typology of industrial districts. *Economic Geography* 72: 293–313.

Maskell, P., Malmberg, A. (2007) Myopia, knowledge development, and cluster evolution. *Journal of Economic Geography* 7: 603–618.

Neffke, F., Henning, M. (2013) Skill-relatedness and firm diversification. *Strategic Management Journal* 34: 297–316.

Neffke, F., Henning, M., Boschma, R. (2011) How do regions diversify over time? Industry relatedness and the development of new growth paths in regions. *Economic Geography* 87: 237–265.

Porter, M. E. (1990) *The Competitive Advantage of Nations*. New York: Free Press.

Porter, M. E. (1998) Clusters and competition: New agendas for companies, governments, and institutions. In Porter M. E. (ed.) *On Competition*, pp. 197–299. Boston: Harvard Business School Press.

Porter, M. E. (2003) The economic performance of regions. *Regional Studies* 37: 549–578.

Porter, M. E., Ramirez-Vallejo, J. (2013) *The New Carolina Initiative*. Harvard Business School, N9-713–462

Rosenthal, S. S., Strange, W. C. (2003) Geography, industrial organization, and agglomeration. *Review of Economics and Statistics* 85: 377–393.

Rosenthal, S. S., Strange, W. C. (2004) Evidence on the nature and sources of agglomeration economies. In Henderson J. V. and Thisse J. F. (eds) *Handbook of Regional and Urban Economics*, vol. 4, pp. 2119–2171. Amsterdam: Elsevier North-Holland.

Saxenian, A. (1994) *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*, Cambridge, MA: Harvard University.

Scherer, F. M. (1982) Inter-industry technology flows and productivity growth. *The Review of Economics and Statistics* 64: 627–634.

Sorenson, O., Audia, P. G. (2000) The social structure of entrepreneurial activity: Geographic concentration of footwear production in the United States, 1940–1989. *American Journal of Sociology* 106: 424–462.

Storper M. (1995) The resurgence of regional economies, ten years later: The region as a Nexus of untraded interdependencies. *European Urban and Regional Studies* 2: 191–221.

Storper, M., Venables, T. (2004) Buzz: Face-to-face contact and the urban economy. *Journal of Economic Geography* 4: 351–370.

Swann, P. (1992) *The Dynamics of Industrial Clusters*. Oxford: Oxford University Press.

Swann, P. (1998) Clusters in the U.S. computing industry. In Swann P., Prevezer M., and Stout D. (eds) *The Dynamics of Industrial Clustering: International Comparisons in Computing and Biotechnology*, pp. 76–105. Oxford: Oxford University Press.

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015

Appendix

Table A1. Similarity matrices used to generate cluster configurations *C*s

Similarity matrix	Number of *C*s	Type of M_{ij}	Definition
LC-Emp	155	Unidimensional	LC of employment [−1, 1]
LC-Est	155	Unidimensional	LC of establishments [−1, 1]
IO	31	Unidimensional	IO link [0, 1]
Occ	93	Unidimensional	Labor occupation link [−1, 1]
COI	31	Unidimensional	Co-agglomeration index
LC-IO-Occ	31	Multidimensional	Average of (standardized) LC-Emp, LC-Est, IO, Occ
COI-IO-Occ	31	Multidimensional	Average of (standardized) COI, IO, Occ
LC	31	Multidimensional	Average of LC-Emp, LC-Est
IO-Occ	31	Multidimensional	Average of (standardized) IO, Occ
LC-Emp-IO	31	Multidimensional	Average of (standardized) LC-Emp, IO
LC-Est-IO	31	Multidimensional	Average of (standardized) LC-Est, IO
LC-Emp-Occ	31	Multidimensional	Average of LC-Emp, Occ
LC-Est-Occ	31	Multidimensional	Average of LC-Est, Occ

Downloaded from http://joeg.oxfordjournals.org/ at University of California, Santa Barbara on July 20, 2015