Old and new theories of industry clusters

Edward J Feser, University of North Carolina at Chapel Hill
Old and new theories of industry clusters

Edward J Feser

University of North Carolina at Chapel Hill*

16 April 1998

Abstract. The paper reviews the broad range of theories and ideas that constitute, often implicitly, the logic behind strategic cluster policies. The title of the paper notwithstanding, there is no theory of industry clusters, per se. Even Porter’s (1990) seminal contribution is more a theory of firm competitiveness than clusters. There is, instead, a variety of older and newer theories of 1) the interrelationships between economic actors that clusters describe, and 2) the implications of such interrelationships for economic growth and development. Industry clusters have proven a useful way of characterizing webs of relationships between and among firms and other institutions. Policy makers designing cluster strategies attempt to leverage such relationships in the interest of growth and development objectives. For their part, regional scholars must strive to specify and test clearer hypotheses about the workings and impacts of such relationships in order to verify the efficacy of cluster policies.

1. Introduction

Despite the intense interest in industry clusters in economic development policy making in Europe and North America, there is presently little consensus about the precise meaning of an industry cluster, the dynamics that underlie cluster growth and development, and the policy initiatives that would best build and strengthen clusters. To be sure, there has been movement toward consensus as the literature deepens, with most progress probably occurring in the area of methodologies for identifying and documenting clusters. But the great variety of development strategies that are either nominally or fully based on some notion of industry clusters testifies to a high degree of malleability in the cluster concept (Van der Laan 1997, Marceau 1997). In Europe, Boekholt (1997, p. 1) writes that the “multitude of cluster initiatives has led to a wide spread confusion of what clusters really are, and in what way they differ from related phenomenon, such as industrial districts, technopoles, networks, and industry-research collaborations.” And in a critique of cluster policy in the U.S., Held (1996, p. 249) notes: “Sadly, in the rush by various governments to employ clusters, some fundamental issues have been slighted, including appropriate research methods and even the definition of the cluster itself.”

As Jacobs and de Man (1996) assert, a degree of flexibility in the cluster concept and a variety in policy approaches can be useful. However, Jacobs and
de Man also go on to provide a careful outline of various dimensions of an industrial cluster in an effort to clarify terms, definitions, and policy implications. Clearly, the concern that “if clusters are everything, maybe they are nothing” is one worth addressing.

The experience of two related concepts that are similarly broad in scope is instructive. Harrison (1992) questions whether the theory of new industrial districts is merely “old wine in new bottles,” noting its many similarities to more traditional and well-worn regional theories. In a paper published two decades earlier, Gilmour (1974, p. 336) has this to say about the concept of external scale economies: “. . .this explanation of agglomeration is so intellectually appealing as to cause us to assume it must be right (it probably is). . .It has never been demonstrated that it is completely invalid, but neither has the converse been demonstrated. More than anything else this noteworthy state of affairs reflects the theory’s difficulty of verification.” Of course, Harrison goes on to conclude that new industrial district theory is, in fact, more than just old wine, while Gilmour does attempt to verify the presence of external scale economies. So all is probably not lost with industry clusters either.

This paper provides a brief review of the broad range of theories and ideas that constitute, often implicitly, the logic behind strategic cluster policies. The title of the paper notwithstanding, there is no theory of industry clusters, per se. Even Porter’s (1990) seminal contribution is more a theory of firm competitiveness than clusters (Kaufman et al. 1994). There is, instead, a variety of older and newer theories of 1) the interrelationships between economic actors that clusters describe, and 2) the implications of such interrelationships for economic growth and development. Industry clusters have proven a useful way of characterizing webs of relationships between and among firms and other institutions. Policy makers designing cluster strategies attempt to leverage such relationships in the interest of growth and development objectives. For their part, regional analysts must strive to specify and test clearer hypotheses about the workings and impacts of such relationships in order to verify the efficacy of cluster policies.

Section 2 describes several ways of classifying the wide range of development initiatives related to industry clusters. Policies differ based on varying definitions of clusters, possible levels of analysis, and degree to which clustering constitutes the central focus. Industry cluster principles are often used merely to improve the implementation of traditional development schemes. But in general, cluster-related development policies represent attempts to leverage synergies arising from economic and spatial interdependence between economic actors. Section 3 summarizes major theories of such interdependence in economic and geographic space, as well as models of the link between
interdependence and regional growth. By way of summary, Section 4 summarizes a set of unresolved theoretical questions and their implications for cluster policy.

2. Clusters in development policy

A number of authors have attempted to clarify the range and appropriateness of different cluster policies (Boekholt 1997, Jacobs and De Man 1996, Roelandt et al. 1997, Rosenfeld 1995, 1997). Naturally, much of the focus is on how clusters are defined. Boekholt (1997) demonstrates how different definitions of clusters imply different development strategies. She develops a typology of cluster initiatives based on how such policies define: 1) the types of collaborative links among cluster firms (e.g., simple buyer-supplier relations versus knowledge/technology transfer); 2) the types of constituent firms and actors included in the cluster (e.g., firms only or firms and supporting institutions); 3) the appropriate level of aggregation (e.g., micro versus macro); 4) the position of firms in the value chain (i.e., horizontal, vertical, or lateral); 5) the appropriate spatial level of intervention (local, regional, national, international); and 6) the specific policy mechanisms employed (general business assistance, network brokering, technology transfer, information provision, and so on). The typology is helpful for classifying the fast growing policy applications, from the development of joint facilities for related warehousing and distribution industries in the Hudson Valley, New York (Held 1996), to the reform of telecommunications industry regulations in Austria (Peneder and Warta 1997). The typology also helps clarify the difference between related concepts, such as clusters and networks (see also Rosenfeld 1997).

Roelandt et al. (1997) suggest another typology based on a range of possible levels of analysis (see Table 1). At the national level, clusters are conceived as broad industry groups linked within the overall macro economy. Relevant types of cluster analysis at this level include the study of patterns of industrial specialization and the examination of general innovation processes as well as the adoption and use of more generic production and management technologies (those applicable to a wide range of industries). Clusters at the industry (or meso) level constitute the extended value chains of given end-market products (as revealed through patterns of inter- and intra-industry linkage). Cluster analysis at the meso level involves best-practice benchmarking and studies of cluster-specific technology adoption and innovation processes. Finally, at the firm (or micro) level, clusters are conceived as one or a few linked or related enterprises along with their important specialized suppliers. Micro-level cluster analysis includes needs assessments regarding networking and other types of inter-firm collaboration, ‘chain analysis and chain management’ (e.g.,
Table 1. Value chain approach at different levels of analysis

<table>
<thead>
<tr>
<th>Level of analysis</th>
<th>Cluster concept</th>
<th>Focus of analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>National (macro)</td>
<td>Industry groups’ linkages in overall economic structure</td>
<td>Patterns of specialization in national/regional economy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Innovation and technology upgrade needs in mega-clusters</td>
</tr>
<tr>
<td>Branch or industry (meso)</td>
<td>Inter- and intra-industry linkages in different stages of production chain of a single end product</td>
<td>Industry benchmarking and SWOT (strengths, weaknesses, opportunities, threats) analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Innovation needs</td>
</tr>
<tr>
<td>Firm (micro)</td>
<td>Specialized suppliers around one or a few core enterprises (inter-firm) linkages</td>
<td>Strategic business development needs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Value chain analysis and chain management</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Need for collaborative innovation projects</td>
</tr>
</tbody>
</table>

Source: Roelandt et al. (1997)

examination of buyer-supplier contracting practices), and design of business development programs (perhaps involving marketing, recruitment, and entrepreneurship strategies). Clearly, the type of analysis implies different types of policy intervention, from establishment of framework conditions and general technology policies at the national level to networking initiatives and business development at the micro level.

Another way to distinguish policy applications of the industrial cluster concept is in terms of those that focus specifically on identified clusters (cluster-specific strategies) and those that use cluster-related principles and techniques to inform traditional development policies (cluster-informed strategies). The difference is subtle but important. Figure 1 summarizes the first type of approach for a hypothetical environmental technologies cluster. Under the cluster-specific policy approach, the objective is to encourage the emergence or development of a distinct, identified cluster. The first step is to map out the cluster itself, which, in the case of environmental technologies, includes four major end-market industries (equipment and supplies, sustainable goods, environmental services, and environmental resources), their numerous
Figure 1. Building an environmental technologies cluster: create favorable conditions for emergence and development. Adapted from a diagram presented by C. DeBresson at the OECD Workshop on Cluster Analysis and Cluster-Based Policies in Amsterdam, The Netherlands, 10-11 October 1997. Information on the environmental technologies cluster is drawn from Kirkpatrick and Gavaghan (1996).
specialized suppliers, providers of advanced services, related and supporting public and quasi-public institutions, industry trade groups, and so forth.

But the broad environmental technologies value chain only constitutes the supply side element in a comprehensive cluster strategy. To influence the development of the cluster, policymakers must also consider the demand side, which is made up of the major users of environmental technologies goods (government, private industry, and consumers). Thus, a cluster strategy designed to build or extend the environmental technologies value chain must include supply side elements (many of which are outlined in Table 1, including encouragement of interindustry linkages, technology transfer, networking, improved innovation systems, etc.) as well as demand side initiatives. In the case of environmental technologies, government has a strong role to play in ensuring adequate demand through environmental regulation, enforcement and resource pricing, as well as purchasing (as it adopts monitoring, prevention, and clean-up systems and as it subsidizes sustainable infrastructure). The provision of assistance with compliance as well as tax and spending incentives would encourage demand in private industry, while education, various incentives, and adequate recycling facilities would help boost demand for environmentally sustainable goods among consumers.

The primary characteristic of the cluster-specific approach is the comprehensive attempt to nurture a given value-chain through a range of carefully crafted demand- and supply-side policy interventions. It is not the individual policy initiatives that are particularly unique, but rather the way in which they are formulated and targeted; that is, in a manner that establishes--and mutually reinforces--the conditions for the growth and development of key or promising end-market sectors and their supporting industries. Note also that cluster policies from this perspective may involve many economic interventions that are not development strategies per se, but are instead traditional functions of government (regulation, enforcement, pricing, education). Indeed, one of the hallmarks of the cluster approach is the implementation of holistic and comprehensive development strategies that account for the full range of factors influencing the success of a given sector or set of sectors. It stands to reason that many of these factors will fall outside the purview of the typical set of individual development initiatives and tools.

In contrast to what I have termed the cluster-specific approach, the principal policy objective from cluster-informed perspective is the improved implementation of individual (or isolated) development initiatives. The logic is that particular development schemes or interventions will be more effective if they take account of economic and spatial interdependencies. Consider an example. In 1995, a state technology planning agency commissioned a study of
industry clusters in North Carolina to determine where to best focus modernization programs (Bergman, Feser, and Sweeney 1996). The agency was particularly interested in better diffusing advanced production technologies and practices through extended supplier chains by targeting programs to specific sectors within given chains. If, for example, end-market sectors play an important role in encouraging investment in new technology among core suppliers, the agency could leverage modernization resources by ensuring best practice techniques among end-market sectors, thus initiating a ‘modernization ripple effect’ within the value chain. Alternatively, the agency could use the information to better diagnose the obstacles to technology adoption among given segments of supplier chains. Finally, the organization planned to identify gaps in particular chains that were preventing technology diffusion, limiting networking, or otherwise reducing the advantages of co-operative relationships between local producers. These gaps would be considered possible areas for targeted business development initiatives (entrepreneurship or recruitment). Thus, from the cluster-informed policy perspective, clusters are primarily an analytical device used to improve the efficacy of narrower policy tools or types of policy tools.

All cluster policies, however classified, have two important and related goals: resource targeting and resource leveraging. The analysis of industry clusters in national or regional economies helps policymakers identify appropriate targets for scarce development resources; rather than a scattershot approach to economic development, clusters provide a useful device for strategic planning and investment. That goal is most apparent in cluster-specific initiatives. But more importantly, clusters presumably hold promise for leveraging development resources through the encouragement of synergies, external economies, and increasing returns in the clusters themselves, a goal most apparent in cluster-informed types of strategies. To properly achieve that leveraging goal, however, we require a solid understanding of how clusters work, what the specific types of synergies are, and how (indeed, if) they can be cultivated. Those questions are presumably the subject of cluster theory. Unfortunately, attention to theory in the industry cluster debate has been relatively limited. If there is a (new) ‘cluster theory,’” is more a complex amalgamation of ‘old’ development theories, than a self-contained model of cluster growth and development.

3. Clusters in regional development theory

Although Porter’s (1990) study has been criticized as vague in application (Rouvinen and Ylä-Anttila 1997), it nevertheless represents the most useful starting point for a discussion of how industrial clusters fit into the broader framework of regional development theory. While one objective of this essay is
to demonstrate that the concept of industrial clusters has been around in various guises for some time, the Competitive Advantage of Nations has clearly galvanized policy interest at both the national and regional level. More importantly, Porter provides a holistic framework for summarizing what has become a general trend in regional analysis. This is the study of how interdependence between firms, industries, and public and quasi-public institutions affects innovation and growth in regional agglomerations. More specifically, the regional agglomeration literature is increasingly focused on the social or cultural dimensions of such interdependence, rather than the more traditional economic or technical relations between firms (Hassink 1997, Malmberg and Maskell 1997).

Porter’s ideas are also consistent with recent developments in growth and trade theory that highlight the role of social increasing returns and the consequent likely spatial concentration of industrial activity. Although these basic ideas also have something of a history (as described below), it is notable that economists traditionally unconcerned with spatial questions have begun finding implications for geography in their models (Martin and Sunley 1996). And though Porter is no regional economist either, his own studies of the determinants of economic competitiveness have led him to highlight the role of location.

Porter’s model of national competitiveness—the “diamond”—is by now well-known and warrants only a brief summary here. Porter used industries’ success in international markets as the primary barometer of the competitive strength of a nation. Accordingly, he traced the success of the firm to four major factors: 1) the nature of firm strategy, structure and rivalry in the country, including attitudes toward competition, market institutions, the degree of local competition, and other cultural and historical factors affecting how firms do business with each other, their workers, and the government; 2) factor conditions, or the basic endowments or conditions on which the firm seeks to compete (e.g., cost-related basic factors such as ready supplies of natural resources or inexpensive, unskilled labor versus knowledge and/or technology related advanced factors); 3) demand conditions or the nature of local demand (e.g., the needs and wants of the consumer for foreign and domestic goods as well as the existence of local industrial demand for related intermediate goods); 4) the presence of related and supporting industries, including suppliers and successful competitors (both to stimulate cooperation, the latter to also stimulate rivalry).

Competitive companies must depend, to a degree, on the competitiveness of their intermediate input suppliers, who must depend on the capabilities of their suppliers, and so on, back through all links in the value chain. But such
companies also depend on service providers (management, marketing, financing, legal, etc.), sources of basic and applied R&D (e.g., universities and/or contract research organizations), capital goods suppliers, wholesalers and distributors, and suppliers of trained workers (again, universities and colleges). Even competitors are important, including direct competitors to the company as well as competitors to the company’s suppliers, since their presence maintains pressure to continually upgrade processes and techniques and to seek new opportunities. Competitors also provide opportunities for co-operation in solving joint problems or addressing industry-wide issues (see also Best 1990).

Thus, the success of an individual company may be partly traced to the size, depth, and nature of the cluster of related and supporting enterprises—both private and public—of which it is a part. Much of Porter’s analysis, which turns on the findings from an impressive set of national case studies, focuses on outlining the basic conditions determining cluster competitiveness. His framework leads naturally to a focus on end-market sectors as the point of departure for studying clusters. But such end-market industries should not be studied in isolation; the critical function of interdependence in the process of economic growth and change—not just in terms of how it has traditionally been viewed, i.e., in technical or input-output terms—is the guiding principle in his study.

The dynamics that presumably characterize industry clusters need not be localized in scope, though Porter argued that clusters tend to be geographically concentrated. Thus we have two core dimensions of the cluster concept: economic and geographic (see Figure 2). Some end-market sectors (e.g., automobiles) may be both economically and geographically clustered (quadrant I in Figure 2). Others may be more or less clustered on either of the two dimensions. For example, we could imagine an extractive industry (e.g., raw timber that is exported with little subsequent processing) as falling into quadrant

<table>
<thead>
<tr>
<th>Geographic Clustering</th>
<th>Economic Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Extractive industry: little processing?</td>
<td>Classic Porterian cluster</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Geographic Clustering</th>
<th>Economic Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Many basic consumer services?</td>
<td>Some advanced producer and consumer services?</td>
</tr>
</tbody>
</table>

Figure 2. Two core cluster dimensions
II: little economic or geographic clustering. Other examples are possible for quadrants III and IV, particularly in the services industries.

To complicate matters, the degree of economic and geographic clustering one observes for a particular end-market industry is relative to space, time, and scale. An end-market sector clustered geographically from a national perspective may not be spatially clustered from a regional or local one (or vice versa). Moreover, a given sector becomes more economically clustered as vertical, horizontal, and lateral linkages and relationships expand and deepen (with growth in related and supporting industries and/or the establishment of stronger ties or networking with existing enterprises). And there is no reason, a priori, to assume that clustering along either dimension need only increase over time, even with economic growth and nominal increases in various elements of the cluster. Changes in the social, cultural, or political environment could lead to altered relations between cluster firms such that the positive synergies described by Porter are reduced. Alternatively, improvements in the transportation or communication infrastructure may lead to some spatial dispersal of cluster firms and a reduction in geographic clustering. Finally, there is the element of scale. True clusters are probably large, perhaps exceeding some threshold, but size alone does not guarantee clustering.

Even with this brief description, it should be apparent how complicated Porter’s industrial cluster concept is likely to be in application. One problem is that there is a distinct normative element to both the geographic and economic dimensions that makes identifying clusters for policy purposes difficult. Economic clusters are not just related and supporting industries and institutions, but rather related and supporting institutions that are more competitive by virtue of their relationships (Rosenfeld 1996, 1997). Enterprises are not clustered in space merely because they are located next door, but rather because they are located next door and because they enjoy the types of interdependence described by the economic clustering dimension. Since there are few direct measures of interfirm relationships outside of input-output relations, which clearly do not capture all of the richness of the dynamics described by Porter, it is not surprising that clusters specified for policy attention often constitute little more than large industries and suppliers in which the nation or region has a particular specialization.

Another problem that significantly complicates nationally or regionally administered development policies is that economic clustering may manifest itself over widely varying geographic scales. How, for example, should a regional agency foster (or leverage) positive synergies between local enterprises and cluster members located in other regions or across the world? It is possible that the competitiveness of a given local enterprise may depend more on a
global cluster than a regional one. Could some regionally-focused cluster strategies hinder competitiveness in this context?

These problems, among others, have been discussed at length elsewhere (Jacobs and de Jong 1992, Jacobs and de Man 1996, Rouvinen and Ylä-Anttila 1997). More important for the purposes of this essay is the fact that the theoretical links between 1) various kinds of interdependence between economic actors; 2) the time and space-dependent economic and geographic clustering dimensions, and 3) regional growth and change, are only weakly specified in the Porter framework. On the whole, the model, descriptive and stylized as it is, is presented almost as a fait accompli. This means that regional scientists must determine what Porter’s analysis adds to deeper and richer (though perhaps less comprehensive) body of literature on interdependence and regional development in order to begin to understand how clusters work, how they change over time, and what the implications of such changes are for regional policy.

This essay attempts to contribute to this effort by summarizing the major regional theories and models relevant to the cluster concept. The discussion follows a three part organizational scheme: 1) theories addressing interdependence and economic space; 2) theories of interdependence and the formation of agglomerations; and 3) theories of the relationship between interdependence/agglomeration and regional growth. There are undoubtedly other ways in which this somewhat disparate material might be effectively organized. The framework used here is based primarily on the two core cluster dimensions.

3.1 Interdependence in economic space

By economic space is meant the non-spatial sphere of relations between firms and other economic actors. Clearly one cannot discuss economic interdependence without reference to Perroux’s (1950) theory of abstract economic space as a field of forces in which relations between firms and their buyers and suppliers take place. For Perroux, there is no reason why physical space should necessarily bear any relationship to economic space; enterprise linkages will extend without spatial limit throughout the globe, at least where they are economically justified. Directing one’s analysis to particular regions will only provide a distorted picture of the growth and development process (geographic space as ‘banal’, though note Perroux 1988). In Perroux’s framework, to understand economic growth and change, analysts need to focus on the role of propulsive industries, those industries that dominate other sectors because of their large size, considerable market power, and role as lead innovators. Propulsive industries (or even individual firms) represent poles of growth which attract, focus, and direct other economic resources (Darwent 1969).
The similarities between the cluster concept and Perroux’s theory of growth poles should be readily apparent (see also Dahmén’s 1984, 1988, notion of ‘development blocks’). The cluster focus on how end-market industries drive the deep and broad value chains of which they are a leading part is consistent with propulsive industries as dominant economic actors. End-market industries in given clusters transmit growth pulses through the cluster through demand for intermediate and capital goods. In addition, because they are composed of internationally competitive, best practice firms, they may play an important role as diffusers of process and product innovations. To the degree, for example, that large original equipment manufacturers (OEMs) can use their market power to dictate (or perhaps strongly encourage, even assist with) technology upgrades and improved manufacturing strategies to their suppliers, such end-market industries might be said to drive, at least in part, overall cluster competitiveness. On the other hand, one can also conceive of market power among some cluster members as exerting a detrimental influence on the overall cluster. For example, short-term, least cost-focused contracting practices of OEMs with their suppliers may actually discourage strategic thinking and investment.

Perroux’s macro-level analysis is not unlike Porter’s in another respect: the nature of the dynamics between economic actors is only described in general terms. Micro or meso level theories and studies of changing contracting practices, best-practice technology diffusion, and networking among related firms and industries are particularly important for better specifying how clusters develop and grow (Helper 1991, Imrie and Morris 1992, Klier 1994). Research on these topics has deepened in recent years and holds considerable promise for providing a clear and tangible picture of how enterprise transactions influence strategic choices, innovation, and competitiveness.

Perroux was particularly concerned with demonstrating that economies are characterized more by imbalance than equilibrium. Technological change is a central feature in his framework. The introduction of a significant innovation can lead to the concentration of market power and influence in the hands of the innovating sector, which is then able to establish a lead or dominant position vis-à-vis other industries and firms, at least through the exhaustion of the economic life of the innovation. Thus, the introduction of new innovations establish new or even competing poles, which may alter the structural relations among enterprises in economic space. The development trajectory of the economy is one of movement from imbalance to imbalance.

Schumpeter’s (1934) theory of innovation and development is related to growth pole theory in some of those respects, though not in terms of any focus on economic space. As pointed out by DeBresson (1996; DeBresson and Hu 1997), Schumpeter suggested how innovations cluster in time: in particular, as a
result of “reductions in uncertainty, entrepreneurial profits for rapid imitators, more ingenuity in times of recession, some periods of the business cycle being more conducive to entrepreneurial activity than others, and so on (DeBresson 1996, p. 149).” Schumpeter described the development process in terms of a series of waves. The entrepreneur who initially exploits an opportunity associated with a radical innovation initiates a flow of productive factors from consumption to investment activities (first wave) and is soon followed by imitating entrepreneurs. A general increase in prices, debt financed increases in output, and a shift back toward consumption activities fuels a general expansion phase (second wave). As the innovation nears the end of its economic life, the economy, in effect, overshoots itself and enters a recessionary period. In this economy of fits and starts, we observe key and related innovations clustered in time.

But we may also think of innovations as clustered in economic space. Technological advances tend to establish paths of further innovation through the adoption of similar learning processes, as a result of the systemic nature of particular technologies, and through cumulative learning processes (Debresson 1996). Clustering of innovations can increasingly be studied through recent national surveys of innovative activity, which trace flows of innovations among sectors in a traditional input-output framework. This is the focus of much of the research on clusters and national systems of innovation (Lundvall 1996; Nelson 1988, 1993).

Indeed, the study of industrial clusters in the context of national systems of innovation is an interesting example of how Porter’s general framework has been received, embellished, and implemented differently in various international contexts. In the United States, industrial clusters are often simply a more sophisticated means of targeting traditional economic development programs than the narrower sectoral schemes pursued in the 1970s and 1980s (see Sternberg 1991), while in Europe they are increasingly viewed as an integral element in broader industrial innovation processes or systems. As a result, in the U.S., industrial clusters are perhaps most frequently used for marketing purposes, whereas in Europe, they help identify and characterize the conduits through which learning, technology, and innovations diffuse (Roelandt et al. 1997, Lagendijk and Charles 1997). These continental variations are a function of differences in the related literatures, theories, and concepts that inform regional economic policymaking in the two places. The systems-oriented view of economic processes is much more common in Europe than the United States, which is still dominated by what might be described as an atomistic perspective of interdependence among economic actors.
3.2 Interdependence in geographic space

Of course, regional scientists and geographers are keenly interested in how and why clustering occurs in geographic space, and particularly how such clustering influences regional development paths. It is important to note that Porter’s framework implies that geographic clustering affects national as well as regional fortunes, i.e., questions of location cannot simply be ignored or assumed away in any understanding of national competitiveness. But I defer my discussion of the linkage between geographic clustering and economic growth and development to the following section. Here, I focus on theories of why proximity matters to firms, i.e., the determinants of spatial clustering. Relevant concepts include Marshallian spatial externalities, agglomeration economies, and new industrial districts (a resurgence of interest in Marshall’s original analysis).

Why should firms seek to co-locate in space? Abstracting from transportation or natural resource considerations (following Weber, 1929, who defined these--along with agglomeration economies--as the three basic industrial location factors), firms presumably derive inherent benefits from locating nearby other enterprises. Two basic conceptual approaches to such benefits dominate the literature: the industrial location theory perspective that builds on Weber and Hoover (1937), where the benefits are called agglomeration economies, and the Marshallian perspective that takes as its point of departure Marshall’s ([1890] 1961) original analysis of external scale economies and their typical presence in what he termed ‘industrial districts.’ In both cases, various specific types of externalities (or, more appropriately, sources of externalities; see Feser 1998) are cited as the reason why firms co-locate. The literatures differ somewhat in their relative emphasis on static versus dynamic externalities. Neither perspective is particularly concerned with distinguishing between pecuniary and technological externalities, an underappreciated ambiguity that has direct implications for the need for, and utility of, some types of cluster strategies. Note that here the focus is mainly on externalities related to proximity among business enterprises (localization effects), rather than externalities associated with general urban advantages (urbanization effects).

In agglomeration theory, the reasons why firms co-locate are usually assumed or implied, rather than carefully specified. In his theory of industrial location, Weber (1929, p. 126) assumes a known measure of the cost savings associated with spatial juxtaposition between producers (termed a ‘function of economy of agglomeration’). He distinguishes between types of juxtaposition in terms of the (spatial) concentration of production within a single plant, the concentration of production across several plants in the same industry, and the concentration of production across multiple plants in multiple industries. (Those
are further refined by Hoover, 1937, into the well-known distinction between localization and urbanization economies.) Weber is not particularly concerned with why such agglomeration economies arise, preferring to suggest that they are simply external varieties of the well-accepted notion of internal scale economies (see Weber, 1929, p. 127). Because his primary aim is to demonstrate how the economies might lead to agglomeration, Weber offers very little guidance as to what agglomeration economies really are.

Subsequent theorists in the industrial location tradition have tried to redress this particular problem to a limited degree. Hoover (1937, p. 98) is also confident that the benefits associated with geographically concentrated production are self-evident enough to warrant little discussion. Interestingly, he does note that they likely “depend on the conditions of the particular industry, and also upon institutions (e.g., the way in which wages are determined, the presence or absence or labor unions, the industrial-promotion and tax policies of the communities involved, etc.”). But that is the closest he comes to going beyond general references to industry and urban size as primary determinant of such economies. Other researchers cite particular advantages of proximity between firms, including increased market power through brokered buying and selling, the better availability and use of specialized repair facilities, shared infrastructure, reduced risk and uncertainty for aspiring entrepreneurs, and better information (Isard 1956, Lichtenberg 1960, Vernon 1960, Carlino 1978).

For the most part, one has to look outside the traditional agglomeration economies literature for any sophisticated discussion of firm interdependence and geographic clustering. One important exception is Chinitz’s (1961) paper on market structure as a key determinant of agglomeration economies. In a brief but rich discussion that essentially anticipates the present-day focus on how firm and industry organization influences regional development paths, Chinitz essentially draws a direct link between what Porter calls ‘firm structure and rivalry’ and regional economic fortunes. Critiquing the agglomeration economies literature’s focus on urban and industry size, Chinitz argues that industrial structure particularly influences learning, innovation, and entrepreneurship, giving diverse, and small-firm rich places like New York a leg up over large-firm, single-industry towns like Pittsburgh. This has become an important theme in the Marshallian new industrial district theory.

But before considering the Marshallian perspective, it is worth noting that in his explication of localization and urbanization economies, Hoover relies partly on Robinson ([1931] 1958), who makes an important distinction between mobile and immobile external economies. Immobile external economies are localized, i.e. dependent on the growth of an industry in a given place. Mobile external economies are, in principle, global in scope. Firms may benefit from
the worldwide development of the industry (usually through diffused technological advances). Interestingly, just as the global nature of some types of firm interdependence contributes to the ambiguity in the cluster concept today, early agglomeration theorists also faced the problem of disentangling local and non-local external economies. It is worth noting that Robinson’s distinction is rarely, if ever, referenced in the agglomeration economies literature after the 1960s (for early discussions of mobile and immobile economies in the regional economics literature see Guthrie 1955 and Nourse 1968).

Marshall ([1890] 1961) defines external scale economies as cost savings accruing to the firm because of size or growth of output in industry generally. Such economies contrast directly with internal scale scale economies, which are the source of increasing returns from growth in the size of plant. Such external economies are essentially spatial externalities, which may be defined generally as economic side-effects of proximity between economic actors. They can be either negative or positive, static or dynamic, pecuniary or technological. The static variety is reversible, whereas dynamic externalities are those associated with the technological advances, increased specialization, and division of labor that accompanies and/or drives growth and development (Young 1928). For the most part, regional scientists are interested in dynamic external economies, though this is not always explicitly stated. A static external economy enjoyed by a firm in a given industrial district might be the lower costs it enjoys for intermediate inputs because of proximity to its suppliers (e.g., as a result of reduced shipping costs). That economy is also pecuniary and imposes no market failure since it is fully reflected in the price mechanism. There is certainly no role for government to encourage geographic clustering in this context. To the degree that such benefits outweigh any costs associated with agglomeration (congestion), enterprises will be inclined to cluster on their own.

Of most relevance for understanding industry clusters are dynamic external economies associated with learning, innovation, and increased specialization. Marshall illustrates the workings of (largely dynamic) external economies with reference to concentrated industrial districts, places where firms enjoy the benefits of large, skilled pools of labor, greater opportunities for intensive specialization (a finer social division of labor), and heightened diffusion of industry-specific knowledge and information (knowledge spillovers). Behind these dynamics is not just the size of the district alone, but social, cultural and political factors, including trust, business customs, social ties, and other institutional considerations (Bellandi 1989). Much of Marshall’s analysis is relevant to Porter’s (1990) discussion of firm structure, strategy and rivalry as one of the four determinants of competitiveness (Peneder 1995). In effect, Marshall provides some of the first hints as to how micro-level business
relationships might influence regional growth and development. But he also
emphasizes how important industrial districts are for small firms, which, through
a social division of labor, may enjoy the same types of benefits large firms earn
through internal scale.

Marshall’s basic ideas are subjected to scrutiny and elaboration in the recent
explosion of literature on new industrial districts (particularly ‘Third Italy’) and
small and medium sized enterprises (SMEs; see Asheim 1996, Park and
Markusen 1995, Park 1997). That literature, in turn, draws on theories of
flexible specialization (Isaksen 1997, Asheim and Isaksen 1997, Heidenreich
1996), though the latter’s focus on substantiating a basic sea-change in the
organization of production is less important for understanding the specific,
micro-level relationships that link firm interdependence to geographic
clustering. In other words, to understand why firms might cluster
geographically, it is not necessary to demonstrate a general shift from the
dominance of mass or ‘Fordist’ production methods to more flexible production
regimes characterized by networks of smaller firms, a deeper social division of
labor, and more cooperative business relations.

Of more significance are recent efforts to clarify the general relationships
between scale and scope economies (e.g., Bellandi 1996), as well as the many
case studies of particular industrial districts that identify not only basic
economic trends in agglomerations of smaller firms, but also social and cultural
behavioral codes that govern relationships between firms in these dynamic
regions (see Humphrey 1995 and related articles in World Development 23, 1).
The study of the ‘social embeddedness’ of economic transactions constitutes a
principle contribution of the new industrial district literature (Harrison 1992),
and holds promise for making clearer the broad institutional factors Porter cites
in his work. Unfortunately, from an empirical perspective, drawing conclusions
based on highly stylized case studies is difficult (Bellini 1996). More attention
to the task of assembling the many disparate studies and comparing findings is
 sorely needed. Such an effort would greatly facilitate more systematic theory-
building.

3.3 Interdependence, agglomeration, and regional growth

What does economic and geographical clustering mean for explaining
observed differential regional rates of growth and development? Porter suggests
that a nation’s competitiveness depends on its economic clusters, which
themselves are likely to be spatially concentrated. But his is not a study of
regional growth and change. Although many regional policy makers clearly
perceive a linkage between the promotion of clusters and positive regional
outcomes, the recent cluster literature offers few explicit clues as to what that
linkage looks like or how it works. Instead, we must turn to a host of mainstream theories.

The discussion of economic interdependence began with a review of Perroux’s (1950) conception of economic space and poles of growth. Here it is appropriate to start with the concept of growth centers, the regional extension of Perroux’s non-spatial ideas. Though writing with only limited reference to Perroux’s work, Hirschman (1958) offers one of the first justifications for a growth center strategy for underdeveloped areas. Indeed, Hirschman effectively provides a comprehensive study of economic interdependence (backward and forward linkages), geographic interdependence (‘growth points’), and the implications for regional growth disparities (‘trickling-down of progress’ and ‘polarization’ effects). Granted, Hirschman, along with subsequent growth center advocates, is concerned with jump-starting stagnant underdeveloped areas with significant publically-directed capital investments in a few key sectors. Most of the current cluster debate is taking place in industrialized countries with already diverse economies and relatively strong effective demand (domestic and/or international). Also, most cluster policy applications focus on the establishment of proper framework conditions for strategic clusters to succeed, rather than strategic capital investments in favored industries.

But at the same time, other observers have noted how commonplace it has become for local and regional agencies within industrialized countries to designate clusters for policy attention that are actually very poorly developed or that constitute the only viable industry in the given region (designations motivated more by limited choice sets or politics than any economic rationale; Held 1996). In that context, given limited resources, the decision to concentrate policy attention on key industries—even if it means groups of related sectors—instead of focusing on basic infrastructural needs or other strategies that would serve best a broad array of industries cannot be taken lightly. And from a national policy perspective, a cluster promotion strategy will arguably benefit some regions over others, such as peripheral areas subject to backwash effects from strong growth in neighboring regions. A possible counter-argument is that no one is advocating a significant diversion of development resources to a cluster-focused strategy. But if this is the case, all of this debate is unnecessary anyway.

The attempt to develop propulsive sectors with strong backward linkages that would start a process of cumulative regional advance was a hallmark of the growth center strategies of the 1960s and 1970s (Higgins 1983). The history of failure in growth center policy is legendary. One of the most important reasons why such strategies often misfired is that too little attention was paid to the economic and social pre-requisites that are necessary—at least as hypothesized in
the vast theoretical literature—for growth centers to work (Malizia and Feser 1998). Just as political and equity considerations often dictated, through a criterion of need rather than potential, the designation of very small and peripheral towns as ‘growth centers,’ so it is the case that clusters identified in practice often bear little resemblance to Porter’s (1990) ideal type.

Thus an important issue for industrial cluster applications is how cluster strategies can be developed while also addressing a traditional goal of regional policy: to reduce geographic disparities in income and employment. By some accounts, this is a new era of the city-state, marked by the ascendance of metropolitan regions as the relevant geographic unit for organizing social, economic, and political life (Ohmae 1995). Industry clusters are likely to be concentrated and fare best in prosperous and powerful regions. As industrial and locational advantage begets advantage through increasing returns, economic activity is further concentrated in select places. And, of course, the object of cluster policy is to leverage these advantages to promote positive synergies and returns.

But what about peripheral and rural areas? Does the cumulative advance of some regions become the cumulative decline of others? In his classic criticism of neoclassical growth and trade theory’s faith in the equilibrating force of labor and capital mobility, Myrdal (1957) argued that the backwash effects suffered by peripheral regions as a result of proximity to growth centers likely outweigh any countervailing spread effects. Although the spread-backwash debate has subsided to a degree, recent developments in growth and trade theory show some important consistencies with the early cumulative causation theories.

By replacing the standard neoclassical growth framework and its focus on factor accumulation and exogenous technological change with models that combine perfect competition at the level of the firm with industry-wide externalities, new growth theory opened the door for geography in mainstream economic models. The externalities highlight the role of human capital in generating long-run growth. Operationalized initially by Romer (1986), endogenous growth theory represents a technical refinement of Young’s (1928) important thesis: increasing returns yield growth models that predict perpetual growth. Building on Arrow (1962), Romer identified knowledge accumulation (learning) as a form of investment that generates social externalities. According to Griliches (1992, p. 29), the new approach emphasizes two points: “(i) technical change is the result of conscious economic investments and explicit decisions by many different economic units and (ii) unless there are significant externalities, spillovers, or other sources of social increasing returns, it is unlikely that economic) growth can proceed at a constant, undiminished rate.
One of the most important analysts of knowledge spillovers was Marshall ([1890] 1961) in his study of industrial districts (Krugman 1991).

As a practical matter, the new growth theory’s focus on technology spillovers, information, and knowledge about process and product technologies as technological externalities, presented a solution to the problem of the incompatibility of increasing returns with perfect competition. The intuition behind the spillovers is that they permit the necessary productivity improvements that ensure continued investment. But importantly for regional analysis, the idea that financially uncompensated information exchanged between firms may be an important source of growth leads logically to the proposition that physical proximity is an important determinant of such externalities. This linkage between geographic proximity, externalities and increasing returns brings mainstream economic growth theory closely into line with many regional models of growth and change, particularly those of the industrial districts and agglomeration economies literatures, and is also consistent with the industry cluster perspective.

For example, in drawing a direct parallel between the new advances in growth theory and the role of cities in national and global economies, Lucas (1988) emphasized the linkages between proximity, externalities, and growth. In his survey of ‘new regional economics’ Glaeser (1994, p. 13) writes: “Lucas followed [Jane] Jacobs and argued that when we are thinking about human capital, knowledge and growth, thinking about cities is almost inescapable. Ideas move quickly in cities…Lucas brought into growth economics the idea that cities may be playing a major role in facilitating the accumulation of knowledge spillovers in the growth process.” The geographically concentrated nature of knowledge spillovers has subsequently been the subject of empirical work based on patent data (Jaffe 1989, Jaffe, Trajtenberg and Henderson 1993).

Lucas also noted the importance of grounding new growth theory by identifying spillover mechanisms outside the bounds of the generalized theoretical model: “The engine of growth in the [new growth] models…is human capital. Within the context of [these models], human capital is simply an unobservable magnitude or force, with certain assumed properties, that I have postulated in order to account for some observed features of aggregative behavior (1988, p. 35).” He argued that unless human capital is better defined, it “would make little difference if we simply re-named this force, say, the Protestant ethic or the Spirit of History or just ‘factor X’ (1988, p. 35).” He charged growth theorists with coming up with a clear and testable explanation for such externalities. In this respect, new endogenous growth theory’s “engine of growth” as “factor X” is not unlike Weber’s (1929) ‘function of economy of
agglomeration”; that latter is also something of a black box, an assumed behavioral phenomenon in a model serving other purposes.

Lucas’s external effects of human capital have to do with influences economic actors have on the productivity of each other. The scope of such effects depends on the “ways various groups of people interact (1988, p. 37).” In particular, these effects could be regarded as either global in nature or purely localized at the level of family or firm. If this were the case, Lucas writes, “a model that incorporated internal human capital effects only plus other effects treated as exogenous technical change would be adequate.” However, there is more likely some middle (geographic) ground, since both individuals and firms typically interact at a larger social scale, i.e. the community or neighborhood, city, and industrial district or complex. If knowledge spillovers are prevalent in industry clusters, then the linkage between clusters and long-run economic growth becomes clearer. The proof is in the identification of technological externalities in industry clusters.

Like the new growth theory, ‘new international economics’ also holds important implications for regional analysis. It is not that trade theory now admits a geographic dimension; trade theory has always been spatial theory (Ohlin 1933, Krugman 1991). Rather, the incorporation of increasing returns in models of trade implies the prospect of a highly concentrated geographic pattern of development (Krugman 1990), including sustained disparities in regional income and employment. Again, the focus is on knowledge-related externalities as sources of increasing returns, particularly in advanced technology industries (Krugman 1996). The process of cumulative advance in regions whose industries have established a competitive lead in given markets has been described as a ‘lock-in effect’ (Arthur 1989, 1990a, 1990b). In principle, the initial lead may be as much a result of luck or historical accident as business acumen. But either way particular ‘locational clusters’ may be able to establish a type of monopoly advantage over industries in other places. How likely or sustained such a process would be is an empirical matter. Nevertheless, because industry clusters, by definition, are not ubiquitous, industry cluster policy would seem to imply at least acceptance of a potential worsening of regional economic disparities. Whether this is desirable is a matter of debate, and brings to the fore a host of other difficult questions, including the social costs of concentrated versus dispersed development and the inherent importance of place (the shoring up of perpetually lagging regions). Answers to these questions turn as much on issues of distribution as efficiency.

4 Summary

While the industry cluster concept does not constitute a new, self-contained model of regional development, it does represent a comprehensive description of
how economic and geographic interdependence are integral to regional growth and development processes. But while clusters provide a provocative and holistic perspective on interdependence, the research and policy activity on clusters to date has probably raised more questions that it has answered. Some of the most important of these are: 1) the sources of technological externalities that drive increasing returns in industry clusters; 2) the role of social and cultural versus economic factors in determining such externalities; 3) the role of proximity as an influence on externalities; 4) the prospects for leveraging technological externalities through policy interventions; and 5) the implications of spatially targeted development policy for the growth prospects of lagging regions. Undoubtedly there are others, but these are arguably central to the cluster debate, particularly the question of policy efficacy. Most of them are also long-standing issues in regional analysis.

Answers to many of the difficult questions related to industry clusters will only be possible through continued careful and theoretically grounded empirical work. This invariably implies some (at least temporary) narrowing of terms and definitions, as well as the better integration of cluster ideas with traditional but still relevant regional theories and models. Failure to base industry cluster initiatives on theory and empirical evidence increases the risk of ineffectual policy, wasted resources, and unintended consequences. But the empirical work must also strive for greater generalizability than has heretofore characterized the literature, or the industry cluster concept faces the same problem identified by Bellini (1996, p. 3) for the notion of industrial districts: without generalization, “historic phenomena turn into concepts and one ends up talking about ideal-types, with a more and more vague connection with reality.”

References


Van der Laan, H. B. M. 1997. Everything you always wanted to know about clusters, but were afraid to ask. Opening address, OECD Workshop on Cluster Analysis and Cluster Policies, Amsterdam, Netherlands, 9-10 October.

