Marburg Geography

Working Papers on
Innovation and Space

Identification of Clusters – An Actor-based Approach

# 02.17

Thomas Brenner

Impressum:

Working Papers on Innovation and Space
Philipps-Universität Marburg

Herausgeber:

Prof. Dr. Dr. Thomas Brenner
Deutschhausstraße 10
35032 Marburg
E-Mail: thomas.brenner@staff.uni-marburg.de

Erschienen: 2017

Identification of Clusters – An Actor-based Approach

Thomas Brenner[1]

[1]Philipps University Marburg, Economic Geography and Location Analysis,

Deutschhausstr. 10, 35032 Marburg, Germany

Abstract:

This paper provides two things. First, it gives an overview on the existing top-down methods for the identification of clusters (Section II). Second, it presents a new method that has been recently introduced by Scholl and Brenner (2016) in a basic version. However, the existing version of this approach is limited and does not take full advantage of its potential. The approach is further developed here and its characteristics and the procedure of its use are presented and discussed in detail (Section III).

Keywords: Cluster, cluster identification, spatial methods

JEL Classifications: C43, R12, L60

Identification of Clusters – An Actor-based Approach

Thomas Brenner
Philipps University Marburg
thomas.brenner@uni-marburg.de

I. Introduction

Regional clusters have become very prominent in science and policy in the last 25 years. One of the topics in this field is the identification of clusters. Scientifically, this identification is necessary to study the emergence, evolution and characteristics of clusters. Policy makers are interested in the identification of clusters in order to support exiting clusters directly. Therefore, many approaches to identify clusters have been put forward and used in the last 25 years.

The literature distinguishes fundamentally between two types of approaches: top-down and bottom-up approaches (cf. Cortright 2006). Bottom-up approaches focus usually on one cluster and examine qualitatively the development and characteristics of the cluster under consideration. Such approaches allow to identify clusters and the involved actors in a case-based way.

Here the focus lies on top-down approaches. Top-down approaches identify regional cluster quantitatively on the basis of secondary data (cf. Cortright 2006). Usually the aim is not to identify one specific cluster but all clusters in a nation, region, and/or industry. The most prominent approaches are based on location quotients and input-output tables.

This paper provides two things. First, it gives an overview on the existing top-down methods for the identification of clusters (Section II). Second, it presents a new method that has been recently introduced by Scholl and Brenner (2016) in a basic version. However, the existing version of this approach is limited and does not take full advantage of its potential. The approach is further developed here and its characteristics and the procedure of its use are presented and discussed in detail (Section III).

II. Methods for the top-down identification of clusters

The scientific community has so far not established a standard approach to identify clusters. Various approaches have been proposed and are used in different publications as well as in practice. This is partly caused by the fact that various definitions of clusters exist and that most of them are rather fuzzy.

The most commonly used definition of clusters originates from Porter (1990, p. 78) stating that clusters are “geographic concentrations of interconnected companies and institutions in a particular field”. The success of this definition in scientific as well as political spheres rests on the fact that it is flexible enough to be applied to most if not all cases of existing clusters (cf. Martin & Sunley 2003, p. 9). However, flexibility on the one hand means ambiguity on the other hand. The definition of clusters is especially unclear on the following issues:

1. Definition of the relevant spatial unit
2. Definition of the relevant economic activity
3. Threshold for the definition of an agglomeration
4. Definition of the required connectivity

As a consequence, the various approaches to identify clusters also differ in these aspects.

Most of the mathematical literature on clusters does not provide a solution to this issue.

Mathematical models in the so-called New Economic Geography literature (cf. Krugman 1991, Fujita et al. 1999, Fujita & Thisse 2003) aim to explain the unequal distribution of specific economic activity in space, but not to identify clusters. Giullio Bottazzi and co-workers have studied in more detail and in comparison to empirical evidence the dynamics of firm location and the resulting distribution of economic activity in space (Bottazzi et al. 2007, Bottazzi & Gragnolati 2015). However, a theory that makes it possible to deduce a clear definition of clusters including answers to the above four issues does so far not exist.

As a consequence, most approaches on the identification of clusters are not theoretically deduced but rather implement the basic considerations on clusters in a situational and practical way. The most common way is to consider employment in industries (according to NACE codes) and administrative regions (Issues 1 and 2 above). With respect to the issues 3 and 4 two approaches can be distinguished (cf. Cortright 2006): The identification of clusters on the basis of concentration and specialization measures (described in Sections II.1, II.2 and II.3) and the identification of clusters on the basis of input-output relationships (described in Section II.4).

II.1 Geographic concentration

Clusters are according to their definition related to the fact that industries are not spread in space but agglomerate in certain locations. Hence, the occurrence of clusters is related to the geographic concentration of industries.

The simplest way to measure geographic concentration of industries is the Gini coefficient proposed by Krugman (1991). A more elaborated measure of the geographic concentration of industries is the Ellison-Glaeser index (Ellison & Glaeser 1997). The Ellison-Glaeser index takes into account the size of firms and compares the spatial distribution of an industry with a 'natural' distribution in order to determine whether the industry is more or less concentrated in space.

The Gini coefficient and the Ellison-Glaeser index are both subject to the so-called Modifiable Area Unit Problem (MAUP) (Openshaw 1984): They use data on predefined regions and the definition of these regions has an impact on the outcome of the calculation. As a solution to this problem Duranton and Overman (2005) developed the Duranton-Overman index, which uses geo-located data and is, therefore, a MAUP-free approach.

The Gini coefficient, Ellison-Glaeser index and Duranton-Overman index are well-established methods to study whether industries are geographically concentrated, so that they provide information about the industries in which cluster can be expected. However, they do not offer an option to identify these clusters.

II.2 Location quotient

The location quotient (LQ) – also sometimes called specialization quotient -, is a measure of regional specialization and is frequently used in the literature to measure the strength of clustering. It is defined as,

$$LQ_{ind,reg}=\frac{\frac{v_{ind,reg}}{v_{ind,sp}}}{\frac{v_{ec,reg}}{v_{ec,sp}}} \tag{1}$$

where *v* is the variable for which clustering is to be studied (usually employment) and *ind* stands for a specific industry (part) of the whole economy (*ec*), while *reg* stands for a specific region within the studied space (*sp*).

The location quotient is very often used to identify clusters in the literature. It is used in papers that identify all clusters in a country (e.g., Isaksen 1996, Braunerhjelm & Carlsson 1999, Sternberg & Litzenberger 2004 and Brenner 2006) as well as in studies of specific regions (e.g., Held 1996 and

Crawley et al. 2013). Furthermore, the location quotient is a central element in the identification of clusters in the European Cluster Observatory and the US Cluster Mapping.

In order to use location quotients for cluster identification, a threshold has to be defined. The location quotient can be calculated for each industry in each region. If the threshold is exceeded by this quotient for an industry-region combination, a respective cluster is assumed to exist. In the scientific and practical literature various thresholds are applied. This means that no general threshold has been established in the literature and the choices made in the different publications are quite arbitrary. The thresholds range from 1 (e.g, Held 1996) up to values of 3 (e.g. Isaksen 1996). The European Cluster Observatory uses a value of 2, while the US Cluster Mapping declares the top 25% regions with respect to the location quotient in each industry group as clusters. By this, the US Cluster Mapping defines the threshold in a way that makes it different for each industry group.

In general, the choice of the threshold should depend on the scale of the regions, the used industry classification as well as the aim of the study (cf. Sternberg & Litzenberger 2004). Brenner (2006) proposed a statistical approach to deduce the threshold from the data. This also leads to different values of the threshold for different industries and for different definitions of the spatial units.

Another issue in the use of the location quotient is the fact that it requires comparably little activity in an industry to reach high LQ values in small regions (Caroll et al. 2008). In contrast, using absolute employment numbers instead of location quotients would favor large regions. Therefore, some approaches, such as Isaksen (1996) and the European Cluster Observatory, combine conditions on absolute employment with conditions on location quotients in the definition of clusters.

Finally, the definition of industries is crucial for the use of location quotients to identify clusters. Porter (2003) studies the correlation between industries by examining the co-location of their employment. The findings are used to aggregate industries that are strongly co-located and build industrial complexes (called industry clusters by Porter) from this. Then, location quotients are calculated for these industrial complexes and clusters are identified on this industrial level, as done in the US Cluster Mapping and European Cluster Observatory. Delgato et al. (2016) have further developed the approach of building industrial complexes. Aggregating connected industries is also a basic aspect of the approach developed by Feser and Bergman (2000), which is described in Section II.4.

II.3 Spatial statistics

Another option to examine the concentration of activities in space is the use spatial correlation measures, such as the Moran's I or G statistics. Some publications (e.g. Feser et al. 2005 and Caroll et al. 2008) have used G statistics (developed by Getis & Ord 1992) for the identification of clusters. The calculation of the respective G statistics is done according to (see, e.g. Feser et al. 2005):

$$G_r = \frac{\sum_s \left(w_{r,s} \left(x_s - \frac{\sum_q x_q}{n-1} \right) \right)}{\left(\frac{\sum_q x_q^2}{n-1} - \frac{\left(\sum_q x_q \right)^2}{(n-1)^2} \right) \sqrt{\frac{n \cdot \sum_q w_{r,q}^2 - \left(\sum_q w_{r,q} \right)^2}{n-1}}} , \qquad (2)$$

where *r* denotes the region under consideration, $w_{r,s}$ represents the spatial weight between regions *r* and *s*, $x_{q/s}$ measures the activity under consideration in region *q/s*, and *n* is the number of regions.

The G statistics measure for each region whether the region itself and the surrounding regions show high or low values of activity *x*. In contrast to the approaches described above, it takes activities in the surrounding regions into account for estimating the cluster strength in a region. How many and how strongly surrounding regions are considered depends on the definition of the spatial weights.

Different options for spatial weights are discussed, for example, by Caroll et al. (2008).

II.4 Input-output relationships

An identification of clusters using one of the above described methods does not consider interactions between the actors in clusters (Issue 4 above). Hence, these approaches focus on the aspect of agglomeration within the definition of clusters, while ignoring social, institutional and relational aspects.

Interactions in cluster can take many forms. However, supplier-buyer interactions are often seen as crucial within clusters (see, e.g. Malmberg & Maskell 2002). Feser and Bergman (2000) have taken up this issue and developed an approach to create industrial complexes by aggregating industries. In their approach national input-output tables are used to consider the links between given industries. Industries with strong input-output links are aggregated and industrial complexes are formed (those do not have to be mutually exclusive). The advantage of these complexes is that strongly connected industries are studied together, so that clusters are identified on the level of interconnected industries and not on an industry classification that is developed for different reasons.

Once the industrial complexes are defined, which is done on the national level due to data availability, the relevance of these complexes in a regional economy can be studied. For example, Sánchez Moral (2009) used local Moran's I coefficients to find out which industrial complexes dominate in the region of Madrid. Titze et al. (2011) developed this approach further in order to identify for the German regions whether horizontal and vertical clusters are present there.

III. Actor-based cluster identification

The actor-based cluster identification is based on the actor-based cluster index, which was developed by Scholl and Brenner (2016) and has so far been applied for cluster identification in the Netherlands (Raspe et al. 2012). In the following the approach will be described in detail (Section III.1), its practical application will be explained (Section III.2), various applications will be shown (Section III.3) and its advantages and disadvantages will be discussed in comparison to the existing methods (Section III.4).

III.1 Description of the approach

The main difference of the actor-based cluster index (Scholl & Brenner 2016) to all existing top-down methods to identify clusters lays in the unit of analysis. While all other top-down methods study regions, the actor-based cluster index uses individual actors and their exact geographic location as basic unit of analysis. This allows for assigning a value to each actor that represents whether and to what extent this actor is located in a cluster. The cluster index takes not only values of one (for being part of a cluster) or zero (for being outside of clusters), but provides a more detailed picture from high values (for being in the center of a cluster) to low values (for being far away from clusters). It can be used as a variable in empirical studies that analyze whether being more or less near to clusters influences other characteristics or processes (for an application see Scholl et al. 2016).

Besides such a utilization, the actor-based cluster index can also be used for the identification of clusters (as done in Raspe et al. 2012). The following description focuses on this use of the index. While Scholl and Brenner (2016) developed a cluster index for firms based on employment data, this index is generalized here to be applicable to all kinds of actors and activity.

Hence, for the calculation of the actor-based cluster index two basic units have to be defined:

- **Unit of analysis**: The set of actors, $a \in A$, for which the cluster index is calculated has to be defined. Usual actors are firms and individuals that are active in a certain industry or technological field. However, all other kinds of actors can also be used
- **Unit of activity**: A variable v that is used to measure the amount of activity that is present in a location has to be defined. A usual variable is industry-specific employment. However, all kinds of measurable activities, such as sales or patents can be used. Furthermore, combinations of variables or rates can also be used.

In addition, the calculation of the actor-based cluster index requires two kinds of data:

- **Locations**: The geographic location has to be known for all actors. Furthermore, the value of variable *v* has to be known for all locations. If variable *v* is a characteristic of actors, the data consists of a set of actors, $a \in A$, their locations, l_a, and the corresponding values v_a. However, the approach is not restricted to such a situation and further data can be used in variable *v*. In this case, a set of locations, $l \in L$, has to be defined and for each location *l* the variable v_l has to be known. The set of locations has to contain the locations of all actors, l_a.
- **Distances**: In addition, the approach requires a distance matrix *D*, which contains the distances $d_{l,m}$ between all pairs of locations, $l,m \in L$. The distances, $d_{l,m}$, might be defined as Euclidean distances, but also all other kinds of distances, such as travel times or even social distances, can be used.

To sum up, the approach requires a set of locations (*L*), their distances (*D*), a variable of activity and its value at each location (v_l), a set of actors (*A*), and their locations (l_a). The cluster index is defined for each actors *a* by:

$$C_a = \sum_{l \in L} \left(v_l \, f\left(d_{l,l_a}\right)\right) , \tag{3}$$

where *f(d)* has to be a function that decreases with *d*. There are various options for the definition of the distance decay function *f(d)*. The variants that are most common in the literature are:

- **Power decay function**: A natural assumption is that the relevance of activities at other locations decreases continuously with distance. A simple mathematical formulation of such a decrease is a power decay function:

  $$f(d) = \frac{1}{d^{\alpha}} .$$

  The most simple version is the power decay function with *α=1*, but other values can be used as well. However, the choice of any such function for *f(d)* has the following problem. All this functions become infinite if the distance goes to zero. One option to solve this problem is excluding all activities v_l at the location of actor *a* from Equation (3). However, activities at the same location can be expected to be very influential. The other option is to define a minimum distance d_{min}, below which the value of *f(d)* does not change any more. This can be justified by arguing that below a certain distance, e.g. 500 meters, the exact distance does not matter anymore. As a consequence, *f(d)* can be defined by (cf. Scholl & Brenner 2016):

  $$f(d) = \frac{1}{max\left(d_{min}^{\alpha}, d^{\alpha}\right)} . \tag{4a}$$

- **Exponential decay function**: The necessity to define a minimal distance can be avoided by the use of an exponential function:

  $$f(d) = \exp(-\alpha \cdot d) , \tag{4b}$$

  where *α* is a parameter defining how fast relevance decreases with distance. Using this function assumes, again, that the relevance of activities for an actor decreases continuously with distance.

- **Radius decay function**: In spatial statistics one approach is to consider all activities that occur within a certain radius around the considered actor (Rosenthal & Strange 2003). In such an approach it is argued that all activities within a distance below a certain threshold matter for an actor, while all activities further away do not matter.

In this case, the function *f(d)* is given by

$$f(d) = \begin{matrix} 1 & \text{if } d \leq r \\ 0 & \text{if } d > r \end{matrix} \quad , \tag{4c}$$

where *r* denotes the threshold defining the range of relevance.

- **Log-logistic decay function**: A more flexible mathematical formula that contains (approximately) the two above cases as extreme cases is the sigmoidal log-logistic decay function (DeVries et al. 2009):

$$f(d) = \frac{1}{1+\exp(s \cdot \log(d/r))} = \frac{1}{1+\left(\frac{d}{r}\right)^{-s}} \quad , \tag{4d}$$

where *r* is a parameter defining the distance at which relevance has decreased to ½ and *s* is a parameter defining the shape of the function. For s→0 the function becomes exponentially decreasing, while for s→∞ the function converges to Function (4c). While the logistic function is more flexible than the other options, it requires to fix two parameters with little empirical knowledge about these parameters.

So far it is rarely studied empirically which is the best fitting function *f(d)*. De Vries et al. (2009) find that a log-logistic decay function fits commuting flows much better than power or exponential decay functions. Duschl et al. (2015) estimate the parameters within the log-logistic decay function empirically and find that in most cases the parameter *s* is rather large, so that the log-logistic decay function becomes quite similar to the radius decay function. Hence, the empirical literature so far provides more evidence of the functions (4c) and (4d).

Given the necessary data (*L*, *D*, v_l, *A* and l_a) and the distance decay function, the cluster index (3) can be calculated for each actor. This results in a value C_a for each actor *a* denoting how much activity *v* is given in the surrounding of this actor. In order to detect clusters, a threshold T_{cl} has to be defined that separates those actors within clusters ($C_a > T_{cl}$) from those actors outside of clusters ($C_a < T_{cl}$). There are several ways to fix this threshold:

- **Theoretical approach**: Dependent on the definition of the activity variable *v*, there might exist theories that allow for the deduction of threshold T_{cl}. As stated above, cluster theory does not provide such a theoretical threshold so far. However, there might be specific aims, e.g. the aim to identify the strongest cluster that imply a certain threshold.
- **Literature approach**: As reported above the literature provides many approaches to detect clusters, most of which are based on thresholds. These thresholds, which are fixed according to the empirical experience with their use, can be transferred to the approach proposed here. To this end, the activity variable *v* has to be chosen in correspondence to the respective literature, so that the same threshold can be applied.
- **Statistical approach**: The distribution of the cluster index values C_a can be used to fix the threshold T_{cl}. The values C_a can be sorted and, for example, the top 5% values can be declared as signifying clusters. Of course, other percentiles can be used as well. Alternatively, the cluster index can be calculated for a randomly chosen benchmark case and the threshold can be defined as the 90th, 95th or 99th percentile of the resulting distribution. Such an approach is proposed by Scholl and Brenner (2016).

Independent of the definition of threshold T_{cl}, clusters are defined in the approach proposed here by calculating for each actor whether or not this actor is part of a cluster. In contrast to the usual approaches, which calculate for each region whether it contains a cluster or not, a cluster is defined here by the actors that it consists of. Hence, the range of a cluster is not predefined by the borders of regions, but follows from the spatial spread of the actors that are classified as being part of it.

III.2 Practical steps to identify clusters

Table 1 summarizes the data and decisions needed for the conduction of a cluster identification using the actor-based cluster index. The underlying steps are explained in details in the following.

Table 1: Data and definitions for actor-based cluster identification

Variable	Content	Comment
A, l_a	Set of actors *a*	Contains all actors for which it is to be calculated whether they belong to a cluster or not. The location of all these actors has to be known.
v	Activity level	Variable describing the activity that is high within the clusters under consideration and low outside.
L	Set of considered locations l	Including all locations of actors and all locations of relevant activities.
D	Distance matrix	Provides the distance between each pair of locations.
$f(d)$	Distance decay function	Mathematical form of how the relevance of activity *v* decreases with the distance from the actor.
T_{cl}	Cluster threshold	Threshold for the cluster index C_a that determines whether an actor *a* is part of a cluster ($C_a > T_{cl}$) or not.

III.2.1 Set of actors

For the definition of the set of actors it is important to consider that the choice of this set has only implications for the locations at which it is calculated whether a cluster is present or not. The approach could also be conducted without defining any actor and only looking at locations. However, the theoretical considerations behind the approach are based on the idea to identify actors that belong to clusters. Actors are part of clusters and not locations. Nevertheless, the approach is quite flexible in this sense, because for the results of cluster identification it does not matter why a location is considered. For the approach it is only important to have a set L_A of locations l_a for which the cluster index is calculated and for which it is finally decided whether they belong to a cluster. As a consequence, there are neither restrictions for the definition of the set of actors nor is the kind of actors chosen relevant for the result (only their location matters).

The usual actors in clusters are firms. Hence, defining a specific set of firms as *A* is a natural choice. Clusters are usually defined on the level of industries. This implies that all firms belonging to one industry can make up the set of actors. However, set *A* may also consist of all firms in a number of related industries.

Furthermore, other organizations can also be included in the set of actors. According to the common cluster theories, universities, research institutes, intermediaries and public organizations play also a role in clusters. All these actors can be included in set *A*.

In addition, the set of actors can also be built on the basis of individuals instead of organizations. Workers, innovators, researchers or similar actors can be used. Again an industry or a specific group of industries can be considered. However, also certain technologies or research topics may define the object of examination.

Independent of the definition of actors, it has to be kept in mind that actors are only a mean to define locations, because locations are part of a cluster only if there are actors that can be part of a cluster. The question of whether an actor-location is part of a cluster is determined mainly by the choice of the activity variable *v* and partly by all other choices that are described and discussed in the following.

III.2.2 Activity variable v

The choice of the activity variable *v* is most central for the result of cluster identification. In the literature employment is used most often. Considering employment in absolute terms favors big cities. Therefore, the location quotient (1), measuring specialization, is used in most approaches. A straight forward transfer of such an approach to the actor-based cluster identification is not possible. Using the location quotient as activity variable, *v=LQ*, would require a value for the total economic activity at each location. If the site of an actor, e.g. a firm, defines a location, there is no other economic activity at this exact location and the total economic activity would equal the industry-specific activity. Even if we define location in a more brought sense, e.g. as a municipality, we run into another problem: The cluster index is calculated according to Equation (3):

$$C_a = \sum_{l \in L} \left(v_l f\left(d_{l,l_a}\right) \right).$$

Hence, small locations with little industry-specific and little total economic activity add up in the cluster index in the same way as locations with a lot of economic activity. The values of ratios, such as the LQ, vary in locations with small total activity much more than in locations with high total activity. Hence, locations with small total activity would influence the cluster index too much if ratios are used.

Therefore, if the activity variable is defined as a ratio, this ratio should be taken after summarizing all locations. For example, a cluster index based on the LQ as a measure of activity would read

$$C_a = \frac{\dfrac{\sum_{l \in L} \left(v_l f\left(d_{l,l_a}\right) \right)}{\sum_{l \in L} \left(w_l f\left(d_{l,l_a}\right) \right)}}{\dfrac{\sum_{l \in L} \left(v_l \right)}{\sum_{l \in L} \left(w_l \right)}}, \qquad (3a)$$

where w_l denotes the total activity at location *l*. Hence, Equation (3a) provides an alternative definition of the cluster index that is more adequate if the cluster examination should be based on a quotient. All kinds of variables can be used for *v* and *w* in Equation (3a), so that all kinds of quotients can be build. The approach can be generalized in three directions.

First, other measures can be used, such as number of firms, sales, patents or publications. The variable *v* would still represent the specific – industry-specific, technology-specific etc. – activity, while *w* would denote the total activity. By this, cluster identification can be done for all kinds of specific activity.

Second, in addition, *w* might not be defined as the overall activity but completely independent of *v*. By this the relationship between two variables can be used as a basis for cluster identification. A simple option is the use of a specific economic activity (industry employment) in relation to population. In this case, *v* would be industry-specific employment and *w* would be population. However, any kind of relation can be used, e.g. defining *v* as the number of patents in a certain technology and *w* as the number of employees in this technology.

Third, one might go beyond that and combine different activities. One option would be the combination of one kind of activity (e.g. employment) from different specific fields (e.g. different industries). This allows to include input-output connections in the analysis by adding to the employment in the industry under consideration the employment of all other industries multiplied by the share of their connection to the analyzed industry. *w* would remain total employment. Through this, a combination of the two standard approaches, based on the location quotient and input-output matrices, is possible, transferring the ideas of Section II.4 to the actor-based cluster identification. Another option is the use of a composite indicator that combines different activities, such as economic, research, scientific and educational activities. As long as activities are summed, *v* can be defined as this sum and then the cluster index can be calculated according to Equation (3a). If activities are multiplied – in composite indicators this is often the case, but there might also be other reasons for a multiplication –, this might be done before calculating the cluster index (inserting $v=v_1v_2$ into Equation (3a)), or after calculating the sum over all locations:

$$C_a = \frac{\dfrac{\sum_{l \in L} \left(v_{1,l}\, v_{2,l}\, f\left(d_{l,l_a} \right) \right)}{\sum_{l \in L} \left(w_l\, f\left(d_{l,l_a} \right) \right)}}{\dfrac{\sum_{l \in L} \left(v_{1,l}\, v_{2,l} \right)}{\sum_{l \in L} \left(w_l \right)}} \tag{3b}$$

Independent of how activity *v* is defined, the set of locations *L* automatically results from the definition of this activity. Set *L* contains all locations at which this activity takes place plus all locations with actors from set *A*, which is usually a subset of the locations with activity *v*.

III.2.3 Distances and distance decay function

The calculation of the cluster index requires the definition of distances. The most usual and simplest definition is the use of Euclidean distances. In order to be more accurate, orthodromic distances represent an alternative option. These distances can be calculated whenever the geographic coordinates of the locations are known. Hence, the use of these distances provides a basic and usually available option. However, there are more complex options.

In order to consider the transport infrastructure, travel times can be used as distances (cf. Duschl et al. 2014). For an actor the relevance of activities at other locations is due to the interaction that takes place between locations. Such interactions are conducted by people. Besides travel times, people also consider in their decisions to interact social, cultural, institutional and organizational aspects. Since those aspects are difficult to measure, measuring true interaction provides an option to approximate the impact of these different kinds of influences. True interaction between locations can be measured by commuting links, by migration, by cooperative activities etc. The combination of such an approach of measuring distance with the actor-based cluster identification might generate interesting insights into the geographical range of clusters.

Finally, different distance measures might be combined. Proximity in some aspects compensates for distance in other aspects (Agrawal et al. 2008). Especially trust and long-standing cooperation as well as organizational proximity are able to bridge geographic distance. Given that the approach proposed here is actor-based, geographic distance and cooperative activities or belonging to the same organization can be combined with distance measures. This would also allow to integrate the aspect of cooperation into the identification of clusters.

After defining the distance measure, a distance decay function has to be chosen. Above it has been argued that a log-logistic decay function (Equation (4d)) and a radius decay function (Equation (4c)) are most supported by the literature so far. However, the choice of a distance decay function should also depend on the activity *v* that is considered. Different activities might have different ranges of their impact. Little is known about this so far, so that hopefully more research will come that provides further information about adequate distance decay functions.

III.2.4 Index threshold and identification of clusters

The steps above allow finally to calculate a cluster index for all considered actors. Different options exist to set a threshold for the identification of those actors that are part of a cluster. These options are described and discussed in Section III.1. The geographic shape of each cluster is then the hull around all nearby actors that are found to be part of a cluster.

However, the actor-based cluster identification offers to go beyond the simple identification of clusters. For each actor a cluster index is calculated. Hence, clusters of different degrees can be identified by using different thresholds, which is, of course, possible also with the traditional approaches. However, using a traditional approach, the regions in which different degrees of clustering are found could be quite scattered in space. In contrast, in the actor-based approach the locations with very high cluster values   form certain areas, which could be called the core of a cluster, while the location with high, but not very high cluster values are located in the surrounding of this core and could be called the connected periphery (see the examples given in Figures 1-5). For each actor it can be determined whether she/he is in the core, in the periphery or not belonging to a cluster. Using a distance that includes social, institutional or other kinds of relationships, the approach can go far beyond a geographic identification of clusters.

III.3 Exemplary applications

In order to show the potential of the actor-based cluster identification, it is applied to the automotive industry in Germany in 2012. The employment in the automotive industry (NACE, Rev. 2: Division 29: Manufacture of motor vehicles, trailers and semi-trailers) is used as variable *v*.

Instead of using the exact geographic location of each actor, all actors are assigned to municipalities. This is done for two reasons: First, complete data is only available on this level. Second, using municipalities allows to use the shapes of these municipalities in the graphics instead of bubbles or points (as done in Raspe et al. 2012 and Scholl & Brenner 2016). Through this clearer pictures can be generated. Especially for demonstration this seem to outweigh the disadvantage of not using exact locations, which matters especially in the big cities. Distances are measured in travel times between the municipalities.

As a first example, the cluster index is calculated for the absolute number of employees using the power and radius decay function and depicting the top 5 (red) and 10% (yellow) in Figure 1. The power decay function allows to identify each single spot, clearly favoring big cites, the radius decay function works better to identify the hot spots of the automotive industry their surrounding.

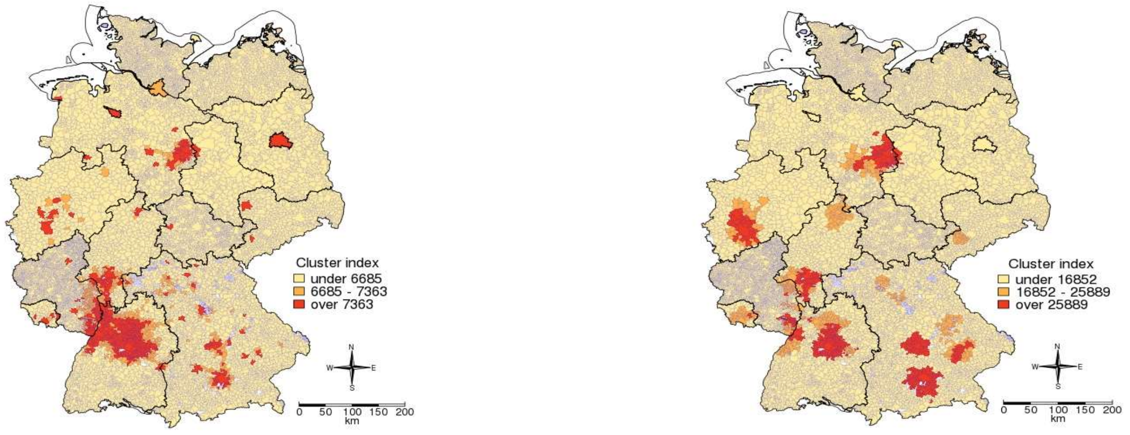

Figure 1: Cluster index for automotive employment in German municipalities using the power decay function, α=1, (left) and the radius decay function, r=30 minutes (right).

As discussed above the radius decay function is supported by empirical evidence. According to this evidence, the relevant radius should be somewhere between 30 and 60 minutes.Hence, radius decay functions with *r=45* minutes and *r=60* minutes are used in Figure 2. Increasing the radius leads to the identification of less and larger clusters. Using a 60 minutes radius leads to the identification of only the three largest clusters: Wolfsburg, Stuttgart and Munich.

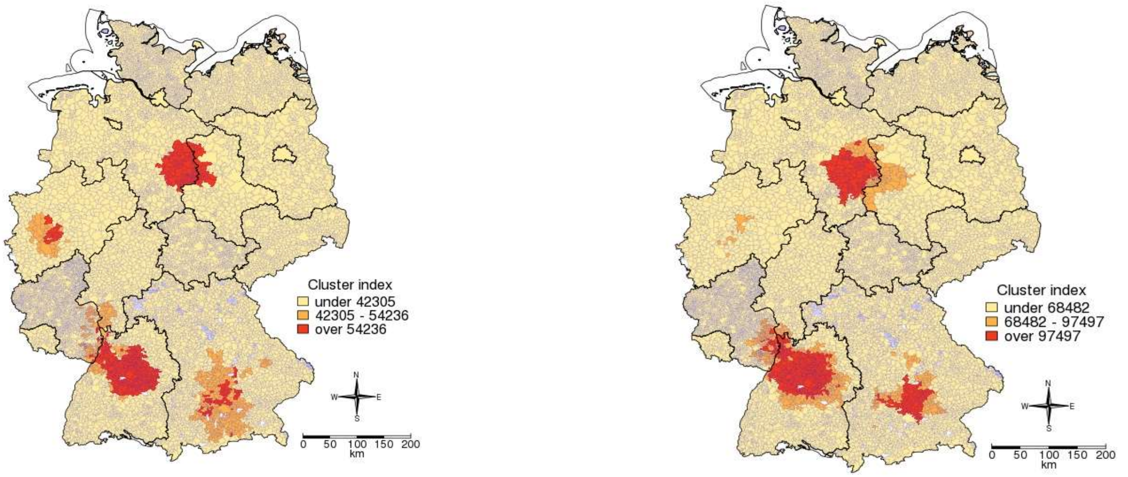

Figure 2: Cluster index for automotive employment in German municipalities using the circular sphere of relevance with r=45 minutes (left) and r=60 minutes (right).

The most flexible distance decay function is the log-logistic function. Using such a decay function leads to quite similar results as using the radius decay function (see Figure 3). Especially using *s=5* and *r=30* minutes leads to a very good identification of the big clusters around Wolfsburg (VW), Cologne (Ford), Rüsselsheim (Opel), Stuttgart (Daimler and Porsche), Ingolstadt (Audi) and Munich (BMW).

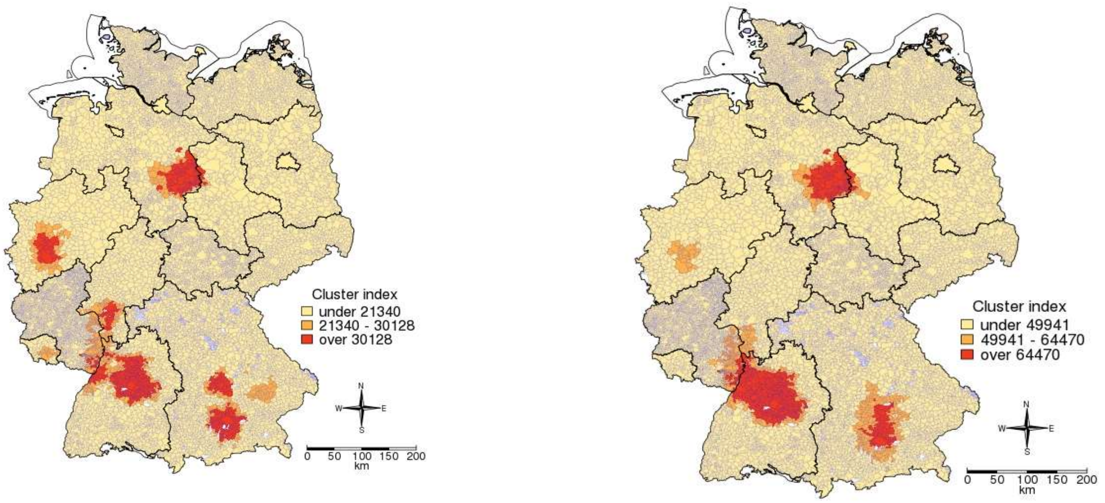

Figure 3: Cluster index for automotive employment in German municipalities using the log-logistic decay function with s=5 and r=30 minutes (left) and r=45 minutes (right).

As mentioned above, the identification of clusters on the basis of absolute employment numbers favors big cities. Therefore, usually the location quotient is used for cluster identification.

Using the LQ changes the picture (see Figure 4). Now places with less total economic activity are favored and especially those locations are identified in which the automotive industry dominates the economic activity (such as, e.g., Wolfsburg), while diverse places such as Stuttgart lose their dominant status. Figure 4 depicts the clusters based on the usual thresholds between 1 and 3 for the location quotient. It can be easily seen that a threshold of 1 would lead to an identification of large areas as clusters. A threshold between 2 and 3, rather nearer to 3, seems to be adequate in the automotive industry.

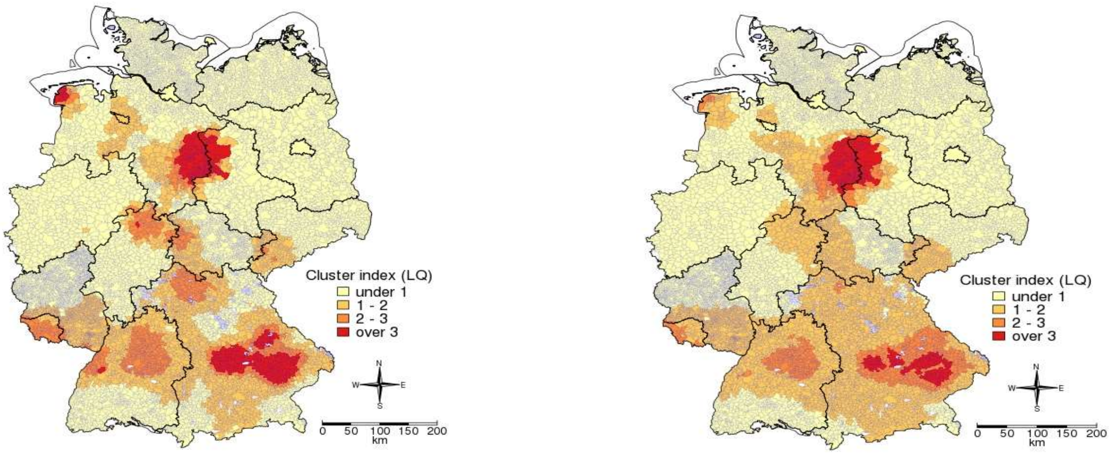

Figure 4: Cluster index based on the location quotient for the automotive industry in German municipalities using the log-logistic decay function with s=5 and r=30 minutes (left) and r=45 minutes (right).

Finally, the above outlined option to include related industries in the cluster identification is tested here. To this end, the variable *v* is defined as the employment in the automotive industry plus the employment in all manufacturing supplier industries that supply more than 2% of the industry's input, each multiplied with the share that it supplies. The resulting value is used to calculate a cluster index based on a location quotient according to Equation (3a). The resulting cluster index is depicted in Figure 5. The picture does not change much because in the automotive industry the supplier industries are located nearby. In other industries this might be different.

The presented applications can be seen as first examples to show the potential of the actor-based cluster identification. As described above, the approach offers many more options. However, using the full potential of the approach is beyond this paper and has to be done in future applications.

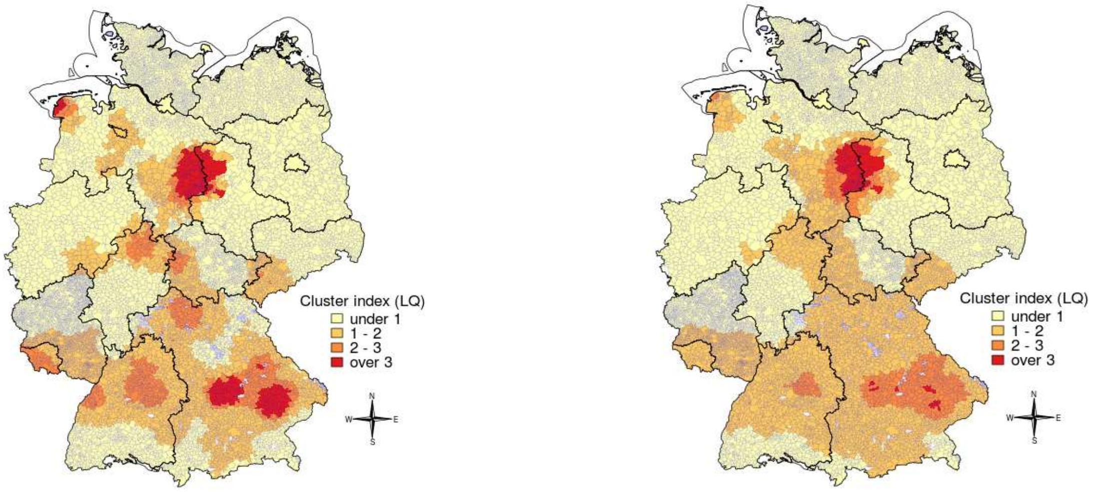

Figure 5: Cluster index based on the location quotient for the automotive industry and its suppliers in German municipalities using the log-logistic decay function with s=5 and r=30 minutes (left) and r=45 minutes (right).

II.4 Advantages and disadvantages

In literature various different approaches are used to identify clusters, as described above. In order to show and discuss the advantages and disadvantages of the actor-based cluster identification the different approaches are compared in the following. Five aspects are considered: data requirements and the four issues that are brought up at the beginning of Section II: Spatial unit, relevant economic activity, agglomeration threshold and connections with clusters. The comparison is summarized in Table 1.

The main disadvantage of the actor-based cluster identification is the requirement of micro-geographic data. All other approaches require only data about the activity under consideration (usually employment, sometimes additional variables) for all regions. The approach based on input-output relationships requires in addition an input-output table on the national level, which is available for most countries. For the actor-based cluster identification all activities have to be geographically located. This data requirement might be partly lowered by using very small regional units (such as municipalities as done above). Furthermore, the actor-based approach requires is a measure of distance between actors.

The main advantage of the actor-based cluster identification is the avoidance of the Modifiable Area Unit Problem (MAUP). All other approaches work with predefined regions, for which the identification checks whether the region is part of a cluster or not. The results of these cluster identification approaches strongly depend on the regional level that is used. The spatial range of detected clusters are to some extent predefined by the regions studied. In contrast, the actor-based approach treats each actor separately. Hence, the exact shape of a cluster is detected by the approach. Contiguous areas are found to build clusters. Furthermore, the results of the cluster identification can be used to distinguish between the core and the periphery of a cluster.

Table 1: Advantages and disadvantages of the various approaches

Aspect	**Location quotient**	**Spatial statistics**	**Input-output relationship**	**Actor-based approach**
Data requirements	Medium (regional data)	Medium (regional data)	Medium (regional data and input-output matrix)	High (micro-geographic data and distances)
Spatial unit	Affected by MAUP (predefined regions)	Affected by MAUP (predefined regions)	Affected by MAUP (predefined regions)	MAUP-free (allows for identifying the spatial range of clusters)
Complexity of the definition of economic activity	Medium-high (usually industries, but industrial complexes and more is possible)	Medium-high (usually industries, but industrial complexes and more is possible)	High (usually industrial complexes on the basis of input-output relations, but more is possible)	Medium-high (usually industries, but industrial complexes and more is possible)
Agglomeration threshold	A threshold is needed (different options can be used)	A threshold is needed (different options can be used)	A threshold is needed (different options can be used)	A threshold is needed (different options can be used)
Considering connectivity within the cluster	Not possible (besides industrial connections considered in the definition of economic activity)	Not possible (besides industrial connections considered in the definition of economic activity)	Not possible (besides industrial connections considered in the definition of economic activity on the basis of input-output relations)	All kinds of connections can be considered and combined, especially on the level of actors

In the standard cluster identification the relevant economic activity is defined according to industry classifications or industrial complexes. By the definition of industrial complexes spatial co-location or input-output relations can be taken into account (as done in the literature: Porter 2003 and Feser & Bergman 2000). Although the identification of clusters on the basis of input-output relationships contains the building of industrial complexes as a central feature, such a step can be easily integrated in all approaches. All approaches do also allow to go beyond this and either include or combine other aspects and variables.

All approaches require some kind of threshold if finally clusters are to be identified. So far the different strands of literature approach this issue differently. However, the various ways to define a threshold can be applied in all approaches. Hence, all approaches face this problem and none of them provides a non-arbitrary solution.

The second main advantage of the actor-based approach is the possibility to include connections within clusters in the identification approach. All other approaches focus on the agglomeration aspect within the cluster definition. The approach based on input-output relations includes supplier-buyer-relations as a central element. However, this is only used to define industrial complexes as the industrial unit of analysis. The input-output relations are not taken from within the clusters but are measured on a national level. As described above, in all approaches the definition of the analyzed industrial complexes might include relationships between industries. However, interactions or connections within clusters are not included in the identification by such an approach. Interactions and connections occur between the cluster actors, such as firms, organizations, institutions and individuals. Hence, only an actor-based approach is able to take into account such interactions and connections.

The definition of distances can easily be used to include these aspects. As stated above, different kinds of proximity might compensate each other. This fact can be directly transferred into the actor-based approach. The actor-based approach requires a definition of the distance between all actors. Usually a geographic distance is calculated. However, all other kinds of distances, such as social, institutional, organisational and cognitive, can be used or combined in a definition of distance for this approach, if respective data is available. It is also possible to use data on actual connections and interactions, such as co-patents or R&D-cooperations. Hence, the actor-based approach is able to consider both aspects of clustering, spatial agglomeration and connectivity, in the identification of clusters. Also this option is not used so far, this is a clear advantage of the actor-based approach, which will be hopefully used in the future.

IV. Conclusions

The approach for identifying clusters that is proposed here has so far rarely been used. It is mathematically easy to conduct and offers a number of advantages compared to the usually approaches, but requires data on the micro geographic level. However, geographically coded data becomes more and more available, so that the proposed approach offers great potential in the future. Especially, it is able to identify clusters without facing the MAUP problem and detecting their spatial range from the data. Furthermore, it allows to include interactions and connections within clusters in the identification of clusters.

The approach is quite generally applicable and offers many options, meaning that it requires various choices by the applicant. The various options are discussed comprehensively above. Most of these options are also available in the common approaches. Hence, this paper also opens up a discussion about what are the adequate values to identify clusters in a top-down approach. So far, the potentials that cluster identification approaches offer are used to a very small extent. Hopefully, the higher availability of data will lead to more elaborated top-down identifications of clusters.

References

Agrawal, A., Kapur, D. & McHale, J. (2008): How do spatial and social proximity influence knowledge flows? Evidence from patent data. *Journal of Urban Economics* 64: 258-269.

Bottazzi, G., Dosi, G., Fagiolo, G. & Secchi, A. (2007): Modeling industrial evolution in geographical space. *Journal of Economic Geography* 7: 651-672.

Bottazzi, G. & Gragnolati, U. (2015): Cities and clusters: Economy-wide and sector-specific effects in corporate location. *Regional Studies* 49: 113-129.

Braunerhjelm, P. & Carlsson, B. (1999): Industry Clusters in Ohio and Sweden, 1975–1995. *Small Business Economics* 12: 297–293.

Brenner, T. (2006): An Identification of Local Industrial Clusters in Germany, *Regional Studies* 40: 991-1004.

Carroll, M.C., Reid, N. & Smith, B.W. (2008): Location quotients versus spatial autocorrelation in identifying potential cluster regions. *Annals of Regional Science* 42: 449-463.

Cortright, J. (2006) Making sense of clusters: regional competitiveness and economic development, Discussion Paper, mimeo, The Brookings Institution Metropolitan Policy.

Crawley, A., Beynon, M. & Munday, M. (2013): Making location quotients more relevant as a policy aid in regional spatial analysis. *Urban Studies* 50: 1854-1869.

Delgato, M., Porter, M.E. & Stern, S. (2016): Defining clusters of related industries. *Journal of Economic Geography* 16: 1-38.

DeVries, J., Nijkamp, P. & Rietveld, P. (2009), Exponential or power distance – decay for commuting? An alternative specification. *Environment and Planning A* 41: 461–480.

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

Duschl, M., Schimke, A., Brenner, T. & Luxen, D. (2014): Firm growth and the spatial impact of geolocated external factors - empirical evidence for German manufacturing firms, *Journal of Economics and Statistics* 234: 234-256.

Duschl, M., Scholl, T., Brenner, T., Luxen, D. & Raschke, F. (2015): Industry-specific firm growth and agglomeration, *Regional Studies* 49: 1822-1839.

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

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., Sweeney S. & Renski H. (2005): A descriptive analysis of discrete US industrial complexes.

Journal of Regional Science 45: 395–419.

Fujita, M., Krugman, P. & Venables, A.J. (1999): The spatial economy: Cities, regions and international trade, Cambridge, Mass.: MIT Press.

Fujita, M. & Thisse, J.-F. (2003): Does geographical agglomeration foster economic growth? And who gains and loses from it? *The Japanese Economic Review* 54: 121-145.

Getis, A., & Ord, J.K. (1992): The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24: 189–206.

Held, J.R. (1996): Clusters as an economic development tool: beyond the pitfalls. *Economic Development Quarterly* 10: 249–261.

Isaksen, A. (1996): Towards increased regional specialization? The quantitative importance of new industrial spaces in Norway, 1970-1990. *Norsk Geografisk Tidsskrift* 50: 113–123.

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

Malmberg, A. & Maskell, P. (2002): The elusive concept of localization economies: towards a knowledge-based theory of spatial clustering. *Environment and Planning A* 34: 429–449.

Martin, R. & Sunley, P. (2003): Deconstructing clusters: chaotic concept or policy panacea? *Journal of Economic Geography* 3: 5-35.

Openshaw, S. (1984): The Modifiable Areal Unit Problem. Geo Books, Norwich.

Paniccia, I. (1998): One, a hundred, thousands of industrial districts. Organizational variety in local networks of small and medium-sized enterprises. *Organizational Studies* 19: 667 –699.

Porter, M.E. (1990): The competitive advantage of nations. New York: The Free Press.

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

Raspe, O., Weterings, A., Geurden-Slis, M. & van Gessel, G. (2012): De ratio van ruimtelijk-economisch topsectorenbeleid. Den Haag: PBL.

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

Sánchez Moral, S. (2009) Industrial clusters and new firm creation in the manufacturing sector of Madrid's metropolitan region. *Regional Studies* 43: 949-965.

Saxenian, A.-L. (1994): Regional Advantage. Cambridge, MA: Harvard University Press.

Scholl, T. & Brenner, T. (2016): Detecting spatial clustering using a firm-level cluster index, *Regional Studies* 50: 1054-1068.

Scholl, T., Brenner, T. & Wendel, M. (2016): Evolving localization patterns of company foundations, *Journal of Evoluationary Economics*, in print.

Sternberg, R. and Litzenberger, T. (2004): Regional clusters in Germany: their geography and their relevance for entrepreneurial activities. *European Planning Studies* 12: 767 –792.