# Three variations on identifying clusters

Alex Hoen[1, 2]

## 1. Introduction

This study searches for a method that identifies clusters. In order to achieve this goal, first the theoretical concept of clusters is analysed. This part distinguishes six types of clusters that are generally used in economic literature. One of these concepts is mostly used in theoretical studies and policy aims, another concept is often used in empirical analyses. Fortunately, a link between these two concepts can be established, which helps to interpret empirical clusters as a proxy for theoretical clusters. Next, the possible empirical methods to identify clusters are investigated. Based on theoretical considerations and an empirical example, a cluster identification method is derived. An application of this method on two data sets, one of which is aggregated but available for a long period and one of which is disaggregated but available for only a few years, leads to some interesting conclusions about the empirical cluster method. Finally, the value of empirical clusters as a proxy for theoretical clusters and policy aims is reconsidered in the conclusions based on the findings in this study.

## 2. A Classification of Clusters

From a theoretical point of view, the variety of clusters can be classified along two dimensions. Firstly, clusters can be distinguished according to the scope or level of analysis: the *micro level* refers to clusters of firms, the *meso level* and the *macro level* refer to clusters of sectors. Secondly, the relation between the entities in a cluster may refer to *innovative efforts* or to *production linkages*. Clusters based on innovative efforts relate to firms or sectors that co-operate in the process of diffusing innovations such as new technologies or products; clusters based on production linkages relate to firms or sectors that form a production or value-added chain.

The distinctions above yield six types of clusters. A short description of each of these types is given in table 1.

*Table 1:six types of clusters*

| | Innovative efforts | Production linkages |
|---|---|---|
| Micro | Diffusion of technologies and knowledge between firms, research institutions, etc. | Suppliers and buyers in a value-added or production chain of firms |
| Meso | Diffusion of technology and knowledge between sectors | Backward and forward linkages between sectors; partial analyses |
| Macro | A split up of the economic system in sectors that diffuse knowledge or technologies | A split up of the economic system in sectors that form value added or productions chains |

[1] Netherlands Bureau for Economic Policy Analyis, P.O. Box 80510, 2508 GM, The Hague, The Netherlands, tel. +31-70-3383497, e-mail: hoen@cpb.nl.

[2] I am very thankful to Paul Arnoldus and Fieke van der Lecq for valuable insights and for many suggestions regarding the ideas in this study and for severely improving my texts.

## 3. An operational definition of a cluster

Most theoretical studies focus on micro clusters diffuse innovations. These clusters are also most relevant for policy issues. After all, cluster policy intends to stimulate the development of new technologies and knowledge, which generates higher economic growth. Empirically, however, most analyses are based on meso clusters of sectors in a value added chain. Hence, there seems to be a tension between theoretical analyses and policy aims on one side and empirical analyses on the other side.

Still, there is a relation between the empirical meso clusters and the theoretical micro clusters. Porter[3] and DeBresson[4] both stress that firms co-operating in a cluster will often be situated in different sectors. Moreover, firms enjoying in combined innovative efforts will probably be linked in a production chain as well. For an innovation, suppliers and buyers in the production chain have to be informed of its consequences. Very often the co-ordination of actions and strategies is required not only in the production process but also in the innovative effort itself.

DeBresson finds empirical support for the hypothesis above. He concludes that the pattern of diffusing innovations resembles the pattern of linkages in an input-output table. Therefore, empirical results of meso clusters of sectors based on linkages can be used to derive conclusions about the co-operation of firms in innovative efforts.

There are four additional reasons why meso clusters are important. Firstly, most policy aims at creating general favourable conditions rather than stimulating specific firms. Stimulating micro clusters entails the risk of degrading into stimulating specific firms, which disturbs the market mechanism in an economy. Hence, policy aims at creating possibilities that can support every firm that helps to achieve the goals of the government, which calls for meso economic analyses.

Secondly, the cluster concept may be useful in presenting the main results of sector studies. Results of analyses at the sectoral level are often very disaggregated. Furthermore, the results of a single sector also depend on the co-operation of this sector with other sectors. Difficulties to interpret the outcomes of sectoral studies arise because of the large amount of details and because the results are displayed out of a context. Clusters provide for the context and they are a way to meaningfully aggregate the sectoral results.

Thirdly, meso clusters can be used in international comparative analyses. Clusters show which sectors co-operate in different countries, thus showing differences between technologies used or between the goods produced. After all, a sector that produces food and buys its inputs from the agricultural sector will produce different goods than a sector that produces food and buys its inputs from the chemical sector.

Lastly, analyses of meso clusters are less difficult to perform than analyses of micro clusters. For the latter, a researcher often needs to rely on interviews or surveys, whereas analyses of meso clusters can be performed using a more objective quantitative method. Hence, the outcomes of analyses of meso clusters appear to be more reliable than the outcomes of analyses of meso clusters.

Because of the reasons above, the analysis below concerns meso clusters based on linkages. Hence, the operational definition of a cluster in this study amounts to:

*A cluster is a set of sectors that use relatively large amounts of each other's products.*

## 4. Empirical methods for identifying meso clusters

[3] Porter, M.E., 1998, *On Competition*, Boston: Harvard Business School Press.

[4] DeBresson, C., 1996, *Economic interdependence and innovative activity: an input-output analysis,* Cheltenham: Edward Elgar.

Roelandt *et al*[5] distinguish two empirical methods for identifying clusters, the *monographic method* and the *input-output method*. De Bresson and Hu[6] add a third method, the *graph method*. Since the innovation data required for this method are not available, the method chosen for this study has to be one of the other two methods.

The monographic method generally uses a cluster chart based on Porter's diamond[7] to identify the most important clusters in an economic system. It can be used to identify innovative clusters as well as value-added chains. Since this method generally involves interviews, surveys, and case-studies, the techniques used are more qualitative than quantitative.

The input-output method is the most quantitative method. Based on the linkages in an input-output table, the sectors that use each other's products are identified and grouped into clusters. The same method can easily be applied for several years or countries, which adds to the objective character of the method. Besides this advantage, there are more positive features associated with the input-output method. Input-output tables are available for most countries, hence data are much easier available than data needed for the monographic method. Furthermore, the input-output method is generally a straightforward method that is easy to apply. For these reasons, the current analysis uses the input-output method to identify clusters of sectors empirically.

## 5. The input-output method

Deriving an input-output method for identifying clusters requires choices concerning three elements: the *relation* between sectors, the *data* to be used, and the *technique* to be applied in grouping the sectors together in clusters.

[5] Roelandt, T.J.A., Den Hertog, P., Van Sinderen, J., Van den Hove, N., 1999, Cluster Analysis and Cluster Policy in the Netherlands, in: OECD, *Boosting Innovation: The Cluster Approach*, Paris: OECD.

[6] DeBresson, C., Hu, X., 1999, Identifying Clusters of Innovative Activity: A New Approach and a Toolbox, in: OECD, *Boosting Innovation: The Cluster Approach*, Paris: OECD.

[7] Porter, M.E., 1990, *The Competitive Advantage of Nations*, New York: Free Press.

*Table 2: Possibilities for the three elements of an input-output method*

| Relation | Data | Technique |
|---|---|---|
| Intermediate deliveries | values / flows | maximising |
| Innovations | input coefficients | decomposability |
| R&D | output coefficients | input-output model |
| | multipliers / spillovers | |
| | make and use tables | |

<u>The relation</u>
A considerable data collecting effort is needed to obtain input-output tables with innovations or R&D data, or the tables have to be estimated based on the tables with intermediate demand and sectoral innovation or R&D data. However, in theoretical studies as well as empirical studies, it is argued that intermediate deliveries are a good proxy for innovations. Hence, the proposed method for identifying clusters of sectors will be based on regular input-output tables with intermediate demand.

<u>The data</u>
For international analyses, it is hard to obtain make and use tables of several countries in the same sector classification. The availability of input-output tables is much better. Hence, the proposed method uses input-output tables. For the specific data, it is not necessary to limit the method to only one type of data. After all, all data have their own interpretation: values refer to the economic importance of a transaction for the economic system as a whole, input coefficients refer to the importance of a transaction for the buyer, and output coefficients refer to the importance of a transaction for the supplier. This study proposes to derive a method which takes these differences into account by using all data.

<u>The technique</u>
Most empirical analyses use a maximising procedure to derive clusters of sector. Generally, a table of data is chosen. Next, the largest element in this table is selected and the two sectors involved are added together in a cluster. This procedure is continued until a fixed number of clusters has been found.

Restricted maximising is a possible refinement of the maximising procedure. The procedure above is only used for these elements that fit certain restrictions concerning, for example, the height of the value of the transaction.

Besides the maximising procedure, it is possible to use decomposability. A decomposable matrix can be split up in groups of rows and columns that have no relation with all other groups. A similar procedure has already been used to find the fundamental structure of input-output tables.[8] Of course, a decomposable matrix leads to a natural split up of sector into clusters.

Finally, the input-output model can be used to derive clusters with a technique resembling hypothetical extraction[9]. It is possible to derive a new input-output matrix is by selecting a cluster division and putting all transactions between sectors in different clusters to zero. Total output with this new matrix and the same final demand will be lower than in the case of the original matrix. The cluster division that has the least loss of total output is the best cluster division.

## 6. The Results of three different methods

For the cluster identification method used here, a procedure is selected that uses both intermediate good flows and input and output coefficients. For the technique, it was not yet decided whether to use a

[8] Simpson D., Tsukui, J., 1965, The Fundamental Structure of Input-Output Tables, an International Approach, *Review of Economics and Statistics*, vol. 47, blz. 434-446

[9] Dietzenbacher, E., Van der Linden, J.A., Steenge, A.E., 1993, The Regional Extraction Method: EC Input-Output Comparisons, *Economic Systems Research*, vol. 5, blz. 185-206.

maximising procedure, decomposability, or the input-output model. This last choice is made by trying these methods empirically and comparing the outcomes.

The method that uses the input-output model proves to be not very well suited for identifying clusters. It appears that the output loss depends on the number of elements put to zero. The less elements put to zero, the less the output loss. Since this number is the least when only one sector is separated from all other sectors, this method tends to result in many 'clusters' of only one sector and one cluster of all other sectors. Hence, three possible techniques remain:

* maximising
* restricted maximising
* decomposability

These three methods are applied to the 1998 input-output table of the Netherlands in 1998 distinguishing 105 sectors. This exercise shows some interesting results, concerning mini-clusters, mega-clusters, inconsistent results, and the use of restrictions. These results are discussed below.

<u>Mini-clusters</u>
Many of the applied methods result in clusters of no more than two sectors. Apparently, these two sectors have a strong link with each other and weak links with all other sectors. The resulting cluster structure is difficult to interpret: just a few mini-clusters are found, most other sectors are assigned to one mega-cluster, which is the second noticeable result:

<u>Mega-clusters</u>
Besides the mini-clusters, many methods yield one mega-cluster that includes most of the sectors. This result is plausible for the maximising method, especially if maximising is based on flows. If two sectors are aggregated into one cluster, their transactions with the other sectors are aggregated. Since their transactions are aggregated, it is likely that the next largest element is found in this new, larger cluster. Furthermore, if two clusters arise, the transaction between these two clusters exists of the aggregated figures of all transactions between the sectors in both clusters. Hence, there is a good chance that the next round aggregates these two clusters into a new one. Although this explains the forming of a mega-cluster, the interpretation remains difficult.

<u>Inconsistent results</u>
Some clusters are found by each method and with each type of data; many other clusters are only found by some methods and with some data. Of course, this is a consequence of the method used: a different method leads to a (slightly) different type of cluster with a (slightly) different interpretation. For example using a maximising procedure based on input coefficients leads to a cluster with most important suppliers whereas the same procedure based on output coefficients stresses the importance of buyers.

<u>Restrictions</u>
Two procedures use restrictions on the data. The restricted maximisation procedure adds together the sectors that share the largest element in the specified data table, but only if these elements satisfy certain restrictions concerning the value of the transaction, the input coefficient and/or the output coefficient. These restrictions guarantee that clusters are based on transactions that are important for the entire economic system, for suppliers, and for buyers.

Similarly, the procedure based on decomposability is applied to a table in which elements that do not satisfy certain restrictions are put to zero. This also guarantees that clusters are based on elements that are important for the economy as a whole, for suppliers, and for buyers. The restrictions are an intrinsic reason for making a matrix decomposable. A regular input-output table generally contains many zeros, but not enough to guarantee decomposability. Due to the restrictions, the transactions that are least important are ignored, which results in the 'core' of the economic system.

Although restrictions are required in the case of the decomposition method, for the maximising procedure they are not necessary. Without the restrictions, the procedure also leads to a satisfying cluster division. It appeared that the restrictions largely influence the outcome of the maximising procedure. The dependence of the results on the restrictions is strong enough to demand for a justification.

A last point concerns the specification of the restrictions. Most empirical analyses that use a restricted maximising procedure do not justify the restrictions they apply to the data. It is very difficult to explain which restrictions should be needed. Many restrictions can be thought of, each leading to different results whereas theoretically it is unclear which restrictions are to be preferred. Therefore, the application of restrictions is to some extent arbitrary.

## 7. The selected method

Most problems described above appear only in the maximising procedures. The results of the procedure based on decomposability turn out to be more harmonious. Few or zero mini-clusters and no mega-clusters were found. Furthermore, the data used in the method (input coefficients, output coefficients, or values) does not influence the results. This result always holds, since the procedure uses the structure of the table after several unimportant elements have been put to zero. A value put to zero implies that both the input coefficient and the output coefficient associated with this transaction become zeroes as well. Therefore, the decomposability technique is chosen for the cluster identification method.

To select the exact method, the restrictions still have to be specified. The proposed way to select the restrictions is by using the distribution of the data. If it is possible to find a statistical distribution that describes the data, the mean and standard deviation lead to appropriate restrictions. In that case, the restrictions resemble statistical testing. However, it turn out that the data are not easily described by a known statistical method. Therefore, the restrictions are set in terms of the empirical distribution of the elements themselves. To do so, a significance level is specified, after which the largest elements are chosen according to this significance level. For example, if the significance level is 5%, the 5% largest elements are used in the analysis; all other elements are put to zero.

## 8. Results

The cluster identification method is applied to two time series. The first series consists of a input-output tables according to an aggregated sector classification of 52 sectors, which is known for a large period: from 1969 to 1992. The second series consists of a more detailed time series of 105 sectors for a short period, from 1995 to 1999. Both series are obtained from Statistics Netherlands, the first series has been adjusted for reasons of consistency in the sector classification over time.

The aggregated series shows a remarkable consistence of clusters over time. Generally, four clusters are identified: the agro-food cluster, the energy cluster, the building cluster and the trade cluster. In most years a metal cluster is found as well, but this cluster seems to have disappeared in the later years. The more detailed tables identify more clusters. In that case we find the following clusters:

* Energy
* Meat
* Chemicals
* Building
* Cars
* Trade
* Financial
* Travel
* Media
* Mail, communication, and banks

Both time series show a large degree of time consistency of the clusters. Clusters tend to be more similar between years that are close to each other. The further the years are apart, the larger are the differences between the identified clusters. Lastly, it appears that more detailed data show more clusters than aggregated data. Aggregation only seems to matter for the number of clusters found, not for the clusters themselves: if a cluster has been identified with aggregated data, this cluster also pops up in the detailed data.

## 9. Conclusions

This paper describes several methods that can be used to identify cluster empirically. The method selected uses an input-output table that is made decomposable to identify clusters. Next, each part of the decomposed table is a cluster, since it contains sectors that use each other's input relatively intensive whereas they only make little use of the products of sectors in other clusters.

However, a discrepancy remains between clusters in theoretical analyses and clusters in empirical analyses. Theoretical analyses generally refer to clusters of firms that are located at the same geographical spot and that co-operate in developing innovative efforts. Empirical analyses generally identify sectors that have large linkages in production. Although for both theoretical and empirical reasons the empirical meso clusters can be interpreted as a proxy for the theoretical micro clusters, the discrepancy is not fully solved. The outcomes also suggest that the meso clusters do not identify all clusters present in an economic system. The number of clusters depends, for example, on the level of aggregation of the data. Therefore, policy aimed at stimulating innovations by stimulating clusters should be based on more than just these meso clusters. Detailed studies of micro clusters are also needed to find all clusters in an economic system and to understand which firms co-operate and how and why they develop innovations. Still, meso clusters provide an elegant way to analyse which sectors work closely together and they provide a general framework which can be used as a starting point for more detailed cluster studies at the micro level.