A STATISTICAL METHODOLOGY TO ESTIMATE CLUSTER OF ENTERPRISES IN THE ANALYSIS OF PORT SYSTEMS

Adriana Monte, Gabriella Schoier

1. Introduction

In this paper we present a solution to the problem of identification of dense clusters of port system enterprises. We consider a modification of an algorithm proposed in social network analysis which has been applied with success in different fields. The basic idea of social network analysis is an assumption of the importance of relationships among interacting units. Theories, models, and applications are expressed in terms of relational concepts or processes. Social network analysis is characterized by a methodology which gathers different techniques such as graphs, matrices, structural and location properties with the notions of centrality, structural balance transitivity and cohesive subgroups; role and position with the notions of structural equivalence, blockmodeling, relational algebra, network position and roles, different statistical methods, etc…

Two classical social network theories provide insights that can help in analysing networks. First, the small-world literature has shown that while most of our acquaintances tend to be acquainted with each other, short acquaintanceship chains (relative to the size of the network) link most pairs in the network. This high degree of local clustering suggests that a practical approach to studying the structure of networks would involve first identifying local clusters and then analysing the relations within or between clusters. Second, we know from work on peer influence models that close units tend to be similar to each other.

The application starts from the results of a research project aimed at identifying the main economic and industrial characteristics of the port system of the Friuli Venezia Giulia Region, Italy, and the role it plays within the economy. Combining a top-down and bottom-up approach, based on interviews and detailed data at firm level, a total number of 480 authorized firms has been considered. The research has been performed in two phases. In the first one, in 2008, all the enterprises of the Regional Port System (RPS) are investigated in order to collect data about the dimension (as sales and employees), economic activity, organization and other aspects. In this phase the sources of data are the enterprise budgets and interviews by phone. In the second phase, in 2010, the aim has been the estimate of input

output tables. The information about income and expenditure flows, by economic activity of the purchasing and supplying enterprises, come from a sample survey (face to face interview). In this case only the enterprises with deposited budget are considered. Information obtained in the second phase to estimate the intermediate consumption has been used to identify clusters of "similar" enterprises. We use relations between sectors, analyzing from which sectors an enterprise buys the input and to which sectors an enterprise sells the output. In paragraph two we described the port system and the data set used in the analysis. In paragraph three the algorithm is presented while in the last paragraph the application has been performed.

2. Problem definition

2.1 The Regional Port System

The Friuli Venezia Giulia (FVG) Region, an Italian region located on the North East of Italy and bordering with Austria and Slovenia, utilizes three ports for goods shipping: Trieste, Monfalcone and Porto Nogaro. They form what we have called the FVG port system. They are all situated in the North Adriatic and, geographically and historically, serve the North East of Italy and the so called Mittle-European countries, that is, Southern Germany, Austria, Hungary and the Check Republic.

Figure 1 – *F.V.G. Region*

Geographically, the port of Trieste is conveniently located. It lies 1,294 marine miles away from Suez (Port Said), equivalent to 2 days and 16 hours navigation at 20 nautical miles, whereas the Northern range ports are 3,527 marine miles away, equivalent to 7 days and 8 hours navigation at 20 nautical miles. The distances from the port of Trieste of some of the main surrounding cities is as follow: Milan 411 km, Munich 480 km, Graz 288 km, Budapest 539 km, Wien 472 km, Prague 865 km, Zagreb 222 km. The port has a considerable draft: 18 meters minimum along the docks. The total available area is 2.3 km^2 (of which 1.8 custom free); the area available for deposit is equal to 925,000 m^2; the docks length is equal to 12 km. The port is divided into five sections, two of which are devoted to industrial activities, including the Trieste-Ingolstadt pipeline.

The throughput in term of tons of the port of Trieste is rather stable between 2002 and 2007, with a large component of liquid fuel. The share of containers is growing but small in absolute terms. Ro-Ro/ferry transport of trucks from Turkey is quite relevant.

The port di Monfalcone, which lies closer to the hinterland than Trieste, has an area of 0.68 km^2 with a draft varying between 9.5 and 11.7 meters. Various industrial activities are located nearby, including the largest Italian shipyard. The port specializes in general cargo and dry bulk cargo such as kaolin, coal, cellulose, cement, grains, wood, machinery, minerals, stones, steel and iron, iron scraps, vehicles. In 2007 the port throughput was 4,411,900 tons.

The port of Nogaro operates nearby the river Corno. It is rather small and with draft of between 4.5 and 7.5 meters. It benefits of a large hinterland area and it serves the metallurgical and wood industries located nearby. In 2007 its throughput has been 1,455,000 tons.

2.2 The data

The port is a delimited area the access of which is allowed by the Port Authority. The enterprises population is formed by all those who have the formal authorization of the Port Authorities. These enterprises are various and the economic activities are disparate. In the RPS coexist enterprises whom activity is sometimes not related with the typical port activity (Danielis, 2011), (Musso,1999). Sometimes the classification of economic activities *Ateco* is not suitable to distinguish between port economic activities and other activities, consequently in the research was adopted an *ad hoc* classification.

The enterprises involved in the Friuli Venezia Giulia RPS were classified into 16 different sector activities, then grouped in three macro sectors: Port sector (P) – the enterprises with a typical port activity (9 types of activity as shipping and customs agents, shippers, marine terminal operators, …), Support to port sector

(SP)– enterprises that operate an inland or maritime transport activity (3 types of activity), No port sector (NP) - various enterprises whose activity is not a port one, but they carry out an activity in construction, manufacturing, trade or services sector.

Some information needed to estimate the input-output tables were collected by a sample survey. As we can see from Table1 the response rate was 50% for the enterprises of P sector and less for the others.

Table 1 – *Firms per macro sector in census analysis and in sample survey.*

Macro sector	Total enterprises	Enterpries with budget	Sample Size	Respondent Enterprises
P - Port sector	244	170	89	44
SP - Support to port sector	71	29	70	6
NP - No port sector	165	104		10
Total	480	303	159	60

In order to estimate intermediate consumption for the input output table, there is a crucial question in the questionnaire used in the sample survey. The enterprises were asked to subdivide the sales into sixteen types of customers (in the sense of their economic activity). A similar question was made for the suppliers. We use these information in order to identify relations between the macro sectors.

3. The Algorithm

Our INWM (Iterative Neighbourhood Weighted Mean) algorithm is based on Moody's RNM (Recursive Neighbourhood Mean) algorithm (see Moody, 2001) and on a modified version the MRNM (Modified Recursive Neighbourhood Mean) algorithm (see Schoier et al. (2004)). Moody's algorithm is realized through a recursive algorithm allowing the units to adjust their behaviour according to the behaviour of their similar.

The principal advantage of our algorithm with respect to the previous one regards the possibility of handling data in a short time; in fact, in order to build up the groups only the matrix of the so called influence variables **Y**, which represents the positions of the units, are considered. This results in a reduced and more flexible structure on which different techniques such as cluster analysis may be used. The novelty relies on the fact that both the preceding algorithms need as input a binary matrix of the adjacencies representing the presence or absence of the relation among the units, while here it needs, as input, a matrix of weights so we

can not only use the piece of information deriving by the presence or absence of the relation but we can measure its strength.

THE INWM ALGORITHM

Given $u_1,\ldots,u_N$ *units on the network on which* $R_1,\ldots,R_k$ *relations are defined:*

Step 1. *Evaluate the matrix of weights on the basis of the relations among the units.*

Step 2. *Assign a uniform random number (between 0 and 1) to each unit for every of the m influence variables so to obtain the influence matrix* ***Y* (0)**.

Step 3. *Re-assign to each element of the matrix* ***Y (t)*** *, t=1,..., the weighted mean of the contacts on the basis of the matrix of weights*

$$Y_{ik}^{(t+1)} = \frac{\sum_{j \in L_i} Y_{jk}^{(t)} N_{ij}}{\sum_{j \in L_i} N_{ij}} \quad k = 1,\ldots, m,\ i = 1,\ldots, N$$

where L_i *is the sub-sample of 1,…,N corresponding to the units which are in relation with i, and* N_{ij} *is the weight between i and j .*

Step 4. *Repeat* Step 3. *n times.*

4. The Application to the Regional Port System

Given a finite set of N units E (i.e. the enterprises) we have considered the relation R:" sector *i* and *j* sell to one or more sectors and among these sectors there is the port one. On the basis of this relation the matrix of weights is built in this way

1→ *i* and *j* sell only to port sector,
0.5→ *i* and *j* sell to different (included the port one),
0 → otherwise.

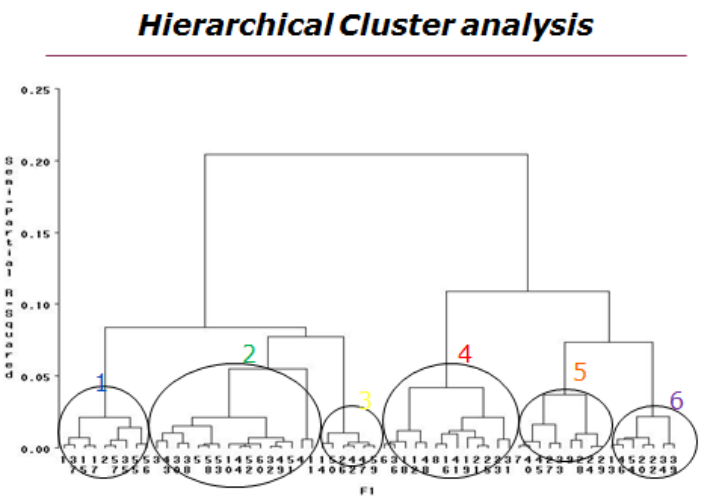
(E,R) is a network (set of units and relation(s) defined over it). We applied the algorithm and obtain a stable solution after three iterations. Five positional variables have been considered in the analysis. On the basis of these positional variables we create a hierarchical decomposition of the database by using the Euclidean distance and the Ward method. The hierarchical decomposition is represented by a dendrogram (see Fig.1).

The algorithm allows to identify six clusters, the smaller one includes six enterprises, the larger one sixteen.

Table 2 – *Positional variables*

	pos 1	pos 2	pos 3	pos 4	pos 5
1	0.534629	0.275460	0.845099	0.060548	0.344219
2	0.754435	0.296487	0.613288	0.324518	0.428758
3	0.631572	0.299255	0.387202	0.813333	0.440788
4	0.018259	0.504354	0.900876	0.944314	0.047119
5	0.351285	0.870331	0.64293	0.605837	0.068903
6	0.185441	0.258709	0.059592	0.725121	0.999659
7	0.698163	0.483717	0.781061	0.879577	0.918197
..	……..	……..	……..	……..	……

Figure 1 – *Hierarchical Cluster Analysis Representation*



The algorithm, we have used to individuate the clusters, is based on the idea that the enterprises are more or less similar according how strong is the relation with the enterprises whose activity is port (P). This relation is considered from the

point of view of the sales to such enterprises. The sale behavior lets us to identify the six clusters highlighted in the dendrogram.

In order to describe the clusters we use a set of variables: percentage of enterprises in P sector, if the enterprises work inside or outside the port, if they are mono activity or not, if they are part of a group of enterprises, where the employees are located, the economic activity of enterprise suppliers.

These information allow us to identify the smaller or higher integration of the port system towards the local economic system, in the sense of existing relations. The description of the six cluster is:

Cluster 1: enterprises working inside the port system, but in strong relation with NP and SP macro sectors and with the economic system outside the port.

Cluster2: The larger group, characterized by firms operating mainly into the port and belonging to enterprise groups, consequently with employed outside the port.

Cluster 3: monoactivity enterprises not part of enterprise groups, very linked with the port system (83% belongs to P macro sector) and their suppliers too.

Cluster 4: very near to cluster 2, but with a stronger connection with SP sectors suppliers and a higher localization of employees into the port.

Cluster 5: this cluster is more heterogeneous than the others, the enterprises work inside and outside the port.

Cluster 6: a restricted group of enterprises strongly integrated with the port, in major part belongig to enterprise groups.

5. Conclusions

In this paper we have presented a solution to the problem of identification of dense clusters in the analysis of the Regional Port System of the Friuli Venezia Giulia region, by considering a modification of an algorithm known from social network analysis. In doing so we have obtained an useful tool to study and profile enterprises in terms of their behaviour inside and outside the Port System. This allows us to build up useful business intelligence for the improvement of the Port System organization.

References

DANIELIS R. 2011. Il sistema marittimo-portuale del Friuli Venezia Giulia. Caratteristiche Strutturali e Interdipendenze Settoriali. In DANIELIS R. (Ed.), *Il sistema marittimo-portuale del Friuli Venezia Giulia. Aspetti economici, statistici, storici*, Trieste: EUT, pp.9-189.

MONTE A. 2011. Il Disegno dell'Indagine sul Sistema Portuale. In DANIELIS R. (Ed.), *Il sistema marittimo-portuale del Friuli Venezia Giulia. Aspetti economici, statistici, storici*, Trieste: EUT, pp.190-217.

MOODY J. 2001. Peer influence groups: identifying dense clusters in large networks, *Social Networks*, 23, pp.261-283.

MUSSO E., BENACCHIO M. E FERRARI C. 1999. Sistemi locali e impatto occupazionale: il caso dei porti, *Atti della XX Conferenza Italiana di Scienze Regionali, Piacenza 5-7ottobre 1999*.

SCHOIER G., MELFI G. 2004. A Different Approach for the Analysis of Web Access Logs. In VICHI M., MONARI P., MIGNANI S., MONTANARI A. (Eds.), *New Developments in Classification and Data Analysis*, Berlin: Springer, pp.211-216.

WASSERMAN S Wasserman S., FAUST K. 1994. *Social Network Analysis: Methods and Applications*, New York: Cambridge University Press.

SUMMARY

A statistical methodology to estimate cluster of enterprises in the analysis of port systems

In this paper we present a solution to the problem of identification of dense clusters in the analysis of port system enterprises. We consider a modification of an algorithm proposed in social network analysis which has been applied with success in different fields. A practical approach to studying the structure of networks would involve first identifying local clusters and then analyzing the relations within or between clusters. The application of methodology to data referring to the Friuli - Venezia Giulia seaport system (SPR) allows to identify four cluster of enterprises, that are integrated in different ways with the port activities.

Adriana MONTE, Dipartimento di Scienze Economiche Aziendali Matematiche e Statistiche, Università degli Studi di Trieste, adriana.monte@econ.units.it
Gabriella SCHOIER, Dipartimento di Scienze Economiche Aziendali Matematiche e Statistiche, Università degli Studi di Trieste, gabriella.schoier@econ.units.it