UNIVERSITÀ DEGLI STUDI DI SALERNO
DIPARTIMENTO DI SCIENZE ECONOMICHE E STATISTICHE

DOTTORATO DI RICERCA IN
INGEGNERIA ED ECONOMIA DELL'INNOVAZIONE
XIII CICLO

TESI DI DOTTORATO:

A SYMBOLIC DATA ANALYSIS APPROACH TO EXPLORE THE RELATION BETWEEN GOVERNANCE AND PERFORMANCE IN THE ITALIAN INDUSTRIAL DISTRICTS

COORDINATORE:
Ch.ma Prof.ssa Alessandra Amendola

TUTOR:
Ch.ma Prof.ssa Alessandra Amendola

CO-TUTOR:
Ch.mo Prof. Giuseppe Giordano

CANDIDATO:
Ilaria Primerano

ANNO ACCADEMICO 2014/'15

To my husband, Massimo,
my love, my life, my everything.

CONTENTS

INTRODUCTION

This work aims at analyzing complex phenomena through appropriate statistical methods that allow considering the knowledge hidden behind the classical data structure. Interesting innovations arise when looking a well-know topic in a new perspective. In other words, it is interesting to gain new knowledge, by continually reframing and reinterpreting events through the integration of new meanings within the complexity of concepts.

Nowadays, with the diffusion of huge repository, streaming data and web-based data, the management of very large dataset have become routine. So, the increasing interest of researchers have moved towards the definition of the most suitable methodologies in order to extract useful and meaningful information from this huge amount of data. The process of acquiring new knowledge from large datasets plays an increasingly important role in several application fields. However, the extraction of meaningful knowledge is still a highly non-trivial task which require to perform specific methodologies and tools. The use of Symbolic Data Analysis methods allows to manage these complex data structures in a suitable way.

Symbolic Data Analysis – SDA – consists of visualizing, classifying and reducing the information retrieved in a Symbolic Data Table. It aims at extending statistics and data mining methods from first-order

units (i.e. micro-data) to second order objects (i.e. concepts). The new data are independent form the initial dataset and preserve the inherent variability of the data.

The main aim of symbolic methods is to model the intent of a concept by describing the classes of individuals that compose its extent defined by a Symbolic Data Object, since that only a rough description of the concept can be reached. Symbolic data arises through a synthesis process of a huge dataset, usually too large to be conveniently managed and analyzed into the standard statistical framework. New statistical methodologies with new ways of thinking about data are required in order to discover latent knowledge.

The starting point of this work is to develop new contextual implications for future studies designed on the process of definition of an operationalization process and the statistical analysis of a theoretical constructs. The main proposal is to move from fist-level units to second-level units by means of statistical tools that allow to consider complex objects as a whole.

The application on real data can be useful to validate the importance of this approach. Specifically, the case of the Italian Industrial District model will be referred as the context of application. In this perspective, Italian firms are considered first-level units, instead the Industrial District concept is the second-level unit.

In recent decades there has been an enormous growth of interest in the notion of Industrial District. Traditionally, these entities have been addressed as territorial aggregations of both a community of people and a population of specialized firms located into determined geographical boundaries. Furthermore, the importance of governance systems is usually referred as the ability of public administration to manage and govern networks, involving all actors of civil society in political decision-making processes of the district area. Moving from the shared assumption that these governance systems are related with the district performance, we aim to explore this relation.

In contrast with the standard approach of Industrial District analysis, here the district is considered in its total complexity.

The challenge of this work is to overcame the classical approach applied in this specific field, mainly based on atomic sample data, by proposing the analysis of the Industrial District considered as a whole.

At this aim, the Symbolic Data framework provides the basis to face this shift of perspective. In fact, the definition of the Symbolic Industrial District allows to analyze this entity in its total complexity.

The main advantage of the proposed approach deals with the possibility to model concepts known only by their extent, without losing meaningful information by managing huge and complex databases. Moreover, the visualization of these concepts take into account the internal variation of the data, thus enriching the interpretation of the results.

This thesis is structured in four chapters organized as follows.

The *first chapter* gives an overview of the concept of Industrial District. Starting from the definition given by Alfred Marshall in the 20s, the different definitions available in literature are discussed. Several definition of Industrial Districts can be found in the vast body of literature on this topic. They mainly differ in the emphasis attributed to the importance of either normative-governance and cultural-cognitive elements for the characterization of districts.

The main goal of this chapter is to underline the peculiar characteristic of the phenomenon in the Italian economic and industrial context. It is a complex concept, fragmented into several important aspects. In particular, this chapter focuses on the structural and relational dimensions of Social Capital within Industrial Districts and the importance of the governance systems. Moreover, the legislative recognition of Industrial Districts in Italy is presented together with the available mappings of Italian Industrial Districts. In order to match literature and reality on this topic, some empirical studies are presented.

In the *second chapter* a new working definition of the Italian Industrial Districts is proposed in order to solve the ambiguities related to the lack of a concrete definition of this concept. Indeed, those socio-economic entities have been defined according to different interpretation of the phenomenon. Almost all these definitions agree in

the basic features of Industrial Districts, related to their territorial localization and productive specialization.

At our purposes, we define the Industrial District as a Complex Object by adopting the typical definition of the SDA approach. Starting from the generic definition of a Symbolic Data Object in this framework of analysis, the Symbolic Industrial District is defined. Therefore, the synthesis process that allows this transformation of the data, from first to second level units is discussed.

In the *third chapter* it is proposed a review of the main explorative SDA methodologies available in the literature. Symbolic Principal Component Analysis and Clustering Methods are presented together with the importance of using symbolic data (as interval-valued data, data distributions, etc.) instead of the classical ones (usually formatted as atomic data).

The SDA methods are the answer to the increasing interest of researchers towards the definition of the most suitable methodologies in order to extract useful and meaningful information from huge datasets. Symbolic Data Analysis has known, in recent years, a great development in terms of applications and methodological innovations.

The *fourth chapter* presents a case study performed on real data. The main research question is to explore the existence of a relation between Governance and Performance in the Italian Industrial Districts considered as a whole. To answer to the research question, an exploratory SDA is presented considering a subgroup of Italian Industrial Districts, suitable operationalized into Symbolic Industrial District.

The main idea is that it is possible to study the Industrial District concept by means of an aggregation of the first-level units in terms of the performance ratios expressed in terms of interval or histogram-valued variables. The study of such new entities by means of exploratory multidimensional data analysis allows to compare Symbolic Industrial Districts, to classify them into homogeneous clusters according to similarity measures and to represent them in a reduced space.

Working on different subsets of the initial dataset, the Symbolic Principal Component Analysis for histogram-valued data highlights, through the factorial maps representations, the main relationships among performance ratios and, reducing the redundancy of the data, allows to discover useful patterns into the data. Furthermore, the hierarchical classification underlines the presence of homogeneous groups of Symbolic Industrial Districts.

Concluding remarks, together with some suggestions for future researches are finally reported.

1 ON THE DEFINITION OF THE ITALIAN INDUSTRIAL DISTRICT. FROM LITERATURE TO REALITY

1.1 Introduction

The purpose of this chapter is to give a detailed overview of the concept of Industrial District. Starting from the definition given by Alfred Marshall in the '20s, the aim is to apply this fragmented concept into reality, trying to underline the peculiar characteristic of the phenomenon in the Italian economic and industrial context.

The starting point are the different definitions available in literature, from different point of view and in many application fields. Several empirical studies on Industrial District have been produced, even if the leitmotiv is not always the same above all for the Italian context. To solve these ambiguities, we try to match the different definitions available in order to operationalize and make comparable the research object.

The attempt is to integrate the existing research carried out by Italian scholars paying particular attention to the system of governance and legislation related to this topic.

In section 1.2 the classic definitions of the Industrial District are introduced. The main features of this phenomenon are discussed in section 1.3. In particular, section 1.4 introduces the structural and relational dimensions of Social Capital within Industrial Districts. Moreover, the importance of governance systems in this framework is addressed in section 1.5. Section 1.6 is concerned with the official recognition of Industrial Districts in Italy and section 1.7 deals with the mapping of Italian Industrial Districts. Finally, section 1.8 underlines some important empirical research on this topic. Some concluding remarks are given in section 1.9.

1.2 The Industrial District: some definitions of the concept

The definition of the concept of the Industrial District has been for decades matter of interest and scientific discussion by researchers belonging to different disciplines: from industrial economics to sociology, economic geography to business administration, and management to industrial policy.

The interest on this topic shown by several researchers is mainly due to the difficulty in the interpretation and definition of the key-success that this form of industrial organization has collected.

Industrial Districts represent an original form of firms agglomeration, a manufacturing system characterized by a strong industrial specialization whose production is targeted to a specific production sector. Traditionally, these agglomerates were made up of small and tiny industries, related by solid connections.

The first definition of the Industrial District dates back to the early of the 19th century in the works of the English economist Alfred Marshall. In his Principles of Economics (1920), he analyzed the importance of *external economies* which makes possible to identify the main benefits of spatial concentration (local) and specialization (sector), in contrast with the internal economies of firm, resulting from the company size.

According to external economies, small businesses succeed in achieve the typical advantages of large-scale production. Due to a

strong concentration into a well-defined geographical area, as underlined in Sforzi and Lorenzini (2002), they are able to promote:

- the reproduction of skills;
- the spread of knowledge;
- the development of support activities in both manufacturing and services thanks to the variety in production;
- the use of specialized equipment;
- the formation of a skilled labor market;
- the development of complementary industries thanks to a diversification in employment.

Marshall puts the basis for the transition from the traditional unit of analysis in the economic-industrial framework, i.e. industry, to an intermediate level labeled as Industrial District. He classifies this phenomenon also as *the concentration of specialized industries in particular localities*.

The emphasis placed by Marshall concerns not only the economies of specialization relevant in this context, but also and above all the structural characteristics of these agglomerations.

The two dominant characteristics of a Marshallian Industrial District are high degrees of vertical and horizontal specialization and a very heavy reliance on market mechanism for exchange (Robertson and Langlois, 1995).

Firms located into Industrial Districts tend to be small in dimension and to focus on a single function in the production chain. The major advantages of Marshallian Industrial Districts arise from contiguity of the firms, which allows easier recruitment of skilled labor. At the same time, rapid exchanges of commercial and technical information through informal channels are speeded-up (Langlois and Robertson, 2002).

In particular, Marshall identifies and defines an *industrial atmosphere* within geographic concentration of skilled workers. Workers appear to be committed to the district rather than to the firm while the district is seen as a *relatively stable community* that promote the development of a strong local cultural identity and the sharing of industrial skills (Alberti, 2010). Regarding so, Marshall writes:

> *In districts in which manufactures have long been domiciled, a habit of responsibility, of carefulness and promptitude in handling expensive machinery and materials becomes the common property of all ... The mysteries of industry become no mysteries; but are as it were in the air, and children learn many of them unconsciously*" (Marshall, 1920).

Marshall underlines the mutual influence between the social and the economic systems, but he did neither very much elaborate on this idea, nor on its social foundations (Belussi, 2001).

However, this issue is very clear in the Italian geographical and industrial context. The main contribution to the definition of the Italian Industrial District is due to the researches of Giacomo Becattini (1979, 1991, 2000, 2004).

In his studies on the topic, he provides a detailed re-reading of the Marshallian concept of the Industrial District. Becattini defines the Industrial District as:

> *a socio-territorial entity which is characterized by the active presence of both a community of people and a population of firms in a natural and historical bounded area (Becattini, 1990)1.*

The Marshallian concept of *industrial atmosphere* has been changed by Becattini into *belong feeling*, underlining the districts communities' tendency to identify themselves with the district, i.e. to feel part of the productive system.

Therefore, the relevance that Industrial District have taken for the Italian economy and society, up to constitute a distinctive element, has also engendered an intense research on such topic. The specificity of the Italian case will be explored in the next sections.

[1] According to Becattini's original language definition, Industrial District is: «un'entità socio-territoriale caratterizzata dalla compresenza attiva, in un'area territoriale circoscritta, naturalisticamente e storicamente determinata, di una comunità di persone e di una popolazione di imprese industriali. Nel distretto, a differenza di quanto accade in altri ambienti (ad esempio la città manifatturiera), la comunità e le imprese tendono, per così dire, ad interpenetrarsi a vicenda» (Becattini, 1998)

1.3 Main features of Industrial Districts

The industrial development model based on districts has found in Italy the ideal conditions for its assertion since the seventies, in a context characterized by the first signals of the large company crisis, after the great industrial development that characterized the sixties.

The affirmation of a new growth process emerges in a context marked characterized by the constant search for new strategies of economic and industrial growth and by the reorganization of the productive sector. It is constituted by the presence of a network of small craft businesses, strongly rooted into the traditional production of restricted geographical areas. These firms gradually reached significant market shares in niche products, based on the potential of specialization and the division of labor between firms belonging to the same sector.

The intuition that similar or complementary firms tend to concentrate into a bounded geographical space, as already said, dates back to Alfred Marshall and his formulation about the importance of external economies.

The Marshallian theories, adapted to the Italian context of the seventies, have kicked off to a huge production of scientific works about Industrial District and, in general, geographical and industrial agglomeration. The district has thus become the object of study by many scholars from economists to sociologists (among others, Brusco, 1989, 1990; Pyke et al, 1990; Signorini, 1994, Dei Ottati, 1995; Becattini, 1998; Bellandi and Sforzi, 2001).

The territory is seen as *the environment in which a network can grow and develop and promote firm innovation* (Boari and Lipparini, 1999). Geographical proximity among firms is essential in this context.

Agglomerations, and particularly clusters and Industrial Districts, have been identified as places in which

> *close inter–firm communication, socio–cultural structures and institutional environment may stimulate socially and territorially embedded collective learning and continuous innovations* (Asheim and Isaksen, 2002).

The idea that all economic behavior in industrial agglomerations are embedded in inter-firm networks (Sabel, 1989; Pyke and Senegenberger, 1992) represent a central argument in district literature. Enterprises located in Industrial Districts are always looking for new forms of cooperation. Following this research line, Brusco and Sabel (1981) affirm that «*the very survival [of the district firm] is linked to the collective efforts of the community to which it belongs and whose property it must defend*».

To investigate the effects of such structure on the innovative performance of firms, it is important to underline the relationships and the collaboration among firms (Boari and Lipparini 1999, Tallman et al. 2004, Muscio 2006) in terms of knowledge flows (Belussi and Pilotti 2003, Noteboom 2004), and mechanism of social capital and trust (Dei Ottati 1994a, 1994b; Lorenz 1988, 1999).

In literature, the importance of knowledge transfer has been widely discussed. This feature is emphasized as a strategic issue for firm competition. In the Italian industrial scenario, the knowledge transfer process inside the district is strictly associated with the social relationships that exist among firms and between these and the local institutions involved.

The localization of a firm within a specific Industrial District can improve the capability of its employees to generate, diffuse and engage *tacit knowledge* (Polanyi, 1958). Thanks to such kind of industrial concentration, skilled personnel can be moved from one firm to another. The interactions between producers and users become more and more easy. The reputation effects are strengthened, the risks of opportunistic behavior decrease and the exchanges of information between competitors are allowed (Desrochers, 2001). The main reason is that tacit knowledge may be exchanged by people involved in its creation or by those being part of the same local or epistemic community.

Firms co–localized within a geographical cluster have an enhanced ability to create knowledge flows and new knowledge (Maskell, 2001). Moreover, tacit knowledge promotes innovation processes (Belussi and Pilotti 2003), thanks to the interaction between people and businesses enabled by the cooperation networks promoted by the local district cultural background.

1.4 Social Capital within Industrial Districts

The common background, typical of the Industrial District definition, is the necessary prerequisite for the process of knowledge flow. At the same time, a well established literature agrees in supporting the idea that the district's competitive assets is made by a combination of local tacit knowledge and external codified knowledge (Becattini and Rullani, 1996), even if the process that lead to such successful combination in not very clear.

The innovative dynamism that characterizes Industrial Districts lies in their ability to integrate external codified knowledge absorbed by distant actors with the tacit local and to spread it towards its members.

Boshma and Lambooy (2002), exploring these dimensions of the phenomenon, point out that Industrial Districts are characterized by a network type of coordination of economic relations, associated with horizontal trust-based relations among local firms and between those firms and institutions. Thanks to short distance, trust among firms means, at the same time, easier access to knowledge. As developed in the theory of *Transaction Costs Economics* by Williamson (1985; 1996) the organization of enterprises is strictly related to the organization of market.

Hybrid structure, such as network, can emerge, since the boundary of the firms can be changed in order to select the coordination mechanism to allocate resources and organize transactions.

A complementary approach, known as *Institutional Economics* (North 1989, 1992; Coase 1992, 1998), based its theory on the individual actor-oriented perspective taking into account the relationships of different actors inside an institutional environment. In this context, the actors to consider are producers, sellers, distributors and buyers and the institution changes according to regions. This means that each region has its own production structure and different paths of development and equally different market structure.

Noteboom (1999) explored another important facet of region specification: the nature of *inter-firm relation* that he defines *alliances*. With this expression he refers not only to physical exchange of assets,

goods or products, but also of knowledge, trust and values (Storper, 1997).

In this framework each firm is considered as a part of a whole, each firs is involved into a complex structure made up of input-output relations, networks and chains whose key features are the social division of labor and the trust based cooperation. This local organization structure, in which all transactions are organized at the district level, contributes to the competitive advantage of the district.

As pointed out by Asheim (1996) cooperation is a very important strategy in order to promote innovation and to achieve a global competitive advantage. The balance between competiveness and competition among Industrial Districts members will lead to a better use of the available resources and thus to innovation (Dei Ottati, 1994; You and Wilkinson, 1994).

Inter-firm networking together with horizontal communication patterns and frequent movement of people are becoming more and more important to enhance learning capabilities (Ludvall & Johnson, 1994).

According to Lipparini and Lorenzoni (1994), the purpose is to create, through networking, *strategic advantages over competitors outside the network*. This means that, inside an Industrial District, a competitive advantage is reached internally through inter-firm cooperation and exploited externally through competition with outside firms.

Industrial Districts include not only specialized firms but also a set of local institutions, which are fundamental for the district competitiveness. They are local based supporting organizations, both private and public, supporting the whole firms in district, like universities, research centers, industrial policy agencies, technical consultancy or professional and trade associations.

Their role will always be conditioned by their context, since they are constantly in contact with different external circles and, at the same time, are close to district firms (Molina-Morales et al, 2011). Further than provide generic and specific services for the firms, local institutions act as repositories of knowledge and opportunities in the territorial networks (Baum and Oliver, 1992) in order to improve competitive capacities. As a consequence, district firms can take

advantage of having a network of ties with these local institutions, since they act as leader actors in processes of innovation and improvement.

Although there are many scholars who share the idyllic vision of the district outlined above, many others emphasize some critical issues.

On one side there are those who share the definition of an Industrial District as a *cohesive social environment* (Becattini, 1990), consisting of companies that share values, rules and common languages. Knowledge and information flow freely among its members, thanks to the cultural proximity, the closeness and the spread of informal relations. Local tacit knowledge of a district differentiates its members from companies located outside its borders. The latter cannot benefit from the onsite externalities (Becattini and Rullani, 1996).

On the other side, many scholars disagree with these assumptions. These criticisms, reported to different disciplines ranging from innovation to economic geography, share the vision of the district as *a community of undifferentiated small businesses* (Morrison, 2008).

According to some empirical studies conducted in Italy (Albino et al, 1999; Grassi and Pagani, 1999; Bellandi 2001; Lissoni, 2001; Breschi e Lissoni, 2001) reality has changed and this change has also affected the district structure. They show that the element characterizing the development of an Italian district can no longer be attributed to shared strategies, but to the individual business strategies. In this context, the strategies and skills of individual firms play a central role in explaining the district dynamics.

The emergence of heterogeneous actors, including the leader firms, influences the district organizational structure. Being equipped with the best technology and higher propensity to investment than small and medium firms typical of the Industrial District, leader firms gain access to a broader set of knowledge and external information (Belussi, 2003). According to Morrison (2008), Leader firms act as *gatekeepers of knowledge* (Allen, 1977) if they are able not only to search external knowledge and to absorb it internally, but also to share the acquired knowledge with all the other district members.

It follows that knowledge is no longer shared freely among actors, but becomes a specific and personal benefit, and it depends both on the disposition of sharing of the leading firm and on the cognitive distance

(Boshma, 2005) between the latter and the other companies in the district.

On the diffused acceptance that physical proximity facilitates the freely knowledge sharing and the learning process in bounded geographical areas, some authors (among others Lissoni, 2001; Breschi and Lissoni, 2001; Capello and Faggian, 2005; Morrison and Rabelotti, 2009), argue that each local agglomeration is characterized by different types knowledge flows. They distinguish between free access flows – informational networks – and the restrictive ones – knowledge flows. These critical approaches agree with the idea that knowledge does not circulate freely among all members of a district, but is constrained within small groups, thus leading the creation of multi-levels network (Lisson and Pagani 2003; Morrison and Rabellotti, 2009).

Industrial District is considered as real network, made up of nodes and links, even if many studies of district networks are silent about this network structure. They are more descriptive than analytical. According to Udo Staber (2001) firm networks are *an important defining characteristic of Industrial District*. It is a *coherent and innovative system of relational contracting* able to bind firms together designed to realize collaborative product development and multiplex inter-organizational alliances.

All the economic actions in Industrial Districts are said to be embedded in a dense network of ties involving individuals, firms and service organizations located in that area.

1.5 The Governance of Industrial Districts

The term *governance* is widely used in very different field of application, even if its definition is still ambiguous leading to multiple interpretations.

The most shared definition of governance deals with the ability of public administration to manage and direct networks, involving all actors of civil society in political decision-making processes (Pastore and Tommaso, 2013).

Following Bagnasco (2009)

> *the term governance suggests that within the complex frameworks of contemporary societies even public policies are formulated through the direct participation of various public and private players who negotiate and reach agreements in order to ensure implementation*.

Concerning the Industrial Districts, all the issues discussed in the previous sections (i.e. inter-firm collaboration, networking and relationships of interdependence, reciprocity and trust, transparency and sharing information between a multiplicity of actors) contribute to define their governance.

Previously introduced by Marshall and then developed by scholars in different disciplines, the interest towards the study of this topic has increased a lot over the years. About the development and features of Industrial Districts, Marshall (1890) underlines the relevance of business relationships established into a specific local environment along with the importance assumed by socio-cultural aspects of this phenomenon.

In recent decades, also the Italian debate on local development in general, and on the Industrial District, in particular, stresses the importance of the process of social and productive integration. It is defined as facilitator of the process of development and systemic organization of the economy and the local society (Garofoli, 2006). This process involves all private and public local actors. Not only the enterprises system is involved, but also public institutions, intermediary organizations and interest associations are implicated in this process. All the stakeholders contributing to the social economy are involved in the governance of their territorial system, each with its own capabilities, skills and knowledge.

In literature on Industrial District, two main basic patterns of governance are identified considering the power distribution between district stakeholders (firms and public or private entities): the horizontal and the vertical governance models.

The horizontal governance model is based on the symmetrical power between firms. Instead, the vertical model supposes a hierarchy between agents, thus an asymmetrical power. Without a hierarchy, governance is obtained through spontaneous relationships: a set of localized formal and informal institutions, both public and private,

regulates what can be considered as an acceptable business behavior in the area. On the other hand, according to the vertical model, relationships between parties are planned and guided by executive organisms and coordination tools that assume the strategic control of the whole district. (Storper and Harrison, 1991; Golinelli, 2005; Provan and Kenis, 2007, Kerstin, Mele et al., 2008).

As already said, this is not the only a unique perspective about Industrial District governance. Another relevant study on this topic concerns the organizational and management theories (Coda, 1989). Considering the Industrial District as the unit of analysis, this topic has been investigated emphasizing the role assumed by particular actors.

Many authors acknowledge the relevance of these actors while addressing them with different labels: *meta-managing actors* (Marelli, 1997 and 1999; Brunetti and Visconti, 1999, borrowing this term from the concept of meta-management introduced by Normann, 1979)*, calaysts or pivots of local developments* (Garofoli, 1991), *social architects* and *collective entrepreneurs* (Alberti, 2001). In general terms, these actors act as agents of development and innovation in charge of the governance of the Industrial District.

Brusco stresses the importance of the role assumed by this agents, noting that *the lack of an apex, a vertex, a hierarchy or a governing center may constitute a major weakness for some Industrial Districts* (Pyke et all., 1990).

Most of the authors involved in the study of this topic pay attention to further themes related to these governing bodies in Industrial District, such as: i) their roles in the assessment of local entrepreneurial dynamics (Coda, 1989); ii) their specific activities in supporting and maintaining a favorable socio-economic context also towards the promotion of the district image and the creation of a shared consensus around central government authorities in the district (Sinatra, 1994); iii) the threats that may affect negatively their performance.

The heterogeneity attributed to these governance bodies lead some authors to narrow the criteria to guarantee their effectiveness. Among others, Molteni and Sainaghi (1997) define some other important requirements. They refer to several aspects as the organizational structure, the membership, the resource availability for supporting projects, the sharing of values and mission, the ability to plan and

control their capabilities, and the directed and undirected supply for specific services.

Dealing with this far-reaching topic, in Italy the law No. 317/1991 is issued in order to identify Industrial District and to define the actors entrusted with their governance. The *District Committee* is thus introduced in the Italian normative context. These actors can be both leading firms and both local institutions.

In this regard, Alberti (2002) identifies some distinctive attributes characterizing the district committee in his role of metamanager in the Industrial Districts. Specificallt, he refers to the legitimacy, the power, the presence and the use of knowledge, and the cohesiveness.

According to Bagnasco (2009), their fundamental task is to combine the different interests and to ensure a unified strategic direction to the district. With regard to this normative perspective, we can find empirical studies that underline the presence of other entities which pay a strategic role within the districts governance model.

Visconti (1996) includes the following actors: local government offices, Provinces, Chambers of Commerce, local banks and other financial institutions and service centers. Even if Industrial Districts are referred as the outcome of spontaneous processes within specific territories due to the interactive relations between firms and local institutions, this formal acknowledgment of the governance grant a significant sharing of factors determining their competitive advantage.

The governance system embedded in a limited territory may affect the local level. Borrowing concepts from the *corporate governance* theories, the governance of Industrial District consists of the relationships among heterogeneous participants, internal and external stakeholders, which determine its direction and its performance, presenting different interests.

Different theoretical perspectives give as much definition of district committee, envisaging their roles, activities, composition, attributes and expected results[2]. Several theories give prominence to the support that the district committee has to internal stakeholders (Stewardship and Stakeholder theories). Others refer to all agents involved in the

[2] For a detailed review of these perspectives from the corporate governance literature refer to Alberti, 2001.

system without differences (Agency theory). Other theories base their conceptualizations on the availability of resources (Resource dependence and resource based theory).

The theory related to the normative aspects of governance is the Legalistic one. Through the interpretation of the Italian Law 317/1991, this perspective maintains that the governing bodies contribute to the performance of the Industrial District fulfilling their mandatory responsibility, as issued by law. It seems that the district committee acts as imposed from the outside, as a referent of the Central Government, and not by the direct expression of local stakeholders' interests.

Next paragraphs deal with the Italian Regional Legislation and the mapping of Industrial Districts located in each region. This recognition of the empirical normative evidence in Italy is very useful to identifies those territorial systems according to their location, governance system and main productive activity.

1.6 Italian Regional Laws

The importance assumed by the Industrial Districts in Italy dates back to the Seventies of the last century, at first in the field of scientific research through the writings of Becattini and other distinguished scholars of local systems of small businesses, after also in the field of industrial and regional policy.

The strong rising in this area in the late '70s and early '80s, reveals significant changes of the competitive scenario, along with the need to innovate processes and products to reach new markets. It is not just a structural change, but also additional factors related to the offer of services, until then considered marginal, begin to take on a key role. This is the case of aspects related to quality, image and additional services to be offered to customers. In other words, product design, commercialization and marketing become very relevant.

Industrial Districts are faced with the urgent need for adaptations of structure, management and trade in a context, as the Italian one, where the measures of industrial policy was not yet able to stimulate the growth of small business systems.

The protagonists of this change are the local government authorities, subordinate or superordinate to Regions and other subjects, who incentivize industrial development. Regions, Municipalities and Provinces, supported by business organizations and by local and national research institutions and training organization, begin several intervention procedures to support local growth.

Each actor, with its expertise, contributes to this development with different tools. Local government authorities develop forms of financial support. Chambers of Commerce provide direct services to enterprises. Research institutions and training organization (as Universities and CNR) support the development of projects by providing the main technical inputs and scientific works needed for their implementation. Business organizations participate as financial supporters of the initiatives and as potential users of the provided services (Caloffi, 2000).

The ultimate goals of the Italian regions is to increase economic development, social cohesion, employment and, in particular, to strengthen the competitiveness of the regional production system, as well as to seek and implement new lines of action.

The Italian Regions are certainly the main focus of investigation to observe the local industrial policies, above all after the Bassanini Law (1988) with which Regions take all the competences for the concessions for industrial benefits. The first presence of the Industrial District in the Italian Legislation dates back to 1991, specifically with the law n. 317, October 5th, 1991, *Interventi per l'innovazione e lo sviluppo delle piccole imprese.* This law defines the Industrial District with art. 36, paragraph 1, as a territorial area characterized by high concentration of firms, which have a particular productive specialization, where there is a special relationship between the presence of businesses and residents[3]. With the following paragraphs, the legislator entrusts to the Ministry of Industry, Trade and Crafts the

[3] The original language definition of the Industrial District, as recited by Law 317/1991, article 36, paragraph 1, is as follows: «si definiscono distretti industriali le aree territoriali locali caratterizzate da elevata concentrazione di piccole imprese, con particolare riferimento al rapporto tra la presenza delle imprese e la popolazione residente, nonché alla specializzazione produttiva dell'insieme delle imprese».

enactment of a special decree containing the benchmarks to allow the legislative intervention of Regions.

Each Region, taking into account of the views expressed by several organisms operating in their area (as regional unions of chambers of commerce, industry, trade and agriculture), has to identify those areas in their territory that could be considered as an Industrial District. Italian Regions are also responsible for granting loans to finance innovative projects implemented in several companies, defining the priorities of the interventions to be implemented, according to the stipulation of a program contract with local industrial development consortia.

These tasks entrusted to Regions are formalized with the Law n. 112 issued on March 31th, 1998: *Conferimento di funzioni e compiti amministrativi dello Stato alle Regioni e agli Enti locali, in attuazione del capo I della L.15/03/1997, n. 59*. Regions are now called to handle the functions related to the granting of concessions, subsidies and benefits of different kind to the industrial sector.

The abovementioned Ministerial Decree is issued on April 21th, 1993: *Determinazione degli indirizzi e dei parametri di riferimento per l'individuazione, da parte delle Regioni, dei Distretti Industriali.* With this Decree, the Ministry of Industry identifies the Local Labor System (as defined by ISTAT[4]) as territorial area of reference. It also establishes some indicators and their quantitative thresholds values to be overcome so that a can be identified as an Industrial District.

The Italian legislative framework around Industrial District is renewed with Law No. 140 issued on May 5th, 1999: *Norme in materia di attività produttive.* Paragraphs 1 to 3 of art. 36 of Law 317/1991 are replaced by paragraphs 8 and 9 art. 6 of Law 140/1999. Thanks to this legislative renewal, the *Local Production Systems* are introduced. They are defined as *homogeneous productive contexts, characterized by a high*

[4] The original language definition of Local Labour System given by Istat is «Entità socio-economica che compendia occupazione, acquisti, relazioni e opportunità sociali. Tali attività, limitate nel tempo e nello spazio, risultano accessibili sotto il vincolo della loro localizzazione e della loro durata, oltreché delle tecnologie di trasporto disponibili, data una base residenziale individuale e la necessità di farvi ritorno alla fine della giornata» (Istat, 1997).

concentration of firms, mainly small and medium-sized, and a particular internal organization[5].

Starting from this definition, also the Industrial District takes on a new definition. It is identified as a *local production system characterized by a high concentration of industrial companies as well as the specialization of business systems.* The Regions have the mission to identify the local production systems in their territory. At the same time, they define which of these systems have to be addressed as Industrial Districts.

Following this line, almost all the Italian Regions, although with different times, have approved its own regional decree. Although there is a considerable gap between the arrangement of regional and industrial policy instruments and their effective implementation, a legislative map of Italian Industrial Districts is here denoted. It follows the chronology of the initiatives implemented in the Italian Regions.

The first Italian Region to know of the importance linked to legal recognition of district areas in its territory is the *Lombardia* Region. A short time after the ministerial decree No. 317/1993, the regional council issues with Regional Council Resolution No. V/43192 dated 17/11/1993 in implementation of Regional Low No. 7 dated 22/02/1993. With these regulatory actions, the Regional Council identifies the Industrial Districts within its geographical area. Instead with Regional Council Resolution No. V/1049 dated 09/02/1994, it determines the content and the procedures in order to implement the related innovative programs development.

The next year, the *Friuli Venezia Giulia* Region issues its own norms. With Regional Council Resolution No. 2179 dated 27/05/1994 and its subsequent amendments the council identifies Industrial Districts.

In 1995, three Regions *Liguria*, *Toscana* and *Marche* shall identify Industrial Districts in their territorial bounds.

With Regional Council Resolution No. 496 dated 17/02/1995, the *Liguria* Region identifies those areas, even if the regulatory process has

[5] The original language definition of Local Production System is «contesti produttivi omogenei, caratterizzati da un'elevata concentrazione di imprese, prevalentemente di piccole e medie dimensioni, e da una peculiare organizzazione interna».

already begun the previous year, with Regional Low No. 43 dated 09/08/1994.

In the same year, in *Toscana* Region, the Regional Council enacts the Resolution No. 35 dated 07/02/1995 to identify such areas. Simultaneously, it approves the procedures for the implementation of the interventions. Following the changes introduced by Law 140/1999, the Region enacts a new resolution of the Regional Council, No. 69 of 21/02/2000 in order to update its map of Industrial Districts.

The *Marche* Region identifies for the first time the district areas located in its territory with Regional Council Resolution No. 225 dated 07/03/1995, while with Regional Council Resolution No. 3236 dated 21/12/1998 further Industrial Districts are defined.

In 1996, other two Italian Regions *Piemonte* and *Abruzzo* emit their resolutions. Although the map of districts located in *Piemonte* has undergone many changes, from the first Resolution No. 722-2183 dated 01/03/1994, to successive aggregations with the Regional Council Resolutions No. 250-9458 dated 18/06/1996 and No. 62-3705 dated 31/08/2001, until the least dated 26/02/2002 No. 227-6665.

In *Abruzzo*, the first two reference regulations are the Regional Council Resolutions No. 742 dated 07/03/1996 and No. 34 dated 23/07/1996. With these resolutions, Industrial Districts have been territorially defined, while the prior interventions to activate within their territories have been identified.

In 1997, other three Regions issue their own regulations in order to identify those areas located within their boundaries. This compliance occurs in *Campania* and *Sardegna.*

The *Campania* Region enacts its first regulation on this topic on 02/06/1997 with Regional Council Resolutions No. 59, definitively approved with Regional Council Resolutions No. 25 dated 15/11/1999. With these regulations the Regional Council identifies the Industrial Districts and, at the same time, prioritizes the criteria to be adopted for the promotion and the fulfillment of development projects.

In *Sardegna* Region, the recognition of Industrial Districts occurs thanks to the Regional Department of Industry, whose councilors on 07/08/1997 enact the Decree No. 377, proceeding the Regional Council Resolution No. 61/120 dated 30/12/1996.

In 1998, the *Veneto* Regional Council locates the Industrial Districts within its boundaries by enacting the Regional Council Resolutions No. 23 dated 03/03/1998 and the No. 79 dated 22/11/1999.

In the last decade also other Italian Regions brings into line their resolutions, even if the normative framework is not still complete.

In 2001, the Industrial Districts are identified in the *Basilicata* Region, according to the Regional Law No. 1 dated 23/01/2001 and with Regional Council Resolutions No. 1433 dated 25/06/2001. In the same year, the *Lazio* Region issues the Regional Law No. 36 dated 19/12/2001, while an year after the Regional Council enacts its Resolution No.135, dated 08/02/2002 in order to organize the local Industrial Districts.

The *Molise* Region recognizes the Industrial Districts and it establishes the procedures in terms of public founding, by issuing the Regional Law No. 8 dated 08/04/2004 and its subsequent amendments. In the same year, also the *Sicilia* Region updates its legislation in terms of Industrial Districts. These areas are identified thanks to the Regional Law No. 17 dated 28/12/2004, with the article No. 56, issued by the council member of the Regional Department for cooperation, trade, crafts and fishing. Subsequent amendments contribute to enhance the definition and the purpose of Industrial Districts (Department Decrees No. 152 dated 01/12/2005 and No. 179 dated 06/02/2008).

The only Industrial District located in the *Trentino Alto Adige* Region, is recognized through the Provincial Law No. 7 dated 24/10/2006.

In *Puglia* Region the legislative recognition occurs with Regional Law No. 23 dated 03/08/2007 and the Regional Council Resolutions No. 91 dated 31/01/2008.

The *Emilia Romagna* Region, instead, deserves a separate discussion. Though this region has not officially recognized the Industrial District by approving legislation, it has issued several measures in favor of such areas. In doing so, the identification of target areas of intervention takes place "bottom-up". In fact, the actors located in the Emilia are left free to organize themselves in order to design interventions in favor of specific bordering areas.

Other Italian Regions are still stationary regarding the issuance of rules and the formal recognition of the Industrial Districts. These Regions are: *Calabria*, *Umbria* e *Valle d'Aosta* (Cresta, 2008).

1.7 Mapping Italian Industrial Districts

The detailed map of the Italian Industrial Districts is not simply to be reconstructed since it lacks of an unambiguous definition. Different criteria are used in the definition of district reality. Therefore, the main core can be traced in the complex interlace that involve both socio-economic features both the production system made of small and medium specialized firms.

The standardization of these characteristics together with the complexity of production systems structure are not easy to be defined on statistical basis, because of their qualitative nature.

In fact, the most typical traits deal with different kind of relationships among firms (social, division of labor, production and value chain), transaction costs, trade organization, social and local regulation shapes, sense of belonging to a local community. All these aspect need a survey to be observed. Only through an ad hoc data collection, the systemic structure of an Industrial District can be traced.

Many studies have been conducted to understand the organizational arrangements of local production and the interrelationships between industries and local systems' production sectors in the Industrial Districts.

Different attempt have been done to map Industrial District in Italy. The normative or legislative perspective presented in the previous section is just one of the maps that we can reconstruct from official sources on Italian Industrial Districts.

Among the most significant empirical studies aiming to map these realities, some are very significant. We present the main surveys regarding the identification of Industrial Districts in the Italian context, paying particular attention to the definition adopted in order to identify the Industrial District.

One of the first quantitative analysis involving Industrial Districts, we find the research leading by Sforzi (1990). Istat now adopts the

detection procedure proposed by Sforzi. Industrial District are identified according to the data collected during the census of the population (1981), on the topic of commuting. Sforzi divides the Italian territory into local labor systems defined by the ratio between the place of residence and the place of work. The local labor systems are the places of everyday life, those places where most of the socio-economic relations are triggered.

Local labor systems evolve over times, according to the characteristics of the population of firms and people living there. Following this definition and with the availability of the data from the 1981 census, 955 local systems are indentified. Sforzi, using the typical tools of cluster analysis, classifies all the local labor systems and identifies only 61 Industrial Districts, as systems where there is a high productive specialization and almost all the firms involved are of small or medium size.

The first map produced by *Istat* dates back to 1995. The data refer to the population census 1991 and the procedure follow those proposed by Sforzi, involving commuting collected data. According to this map, in Italy there are 199 Industrial Districts.

Other attempts concern the studies conducted by Moussanet and Paolazzi (1992) for the Italian newspaper *IlSole-24Ore*. They produce a collection of articles that records the journey they carry out in Italy to discover of the Industrial Districts, accompanied by photographs, for 65 districts in total.

Unioncamere, together with the *Istituto Tagliacarne* and in collaboration with the *Censis*, produces its map of Italian Industrial Districts. They collect data from previous research (Istat and Il sole 24 ore) and by a direct survey carried out by the Chambers of Commerce in order to obtain a very large database. In fact, they identify 224 *firms concentration areas.*

Other researches on this topic have been conducted. Garofoli (1995) identifies 101 Italian Industrial Districts. Cnel-Ceris/Cnr (1997) makes a distinction between real district and legal districts (just defined by law) and produces a final map of 90 districts. Furthermore, the *Club dei Distretti (Osservatorio Nazionale dei Distretti Italiani)* identifies these realities according to the information gained by the associated districts.

All those maps are different in the number of Industrial Districts identified, and so on the criteria adopted by researchers. This is due to the huge amount of aspects the concept of Industrial District brings within itself. They are: productive specialization and local employment; social, local and economic organization (production chain or network of firms), governance system (hierarchic or horizontal).

Many studies have been realized in Italy on the topic of Industrial District, because they represent a very important aspect of the Italian economic system. Not only maps, but also research involving important aspects as financial performance and/or different kind of relations are carried out by several scholars in order to study these complex social and economic realities.

The following section points out some of these studies. The aim is to match theoretical and empirical studies on Italian Industrial District.

1.8 Matching literature and reality on Industrial Districts: empirical evidence in the Italian context

In this section we present some empirical studies carried-out on Italian Industrial Districts. Over the last decades, there has been a considerable increase in interest on this topic in different disciplines. In this section we consider just a very small part of these works, considering only the Italian territory as the application context.

A particular case-study is proposed by Boschma and Ter Wal (2007), about the Industrial District of Barletta, a footwear cluster in Southern Italy. In their study, they focuses the attention on the network of inter-firm knowledge exchange to verify how a firm's absorptive capacity affects its position in local and non-local networks, i.e. ties among firms within or beyond the cluster's boundaries. Exploiting the typical tools of Social Network Analysis – SNA – (Wasserman and Faust, 1994), they investigate the structure of the knowledge relationships of the district. They demonstrate that the position of a firm in these networks affects its innovative performance. A strong local network position of firms impacted positively on their innovative performance.

A wide number of scholars, in the field of industrial and regional studies (Capello, 1999; Lazerson and Lorenzoni, 1999; Breschi, and

Lissoni, 2001), consider that informal relations are the key channel to transfer both knowledge and information in industrial clusters/districts. Following this view, some empirical studies have focus on the importance of networking and informal relations in spreading knowledge in localized clusters.

Morrison (2008) carries out an empirical study on the *Murgia Sofa District*, located between Basilicata and Puglia. He explores several issues in order to analyze the role of the leading firms in identifying, absorbing and diffusing innovation-related knowledge. He focuses on informal contacts between expert technicians of the leader firms, in order to map the *knowledge socialization* network in terms of know-how flows (technical advice) and exchange of declarative knowledge (generic information). Using typical tools of SNA, Morrison compares the two different knowledge network taking into account the role played by leader firms to verify if they act as *gatekeeper*. The analysis shows different structural characteristics of the two examined networks. Interactions with external actors are developed and maintained by specific departments within the leaders and those relationships depend on the content of the exchange (i.e. information, knowledge). Clearly, information circulates better than the exchange of know-how that is limited to the participation of some actors.

Morrison points out that the community of informal links seems to be rather small and know-how sharing is also quite limited. He suggests that knowledge from leaders do not circulate equally among all local partners. Leading firms play a central role in training the Industrial Districts. They are the core of supply chains and of multiple-level networks of information and knowledge. Morrison emphasizes the dual aspect of leader firms. On one side they act as knowledge gatekeepers, on the other side, their central position in the knowledge network makes its development depends on the strategy of a few dominant players, for this reason potential internal conflict may arise.

Morrison and Rabellotti (2009) conduct a similar study using methods of Network Analysis. Their application concerns the Italian wine district of North *Piemonte* Region, also known as *Colline Novaresi*. They aim to reconstruct the internal informal networks developed by the wineries within the considered area and the external informal networks among distant actors (both geographic and socio-cognitive

distance). Leaving from sociological studies focusing on economic action as embedded in social structures (Granovetter, 1985; 2005, Powell, 1990; Burt, 1992; Uzzi, 1997), the authors have considered not only the dimension of connectivity in the district, but also many other aspects that help to define its structure. in particular, they consider the degree of closure (Colemann, 1988), the reciprocity, the strength of ties (Granovetter, 1973) and the core-periphery structure (Borgatti and Everett, 1999).

Within networks with high degree of closure the probability of knowledge exchange between interconnected agents increase (Colemann, 1988). At the same time, individuals placed in cohesive communities have higher probability of access to the same information sources (Uzzi, 1997) and, therefore, redundant information are shared. In this regard, the strength of ties has to be verified (Granovetter, 1973).

Empirical findings about the study *Colline Novaresi* district underline the importance of the conceptual distinction between information and knowledge in the analysis of networks in local production systems. The *Colline Novaresi* district, in fact, presents two clear different network structures: the information network is dense, non-mutual and distributed, and actors are linked by weak ties. Instead, the knowledge network is rather sparse, with a core of actors connected through strong and mutual ties. This means that while information is easily accessible by almost everyone. Knowledge flows, in terms of technical advices, are restricted to a strongly connected community of local entrepreneurs, as pointed out by the core-periphery structure. The core of this district is composed by small wineries, less innovative and opened toward external knowledge compared to the periphery, whose firms are larger, more innovative and opened toward external sources of knowledge. These latter are able to access knowledge inputs through either their outside linkages or by developing internal resources.

A further prevalent assumption in the literature on industrial agglomerations, concerns the creation of innovation process through the exchange of knowledge and the trust relationships developed within well-defined geographical clusters. Many scholars, through their

research, show that concentration fosters innovativeness of firms (Bell, 2005; Noteboom, 2006; Baoari and Lipparini, 2009).

In this regard, Lazaretti and Capone (2009) affirms that the *concentrations of firms encourage their innovation capacity because of a more widespread diffusion of knowledge, the presence of social capital and trust, and a more efficient ability to network.* Their empirical study concerns the Tuscan shipbuilding industry of pleasure and sporting boats. They analyze three project networks composed by economic, non-economic and institutional actors, financed in the framework of a regional project to support technological innovation and transfer. Their empirical result confirms the hypothesis that industrial cluster effects influence the innovation capability of a network. Through a performance analysis, i.e. the analysis of financial statements, they also verify that different behaviors in terms of innovation within the industrial clusters lead to different performances in their firms.

These studies are only some examples of the empirical researches carried out in the Italian context on this topic.

1.9 Concluding remarks

Considering what we have stated before, it is possible to affirm that there is an obvious difficulty related to the conceptualization, the definition and the individualization of Industrial Districts. Different aspects contribute to the definition of such complex socio-economic realities. All the available definitions, in fact, adopt a different point of view: from the legislative to the social, organizational or managerial model.

In the Italian context, the assumed definitions, in fact, lead to several maps of the districts, identified by different criteria

Therefore, Industrial Districts are, by nature, highly complex structures. They are also dynamic entities, because their geographic boundaries and the characteristics of the resident population change the years. In line with these transformations, even their governance systems suffer the consequences arising from those structural and environmental adjustments.

The difficulties related to the realization of a unique map of Italian districts that consider jointly all the aspects covered in literature, is still a very important subject of the scientific debate. This has driven our interest in complex statistical methods of analysis. The high complexity of the district structure and definition, as underlined in this chapter, lead us to identify and study the Industrial District following the typical methods of the Symbolic Data Analysis. The Symbolic Industrial District will be defined and analyzed in this framework.

The use of real data on Italian Industrial Districts will be the common threat of the whole dissertation. Those real data, as presented in the next chapters, highlights the applicability and the practical interpretation of the method described afterwards.

2 THE ITALIAN INDUSTRIAL DISTRICT AS A COMPLEX OBJECT. A SYMBOLIC DATA ANALYSIS APPROACH.

2.1 Introduction

The Italian Industrial Districts are characterized by several aspects. Those socio-economic entities have been defined according to different interpretation, even preserving their basic features of territorial localization and productive specialization. Actually it is still difficult to find a clear and unambiguous definition of the Industrial District.

The lack of a concrete definition carry us towards a new working definition of the Industrial District as a Complex Object, adopting the typical definition within the approach of Symbolic Data Analysis. Through this device we are able both to indentify both to analyze the district.

In the following paragraphs the Symbolic Data framework is introduced in section 2.2, while some main formal definitions are given in sections 2.3. The Symbolic Data Table is introduced in section 2.4. Starting from the generic definition of a Symbolic Data Object in the framework of Symbolic Data Analysis approach, given in section 2.5, the

working definition of the Symbolic Industrial District is proposed in sections 2.6. Section 2.7 presents the process applied to build up the Symbolic Data Table of Industrial Districts. Some concluding remarks are given in the section 2.8.

2.2 Foreword to define a Symbolic Object

In statistical data analysis, both in classical techniques and in multivariate methods, the identification of the basic unit of analysis plays a prominent role. A very important aspect of statistical units deals with their definition. In general, it is defined as the elementary observation of a population for which data on one or more phenomena under study are collected, examined and complied. Usually, the statistical units under investigation are single entities, such as a single individual, described by a set of variables (qualitative - categorical and/or quantitative - numerical) on which it takes only a single value.

The statistical units are the elementary building block for identify statistical aggregates. Sometimes it is useful for research purpose to aggregate these elementary units in main units. Making this, we need some rules to link the examined statistical units to the classification categories for which one would like to apply the synthesis process. This means that one could aggregate single individuals according to one or more characteristics. For example, individuals could be aggregated according to the economic activity of the firms for which they are employed. Otherwise the statistical units may be decomposed. For example a family can be considered as a basic unit if we are interested in studying some aspects concerning its membership (i.e. the number of members or the average monthly expenses), but it is also composed by several statistical units, its members, that we take into account if we are interested in their individual characteristics (i.e. the age or the profession of each single member).

A large part of statistical data methods involve statistical units associated with various aspects of a particular phenomenon. Even if we are interested on a single aspect of an object, observations are always multivariate in character. Therefore, in order to carry out a good

statistical survey it is always required to specify some fundamental aspects:

- to identify the phenomenon under study explaining what observable features (i.e. variables) are relevant to answer the research questions;
- to establish all the aspects of the phenomenon in which we are particularly concerned;
- to identify the target population and thus the basic statistical units on which the survey will be performed.

In recent years, with the advent of computers, very large datasets have become routine (Billard and Diday, 2003). If in simple datasets, the statistical units are in one-to-one correspondence with the data value, nowadays with the diffusion of complex datasets, multiple measurement are made for each unit and their analysis need high specialized methods and models. For instance, when we are interested in studying a concept rather than a single individual, the units of interest in a large database are not the individual data (the *microdata*) but some second level entities, than the intrinsic variability within each group of interest should be taken into account in order to provide aggregated generalized descriptions (Noirhomme-Fraiture and Brito, 2011).

Several techniques are available in literature to address the complexity of such kind of data. The main aim is to summarize large datasets into smaller and more manageable ones, without losing the knowledge included in the original dataset. Such as, the data in the resulting summary datasets assume different forms of realization. The data are no longer formatted as single values such as in the case of classical data, but are represented by lists, intervals, distributions and the like. These summarized data are known as *Symbolic Data*, while the *Symbolic Data Analysis – SDA –* is the reference framework of analysis (Diday, 1987, 1989, 1990, 1991, 1995). Thanks to a summarization process the information contained in huge datasets are reduced in a shorter set of new statistical units, known as *Symbolic Objects*.

Behind any summarization process there is the notion of a Symbolic Concept. Thus, any aggregation is necessary tied to the

concept with regard to the specific aim of the ensuing analysis (Billard and Diday, 2003). When the data describing the concept of interest have been obtained by contemporaneous or temporal aggregation of observations, or when we are dealing with a concept already specified by an expert or put in evidence by clustering method, the new elements appeared can no longer be described by usual qualitative or quantitative variables (Noirhomme-Fraiture and Brito, 2011).

A new kind of variables, known as *Symbolic Variable*, is introduced. Its main advantage consists in the fact that these variables take into account the internal observed variability in describing the entities of interest. Even the use of standard statistical techniques is often inappropriate to analyze the Symbolic Data Objects. In fact, in recent years, several methods of analysis, mainly explorative data analysis and data mining, have been extend from standard data to symbolic data (this subject will be discussed in chapter 3).

Several examples of symbolic data are available in different domain. The most popular in literature concern medical records, credit card purchases, species of birds, sports team, geographic boundaries and so on.

Let us consider as a simple example a dataset comprising medical records. For each patient in the dataset several information are recorded, such as geographical location variables (as region and city), demographic variables (as gender, age, marital status), basic medical and health variables (as weight, pulse rate, blood pressure) and, for a given prognosis, we have also information about treatments and other variables related to the disease. Those variables could be recoded in different ways according to the nature of the variable itself. Disease variables could be recoded as categories of a single variable with different modalities concerning the pathologies (as heart, cancer, cirrhosis). For example, cancer variables could be recoded as a list including all possible categories of cancer. This is an example of *multi-valued variable* Other variables, such as weight or age, could be recoded as being in an interval of values. These variables are *interval-valued variables*. Other variables, as propensity to diabetes, may be a histogram, an empirical distribution function, a probability distribution or a model. These variables are known as *modal variable*.

These variables can describe a group of individuals or concepts. This group become the new observations of a dataset.

This kind of data is much more complex than the standard one, since they are characterized by internal variation and may be structured. As a consequence, more complex data tables, called *Symbolic Data Table* are used to present Symbolic Data Objects (Diday, 2008).

2.3 Kinds of Symbolic Variables

In many domains the observations may be much more complex than the standard ones. Symbolic Data Objects, introduced by E. Diday and colleagues in a series of papers and books published in the second half of the twentieth century, are defined as a new representation of complex data. They are new statistical units that summarize groups of observations according to some common characteristics. Each generated symbolic data is defined by the Cartesian product of the modalities of the common variables.

Given a standard dataset, that is data organized in a rectangular matrix or an array where each cell contains the value of a variable *j* for an individual *i*, the first step in Symbolic Data Analysis – SDA – is to aggregate first-level units – individuals – in higher-level units – classes categories or concepts – aiming to describe them by taking care of their internal variation (Diday, 2008).

A second-level unit is called *class* when it is associated with a set of individuals. If it is deals with a value of a categorical variable, than it is called *category*. Instead, it is a *concept* if it has an *extent* and an *intent*. The extent of a concept is the set of first-level units that satisfy the concept, while its intent is the way used to find the extent of the concept. More accurately, the intent of a concept is modeled mathematically by a generalization process applied to a set of individuals considered to belong to the extent of the concept (Diday, 2008). This generalization process produce a Symbolic Object by taking care of the internal variation of the description of the individuals involved.

One of the most important things in summarizing a dataset is the *description* of the Symbolic Object under study. Most algorithms of SDA present the Symbolic Object, i.e. description of a class of individuals, as the result of a generalization process (output). A different process occurs when the starting point is this description. In such case it is interesting to find the individuals which are consistent with the description of the concept under study.

A formal definition could be found in Bock and Diday (2000). Let denote with Ω a dataset consisting of n individuals, with $\mathcal{D}$ the set of the d descriptions of individuals and the description of C classes of individuals, and with Y a mapping defined from Ω into $\mathcal{D}$ which associate each individual $\omega \in \Omega$ a description $d \in \mathcal{D}$ by using a vector of variables Y_j of a single quantitative variable into a set of values with associate weights. The description of an individual ω is called *individual description*, while the description of class of individuals C is called *intensional description* (Bock and Diday, 2000).

Moreover, as in Diday (2003), let denote with p be the number of variables available for each individual $i \in \Omega = \{1, \dots, n\}$ in a standard $n \times p$ matrix $\mathbf{X} = (xij)$, with p and n extremely large, where Y_j is the jth variable, with $j = 1, \dots, p$, and $Y_j = xij$ is the value assumed by the jth variable for the ith individual in the matrix. Let the domain of Y_j be $\mathcal{Y}_j$ so that $\mathbf{X} = (Y_1, \dots, Y_p)$ takes values in $\mathcal{X} = \times_{j=1}^{p} \mathcal{Y}_j$.

Therefore, if a classical data point is a single point in a p-dimensional space $\mathcal{X}$, than a symbolic data point can be represented as an hypercube in a p-dimensional space or a Cartesian product of distributions (Diday, 2003).

Different kinds of symbolic variables have been defined. Mainly, according to the type of the classical variables (numerical or categorical) we distinguish three major types of symbolic data:

- interval-valued variables;
- multi-valued variables;
- modal variables.

A symbolic dataset may also include special cases of the classical qualitative and quantitative called *single-valued* variables.

As a result, we consider a basic universe of individuals for which a classical single-valued variable is known and a system of classes of these individuals. It is possible to characterize the behavior of these classes with respect to the specified variable by defining a new aggregated variable able to specify for each class the range of values realized by the single-valued variable.

Notationally, we consider a basic set of N entities $E = \{1, \dots, N\}$. The entities u of E are called objects or data units and E is called *object set*. The object-set could represent:

- a universe of n individuals k, i.e. elementary objects, $E = \Omega = \{1, \dots, n\}$ with $N = n$;
- a subset or a sample from the universe of individuals $E \subseteq \Omega$ with $N \leq n$;
- a system of **C** classes of individuals $k \in \Omega$, i.e. aggregated objects, $E = \{C_1, \dots, C_m\}$ with $C_i \in \Omega$ and $N = m$ (*two-level paradigm*).

Both in the first and the second case the k individuals are also called *first order objects*, while in the third case the C classes are also referred to as *second order objects*

The properties of each entity $u \in E$ are described by p symbolic variables Y_j (with $j = 1, \dots, p$). A symbolic variable Y with domain $\mathcal{Y}$ is a mapping of the set E to a range $\mathcal{B}$ of its domain $\mathcal{Y}$ ($E \to \mathcal{B}$). Depending on the specification of $\mathcal{B}$ in terms of $\mathcal{Y}$, different symbolic variables can be defined.

> *A variable Y defined for all individuals/elements k of a set E is termed set-valued with the domain $\mathcal{Y}$ if it takes its values $Y(k)$ in the system $\mathcal{B} = \mathcal{P}(\mathcal{Y}) = \{U | U \subseteq \mathcal{Y}\}$ of all nonempty subsets U of the set $\mathcal{Y}$: $Y(k) \subseteq \mathcal{Y}$ for all $k \in E$, possibly with some constraints on the size or structure of U* (Bock and Diday, 2000).

In the case of single-valued variables, denoted with $\tilde{Y}$, $\mathcal{B} = \mathcal{Y}$ and so $|Y(k)| = 1$ for all elements $k \in E$. Multi-valued variables and interval variables are common types of set-valued variables.

A set-valued variable Y is defined a *multi-valued variable* if its values $Y(k)$ are all finite subsets of the underlying domain $\mathcal{Y}$. Formally, this means that $\mathcal{Y}: |Y(k)| < \infty$ for all elements $k \in E$. As in the classical case, multi-valued variable could be either categorical or quantitative. We have a *categorical multi-valued variabl*e if the variable Y present a finite range of categories in the domain $\mathcal{Y}$ such that all values $Y(k)$ are finite as well. Instead, we have a *quantitative multi-valued variable* if the values $Y(k)$ are finite set of real numbers $Y(k) \subset \mathbb{R}$.

A set-valued variable Y is called an *interval variable* if for all elements $k \in E$ the subset U on the underlying domain $\mathcal{Y}$ is either an interval of $\mathbb{R}$ endowed with the natural ordering $\leq$ or another totally ordered domain with an order $\preccurlyeq$ on $\mathcal{Y}$ (in the case of ordinal variables).

Another important type of symbolic variable is termed *modal variable*.

> *A modal variable Y with domain $\mathcal{Y}$ defined on a set $E = \{a, b, \dots\}$ of higher-order objects is a mapping $Y: E \to \mathcal{B} = \mathcal{M}(\mathcal{Y})$ from E into the family $\mathcal{M}(\mathcal{Y})$ of all non-negative measures π on $\mathcal{Y}$, with values $Y(a) = \pi_a$* (Bock and Diday, 2000).

A modal variable assigns to each object $a \in E$ both the category set $Y(a) \subseteq \mathcal{Y}$ *of* the domain $\mathcal{Y}$, both for each category $y \in Y(a)$, a frequency distribution, a probability distribution or a weighting $w(y)$.

If we consider a categorical variable with a discrete frequency distribution displayed as a bar diagram, we are dealing with a *diagram variable* in the symbolic framework. Instead, if we consider a quantitative variable, the property of the aggregated class can be described by a symbolic variable whose values could be specified by a normal distribution or an empirical distribution function or a histogram describing the empirical distribution of the new variable in each class. The latter is usually referred to as a *histogram variable* in SDA.

Moreover, variables can also be defined:

- *taxonomic* when the underlying domain $\mathcal{Y}_j$ of a variable Y_j is ordered into a hierarchical or tree structure;

- *hierarchically dependent* or *mother-daughter variables* when the outcome of a variable depends on the actual value realized for another variable (i.e. *Z* is a *non-applicable variable* for certain categories *y* of another variable *Y*);
- *logical dependent* if a variable depends logically or functionally on the values assumed by another previous variable (i.e. a *rule* is defined on *Y* such that *Z* may assume specified values).

2.4 The Symbolic Data Table

In contrast with the classical data table, for which a single specific value is assigned for each *xij* in a matrix **X**, the entries of a symbolic data table are not restricted to a single value. This means that, as introduced in the previous section, each cell of a symbolic data table can contain more than a single value, but rather it measures the symbolic variable with a $(x \pm \delta)$ value.

A symbolic data table is organized as a classical data matrix, but each cell contains *symbolic data*, each raw corresponds to the *symbolic description* of a group or a concept of interest and each column corresponds to a *symbolic variable* (Noirhomme-Fraiture and Brito, 2011).

Therefore, following the terminology introduced in the previous section, each object included in the object set ($u \in E$) is recorded in a symbolic data matrix or a symbolic data array $\underline{\boldsymbol{X}}$ of dimension $N \times p$. The generic element of this data table is the value of the *j-th* symbolic variable Y_j observed for the entity u. It is denoted with ξ_{uj}. Each object u in the matrix is described by a *symbolic data vector X(u)* $= (y_1(u), \dots, y_p(u))'$*:*

$$x_u = X(u) = \begin{pmatrix} \xi_{u1} \\ \vdots \\ \xi_{uj} \\ \vdots \\ \xi_{up} \end{pmatrix} = (\xi_{u1} \quad \dots \quad \xi_{uj} \quad \dots \quad \xi_{up})' \qquad [1]$$

Therefore, the [1] is introduced as a raw in the symbolic data table $\underline{\boldsymbol{X}}$ that will be arranged as follows:

$$\underline{\boldsymbol{X}} = \begin{pmatrix} x_1' \\ \vdots \\ x_u' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} \xi_{11} & \xi_{12} & \cdots & \xi_{1j} & \cdots & \xi_{1p} \\ \xi_{21} & \xi_{22} & \cdots & \xi_{2j} & \cdots & \xi_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \xi_{u1} & \xi_{u2} & \cdots & \xi_{uj} & \cdots & \xi_{up} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \xi_{N1} & \xi_{N2} & \cdots & \xi_{Nj} & \cdots & \xi_{Np} \end{pmatrix} = \left(\xi_{uj}\right)_{N\times p} \qquad [2]$$

where each ξ_{uj} may be either an interval, or an histogram or a probability distribution according to the type of variable it refers to.

In general, we affirm that, unlike classical data, symbolic data can contain internal variation and can be structured. The presence of those specific characteristics request new techniques of analysis, which differ from the standard ones. However, classical data are represent as special case of symbolic data, e.g. a classical point $x = a$ is the equivalent of the symbolic interval $\xi = [a, a]$ (Billard and Diday, 2004).

In other words, suppose we have a classical data table where individuals are described by some variables. Through a generalization process, individuals could be aggregated according to the categories of one or more of those variables, thus defining a Symbolic Object. The resulting data table is a symbolic data matrix that will be analyzed using symbolic data analysis methods.

As in Diday 2008 and represented in Table 2.1, let consider a classical 6×3 data matrix $\mathbf{X} = (x_{ij})$ for $n = 6$ individuals and $p = 3$ single-valued categorical variables (Concepts with two categories C_1 and C_2; Y_1 with two categories a, b and c; and Y_2 with categories coded in 1, 2 and 3). It is possible to reduce the dimension of the classical matrix into a symbolic 2×2 array in Table 2.2, by establishing two simple rules:

$$\begin{aligned} &[Y_1 = a] \Rightarrow [Y_2 = 2] \\ &[Y_2 = 1] \Rightarrow [Y_1 = b] \end{aligned} \qquad [3]$$

Table 2.1 Example of a classical n × p matrix

Individuals	Concepts	Y_1	Y_2
I_1	C_1	a	2
I_2	C_1	b	1
I_3	C_1	c	2
I_4	C_2	b	1
I_5	C_2	b	3
I_6	C_2	a	2

Table 2.2 A symbolic data table introduced from table 2.1

	$\underline{Y_1}$	$\underline{Y_2}$
C_1	$\{a, b, c\}$	{1,2}
C_2	$\{a, b\}$	{1,2,3}

The introduction of the rules in [3] allows the definition of the concepts C_1 and C_2 in a symbolic data table by two categorical multi-valued variables ($\underline{Y_1}$ and $\underline{Y_2}$). These concepts represent the Symbolic Objects that emerge after the summarizing process.

The process of building Symbolic Objects may have as starting point dynamic queries to a relational database thought which it is possible to extract groups of individuals with common characteristics. Furthermore, it may starts ones defined the common variables and modalities of different surveys.

This lead us to a more rigorous definition of Symbolic Objects developed by E. Diday and his colleagues (Diday, 1987, 1989, 1990; Bock and Diday, 2000; Stephan et al., 2000), as presented in the next section.

2.5 The Symbolic Object: definition and properties

Symbolic Data Objects are a new way of representing complex data. As introduced before, they are the representation of a concept by its

intent and provide a way for finding its extent. The intent of a concept is mathematically modeled by a generalization process applied to a set of individuals belonging to its extent (Diday, 1995).

Symbolic Objects are able to summarize the original symbolic data table in an explorative way by expressing concepts in terms of descriptions based on properties concerning the initial variables or meaningful variables (Diday, 2008). Therefore they are independent from the original database and so the matching with any new individual is easy to be identified.

The formal definition of the Symbolic Data Object is the following:

> *a Symbolic Object is a triple* $s = (a, R, d)$ *where "*d*" is a description, "*R*" is a relation between descriptions, and "*a*" is a mapping defined from* Ω *in* $\mathcal{L}$ *depending on R and d (Bock and Diday, 2000).*

For instance, a Symbolic Object s is defined by:

- description $d \in D$ from a given symbolic table, where D represents the set (vector) of descriptions of individuals or of groups of individuals;
- a matching relation R for comparing d to the description of an individual. It could be a single comparison operator $R \in \{=, \neq, \equiv, \leq, \geq, \subseteq, \supseteq, \in, \notin, \dots\}$, an implication, a kind of matching, or a logical combination of such operators;
- a mapping or a *membership function* $a: \Omega \to \mathcal{L}$ which maps individuals $\omega \in \Omega$ onto the space $\mathcal{L}$, such that $a(\omega) = [y(\omega)\ R\ d]$.

Consider a general situation with an universe Ω of individuals ω, a space of descriptions D containing all the descriptions d, containing all description vectors d in the case of p classical variables or all description systems D in the case of p symbolic variables of individuals and classes of individuals and p variables Y_j which could be typically multi-valued, interval type or modal variables. If the initial database contain p variables, we denote:

$$y(\omega) = \left(y_1(\omega), \dots, y_p(\omega)\right)$$
$$D = (D_1, \dots, D_p)$$
$$d \in D: d = (d_1, \dots, d_p) \quad [4]$$
$$R = (R_1, \dots, R_p)$$

The descriptions d, taken by a mapping $y(\omega)$ which associate to each $\omega \in \Omega$ a description $d \in D$ by using a vector of variables Y_j given on a set-object E, are usually recorded in a symbolic data table. A set of coherent descriptions constitutes the set of objects on which any SDA algorithm applies (Diday, 1998).

Thus, a special case of Symbolic Object is defined as an *assertion.* An assertion is a conjunction of corresponding events (Bock and Diday, 2000). Assertions, also termed *queries*, are of utmost importance when aggregating individuals into classes from the initial database. It is defined by $s = (a, R, d)$ and it is written as the corresponding product of relations:

$$a = \bigwedge_{j=1}^{p} \left[Y_j R_j D_j\right] \quad [5]$$

where each condition or statement $\left[Y_j R_j D_j\right]$ or $\left[Y_j R_j d_j\right]$ is an event.

Each raw of a symbolic data matrix provides an assertion called *individual Symbolic Object* when in the [5] $R_j = \ll = \gg$:

$$a(w) = \bigwedge_{j=1}^{p} \left[Y_j = d_j\right] \quad [6]$$

The extent of an individual Symbolic Object is the set of individual descriptions defined in the symbolic data table which realize the same given data vector.

For instance, consider a binary data matrix describing the economic structure of $N = 10$ regions $(u \in E = \{A, B, \dots, J\})$ with

respect to $p = 6$ binary variables ($Y = (Y_1, \dots, Y_6)$) showing the presence or absence of different industrial sectors, recorded as 1 or 0 respectively. Consequently, each region is described by the corresponding raw in the matrix, i.e. a binary data vector.

Suppose we are interested in a subset of the cities which have specific industrial sectors (e.g. $Y_1 = 1, Y_2 = 1, Y_6 = 0$). We can formulate a query or an assertion of the type:

$$a = [Y_1 = 1] \wedge [Y_2 = 1] \wedge [Y_6 = 0] \qquad [7]$$

With the statement in the [7] we can compare the variables of interest (Y_1, Y_2, Y_6) with the required profile $z = (z_1, z_2, z_6) = (1,1,0)$.

The same query can be written in an equivalent form considering a subset D of all binary vectors of Y which realize the statement in the [7]:

$$\begin{aligned} D &= D_1 \times D_2 \times D_3 \times D_4 \times D_5 \times D_6 = \\ &= \{1\} \times \{1\} \times Y_3 \times Y_4 \times Y_5 \times \{0\} \end{aligned} \qquad [8]$$

Then the assertion takes the form: $a = [Y \in D]$.

The query induces on the object set E a *truth function* $a(.)$ which takes values 1 and 0, i.e. $true$ and $false$, if the data vector $Y(u)$ fulfill or not a. Formally, we have:

$$\begin{aligned} a(u) &= [Y \in D] = \\ &= [Y_1 = 1] \wedge [Y_2 = 1] \wedge [Y_6 = 0] = \begin{cases} 1 \\ 0 \end{cases} \end{aligned} \qquad [9]$$

where "∧" denotes the conjunction *"and"* for truth functions.

Returning to the example above introduced, the requirement "*all regions which fulfill a*" is a conceptual entity and, thus it is a Symbolic Data Object, if formalized as in the [9]. The extension of this assertion object, denoted with $Ext(a)$, contains all the elements in the subset which realize the requirement. In this example, supposing that only three regions fulfill a, formally we can write:

$$Ext(a) = \{u \in E | Y(u) \in D\} = \\ = \{u \in E | a(u) \in 1\} = \{A, C, G\} \qquad [10]$$

In general, given an object set E and a set of r events v, the extension of an assertion in the [10] is written as follows:

$$Ext(a) = \{u \in E | [Y_j(u) R_j d_j] = 1 \; for \; all \; v = 1, \dots, r\} \qquad [11]$$

As in the given example, sometimes, when we are looking for some specific properties of the elements into a database. These do not concern all the recorded variables $Y = (Y_1, \dots, Y_p)'$, but only with a subset t of them, usually referred to as the *set of indexes* and denoted by $J = \{j_1, \dots, j_t\} \subset \{1, \dots p\}$. In such cases, we have to introduce a *filtering operator* h in order to select the desired variables Y_j with $j \in J$ from Y and the corresponding descriptions d_j with $j \in J$ from d. This filtering operator is formally specified by:

$$h_J(y) = h(y) = h(y_1, \dots, y_p) = \\ = (y_j | j \in J) = (y_{j1}, \dots, y_{jt}) \qquad [12]$$

Then, the extension function of a Symbolic Object in [9] and [10] is formalized as follows:

$$a(u) = [Y_{j1}(u) R_{j1} d_{j1}] \wedge \dots \wedge [Y_{jt}(u) R_{jt} d_{jt}] = [Y(u) R_J d] \qquad [13]$$

where the relation $h_J(R) = R_J$ refers only to the selected element and is defined by $[Y(u) R_J d] = \wedge_{j \in J} [Y_j R_j d_j]$.

What have been so far described is just one of the simplest elements in the SDA framework.

Basically, two main kinds of Symbolic Objects can be defined: the *Boolean Symbolic Object* – BSO – and the *modal (or probabilistic) Symbolic Object* – MSO.

> *When $\mathcal{L} = \{0,1\}$, i.e. when a is a binary mapping, then s is a Boolean Symbolic Object. When $\mathcal{L} = [0,1]$, then s is a Modal Symbolic Object* (Billard and Diday, 2004).

This means that, if we have a binary relation between the descriptor of the object and the definition domain that is defined to have values true or false, i.e. $[Y(\omega)RD] \in \mathcal{L} = \{0,1\}$, then we are dealing with a *BSO*. In such case, the cells of a symbolic data table contains both qualitative and quantitative single-valued and multi-valued, or intervals. For example, suppose we have a datasets of firms and we are interested in describing only those firms which are located in particular regions (i.e. Region A and E).

We can define a BSO as $s = (region \in \{A, E\})$. The extent of a BSO is defined by $Ext\ (a) = \{\omega \in \Omega\ /\ a\ (\omega) = true\}$.

Moreover, many practical cases do not fit exactly in the Boolean framework, but we have a degree of belonging of an elementary object with a description that must be measured on a continuous scale, i.e. $[Y(\omega)RD] \in \mathcal{L} = [0,1]$. In these cases the MSO occurs and the symbolic table contains, mainly, multi-valued data types with weights. For instance, returning to the example of firms, suppose we are interested in describing the firms located with a different weight (0.25 and 0.75) in the regions A and E. We can define a MSO as $s = (region \in \{[0.25]A, [0.75]E\})$. Those weights have different meanings: *probabilities, capacities and possibilities*, according to the underlying application field[6].

2.6 The working definition of Industrial District as a Symbolic Data Object

The Industrial District is a very complex entity, as underlined in the previous chapter. According to several definitions[7], it has been

[6] For further details on Modal Symbolic Objects, refer to Bock and Diday (2000).

[7] A detailed overview about Industrial District has been addressed in Chapter 1. For further insights on this topic, refer to the texts of A. Marshall and G. Becattini and other famous scholars of this important issue.

addressed as a territorial area in which common characteristics of firms and people living there contribute to define it as a community.

Considering the syntax introduced in this chapter about the definition of a Symbolic Object in the framework of SDA and considering the common elements emerged in the several available definition of the Industrial District, it is possible to give a new working definition of this entity, as a Symbolic Industrial District.

In fact, an Industrial District is made up of specialized firms on a particular industrial sector, that are located within the boundaries of the same geographical area, which share tangible and intangible aspects of the population living there, of the local infrastructure and of the governance systems established for its management.

Thus, the starting point for a new working definition of Industrial District in the SDA framework is represented by the firms that compose it. Those firms can be found according to their economic activity identified through an activity code, i.e. the *Ateco 2007 Code*[8].

Consider the universe Ω of all the Italian firms (ω) for which several aspect can be recorded: the activity codes (Ateco 2007), the localizations (geographical area, region, province, city), the data extracted from their annual financial statements or financial report (profitability indicator ratios, financial ratios), and so on. On the other hand we have Italian Industrial Districts (ID) that are defined according to their productive specialization (Ateco 2007), the Italian province in which they are located, the normative law that has been approved to detect and regularize its existence in the regional boundaries (see section 1.6 in Chapter 1 for details), some other aspects linked to its governance system (the presence of a district Committee, of an authority reference, of facilitators or expert in the implementation of project, of governance instrument as the district development pacts). Consequently, by taking into account the theoretical definitions given in

[8] As part of the administrative simplification, in Italy since January 2008, the new ATECO 2007 classification of economic activities has been adopted. This is a unique table of classification containing the codes for the different economic activities of firms, common to all the organizations that, for several reasons, deal with the classification of firms, as Istat, the Revenue Agency, the Chambers of Commerce and others.This classification is the national version of the European nomenclature, Nace Rev. 2, published in the Official Journal of 20 December 2006 (Regulation (EC) no 1893/2006 of the European Parliament and of the Council of 20 December 2006).

Chapter 1 whose main aspect are reported above, each firm can be assigned to a specific Industrial District.

Those data are organized in a classical $n \times p$ matrix $\mathbf{X} = (xij)$, where the raw entries are the firms defined by different type of variables among which we find the Italian ID of belonging.

The complexity of this data structure involves a methodological choice: the search for the most suitable method for the reduction of its dimensionality. In this context, the concept of Industrial District is well suited to be regarded as a second order object in the SDA framework, that is as a symbolic description defined by a set of symbolic variables. This means that it is possible to group the firms in the initial database according to the different variables that characterize them, and as consequence, also according to the Italian Industrial District.

Known all the above mentioned aspects, the Symbolic Industrial District can be defied. Denoting with Y_{Prov} the list of all Italian provinces and with $Y_{Ateco07}$ the list of all the Ateco 2007 codes that characterize a specific industrial sector, then according to the query in the [7], a generic Symbolic Industrial District ID_i can be now formally defined according to the following statement:

$$ID_i = [Y_{Prov} = Province_i] \wedge [Y_{Ateco} = Ateco07_j] \qquad [14]$$

The corresponding *truth function*, as expressed in the [9], will now take the form:

$$ID_i(u) = [Y_{Prov} = Province_i] \wedge [Y_{Ateco} = Ateco07_j] = \begin{cases} 1 \\ 0 \end{cases} \qquad [15]$$

Its extension will assume as result the subset of all the firms in the initial dataset that fulfill the requirement in the [14]. As in the [10], this is formally written as:

$$Ext(ID_i) = [u \in E | ID_i(u) \in 1] = \{firm_1, \dots, firm_n\} \qquad [16]$$

In this way, the firms that are part of an Italian Industrial District are grouped according to the main characteristics of this entity. As far as the variables in the initial database, depending on the type of

variable we are dealing with, they will assume the typical forms of symbolic variables, as introduced in the section 2.3 of this chapter.

The following section describes the procedure adopted in order to obtain the symbolic data table on which the methods of SDA will be performed.

2.7 The data mining to build up a Symbolic Data Table of Italian Industrial Districts

The Symbolic Industrial District definition we propose, is useful to realize a database containing homogeneous information for each entity, here identified as a complex unit of analysis.

The data collection is carried out by the extraction of secondary data from the database *Analisi Informatizzata delle aziende italiane – Aida*[9], which provide detailed information about the Italian firms whose turnover is at least 100,000 Euro, as income statement, balance sheet, financial ratios, trade description, ownership and management, and so on. The research program Aida has 170 filters, such as the company name, the Ateco 2007 codes, the geographic area and the financial ratios. All these filters can be use separately, one at the same time, or in a sequence by specifying the adequate Boolean Operators (And, Or, Not). The selection of the most appropriate filters is of utmost importance, since a well-expressed query is the starting point for the data mining. Once defined the query, the program extract all the firms whose characteristics fulfill the requirement. The companies mined are displayed as a matrix or an array including all the available variables (anagraphic data and financial ratios).

As far as the Italian Industrial Districts this procedure of data extraction is carried out through the formulation of the query defined in the [14]. In this way it is possible to obtain for each Industrial District detailed information about all the firms that are located into its

[9] *Aida* – Analisi Informatizzata delle aziende - is a repository containing comprehensive information on Italian Companies, according their annual financial statements with up to 5 years of history. It covers one million companies in Italy providing detailed accounts following the scheme of the 4th Directive CEE. For more detail see the official web site at http://www.bvdinfo.com/it-it/home.

geographical boundaries. Their main production activity is linked to the specified Ateco 2007 codes. Furthermore, it is possible to select the time series up to 10 years earlier than the current one.

As a result, for each Italian ID and for each selected year we have a single-valued variable recorded in the standard data table. This procedure will be repeated as many time as the number of Italian ID for which it is possible to define the query expresses in the previous section in order to create a single database[10] .

Once the extraction procedure has ended we have a huge database in which by raw we have the firms and, among all the recorded variables, now we find a new variable, i.e. the Italian ID associated with each firm. The high complexity of this data structure leads to a methodological choice related to the reduction of its dimension, in order to realize a new and more manageable database on which several methods of analysis could be performed.

In order to deal with this important issue, the SDA framework offers the right procedure of reduction, thus we are transforming a standard data table into a symbolic data table, as defined in the section 2.4.

The first step of this procedure consist in codifying the recoded variables observed for each firm into symbolic variables whose unit of reference is the concept of Italian ID, as emerged in the query formulated for the Aida database. As the first moment of a general strategy of synthesis and representation of symbolic data, the available financial variables can be treated as *set-valued variables*, and specifically as *interval-valued variables.*

Consider the example tables 2.1 and 2.2, suppose that the individuals are the firms, the concepts are the Italian Industrial Districts and the variable are the financial ratios (as, for instance the *Return on Equity* – ROE – and the *Return on Investments* – ROI). It is possible to define the symbolic data table in which by raw we have the Symbolic IDs (second-order units) defined by the subset of firms (first-order units) which are part of each Italian ID, and by column the interval-valued variables which describe each Symbolic ID by the two

[10] See Chapter 4 for the selection of the Italian Industrial Districts.

extreme values (low and upper) of the range of values recorded for each considered ratio in the initial database, denoted as $[l_{ij}, u_{ij}]$.

Thus, we obtain the following data table:

Table 2.3 Interval data table of the Italian Industrial Districts

	ROE	*ROI*	…	Y_j	…	Y_p
ID_1	$[l_{1ROE}, u_{1ROE}]$	$[l_{1ROI}, u_{1ROI}]$	…	$[l_{1j}, u_{1j}]$	…	$[l_{1p}, u_{1p}]$
ID_2	$[l_{2ROE}, u_{2ROE}]$	$[l_{2ROI}, u_{2ROI}]$	…	$[l_{2j}, u_{2j}]$	…	$[l_{2p}, u_{2p}]$
⋮	⋮	⋮	…	⋮	…	⋮
ID_i	$[l_{iROE}, u_{iROE}]$	$[l_{iROI}, u_{iROI}]$	…	$[l_{ij}, u_{ij}]$	…	$[l_{ip}, u_{ip}]$
⋮	⋮	⋮	…	⋮	…	⋮
IID_n	$[l_{nROE}, u_{nROE}]$	$[l_{nROI}, u_{nROI}]$	…	$[l_{nj}, u_{nj}]$	…	$[l_{np}, u_{np}]$

where the i-th Symbolic ID is represented by a *p-tuple* of intervals, as:

$$I_i = (I_{i1}, \dots I_{ip}) \qquad \text{with } I_{ij} = [l_{ij}, u_{ij}] \qquad [17]$$

where $i = 1, \dots, n$ and $j = 1, \dots, p$.

Each Symbolic Industrial District is here described by interval-valued performance ratios. Chapter 4 will presents the empirical analysis performed on such new structure of the data. Moreover, the main aspects of the synthesis process that allows us moving from first to second level analysis will be discussed, according to their application on real data. observed for a subset of Italian IDs.

2.8 Concluding Remarks

Symbolic data are the result of several data reductions carried out in order to answer a particular research question. It does not involve individual units, but rather it focuses on collections, classes or groups. The Symbolic Data framework can be seen as an extension of classical data to particular kind of structured data, where each observation is described by multi-valued variables. This is not just a theoretical conjecture able to define a new type of data.

As underlined in this chapter, the definition of symbolic data and their tabular representation allows to consider all the original data, moving from the classical data framework to the symbolic one. In this context, a theoretical construct can be operationalized in order to perform advanced statistical analysis.

The added value of this work is to consider a well-known theoretical definition of a topic, in this specific case the Italian Industrial District, and give it a new definition that allows to a new quantitative treatment. This means that, we are no longer dealing with Statistics of atomic data, but rather with Statistics of knowledge. The Italian Industrial District is here defined as a concept, with its intent and extent, as a typical complex object in the symbolic data framework.

In order to find symbolic versions of common statistical techniques, several methods to handle symbolic data have been developed over the past decades. The following chapter will give an overview of the main exploratory multidimensional methods for the analysis of symbolic data. In particular, the analysis of interval-type variables, in which the endpoints (lower and upper values) as well as the centers are taken into account. At the same time some references will also be addressed to methodologies for the analysis of histogram-valued data.

3 *STATISTICAL METHODS FOR SYMBOLIC DATA ANALYSIS*

3.1 Introduction

Symbolic Data Analysis – SDA – has known in recent years a great development in terms of applications and methodological innovations. From one side, the growing interest in this framework is linked to the increase of huge datasets available in different fields. On the other side it is related to researchers' need to adopt new procedures of analysis able to manage complex data structures.

In fact, with the arrival of huge repository, streaming data. web-based data, the management of very large dataset have become routine. Thus, the increasing interest of researchers have moved towards the definition of the most suitable methodologies in order to extract useful and meaningful information from this huge amount of data. Unlike classical data, for which a considerable annals of statistical methodologies have been produced over the last century, symbolic data and their statistical analysis are quite new. Until now, only few methodologies are available, even considering descriptive statistics and

advanced methods as the Regression Model, the Principal Component Analysis and the Clustering methods.

The aim of this chapter is to give an overview of the main symbolic data methodologies available in the literature, starting from the origins of SDA up to the current statistical techniques introduced in this field.

Section 3.2 describes the historical and philosophical background of SDA, while section 3.3 underlines the importance of using symbolic data, as interval-valued data, data distributions, list and similar instead of the classical ones usually formatted as single-valued data. Sections 3.4 and 3.5 provide a general overview on the SDA structure and the ways it allows to model concepts through the use of Symbolic Data Objects. Section 3.6 presents a short review of the main explorative SDA methodologies available in literature: Principal Component Analysis and Clustering Methods. Concluding remarks are given in section 3.7.

3.2 The history of Symbolic Data Analysis

The historical idea of SDA can be found in *the Aristotle Organon* (IV B.C.) where he clear distinguishes between *first order individuals*, i.e. a unit associated to a single individual in the world, and *second order individuals*, i.e. a unit associated with a class of individuals (Diday, 2002). One of the main aim of the SDA is to analyze second order individuals toward the extension of standard data methods to complex units. As introduced in the previous chapter as regards the working definition of the Italian Industrial District – ID, we will consider firms as first order individuals and IDs as second order individuals. The ID is described by the summary of the values taken by its firms, formatted as intervals, histograms, probability distributions and so on depending on the random variables in the original data matrix. Thus the ID can be defined as a *Symbolic Object*.

This aspect is directly linked to another fundamental goal of SDA: the extraction of knowledge from a huge dataset through the definition of a Symbolic Data Object. In general terms a concept is defined by an *intent* and an *extent*. The intent consists of a set of properties. The extent is defined by the set of individuals that satisfy the intent. Therefore, obviously, it is quite impossible to detect all the properties

of a concept and its complete extent in the real world. As a consequence, the Symbolic Data Objects are just approximations of the reality. Like concepts in our mind, a Symbolic Object models a concept by using a description (i.e. the intent) and a mapping function through which it is possible to found its correspondence in the real world (i.e. the extent). [11]

The Aristotelian tradition is just one of the main four tendencies from which the Symbolic Object gets its advantages. Indeed, while it deals with the Aristotelian definition of concepts as far as the explanatory power of a logical conjunction of several specific properties, it also considers the Adansonian tradition as regards its extent. Following this tradition the concepts represented by a Symbolic Data Object are polytheistic (Diday, 2002) because all its members are similar since they must satisfy at best some properties.

Other important influences can be found in Rosch (1978) and Wille (1982) points of views. Following the fist one, it is derived the definition of the membership function of a symbolic data. In fact, according to Rosch, concepts are represented by classes defined in terms of prototypes, also in SDA a mapping function provides prototypical instances whose attributes are the most representative ones. Instead, the so called Complete Symbolic Object originates its properties from the philosophical tradition of Wille. In his opinion, the intent of concepts must be able to describe all properties applicable for the members that constitute its extent. In the SDA framework, it means that a Complete Symbolic Object constitutes a Galois lattice on symbolic data, as proved, among others, by Diday (1991,1998) and Brito (1994).

Moreover, the development of SDA methodologies has born under the influence of several important fields. Specifically, for different aspects, the major influences come from the Explorative Data Analysis whose leading pillars can be found, among others, in Tuckey (1958), Benzécri (1973), Saporta (1990); Artificial Intelligence and Learning Machine as regards the language and the mathematical models implemented for the representation of complex knowledge; Numerical taxonomy as in biology where the species can be considered as

[11] For a detailed overview on the formal definition of a Symbolic Object, in general, and of the Industrial District as a Symbolic Object, in particular, refer to Chapter 2.

concepts that can be modeled through the definition of a Symbolic Object and, above all, Classification Data Analysis, either in the *bottom up* (see Jussieu, 1974) either in the *top down* approach (see the Adanson's algorithm on Sequential Agglomerative Hierarchical Clustering, 1759), and the Ward's method about the minimum variance criterion (1963), the classes obtained, respectively monotheistic and polytheistic classes, and their descriptions can be interpreted as modeled by a Symbolic Object[12].

3.3 Advantages of using a Symbolic Data Analysis approach

In recent times and in several fields, statistical units turned to more complex shapes compared with the standard ones. The need to analyze those new type of data has led to the development of SDA. This new data are no longer formatted so that each individual takes exactly one value for each variable, and new forms of variables have been formally introduced (Bock and Diday, 2000)[13]. These new data are characterized by two levels of units and their tabular representation is the so called symbolic data table, where first order units (i.e. individuals, firms, patients) are aggregated in second order units (concepts), while variables are of multi-valued type. The main aim is to model the intent of a concepts by describing the classes of individuals that compose its extent defined by a Symbolic Data Object, in the consideration that only a rough description of the concept can be reached.

In comparison with classical approaches, the use of SDA methods has several advantages. First of all they allow to manage complex data structure, since data are synthesized in Symbolic Objects through a generalization process. In addition, the visualization of those data is more immediate and practical, considering both the tabular (the symbolic data table) and the graphical representations (data are no

[12] For more details on the philosophical background of SDA, please refer to the pioneer works of Diday, for a complete overview on SDA definitions and methods see Bock and Diday, 2000

[13] The formal definition of symbolic variables has been addressed in chapter 2 of this dissertation.

longer a point in a geometrical space, but the displayed objects assume a particular shape). Given as input a symbolic data table, the Symbolic Data Object obtained as output gives an explanation of the results in a user-friendly language, while its graphical description takes into account its internal variation (Bock and Diday, 2000).

Moreover, as said before, symbolic data be easily converted in terms of assertion (i.e. query), and so they can be used to spread a concept among different databases. By definition Symbolic Objects are independent from the initial data table. Thanks to this important property they are able to identify any matching individual described in any data table. Symbolic Data Objects are also able to define the same concept by combining several properties of different variables, even if these latter originate from different arrays or refer to different underlying populations.

Another important advantage of using SDA consists in the fact that it is possible to extract symbolic data from different standard datasets, instead of merging them, and then apply a SDA to the whole set of the Symbolic Objects (Bock and Diday, 2000). These latter aspects are particularly interesting if we are interested in exploratory data analysis, and therefore they are of fundamental importance for this work. In fact, the symbolic data table of Italian Industrial Districts used for this dissertation has been built taking into account these features of the Symbolic Objects (for more details see Chapter 2 and Chapter 4).

As of some important statistical properties, it is possible to measure the quality, the robustness and the reliability of Symbolic Data Objects (see Diday, 2008).

The strong interest shown towards the Analysis of Symbolic Data is equally evident in the increased production of new tools, whose development consists both in the introduction of new algorithms in popular and well-known statistical software, both in the production of ad-hoc software.

In the first case, we are dealing, with new packages implemented in the **R** software, such as:

- RSDA: R to Symbolic Data Analysis (Oldemar Rodriguez, 2014);

- symbolicDA: Analysis of Symbolic Data (Dudek, Pelka and Wilk, 2015);
- HistDAWass: Histogram Valued Data Analysis (Irpino, 2015);
- GraphPCA: Graphical tools of histogram PCA (Brahim and Makosso-Kallyth, 2014);
- MAINT.Data: Model and Analyse Interval Data (Duarte Silva and Brito, 2015);
- ISDA.R: Interval Symbolic Data Analysis for R (Queiroz Filho and Fagundes, 2012);
- iRegression: Regression methods for interval-valued variables (Lima Neto, 2012).

Instead, in the second case, new computational tools are now available for the analysis of symbolic data. The most important are the *SYR Software* of the *Syrokko* company[14] (Afonso et al., 2013) and the *SODAS Software Package*[15] (Diday and Esposito, 2003; Diday and Noirhomme-Fraiture, 2008).

The SDA theoretical framework is likely to be applied in various fields of study, where the researcher is not only concerned with the data itself, but he turns his interest in the latent knowledge that they hide in the form of concepts or categories of the concepts themselves.

3.4 Statistical methods for the analysis of Symbolic Data

Some important aspects of the data structure are considered in order to apply the symbolic data tools. First of all, we need two levels of units (individuals and concepts). If we consider the second level, we are dealing with class of individuals defined by the extent of a concept through its symbolic description. This description takes into account the variation of the individuals constituting its extent.

[14] http://www.syrokko.com/

[15] The SODAS 2 software is the result of the European project "ASSO" - Analysis System of Symbolic Official data (2001-2004)

The underlying process that lead to a Symbolic Data Analysis has been traced by Diday in only eight steps, as showed in Figure 3.1 (Diday, 2008).Starting from a given relational database (step 1) on which a set of modalities based on the categorical variables it contains are defined by an expert (step 2), it is possible to built a symbolic data table (step 7) on which symbolic data tools are applied (step 8).

The steps that agree this transformation starting with the specification of a detailed query (step 3). This query lead to a data table (individuals per variables) in which the categories (second column in the figure) are associated with each individual (first column in the figure) described by several variables (see chapter 2). Each category is associated with its extent, i.e. it is defined by the set of individuals that will satisfy it. This class of individuals is considered as the extent of the concept, which objectify this category (step 4). In order to associate a description with any subset of individuals, a generalization process is defined (step 5) and applied (step 6), thus producing a description of each concept, while a symbolic data table is defined, where the concepts are the units and the variables are symbolic variables that describe them (step 7).

When a symbolic data table is acquired, the next step is to conduct the most appropriate statistical analysis (step 8).

Since the first pioneer papers published on this new field by Edwin Diday, several approaches have been investigated. Different approaches have been considered by different authors in order to extend classical methods to symbolic data, from descriptive statistics to multivariate data analysis.

SDA extends the standard exploratory data analysis and data mining to the case where the units are concepts described by symbolic data (Diday, 2008).

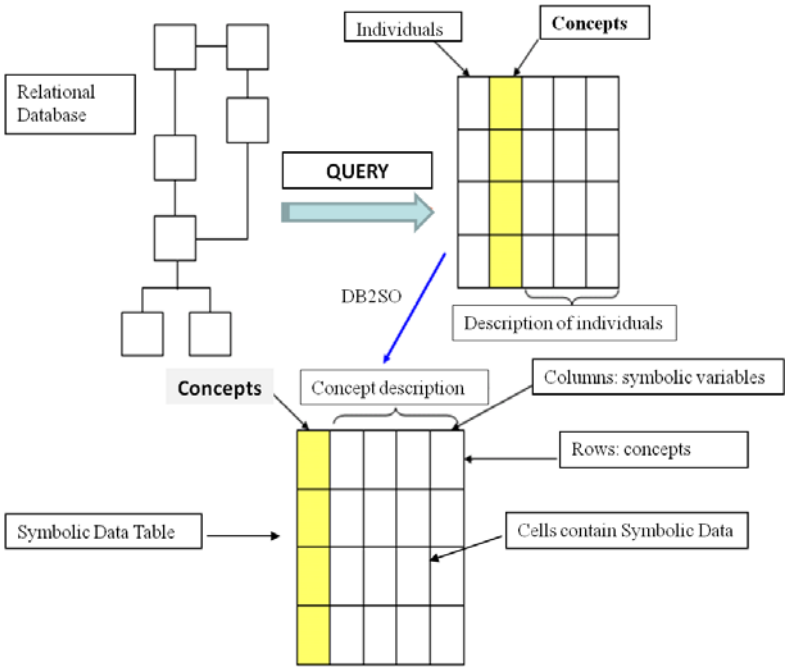

Figure 3.1 – SDA in eight steps
Source: Diday, 2008

Formally, the mathematical framework of a Symbolic Data Analysis has been summarized by Diday in the following way:

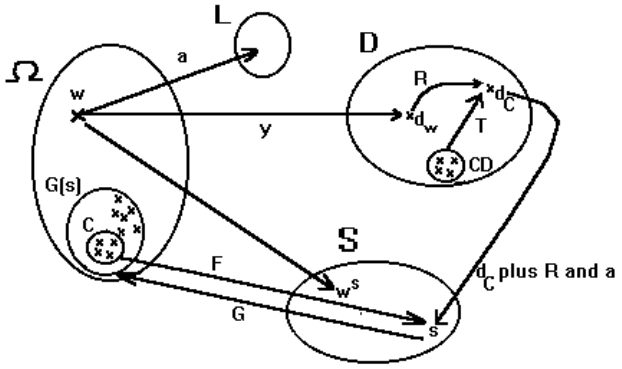

Figure 3.2 – The mathematical framework of SDA
Source: Diday, 2002

Following the notation introduced in Chapter 2, in figure 3.2 it is denoted with Ω the set of individuals, D is the description set, $L = \{true, false\}$ or $L = \{0, 1\}$ is the truth function, and S is the set of Symbolic Objects. Moreover, y is the description function while a is a membership function from Ω in L. R is the comparison relation. T, F and G are, respectively, the generalization mapping, the intension mapping and the extension mapping. $d(\omega)$ denotes the individual description given by $y(\omega)$; ω^s is an individual Symbolic Object obtained by $F(\omega) = (a, R, y(\omega))$. d_C is the description of a class $C \in \Omega$; s is the intentional Symbolic Object given by $F(C) = (a, R, d_C)$ and $G(s)$ is the extension of the Symbolic Object s (see Chapter 2).

3.5 Modelization of Concepts through the use of Symbolic Objects

From a statistical point of view there are two main ways to model the intent of a concept in order to obtain its extent giving a description of the classes of individuals that constitute its extent.

The first way consists in the description of such classes by a joint distribution of all variables , while following the second way a class is described by the endpoints associated with all the available descriptive variables.

Some difficulties arise from both ways. As in data mining, as the number of variables increase, their joint distribution tends to realize an empty space. While, if the endpoints of variables are considered, variables are not hidden in a joint distribution and the obtained results are easier to interpret, even if in this case some joint information are lost. In order to model concepts through the descriptions of their intent and the definition of the Symbolic Object, four different spaces have to be considered:

- the space of Individuals;
- the space of concepts that constitute the *Real World*;
- the space of Descriptions;
- the space of Symbolic Objects that constitute the *Modelled World.*

In the so called Modeled World the space of descriptions model individuals or classes of individuals of the Real World, while the space of Symbolic Objects models concepts. Instead, individuals and concepts in the Real World are considered as *lower and higher-level objects* (Diday, 2008).

Let consider a concept whose extent can be known is considered as the starting point, then each individual in the extent of the concept is described in the description space. If, instead, the starting point is a given class of individuals associated with the concept, then each individual in the class is described in the description space. In both cases, the concept can be modeled in the set of Symbolic Objects taking into account the obtained descriptions generalized in the space of description with an operator T, choosing the matching relation R in relation with T (see Figure 3.2 and Chapter 2).

The membership function which allow to define the Symbolic Object is able to match the description of the individuals and the description of the extent of the concept.

Moreover, if the individuals are unknown, then Symbolic Objects can be used in order to find them by exploiting the extent of a concept.

Following Diday (2008), it is possible to assert that SDA refers to classes of Symbolic Data Objects where the matching relation R is fixed, the description d varies among a finite set of comparable description and the membership function a is, by definition, the result of the comparison of the description of an individual ω to the description d.

3.6 A review of the exploratory methods for complex data structure

Several efforts have been made in order to address classical statistical methods to interval-valued data and the other typologies of symbolic variables. As regards the basic descriptive univariate and bivariate statistics, refer, among others, to the pioneer works published by De Carvalho (1994, 1995), Bertrand and Goupil (1999), Billard and Diday (2003, 2006).

Instead, for the more suitable graphical tools for visualizing complex data structures in order to identify relevant information in

large and complex information spaces, refer to i) Noirhomme-Fraiture and Rouard (1997, 2000) for the *Zoom Star Representation*, and ii) Lauro et al. (2003, 2004) for the *Symbolic Object Representations in Parallel Coordinate.*

Although not explicitly addressed in this review, many other traditional methods have been extended to Symbolic Data. Just to mention some of the most interesting ones, for the *Linear Regression Model* of interval-valued variables based on Center method, refer to Billard and Diday (2000), while Neto and De Carvalho (2008, 2010) proposed an approach based on center-range method. Billard and Diday (2006) propose a Regression model based on the first and second order moments for histogram-valued variables. Irpino and Verde (2008) developed a linear regression model based on the exploitation of the properties of a decomposition of the Wasserstein distance.

Another useful extension of SDA deals with *Discriminant Analysis*. In this framework, an important contribution can be found in Duarte Silva and Brito (2006) who suggest a three distance based approach. Instead, a *Symbolic Factorial Distriminant Analysis* has been proposed by Lauro and Palumbo, (2000).

More recently, Giordano and Brito (2014) extend SDA in the framework of *Social Network Analysis* and *Graph Theory* by representing social network as a complex data objects.

This section presents a short review of some multivariate exploratory methods proposed for Symbolic Data Analysis. In this framework, the exploratory multivariate methods are extensions of well-known classical theory.

In particular, this review will focus on Principal Component Analysis (whose classical origins can be traced back to Pearson, 1901 and Hotelling, 1933) and Clustering methods extended to complex data structure, also paying attention to the similarity and dissimilarity measures between Symbolic Data Objects.

3.6.1 Principal Component Analysis for Symbolic Data

Dealing with large multivariate datasets, the proper statistical methods used to analyze those data, are the Multidimensional Data

Analysis. Among those methods, the most common is the *Principal Component Analysis* (PCA). Its goal is to extract important information from a large dataset in order to discover and visualize useful patterns among the observations in a new and reduced space. The dimensionality reduction is one of the most important aspect when the aim is to analyze large datasets in which observations are described by several inter-correlated quantitative dependent variables.

By means of PCA the structure of a multidimensional dataset is reduced from the p original variables by a smaller number, q, of orthogonal variables listed in descending order according to their variance, called the Principal Components, which are linear combinations of the original variables, where usually $q \ll p$. More specifically, the aim is to find those Principal Components which, together, are able to explain most of the variance-covariance structure of the initial variables.

According to its classical framework, PCA is performed on a data matrix of type $X = x_{ij}$, where the generic element x_{ij} is a single value taken for the $i - th$ observation on the $j - th$ variable.

In the SDA framework, the first element to take into account is the input data table. As described in Chapter 2, the symbolic data, organized into a Symbolic Data Table of type $\underline{X} = \xi_{ij}$, presents a complex structure. Considering Interval data, in fact, the generic element of the symbolic data table is an interval value $\xi_{ij} = \left[\underline{x}_{ij}, \overline{x}_{ij}\right]$ of the feature *j* for the object i.

As in the classical framework, the aim of a *Principal Component Analysis on Symbolic Objects* is related to the dimensionality reduction of the space with the minimum loss of information. The units are concretely different. Their representations are no more *points* but become *rectangles* in a multidimensional space. It is not enough to worry about the *location* of points onto a factorial space, but it is necessary to consider the *size* (or volume) of a box and its *shape* (e.g. small or large in one or more dimensions). Thus, performing a Symbolic PCA leads to a reduced number of new interval features, known as *interval principal components* whose visualization are rectangles in the two-dimensions factorial plan (Chouakria et al., 2000).

According to statistical treatment of symbolic data, i.e. the technique, it is possible to identify different approaches. The first one

considers the input and output as Symbolic, while the treatment is made with classical data analysis techniques as proposed by Cazes et al. (1997) and Chouakria et al. (1997). It is a 2-step analysis: symbolic data are coded according to the numerical values of a rectangle's vertices or its centers on which a classic PCA is performed, then classical results are transformed into a symbolic description. This is the case of *Vertices Principal Component Analysis* (V-PCA) and *Centers Principal Component Analysis* (C-PCA). The extreme vertices projections of a Symbolic Data Object on the principal axes in a two-dimensional space define a rectangle called *Maximum Covering Area Rectangle* (MCAR). Each MCAR is coherent with the *hypercube* associated to each Symbolic Object in the input data table.

Instead, considering the input, the treatment and the output as symbolic Lauro and Palumbo (2000) proposed a different approach called *Symbolic Object Principal Component Analysis* (SO-PCA). The main idea of their method consists in the maximization of the between Symbolic Objects variance matrix. Starting from the V-PCA, the authors, in order to overcome the drawbacks of the previous approach, propose to treat the vertices of intervals no more as independent units described by points by introducing suitable constraints for the vertices belonging to the same object such that the units are considered as complex data representation. They also propose another approach known as *Principal Component Analysis on the Range Transformation* (RT-PCA) of interval variables. In order to take into the Symbolic Objects structural elements, the Authors use a range transformation that shows useful information to study size, shape and location of Symbolic Data Objects in the factorial space. Another approach to PCA of symbolic data described by symbolic intervals can be found in Lauro et al. (2008).

The *Symbolic-Symbolic-Symbolic* approach is also implemented in the SODAS/ASSO software. In both approaches Symbolic Objects are visualized in the factorial space as RMCA as a cohesive set of vertices.

Moreover, considering the intervals algebra theorems (Moore, 1966), others methods have been proposed codifying the numerical intervals by its centre, also addressed as *midpoints* and *radii*. Such approach, developed by Palumbo and Lauro (2003) are referred to as *Midpoints Radii Principal Component Analysis* (MR-PCA).

Considering both the interval linear algebra and the Symbolic-Symbolic-Symbolic paradigm Gioia and Lauro (2006) proposed an approach called *Interval Principal Component Analysis* (I-PCA). They suppose that interval variables have been previously standardized (Gioia and Lauro, 2005) and propose to perform a PCA of a interval correlation matrix in order to determine the interval of solutions on each set of units-point, i.e. the set of axes that maximizes the sum of square projections of a set of points in the plane and the their variances.

The extension of PCA to histogram variables has been proposed, among others, by Rodriguez et al. (2000). In their approach, each observation described by histogram variables is represented by a succession of the maximum number of intervals taken by some variables in the input symbolic data table. Onto the factorial plain, the empirical distribution function associated with each histogram is represented. While, Ichino (2008) addresses this method by proposing the use of a quantile representation of the data.

Makosso-Kallyth and Diday (2012) propose two adaptations of I-PCA to histogram data. The first consists in a three steps analysis: i) to code of bins of histograms, ii) to perform an ordinary PCA of means of variables and iii) to represent the dispersion of concepts through the transformation of histograms into intervals. For the second methodology they propose the angular transformation for the use of the previous three steps.

Instead, Le-Randemacher and Billard (2013) propose an algorithm to construct histogram values for the principal components of interval-valued observations.

3.6.2 Dissimilarity measures between Symbolic Data

Comparing and classifying Symbolic Objects is an important step of symbolic data analysis. It can be useful either to cluster some SOs or to discriminate between them. Several proximity measures (similarity and dissimilarity) for different type of symbolic variables have been proposed and investigated.

In particular, considering two Symbolic Data Object $u, v \in \Omega$ Gowda and Diday (1991) introduce a dissimilarity measure $D(u, v)$ for two

Symbolic Objects with three components based on position D_π, span D_s and content D_c. The dissimilarity between the two data object for the $j-th$ symbolic variable, can be defined as:

$$D(u,v) = D_\pi(u_j, v_j) + D_s(u_j, v_j) + D_c(u_j, v_j) \qquad [18]$$

where the component based on position indicates the relative positions of two variables values on a real line. The component based on span indicates the relative sizes of the variable values without referring their common parts. The component based on content measures of the non-common parts between two variables values.

Ichino and Yaguchi (1994) presented the generalized Minkowski metrics for variables expressed with different units of measurement. For an overview of the dissimilarity measures among Symbolic Data Objects, refer to Bock (2000) and Esposito et al. (2000, 2008).

Since similarity measures are defined as the inverse function of its corresponding dissimilarity measure, here only the most widely used dissimilarity and distance measures are presented. In particular, we focus on distances measures defined between Symbolic Data Objects according to the *Hausdorff* and the *Wasserstein* metrics. In mathematics, both functions measure the distance between objects on a given metric space. The first one deals with subsets of data, while the second one is defined by probability distributions.

The Hausdorff distance measure how far two subsets are from each other by measuring the distance of each point in a subsets to each point in the other subset.

In this context, Chavent and Lechevallier (2002), by using the L_1 C*ity-Block* distance, propose a *symbolic Hausdorff distance* $d_{Hj}(u, v$ between Symbolic Data Object described by intervals ξ_u, ξ_v. For the variable Y_j the proposed distance function is written as:

$$d_{Hj}(u,v) = \max\left\{\left|\underline{x}_{uj} - \underline{x}_{vj}\right|, \left|\overline{x}_{uj} - \overline{x}_{vj}\right|\right\} \qquad [19]$$

It is the maximum distance of a set of points near to another set. In other words, it is the longest among all the distances from a point in one subset to the closest point in the other subset.

An extension of the dynamic clustering algorithm based on non-adaptive Hausdorff distances proposed in (Chavent and Lechevallier, 2002) can be found in De Carvalho et al. (2006). The Authors propose a new partitional dynamic clustering method for interval data based on the use of an *adaptive Hausdorff distance* at each iteration.

Irpino and Verde (2006), propose the use of the Wasserstein metric to measure the distance between intervals by means of midpoints and radii. Following this approach, since intervals are supposed uniformly distributed, they may be expressed as the function of its midpoint and radius. So, the *squared Euclidean distance* between homologous points of two intervals of reals described by the midpoint-radius notation, can be defined as follows:

$$d_W^2(u,v) = (m_u - m_v)^2 + \frac{1}{3}(r_u - r_v)^2 \qquad [20]$$

Proximity measures for data described by distributions based on Hausdorff and Wasserstein metrics are proposed in Irpino et.al (2006), respectively, for samples data described by histograms partitioned into bins and distribution functions of two random variables taking into account their empirical distribution functions and not only their frequency functions. This means that also the upper and the lower limits are distribution functions.

Once obtained a distance matrix between Symbolic Data Objects it is possible to perform a Cluster Analysis in order to explore the similarities among them.

3.6.3 Clustering of Symbolic Data

Clustering is one of the most popular task in knowledge discovery. The aim is to classify the investigated statistical units in the initial dataset into *C clusters* which have to be internally as homogenous as possible and externally as distinct from each other as possible.

In this section Clustering Methods for complex data structures are considered. As in the classical method, in the Symbolic Data framework, this method aims to detect homogeneous groups of objects such that

the objects belonging to the same group show a high similarity whereas objects from different groups have a high degree of dissimilarity.

Cluster Analysis is an exploratory data analysis tool whose approaches can be classified according to several criteria related to the type of data, the type of classification structure, the type of proximity (dissimilarity) measure, the type of algorithm, and so on.

The most common classification structures are: hierarchies, pyramids and partitions. Hierarchical and pyramidal clustering methods produce a structure of nested cluster. Each level corresponds to a partition of the dataset if considering a hierarchy, or a family of overlapping clusters (not necessarily disjoint) if considering a pyramid. More specifically, a hierarchical classification method is constructed in a recoursive way either from successive splitting of classes, i.e. *divisive or top-down clustering* or by successive agglomeration of classes, i.e. *agglomerative or bottom-up clustering*. The first starts considering all objects in a unique class, then proceeds in dividing successively each class into smaller ones until a stopping rule prevents further divisions. The agglomerative algorithm reverses the previous process considering each unit as a single class and merging successively two classes on the bases of a similarity measure. Instead, the partitional cluster is a non hierarchical method that produces, through an iterative process, a partition of the initial dataset on a fixed number of disjoint classes.

In order to extent classical models to complex data structures, making it possible to interpret the results of a clustering methods within the same formalism of the input symbolic data table since symbolic variables allow describing classes taking into account their internal variability (Diday, 1987), a multitude of methods have been proposed.

Concerning hierarchical clustering Chavent (1998) proposes a criterion-based divisive clustering where the division is based on the optimization of a suitable generalization of the classical *within-clusters variance criterion* to the case of symbolic data. The output of this method is a hierarchy of the symbolic dataset and also a decision tree.

In contrast with divisive methods, Brito (1994, 1995, 2000) develops an agglomerative algorithm by using hierarchical-pyramidal classification structures. This algorithm allows to cluster a set of objects described by symbolic variables by means of the criterion of the

intension-extension duality. As output, a class is formed if it can be represented as a complete assertion object. Following this method, pyramidal clusters are defined as families of nested overlapping classes. This means that a class can belong to two different clusters. The method has been extended to the case in which specific hierarchical rules there exist (Brito and De Carvalho, 1999) and to dependency rules (De Carvalho, Verde and Lechevallier, 1999). Another approach for symbolic data described by interval variables has been introduced by Polaillon (2000) who develops a pyramidal cluster using Galois Lattice reduction.

In this context, it should also be mentioned the clustering method of histogram symbolic variables. Among others, Irpino and Verde (2006) present a new distance, based on the Wasserstein metric, in order to cluster a set of data described by continue distributions. They propose a hierarchical agglomerative clustering algorithm of histogram data using the Ward criterion (Ward, 1963).

As regards to dynamic clustering methods, Souza and De Carvalho (2004) have introduced an adaptive and non-adaptive partitioning cluster methods for interval data based on City–Block distances. They present two dynamic clustering methods for partitioning a set of Symbolic Objects where each object is represented by a vector of intervals. In both methods, the prototype of each cluster is represented by a vector of intervals, where the bounds of the intervals for a variable are, respectively, the median of the set of lower bounds and the median of the set of upper bounds relative to the intervals observed for the same variable of the objects belonging to the cluster.

Moreover, Lechevallier, Verde and De Carvalho (2006) propose a Symbolic Clustering methodology for large datasets that integrates the *Kohonen Sefl Organizing Maps* (SOM) with a *Dynamic Clustering algorithm of Symbolic Data* (SCLUST) considering two types of symbolic variables: multi-categorical and interval ones. It is a generalization of the standard Dynamic Clustering Method to cluster a set of concepts modeled by Symbolic Data Objects into a few number of homogeneous groups.

3.7 Concluding Remarks

Since Symbolic data may arise through a synthesis process of a huge dataset, usually too large to be conveniently managed and analyzed into the standard framework, it is possible to consider Symbolic data as a method for complex data structures.

New statistical methodologies with new ways of thinking about data are thus required in order to discover latent knowledge. Summary statistics and other common explorative methodologies, such as Principal Component Analysis, Clustering and Regression Model, have been extended to symbolic data. A variety of analytical properties make the analysis of symbolic data appealing (refer to Chapter 2), such as internal variation and logical dependences among observations.

As emphasized in this chapter, there are many advantages of performing SDA on real data are. This methods allow to extract knowledge from classical big datasets by reducing the number of both units and variables. At the same time, the output of symbolic explorative analysis leads to non-trivial interpretations. Furthermore, the visualization of Symbolic Data Object in a reduced space facilitates the recognition of new pattern and regularities in the data.

The following chapter will present an application of the reviewed methodologies for symbolic data to the symbolic concept of Italian Industrial District, as it has been defined in Chapter 2. This work will aim to present deep implications for future studies on the operationalization process and the statistical analysis of a theoretical construct, in general, and specifically of the Italian Industrial District.

4 EXPLORING THE RELATION BETWEEN GOVERNANCE AND PERFORMANCE IN ITALIAN INDUSTRIAL DISTRICTS

4.1 Introduction

This chapter presents a case study performed on real data in order to explore the existence of a relation between Governance and Performance in the Italian Industrial Districts by means of an Exploratory Symbolic Data Analysis. The structure of this chapter goes over the different notions developed in the previous chapters.

The aim is to propose a procedure for statistical analysis of a theoretical construct suitably operationalized. In this particular case, the theoretical construct is the Italian Industrial District – ID – whose operationalization has been done in the framework of Symbolic Data Analysis – SDA. Thus each Industrial District is represented as a complex data object in the SDA approach.

The main idea is that to study this concepts by means of an aggregation of the first-level units in terms of the performance ratios expressed in terms of interval or histogram-valued variables. A symbolic data table is obtained, where any raw is a Symbolic Industrial

District, while columns hold symbolic financial ratios. The study of such new entities by means of exploratory multidimensional data analysis allows to compare Industrial Districts, to classify them into homogeneous clusters according to similarity measures and to represent them in a reduced space.

In section 4.2 the research question is expressed in terms of the specific context of analysis. Section 4.3 points out the value added of this work. Section 4.4 gives the details about the subgroup of Italian Industrial Districts on which the exploratory symbolic data methods have been performed. The Data Structure described in Section 4.5 traces back the eight steps of a SDA approach in order to reach a Symbolic Data Table of the Industrial Districts. In section 4.6 the statistical methods performed for the analysis are detailed, while the main results are presented in section 4.7. Concluding remarks are given in section 4.8.

4.2 The Research Question

Moving from the theoretical framework discussed in Chapter 1, interesting research questions can be formulated about of the gap between theory and reality. Most studies underline how the efficiency of firms located into an Industrial District depends on their environmental social, cultural context and both on the governance structures. In this context, the Industrial District is supposed as an economically efficient entity consisting of a set of firms. Many difficulties appears when we try to analyze the economical performance of an Industrial District. Previous researches have been carried out looking at the Industrial District as a set of firms, obtained by a sample strategy. Firms are so considered as the statistical units of analysis, while results are extended to districts by means of summary statistics. What is worth to notice, is that, until now, the Industrial District has not been considered as a whole. Looking at this complex entity in a new perspective, i.e. considering it in a new framework of analysis through specific tools and skills, it could lead to interesting developments, such as comparison among territories, industrial sectors and their dynamics.

This work considers this complexity proposing a quantitative methodology that could enable researchers to analyze the Industrial District by considering its own properties.

At this point, conscious of the implications that this proposal could have on this complex topic, the following research question is specified in terms of an hypothetical assumption:

> *The efficiency of Industrial Districts depends on the environment. There exists a relation between the district governance and the economical and financial performance of the district as a whole*.

In order to answer this question, important methodological choices have to be addressed. First of all, the working definition of the concept and the suitable statistical methodologies have to be identified. The first one can be found in Chapter 2, where the working definition of the Italian Industrial District has been given according to the typical framework of Symbolic Data Analysis, whose review has been presented in Chapter 3. In order to pull together findings appropriately, this chapter presents a case study on some Italian Industrial Districts selected according to the research of Mediobanca-Unioncamere (2013).

The main goal is to explore the relation between Governance and Performance in Italian Industrial Districts, where this latter are newly defined as concepts.

Interesting innovations arise when looking a know topic in a new perspective. In other words, it is interesting to gain new knowledge, by continually reframing and reinterpreting events through the integration of new meanings within the complexity.

This work stands out as the starting point to develop new contextual implications for future studies designed on the process of definition of an operationalization process and the statistical analysis of a theoretical construct, in general, and specifically of the Italian Industrial District.

4.3 Framework and value added

The concept of Industrial District presents many difficulties, both in the definition and in the analysis. The research object is strictly connected to political, social, historical, and contextual issues, providing a wide range of possibilities for research questions. In this work the governance-performance relation is investigated. To give an answer to the defined research question, a new proposal for the treatment of Italian Industrial Districts is discussed.

The value added of this work consist, mainly, on the newly working definition of this complex data, as detailed in Chapter 2.

The main proposal is to move from fist-level units to second-level units by means of statistical tools that allow taking into account this latter as a whole. The application on real data may validate the importance of this approach. Specifically, consider Italian firms as first-level units and Industrial District as second-level units. The challenge is to overcame the classical approach applied in this specific filed, mainly based on atomic sample data. The Symbolic Data framework provides the fundamentals to face this transition.

In fact, the definition of the Symbolic Industrial District allows to consider this entity in its total complexity. As for the classical approach on the topic, the starting point is given by firms. The greatly differences is the statistical treatment of these data. In previous analysis only summary statistics have been taken into consideration when describing the Industrial District efficiency, mostly in terms of single-valued financial ratios. Here the Symbolic Industrial District is described by multi-valued variables. These variable keep the distributions of those in the initial dataset, without losing relevant information about the variability of the data. Internal variability of district-level financial ratios is thus considered. At the same time, symbolic exploratory multidimensional analysis allow to point out new patterns of the data, by fostering comparison among those new-defined units.

4.4 Governance and Performance of Italian Industrial Districts

Several definitions of Industrial Districts have been suggested in the reference literature. All of these definitions would like to point out relevant features of those entities, above all for the Italian context.

The consolidated district model represents the milestone of the made in Italy. Industrial Districts are presented as entities able to deal with the challenges of modernity by means of their internal structure. They benefit from a business and social cohesion much stronger than elsewhere. They are the roots on which the Italian economy has based its development over the last seventy years. They lie in the territory their deepest expression, also linked to the values, the knowledge and the tradition of the local community.

As in the past, nowadays the district areas are still presented as the excellence of the Italian production. They are the support on which is based, in particular, the Italian manufacturing sector. Behind their successes there is an organizational and governance model that has characterized the history of the districts. In many districts, the main focus is not a brand firm, but there is a collectivity of qualified supplier in the production chain strongly linked to the territory. All this, thanks to the Local Governance structures that ensure the sharing and exchange of goods, skills, know-how and human capital.

Given the strong influence of the Italian District Model, both for the local and foreign economy, a lot of studies have been made in order to census and analyze those objects. One of the main problems encountered in the analysis of Industrial Districts is due to the lack of a precise and shared mapping. Institutional organizations and private agencies carry out research surveys on to address this purpose. Among the latest empirical researches in the Italian context, we fine those published by ISTAT, CNEL, CNR, IPI, Mediobanca-Unioncamere. Furthermore, significant contributions have been also advanced by private organizations, such as Osservatorio Nazionale dei Distretti Industriali – ODI, Banca Intesa Sanpaolo, Fondazione Edison and IlSole24ore. All these researches deal with different results findings, thus generating a lot of confusion on this topic. In fact, it is possible to identify several maps of Italian Districts. They diverge both in terms of

number of districts identified, both in terms of the criteria contemplated to this purpose.

In this work we consider the 59 Industrial Districts recognized by Mediobanca-Unioncamere (2013) [16]. They represent a reliable synthesis, albeit not complete, of the Italian Industrial Districts.

Since the aim of this work is to explore the governance-performance relation in the Italian Industrial Districts, it is helpful to give the reader the considered definition of those two features.

The term Performance is here conceived in terms of profitability and financial ratios. Indeed, performance information are those issues dealing with the amount and value of money, wealth, debt, and investment of all district firms. They can be extracted from the financial statement of firm. For this purpose, the Aida database is queried.

The Governance structure is here considered as a totality of several organizations committed to support the district activities. They are different kinds of structures complied with:

- the government of the district itself, as required by the reference legislation (e.g. District Committee, Coico, Asdi);
- the support of development policies (e.g. Grant Foundations, District Observatory, shared service centers, associations, Consortium, Universities, R&D centers)
- the supply of services linked to districts' strategic choices;
- the agreement on ad-hoc planning tools of regional economy with regard to the specific manufacturing sector (e.g. development plans).

All these information have been put together into the final dataset. Thus, it will hold both performance and governance data. The process of database creation is defined below according to the eight steps specified for SDA.

[16] In Appendix A 59 Industrial Districts considered in this case study are presented according to their main productive specialization.

4.5 The Symbolic Industrial District

The process of Industrial Districts database creation follows the scheme proposed by Diday (2008) that synthesize the SDA in eight steps (see section 3.4). Some of these steps have been addressed in Chapter 2 (see sections 2.6 and 2.7), defining the Industrial District as a complex data. In synthesis, the relational Database considered for the extraction of performance data is the *Aida* database.

At this purpose the conceptual categories of districts, analyzed according to the research of Mediobanca-Unioncamera (2013), are used in the specification of the query in the [14].

Each query is associated to a specific Industrial District, thus 59 queries have been formulated and the same number of matrices have been extracted from the *Aida* database. Each matrix contains some balance sheets items of all the firms answering the query, observed in four years 2009-'12.

Querying the *Aida* database leads to a data matrix in which each concept (Industrial District) is associated with each first-level units (firms) described by several variables (performance ratios).

The structure of the raw database, X, is as follows: 16311 first-level units, i.e. Italian firms arranged according to 59 concepts, i.e. Industrial Districts, and 112 single-valued variables, i.e. 28 financial statement items and performance ratios observed in 2009-'12.

This database has been accurately treated before proceeding with the next steps of the analysis. Since it deals with secondary data, particular importance has been given to missing data. As one might expect when working on data from the public financial statements of firms, many missing values have been found. In order to obtain a database as clean as possible, the variables with a considerable number of missing values have been removed.

The same has been done as regards to units. Firms with missing values for a consistent number of variables have been removed. Instead, an imputation process has been defined for those variables with few missing values. In particular, the imputation of missing data has been done at firm-level.

Since data in the raw dataset are observed for each variable for four years, the missing values have been replaced with the expected

values. As instance, consider a missing value in correspondence of a firm for the variable observed in 2009. The missing value is replaced by the average value, obtained as an arithmetic mean of the values observed for that firm on the same variable in 2010-'12.

Once concluded the pre-treatment of the data, the dataset holds 15047 firms arranged according to 59 Industrial Districts, and 22 financial statement items and performance ratios observed in 2009-'12. An extract of this dataset is presented in the following Table 4.1:

Table 4.1 – An extract of the Industrial District dataset

Firms	Ind_District	ROA2012	ROA2011	ROA2010	ROA2009	…
1	Alessandria	7,76	10,35	11,8	6,09	…
2	Alessandria	9,43	11,4	6,74	-6,93	…
3	Alessandria	-1,4	-3,1	-1,77	-6,45	…
-	-	-	-	-	-	…
15045	Vigevanese	2,16	2,87	2,92	4,22	…
15046	Vigevanese	-5,7	-0,44	0,13	-3,18	…
15047	Vigevanese	4,51	5,11	2,66	3,28	…

As pointed out in the previous section, we are interested in second-level units, i.e. District level analysis. At this purpose the district as a whole has to be introduced. Thus the Symbolic Industrial District is defined in the symbolic data framework. In this framework, each concept is associated with its extent.

This means that each Industrial District is defined by the set of firms that satisfy its extent. In other words, this subset of firms is considered to be the extent of the Symbolic Industrial District which operationalize the concept.

A symbolic data table is defined, where Symbolic Industrial Districts are the units described by multi-valued variables. The Figure 4.2 shows the transformation process that allows the definition of the Symbolic Industrial District. This process starts with the specification of the queries for Aida database and it ends with the organization of data into the symbolic data table. In the resulting symbolic data table, each raw is a Symbolic Industrial District, each column is a multi-valued performance ratio and each cell contains symbolic data.

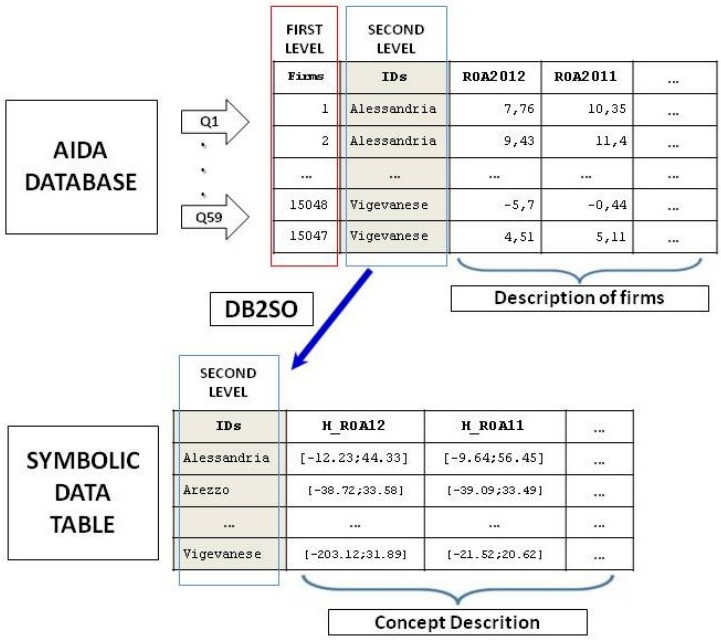

Figure 4.1 – The Symbolic Industrial District

The fine dataset is structured as follows: the units are the 59 Symbolic Industrial Districts described by 88 symbolic variables (i.e. the distributions of the 22 financial statement items observed in 2009-'12).

The Table 4.2 presents an extract of this dataset. Moreover, qualitative information observed at this level of analysis can be added. Specifically, governance attributes (four variables indicating the presence/absence of District Committee, strategic development plan, service center and reference institution), territorial location (Italian Region and geographical subdivision) and industrial sector (manufacturing sector and macro-sector).

Table 4.2 – An extract of the Symbolic Data Table of Industrial Districts

Ind_District	ROA_12	ROI_12	ROE_12	...
Alessandria	[-12.23;44.33]	[-24.54;28.75]	[-67.75;74.46]	...
Arezzo	[-38.72;33.58]	[-28.39;29.98]	[-90.15;85.14]	...
Barletta	[-68.27;28.31]	[-120.94;20.88]	[-5.09;20.88]	...
-	-	-	-	...

In order to achieve our aim and to underline the main results of this research, the following methodologies will be performed on a subset of the fine database. Aiming to explore the main relationships in the data by means of exploratory symbolic methods, a variable selection procedure is applied. As for the performance variables, among all the financial statement items, we consider only those measuring profitability and financial aspects observed in 2012. Respectively, as for profitability ratios we consider: *Return on Investment –ROI*, *Return on Equity – ROE*, *Return on Assets – ROA*, *Return on Sales – ROA* – and *Ebitda/sales* (i.e. *Earning before interests, taxes depreciation and amortization*). As for financial ratios we consider: *Financial Autonomy*, *Finance Expense*, Solvency, *Liquidity* and *Leverage*.

At the same time, a constraint for the observations is defined. The criterion used to cut the dataset is strictly related with the symbolic data framework. As for symbolic concepts, we consider only those districts whose extent is made up of at least 20 first-level units (i.e. the output of the query imposed on the relational database).

The Figure 4.1 shows the bar-plot of the smaller IDs.

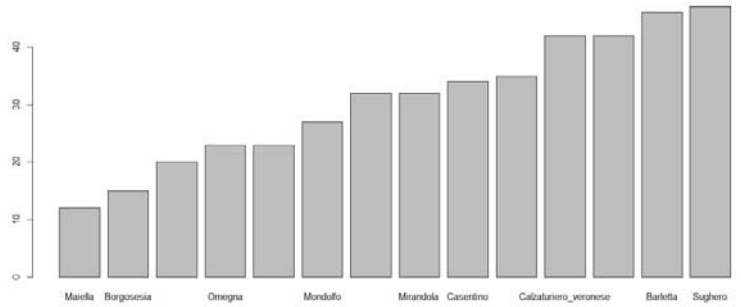

Figure 4.2 – The IDs distribution according to the number of firms
Cutting threshold is set to 20 first-level units

The final database, denoted with $\underline{X}$, of size 57×18, holds by row the 57 Symbolic Industrial Districts and by column 10 multi-valued profitability and financial ratios and 8 additional district attributes concerning governance, geographical location and manufacturing sector. The table 4.3 presents an extract of the final dataset.

Table 4.3 – An extract of the Final Symbolic Data Table of IDs

Sym_Ind_Dis	ROA_12	ROI_12	…	GOV
Alessandria	[-12.23;44.33]	[-24.54;28.75]	…	{Committee;…;Plan}
Arezzo	[-38.72;33.58]	[-28.39;29.98]	…	{Committee;…;PlanNr}
Barletta	[-68.27;28.31]	[-120.94;20.88]	…	{Committee;…;Plan}
-	-	-	…	

Each row of the final database is a Symbolic Industrial District. The Figure 4.3 shows the description of the "Fermo" Symbolic District performance according to the distributions of the considered financial ratios.

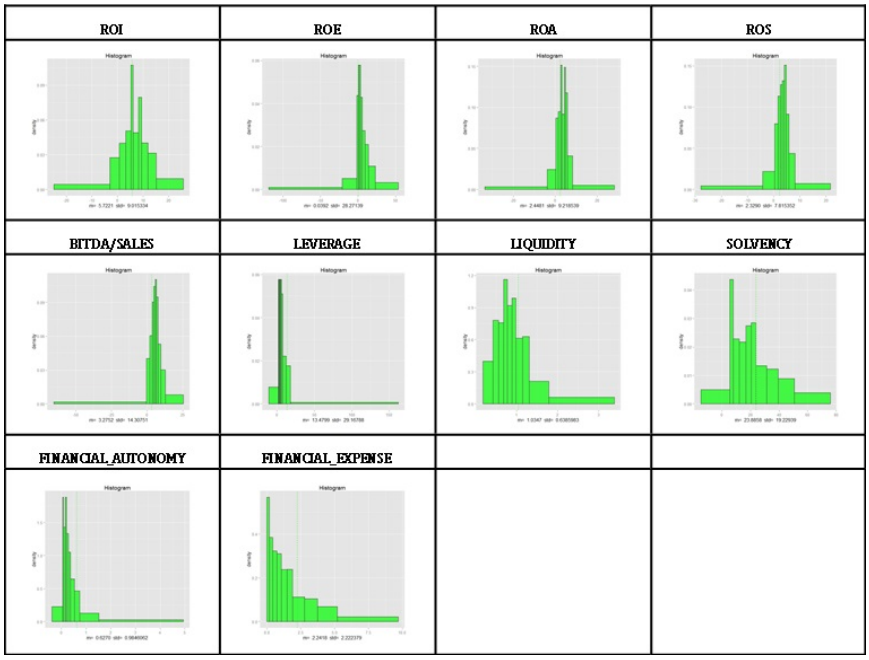

Figure 4.3 – The Fermo Symbolic Industrial District

4.6 The analysis of Symbolic Industrial Districts

The classical data analysis of the financial ratios, conducted at the firm level, allows to know the average position of a group of firms by

considering a specific classification (e.g. the relevant sector) through the arithmetic average (simple or weighted) and or the median. Moreover, by a detailed comparison between the quantiles of the excellent firms (best-practice), it allows to compare the performance of a firm with the others in the group, thus identifying the actual performance levels that it has been able to accomplish.

Since there is a multiplicity of financial statement ratios and all of them have to be considered in this analysis, both their comprehensive analysis and comparison is difficult when considered them separately. For a correct interpretation of income and financial data derived from firms financial statements, it could be useful to consider more financial ratios simultaneously on the same data and, at the same time, to compare their results with the corresponding ones observed for other companies.

From the statistical point of view we have a double solution. This problem can be solved by synthesizing information of several indices through their simultaneous reading. Instead the comparison among firms can be made by considering simultaneously all the ratios for all firms. The statistical methodologies that allow the simultaneous analysis of multiple variables observed are the multivariate statistical methods.

As stated above is reliable for a classic treatment of single-valued data observed at first-level units. Rather, in this work, the interest is focused on second-level units, specifically on the district level analysis.

The aim is to provide tools that consider the second-level units as a whole, not reducing them to a point onto a space, but preserving its internal variability. In this work, the reference unit of analysis is the Industrial District, but the operationalization of theoretical features, not only of economic type, can be extended to other kind of classification variables, if they are properly defined and treated (see Chapter 2).

Such perspective move from classical to symbolic data, so the researchers can manage complex data structures. The Industrial District as a whole is considered as the row of the symbolic data table.

The Symbolic Industrial Districts are represented by several multi-valued numerical descriptors defining their performance. Among a wide range of interests, we focus on the analysis of the performance situation of these complex data objects, taking into account all the

available ratios. We are also interested in the comparison among Symbolic Industrial Districts in order to highlight the position of each one of them in relation with the investigated collective.

In order to synthesize and reduce the multi-valued financial and profitability ratios, the Symbolic Principal Component Analysis provides a valuable support to identify those seminal ratios that emphasize the differences between district performance. These ratios are the one that show an high correlation with the most meaningful principal components (they explain the higher proportion of variance).

A further advantage of PCA lies in the possibility to include additional information on the achieved results. Although the PCA is a method of analysis designed for quantitative variables, it allows the projection of additional variables on the factorial plane, also nominal-type variables. Supplementary variables do not participate in the principal components identification. They are simply projected onto the factorial space to increase the interpretation of the phenomenon highlighting the latent factors. The same procedure can be applied when performing a Symbolic PCA. In this work we consider Governance attributes as supplementary variables in order to explore the relation structural features of the Symbolic Industrial District and its performance.

Moreover, the Symbolic PCA, as a method of transformation and reduction of variable and observation space, is very useful also for further statistical analysis. Considering the output of Symbolic PCA, a clustering method may be performed on complex data. In this case, the interest is focused on the identification of homogeneous groups of symbolic districts with respect to financial ratios appropriately summarized in the main components. This allows to identify homogeneous groups of Symbolic Industrial Districts with similar behavior and/or excellent performance. This information can be extremely useful, not only in exploratory terms, but also in terms of quantitative benchmarking.

In the following section, referring to the SDA methods reviewed in Chapter 3, the explorative analyses of the identified Symbolic Industrial Districts will be presented.

4.7 Main findings of the case study

In order to answer to the research question, explorative multivariate analysis have been performed. This section presents the main summary results of the Symbolic Industrial Districts analysis.

Moving from the shared idea that high performance in Industrial Districts are related with the presence of well-established governance structures, we are looking for the right methods to explore this relation.

We aim to analyze the Symbolic Industrial District performance considering, at the same time, several ratios by means of exploratory data analysis. These methods allow us to project supplementary information useful for the interpretation of the resulting factorial axes. Consequently, we are able to explain the governance-performance relation within the complex objects under study.

In particular, we carry out a Symbolic Principal Components for the multi-valued performance variables (histogram type) considered in the final database $\underline{X}$ (see section 4.5). Furthermore, a clustering method is presented in order to underline the different composition of the resulting groups. Specifically, supplementary variables are used for the description of each cluster.

As specified in section 4.5, at first, we consider a subset of the Symbolic Industrial District t. These second-level units we analyze are the 57 Symbolic Industrial Districts.

The Histogram PCA shows the following results. The cumulative percentage of variance of the first two components spans the 75% of the total variance (see Table 4.4). So we decide to retain these two components for further investigations. Table 4.5 and Figure 4.4 show the correlation between the symbolic performance ratios and the first two components. We notice that on the first axis there is a contraposition of financial versus profitability ratios. In particular, the classical *size effect* appears, indeed the districts characterized by high level of performance are on the left side, opposed to those districts that show high levels of debts. Onto the second axis a contraposition between trading profitability and financial solvency ratios appears. In this sense, the *shape effect* appears.

Table 4.4 – Eigenvalues of the first five Principal Components

	Eigenvalue	Percentage of Variance	Cumulative Percentage of Variance
comp 1	12590,43	48,63	48,63
comp 2	7017,95	27,11	75,74
comp 3	3116,68	12,04	87,78
comp 4	921,58	3,56	91,34
comp 5	673,51	2,60	93,94

Table 4.5 – Correlation of the symbolic performance variables with PCs

	Component 1	Component 2
EBITDA_sales	-0,68	-0,11
ROS	-0,84	0,17
ROI	-0,53	0,74
ROA	-0,79	0,53
ROE	-0,55	0,77
Liquidity	-0,88	-0,23
Leverage	0,64	0,57
FinacialExpense	0,49	-0,35
FinancialAutonomy	-0,73	-0,60
Solvency	-0,74	-0,62

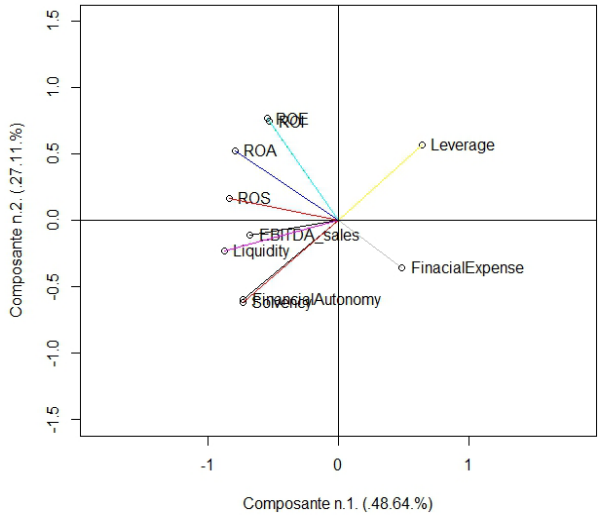

Figure 4.4 – Correlation circle of symbolic performance variables (75.74%)

The factorial plan of the symbolic Industrial Districts, here represented as MCAR, shows the presence of some peculiar districts.

Notice that the dimension of the RCMA is due to the variability of the data into each district. In particular, we underline the districts, which present a high variability together with those homogeneous districts which show a high level of profitability. The Casarano Industrial District shows high leverage ratio, but also high heterogeneity among its firms. The Civita Castellana Industrial District is characterized, above all, for high financial expense ratio. Gallaratese and Alessandria Industrial Districts show high financial autonomy and high level of sales revenue. In general, we observe that the first component divides the districts which show best performance (on the left side) against those which show solvency issues (on the right side).

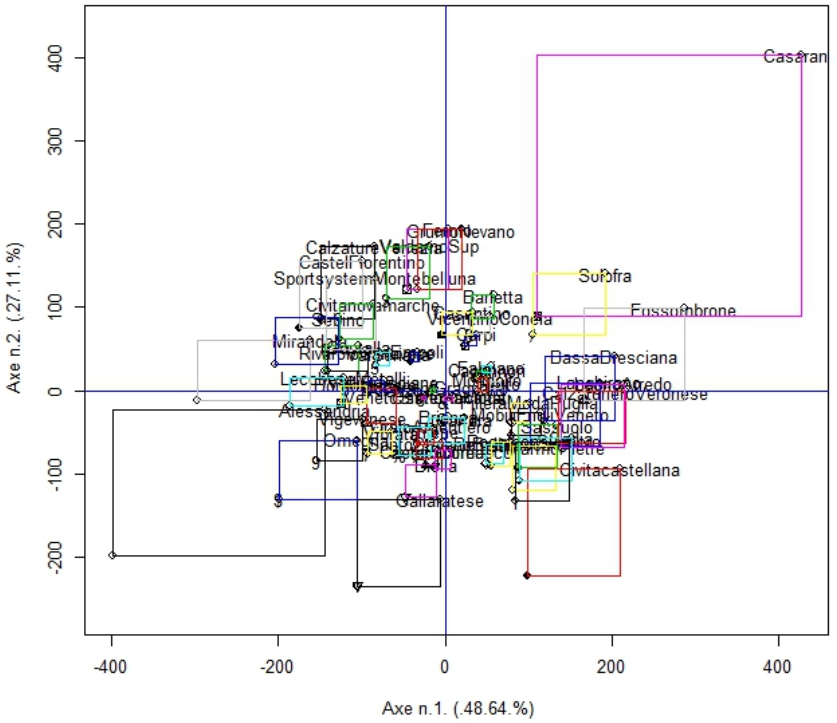

Figure 4.5 – Factorial Plan of the Symbolic Industrial Districts (75.74%)

Figure 4.6 shows the symbolic Industrial Districts represented by their midpoints together with the projection of the supplementary information as in the classical PCA.

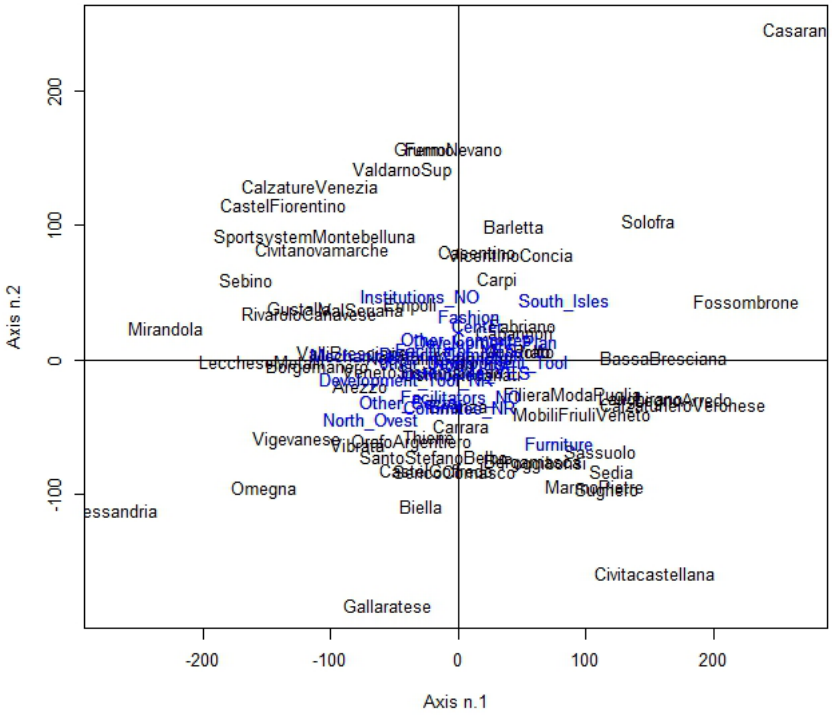

Figure 4.6 – Factorial Plan with Symbolic Industrial Districts and supplementary variables (75.74%)

In particular, the projection of supplementary variables allows to explore the relation between those attributes and the districts performance. Considering the governance attributes, we can affirm that the districts showing a high level of leverage ratio are those characterized by the agreements of strategic development plans (Development_Plan). In contrast, districts characterized by high level of solvency index are those where development tools have not been observed (Development_Tool_NR). Regarding the District Committee and Facilitators, we can affirm that their presence (District_Commitee and Facilitarors_YES)/absence (Committee_NR and Facilitarors_NO) affect, respectively, those districts opposed on the second axis.

Furthermore, the lack of institutions (Institutions_NO) is related with a high level of profitability ratios. On the contrary, their presence (Institutions_YES) seems to not affect the performance of the districts.

In order to gain a better visualization of these attributes the Figure 4.7 presents a focus of the supplementary variables projections on the factorial plan.

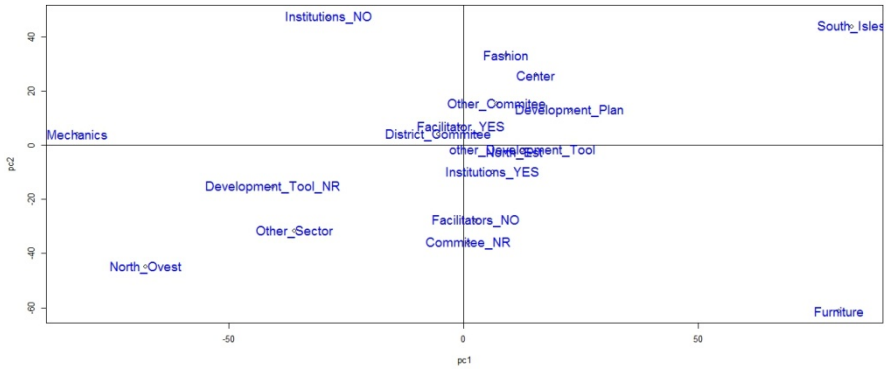

Figure 4.7 – Zooming of supplementary variables projection

The Hierarchical Clustering of the Symbolic Industrial Districts highlights the presence of homogeneous groups of districts. In particular their characterization is related both on performance and supplementary information. We obtain a partition into six groups that will be characterized as follows. Specifically, Table 4.6 shows the attribution of each ID to each cluster.

Table 4.6 – Clusters of Industrial Districts

Cluster	Industrial Districts
Cluster 1	Alessandria, Arezzo, Borgomanero, Empoli, Gallaratese, Gustalla, Lecchese Metalli, Mirandola, Omegna, Rivarolo Canavese, Sebino, Valli Bresciane, Val Seriana, Vibrata, Vigevanese
Cluster 2	Barletta, Capannori, Carpi, Casentino, Fabriano, Mondolfo, Prato, Vicentino Concia
Cluster 3	Bassa Bresciana, Calzaturiero Veronese, Casarano, Fossombrone, Langhirano, Legno Arredo, Solofra

Cluster 4	Bergamasca, Civita castellana, Filiera Moda Puglia, Marmo Pietre, Mobili Friuli Veneto, Poggibonsi, Sassuolo, Sedia Friuli, Sughero
Cluster 5	Biella, Brianza, Carrara, Castel Goffredo, Nocera Inferiore Gragnano, Orafo Argentiero, Osimo Recanati, Santo Stefano Belbo, Serico Comasco,Thiene, Veneto Sistema Moda
Cluster 6	Calzature Venezia, Castel Fiorentino, Civitanova Marche, Fermo, Grumo Nevano, Sport system Montebelluna, Valdarno Superiore

Furthermore, tables from 4.7 to 4.10 show the characterization of each cluster according to governance attributes (institutions, District Committee, Facilitators, Development Plans), instead table 4.11 shows the manufacturing sectors.

Concerning the governance attributes considered in this work, they characterize the analyzed districts as follows. Cluster 1 is mainly composed of highly managed districts, since we find Institutions (86.7%), District Committee or other forms of government (66.7%) and Facilitators (66.7%).

Furthermore, considering jointly the modalities observing the presence of these support tools, we find in almost all the clusters the districts whose success is mainly due to these aspects. Specifically, the agreement of development plans characterize Cluster 5 (63.6%).

As regards the productive specialization here considered in Macro-sectors, mechanics districts are mainly grouped in Cluster 1 (40%), indeed the Furniture districts are grouped in Cluster 4 (44%). The Fashion districts are disclosed among the clusters, mainly characterizing Cluster 6(100%), Cluster 2 (75%) and Cluster 3 (57%).

Table 4.7 – Institutions in IDs Cluster

a. % cluster/modality

DI_Clust	Inst_NO	Inst_YES
1	13.3	86.7
2	12.5	87.5
3	0.0	100.0
4	11.1	88.9
5	27.3	72.7
6	42.9	57.1

b. % modality/cluster

DI_Clust	Inst_NO	Inst_YES
1	20.0	27.7
2	10.0.	14.9
3	0.0	14.9
4	10.0	17.0
5	30.0	17.0
6	30.0	8.5

Table 4.8 – District Commitee in IDs Cluster

a. % cluster/modality

DI_Clust	DisCom	Other	NR
1	53.4	13.3	33.3
2	50.0	50.0	0.0
3	43.0	28.5	28.5
4	44.5	22.2	33.3
5	45.0	45.0	1.0
6	42.9	57.1	0.0

b. % modality/cluster

DI_Clust	DisCom	Other	NR
1	29.7	10.5	45.4
2	14.8	21.1	0.0
3	11.1	10.5	18.2
4	14.8	10.5	27.3
5	18.5	26.3	9.1
6	11.1	21.1	0.0

Table 4.9 – Facilitators in IDs Cluster

a. % cluster/modality

DI_Clust	Fac_NO	Fac _YES
1	33.3	66.7
2	0.0	100.0
3	14.3	85.7
4	44.4	55.6
5	9.0	91.0
6	0.0	100.0

b. % modality/cluster

DI_Clust	Fac _NO	Fac _YES
1	45.4	21.7
2	0.0	17.4
3	9.1	13.0
4	36.4	10.9
5	9.1	21.7
6	0.0	15.2

Table 4.10 – Development Plan in IDs Cluster

a. % cluster/modality

DI_Clust	DevPlan	Other	NR
1	13.3	33.3	53.4
2	50.0	25.0	25.0
3	42.9	42.9	14.2
4	44.5	33.3	22.2
5	63.6	0.0	36.4
6	57.1	28.6	14.3

b. % modality/cluster

DI_Clust	DevPlan	Other	NR
1	8.3	33.4	44.4
2	16.7	13.3	11.1
3	12.5	20.0	5.6
4	16.7	20.0	11.1
5	29.1	0.0	22.2
6	16.7	13.3	5.6

Table 4.11 – Manufacturing Sector in IDs Cluster

a. % cluster/modality

DI_Clust	Furniture	Fashion	Mechanics	Other
1	6.6	26.7	40.0	26.7
2	0.0	75.0	12.5	12.5
3	28.6	57.1	0.0	14.3
4	55.6	22.2	0.0	22.2
5	18.2	45.4	9.1	27.3
6	0.0	100.0	0.0	0.0

b. %modality/cluster

DI_Clust	Furniture	Fashion	Mechanics	Other
1	10.0	14.3	75.0	36.3
2	0.0	21.4	12.5	9.1
3	20.0	14.3	0.0	9.1
4	50.0	7.1	0.0	18.2
5	20.0	17.9	12.5	27.3
6	0.0	25.0	0.0	0.0

The dendrogram in Figure 4.8 shows how districts cluster together. What is interesting to notice is the visualization of these cluster onto the factorial plan. At this purpose, the Figure 4.9 shows the projection of the Symbolic Industrial Districts onto the factorial plan, where these objects take different colors according to their belonging cluster. The hierarchical clustering confirms the characterization of IDs both in terms of performance and governance.

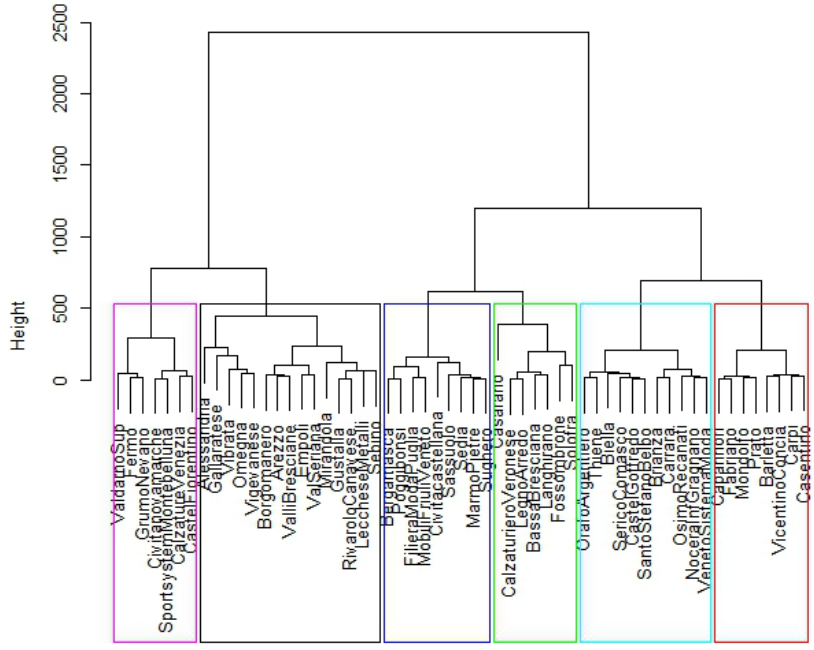

Figure 4.8 – Dendrogram of Symbolic Industrial Districts

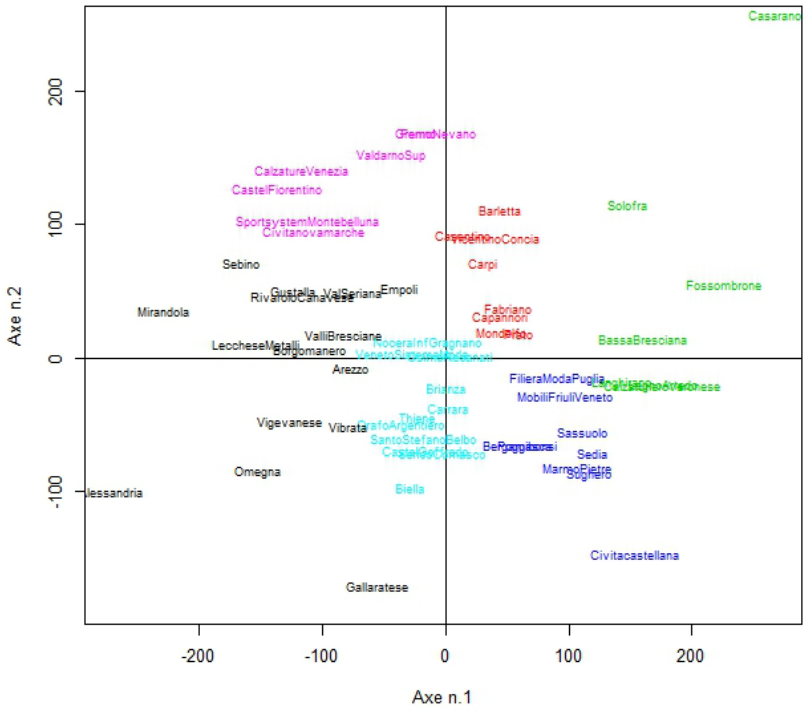

Figure 4.9 –Clusters of Symbolic Industrial Districts projected onto the factorial plan (75.7%)

Additional analysis have been carried out on Symbolic Industrial Districts. The following applications will help us to illustrate the importance of the proposed approach.

We also aim to explore at what extent the dimension of firms involved in Industrial Districts affect the performance and the governance system of the Symbolic Industrial District as a whole.

In order to find the best criteria to cut the initial dataset, the starting point is the European Commission Recommendation No. 301 dated 06/05/2003, effective from 01/01/2005, concerning the definition of micro, small and medium-sized firms. Following the criteria established by the European legislation, we are able to define a cutting threshold criterion, that lead us to introduce another important aspect of this topic: the size of the firms operating in Industrial

Districts. This aspect is generally defined in terms of number of employee - Y_{size} and Sales revenue in thousands euro - Y_{income}.

The criteria used to cut the initial dataset into four subsets according to the dimension of firms (i.e. micro, small, medium and large) operating in IDs dimensions, denoted with X_{micro}, X_{small}, X_{medium} and X_{large}, is specified as follows:

$$\forall \omega_i \in X \subset \Omega,$$

$$\omega_i \in X_{micro} \Leftrightarrow Y_{size} \in [0; 10[\wedge Y_{income} \in [0; 2{,}000]$$

$$\omega_i \in X_{small} \Leftrightarrow Y_{size} \in [10; 50[\wedge Y_{income} \in]2{,}000; 10{,}000] \qquad [21]$$

$$\omega_i \in X_{medium} \Leftrightarrow Y_{size} \in [50; 250[\wedge Y_{income} \in]10{,}000; 50{,}000]$$

$$\omega_i \in X_{big} \Leftrightarrow Y_{size} > 250 \wedge Y_{income} > 50{,}000$$

Applying the criteria specified in the [21] we obtain datasets of different size, according to the number of firms that satisfy them. Specifically: X_{micro} contains 2908 firms, X_{small} contains 9052 firms, X_{medium} contains 2501 firms and X_{large} contains 533 firms.

In order to study the relation of governance and performance in Symbolic Industrial District, considering this important aspect, the procedure showed in Figure 4.1., together with the cutting threshold criteria, is reiterated for each subset. Figures 4.10 to 4.13 show, respectively, the bar-plot of the smaller IDs in each resized dataset.

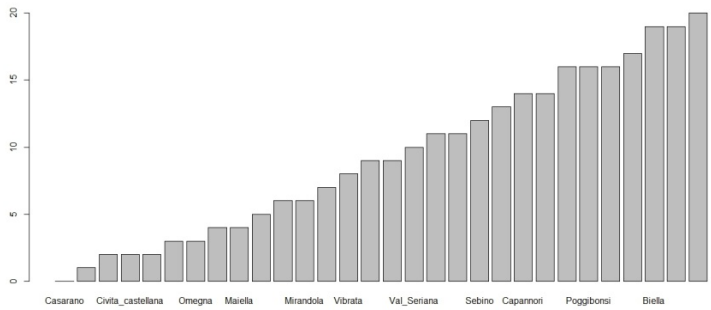

Figure 4.10 – The IDs distribution according to the number of micro firms
Cutting threshold is set to 20 first-level units

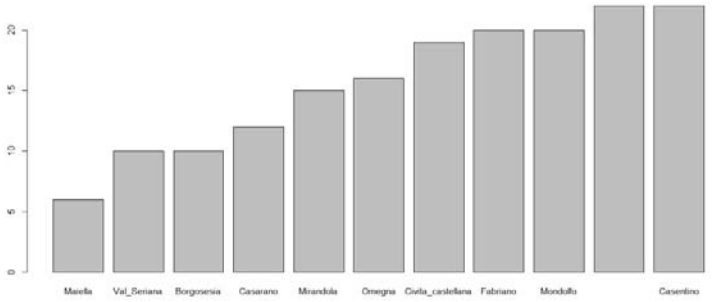

Figure 4.11 – The IDs distribution according to the number of small firms. Cutting threshold is set to 20 first-level units

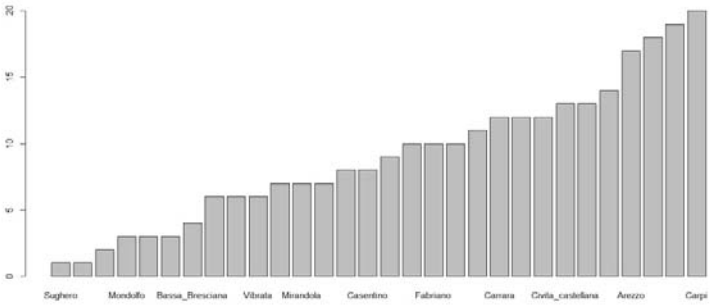

Figure 4.12 – The IDs distribution according to the number of medium firms. Cutting threshold is set to 20 first-level units

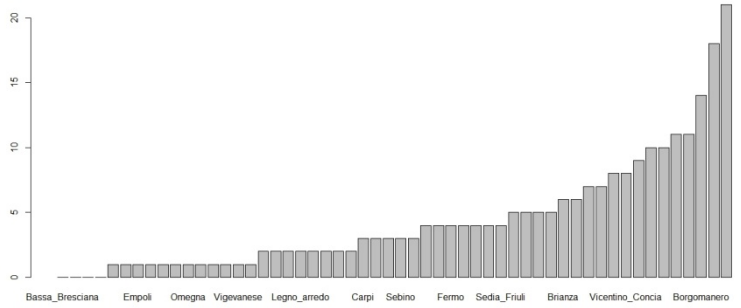

Figure 4.13 – The IDs distribution according to the number of large firms. Cutting threshold is set to 20 first-level units

As result, we obtain four Symbolic Industrial Districts data tables, denotes as $\underline{X}_{micro}$, $\underline{X}_{small}$, $\underline{X}_{medium}$, $\underline{X}_{large}$, holding, respectively, 30, 51, 30 and 6 Symbolic Industrial Districts.

We assert that the dimension of firms operating into the boundaries of an ID may affect the governance system and the economical performance of the district itself. Because of the small number of observations in $\underline{X}_{large}$, we carry out a Symbolic Principal Component Analysis for histogram-type variables only on the other three subsets. The projection of the supplementary variables onto the resulting factorial plans is considered in order to underline different governance-performance patterns. At this purpose, here only the graphical representations of the exploratory analysis performed on each symbolic data table are showed in order to underline the meaningful of the proposed application on different subset of the data.

For each subset we retain the first two principal components, since they span a high percentage of the total variance.

As regards Symbolic IDs of micro-sized firms, $\underline{X}_{micro}$, the first two principal components spans the 69.27% of the total variance. As for $\underline{X}_{small}$ they span the 72.73% of the total variance, while for $\underline{X}_{medium}$. they span the 74.4% of the total variance (see Table s 4.12 to 4.4.14).

Table 4.12 – Eigenvalues of the first five Principal Components considering only micro-sized firms in Symbolic IDs

	Eigenvalue	Percentage of Variance	Cumulative Percentage of Variance
comp 1	14306.85	53.85	53.85
comp 2	4095.34	15.42	69.27
comp 3	3864.80	14.55	83.82
comp 4	2057.87	7.75	91.57
comp 5	826.25	3.11	94.68

Table 4.13 – Eigenvalues of the first five Principal Components considering only small-sized firms in Symbolic IDs

	Eigenvalue	Percentage of Variance	Cumulative Percentage of Variance
comp 1	11528.37	45.98	45.98
comp 2	6708.62	26.75	72.73
comp 3	3105.97	12.39	85.12
comp 4	1231.94	4.91	90.03
comp 5	923.61	3.68	93.71

Table 4.14 – Eigenvalues of the first five Principal Components considering only medium-sized firms in Symbolic IDs

	Eigenvalue	Percentage of Variance	Cumulative Percentage of Variance
comp 1	11169.21	47.81	47.81
comp 2	6212.93	26.60	74.41
comp 3	3012.67	12.89	87.3
comp 4	1248.47	5.34	92.64
comp 5	829.58	3.55	96.19

Tables from 4.15 to 4.17 show the correlation between the symbolic performance ratios and the first two components. The correlation circles obtained for each subset of symbolic IDs are, respectively, presented in Figure 4.14, Figure 4.16 and 4.18. Notice that, as in the previous analysis, on the first axis there is a contraposition of financial versus profitability ratios.

Table 4.15 – Correlation of the symbolic performance variables with PCs considering only micro-sized firms in Symbolic IDs

	Component 1	Component 2
EBITDA_sales	0.70	0.65
ROS	0.89	0.37
ROI	0.58	-0.12
ROA	0.94	0.10
ROE	0.73	-0.01
Liquidity	0.86	-0.31
Leverage	-0.55	-0.46
FinacialExpense	-0.24	0.77
FinancialAutonomy	0.78	-0.06
Solvency	0.80	-0.12

Table 4.16 – Correlation of the symbolic performance variables with PCs considering only small-sized firms in Symbolic IDs

	Component 1	Component 2
EBITDA_sales	-0.69	-0.14
ROS	-0.87	0.12
ROI	-0.67	0.63
ROA	-0.84	0.43
ROE	-0.61	0.71
Liquidity	-0.81	-0.21
Leverage	0.47	0.59
FinacialExpense	0.32	-0.50
FinancialAutonomy	-0.68	-0.67
Solvency	-0.65	-0.67

Table 4.17 – Correlation of the symbolic performance variables with PCs considering only medium-sized firms in Symbolic IDs

	Component 1	Component 2
EBITDA_sales	0.57	0.47
ROS	0.87	0.19
ROI	0.93	-0.01
ROA	0.95	-0.01
ROE	0.93	-0.25
Liquidity	0.28	0.76
Leverage	0.28	0.75
FinacialExpense	-0.55	0.18
FinancialAutonomy	-0.06	0.92
Solvency	0.00	0.94

The MCAR of the symbolic Industrial Districts obtained for each subset are shown in Figure 4.15, Figure 4.17 and Figure 4.19. As for $\underline{X}_{micro}$ and $\underline{X}_{medium}$ we observe that the first component divides the districts which show solvency issues (on the left side) against those which show best performance (on the right side). As for $\underline{X}_{small}$ the interpretation of the first component is the opposite: on the left side we find those IDs which show best performance against those which are characterized by better overall financial outcomes (on the right side).

Furthermore, Figures from 4.20 to 4.22 show, for each subset, the projections of supplementary variables onto the factorial plans. All these three plans show an important relation of the supplementary attributes in the interpretation of the components.

In particular, as for the $\underline{X}_{micro}$ and $\underline{X}_{medium}$ an opposition between Mechanics and Furniture IDs appears. Moreover, looking at Figure 4.20 this opposition is also concerned with the agreement of development plans and the presence of district committee, which characterize Furniture IDs, versus the absence of those governance attributes in the Mechanics ones. Merging this interpretation with the previous ones concerning performance, we can affirm that there is a relation between governance attributes and financial performance, since districts characterized by high level of solvability ratios are also those that show high concentration of governance attributes. This is also true for the representation of $\underline{X}_{small}$ principal components. Here, also the geographical location appears as a discriminant feature of the districts.

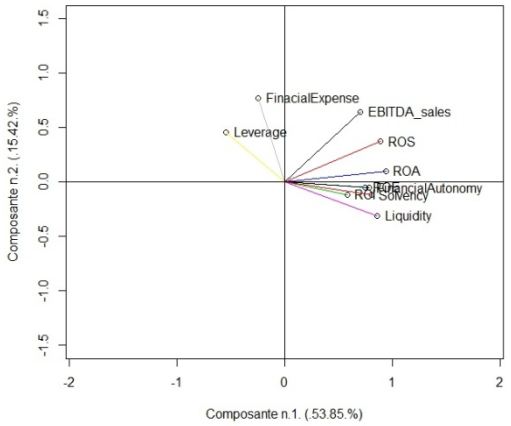

Figure 4.14 – Correlation Circle of symbolic performance variables considering only micro-sized firms in Symbolic IDs (69.27%)

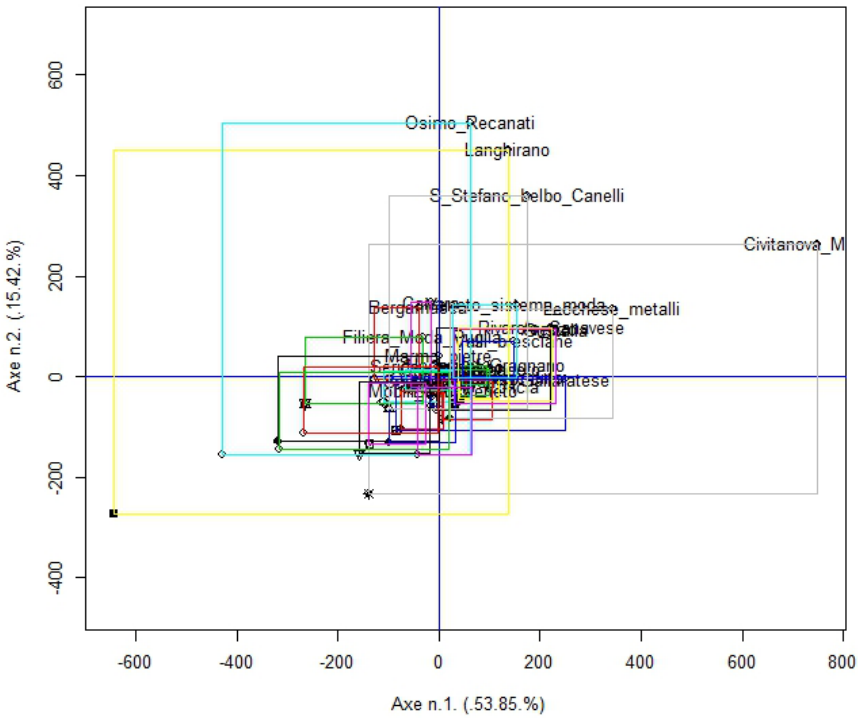

Figure 4.15 – Factorial Plan of Symbolic IDs considering only micro-sized firms in Symbolic IDs (69.27%)

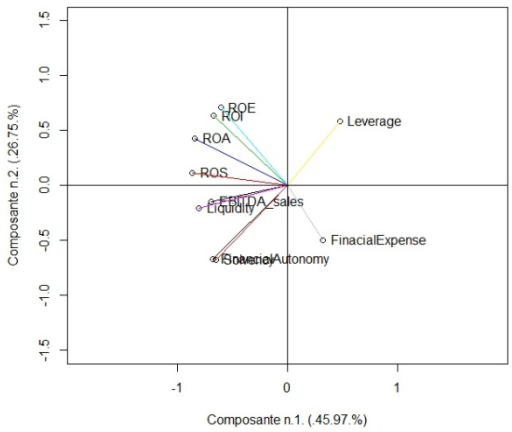

Figure 4.16 – Correlation Circle of symbolic performance variables considering only small-sized firms in Symbolic IDs (72.73%)

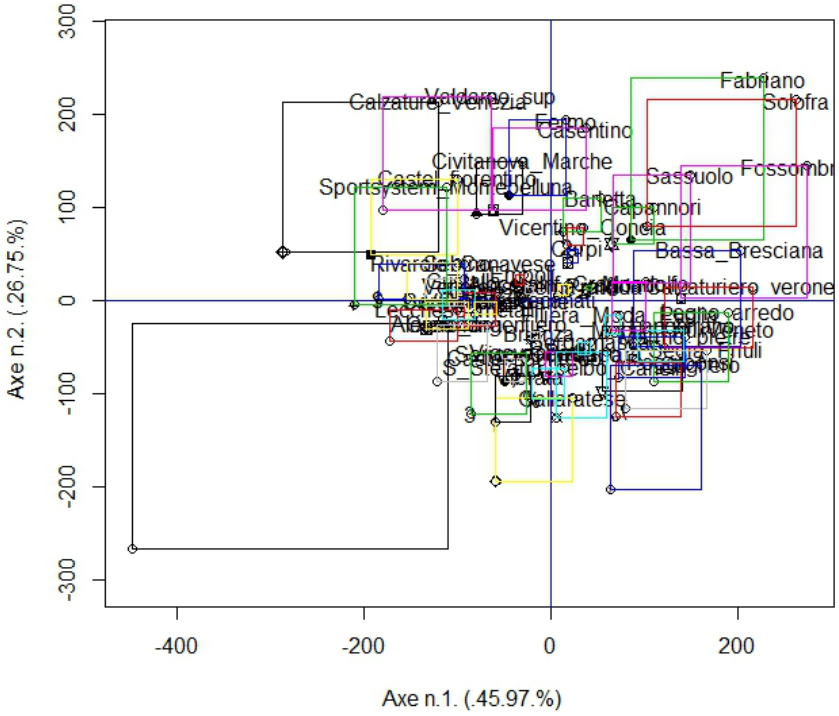

Figure 4.17 – Factorial Plan of Symbolic IDs considering only small-sized firms in Symbolic IDs (72.73%)

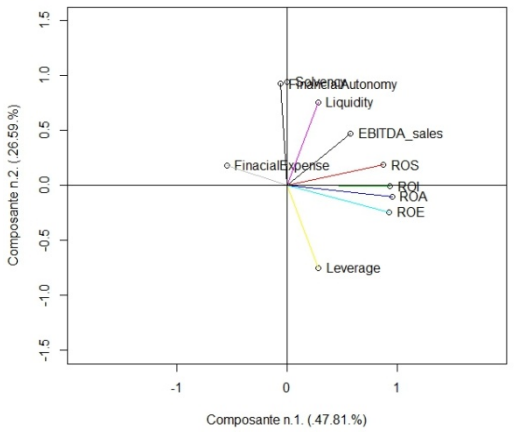

Figure 4.18 – Correlation Circle of symbolic performance variables considering only medium-sized firms in Symbolic IDs (74.4%)

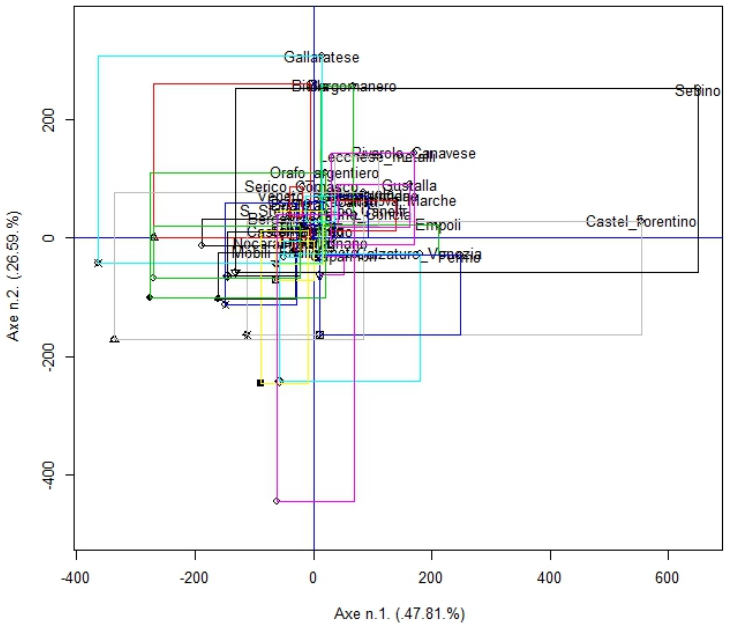

Figure 4.19 – Factorial Plan of Symbolic IDs considering only medium-sized firms in Symbolic IDs (74.4%)

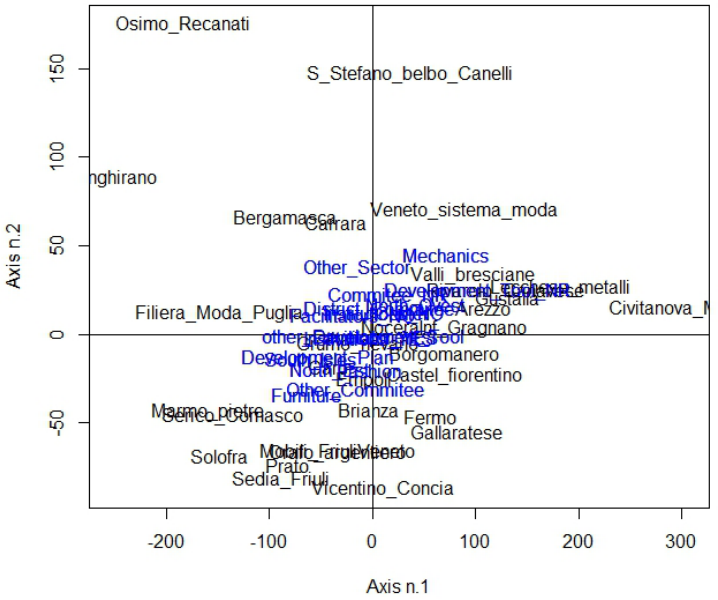

Figure 4.20 – Factorial Plan with supplementary variables considering only micro-sized firms in Symbolic IDs (69.27%)

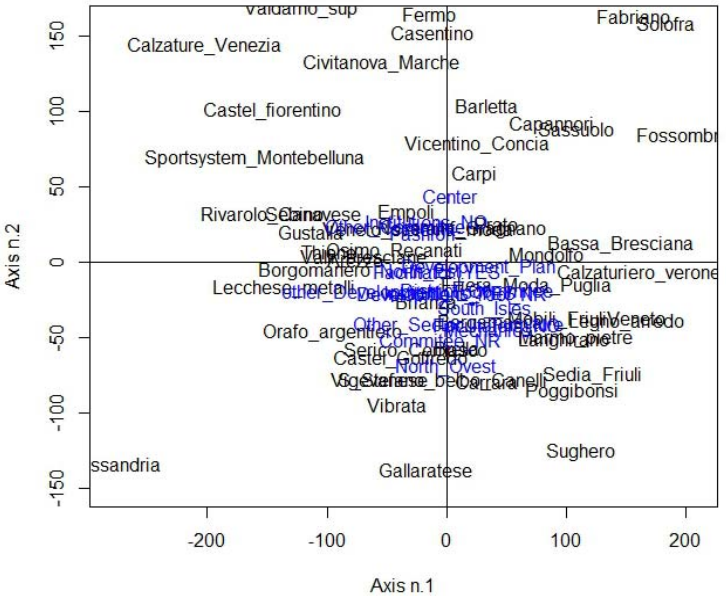

Figure 4.21 – Factorial Plan with supplementary variables considering only small-sized firms in Symbolic IDs (72.72%)

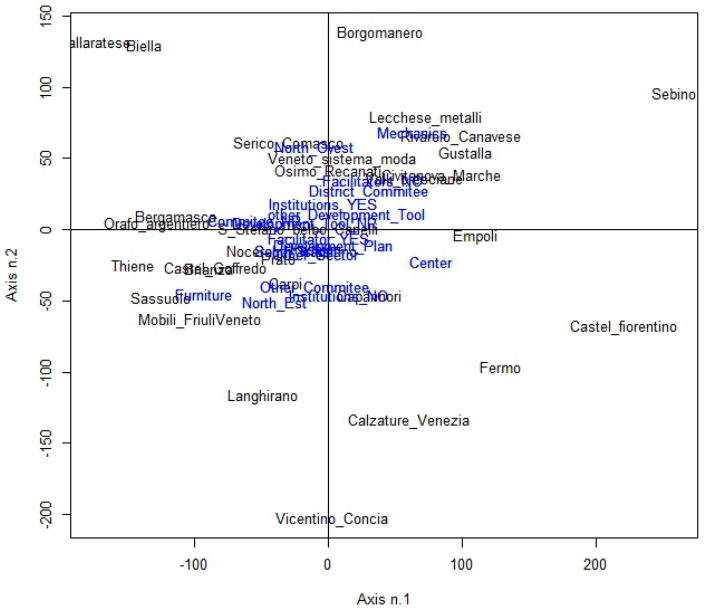

Figure 4.22 – Factorial Plan with supplementary variables considering only medium-sized firms in Symbolic IDs (87.62%)

4.8 Concluding Remarks

The case studies carried out on real data have underlined the usefulness of considering second-level units when dealing with complex data structures. The district-level analysis of performance ratios have lead towards interesting interpretation of the results.

The Symbolic Industrial District here defined and analyzed seems to be consistent, even after the cutting thresholds of the initial dataset. The analysis performed on different reduced size datasets underline the attitude of the method to consider the variability inherent to the data. This important feature is much more evident if we look at the graphical representations of Districts onto the several factorial plans showed in this chapter. Furthermore, the analyses show a relation between peculiar attributes of governance and economical-financial performance of the districts.

In the awareness that there are many other dimensions to be considered in this specific framework, here we have considered the

normative governance to explore this relation. Certainly, the latent aspects of the phenomenon cannot be taken into account when dealing with secondary data. Moreover, other interesting aspects, not only those related to the management system can be considered, such as import/export exchange.

The results show the high capability of the proposed method in extracting knowledge from a complex phenomenon, not univocally defined and measured in the current literature.

In particular, it may help to solve problems related to the quantitative analysis of concepts organized in structured data, as is the case of the Italian Industrial Districts. Moreover, it can be used as a benchmark for monitoring the dynamics of complex data structures over the years.

CONCLUSIONS

In this work we start considering that the real word is made of several concepts whose pure definition depends on the different perspective of analysis chosen by researchers. This means that there is an obvious difficulty related to the conceptualization, definition and individualization of concepts. So any theoretical construct, conveniently operationalized, can be considered as a new level of analysis.

As in the case of the Italian Industrial Districts presented in this work, adopting different points of view, we have highlighted several aspects that contribute to the definition of such complex socio-economic realities. Industrial Districts are, naturally, complex structures. In the Italian context, several definitions of this topic have lead to the production of different maps of the districts, identified by different criteria. The difficulties related to the realization of a unique map of Italian districts that include all their fundamental aspects is still a very important subject of the scientific debate.

The high complexity of the district structure and the lack of an unambiguous definition of IDs have driven our interest towards complex statistical methods. At this purpose, a new operationalization concept of Industrial District has been proposed in the framework of Symbolic Data Analysis.

Besides the review of the developments of SDA from the beginnings until nowadays, the added value of this work is to consider a well-known theoretical definition of a topic, in this specific case the Italian Industrial District, and give it a new definition that leads to a suitable quantitative treatment. This means that, we are no longer dealing with Statistics of atomic data, but rather with Statistics of knowledge.

The Italian Industrial District have been defined as a concept, with its intent and extent, as a typical complex object in the symbolic data framework.

As emphasized in this work, several analytical properties make the analysis of symbolic data appealing for researchers. The definition of symbolic data of Industrial Districts allowed to move from the classical data framework to the symbolic one, considering, at the same time, all the original data.

Many advantages arise when performing SDA on real data. In particular, these methods allow to extract knowledge from huge datasets by a process of reduction of both units and variables. At the same time, the output of symbolic explorative analysis, together with the visualization of Symbolic Data Object in a reduced space, lead to non-trivial interpretations of the results and facilitate the recognition of new patterns and regularities in the data.

This work gives deep implications for future studies on the operationalization process and the statistical analysis of a theoretical construct, in general, and specifically of the Italian Industrial District.

Afterwards, some suggestions for future researches are proposed. First of all, since inter-firm relationships are considered as one of the main features of Industrial Districts: it will be interesting to investigate the informal relations among firms nested into an Industrial District.

A survey on first-level units will foster the appearance of latent dimensions of trustee and knowledge exchange among district's firms. Following this purpose, different relational aspects may be investigated by means of Social Network Analysis tools, creating a multiple structure of relations, both among firms and between these and the local institutions that most contribute in the management of the Industrial District.

Indeed, we suppose that these dimensions may assume a strategic role in the definition of district governance. So doing, the district-level

analysis, presented in this work, will turn out towards higher-up descriptions of the investigated concept.

Moreover, other methods of analysis could be performed in order to consider also the multilevel network structure of the Industrial District as a whole.

APPENDIX A

Table A.1 – Industrial Districts according to Mediobanca-Unioncamere

Industrial District	**Manufacturing Sector**
Alessandria	Jewelry
Arezzo	Jewelry
Barletta	Leather_Footwear
BassaBresciana	Textile_Clothig
Bergamasca	Textile_Clothig
Biella	Textile_Clothig
Borgomanero	Mechanics
Borgosesia	Textile_Clothig
Brianza	Wood_Furniture
CalzatureVenezia	Leather_Footwear
CalzaturieroVeronese	Leather_Footwear
Capannori	Paper Machine
Carpi	Textile_Clothig
Carrara	Pottery
Casarano	Leather_Footwear
Casentino	Textile_Clothig
CastelFiorentino	Leather_Footwear
CastelGoffredo	Textile_Clothig
CivitaCastellana	Pottery
Civitanovamarche	Leather_Footwear
Empoli	Textile_Clothig

Fabriano	Mechanics
Fermo	Leather_Footwear
FilieraModaPuglia	Textile_Clothig
Fossombrone	Wood_Furniture
Gallaratese	Textile_Clothig
GrumoNevano	Textile_Clothig
Gustalla	Mechanics
Langhirano	Food
LeccheseMetalli	Mechanics
LegnoArredo	Wood_Furniture
Maiella	Textile_Clothig
MarimoPietre	Pottery
Mirandola	Bio-Medical
MobiliFriuliveneto	Wood_Furniture
Mondolfo	Textile_Clothig
NoceraInfGragnano	Food
Omegna	Household
OrafoArgentiero	Jewelry
OsimoRecanati	Mechanics
Poggibonsi	Wood_Furniture
Prato	Textile_Clothig
RivaroloCanavese	Mechanics
SantoStefanoBelbo	Food
Sassuolo	Pottery
Sebino	Plastic
Sedia	Wood_Furniture
SericoComasco	Textile_Clothig
Solofra	Leather_Footwear
SportSystemMontebelluna	Leather_Footwear
Sughero	Cork
Thiene	Textile_Clothig
ValdarnoSup	Leather_Footwear
ValliBresciane	Mechanics
ValSeriana	Textile_Clothig
VenetoSistemaModa	Textile_Clothig
Vibrata	Textile_Clothig
VicentinoConcia	Leather_Footwear
Vigevanese	Leather_Footwear

BIBLIOGRAPHY

Adanson, M., (1757). *Histoire Naturelle du Sénégal – Coquillages*, Bauche, Paris

Afonso, F., Haddad, R., Toque, C., Eliezer E.-S., Diday, E., (2013), User Manual of the SYR Software, Syrokko internal publication, 70pp., internal document

Alberti, F., (2010). The concept of Industrial District: main contributions, *International Newtork for SMEs*.

Albino V., Garavelli A. C., and Schiuma, G., (1999). Knowledge transfer and inter-firm relationships in Industrial Districts: the role of the leader firms, *Technovation*, n°19, p. 53-63.

Allen, T.J., (1977). *Managing the flows of technology: technology transfer and the dissemination of technological information within the R&D organization,* MIT Press, Cambridge, MA.

Asheim, B., and Isaksen, A., (2002). Regional innovation systems: The integration of local 'sticky' and global 'ubiquitous' knowledge, *The Journal of Technology Transfer*, no. 1, January: 77–86.

Asheim, B.T., (1996). Industrial Districts as learning regions: A condition for prosperity?, *European Planning Studies*, vol.4, nº4, p. 379-400.

Baum, J. y Oliver, C. (1992). Institutional embeddedness and dynamics of organizational populations, *American Sociological Review*, 57:540-559.

Becattini, G., (1979). Dal 'settore' industriale al 'distretto' industriale. Alcune considerazioni sull'unità d'indagine dell'economia industriale, *Rivista di economia e Politica Industriale,* 5:7-21.

Becattini, G., (1990). The marshallian Industrial District as a socio-economic notion in Pyke F., Becattini G., Sengenberger W. (eds.), *Industrial Districts and Inter-Firm Cooperation in Italy*, ILO, Geneva, p. 37-51.

Becattini, G., (1991). The Industrial District as a creative milieu, in Benko, G., M. Dunford (eds.), *Industrial change and regional development*, Belhaven Press, London, p. 102-114.

Becattini, G., (1998). L'industrializzazione leggera del Mezzogiorno, in G. Becattini, *Distretti industriali e Made in Italy*, Bollati Borighieri, Torino, p. 146-187.

Becattini, G., (2000). *Il bruco e la farfalla - Prato: una storia esemplare dell'Italia dei distretti*, Firenze, Le Monnier.

Becattini, G., (2004). *Industrial Districts: A new approach to industrial change*. Cheltenham, UK: Edward Elgar.

Becattini, G., and Rullani, E., (1996). Local systems and global connections: the role of knowledge" in Cossentino F., Pyke F. and Sengenberger W., (Eds.), *Local and regional response to global pressure: the case of Italy and its Industrial Districts*, IILS, Geneva, p.159-174.

Bell, G.G., (2005). Clusters, networks, and firm innovativeness, *Strategic Management Journal* 26, no. 3, p.287–295.

Bellandi, M., (2001). Local development and embedded large firms, *Entrepreneurship & Regional Development,* 13 (3), 189-210.

Bellandi,M., Sforzi, F., (2001). La molteplicità dei sentieri di sviluppo locale, in G. Becattini et al., *Il caleidoscopio dello sviluppo locale*, Rosenberg & Sellier, Torino.

Belussi, F., (2001). *Local Production Systems/Industrial Districts as Hyper-Networks: A Post-Marshallian Interpretative Frame*, EAEPE (European Association for Evolutionary Political Economy) Conference: Comparing Economic Institutions, November, 8-11, 2001, Siena, Italy.

Belussi, F., and Pilotti, L. (2003). Knowledge creation and codification in Italian Industrial Districts. In Belussi, F., Gottardi, G., and Rullani, E., (eds.), *Technological evolution of Industrial Districts*, , Dordrecht: Kluwer, p. 139–172.

Benzécri, J.P., (1973). *L'Analyse des données*, Vol. 1, 2, Dunod, Paris

Bertrand, P., and Goupil, F., (1999). Descriptive statistics for symbolic data, In: *Symbolic official data analysis*, Springer, p. 103-124.

Billard, L., Diday, E., (2003). From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis, *Journal of the American Statistical Association*, 98(462):470-487.

Billard, L., Diday, E., (2004). *Symbolic Data Analysis: Definitions and Examples*, Université Paris-Dauphine, working paper

Billard, L., Diday, E., (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Wiley, Chichester.

Boari, C., and Lipparini, A. (1999). Networks within Industrial Districts: Organising knowledge creation and transfer by means of moderate hierarchies. *Journal of Management and Governance* 3: 339–360.

Bock, H.H., (2000). Symbolic Data, in Bock, H.H., Diday, E., (eds.), *Analysis of Symbolic Data: Explanatory methods for Extracting Statistical Information From Complex Data*, Studies in Classification, Data Analysis and Knowledge Organization, Springer-Verlag, Berlin, p. 39-49.

Bock, H.H., Diday, E., (2000). Symbolic Objects, in Bock, H.H., Diday, E., (eds.), *Analysis of Symbolic Data: Explanatory methods for Extracting Statistical Information From Complex Data*, Studies in Classification, Data Analysis and Knowledge Organization, Springer-Verlag, Berlin, p.54-75.

Bock, H.H., Diday, E., eds., (2000) *Analysis of Symbolic Data: Explanatory methods for Extracting Statistical Information From Complex Data*, *Studies in Classification, Data Analysis and Knowledge Organization,* Springer-Verlag, Berlin.

Borgatti, S.P., Everett, M.G., (1999). Models of core/periphery structures, *Social Networks*, 21, p. 375-395.

Boschma, R.A., (2005). Proximity and innovation: a critical assessment, *Regional Studies,* 39(1), 1-14.

Boschma, R.A., and Lambooy J.G., (2002). Knowledge, market structure, and economic coordination: dynamics of Industrial Districts, *Growth and Change* 33 (3), 291-311

Boschma, R.A., Ter Wal, A.L.J., (2007). Knowledge Networks and Innovative Performance in an Industrial District: the Case of a Footwear District in the South of Italy. *Industry & Innovation*, 14(2):177-199.

Brahim, B., and Makosso-Kallyth, S., (2014). GraphPCA: Graphical tools of histogram PCA. R package version 1.0. http://CRAN.R-project.org /package=GraphPCA.

Breschi, S., Lissoni, F., (2001a). Knowledge spillovers and Local innovation systems: A critical Survey, *Industrial and Corporate Change*, 10(4):975-1005.

Breschi, S., Lissoni, F., (2001b). Localised Knowledge Spillovers vs. Innovative Milieux: Knowledge 'Taciteness' Reconsidered, *Papers in Regional Science*, 90:255-273.

Brito, P., (1994). Use of Pyramids in symbolic data analysis, In: Diday, E., Lechevallier, Y., Schader, M. (eds), IFCS-93, 378-386.

Brito, P., (1995). Symbolic objects: order structure and pyramidal clustering, Annals of Opertions Research, 55 (2), 277-297.

Brito, P., (2000). Hyerarchical and Pyramidal Clustering with complete Symbolic Objects, in Bock, H.H., Diday, E., eds., (2000) *Analysis of Symbolic Data: Explanatory methods for Extracting Statistical Information From Complex Data*, Studies in Classification, Data Analysis and Knowledge Organization, Springer-Verlag, Berlin, pp. 312-323.

Brito, P., De Carvalho, F.A.T., (1999). Symbolic Clustering in the presence of hierarchical rules , In: Studies and Research Proceeding of the Conference

Knowledge Extraction and Symbolic Data Analysis, Luxemburg, Office for Official Publications of the European Communities, 119-128.

Brusco, S., (1989). *Piccole imprese e distretti industriali. Una raccolta disaggi.* Torino: Rosenberg & Sellier.

Brusco, S., (1990). The idea of the Industrial District: its genesis. In Industrial Districts and inter-firm co-operation in Italy, Geneva: *International Institute for Labour Studies*, pp. 10-19.

Brusco, S., e Sabel, C., (1981). Artisan production and economic growth, in F. Wilkinson (a cura di), *The dynamics of labour market segmentation*, London, Academic Press, pp. 99-113.

Burt, R. S., (1992). *Structural holes: the social structure of competition*, Cambridge, Ma, Harvard University Press.

Capello, R., (1999). Spatial Transfer of Knowledge in High Technology Milieux: Learning Versus Collective Learning Processes, *Regional Studies*, 33, 353-65.

Capello, R., Faggian, A., (2005). Collective Learning and Relational Capital in Local Innovation Processes, *Regional Studies*, 39(1), 75-87.

Cazes, P., Chouakria, A., Diday, E., Schektman, Y., (1997). Extension de l'analyse en composante principales ´a des donn`ees de type intervalle, Rev. Statistique Appliquée, Vol. XLV Num. 3 pag. 5-24, Francia

Chavent, M., (1998), A monothetic clustering algorithm, *Pattern Ricognition Letters*, 19, 989-996.

Chavent, M., Lechevallier, Y., (2002). Dynamical clustering algorithm of interval data: optimization of an adequacy criterion based on Hausdorff distance. In Sokolowsky and Bock (eds), *Classification, Clustering and Data Analysis*, Springer, 53-59

Chouakria, A., Cazes, P., Diday, E., (2000). Symbolic Principal Component Analysis, in Bock, H.H., Diday, E., eds., *Analysis of Symbolic Data: Explanatory methods for Extracting Statistical Information From Complex Data*, Studies in Classification, Data Analysis and Knowledge Organization, Springer-Verlag, Berlin, p. 200-211.

Coase, R. H., (1992). The Institutional Structure of Production, *The American Economic Review*, Vol. 82, No. 4, p. 713-719.

Coase, R. H., (1998). The New Institutional Economics, *The American Economic Review*, Vol. 88, No. 2, Papers and Proceedings of the Hundred and Tenth Annual Meeting of the American Economic Association, p. 72-74.

Coleman, J., (1988). Social capital in the creation of human capital, *American Journal of Sociology*, 94, 95-120.

Cresta, A., (2008). *Il ruolo della governance nei distretti industriali: un'ipotesi di ricerca e classificazione*. Franco Angeli, Milano.

De Carvalho, F. A. T., Verde, R., Lechevallier, Y., (1999). A dynamical clustering of symbolic objects based on a context dependent proximity measure, in

Proceedings of the IX International Symposium on Applied Stochastic Models and Data analysis, Lisboa.

De Carvalho, F.A.T., (1994). Proximity coefficients between Boolean symbolic objects, In: Diday, E., Lechevallier, Y., Schader, M. et al. (eds.), IFCS-93, 387-394.

De Carvalho, F.A.T., (1995). Histohgrams in symbolic data analysis, *Annals of Operations Research*, J. C. Baltzer A.G. Science Publishers, 55, 299-322.

De Carvalho, F.A.T., Souza, R.M., Chavent, M., Lechevallier, Y., (2006). Adaptive Hausdorff distances and dynamic clustering of symbolic interval data, *Pattern Recognition Letters*, 27(3), 167-179.

Dei Ottati, G., (1994a). Trust, interlinking transactions and credit in the Industrial District, *Cambridge Journal of Economics* 18, no. 6: 529–546.

Dei Ottati, G., (1994b). Cooperation and competition in the Industrial District as an organization model, *European Planning Studies,* 4:463-483.

Dei Ottati, G., (1995). *Tra mercato e comunità: aspetti concettuali e ricerche empiriche sul distretto industriale*, Franco Angeli, Milano.

Desrochers, P., (2001). Geographical Proximity and the Transmission of Tacit Knowledge, *The Review of Austrian Economics*, 14:1, 25–46

Diday, E., (1987). The symbolic approach in clustering and related methods of Data Analysis: the basic choices, In *Classification and Related Methods of Data Analysis*, Proceedings of IFCS '87, H.H. Bock, ed., Aachen, July 1987, North Holland, Amsterdam, pp. 673-684.

Diday, E., (1989). Introduction à l'Approche Symbolique en Analyse des Données, *RAIRO (Revue d'Automatique, d'Informatique et de Recherche Opérationnelle),* 23(2):193-236.

Diday, E., (1990). Knowledge Representation and Symbolic Data Analysis, in Schader, M., Gaul, W., eds., *Knowledge Data and Computer-Assisted Decisions*, Springer Verlag, Berlin, pp. 17-54.

Diday, E., (1991). Des Objects de l'Analyse des données à ceux de l'Analyse des Connaissances, In Kodratoff, Y. and Diday, E., eds., *Induction Symbolique et numerique*, Cepáudes, Toulouse, pp. 9-76.

Diday, E., (1995) Probabilist, possibilist and belief objects for knowledge analysis, *Annals of Operations Research*, 55:227-276.

Diday, E., (1998). *L'Analyse des Données Symboliques: un cadre théorique et des outils*, Cahiers du CERMADE, Paris, n. 9821.

Diday, E., (2002). An introduction to Symbolic Data Analysis and the Sodas software, *Journal of Symbolic Data Analysis*, Vol. 1, n° 1. International Electronic Journal

Diday, E., (2008). The State of the Art in Symbolic Data Analysis: Overview and Future, in Diday, E., Noirhomme-Fraiture, M. eds., (2008) *Symbolic Data Analysis and the Sodas Software*, Wiley, Chichester, p. 3-41.

Diday, E., Esposito, F., (2003). An introduction to symbolic data analysis and the SODAS software, *Intelligent Data Analysis*, 7:583–602

Diday, E., Noirhomme-Fraiture, M., eds., (2008). *Symbolic Data Analysis and the Sodas Software*, Wiley, Chichester.

Duarte Silva, A.P., Brito, P., (2006). Linear Discriminant Analysis for interval data, Comput Stat, 21 (2), 231-250.

Duarte Silva, P., Brito, P., (2015). MAINT.Data: Model and Analyse Interval Data. R package version 0.5.1. http://CRAN.R-project.org/package=MAINT.Data.

Dudek, A., Pelka, M., and Wilk, J., (2015). symbolicDA: Analysis of Symbolic Data. R package version 0.4-2. http://CRAN.R-project.org/package=symbolicDA.

Esposito, F., Malerba, D., Appice, A., (2008). Dissimilarities and matching, in Diday, E., Noirhomme-Fraiture, M. Symbolic Data Analysis and the Sodas software, Chinchester, Wiley, p. 123-148.

Esposito, F., Malerba, D., Tamma, V., (2000). Dissimilarity measures for Symbolic Objects, in Bock, H.H., Diday, E., eds., *Analysis of Symbolic Data: Explanatory methods for Extracting Statistical Information From Complex Data*, Studies in Classification, Data Analysis and Knowledge Organization, Springer-Verlag, Berlin, pp. 166-186.

Gioia, F., and Lauro, C. N., (2005). Basic statistical methods for interval data, *Statistica applicata*, 17.1, p. 75-104.

Gioia, F., and Lauro, C. N., (2006). Principal component analysis on interval data, *Computational Statistics,* 21.2: 343-363.

Giordano G., Brito P., (2014). Social Networks as Symbolic Data, in Vicari D., Okada A., Ragozini G., Weihs C., eds., *Analysis and Modeling of Complex Data in Behavioral and Social Science*, Springer: Heidelberg, pp. 133-142.

Gowda, K.C., Diday, E., (1991), Symbolic clustering using a new dissimilarity measure, *Pattern Recognitions*, 24, (6), 567-578.

Granovetter, M., (1973). The strength of weak ties, *American Journal of Sociology*, 78(6), 1360-1380.

Granovetter, M., (1985).Economic action and social structure: the problem of embeddedness, *American Journal of Sociology*, 91(3), 481-510.

Granovetter, M., (2005). The Impact of Social Structure on Economic Outcomes, *Journal of Economic Perspectives,* 19 (1), 33-50.

Grassi, M., and Pagani, R., (1999). Sistemi produttivi localizzati e imprese leader, *Economia e politica industriale* 103, 241-72.

Hotelling, H., (1933). Analysis of a Complex of Statistical Variables Into Principal Components, *Journal of Educational Psychology*, volume 24, pages 417-441 and 498-520.

Ichino, M., (2008). Symbolic PCA for histogram-valued data, In Proceedings of IASC2008, Joint Meeting of 4th World Conference of the IASC and 6th Conference of the Asian regional Section of the IASC on Computational statistics & Data Analysis, Yokohama, Japan.

Ichino, M., Yauchi, H., (1994). Generalized Minkowski metrics for mixed festures type data analysis, *IEEE Transactions on Systems, Man and Cybernatics*, 24 (4), 698-708.

Irpino, A., (2015). HistDAWass: Histogram-Valued Data Analysis. R package version 0.1.3, http://CRAN.R-project.org/package =HistDAWass.

Irpino, A., and Verde, R., (2006). *A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data*. Data Science and Classification (Eds. Batanjeli, Bock, Ferligoj, Ziberna), Springer, Berlin, pp. 185-192.

Irpino, A., and Verde, R., (2008). Dynamic clustering of interval data using a Wasserstein-based distance., *Pattern recognition letters,* 29.11: 1648-1658.

Irpino, A., Lauro, C.N, Verde, R., (2003). Visualizing symbolic data by closed shapes, *Between Data Science and Applied Data Analysis,* Springer Berlin Heidelberg, 244-251.

Irpino, A., Verde, R., Lechevallier, Y., (2006). Dynamic clustering of histograms using Wasserstein metric, In: COMPSTAT. 2006. p. 869-876.

Istat, (1997). *I sistemi locali del lavoro 1991*, Argomenti n. 10, Roma

Jussieu, A. L., (1974). *Taxonomuy. Coup d'oeil sur l'histoire et les principles des classifications botanique*, Dictionaire d'Histoire Universelle.

Langlois, R. N., and Robertson, P. L., (2002). *Firms, markets and economic change: a dynamic theory of business institutions*. Routledge.

Lauro, N. C., Verde, R., Irpino, A, (2008). Principal component analysis of symbolic data described by intervals, in: Diday, E., and Noirhomme-Fraiture, M., eds., *Symbolic Data Analysis and the SODAS Software*, p. 279-312.

Lauro, N.C, Palumbo, F., Iodice D'Enza, A., (2003). New graphical symbolic objects representations in parallel coordinates, in: Between Data Science and Applied Data Analysis, Schader M. et al. eds., GfKl, Springer Verlag, Heidelberg, *Studies in Classification, Data Analysis, and Knowledge Organization*, p. 288-295.

Lauro, N.C., Palumbo F., (2000). Principal component analysis of interval data: A symbolic data analysis approach, *Computational Statistics*, 15, 1, 73-87.

Lauro, N.C., Palumbo, F., Iodice D'Enza, A., (2004). Visualizzazione ed ordinamento di oggetti simbolici, in Lauro, C.N., D'Avino, C., eds., *Data Mining e Analisi Simbolica*, FrancoAngeli, Milano, p.125-154.

Lazerson M.H., Lorenzoni G., (1999). The firms feed Industrial Districts: A return to the Italian source, *Industrial and Corporate Change*, 8, 235-266.

Lazzaretti L., Capone F., (2009). *Industrial District effect and innovation in the Tuscany shipbuilding industry*, IERMB Working Papers in Economics, n° 09.03.

Lechevallier, Y., Verde, R., De Carvalho, F.A.T., (2006). Symbolic clustering of large datasets, Data Science and Classification, Springer Berlin Heidelberg, 193-201.

Le-Rademacher, J., and Billard, L., (2013). Principal component histograms from interval-valued observations, *Computational Statistics* 28.5, 2117-2138.

Lima Neto, E.A. with contribution from Claudio A. Vasconcelos (2012). iRegression: Regression methods for interval-valued variables. R package version 1.2. http://CRAN.R-project.org/package =iRegression.

Lipparini, A., Lorenzoni, G. (1994). Strategic sourcing and organizational boundaries adjustment: A process-based perspective. Paper presented at the workshop on "The changing boundaries of the firm", European Management and Organisations in Transition (EMOT), European Science Foundation, Como, October 1994.

Lissoni F., (2001). Knowledge codification and the geography of innovation: the case of Brescia mechanical cluster, *Research Policy* 30, 1479-1500.

Lorenz, E., (1988). Neither friends nor strangers: Informal networks of subcontracting in French industry. in *Trust. Making and breaking cooperative relations*, ed. Diego Gambetta, 194–210. Oxford: Basil Blackwell.

Lorenz, E., (1999). Trust, contract and economic cooperation. *Cambridge Journal of Economics* 23: 301–315.

Makosso-Kallyth, S., Diday. E., (2012). Adaptation of interval PCA to symbolic histogram variables, *Advances in Data Analysis and Classification* 6.2: 147-159.

Marshall, A., (1920). *Principles of Economics*, *(Revised ed.,* first ed. 1890), Macmillan and Co., London.

Mártinez-Cháfer, L., Capò-Vicedo, J., Molina-Morales, F.X., (2011). The Role of Local Institutions in the Transmission of Information and Knowledge in Industrial Districts. A Social Networks Analysis, *European Regional Science Association.*

Maskell, P., (2001). Towards a knowledge–based theory of the geographical cluster. *Industrial and Corporate Change* 10, no. 4: 921–943.

Mediobanca e Unioncamere (2013), Le medie imprese industriali italiane (2002-2011)

Moore, R.E., (1966). Interval Analysis, Prentice Hall, Englewood Cliffs, NJ.

Morrison, A., (2008). Gatekeepers of knowledge within Industrial Districts: who they are how they interact, *Regional Studies*, vol.42, nº6, p. 817-835.

Morrison, A., Rabellotti, R., (2009), Knowledge and Information Networks in an Italian Wine Cluster, *European Planning Studies*, vol.17, nº7, p. 983-1006.

Muscio, A., (2006). Patterns of innovation in Industrial Districts: An empirical analysis. *Industry and Innovation* 13, no. 3: 291–312.

Neto, E.A.L, and Carvalho, F.A.T, (2010). Constrained linear regression models for symbolic interval-valued variables, *Computational Statistics & Data Analysis* 54.2, 333-347.

Neto, E.A.L, and Carvalho, F.A.T., (2008). Centre and Range method for fitting a linear regression model to symbolic interval data, *Computational Statistics & Data Analysis* 52.3, 1500-1515.

Noirhomme-Fraiture M, Brito P., (2011). Far beyond the Classical Data Models: Symbolic Data Analysis, *Statistical Analysis and Data Mining*, 4(2):157–170.

Noirhomme-Fraiture, M., Rouard, M., (1997). Zoom Star: a solution to complex statistical objects representation, In: St. Howard, J. Hammond, G. Lindgaard (eds.): proc. *INTERACT 97*, Sydney.

Noirhomme-Fraiture, M., Rouard, M., (2000). Visualizing and Editing Symbolic Objects, In. in Bock, H.H., Diday, E., eds., (2000) *Analysis of Symbolic Data: Explanatory methods for Extracting Statistical Information From Complex Data*, Studies in Classification, Data Analysis and Knowledge Organization, Springer-Verlag, Berlin, pp.125-138.

North, D. C., (1992). *Transaction costs, institutions and economic performance*, International Center for Economic Growth, San Francisco, California.

North, D. C., Weingast, B. R., (1989). Constitutions and Commitment: The Evolution of Institutional Governing Public Choice in Seventeenth-Century England, The Journal of Economic History, Vol. 49, No. 4, pp. 803-832.

Noteboom, B., (1999). *Inter-firm alliances. Analysis and Design,* London: Routledge.

Noteboom, B., (2004). *Interfirm collaboration, learning and networks: An integrated approach*. London: Routledge.

Noteboom, B., (2006). Innovation, learning and cluster dynamics. In *Cluster and regional development*, eds. B. Asheim, P. Cooke and R. Martin, pp. London: Routledge.

Pearson, K., (1901). On lines and planes of closest fit to systems of points in space, *Philosophical Magazine*, Series 6, vol. 2, no. 11, pp. 559-572.

Polaillon, G., (2000). Pyramidal Classification foe Interval Data Using Galois Lattice Reduction, in Bock, H.H., Diday, E., eds., (2000) *Analysis of Symbolic Data: Explanatory methods for Extracting Statistical Information From Complex Data*, Studies in Classification, Data Analysis and Knowledge Organization, Springer-Verlag, Berlin, pp. 324-341

Polanyi, M., (1958). *Personal knowledge: towards a post-critical philosophy*, University of Chicago Press, Chicago

Powell, W.W., (1990). "Neither market nor hierarchy: Networks forms of organisation. In Barry Staw and L.L. Cummings, (eds.), *Research in Organisational Behaviour*, Greenwich, CT: JAI Press, 295-336.

Provan, K.G., Kenis P., (2007). *Modes of Network Governance: Structure, Management, and Effectiveness*. Oxford University Press, Oxford.

Pyke, F., Becattini, G., Sengenberger, W., Eds., (1990). *Industrial Districts and Inter-firm Co-operation in Italy*. Geneva: International Institute for Labour Studies.

Pyke, F., Sengenberger, W., (1992). Industrial Districts and local economic regeneration. *Geneva:International Institute for Labour Studies*.

Queiroz Filho, R. J. A., Fagundes, R. A. A., (2012). ISDA.R: interval symbolic data analysis for R. R package version 1.0. http://CRAN.R-project.org/package=ISDA.R.

Rabellotti, R., (2004). How globalisation affects Italian Industrial Districts: The case of Brenta, in H. Schmitz, (ed.), *Local Enterprises in the Global Economy: Issues of Governance and Upgrading*, Cheltenham: Edward Elgar.

Robertson, P. L., and Langlois, R. N., (1995). Innovation, networks, and vertical integration. *Research policy*, *24*(4), 543-562.

Rodriguez, O, Diday, E., Winsberg, S., (2000). Generalization of the principal component analysis, In Proceedings 4th European Conference on Principles and Practice of Knowledge Discovery in Databases; Workshop on Symbolic Data Analysis, Lyion, 14.

Rodriguez, O., R. with contributions from Olger Calderon and Roberto Zuniga (2014). RSDA: RSDA - R to Symbolic Data Analysis. R package version 1.2. http://CRAN.R-project.org/package=RSDA.

Rosch, E., (1978). Principle of categorization, In. Rosch, E., Loyd, B. (eds.), *Cognition and categorization*, Erlbaum, Hillsdale, N.J., 27-48.

Sabel, Charles F., (1989). "Flexible Specialization and the Re-emergence of Regional Economies," in Hirst, P. and Zeitlin, J., eds., *Reversing Industrial Decline?, Industrial Structure and Policy in Britain and Her Competitors.* Oxford: Berg, pp. 17-70.

Saporta, G., (1990). *Probabilitiés, analyse des données et statistique*, Editions TECHNIP, Paris.

Sforzi, F., (1990). The Quantitative Importance of Marshallian Industrial Districts in the Italian Economy, in F. Pyke, G. Becattini, W. Sengenberger (eds.), Industrial Districts and Inter-firm Co-operation in Italy, Geneva, International Institute for Labour Studies, p. 75-107.

Sforzi, F., (1997). *I sistemi locali in Italia*. Roma: ISTAT.

Sforzi, F., Lorenzini, F., (2002). I distretti industriali, in *L'esperienza italiana dei distretti industriali*, IPI-Ministero delle Attività Produttive IPI, Roma

Signorini, L.F., (1994). Una verifica quantitativa dell'effetto distretto. *SviluppoLocale,* no. 1: p. 31–70.

Souza, R.M.C.R., and De Carvalho, F.A.T., (2004). Clustering of interval data based on city-block distances. *Pattern Recognition Letters*, 25 (3), 353–365.

Staber, U., (2001). The Structure of Networks in Industrial Districts, International Journal of Urban and Regional Research, 25, pp. 537-552.

Storper, M. and Harrison, B., (1991). Flexibility, hierarchy and regional development: the changing structure of industrial production systems and their forms of governance in the 1990s, *Research policy,* 20 (5), 407-422.

Storper, M., (1997). *The Regional World: Territorial development in a global economy*, New York: Guildford Press.

Tallman, S., Jenkins, M., Henry, N., Pinch, S., (2004). Knowledge, cluster and competitive advantage. *Academy of Management Review* 29: 258–271.

Tuckey, J.W., (1958). *Exploratory Data Analysis*, Addison Wesley, Reang, Mass.

Uzzi, B., (1997). Social Structure and Competition in Interfirm Networks: The Paradox of Embeddedness, *Administrative Science Quarterly*, 42, 35 – 67.

Ward, J.H., (1963). Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, vol. 58, 238-244.

Wasserman, S., Faust K., (1994). *Social Network Analysis. Methods and Applications*. Cambridge: Cambridge University Press.

Wille, R., (1982). Restructuring lattice theory: an approach based on hierarchies of concepts, In: Rival, I. (eds.), Ordered Sets, Reidel, Dordrecht-Boston, 445-470.

Williamson, O.E., (1985). *The economic institutions of capitalism*, New York: The Free Press.

Williamson, O.E., (1996). *The mechanisms of governance*, New York: Oxford University Press.

ACKNOWLEDGEMENT

Here I am, at the goal of this important university career. Who would have thought about this: I am a Ph.D. holder! So it is real, sometimes our dreams come true!

These years of crazy and desperate studies have been for me an amazing and wonderful experience. At the end of this thesis, I should like to extend my warmest thanks to the people who have supported me during this period of my life, promoting my professional and personal growth.

My sincere thanks to all people engaged, in different ways, in the realization of this work. First and foremost, I wish to thank my advisor, prof. Giuseppe Giordano for his trust, his valuable guidance, for the scholarly inputs and the consistent encouragement that I have received throughout the research work. Thanks for prompting me beyond my limits and encouraged me to do always better and more, to look ahead my limitations, to open my views to new and wonderful horizons. The experience to follow my doctoral programme under his guidance and to learn from his from his research knowledge has been a great opportunity for me. I would also like to thank Prof. Alessandra Amendola, coordinator of the Ph.D. in *Ingegneria ed Economia dell'Innovazione*, for her constant

readiness, for the encouragement and the patience showed during these years. I would like to express my sincere gratitude to Prof. Maria Rosaria D'Esposito for the significant contribution to my academic training, through the opportunity to participate in several academic activities, she has impeccably organized. I also thank Prof. Maria Prosperina Vitale for her valuable suggestions and concise comments, for conveying to me her commitment to work and for letting me to appreciate the extreme organizational rigor in research activities. I would like to express appreciation to all members of the Department of Economics and Statistics and my study-colleagues for their constant incentives and, above all, for having make enjoyable the years spent full-time in this department.

I especially thank my family, my parents, my in-laws and my grandparents. Words cannot express how grateful I am to all of them for their unconditional love and care on my behalf. Thanks to my brother Alessio that, even kilometers away, will always be for me an important point of reference. Thanks to my brothers and sisters in-laws for having supported and beared me, understanding my mood swings. Thanks to my little precious nephew Andrea, auntie love, for giving me the joy of a deep and indissoluble bond.

I also thank my friends (too many to list here but you know who you are!) for their support and friendship. Big thanks to my old friends, the ones on which I can always rely on, those who have never left me. Thanks to my new friends, known during these extraordinary years, with whom I have shared joys and pains. Thanks friends, for making me always smile, even when everything seemed to go wrong.

Thanks to those who always protect me. Those who, facing the balcony of heaven, guide me along my life journey. Especially, thanks to Giovanni, whose memory will always be alive in my thoughts and my heart.

Last but not least, my deep gratitude goes to my beloved husband Massimo, who has spent sleepless nights with me and always support me in the moments when nobody else might answer my queries. Thanks Massimo, for your love and you patience. Thanks because you believe in me, you encouraged me in my decisions and, without hesitation, you appreciate them. Thank you

for choosing me when, more than eleven years ago, I was just a little girl, and for helping me to become a better woman.

Deepest thanks to everybody, this thesis would not come to a successful completion, without the help I have received from all of you!

Ed eccomi giunta alla fine di questo importante percorso di studi. Chi l'avrebbe mai detto: io, dottore di ricerca! Allora è proprio vero che a volte i nostri sogni diventano realtà!

Questi anni di studio matto e disperato, nonostante tutto, sono stati per me un'incredibile e meravigliosa esperienza. A conclusione di questo lavoro di tesi desidero porre i miei più sentiti ringraziamenti alle persone che mi hanno accompagnato in questo periodo della mia vita contribuendo alla mia crescita professionale e personale.

Un sincero ringraziamento va alle persone che hanno partecipato, a vario titolo, alla realizzazione di questo lavoro. Un ringraziamento particolare al prof. Giuseppe Giordano per la fiducia dimostratami durante questi anni, per la sua guida, per gli input di ricerca e per il costante incoraggiamento. Grazie per avermi spinto oltre i miei limiti e per avermi guidata in questa avventura spronandomi a fare sempre meglio e di più, a guardare oltre i miei confini, ad aprire le mie prospettive verso nuovi e meravigliosi orizzonti. La sua guida e la sua esperienza di ricerca sono stati per me una grande opportunità in questo percorso di dottorato. Ringrazio la prof.ssa Alessandra Amendola, coordinatrice del Dottorato in Ingegneria ed Economia dell'Innovazione, per la sua costante disponibilità, per l'incoraggiamento e la pazienza dimostratami in questi anni. Ringrazio la Prof.ssa Maria Rosaria D'Esposito per aver contribuito alla mia formazione dandomi la possibilità di partecipare alle attività formative da lei organizzate in maniera impeccabile. Ringrazio la Prof.ssa Maria Prosperina Vitale per i suoi preziosi consigli, per avermi trasmesso la sua dedizione al lavoro e per avermi fatto apprezzare l'estremo rigore organizzativo nelle attività di ricerca. Ringrazio tutti i membri del Dipartimento di Scienze

Economiche e Statistiche e i miei colleghi per i continui stimoli e, soprattutto, per aver reso piacevoli gli anni trascorsi a tempo pieno in questo dipartimento.

Un doveroso ringraziamento alla mia famiglia, ai miei genitori, ai miei suoceri e ai miei nonni per il loro sincero affetto. Le parole non bastano per esprimere la mia gratitudine per il loro amore incondizionato. Grazie a mio fratello Alessio che, anche a chilometri di distanza, sarà sempre per me un importante punto di riferimento. Grazie ai miei cognati per avermi supportata e sopportata, comprendendo i miei sbalzi d'umore. Grazie al mio piccolo e amato nipotino Andrea, amore di zia, per avermi donato la gioia di un legame profondo e indissolubile.

Grazie ai miei amici (troppi per essere elencati qui) per l'amicizia e il supporto. Un grazie di cuore agli amici di sempre, quelli sui quali è sempre possibile fare affidamento, quelli che non mi hanno mai abbandonata. Grazie ai nuovi amici, quelli conosciuti grazie a questo straordinario percorso, con i quali ho condiviso gioie e dolori. Grazie amici, per avermi fatto sempre sorridere, anche quando tutto sembrava andar male.

Grazie a chi mi protegge sempre, a quanti, affacciati al balcone del cielo, mi guidano lungo il cammino della vita. Un grazie va in particolare a Giovanni, il cui ricordo è sempre vivo nei miei pensieri e nel mio cuore.

Il mio più grande ringraziamento va al mio amato marito Massimo, per le notti insonni trascorse insieme e per avermi sempre capita ed incoraggiata, anche quando nessuno riusciva a rispondere alle mie domande. Grazie Massimo per il tuo amore e per la tua pazienza. Grazie per aver sempre creduto in me, per aver appoggiato le mie scelte e per averle sempre accettate senza esitare. Grazie per avermi scelta quando più di undici anni fa ero solo una ragazzina e per aver contribuito a rendermi una donna migliore.

Grazie di cuore a tutti, questo lavoro non sarebbe stato possibile senza il vostro importante sostegno.