POLITECNICO DI MILANO
Facoltà di Ingegneria dei Sistemi
Corso di Laurea in Ingegneria Gestionale

DATA ANALYSIS FOR NEW PRODUCT DEVELOPMENT PROCESSES

Relatore: Prof. Ing. Marco GARETTI

Correlatore: Dott. Sergio TERZI

Ing. Monica Rossi

Tesi di laurea di:

MªEsperanza DE ARAMBURU MERA

matr.764551

Academic Year 2011-2012

POLITECNICO DI MILANO
Facoltà di Ingegneria dei Sistemi
Corso di Laurea in Ingegneria Gestionale

DATA ANALYSIS FOR NEW PRODUCT DEVELOPMENT PROCESSES

Relatore: Prof. Ing. Marco GARETTI

Correlatore: Dott. Sergio TERZI

Ing. Monica Rossi

Tesi di laurea di:

Mª Esperanza DE ARAMBURU MERA

matr.764551

Academic Year 2011-2012

For my parents

Agradecimientos

Deseo agradecer y expresar mi más profundo reconocimiento a todas las personas que, de cualquier modo, han seguido mi camino y me han permitido y estimulado sea en mis estudios como en la realización de este proyecto final de carrera.

En primer lugar a mis padres, a quienes dedico esta obra, sin los cuales no habría sido capaz de llegar a este punto. Pero no hago referencia solo al soporto económico, que seguramente ha sido indispensable. Hablo de aquella ayuda, a veces tácita y que ha llegado siempre desde el corazón: todas aquellas ocasiones en las que me han estimulado, en las que han sabido utilizar las mejores palabras de motivación para dar de mí lo mejor; pero sobre todo, aquellos momentos en los que realmente se han preocupado de hacerme entender la importancia de mis acciones.

Agradecimientos de corazón a mis hermanos, Tere y Enrique, que han sabido tolerarme, estando siempre a mi lado tanto en los buenos momentos como en otros difíciles y con los cuales he convivido bonitos y divertidos recuerdos.

Mis más sinceros agradecimientos al Profesor Sergio Terzi que me ha ayudado a la realización de este trabajo de fin de carrera, en particular por haberme seguido en el desarrollo del mismo, dando siempre los mejores consejos y confrontes, y que me han ayudado a emprender, en cada uno de los casos, el camino más apropiado.

También me gustaría agradecer con estas palabras a la Ing. Monica Rossi que, además de la ayuda que cada día me ha ofrecido, ha crecido entre nosotras una amistad. Por todo el tiempo que ha sabido dedicarme, por su apoyo y los mejores consejos que me ha regalado durante el estudio y desarrollo del proyecto.

Agradezco a María José, Alejandro y Claudia porque me han hecho vivir momentos inolvidables. Me han comprendido y ayudado en los momentos delicados, sabiendo regalarme una sonrisa cuando era necesaria.

También agradezco con estas palabras a Aline, Alessandra y Jose, porque siempre han estado a mi lado cuando lo necesitaba. Me han hecho ver la importancia de sonreír y han sabido llenar cada momento de felicidad.

Desde aquí muestro también mi gratitud hacia Cristina, Aurora, Javier, Enrique, Alexia, Marta, Lola, Alicia y Manolo, con los cuales he vivido divertidos y preciosos días de vacaciones; pero sobre todo les agradezco haberme hecho entender la importancia de la palabra amistad.

INDICE

INDICE FIGURE

INDICE TABLE

ACRONYM

SWOT: Strengths, Weakness, Opportunities and Threats

R&D: Research and Development

PLM: Product Lifecycle Management

CAD: Computer Aided Design

CAM: Computer Aided Manufacturing

PDM: Product Data Management

DNP: Development New Product

NACE: Statistical Classification of Economic Activities in the European Community

IT: Information Technology

GeCo: Gestione dei Processi Collaborativi di Progettazione

PTC: Parametric Technology Corporation

CLIMB: Chaos, Low, Intermediate, Mature and Best maturity level

EDA: Exploratory Data Analysis

CDA: Confirmatory Data Analysis

IBM: International Business Machines Corporation

GA: Global Average

SD: Standard Deviation

SQRT: Square root

ARLs: Average Run Lengths

UN Statistical Commission: United Nations Statistical Commission

UNSTAT: United Nations Statistics Division

MTO: Make To Order

MTS: Make To Stock

BTO: Built To Order

BTS: Built To Stock

ETO: Engineer To Order

ATO: Assembly To Order

Executive Summary

1. Introduction

The development of new products has become an essential way for businesses. The rate of evolution of both technology and society, and competition, has led to huge changes which affect the habits, styles, preferences and needs of consumers.

The beginning of new products is developed over a process where it can and must make fundamental decisions.

It is necessary to establish a process for new development products because for companies is a crucial aspect, for so many reasons. First we must take into account that the company has to offer new products to market constantly. These products present different tastes and preferences which change and are in a continuous evolution. So we must be aware of these needs and develop the company's offer. The design processes of new products are conducted simultaneously for many new ideas and for that we must considerer the big number of losers throughout the process. Therefore, if many new ideas of new products will not come to develop, we must establish systems which can generate new ideas continuously.

For that managers should to focus on creating new products to get a fast position in the market and to satisfy consumers' needs. In this way, they must take into consideration the strategic decisions, promotion, distribution, price and publicity, as well as the creation of the brand, label and packaging.

Of course next to this huge project we must study the marketing production and strengths opportunities, weakness and threats of new product, especially in turbulent markets.

On the other side it must take into account the rising cost of new products development. The different stages of development process require more recourses, as we progress with it. For that we have to take into consideration that new products development process must make possible identify unfeasible products or products that have a big probability of failure in early stages where it has not yet been used fully a lot of resources. The formalization of new product development process is necessary because the success of new ideas is also related strongly to attitudes and organizational involvement in management practice. This will avoid a lack of interest to top managements, lack of allocation of resources for the new product or even contrary situations of excess interest and involvement that in both case (one for default and the other for excess) they can lead us take to wrong decisions.

As we can see, there are two parallel aspects that are involved in this process: one of them involves product engineering; and the other one makes market analysis. Marketing managers consider new product development as the first step in product lifecycle management.

The generation of new products and services must be performed continuously. For that, companies must take into account the needs of new ideas and shall establish appropriate procedures. The main procedures to create new ideas can be divided: ones where costumers are the source and the other where experts are the sources.

Nevertheless ideas for new products can be obtained from basic research for example using a SWOT analysis (Strengths, Weaknesses, Opportunities & Threats), Market and consumer trends, company's R&D department, competitors, focus groups, employees, salespeople, corporate spies, trade shows, or Ethnographic discovery methods (searching for user patterns and habits) may also be used to get an insight into new product lines or product features. If the company has a specific department where it is developed new product, we must ask its participation and involvement in this process and in the next stages of product development.

In the analysis realized we must establish the release plan with relevant decisions on final configuration, communication (advertising, promoting, etc.), distribution (selection and design channels…) and pricing. This will configure the product launch plan which should be implemented on time and form. We should have an instrumentalist with appropriate control actions because we try to increase the efficiency and effectiveness of the company.

For our study we use a methodology that is a system of principles, practices, and procedures applied to a specific branch of knowledge.

A lot of different methodologies should be used in the PLM framework according to the different needs of the various phases of product lifecycle. Methodologies operate on product data (that are themselves distributed along the product lifecycle) to obtain specific results in the diverse stages of product life. Many methodologies are playing an important role in supporting the product lifecycle needs. Methodologies generally require a large amount of information and product data that can be provided by the PLM system, but they also require human knowledge to be correctly applied and used. From this point of view, as competences are not resident in a single actor, a good level of collaboration and coordination among diverse actors is necessary. PLM integrates people, data, processes and business systems and provides a product information backbone for companies and their extended enterprise. PLM systems help organizations in coping with the increasing complexity and engineering challenges of developing new products for the global competitive markets.

The core of PLM (product lifecycle management) is in the creations and central management of all product data and the technology used to access this information and knowledge. PLM as a discipline emerged from tools such as CAD, CAM and PDM, but can be viewed as the integration of these tools with methods, people and the processes through all stages of a product's life. It is not just about software technology but is also a business strategy.

The purpose of our work is defining the maturity of development and design process of different Italian companies. We must find correlations in the manner of made something and the level of performance obtained. We will be in a different stage of maturity in correlation with efforts made by the different levels of the company.

The aim of the assessment model for DNP process is to provide a 'picture' of enterprime that will be starting point for the AS-IS analysis.

A questionnaire and a Radar model were developed as tools for the assessment process. The questionnaire is composed by multiple choice close questions and by tables with a conveniently defined score able to represent the results of the enterprise directly on the radar, on the five maturity levels.

2.3 Training and competence

	Risposta	Punteggio
2.3.1 The design is heavily based on skills and competence of the actors involved (technicians, designers, Methodists, etc...). On average, how does the company support the training and consolidation of its skills?		
a. Any engineer / designer is personally responsible for developing and maintaining his/ her skills.		
b. The company urges the development of strong technical skills, with an approach to training on the job.	1	
c. The company promotes a multidisciplinary degree, supporting the knowledge management with formal programs (ex. training plans, rotating between project team, etc,…).		
d. Other (specify).		2
2.3.2 Inside the organization are there figures that support the training?		
a. There are not figures that support development skills, each technician/ designer is free to form individually.		
b. The technician/ designer is encouraged to develop his/her own skills from his/ her supervisor.		
c. There are methods one-to-one (the person is related individually with a more experienced person, as a tutor, coach, and mentor).	1	
d. Other (specify).		3
2.3.3 In the case that it exists a training courses how are the skills valued?		
a. Through a 'visual' evaluation it is verified improvements of individual behavior.	1	
b. It is applied test before the training and test at the end of the course.		
c. It is used KPI to assess the impact of training on business development strategies.		
d. Other (specify).		1

Figure I. Example of evaluation of the area Skills and competencies, composed by 3 questions

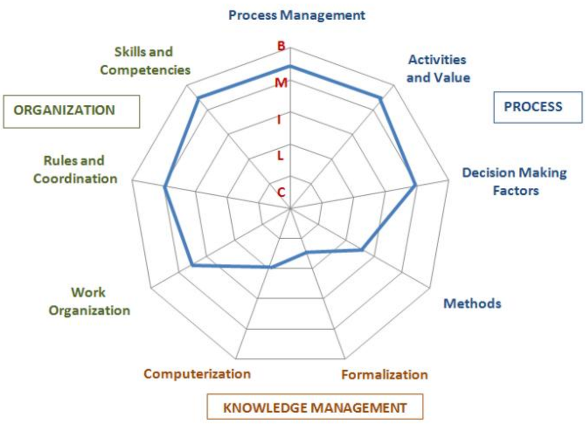

Figure II. Example of Radar Chart of a company

The maturity level of NDP process is based on five levels: Best Practice, Mature Practice, Intermediate Practice, Low Practice and Chaos. The Radar Chart is divided in three main areas: Organization, Process and Knowledge Management. Each of one is subdivided into different elements or sub-areas.

We must work on how to represent the data, how to benchmark companies and how to find best practices. With the radar chart (Figure II) we represent a 'picture' of the company development process, made of 9 areas of interest. For our study of data analysis we use statistic instruments, the average and standard deviation, to analyze the tendency of the companies in function of the different classification that we have created.

We separate the companies in four type ways. The first one has been created a classification in function of the number of employees. We see that we find three relevant groups: medium, big and macro Italian companies.

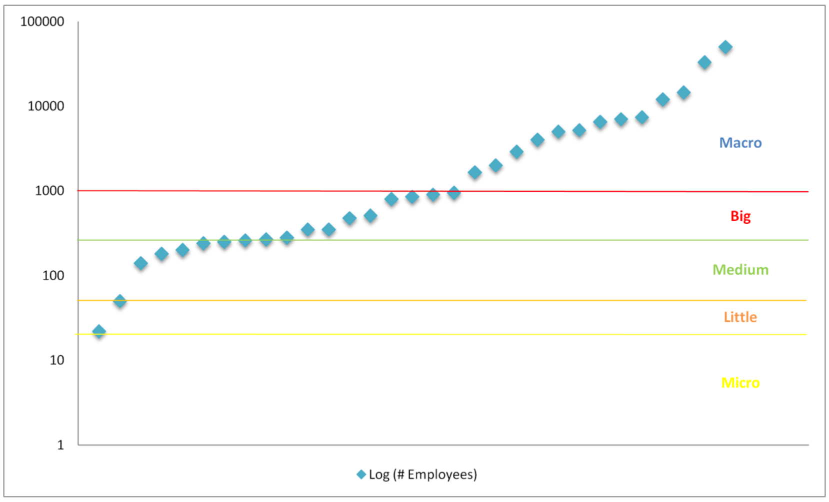

Figure III. Sample distribution of Log scale (y) and Log employees (x)

The next one is the classification in function of the different activities that diverse Italian companies have involved. We use NACE criteria to do it. NACE is the European standard classification of productive economic activities. In this way, our classification requires these codes:

2540	Manufacture of weapons and ammunition
2573	Manufacture of tools
2599	Manufacture of other fabricated metal products n.e.c.
2540	Manufacture of weapons and ammunition
2611	Manufacture of electronic components
2630	Manufacture of communication equipment.
2651	Manufacture of instruments and appliances for measuring, testing and navigation
2712	Manufacture of electricity distribution and control apparatus
2740	Manufacture of electric lighting equipment
2751	Manufacture of electric domestic appliances
2790	Manufacture of other electrical equipment
2811	Manufacture of engines and turbines, except aircraft, vehicle and cycle engines
2812	Manufacture of fluid power equipment
2829	Manufacture of other general-purpose machinery n.e.c.
2932	Manufacture of other parts and accessories for motor vehicles
3030	Manufacture of air and spacecraft and related machinery
3299	Other manufacturing n.e.c.
4299	Construction of other civil engineering projects n.e.c.
7410	Specialised design activities

Figure IV. Classification NACE of our Italian companies

The third type of classification is in function of the market. We create four type of way to satisfy customer demand:

- Make to Order
- Make to Stock
- Catalogue
- Shelf

And the last type to put the diverse Italian companies into groups is in function of the market position. We separate different companies in three types using a ratio:

$$\frac{Net\ Sales}{Average\ Number\ of\ Employees}$$

And depending of this value, we have:

- If Ratio €[0,20): company is inefficient
- If Ratio €[20,50): company efficient
- If Ratio €[50,infinity): company huge efficient

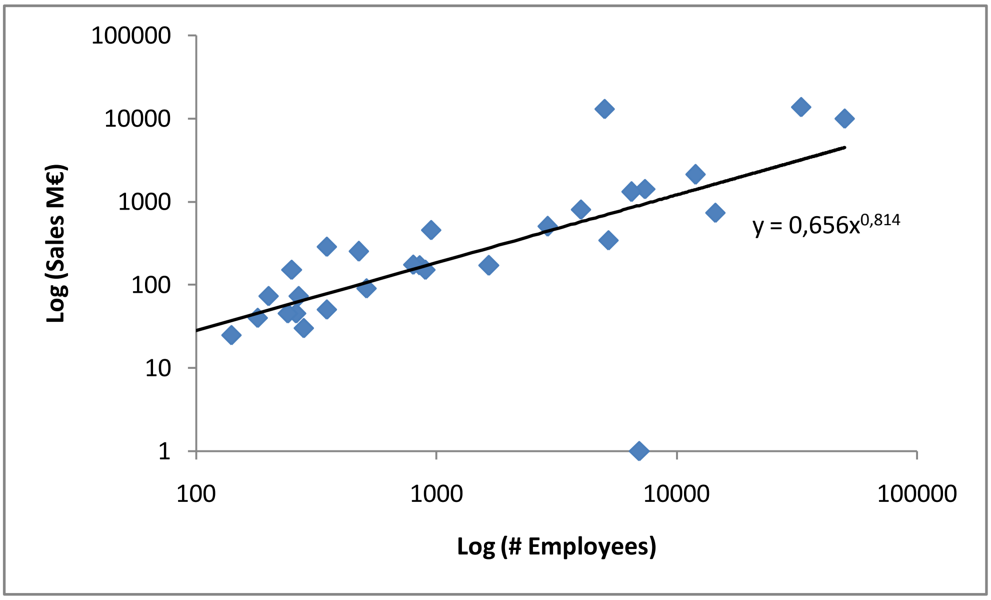

Figure V. Sample distribution Log Sales (y) with Log # Employees (x)

We use a rating process where at the end we see that Italian Companies have a behavior that is not according with the different classification that we have created. Maturity levels increase with the level of employees but we can see the same for the level of sales.

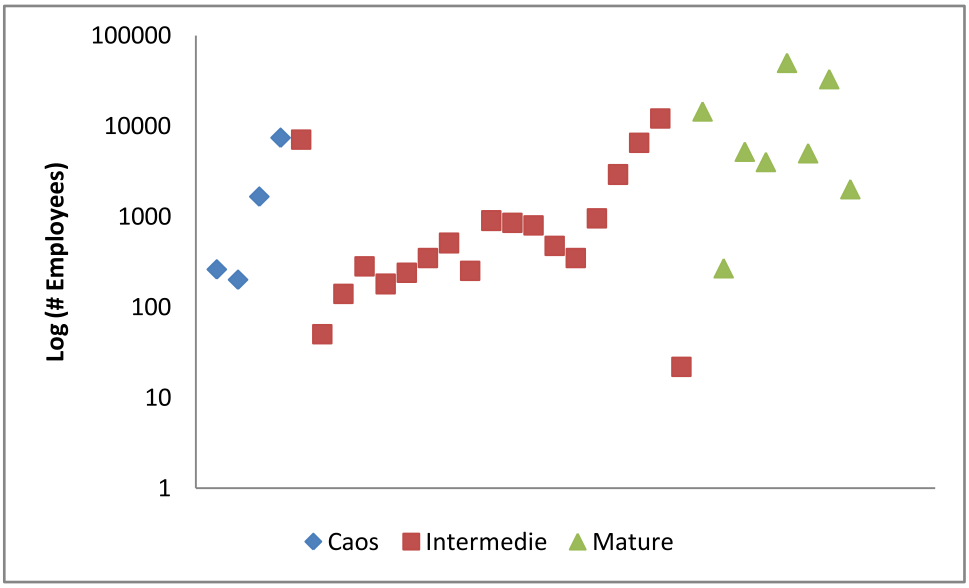

Figure VI. Sample distribution Log Sales (y) Maturity levels (x)

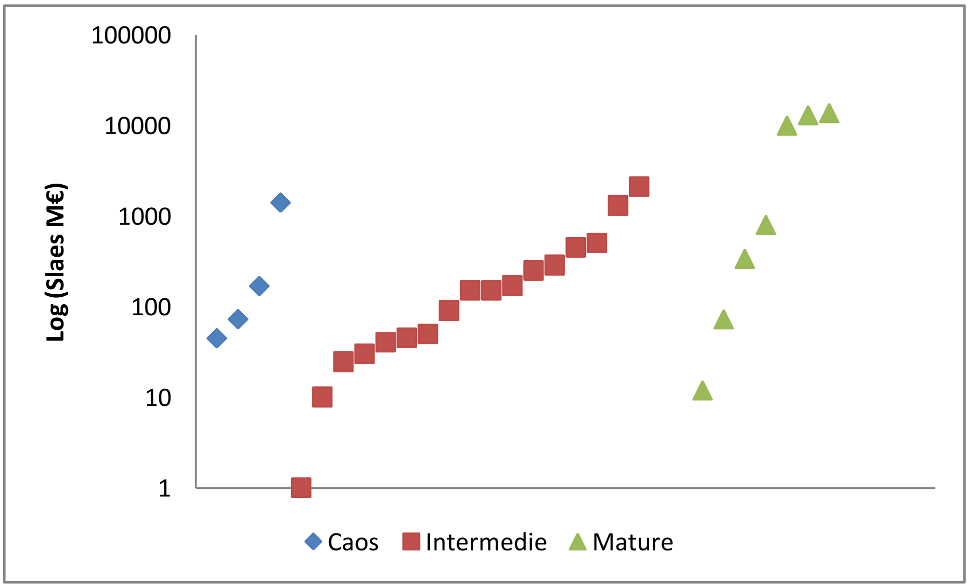

Figure VII. Sample distribution Log Sales (y) Maturity levels (x)

How we say firstly, leader of the market is for companies not that produce a lot of level of sales, despite of have an elevate number of resources and they know how to coordinate them.

It is important follow the mission to coordinate human resources to provide efficient and effective personnel management to departments which empower the users and facilitate the delivery of services to the community.

Chapter 1

Literary criticism

1. INTRODUCTION

New Product Development (NPD) process has assumed a critical role in the modern global competition. Managing NPD process efficiently and effectively is more and more important for the success of the whole company. Many approaches, methods and tools have been created in the last 50 years for supporting this work, but how they are used, known and applied by the companies is still not clear. The general feeling is that there is a relevant gap between the literature theory and the industrial practice.

The global world is more competitive than ever. Markets are global at 360°: it is no more a matter of production delocalization, low cost suppliers, and rich Western costumers. Nowadays design, creation, development, engineering activities can be spread around the world exactly like sourcing, production, distribution, and also consumption and utilization. This is the game to be played today. Due to these global changes, the pressure on the companies (especially at national level) has dramatically increased in the last years, pervading all the processes, activities and organizational structures.

In this context, the backbone of the company competition is more and more its capacity to generate and put into action innovation and value creation: in other terms is its New Product Development process, how it is managed, its effectiveness and its efficiency. As it has been widely said and written (e.g. Womack et al., 1990; Bayus, 1994; Griffin, 1997; Chesbrough et.al., 2006) winning companies are those that are able to provide the right product (with the highest quality, the lowest cost, the most advanced technologies, well-customized, well-branded, eco-compatible, sustainable, etc.), at the right time (with the shortest time-to-market, the smallest time-to-volume, the perfect market life span, etc.), in the right place (around the world, to global and local customers).

Being totally honest, these challenges are not new. They exist since the beginning of the modern industrial era. Certainly, part of the novelty stays in the changed time-scale (e.g. days / weeks instead of months / years), while part stays in the increased degrees of complexity to be concurrently managed (e.g. millions of products and components instead of thousands, multidisciplinary instead of mono, distributed instead of localized, etc.). In any case, many strategic answers and possible solutions can be already retrieved from the literature and from different past experiences. For example, solutions for developing value-added and customer-oriented products can be founded in several “old methods”, like Value Analysis and Engineering (Miles, 1961), or Quality Function Deployment (Akao, 1990). Cost attention is the native focus of Design To Cost techniques (Michaels and Wood, 1996), while product performances (e.g. manufacturability, availability, serviceability, environmental-friendliness, etc.) could be achieved with well-quoted Design for (Design for Manufacturing, Maintenance, Service, Environment, etc.) rules (e.g. Huang, 1996). Organizational and multi-layered approaches and business models have been widely described, from traditional sequential design, to multidisciplinary concurrent engineering (e.g. Prasad, 1996), till lean development (e.g. Liker and Morgan, 2006). At the same time, information technologies (IT) have provided a list of evolutions and revolutions, moving the design experience of the technical departments from the traditional paper-based to computerized 2D and 3D solutions, more and more virtually enabled (e.g. Terzi et al., 2010). Moreover, many collaborative functionalities and platforms have been established in the last 2 decades and nowadays PLM (Product Lifecycle Management) is the reference acronym of such a comprehensive IT market, composed by hundreds of tools and providers, which physically could support a globally distributed NPD process (e.g. Terzi and Garetti, 2008).

The literature state of art provides a list of possible solutions for improving the effectiveness and efficiency of a NPD process: methods, organizational approaches, strategies, techniques, software tools and platforms. But the general feeling is that there is a relevant gap between the literature theory and the industrial practice. Many national companies are facing the modern market challenges, but they don’t know that some methods as well some IT tools are already well-defined and available on the shelves of the possible solutions.

In such a context (at least at Italian level), there is room for a specific research initiative, focused on the existing gap. This initiative is the GeCo Observatory, (Osservatorio GeCo - Gestione dei processi Collaborativi di progettazione, Observatory on the management of design and development collaborative processes). After a list of unsuccessful tentative with public institutions, GeCo has been launched as private sponsored initiative. In fact, GeCo is financially supported by some actors of the NPD process, like PLM vendors and consulting companies experts in the field(namely:

Jmac Europe, PTC, Siemens, EnginSoft, Holonix and PLM Systems), even if it is totally independent in its activities. The research team of GeCo is leaded by the authors and it is composed by scientists coming from different national universities, acting as interested researchers, working for free (namely from: Politecnico di Milano, University of Bergamo, Technical University of Ancona-Marche, University of Brescia, University of Florence, University of Rome "Tor Vergata", University of Salento). GeCo activities have been officially started on February 2012. GeCo aims to be a durable initiative, based on yearly researches. The plan for the first year (2012) is to conduct an explorative empirical research entitled Design Best Practices and Knowledge Management in the NPD process of Italian Companies. This research wants to understand the main issues and challenges currently implemented in Italian companies for improving the NPD process. The research, being explorative, doesn't have any statistical objective, but it is conducted on a list of selected case studies, coming from different sectors (e.g. mechanical, aerospace, textile, fashion, food, etc.), with different dimensions (e.g. employees, turn over) and backgrounds. The only requirement to be strictly fulfilled by a case is that it must be an industrial company, designing, developing and producing a product inside or outside Italy.

The explorative research is based on a reference questionnaire, composed by 33 questions. Data are collected and analyzed according to a reference framework, which gives also the possibility to perform a benchmarking evaluation. Data are collected through cases studies, according to a reference framework structured in 9 areas, divided in 3 parts: Organization, Process and Knowledge Management.

The first part, Organization, which concerns all the people involved in daily company's activities. Core areas are division of labour and tasks (Work Organization), coordination of people and activities, roles of engineers and designers (Roles and Coordination), skills and expertise of the involved practitioners (Skills and Competences).

The second part, or Process perspective, investigates how NPD is practically performed. This part is made of four areas: design rules and methodologies (Methods), control mechanisms and improvement of the process (Process Management), and how decisions are taken every day basing both on internal strategy (Decision Making Factors) and competitors and customers (Activities and Value).

The third part, Knowledge Management, is related to how companies create, share, represent and re-use their tacit and explicit knowledge. This perspective is made by two areas: Formalization (how knowledge is formalized and shared) and Computerization (how IT tools and platforms are used for supporting knowledge storing, sharing and reusing along the NPD process).

For each of the area it is given a % score, obtained weighting the given answers with respect of the better achievable solution. 5 possible Maturity Levels have been identified (Figure 1), basing on the reached % value, named CLIMB (Chaos, Low, Intermediate, Mature, Best Practice):

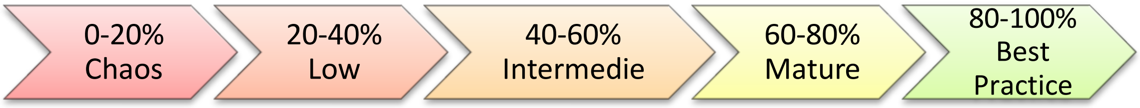

Figure 1. CLIMB Maturity Levels

1. Chaos: NPD is usually chaotic and slightly structured.
2. Low: NPD has a simple formalization and it is barely planned and con-trolled.
3. Intermediate: NPD is structured and planned. Standard solutions are normally applied.
4. Mature: NPD is structured, planned, controlled and measured at its different layers, often through specific quantitative techniques.
5. Best Practice: the organization reached all the previous stages and NPD continuously improves thanks to the analysis of variance of its results. The improvement of NPD performance is reached through incremental and innovative actions.

The collected data of the framework can be represented using a radar chart, as shown in Figure 2.

Since February 2012, 30 Italian manufacturing companies have been assessed so far.

Interviewed companies are from different manufacturing sectors, like aerospace (4), automotive (2), electronics (5), fashion and apparel (4), home appliances (4), machinery (6), mechanics (1), military (1), oil and gas (2), safety (1).

The objective is to collect additional data in order to obtain a more comprehensive picture of the current situation within the Italian market. Moreover it will be interesting to analyze the dynamics and behaviors of different kinds of companies.

Finally, other cases will be collected and analyzed with the final objective of identifying recurring approaches and real best practices.

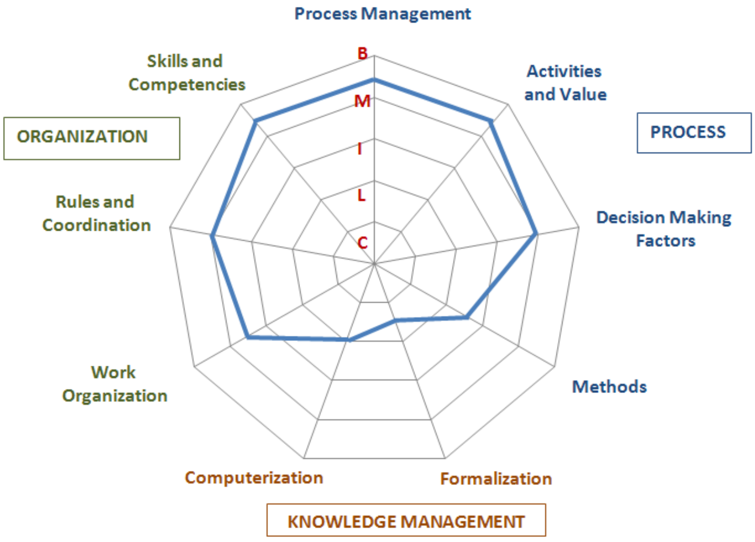

Figure 2. Radar Chart of the Framework

1.1. What is data analysis?

The numerical results provided by a data analysis are usually simple: It finds the number that describes a typical value and it finds differences among numbers. Data analysis finds averages, like the average income or the average temperature, and it finds differences like the difference in income from group to group or the differences in average temperature from year to year. Fundamentally, the numerical answers provided by data analysis are that simple. (Joel H. Levine,1996)

Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.(Wikipedia).

Some people divide data analysis into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA).

Descriptive statistics is the discipline of quantitatively describing the main features of a collection of data.

EDA focuses on discovering new features in the data and CDA on confirming or falsifying existing hypotheses.

Predictive analytics focuses on application of statistical or structural models for predictive forecasting or classification, while text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data.

All are varieties of data analysis.

1.1.1. Descriptive statistics

Descriptive statistics is the discipline of quantitatively describing the main features of a collection of *data* (Prem S. Mann, 1995). Descriptive statistics are distinguished from inferential statistics (or inductive statistics), in that descriptive statistics aim to summarize a data set, rather than use the data to learn about the population that the data are thought to represent. This generally means that descriptive statistics, unlike inferential statistics, are not developed on the basis of probability theory (Y. Dodge, 2003). Even when a data analysis draws its main conclusions using inferential statistics, descriptive statistics are generally also presented.

Descriptive statistics provides simple summaries about the sample and about the observations that have been made. Such summaries may be either quantitative, i.e. summary statistics, or visual, i.e. simple-to-understand graphs. These summaries may either form the basis of the initial description of the data as part of a more extensive statistical analysis, or they may be sufficient in and of themselves for a particular investigation.

The use of descriptive and summary statistics has an extensive history and, indeed, the simple tabulation of populations and of economic data was the first way in which the topic of statistics appeared. More recently, a collection of summarization techniques has been formulated under the heading of exploratory data analysis: an example of such a technique is the box plot. (Wikipedia).

1.1.2. Exploratory data analysis

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics in easy-to-understand form, often with visual graphs, without using a statistical model or having formulated a hypothesis. Exploratory data analysis was promoted by John Tukey (1977) to encourage statisticians visually to examine their data sets, to formulate hypotheses that could be tested on new data-sets.

Tukey's EDA was related to two other developments in statistical theory: Robust statistics and nonparametric statistics, both of which tried to reduce the sensitivity of statistical inferences to errors in formulating statistical models. Exploratory data analysis, robust statistics, nonparametric statistics, and the development of statistical programming languages facilitated statisticians' work on scientific and engineering problems. (Conversation with John W. Tukey and Elizabeth Tukey, Luisa T. Fernholz and Stephan Morgenthaler,*Statistical Science*, Volume 15, Number 1 (2000), 79–94).

The objectives of EDA are to: suggest hypotheses about the causes of observed phenomena, assess assumptions on which statistical inference will be based, support the selection of appropriate statistical tools and techniques and provide a basis for further data collection through surveys or experiments.

1.1.3. Statistical hypothesis testing

A statistical hypothesis test is a method of making decisions using data, whether from a controlled experiment or an observational study (not controlled). (Ronald Fisher, 1925).

Hypothesis testing is sometimes called confirmatory data analysis, in contrast to exploratory data analysis. In frequency probability, these decisions are almost always made using null-hypothesis tests. These are tests that answer the question *Assuming that the null hypothesis is true, what is the probability of observing a value for the test statistic that is at least as extreme as the value that was actually observed*? (Cramer, Duncan and Dennis Howitt, 2004.) More formally, they represent answers to the question, posed before undertaking an experiment, of what outcomes of the experiment would lead to rejection of the null hypothesis for a pre-specified probability of an incorrect rejection. One use of hypothesis testing is deciding whether experimental results contain enough information to cast doubt on conventional wisdom.

Statistical hypothesis testing is a key technique of frequents statistical inference. The Bayesian approach to hypothesis testing is to base rejection of the hypothesis on the posterior probability (M. Schervish, 1996) (David H. Kaye and David A. Freedman, 2011). Other approaches to reaching a decision based on data are available via decision theory and optimal decisions.

1.1.4. Predictive analytics

Predictive analytics encompasses a variety of statistical techniques from modeling, machine learning, data mining and game theory that analyze current and historical facts to make predictions about future events. (Charles Nyce, 2007).

In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions.

The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting it to predict future outcomes. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions. (Wikipedia).

To create an organized piece of word we must use data analysts with lines and tables. The first of these two ideas, the straight line, is the kind of thing I can construct on a graph using a pencil and a ruler, the same idea I can represent algebraically by the equation "y = mx + b".(Joel H. Levine, 1996).

This first idea, the straight line, is the best tool that data analysts have for figuring out how things work. The second idea is the table or, more precisely, the "additive model". The first idea, the line, is reserved for data we can plot on a graph, while this second idea, the additive model, is used for data we organize in tables.

But usually data analysis is really complicated. What do we do with a difficult problem? This may be the single most important thing we teach in data analysis: Common sense would tell you that what you tackle a difficult problem with a difficult technique. Common sense would also tell you that the best data analyst is the one with the largest collection of difficult "high powered" techniques. But common sense is wrong on both points: In data analysis the real "trick" is to simplify the problem and the best data analyst is the one who gets the job done, and done well, with the simplest methods.

Data analysts do not build more complicated techniques for more complicated problems — not if we can help it. The payoff of these techniques is extending the ability of simple tools, of the line and the table, to make sense of a complicated world.

And what are the Rules of data analysis? Some of the rules are clear and easy to state, but these are rather like the clear and easy rules of writing: Very specific and not very helpful — the equivalent of reminders to dot your "i's" and cross your "t's". The real rules, the important ones, exist but there is no list — only broad strategies with respect to which the tactics must be improvised. Nevertheless it is possible to at least name some of these "rules." We'll try the list from different angles:

1. Look At the Data / Think About the Data / Think About the Problem / Ask what it is you Want to Know

Think about the data. Think about the problem. Think about what it is you are trying to discover. That would seem obvious, "Think." But, perhaps, it is the most important step and often omitted as if, somehow, human intervention in the processes of science were a threat to its objectivity and to the solidity of the science. But, no, thinking is required: You have to interpret evidence in terms of your experience. You have to evaluate data in terms of your prior expectations (and you had better have some expectations). You have to think about data in terms of concepts and theories, even though the concepts and theories may turn out to be wrong.

2. Estimate the Central Tendency of the Data.

The "central tendency" can be something as simple as an average. Or it can be something more complicated like a rate. Or it can be something sophisticated, something based on a theory. And why would we have thought to estimate something as specific as a rate or as a theory? Because we have thought about the data, about the problem, and about where we were going (Rule 1).

3. Look at the Exceptions to the Central Tendency

If we've measured a median, look at the exceptions that lie above and below the median. If we've estimated a rate, look at the data that are not described by the rate. The point is that there is always, or almost always, variation: We may have measured the average but, almost always, some of the cases are not average. We may have measured a rate of change but, almost always, some numbers are large compared to the average rate, some are small. And these exceptions are not usually just the result of embarrassingly human error or regrettable sloppiness: On the contrary, often the exceptions contain information about the process that generated the data. And sometimes they tell us that the original idea (to which the variations are the exception) is wrong, or in need of refinement. So, look at the exceptions which, as we can see, brings us back to rule 1, except that this time the data we look at are the exceptions.

The focus is thinking about the world behind the numbers and let good sense and reason guide the analysis. (Joel H. Levine, 1996).

1.2. Data description

1.2.1. Type of data

Data can be of several types:

Quantitative data is data measured or identified on a numerical scale. Numerical data can be analyzed using statistical methods, and results can be displayed using tables, charts, histograms and graphs. Quantitative data involves amounts, measurements, or anything of quantity. After the data is collected the researcher will make an analysis of the quantitative data and produce statistics. (Wikipedia)

Categorical data is a statistical data type consisting of categorical variables, used for observed data whose value is one of a fixed number of nominal categories, or for data that has been converted into that form, for example as grouped data. More specifically, categorical data may derive from either or both of observations made of qualitative data, where the observations are summarized as counts or cross tabulations, or of quantitative data, where observations might be directly observed counts of events happening or they might counts of values that occur within given intervals. Often, purely categorical data are summarized in the form of a contingency table. However, particularly when considering data analysis, it is common to use the term "categorical data" to apply to data sets that, while containing some categorical variables, may also contain non-categorical variables. (Erling B. Andersen, 1980)

Qualitative data is a pass/fail or the presence or lack of a characteristic. It describes items in terms of some quality or categorization that in some cases may be 'informal' or may use relatively ill-defined characteristics such as warmth and flavor; such subjective data are sometimes of less value to scientific research than quantitative data. However, qualitative data can include well-defined concepts such as gender, nationality or commodity type. Qualitative data can be binary (pass-fail, yes-no, etc.) or categorical data. (Y. Dodge, 2003)

1.2.2. The process of data analysis

Data analysis is a process, within which several phases can be distinguished: data cleaning, initial data analysis and main data analysis. (H.J. Adèr, 2008).

Data cleaning

Data cleansing, data cleaning, or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and then replacing, modifying, or deleting this dirty data. (Wikipedia).

After cleansing, a data set will be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores.

Data cleansing differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at entry time, rather than on batches of data.

The actual process of data cleansing may involve removing typographical errors or validating and correcting values against a known list of entities. The validation may be strict (such as rejecting any address that does not have a valid postal code) or fuzzy (such as correcting records that partially match existing, known records).

The motivation, in the business world, is that incorrect data can be costly. Many companies use customer information databases that record data like contact information, addresses, and preferences. For instance, if the addresses are inconsistent, the company will suffer the cost of resending mail or even losing customers.

The existence of anomalies in data motivates the development and application of data cleansing methods. We are able to define data cleansing and specify how to measure the success of cleansing erroneous data.

To be processable and interpretable in an effective and efficient manner, data has to satisfy a set of quality criteria. Data satisfying those quality criteria is said to be of high quality. In general, data quality is defined as an aggregated value over a set of quality criteria (F. Naumann, 2002). Starting with the quality criteria we describe the set of criteria that are affected by comprehensive data cleansing and define how to assess scores for each one of them for an existing data collection. (F. Naumann, 2002 and D. Tonn, 2000).

To measure the quality of a data collection, scores have to be assessed for each of the quality criteria. The assessment of scores for quality criteria can be used to quantify the necessity of data cleansing for a data collection as well as the success of a performed data cleansing process on a data collection. Quality criteria can also be used within optimization of data cleansing by specifying priorities for each of the criteria which in turn influences the execution of data cleansing methods affecting the specific criteria.

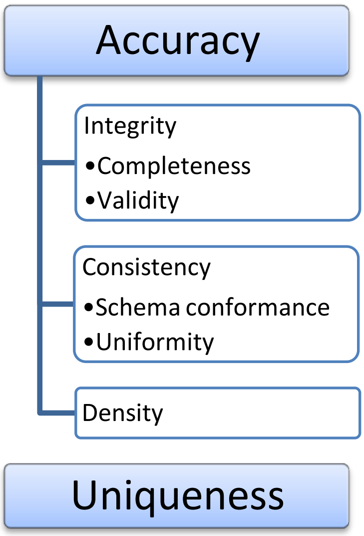

Figure 3. Hierarchy of data quality criteria

The quality criteria defined for comprehensive data cleansing form a hierarchy. This hierarchy results from quality criteria being sub-divided into finer grained quality criteria, i.e., criteria being used as short-cut for a set of quality criteria. Figure 3 shows the resulting hierarchy of the criteria defined below. For each of the criteria we describe how to assess the quality score for a given data collection. Here we assume each collection consisting of only one relational instance.

High-quality data needs to pass a set of quality criteria. Those include:

- *Accuracy*: is defined as the quotient of the number of correct values in the data collection and the overall number of values. Here we use the term accuracy in a different way. Accuracy is described as an aggregated value over the quality criteria Integrity, Consistency, and Density. We define the quality criteria Integrity, Consistency, and Density in the following. (F. Naumann, 2002)
- *Integrity*: is further divided into the criteria Completeness and Validity and therefore again an aggregated value over quality criteria. Intuitively, an integral

data collection contains representations of all the entities and only of those, i.e., there are no invalid tuples or integrity constraint violations as well as no missing tuples. (A. Motro, 1989).

- *Completeness*: achieved by correcting data containing anomalies. Achieving this form of completeness is not a primary data cleansing concern but more of a data integration problem. We achieve completeness within data cleansing by correcting tuples containing anomalies and by not just deleting these tuples if they are representations of entities. (Heiko Müller and Johann-Christoph Freytag, 2003).
- *Validity*: approximated by the amount of data satisfying integrity constraints. The identification of invalid tuples is complicated and sometimes impossible because of the inability or high cost for repeating measurements to verify the correctness of a measured value. Validity can be approximated using the integrity constraints. Integrity constraints represent our understanding of the regularities. Tuples violating integrity constraints are therefore regarded as invalid tuples. Constraint violations arise within systems that do not enforce integrity constraints completely. This might be because of system limitations or on demand of the user, probably because of performance issues. We approximate validity as the quotient of tuples satisfying all integrity constraints. (Heiko Müller and Johann-Christoph Freytag, 2003).
- *Consistency*: concerns contradictions and syntactical anomalies. Intuitively a consistent data collection is syntactically uniform and free of contradictions. (Heiko Müller and Johann-Christoph Freytag, 2003).
- *Uniformity*: directly related to irregularities and in compliance with the set 'unit of measure'.
- *Density*: the quotient of missing values in the data and the number of total values ought to be known. There still can be values or properties non-existent that have to be represented by null values having the exact meaning of not being known. These are no downgrades of data quality. It would be a downgrade if we try to estimate a value for them. (Heiko Müller and Johann-Christoph Freytag, 2003).
- *Uniqueness*: is the quotient of tuples representing the same entity. A collection that is unique does not contain duplicates. Recalling the definition of accuracy as a collection not containing any anomalies except duplicates, a data collection being accurate and unique does not contain any of the anomalies. This describes

a data collection not being in need for data cleansing. (Heiko Müller and Johann-Christoph Freytag, 2003).

A Process Perspective on Data Cleansing

Comprehensive data cleansing is defined as the entirety of operations performed on existing data to remove anomalies and receive a data collection being an accurate and unique representation. It is a (semi-)automatic process of operations performed on data that perform, preferable in this order, (i) format adaptation for tuples and values, (ii) integrity constraint enforcement, (iii) derivation of missing values from existing ones, (iv) removing contradictions within or between tuples, (v) merging and eliminating duplicates, and (vi) detection of outliers, i.e., tuples and values having a high potential of being invalid. Data cleansing may include structural transformation, i.e. transforming the data into a format that is better manageable or better fitting. The quality of schema though is not a direct concern of data cleansing and therefore not listed with the quality criteria defined above.

The process of data cleansing comprises the three major steps (i) auditing data to identify the types of anomalies reducing the data quality, (ii) choosing appropriate methods to automatically detect and remove them, and (iii) applying the methods to the tuples in the data collection. Steps (ii) and (iii) can be seen as specification and execution of a data cleansing workflow. We add another task (iv), the post-processing or control step where we exam the results and perform exception handling for the tuples not corrected within the actual processing. (J.I. Maletic and A. Marcus, 2000). (V. Raman and J.M. Hellerstein, 2001).

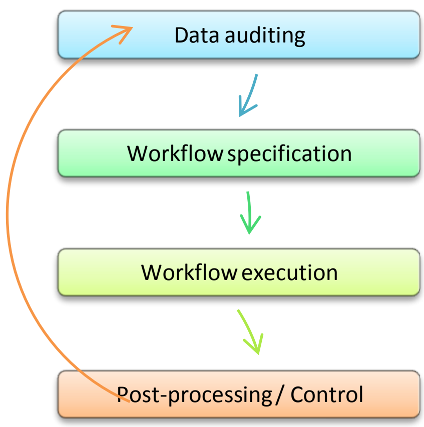

Figure 4. Data cleansing process.

The process of data cleansing normally never finishes, because anomalies like invalid tuples are very hard to find and eliminate. Depending on the intended application of the data it has to be decided how much effort is required to spend for data cleansing.

Data Auditing

The first step in data cleansing is auditing the data to find the types of anomalies contained within it. The data is audited using statistical methods and parsing the data to detect syntactical anomalies. The instance analysis of individual attributes (data profiling) and the whole data collection (data mining) derives information such as minimal and maximal length, value range, frequency of values, variance, uniqueness, occurrence of null values, typical string patterns as well as patterns specific in the complete data collection (functional dependencies and association rules). (E. Rahm and Hong Hai Do, 2004).

The results of auditing the data support the specification of integrity constraints and domain formats. Integrity constraints are depending on the application domain and are specified by domain expert. Each constraint is checked to identify possible violating tuples. For one-time data cleansing only those constraints that are violated within the given data collections have to be further regarded within the cleansing process. Auditing data also includes the search for characteristics in data that can later be used for the correction of anomalies.

As a result of the first step in the data cleansing process there should be an indication for each of the possible anomalies to whether it occurs within the data collection and with which kind of characteristics. For each of these occurrences a function, called tuple partitioner, for detecting all of its instances in the collection should be available or directly inferable.

Workflow Specification

Detection and elimination of anomalies is performed by a sequence of operations on the data. This is called the data cleansing workflow. It is specified after auditing the data to gain information about the existing anomalies in the data collection at hand. One of the main challenges in data cleansing insists in the specification of a cleansing workflow that is to be applied to the dirty data automatically eliminating all anomalies in the data.

For the specification of the operations intending to modify erroneous data the cause of anomalies have to be known and closely considered. The causes for anomalies are manifold. Typical causes for anomalies are impreciseness in measurement or systematic errors in experimental setup, false statements or lazy input habits, inconsistent use of abbreviations, misuse or misinterpretation of data input fields incorrect or careless interpretation of the analysis results, or even be a consequence of anomalies in the data analyzed, leading to invalid tuples results and to a propagation of errors. For the specification of correcting methods the cause of error has to be considered. If for example we assume an anomaly to result from typing errors at data input the layout of the keyboard can help in specifying and assessing the generation of possible solutions. The knowledge about the experiments performed also helps identify and correct systematic errors.

Syntax errors are normally handled first because the data has to be automatically process to detect and remove the other types of anomalies which is additionally hindered by syntax errors. Otherwise there is not specific order in eliminating anomalies by the data cleansing workflow.

Another step is defined after specifying the cleansing workflow and before its execution, the verification. Here, the correctness and effectiveness of the workflow is tested and evaluated. We assume this verification step to be an integral part of the workflow specification. (E. Rahm and Hong Hai Do, 2004).

Workflow Execution

The data cleansing workflow is executed after specification and verification of its correctness. The implementation should enable an efficient performance even on large sets of data. This is often a trade-off because the execution of a data cleansing operation can be quite computing intensive, especially if a comprehensive and 100% complete elimination of anomalies is desired. So we need a heuristics to can achieve the best accuracy while still having an acceptable execution speed. (H. Galhardas, D. Florescu, D. Shasha, E. Simon and C.-A. Saita, 2001).

There is a great demand for interaction with domain experts during the execution of the data cleansing workflow. In difficult cases the expert has to decide whether a tuple is erroneous or not and specify or select the correct modification for erroneous tuples from a set of solutions. The interaction with the expert is expensive and time consuming. Tuples that cannot be corrected immediately are often logged for manual inspection after executing the cleansing workflow.

Post-Processing and Controlling

After executing the cleansing workflow, the results are inspected to again verify the correctness of the specified operations. Within the controlling step the tuples that could not be corrected initially are inspected intending to correct them manually. This results in a new cycle in the data cleansing process, starting by auditing the data and searching for characteristics in exceptional data that allow us to specify an additional workflow to cleanse the data further by automatic processing. This might be supported by learning sequences of cleansing operations for certain anomalies.

Methods used for Data Cleansing

There exists a multitude of different methods used within the data cleansing process. In this section we intend to give a short overview for the most popular of them.

Parsing

Parsing in data cleansing is performed for the detection of syntax errors. A parser for a grammar G is a program that decides for a given string whether it is an element of the language defined by the grammar G. In the context of compilers for programming languages the strings represent computer programs. In data cleansing the strings are either complete tuples of a relational instance or attribute values from as domain. Those strings that represent syntax errors have to be corrected. This can be done for example using edit distance functions choosing the possible correction with the minimal edit distance. (A.V. Aho and J.D. Ullman, 1979).

The existence and amount of syntax errors in a data collection depends on the extend of schema enforcement in the environment where the data is kept. If the data is contained in flat files there exists the possibility of lexical errors and domain errors. In this case, a grammar derived from the file structure is used and the stings represent complete tuples. Data being managed by database management systems is not expected to contain lexical or domain errors. Still, domain format errors can exists for each of the attributes. The grammar used for parsing is the domain format G(A). The specification of domain formats can be supported by pattern learning. They use a sample set of values to deduce the format of the domain. They also generate a discrepancy detector which is used for anomaly detection. (V. Raman and J.M. Hellerstein,1979).

Data Transformation

Data transformation intends to map the data from their given format into the format expected. The transformations affect the schema of the tuples as well as the domains of their values. Schema transformation is often performed in close conjunction with data cleansing. The data from various sources is mapped into a common schema better fitting the needs of the intended application. The correction of values has to be performed only in cases where the input data does not conform to its schema leading to failures in the transformation process. This makes data cleansing and schema transformation supplemental tasks. (S. Abiteboul, S. Cluet, T. Milo, P. Mogilevsky, J, Siméon and S. Zohar, 1999).

Standardization and normalization are transformations on the instance level used with the intension of removing irregularities in data. This includes simple value conversion or translating functions as well as normalizing numeric values to lie in a fixed interval given by the minimum and maximum values. (K.-U. Sattler and E. Schallehn, 2001).

Integrity Constraint Enforcement

In general, integrity constraint enforcement ensures the satisfaction of integrity constraints after transactions modifying a data collection by inserting, deleting, or updating tuples have been performed. The two different approaches are integrity constraint checking and integrity constraint maintenance. Integrity constraint checking rejects transactions that, if applied, would violate some integrity constraint. Integrity constraint maintenance is concerned with identifying additional updates (i.e. repair) to be added to the original transaction to guarantee that the resulting data collection does not violate any integrity constraint. (Enric Mayol and Ernest Teniente, 1999).

Duplicate elimination

Every duplicate detection method proposed requires an algorithm for determine whether two or more tuples are duplicate representations of the same entity. For efficient duplicate detection every tuple has to be compared to every other tuple using this duplicate detection method. In (M.A. Hernandez and S.J. Stolfo, 1995) a fast method (sorted neighborhood method) is developed to reduce the number of required comparisons. The tuples are sorted by a key, possibly constructed from the attributes of the relation, that hopefully brings duplicate tuples close to another. Then only the tuples within a small window (floating over the relation) are compared with each other to find duplicates. The identification whether two tuples are duplicates is done using rules based on domain specific knowledge. In order to improve accuracy, the results of several passes of duplicate detection can be combined by explicitly computing transitive closure of all discovered pair wise duplicate tuples. Not much is said about how the duplicates are merged. This approach is further extended on tuples with syntactically structured attributes, i.e., not conform with the domain format. Because of format errors, tuples that are duplicates might not be close together after sorting. Therefore, the attribute values are tokenized into lexical units and the tokens are then sorted within each attribute before the whole relation is sorted. (Mong Li Lee, Hongjun Lu, Tok Wang Ling and Yee Teng Ko, 1999).

Statistical Methods

These methods can be used for the auditing of data as well as the correction of anomalies. Detection and elimination of complex errors representing invalid tuples go beyond the checking and enforcement of integrity constraints. They often involve relationships between two or more attributes that are very difficult to uncover and describe by integrity constraints. This can be viewed as a problem in outlier detection, i.e., minorities of tuples and values that do not conform to the general characteristics of a given data collection. (J.I. Maletic and A. Marcus, 2000)

By analyzing the data using the values of mean, standard deviation, range, or clustering algorithms a domain expert may find values that are unexpected indicating possible invalid tuples. They can then be analyzed more detail. The correction of such errors is often impossible (besides simply deleting them) because the true values or unknown. Possible solution includes statistical methods like setting the values to some average or other statistical value. Outliers can also be detected as violations of association rules or other existing patterns in the data. (R. Agrawal, T. Imielinski and A. Swami, 1993).

Another anomaly handled by statistical methods is missing values. Missing values are handled based on filling-in (imputing) one or more plausible values. The generation of imputations requires computationally intensive data augmentation algorithms. (J.L. Schafer, 1997). (M.A. Tanner, 1996).

Challenges and Open Problems

In this section we outline some open problems and challenges in data cleansing that are not satisfied until now by the existing approaches. This mainly concerns the management of multiple, alternative values as possible corrections, keeping track of the cleansing lineage for documentation efficient reaction to changes in the used data sources, and the specification and development of an appropriate framework supporting the data cleansing process. (Heiko Müller and Johann-Christoph Freytag, 2003).

Error Correction and Conflict Resolution

The most challenging problem within data cleansing remains the correction of values to eliminate domain format errors, constraint violations, duplicates and invalid tuples. In many cases the available information and knowledge is insufficient to determine the correct modification of tuples to remove these anomalies. This leaves deleting those tuples as the only practical solution. This deletion of tuples leads to a loss of information if the tuple is not invalid as a whole. This loss of information can be avoided by keeping the tuple in the data collection and mask the erroneous values until appropriate information for error correction is available. The data management system is then responsible for enabling the user to include and exclude erroneous tuples in processing and analysis where this is desired.

In other cases the proper correction is known only roughly. This leads to a set of alternative values. The same is true when dissolving contradictions and merging duplicates without exactly knowing which of the contradicting values is the correct one. The ability of managing alternative values allows deferring the error correction until one of the alternatives is selected as the right correction. Keeping alternative values has a major impact on managing and processing the data. Logically, each of the alternatives forms a distinct version of the data collection, because the alternatives are mutually exclusive. It is a technical challenge to manage the large amount of different logical versions and still enable high performance in accessing and processing them.

When performing data cleansing one has to keep track of the version of data used because the deduced values can depend on a certain value from the set of alternatives of being true. If this specific value later becomes invalid, maybe because another value is selected as the correct alternative, all deduced and corrected values based on the now invalid value have to be discarded. For this reason the cleansing lineage of corrected values has to maintain. By cleansing lineage we mean the entirety of values and tuples used within the cleansing of a certain tuple. If any value in the lineage becomes invalid or changes the performed operations have to be redone to verify the result is still valid. The management of cleansing lineage is also of interest for the cleansing challenges described in the following two sections.

Maintenance of Cleansed Data

Cleansing data is a time consuming and expensive task. After having performed data cleansing and achieved a data collection free of errors one does not want to perform the whole data cleansing process in its entirety after some of the values in data collection change. Only the part of the cleansing process should be re-performed that is affected by the changed value. This affection can be determined by analyzing the cleansing lineage. Cleansing lineage therefore is kept not only for tuples that have been corrected, but also for those that have been verified within the cleansing process as being correct. After one of the values in the data collection has changed, the cleansing workflow has to be repeated for those tuples that contains the changed value as part of their cleansing lineage.

The broad definition of require the collection and management of a large amount of additional meta-data to keep track of cleansing lineage. Efficient ways of managing the cleansing lineage have to be developed. It is also of interest to determine which additional information resulting from the initial workflow execution has to be collected in order to be able to speed-up ensuing cleansing workflow executions. (Heiko Müller and Johann-Christoph Freytag, 2003).

Data Cleansing in Virtually Integrated Environments

In virtually integrated sources like IBM's DiscoveryLink (L.M. Haas and al., 2001), the cleansing of data has to be performed every time the data is accessed, which considerably decreases the response time and efficiency. . By collecting and managing appropriate metadata like cleansing lineage and performed operations in a data cleansing middleware the performance could be increased considerably. This could also prevent unnecessary data cleansing if the data in the sources does not change between accessing the sources. There still remains the trade-off between collecting huge amounts of metadata and materializing the complete integrated data collection. The middleware should only collect as much data as necessary but still enable fast cleansing of data. (Heiko Müller and Johann-Christoph Freytag, 2003).

Data Cleansing Framework

In many cases it will not be possible to describe the whole data cleansing graph in advance. This makes data cleansing an iterative, interactive and explorative task. The whole data cleansing process is more the result of flexible workflow execution. Process specification, execution and documentation should be done within a data cleansing framework which in turn is closely coupled with other data processing activities like transformation, integration, and maintenance activities. The framework is a collection of methods for error detection and elimination as well as methods for auditing data and specifying the cleansing task using appropriate user interfaces. Data cleansing should be tightly integrated with the data maintenance within the same framework. (Heiko Müller and Johann-Christoph Freytag, 2003).

1.2.2.1. Initial data analysis

The most important distinction between the initial data analysis phase and the main analysis phase, is that during initial data analysis one refrains from any analysis that are aimed at answering the original research question. The initial data analysis phase is guided by the following four questions: (Àder, 2008).

Quality of data

Data are of high quality "if they are fit for their intended uses in operations, decision making and planning" (J. M. Juran, 1970).

Alternatively, the data are deemed of high quality if they correctly represent the real-world construct to which they refer. Furthermore, apart from these definitions, as data volume increases, the question of internal consistency within data becomes paramount, regardless of fitness for use for any external purpose.

There are a number of theoretical frameworks for understanding data quality. A systems-theoretical approach influenced by American pragmatism expands the definition of data quality to include information quality, and emphasizes the inclusiveness of the fundamental dimensions of accuracy and precision on the basis of the theory of science (Ivanov, 1972).

One framework seeks to integrate the product perspective (conformance to specifications) and the service perspective (meeting consumers' expectations) (Kahn et al. 2002).

Another framework is based in semiotics to evaluate the quality of the form, meaning and use of the data (Price and Shanks, 2004).

One highly theoretical approach analyzes the ontological nature of information systems to define data quality rigorously (Wand and Wang, 1996).

A considerable amount of data quality research involves investigating and describing various categories of desirable attributes (or dimensions) of data. These lists commonly include accuracy, correctness, currency, completeness and relevance.

In practice, data quality is a concern for professionals involved with a wide range of information systems, ranging from data warehousing and business intelligence to customer relationship management and supply chain management. Incorrect data – which includes invalid and outdated information – can originate from different data sources – through data entry, or data migration and conversion projects. (W. Eckerson, 2002).

Problems with data quality don't only arise from incorrect data. Inconsistent data is a problem as well. Eliminating data shadow systems and centralizing data in a warehouse is one of the initiatives a company can take to ensure data consistency.

The market is going some way to providing data quality assurance. Most data quality tools offer a series of tools for improving data, which may include some or all of the following:

1. Data profiling: initially assessing the data to understand its quality challenges
2. Data standardization: a business rules engine that ensures that data conforms to quality rules
3. Geocoding: for name and address data.
4. Matching or Linking: a way to compare data so that similar, but slightly different records can be aligned. Matching may use "fuzzy logic" to find duplicates in the data. It might be able to manage 'householding', or finding links between husband and wife at the same address, for example. Finally, it often can build a 'best of breed' record, taking the best components from multiple data sources and building a single super-record.

5. Monitoring: keeping track of data quality over time and reporting variations in the quality of data. Software can also auto-correct the variations based on pre-defined business rules.
6. Batch and Real time: Once the data is initially cleansed (batch), companies often want to build the processes into enterprise applications to keep it clean.

The choice of analyses to assess the data quality during the initial data analysis phase depends on the analyses that will be conducted in the main analysis phase. (Wikipedia).

Quality of measurements

The quality of the measurement instruments should only be checked during the initial data analysis phase when this is not the focus or research question of the study. One should check whether structure of measurement instruments corresponds to structure reported in the literature. There are two ways to assess measurement quality:

- Confirmatory factor analysis
- Analysis of homogeneity (internal consistency), which gives an indication of the reliability of a measurement instrument. During this analysis, one inspects the variances of the items and the scales, the Cronbach's α of the scales, and the change in the Cronbach's alpha when an item would be deleted from a scale. (Àder, 2008).

Initial transformations

After assessing the quality of the data and of the measurements, one might decide to impute missing data, or to perform initial transformations of one or more variables, although this can also be done during the main analysis phase. (Àder, 2008). Possible transformations of variables are:

- Square root transformation (if the distribution differs moderately from normal)
- Log-transformation (if the distribution differs substantially from normal)
- Inverse transformation (if the distribution differs severely from normal)
- Make categorical (ordinal / dichotomous) (if the distribution differs severely from normal, and no transformations help)

Data transformation refers to the application of a deterministic mathematical function to each point in a data set — that is, each data point z_i is replaced with the transformed value $y_i = f(z_i)$, where f is a function.

Did the implementation of the study fulfill the intentions of the research design?

One should check the success of the randomization procedure, for instance by checking whether background and substantive variables are equally distributed within and across groups.

If the study did not need and/or use a randomization procedure, one should check the success of the non-random sampling, for instance by checking whether all subgroups of the population of interest are represented in sample. Other possible data distortions that should be checked are:

- Dropout (this should be identified during the initial data analysis phase)
- Item noresponse (whether this is random or not should be assessed during the initial data analysis phase)
- Treatment quality (using manipulation checks). (Adèr, 2008)

Characteristics of data sample

In statistics, a sample is a subset of a population. Typically, the population is very large, making a census or a complete enumeration of all the values in the population impractical or impossible. The sample represents a subset of manageable size. Samples are collected and statistics are calculated from the samples so that one can make inferences or extrapolations from the sample to the population. This process of collecting information from a sample is referred to as sampling.

A complete sample is a set of objects from a parent population that includes all such objects that satisfy a set of well-defined selection criteria.

An unbiased sample is a set of objects chosen from a complete sample using a selection process that does not depend on the properties of the objects.

The best way to avoid a biased or unrepresentative sample is to select a random sample, also known as a probability sample. A random sample is defined as a sample where each individual member of the population has a known, non-zero chance of being selected as part of the sample. Several types of random samples are simple random samples, systematic, stratified random samples, and cluster random samples.

A sample that is not random is called a non-random sample or a non-probability sampling.

In any report or article, the structure of the sample must be accurately described. It is especially important to exactly determine the structure of the sample (and specifically the size of the subgroups) when subgroup analyses will be performed during the main analysis phase. The characteristics of the data sample can be assessed by looking at:

- Basic statistics of important variables
- Scatter plots
- Correlations
- Cross-tabulations (Adèr, 2008).

Final stage of the initial data analysis

During the final stage, the findings of the initial data analysis are documented, and necessary, preferable, and possible corrective actions are taken.

Also, the original plan for the main data analyses can and should be specified in more detail and/or rewritten.

1.2.2.2. Main data analysis

In the main analysis phase analyses aimed at answering the research question are performed as well as any other relevant analysis needed to write the first draft of the research report. (Adèr, 2008).

Exploratory and confirmatory approaches

In the main analysis phase either an exploratory or confirmatory approach can be adopted. Usually the approach is decided before data is collected. In an exploratory analysis no clear hypothesis is stated before analyzing the data, and the data is searched for models that describe the data well. In a confirmatory analysis clear hypotheses about the data are tested.

Exploratory data analysis should be interpreted carefully. When testing multiple models at once there is a high chance on finding at least one of them to be significant, but this can be due to a type 1 error.

Type I error (or, error of the first kind) and Type II error (or, error of the second kind) are precise technical terms used in statistics to describe particular flaws in a testing process, where a true null hypothesis was incorrectly rejected (Type I error) or where one fails to reject a false null hypothesis (Type II error).

It is important to always adjust the significance level when testing multiple models with, for example, a *bonferroni* correction. Also, one should not follow up an exploratory analysis with a confirmatory analysis in the same dataset. An exploratory analysis is used to find ideas for a theory, but not to test that theory as well. When a model is found exploratory in a dataset, then following up that analysis with a confirmatory analysis in the same dataset could simply mean that the results of the confirmatory analysis are due to the same type 1 error that resulted in the exploratory model in the first place. The confirmatory analysis therefore will not be more informative than the original exploratory analysis. (Adèr, 2008).

Stability of results

It is important to obtain some indication about how generalizable the results are. While this is hard to check, one can look at the stability of the results. Are the results reliable and reproducible? There are two main ways of doing this: cross-validation and sensitivity analysis. (Adèr, 2008).

Cross-validation

Cross-validation, sometimes called rotation estimation, is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds. (Wikipedia).

Sensitivity analysis

It is a procedure to study the behavior of a system or model when global parameters are (systematically) varied. (Wikipedia).

1.2.3. Data set

A data set (or dataset) is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the data set in question. It lists values for each of the variables, such as height and weight of an object. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows. (Wikipedia).

A data set has several characteristics which define its structure and properties. These include the number and types of the attributes or variables and the various statistical measures. (Jan M. Żytkow and Jan Rauch, 1999).

In the simplest case, there is only one variable, and then the data set consists of a single column of values, often represented as a list. In spite of the name, such a univariate data set is not a set in the usual mathematical sense, since a given value may occur multiple times. Normally the order does not matter, and then the collection of values may be considered to be a multiset rather than an (ordered) list.

The values may be numbers, such as real numbers or integers but may also be nominal data. More generally, values may be of any of the kinds described as a level of measurement. For each variable, the values will normally all be of the same kind. However, there may also be "missing values", which need to be indicated in some way.

A statistical parameter is a data that indexes a family of probability distributions. It can be regarded as a numerical characteristic of a population or a model. (B.S. Everitt, 2002).

The most common statistics parameters are the measures of central tendency. The term central tendency relates to the way in which quantitative data tend to cluster around some value. A measure of central tendency is any of a number of ways of specifying this "central value". (Siddharth Kalla, 2011).

In the simplest cases, the measure of central tendency is an average of a set of measurements, the word average being variously construed as mean, median, or other measure of location, depending on the context. However, the term is applied to multidimensional data as well as to univariate data and in situations where a transformation of the data values for some or all dimensions would usually be considered necessary: in the latter cases, the notion of a "central location" is retained in converting an "average" computed for the transformed data back to the original units. In addition, there are several different kinds of calculations for central tendency, where the kind of calculation depends on the type of data (level of measurement). (Y. Dodge, 2003).

Both "central tendency" and "measure of central tendency" apply to either statistical populations or to samples from a population.

The following basic measures of central tendency may be applied to individual dimensions of multidimensional data, after transformation, although some of these involve their own implicit transformation of the data: arithmetic mean, median, mode, geometric mean, harmonic mean, weighted mean, distance-weighted estimator, truncated mean, midrange, midhinge, trimean, Winsorized mean. (Wikipedia).

Many different descriptive statistics can be chosen as a measure of the central tendency of the data items. These include the arithmetic mean, the median, and the mode.

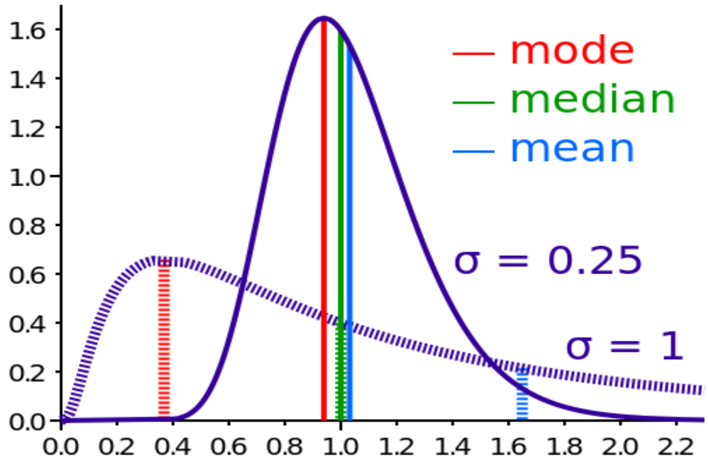

Figure 5. Comparison mean median mode

Other statistics, such as the standard deviation and the range, are called measures of spread and describe how spread out the data is.

For our study we use arithmetic mean and standard deviation.

Arithmetic mean

The arithmetic mean, or simply the mean or average when the context is clear, is the central tendency of a collection of numbers taken as the sum of the numbers divided by the size of the collection. An average is a measure of the "middle" or "typical" value of a data set. (Wikipedia).

Suppose we have sample space$\{\alpha_1, \dots, \alpha_n\}$. Then the arithmetic mean A is defined via the equation:

$$A := \frac{1}{n}\sum_{i=1}^{n} \alpha_i = \frac{\alpha_1 + \alpha_2 + \cdots + \alpha_n}{n} = \mu$$

The arithmetic mean may be misinterpreted as the median to imply that most values are higher or lower than is actually the case. If elements in the sample space increase arithmetically, when placed in some order, then the median and arithmetic average are equal. When one looks at the arithmetic average of a sample space, one must note that the average value can vary significantly from most values in the sample space. (Jyotiprasad Medhi, 1992).

An early meaning (c. 1500) of the word *average* is "damage sustained at sea". The root is found in Arabic as *awar*, in Italian as *avaria*, in French as *avarie* and in Dutch as *averij*. Hence an *average adjuster* is a person who assesses an insurable loss.

Marine damage is either *particular average*, which is borne only by the owner of the damaged property, or general average, where the owner can claim a proportional contribution from all the parties to the marine venture. The type of calculations used in adjusting general average gave rise to the use of "average" to mean "arithmetic mean".

However, according to the Oxford English Dictionary, the earliest usage in English (1489 or earlier) appears to be an old legal term for a tenant's day labour obligation to a sheriff, probably anglicized from "avera" found in the English Domesday Book (1085). This pre-existing term thus lay to hand when an equivalent for *avarie* was wanted.

Standard deviation

The standard deviation (represented by the symbol sigma, σ) shows how much variation or "dispersion" exists from the average (mean, or expected value). A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data points are spread out over a large range of values. (Wikipedia).

Standard deviation may serve as a measure of uncertainty.

The practical value of understanding the standard deviation of a set of values is in appreciating how much variation there is from the "average" (mean).

The standard deviation of a random variable, statistical population, data set, or probability distribution is the square root of its variance. It is algebraically simpler though practically less robust than the average absolute deviation. (Carl Friedrich Gauss, 1816) (Helen Walker, 1931). A useful property of standard deviation is that, unlike variance, it is expressed in the same units as the data.

Consider two numbers, a_1 and a_2. Suppose $a_1 > a_2$. Their mean number μ is the midpoint and the standard deviation σ is the distance from each of the numbers to μ. So μ and σ satisfy the equations:

$$\sigma = \alpha_1 - \mu = \mu - \alpha_2$$

Solving the equations gives:

$$\mu = \frac{\alpha_1 + \alpha_2}{\sigma}$$
$$\sigma = \frac{\alpha_1 - \alpha_2}{\mu}$$

One can find the standard deviation of an entire population in cases (such as standardized testing) where every member of a population is sampled. In cases where that cannot be done, the standard deviation σ is estimated by examining a random sample taken from the population.

An estimator for σ sometimes used is the standard deviation of the sample, denoted by s_N and defined as follows:

$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\alpha_i - \mu)^2}$$

This estimator has a uniformly smaller mean squared error than the *sample standard deviation* (see below), and is the maximum-likelihood estimate when the population is normally distributed. But this estimator, when applied to a small or moderately sized sample, tends to be too low: it is a biased estimator. (Wikipedia).

The standard deviation of the sample is the same as the population standard deviation of a discrete random variable that can assume precisely the values from the data set, where the probability for each value is proportional to its multiplicity in the data set. (Wikipedia).

The most commonly used estimator for σ is an adjusted version, the sample standard deviation, denoted by s and defined as follows:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\alpha_i - \mu)^2}$$

Where $\{\alpha_1, \ldots, \alpha_N\}$ are the observed values of the sample items and μ is the mean value of these observations. This correction (the use of $N - 1$ instead of N) is known as Bessel's correction. The reason for this correction is that s^2 is an unbiased estimator for the variance σ^2 of the underlying population, if that variance exists and the sample values are drawn independently with replacement. Additionally, if N = 1, then there is no indication of deviation from the mean, and standard deviation should therefore be undefined. However, *s* is *not* an unbiased estimator for the standard deviation σ; it tends to underestimate the population standard deviation.

The term *standard deviation of the sample* is used for the uncorrected estimator (using N) while the term *sample standard deviation* is used for the corrected estimator (using $N - 1$). The denominator $N - 1$ is the number of degrees of freedom in the vector of residuals.

The term *standard deviation* was first used in writing by Karl Pearson in 1894, following his use of it in lectures. This was as a replacement for earlier alternative names for the same idea: for example, Gauss used *mean error*.

Relationship between standard deviation and mean

The mean and the standard deviation of a set of data are usually reported together. In a certain sense, the standard deviation is a "natural" measure of statistical dispersion if the center of the data is measured about the mean. This is because the standard deviation from the mean is smaller than from any other point. (Wikipedia).

$$\sigma(\mu) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\alpha_i - \mu)^2}$$

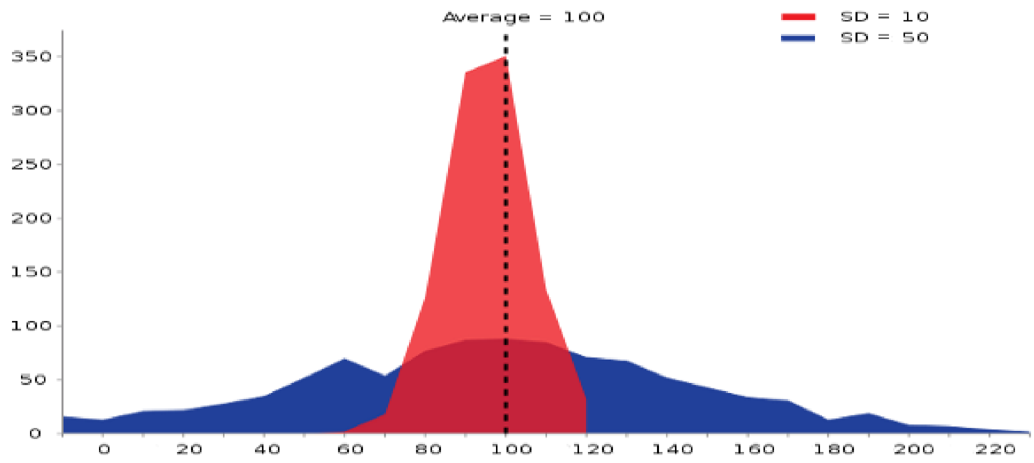

Figure 6. Example of two sample populations with the same mean and different standard deviations

1.2.4. Tools of quality

Take some time to carefully review all of the data we have collected from our experiment is important to obtain the best results.

The Seven Basic Tools of Quality is a designation given to a fixed set of graphical techniques identified as being most helpful in troubleshooting issues related to quality. (Douglas Montgomery, 2005). They are called *basic* because they are suitable for people with little formal training in statistics and because they can be used to solve the vast majority of quality-related issues. (Kaoru Ishikawa, 1985).

The seven tools are: cause-and-effect (also known as the "fish-bone" or Ishikawa) diagram, check sheet, control chart, histogram, pareto chart, scatter diagram and stratification (alternately, flow chart or run chart). (R. Tague Nancy, 2004).

1.2.4.1. Cause- and- effect diagram

Ishikawa diagrams (also called fishbone diagrams, herringbone diagrams, cause-and-effect diagrams, or Fishikawa) are causal diagrams created by Kaoru Ishikawa in 1968 that show the causes of a specific event. Common uses of the Ishikawa diagram are product design and quality defect prevention, to identify potential factors causing an overall effect. Each cause or reason for imperfection is a source of variation. Causes are usually grouped into major categories to identify these sources of variation.

The categories typically include:

- People: Anyone involved with the process
- Methods: How the process is performed and the specific requirements for doing it, such as policies, procedures, rules, regulations and laws
- Machines: Any equipment, computers, tools etc. required to accomplish the job
- Materials: Raw materials, parts, pens, paper, etc. used to produce the final product
- Measurements: Data generated from the process that are used to evaluate its quality
- Environment: The conditions, such as location, time, temperature, and culture in which the process operates

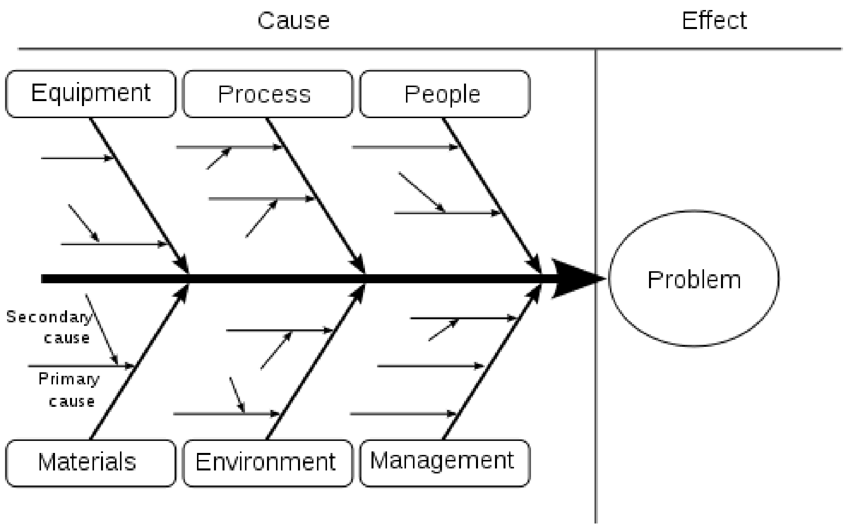

Figure 7. Ishikawa Fishbone Diagram

Causes in the diagram are often categorized, such as to the 6 M's.

- Machine (technology)
- Method (process)
- Material (Includes Raw Material, Consumables and Information.)
- Man Power (physical work)/Mind Power (brain work): Kaizens, Suggestions
- Measurement (Inspection)
- Milieu/Mother Nature (Environment)

In a discussion of the nature of a cause it is customary to distinguish between necessary and sufficient conditions for the occurrence of an event. A necessary condition for the occurrence of a specified event is a circumstance in whose absence the event cannot occur. A sufficient condition is a circumstance in whose presence the event must occur. Ishikawa diagrams have been criticized for failing to make the distinction between necessary conditions and sufficient conditions. It seems that Ishikawa was not even aware of this distinction. (Frank Hutson Gregory, 1992).

1.2.4.2. Check sheet

The check sheet is a form (document) used to collect data in real time at the location where the data is generated. The data it captures can be quantitative or qualitative. When the information is quantitative, the check sheet is sometimes called a tally sheet. (Frank Hutson Gregory, 1992).

The defining characteristic of a check sheet is that data is recorded by making marks ("checks") on it. A typical check sheet is divided into regions, and marks made in different regions have different significance. Data is read by observing the location and number of marks on the sheet.

Kaoru Ishikawa identified five uses for check sheets in quality control: (Kaoru Ishikawa, 1986).

- To check the shape of the probability distribution of a process
- To quantify defects by type
- To quantify defects by location
- To quantify defects by cause (machine, worker)
- To keep track of the completion of steps in a multistep procedure (in other words, as a checklist)

Check sheets are not limited to those described above. Users should employ their imaginations to design check sheets tailored to the circumstances. (Kaoru Ishikawa, 1986).

1.2.4.3. Control chart

Control charts, also known as Shewhart charts or process-behavior charts, in statistical process control are tools used to determine whether a manufacturing or business process is in a state of statistical control. (Wikipedia).

If analysis of the control chart indicates that the process is currently under control (i.e. is stable, with variation only coming from sources common to the process) then no corrections or changes to process control parameters are needed or desirable. In addition, data from the process can be used to predict the future performance of the process. If the chart indicates that the process being monitored is not in control, analysis of the chart can help determine the sources of variation, which can then be eliminated to bring the process back into control. A control chart is a specific kind of run chart that allows significant change to be differentiated from the natural variability of the process.

The control chart can be seen as part of an objective and disciplined approach that enables correct decisions regarding control of the process, including whether to change process control parameters. Process parameters should never be adjusted for a process that is in control, as this will result in degraded process performance. (William McNeese, July 2006). A process that is stable but operating outside of desired limits needs to be improved through a deliberate effort to understand the causes of current performance and fundamentally improve the process. (Donald J. Wheeler, 2000).

A control chart consists of:

- Points representing a statistic (e.g., a mean, range, proportion) of measurements of a quality characteristic in samples taken from the process at different times [the data]
- The mean of this statistic using all the samples is calculated (e.g., the mean of the means, mean of the ranges, mean of the proportions)
- A center line is drawn at the value of the mean of the statistic
- The standard error (e.g., standard deviation/ sqrt (n) for the mean) of the statistic is also calculated using all the samples
- Upper and lower control limits (sometimes called "natural process limits") that indicate the threshold at which the process output is considered statistically 'unlikely' are drawn typically at 3 standard errors from the center line

The chart may have other optional features, including:

- Upper and lower warning limits, drawn as separate lines, typically two standard errors above and below the center line
- Division into zones, with the addition of rules governing frequencies of observations in each zone
- Annotation with events of interest, as determined by the Quality Engineer in charge of the process's quality

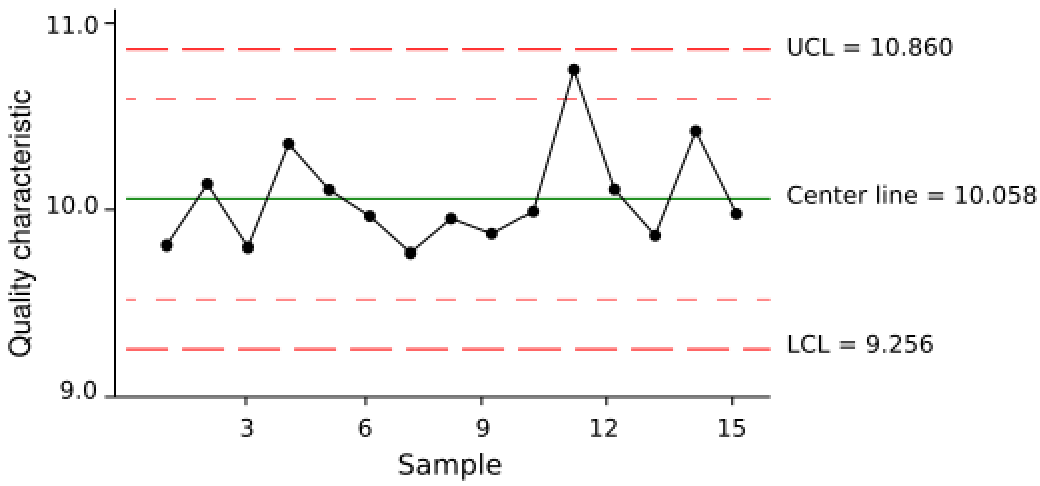

Figure 7. Control Chart

If the process is in control (and the process statistic is normal), 99.7300% of all the points will fall between the control limits. Any observations outside the limits, or systematic patterns within, suggest the introduction of a new (and likely unanticipated) source of variation, known as a special-cause variation.

This makes the control limits very important decision aids. The control limits tell you about process behavior and have no intrinsic relationship to any specification targets or engineering tolerance. In practice, the process mean (and hence the center line) may not coincide with the specified value (or target) of the quality characteristic because the process' design simply cannot deliver the process characteristic at the desired level.

The purpose of control charts is to allow simple detection of events that are indicative of actual process change. This simple decision can be difficult where the process characteristic is continuously varying; the control chart provides statistically objective criteria of change. When change is detected and considered good its cause should be identified and possibly become the new way of working, where the change is bad then its cause should be identified and eliminated.

Several authors have criticized the control chart on the grounds that it violates the likelihood principle. However, the principle is itself controversial and supporters of control charts further argue that, in general, it is impossible to specify a likelihood function for a process not in statistical control, especially where knowledge about the cause system of the process is weak.

Some authors have criticized the use of average run lengths (ARLs) for comparing control chart performance, because that average usually follows a geometric distribution, which has high variability and difficulties.

Some authors have criticized that most control charts focus on numeric data. Nowadays, process data can be much more complex, e.g. non-Gaussian, mix numerical and categorical, missing-valued. (H. Deng, G. Runger and E. Tuv, 2012).

1.2.4.4. Histogram

A histogram is a graphical representation showing a visual impression of the distribution of data. It is an estimate of the probability distribution of a continuous variable and was first introduced by Karl Pearson in 1985. A histogram consists of tabular frequencies, shown as adjacent rectangles, erected over discrete intervals (bins), with an area equal to the frequency of the observations in the interval. The height of a rectangle is also equal to the frequency density of the interval. The total area of the histogram is equal to the number of data. A histogram may also be normalized displaying relative frequencies.

In a more general mathematical sense, a histogram is a function m_i that counts the number of observations that fall into each of the disjoint categories (known as *bins*), whereas the graph of a histogram is merely one way to represent a histogram. Thus, if we let n be the total number of observations and k be the total number of bins, the histogram m_i meets the following conditions:

$$n = \sum_{i=1}^{k} m_i$$

A cumulative histogram is a mapping that counts the cumulative number of observations in all of the bins up to the specified bin. That is, the cumulative histogram M_i of a histogram m_j is defined as:

$$M_i = \sum_{j=1}^{i} m_j$$

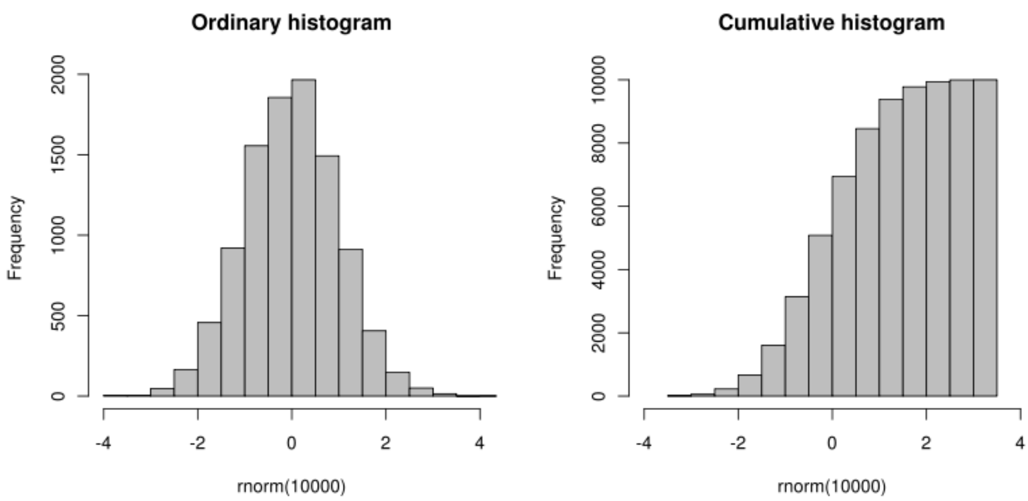

Figure 8. Normal vs Cumulative histogram

There is no "best" number of bins, and different bin sizes can reveal different features of the data. Some theoreticians have attempted to determine an optimal number of bins, but these methods generally make strong assumptions about the shape of the distribution. Depending on the actual data distribution and the goals of the analysis, different bin widths may be appropriate, so experimentation is usually needed to determine an appropriate width. There are, however, various useful guidelines and rules of thumb.

The number of bins k can be assigned directly or can be calculated from a suggested bin width h as:

$$K = \frac{\max x - \min x}{h}$$

1.2.4.5. *Pareto chart*

A Pareto chart, named after Vilfredo Pareto, is a type of chart that contains both bars and a line graph, where individual values are represented in descending order by bars, and the cumulative total is represented by the line. (Wikipedia).

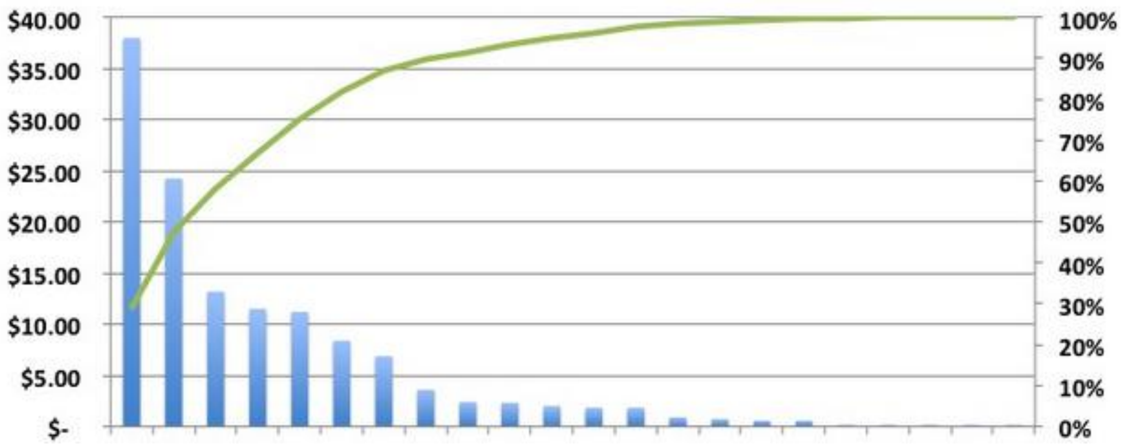

Figure 9. Pareto Chart

The left vertical axis is the frequency of occurrence, but it can alternatively represent cost or another important unit of measure. The right vertical axis is the cumulative percentage of the total number of occurrences, total cost, or total of the particular unit of measure. Because the reasons are in decreasing order, the cumulative function is a concave function.

A pareto chart can be constructed by segmenting the range of the data into groups (also called segments, bins or categories). The purpose of the Pareto chart is to highlight the most important among a (typically large) set of factors. In quality control, it often represents the most common sources of defects, the highest occurring type of defect, or the most frequent reasons for customer complaints, and so on. (http://www.isixsigma.com/tools-templates/pareto/pareto-chart-bar-chart-histogram-and-pareto-principle-8020-rule/)

1.2.4.6. Scatter plot

A scatter plot or scatter graph is a type of mathematical diagram using Cartesian coordinates to display values for two variables for a set of data. (Wikipedia).

The data is displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis. (Jessica M. Utts, 2005). This kind of plot is also called a *scatter chart*, *scattergram*, *scatter diagram* or *scatter graph*.

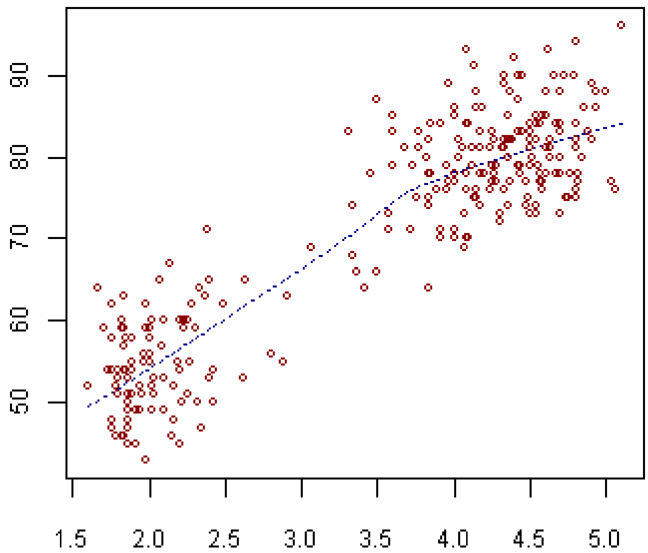

Figure 10. Scatter plot

Graphical statistical methods have four objectives: the exploration of the content of a data set, the use to find structure in data, checking assumptions in statistical models and communicate the results of an analysis. (G. Jacoby William, 1997).

If one is not using statistical graphics, then one is forfeiting insight into one or more aspects of the underlying structure of the data.

A scatter plot can suggest various kinds of correlations between variables with a certain confidence interval. Correlations may be positive (rising), negative (falling), or null (uncorrelated). If the pattern of dots slopes from lower left to upper right, it suggests a positive correlation between the variables being studied. If the pattern of dots slopes from upper left to lower right, it suggests a negative correlation. A line of best fit (alternatively called 'trendline') can be drawn in order to study the correlation between the variables. An equation for the correlation between the variables can be determined by established best-fit procedures. A scatter plot is also very useful when we wish to see how two comparable data sets agree with each other. The more the two data sets agree, the more the scatters tend to concentrate in the vicinity of the identity line; if the two data sets are numerically identical, the scatters fall on the identity line exactly.

One of the most powerful aspects of a scatter plot, however, is its ability to show nonlinear relationships between variables. Furthermore, if the data is represented by a mixture model of simple relationships, these relationships will be visually evident as superimposed patterns.

1.2.4.7. Stratified sampling

A stratified sample is a probability sampling technique in which the researcher divides the entire target population into different subgroups, or strata, and then randomly selects the final subjects proportionally from the different strata. This type of sampling is used when the researcher wants to highlight specific subgroups within the population. (Hunt, Neville and Tyrrell, Sidney, 2001).

Stratification is the process of dividing members of the population into homogeneous subgroups before sampling. The strata should be mutually exclusive: every element in the population must be assigned to only one stratum. The strata should also be collectively exhaustive: no population element can be excluded. Then simple random sampling or systematic sampling is applied within each stratum. This often improves the representativeness of the sample by reducing sampling error. It can produce a weighted mean that has less variability than the arithmetic mean of a simple random sample of the population.

It is important to note that the strata used in stratified sampling must not overlap. Having overlapping subgroups will give some individuals a higher chance of being selected as subjects in the sample. If this happened, it would not be a probability sample.

There are many situations in which researchers would choose stratified random sampling over other types of sampling. First, it is used when the researcher wants to highlight a specific subgroup within the population. Stratified sampling is good for this because it ensures the presence of key subgroups within the sample.

Researchers also use stratified random sampling when they want to observe relationships between two or more subgroups. With this type of sampling, the researcher is guaranteed subjects from each subgroup are included in the final sample, whereas simple random sampling does not ensure that subgroups are represented equally or proportionately within the sample.

Researchers who are interested in rare extremes of a population often use stratified random sampling because it can be representatively sampled even the smallest and most inaccessible subgroups of the population. Simple random sampling does not allow this.

Using a stratified sample will always achieve greater precision than a simple random sample, provided that the strata have been chosen so that members of the same stratum are as similar as possible in terms of the characteristic of interest. The greater the differences between the strata, the greater the gain in precision.

Another advantage that stratified random sampling has over simple random sampling is that is guarantees better coverage of the population. The researcher has control over the subgroups that are included in the sample, whereas simple random sampling does not guarantee than any one type of person will be included in the final sample.

One main disadvantage of stratified random sampling is that is can be difficult to identify appropriate strata for a study. A second disadvantage is that it is more complex to organize and analyze the results compared to simple random sampling.

The Seven Basic Tools stand in contrast to more advanced statistical methods such as survey sampling, acceptance sampling, statistical hypothesis testing, design of experiments, multivariate analysis, and various methods developed in the field of operations research. (Kaory Ishikawa, 1985).

1.3. The assessment process

New products and services are the lifeblood of all businesses. Investing in their development isn't an optional extra - it is crucial to business growth and profitability. Identifying where products or services are in their lifecycle is central to our profitability. Effective research into our markets and competitors will help us do this. Ideally, we should always have new products or services to introduce as others decline so that at least one part of our range is showing a sales peak.

The beginning of new products is developed over a process where it can and must make fundamental decisions.

New products and services have to offer benefits that meet our customers' needs. We need to discover what these are. But we mustn’t only meet our customers' needs; we have to do so in a way that is better than the alternatives offered by the competition.

It is necessary to establish a process for new development products because for companies is a crucial aspect, for so many reasons. First we must take into account that the company has to offer new products to market constantly. These products present different tastes and preferences which change and are in a continuous evolution. So we must be aware of these needs and develop the company’s offer. The design processes of new products are conducted simultaneously for many new ideas and for that we must considerer the big number of losers throughout the process. Therefore, if many new ideas of new products will not come to develop, we must establish systems which can generate new ideas continuously.

For that managers should to focus on creating new products to get a fast position in the market and to satisfy consumers’ needs. In this way, they must take into consideration the strategic decisions, promotion, distribution, price and publicity, as well as the creation of the brand, label and packaging.

The purpose of our work is defining the maturity of development and design process of different companies. We must find correlations in the manner of made something and the level of performance obtained. We will be in a different stage of maturity in correlation with efforts made by the different levels of the company.

The aim of the assessment model for DNP process is to provide a ‘picture’ of enterprise that will be starting point for the AS-IS analysis.

A questionnaire and a Radar model were developed as tools for the assessment process.

2.3 Training and competence

2.3.1 The design is heavily based on skills and competence of the actors involved (technicians, designers, Methodists, etc...). On average, how does the company support the training and consolidation of its skills?	Risposta	
a. Any engineer / designer is personally responsible for developing and maintaining his/ her skills.		
b. The company urges the development of strong technical skills, with an approach to training on the job.	1	
c. The company promotes a multidisciplinary degree, supporting the knowledge management with formal programs (ex. training plans, rotating between project team, etc,...).		Puntegg
d. Other (specify).		2
2.3.2 Inside the organization are there figures that support the training?	Risposta	
a. There are not figures that support development skills, each technician/ designer is free to form individually.		
b. The technician/ designer is encouraged to develop his/her own skills from his/ her supervisor.		
c. There are methods one-to-one (the person is related individually with a more experienced person, as a tutor, coach, and mentor).	1	Puntegg
d. Other (specify).		3
2.3.3 In the case that it exists a training courses how are the skills valued?	Risposta	
a. Through a 'visual' evaluation it is verified improvements of individual behavior.	1	
b. It is applied test before the training and test at the end of the course.		
c. It is used KPI to assess the impact of training on business development strategies.		Puntegg
d. Other (specify).		1

Figure 11. Example of the area Skills and Competences, composted by 3 questions

Organizzazione	Punteggi	Correzio
Organizzazione del Lavoro	19,5	92,8571429
Ruoli e Coordinamento	9,5	79,1666667
Formazione e Competenze	6	66,6666667

Figure 12. Example of evaluation of macro area Organization, composted by 3 areas A1, A2 and A3.

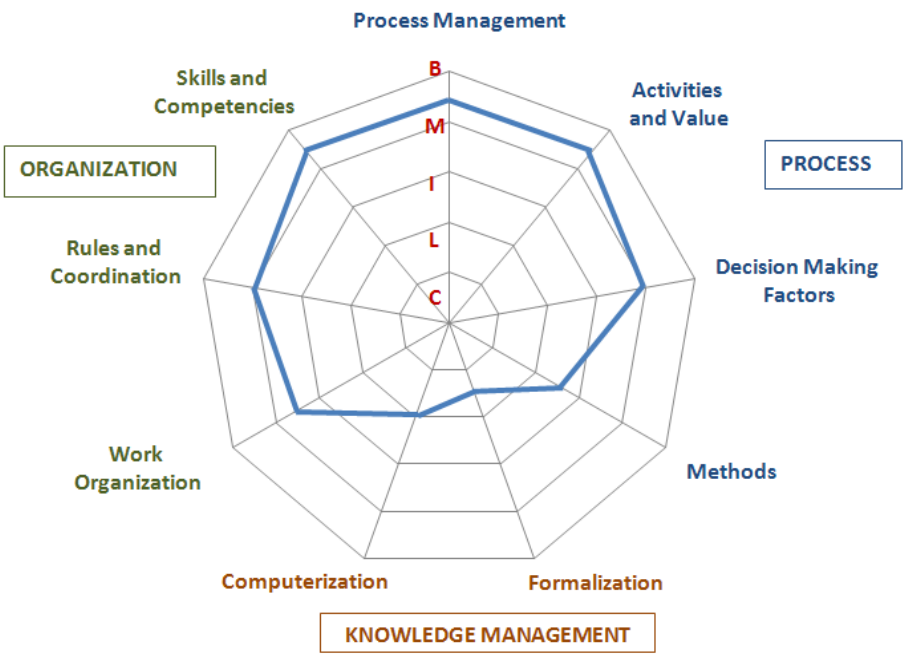

Figure 13. Example of Radar Chart of a company

One person or a group can answer the questionnaire, it is up to the top management (or to the responsible of the contact with GeCo) to decide whether involve a single person, a group of people (for discuss and agree all the questions, exchanging experience and information), or involve different people for each area of the questionnaire (for example more skilled people on PLM systems knows everything about knowledge computerization but is not aware about process management, and vice versa). Usually people involved are project managers from design department.

The questionnaire is composed by multiple choice close questions and by tables with a conveniently defined score able to represent the results of the enterprise directly on the radar, on the five maturity levels.

The maturity level of NDP process is based on five levels: Best Practice, Mature Practice, Intermediate Practice, Low Practice and Chaos.

The Radar Chart is divided in three main areas: Organization, Process and Knowledge Management. Each of one is subdivided into different elements or sub-areas.

- Organization:
    - Work organization;
    - Roles and Coordination;
    - Education and Expertise;
- Process:
    - Process management;
    - Activities and Value;
    - Decision-making factors;
    - Methods;
- Know-how management
    - Computerization;
    - Formalization;

Each area is evaluated through a proper group of questions with multiple answer possibility. To each answer it is associated a value (from 1 to 3, not necessarily integer); also possible alternative answers can be added with a proper evaluation (scale 1-3, not necessarily integer value). The higher is the value the better is the performance; evaluations are based on well proved literature basis. The macro area is then evaluated considering the relative areas with their associated questions. All the values are then normalized to 0-100% scale.

Then, the normalized values are classified according 5 classes of maturity that we defined CLIMB (Figure 14), it means from chaotic management to best practice: Chaos, Low, Intermediate, Mature, Best practice. Each class has 20% scale range. A company can be best practice according skill and competencies area and be intermediate or low in computerized area. We introduced a radar chart for a unique representation of company's performance, it includes 9 areas, 5 maturity levels and the line representing as-is maturity level. (Figure 13)

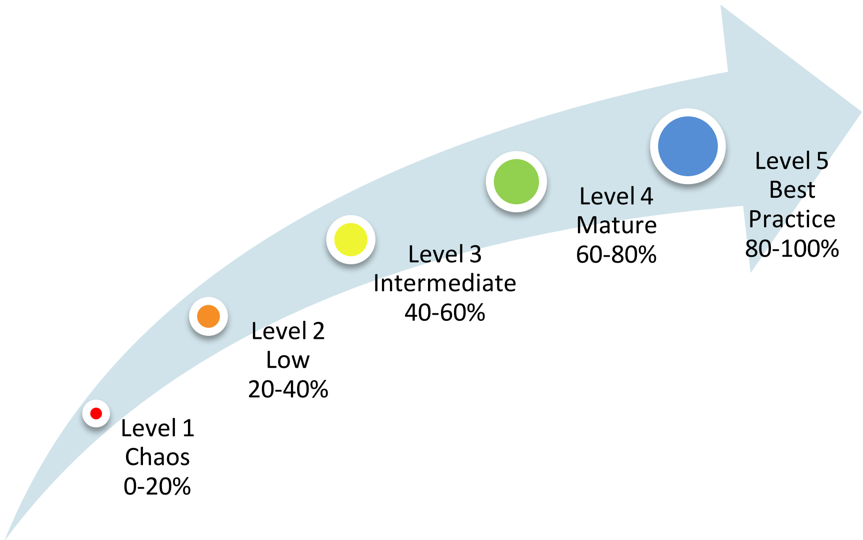

Figure 14. CLIMB Maturity levels

Each level can be described like:

- Level 1 – Chaos (0-20%): this is the initial maturity level, the development process is usually chaotic and little structured. The organization doesn't have a stable environmental development. The success of the company derives exclusively from the education and expertise of the staff because there is not a well-established process.
- Level 2 – Low (20-40%): the organization defines the temporal goal of the NPD process. Afterwards the projects developed by the organization guarantee the realization of the requirements. The process is planed, done, measured and controlled.

- Level 3 – Intermediate (40/60%): an organization reaches the third level when all the goals of the previous level (level 2) are reached including the goal. At this stage the NPD process is well structured inside the organization.
- Level 4 – Mature (60-80%): an organization reached all the goals of the previous levels (levels 1, 2, 3). Moreover it defines the processes that mostly contribute to the performance and the creation of the NPD process. They are observed through statistical/mathematical tools and through specific quantitative techniques.
- Level 5 – Best practice (80-100%): an organization reached all the previous stages (level 1, 2, 3, 4) and the process continuously improves thanks to the analysis of variance of the results. The improvement of the process results is reached through incremental and innovative of the used technologies.

Hence, the final result of the analysis is an evaluation of the 9 areas for the whole sample of companies. So, until now we can only provide a punctual comparison between single areas but not comparing companies considering the all areas as a whole, we can't rate companies. Rating should be important to identify which companies are actually best practices considering the whole vision and will be useful for benchmark companies, that are always interested in understand their own position in the marketplace.

Then we used Excel to do some basic data analysis tasks to see whether it is a reasonable alternative to using a statistical package for the same tasks. Excel is the widely used statistical package, which serves as a tool to understand statistical concepts and computation to check our hand-worked calculation in solving our homework problems.

Then firstly we use the questionnaire. The next step is using a control chart where we see, for all companies, what is the status of each company with regards to the other companies and the average. It was calculated the average and the control limits.

The upper control limit for all part in the assessment is 100%. This value is the best asset that a company can obtain and it is in Level 5, which is best practice.

The lower control limit depends for each part in the assessment.

Then we use a histogram to show the frequency of the standard deviation. This histogram is divided in thirteen part, every part with the same interval that the other.

Chapter 2

Data analysis

2.1. Classification of groups

All observations that are to be described in terms of statistics require systematic classification. Classifications partition the universe of statistical observations according to sets that are as homogeneous as possible with respect to the characteristics of the object of the statistical survey.

At the begging, we have to find into data meaningful or useful groups. If meaningful groups are the goal, the resulting groups should capture the 'natural' structure of the data. In this first case, data groups allow create documents for browsing, to find dates which have a similar functionality, and to provide a grouping of spatial locations. However, in other case, having a cluster analysis is only a useful starting point for other purposes, e.g., data compression or efficiently. Whether for understanding or utility, our focus is finding a classification of dates which can be used in a wide variety of fields.

We have to discover useful, but non-obvious, information or patterns in large collections of data. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups.

Statistical classifications are characterized by an exhaustive coverage of the observed universe; mutually exclusive categories: each element should be classified in only one category of the classification; and a methodological principles which allow the consistent allocation of the elements to the various categories of the classification.

Then, firstly we have to create a classification of dates. The best definition depends on the type of data and this allows us find the best results of our subsequent analysis.

For our analysis, we use four different classifications of dates. We divided the companies for their size, sector and type of market that they provide and the trend of market.

2.1.1. Size of the company

There is no unanimity among economists to determinate what large or small is a company. There is no single criterion to measure the size of the company. The main indicators are: sales volume, equity, number of employees, benefits, etc. (Wikipedia).

The most usually used is the number of workers. This criterion delimits the companies' dimension as: (Figure 15)

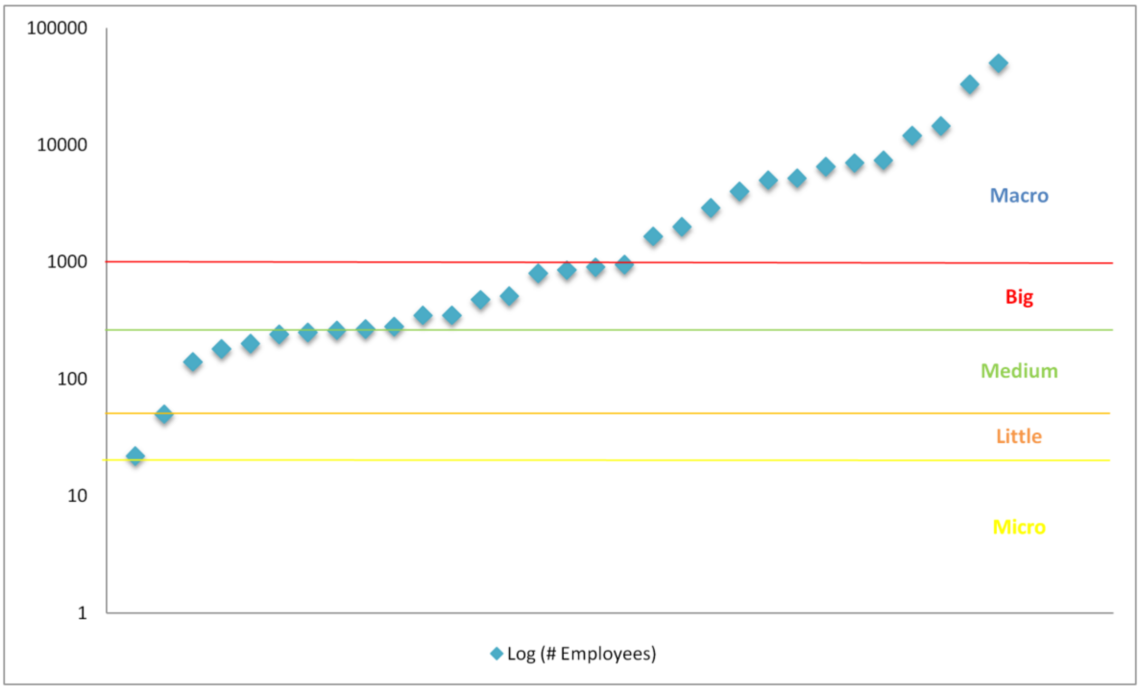

Figure 15. Sample- distribution of number of employees in Log scale (y) in 30 companies (x)

- Micro business: if there is 10 or less employees
- Little business: if the company has between 11 and 49 employees
- Medium business: if there are between 50 and 249 employees
- Big business: if the company has between 250 and 1000 employees
- Macro business: if there are more than 1000 employees

Type of Company	Micro	Little	Medium	Big	Macro
Number of employees	<10	11-49	50-249	250-999	>1000

Table 1. Classification the company in order to the number of employees

2.1.2. *Sector: NACE, European Classification of Economic Activities*

NACE is the acronym used to designate the various statistical classifications of economic activities developed since 1970 in the European Union. NACE provides the framework for collecting and presenting a large range of statistical data according to economic activity in the fields of economic statistics (e.g. production, employment, national accounts) and in other statistical domains. (europa.eu)

NACE is the European standard classification of productive economic activities. NACE presents the universe of economic activities partitioned in such a way that a NACE code can be associated with a statistical unit carrying them out.

It is a basic element of the international integrated system of economic classifications, which is based on classifications of the UN Statistical Commission (UNSTAT), Eurostat as well as national classifications; all of them strongly related each to the others, allowing the comparability of economic statistics produced worldwide by different institutions.

One NACE code is assigned to each unit recorded in statistical business registers, according to its principal economic activity. The principal activity is the activity which contributes most to the value added of the unit. The assignment of the NACE code is helped by: the explanatory notes of NACE, decisions taken by the NACE management committee, correspondence tables and reference to other classification systems.

In this way, our classification requires these codes:

2540	Manufacture of weapons and ammunition
2573	Manufacture of tools
2599	Manufacture of other fabricated metal products n.e.c.
2540	Manufacture of weapons and ammunition
2611	Manufacture of electronic components
2630	Manufacture of communication equipment.
2651	Manufacture of instruments and appliances for measuring, testing and navigation
2712	Manufacture of electricity distribution and control apparatus
2740	Manufacture of electric lighting equipment
2751	Manufacture of electric domestic appliances
2790	Manufacture of other electrical equipment
2811	Manufacture of engines and turbines, except aircraft, vehicle and cycle engines
2812	Manufacture of fluid power equipment

2829	Manufacture of other general-purpose machinery n.e.c.
2932	Manufacture of other parts and accessories for motor vehicles
3030	Manufacture of air and spacecraft and related machinery
3299	Other manufacturing n.e.c.
4299	Construction of other civil engineering projects n.e.c.
7410	Specialised design activities

Table 2. Different NACE Sectors used for our analysis

The different colors represent the subgroup that we use for our analysis. Because in some case, each company require a type of group that is similar for other one and it is more useful to put in groups.

2.1.3. Type of market

Markets can be analyzed via the product itself, or end-consumer, or both.

We have created four groups to create a classification for the different companies we have to analyze. In this way we distingue the next type of business production strategic:

- Make to Order (MTO)
- Make to Stock (MTS)
- Catalogue
- Shelf

Make to Order

The make to order strategy, also built to stock (BTO), only manufactures the end product once the customer places the order. This creates additional wait time for the consumer to receive the product, but allows for more flexible customization compared to purchasing from retailers' shelves. (Wikipedia).

MTO strategy relieves the problems of excessive inventory that is common with the traditional make to stock (MTS) strategy. (investomedia.com)

Related approaches to MTO include the Engineer to Order (ETO) approach, where after an order is received, a part of or the whole design is done, as well as the Assemble to Order (ATO). (Wikipedia).

Make to Stock

Build to stock, or make to stock, often abbreviated as BTS or MTS, is a build-ahead production approach in which production plans may be based upon sales forecasts and/or historical demand. (Wikipedia).

MTS literally means to manufacture products for stock based on demand forecasts, which can be regarded as push-type production. MTS has been required to prevent opportunity loss due to stock-out and minimize excess inventory using accurate forecasts. In the industrialized society of mass production and mass marketing, this forecast mass production urged standardization and efficient business management such as cost reduction. (investopedia.com)

If demand for the product can be accurately forecasted, the MTS strategy can be an efficient choice. It is frequently considered as an appropriate solution for products where there are few product lines and long changeover times between costly products. (investomedia.com)

Catalogue

A catalog merchant is a form of retailing. Unlike a self-serve retail store, most of the items are not displayed; customers select the products from printed catalogs in the store and fill out an order form. The order is brought to the sales counter, where a clerk retrieves the items from the warehouse area to a payment and checkout station. The catalog merchant has generally lower prices than other retailers and lower overhead expenses due to the smaller size of store and lack of large showroom space. (Wikipedia).

From the consumer's point of view, there are potential advantages and disadvantages. The catalog showroom approach allows customers to shop without having to carry their purchases throughout the store as they shop. Possible downsides include that customers may be required to give their contact information when an order is placed, take the time to fill out order forms, and wait a period of time for their order to be available for purchase. This wait may be days long, one of the chief vulnerabilities of the catalog showroom approach. Nowadays, companies are turning their print catalogs to sell their products into online catalogs. This new trend is intended for consumers to shop on line and enjoy several discounts and some other advantages. A catalog supply website includes prices, discounts, tools, shipping options, different methods of payment, and more.

Shelf

Literally marketing shelf is referred to the techniques of disposal of the goods on the shelves of shops and supermarkets. This allows has a habitual or random form of capture, getting him to buy and increase sales.

There are two components. The scientific component uses analytical and statistical studies to plan and implement strategies, pricing and customer politics. The artistic cover all those creative ideas that capture the ‘attention’ of potential customers and influence their consumer choices. Shelf placement can have an impact on the way customers perceive a brand. They expect things on top to be better in general. It can be begun to change the image of a brand just by its placement on the shelf. Also it is used to find and attract new customers to a brand. (knowledge.insead.edu)

We must consider that a company can produce and allocate their product or service in different type of market (e.g. it is possible a company produce make to order and also use a catalogue).

2.1.4. Trend analysis

Trend Analysis is the practice of collecting information and attempting to spot a pattern, or *trend*, in the information. In statistics, trend analysis often refers to techniques for extracting an underlying pattern of behavior in a time series which would otherwise be partly or nearly completely hidden by noise. (Wikipedia)

We must considerer a good indicator to have a correct evaluation of a trend analysis. For our case of study, we use a financial perceptual, using quantitative data.

Trend analysis usually involves choosing one fiscal period as a base period and then expressing subsequent quantities as a percentage of the data associated with this base period. In the case of an income statement, changes in all items could be assessed in relation to the base period. Significant changes can then be investigated further. (Texas State Auditor's Office, 1995)

Our period is one year and we consider sales and resources used in the period of time base considered.

When we talk about sales, we consider the turnover after the sale. In the same way and to simplify the case of study, we use the number of employees how resources used by the company to create these sales.

Our financial ratio is called sale-per-employee and is described in the next formula:

$$\frac{Net\ Sales}{Average\ Number\ of\ Employees}$$

This number indicates how the sales/employee ratio is calculated: a company's annual sales divided by its total employees. Annual sales and employee numbers are easily located in published statements and annual reports. While it does have its limitations, this ratio does give investors some sense of a company's productivity and financial health. The sales-per-employee ratio provides a broad indication of how expensive a company is to run. Interpreting the ratio is fairly straightforward: companies with higher sales-per-employee figures are generally considered more efficient than those with lower figures. A higher sales-per-employee ratio indicates that the company can operate on low overhead costs, and therefore do more with few employees, which often translates into healthy profits. (investopedia.com)

Companies that concentrate on selling and distributing products will typically enjoy much higher sales-per-employee figures than firms that manufacture goods. Manufacturing is typically very labor intensive, while sales and marketing activities rely on fewer people to generate the same sales numbers. In manufacturing, each employee can usually assemble only a certain number of products. Increasing production requires more employees. By contrast, marketing and sales activities can increase without necessarily adding staff. We should also be careful about employee numbers stated in the financial reports. Some companies employ sub-contractors, which are not counted as employees. This kind of discrepancy can put a wrinkle in our analysis and comparison of sales-per-employee figures.

For our analysis we divide the report in 3 types of companies depending the number of this value. We see that there are companies which use inefficiently their resources, companies which use their resources in a normal way and at the end, we distingue companies which use their resources more efficiently. The limits to separate these 3 types are:

- If Ratio €[0,20): company is inefficient
- If Ratio €[20,50): company efficient
- If Ratio €[50,infinity): company huge efficient

We can see in the next graphics how the tendency of the different companies is:

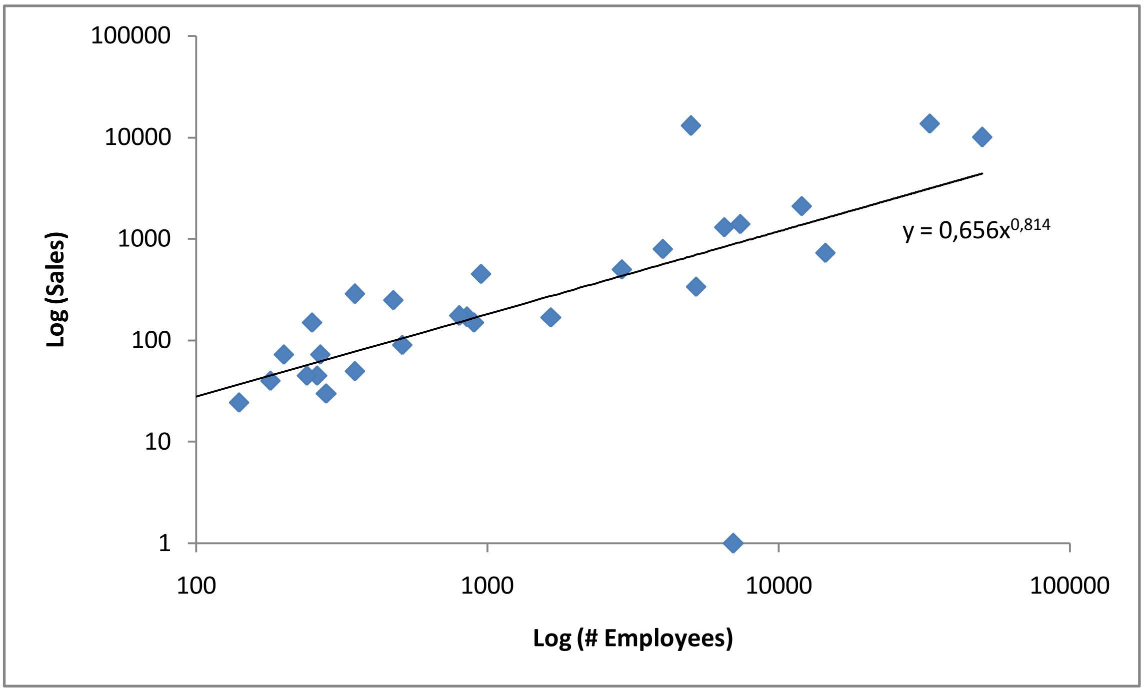

Figure 16. Sample-distribution of log sales (y) and log employees (x)

2.2. Analysis

As already introduced, the data are mostly collected thanks to face to face meetings; also phone interviews and "not-simultaneous" options are possible, even if effectiveness (in terms of understanding the real scope of each question) and involvement (interviews feeling to be part of the project) can decrease a bit.

Interviewed companies can be SME's or multinational companies, and can belong to no matter what technological sector; the only "must" is to have the new product development process inside the company -or at least in one site of the company group- about which the interviewed is very well aware. The interviewed sample is described with some graphs shown in the following: the 30 companies have a huge variety in terms of number of employees (Figure 15), and there is kind of linear distribution between sales and number of employees within the sample, even if some outlier exist (Figure 16).

For our analysis we will use the average and the deviation standard.

At first we study the average for all companies looking for correlations and similitude in companies for the different stage that we can analyze with the questionnaire.

Then we separate companies in the different sectors that we have defined early and we will try to find if there are some correlations in these dates for the different stages that we divide the company.

2.2.1. Average

2.2.1.1. Average for all companies

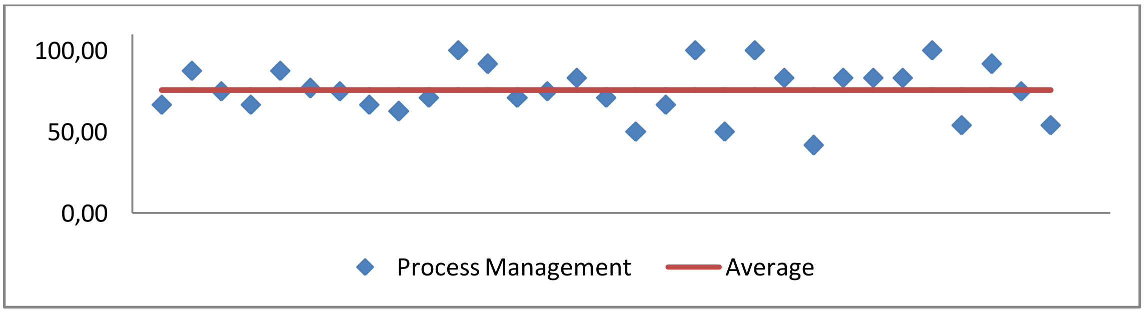

Figure 17. Process Management Area, whole sample with respect of the global average GA (straight line)

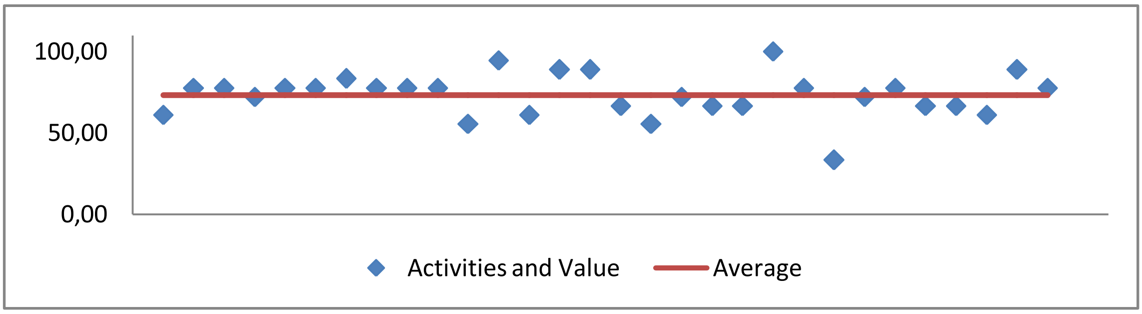

Figure 18. Activities and Value Area, whole sample with respect of the GA

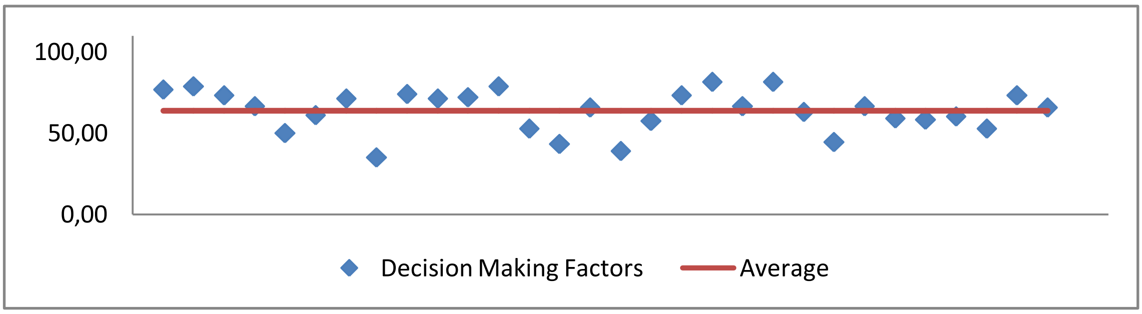

Figure 19. Decision Making Factors Area, whole sample with respect of the GA

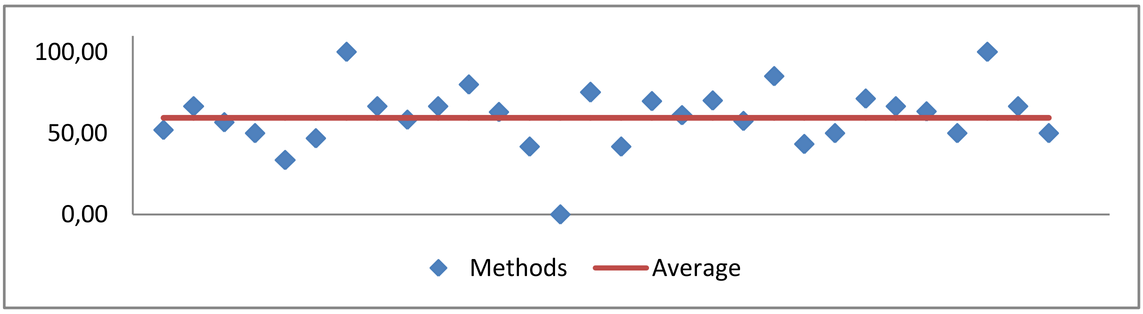

Figure 20. Methods Area, whole sample with respect of the GA

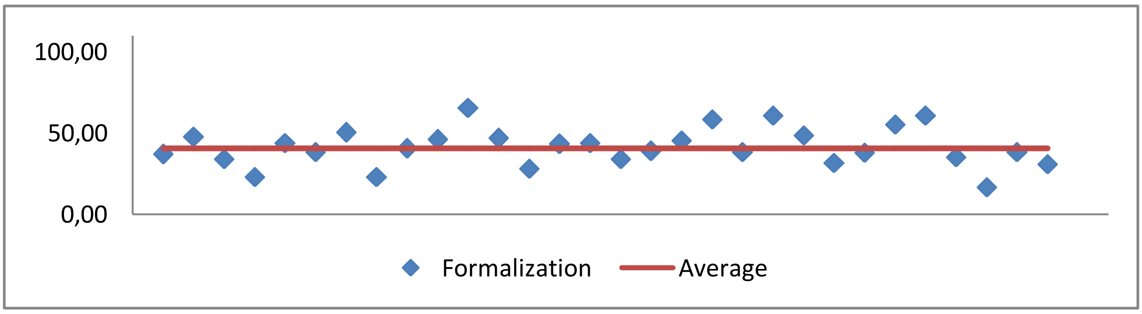

Figure 21. Formalization Area, whole sample with respect of the GA

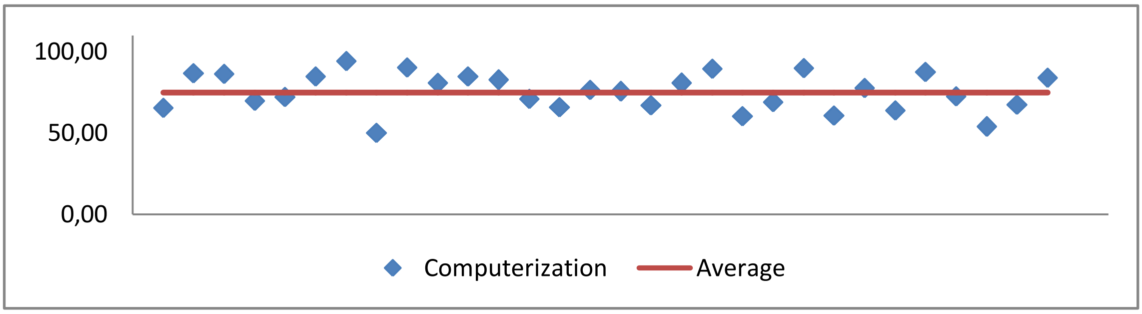

Figure 22. Computerization Area, whole sample with respect of the GA

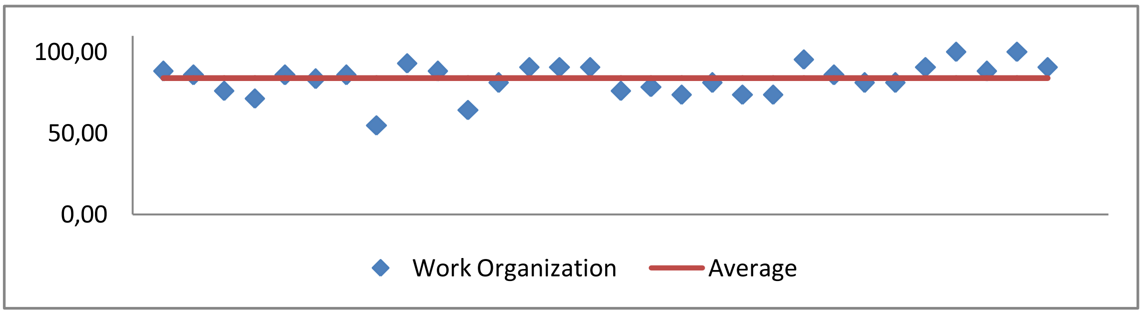

Figure 23. Work Organization Area, whole sample with respect of the GA

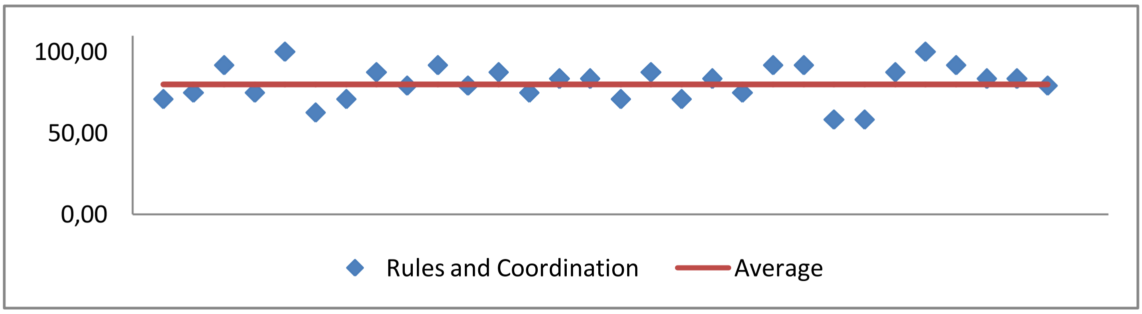

Figure 24. Rules and Coordination Area, whole sample with respect of the GA

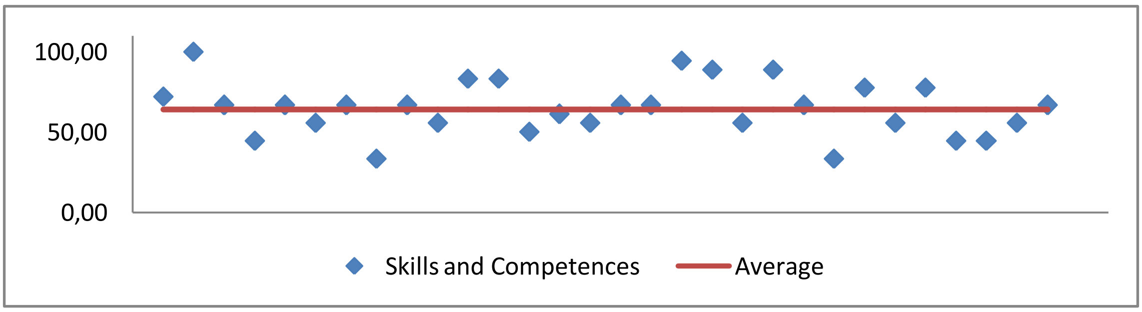

Figure 25. Skills and Competences Area, whole sample with respect of the GA

Summary of Standard Deviation for all companies

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	75,59	73,30	63,86	59,24	40,66
Standard Deviation	15,73	13,02	12,38	19,70	12,08

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	74,65	83,79	79,97	64,16
Standard Deviation	11,95	10,24	11,71	17,08

Table 3. Summary of Standard Deviation for all companies

The higher level of maturity is achieved in the Organization part, area of Work Organization, in which average scores overpass the 80%. In Italian companies, designers assume relevant importance within NPD, and the coordination and cooperation between them imply the goodness of the work environment. The dispersion of values in this area is small, in fact only few cases attest in Intermediate level of maturity (Figure 23). The other two areas of Work Organization present different values. While Rules and Coordination has a good value (Figure 24), near 80%, the other one Skills and Competence presents low values. The dispersion of values in this area is big, in fact many cases attest in Intermediate level of maturity (Figure 25). Even if companies understand that enhancement of individual skills and competences results in a more agile and mature organization, with better product performance, the achieved maturity level is still Intermediate and, in few cases, Low.

Interviewed companies demonstrate to be Mature in performing their NPD process, which is often realized through formalized steps, activities, and tasks, accomplished in order to define the specifications the new products will satisfy. Companies are aware that the strict control of the NPD process is crucial, such as its continuous monitoring and improvement. In fact, they attest around 75% in Process Management (Figure 17), even if some companies result to be less structured and still Intermediate. As above mentioned, the NPD process can be supported by a huge variety of tools and methods (Methods), such as Design for X techniques, Design To Cost, etc. The assessed sample is in the border line between Intermediate and Mature, while some companies are even Low and Chaos (Figure 20). Standard Deviation is very high in this area, around 20%, showing this dispersion between different Italian companies.

Moreover, the NPD process requires a large number of decisions to be taken every day considering both internal (Decision Making Factors) and external (Activities and Value) factors. Interviewed companies are actually paying relevant attention on customer value (Figure 18), except few Intermediate cases and one Low. In average, lower attention is paid on the whole product lifecycle, which not heavily affects companies' strategy and early decisions. This is demonstrated by the 63, 86% average level gained by the sample (Figure 19): most of the companies seem to be not aware of the simple fact that the decisions taken at the early NPD stages affect the whole product lifecycle. The lower level of maturity corresponds to the area Decision Making Factors. But also we can see that there are a lot of companies that follow this tendency because standard deviation is very low.

The third investigated part is Knowledge Management, which results to be still precarious. In fact, even if the level of Computerization is in average Mature (Figure 22), the Formalization is still in the middle between Low and Intermediate (Figure 21). Companies know that maintaining and protecting the internal know-how is crucial. Every day a "piece of knowledge" is created, shared, retrieved, and displayed, while huge amount of data must be handled to improve NPD efficiency.

In order to preserve data, these should be formalized and represented in a way understandable by each practitioner, and easy to be re-used. To obtain these results PLM (Product Lifecycle Management) / PDM (Product Data Management) software are suitable to be implemented. But even if software and IT tools are quite diffuse within companies and the level of computerization is quite high, the communication between different software is critical and proper utilization of PDM / PLM softwares is still missing. Companies are sometimes not even aware about the meaning and benefits of PDM / PLM systems.

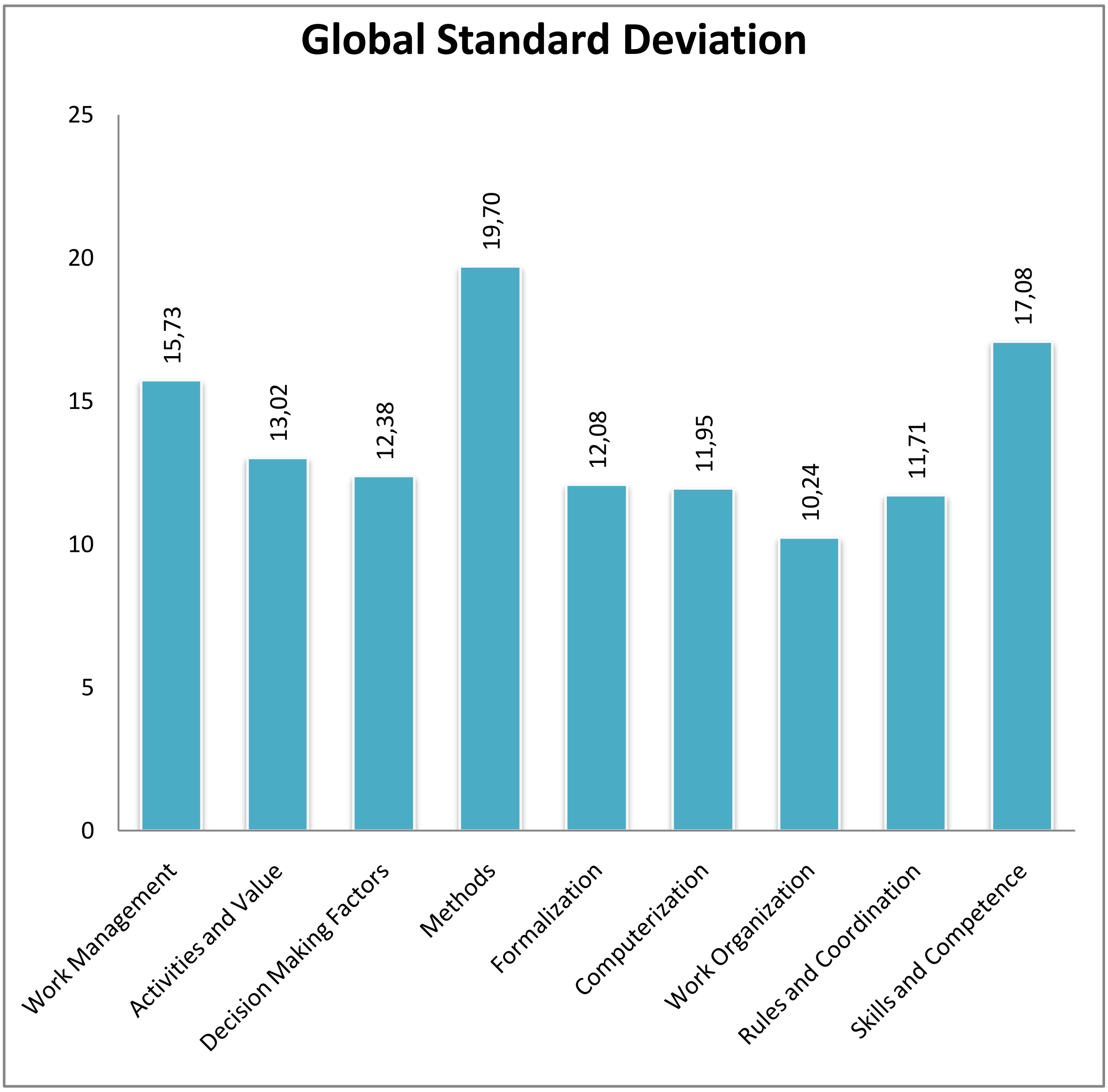

Figure 26. Global standard deviation for all companies

2.2.1.2. Average and Standard Deviation depending on the size of the company

Little companies

In this particular case we have only one company and it is not important analyze it. In spite of we show a table with the average and the standard deviation.

Summary

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	50,00	66,67	66,67	57,58	38,25
Standard Deviation	0	0	0	0	0

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	60,43	73,81	75,00	55,56
Standard Deviation	0	0	0	0

Table 4. Summary of Standard Deviation for the little company

Medium Companies

We use for the next analysis the average for each area and also the global average. In this way we see if the division can allow us increase the stage or otherwise decrease this area of study.

If the red line is on the green line, this represents that an increasing.

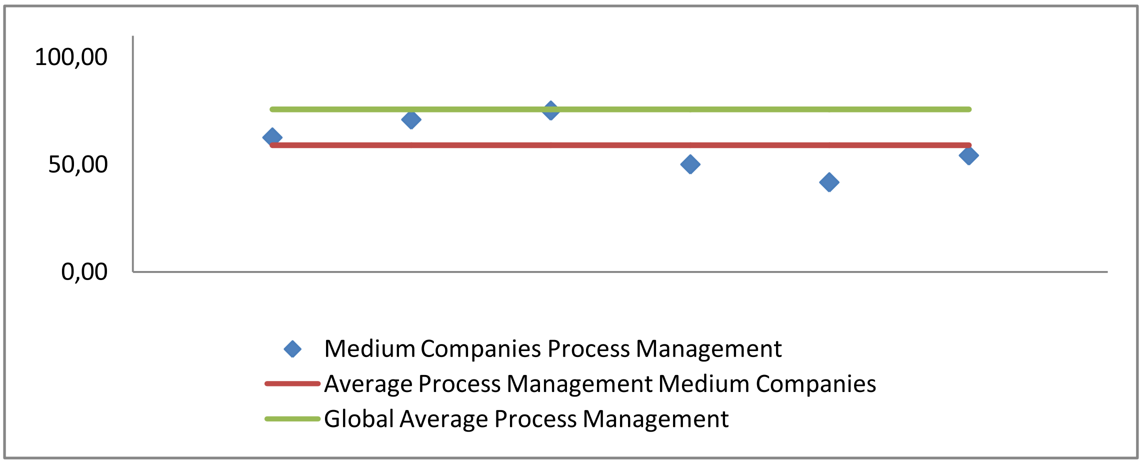

Figure 27. Process Management Area, whole sample with respect of the GA

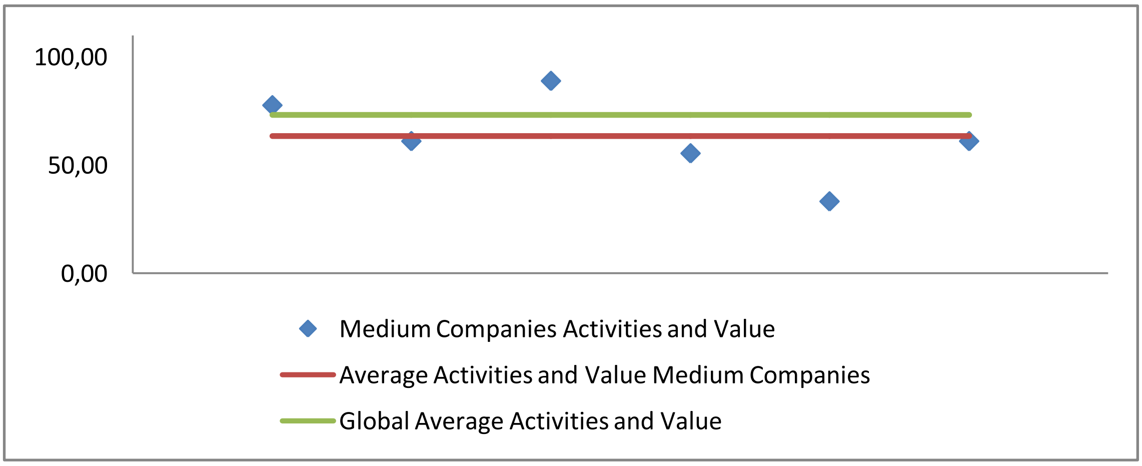

Figure 28. Activities and Value Area, whole sample with respect of the GA

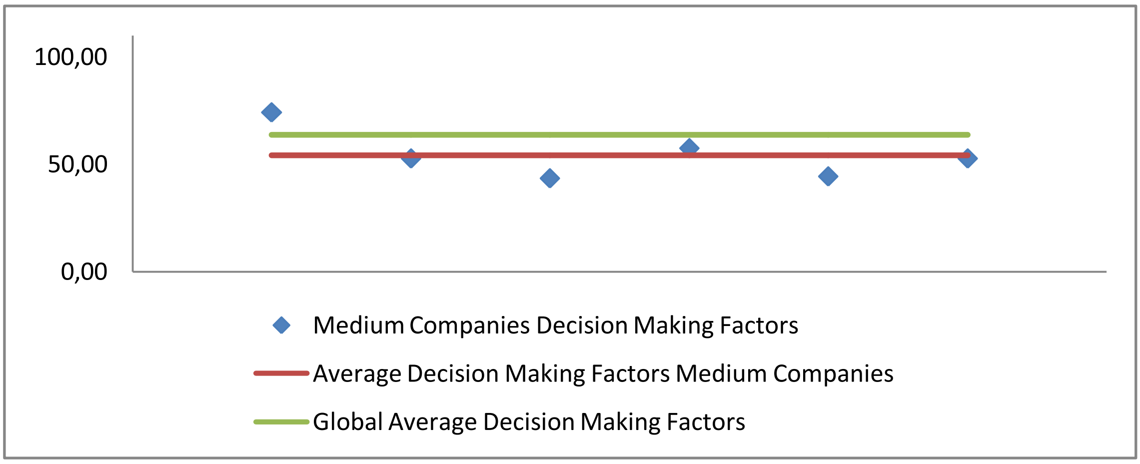

Figure 29. Decision Making Factors Area, whole sample with respect of the GA

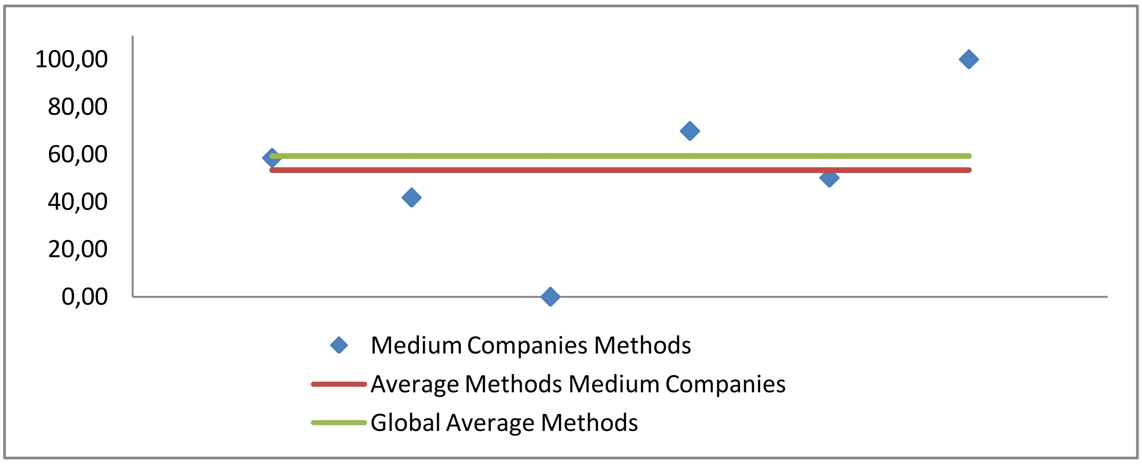

Figure 30. Methods Area, whole sample with respect of the GA

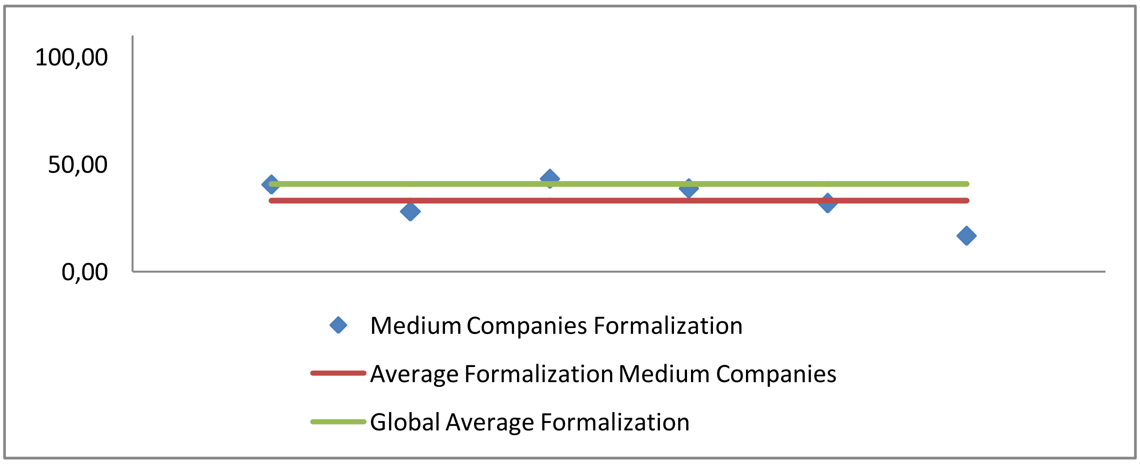

Figure 31. Formalization Area, whole sample with respect of the GA

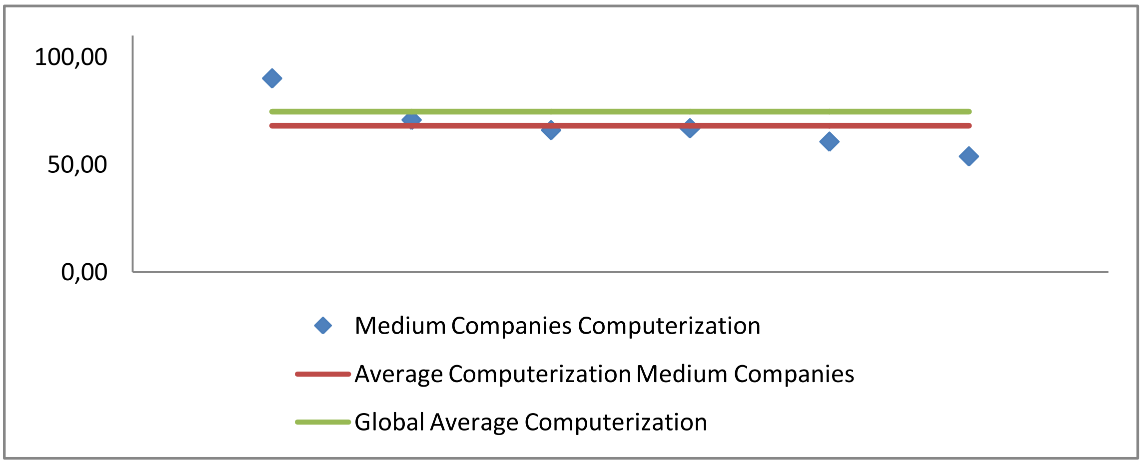

Figure 32. Computerization Area, whole sample with respect of the GA

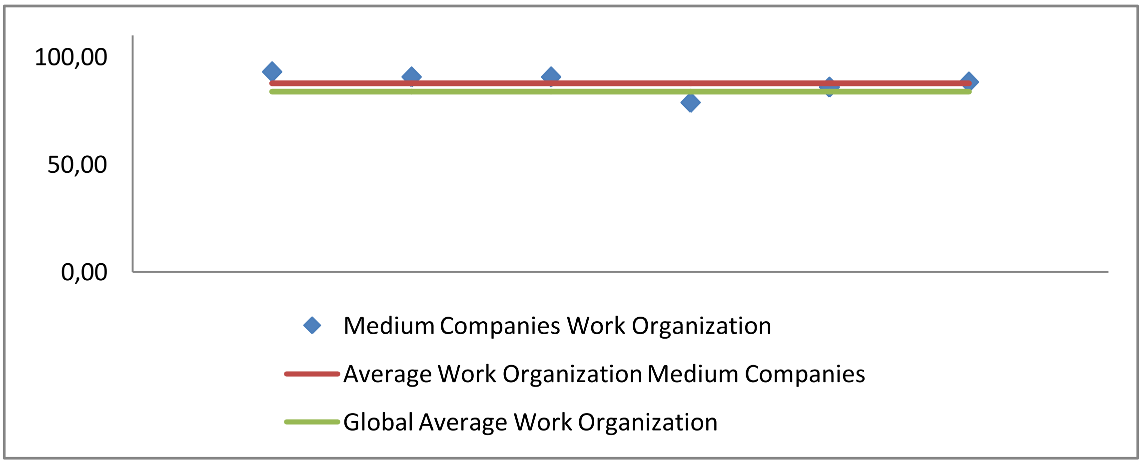

Figure 33. Work Organization Area, whole sample with respect of the GA

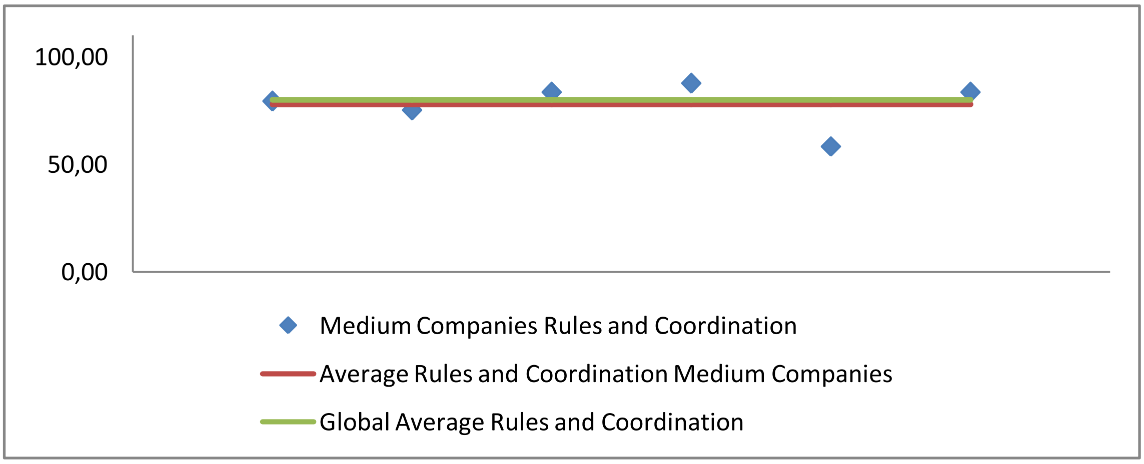

Figure 34. Rules and Coordination Area, whole sample with respect of the GA

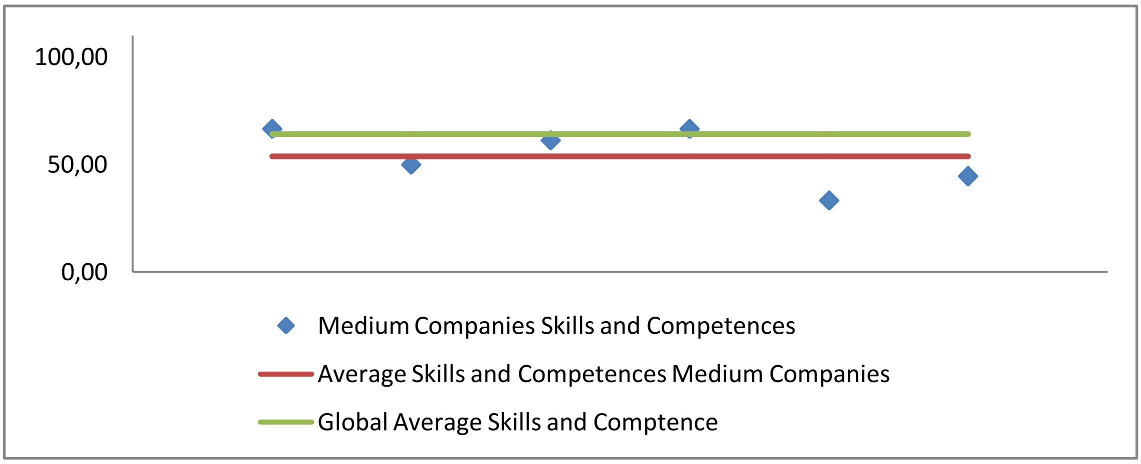

Figure 35. Skills and Competence Area, whole sample with respect of the GA

We can see in medium companies, most of case decrease the average in spite of GA. It is a worst case and this companies decrease the level of each stage.

Big Companies

Also in this case we use the average for each area and also global average. In this way we see if the division can allow us increase the stage or otherwise decrease this area of study.

How in the before case, if the red line is on the green line, this represents that an increasing.

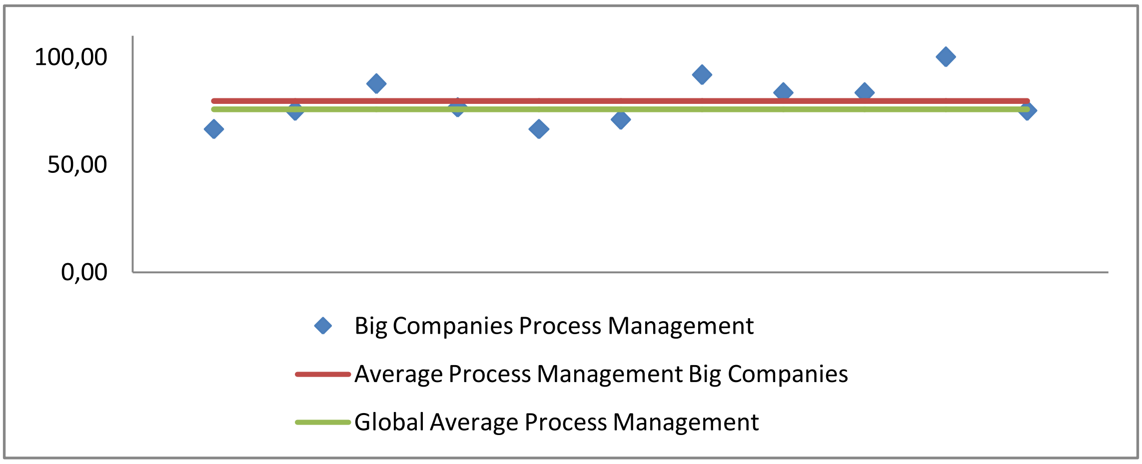

Figure 36. Process Management Area, whole sample with respect of the GA

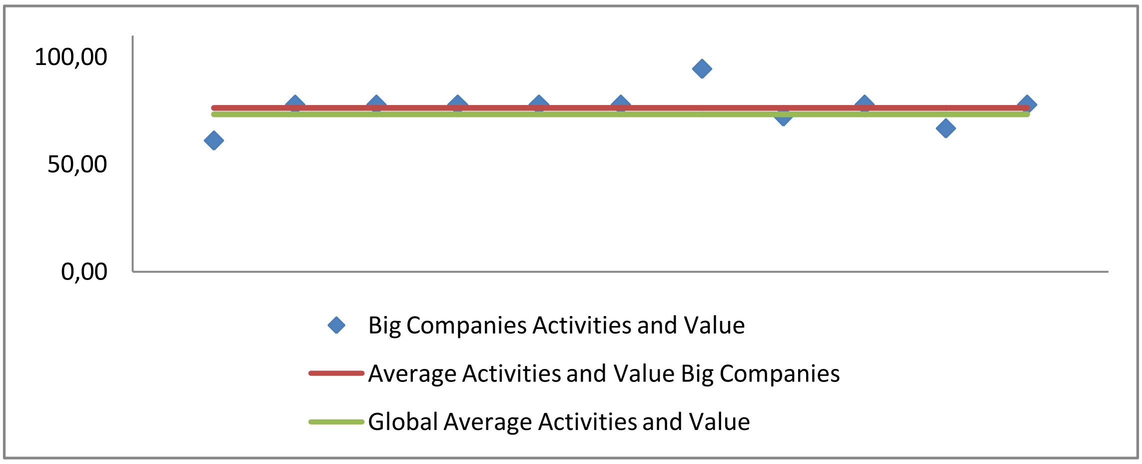

Figure 37. Activities and Value Area, whole sample with respect of the GA

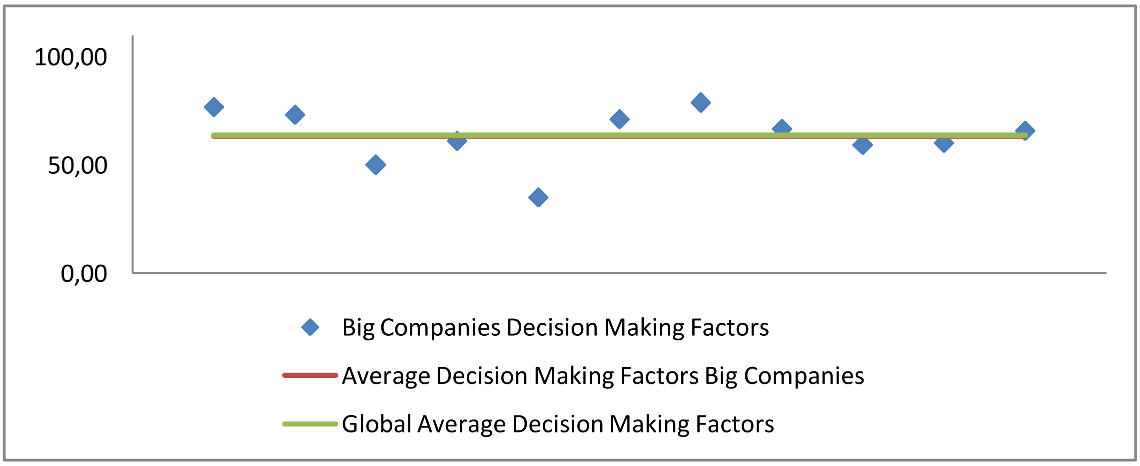

Figure 38. Decision Making Factors Area, whole sample with respect of the GA

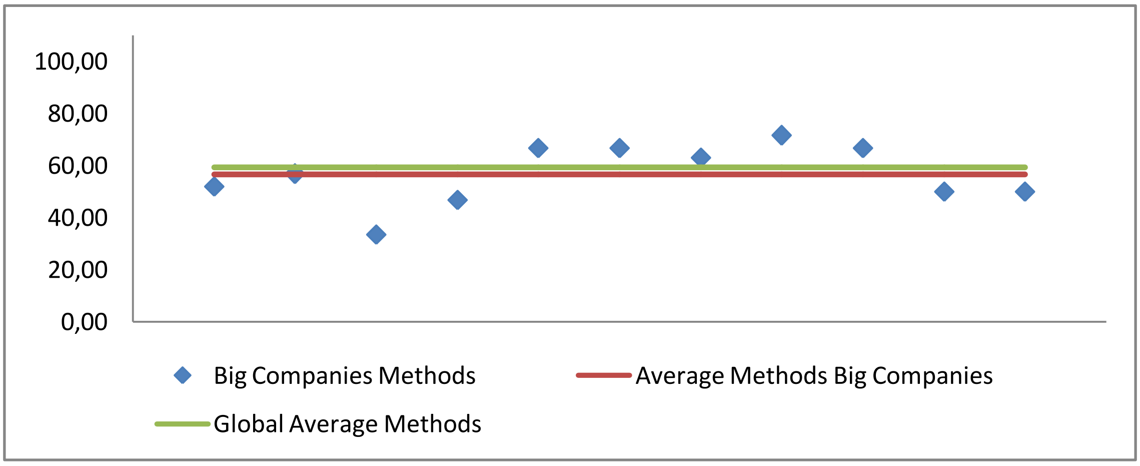

Figure 39. Methods Area, whole sample with respect of the GA

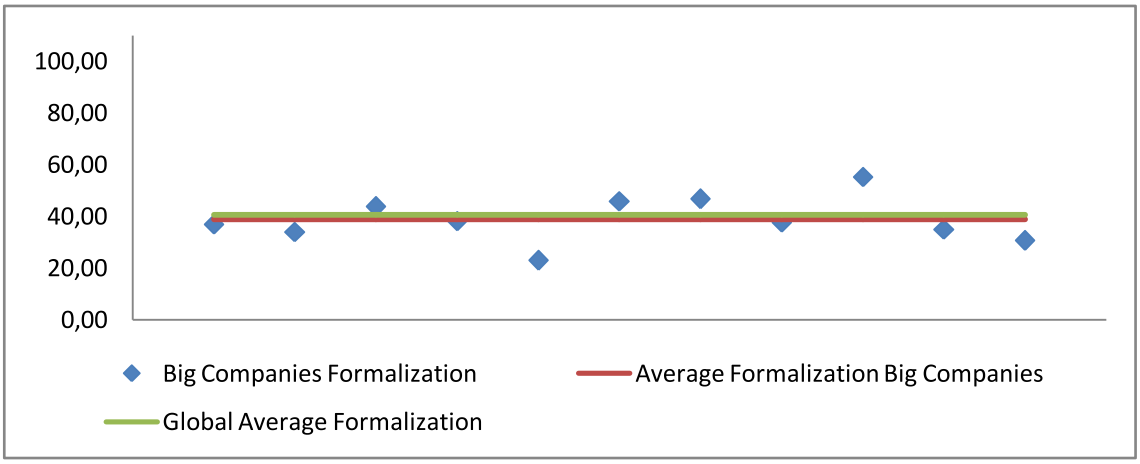

Figure 40. Formalization Area, whole sample with respect of the GA

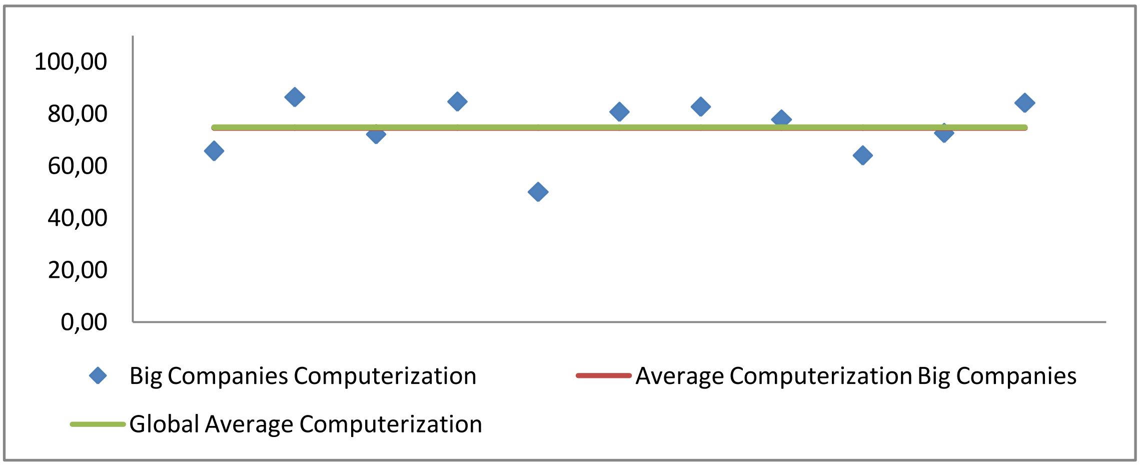

Figure 41. Computerization Area, whole sample with respect of the GA

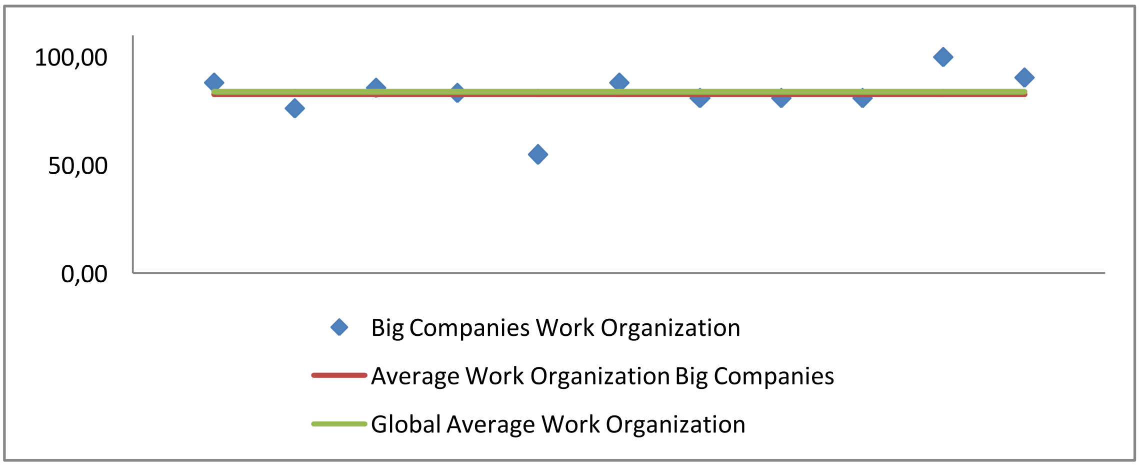

Figure 42. Work Organization Area, whole sample with respect of the GA

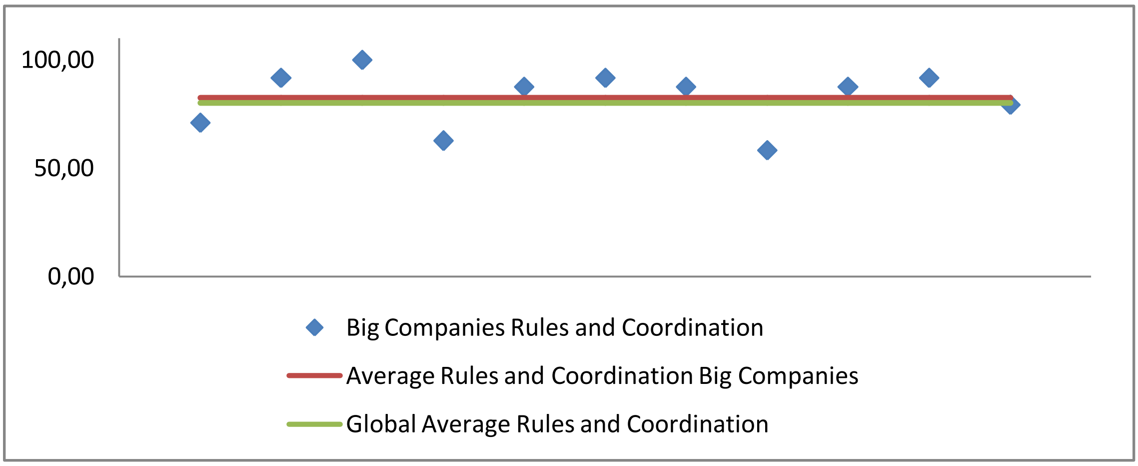

Figure 43. Rules and Coordination Area, whole sample with respect of the GA

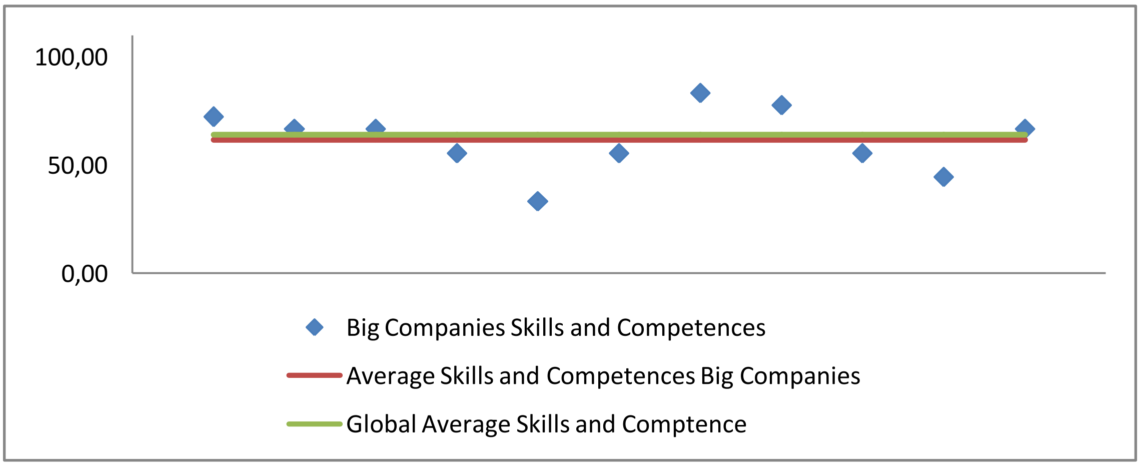

Figure 44. Skills and Competences Area, whole sample with respect of the GA

For this case of study we can see that there isn't a relevant improvement. Each area is different for the others and in some case there is an increasing in spite of GA and in other cases not.

Macro Companies

Also in this case we use the average for each area and also global average. In this way we see if the division can allow us increase the stage or otherwise decrease this area of study.

How in the before case, if the red line is on the green line, this represents that an increasing.

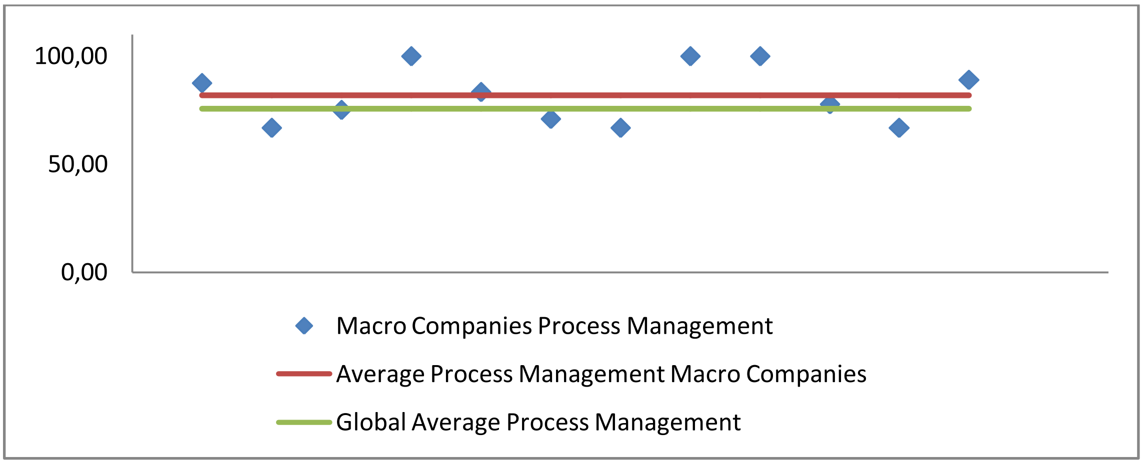

Figure 45. Process Management Area, whole sample with respect of the GA

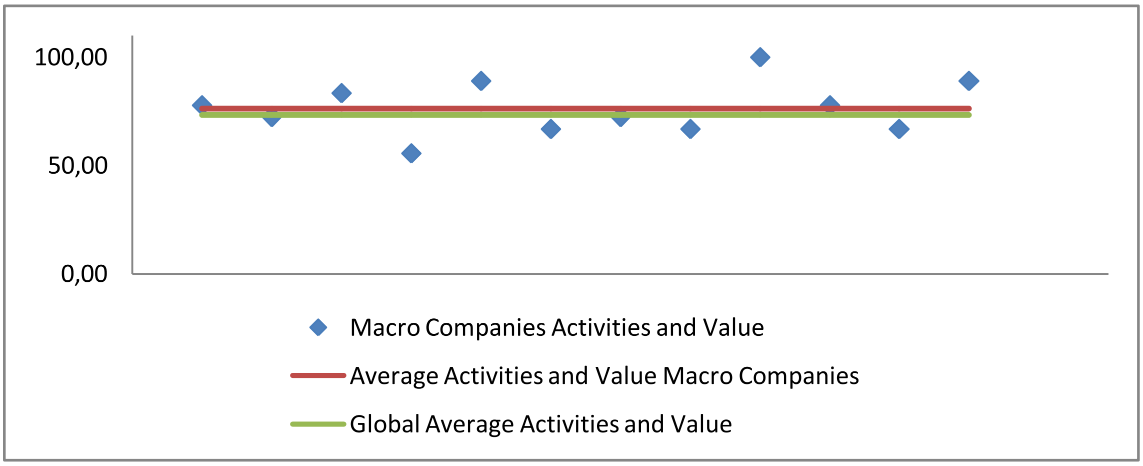

Figure 46. Activities and Value Area, whole sample with respect of the GA

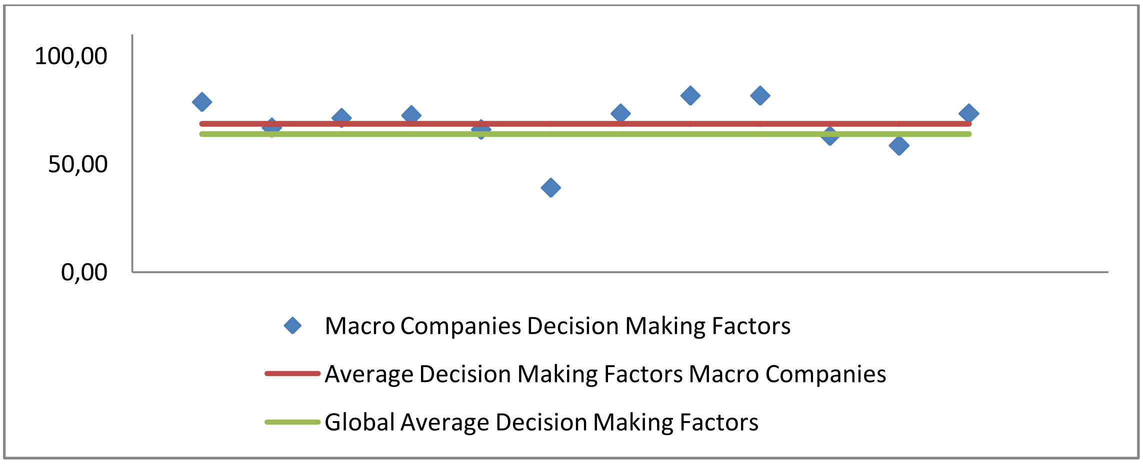

Figure 47. Decision Making Factors Area, whole sample with respect of the GA

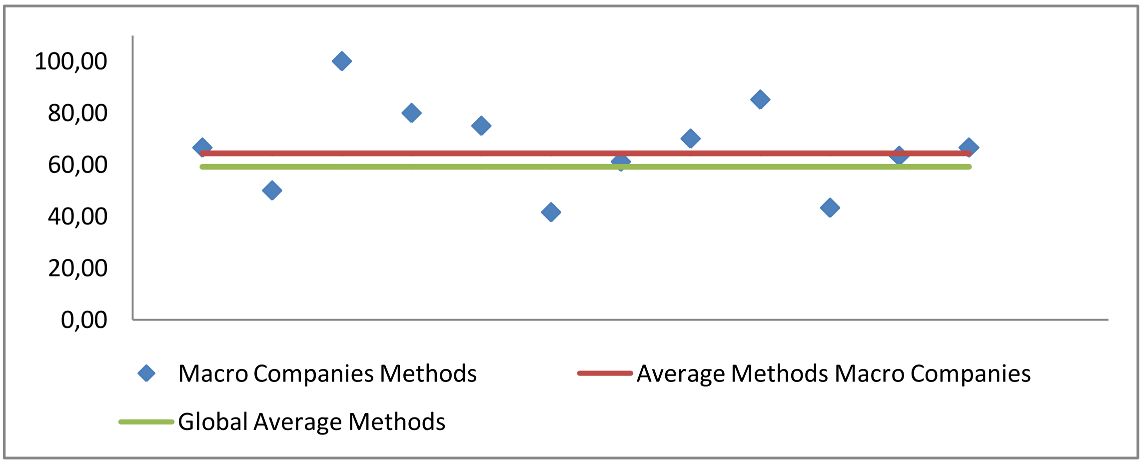

Figure 48. Methods Area, whole sample with respect of the GA

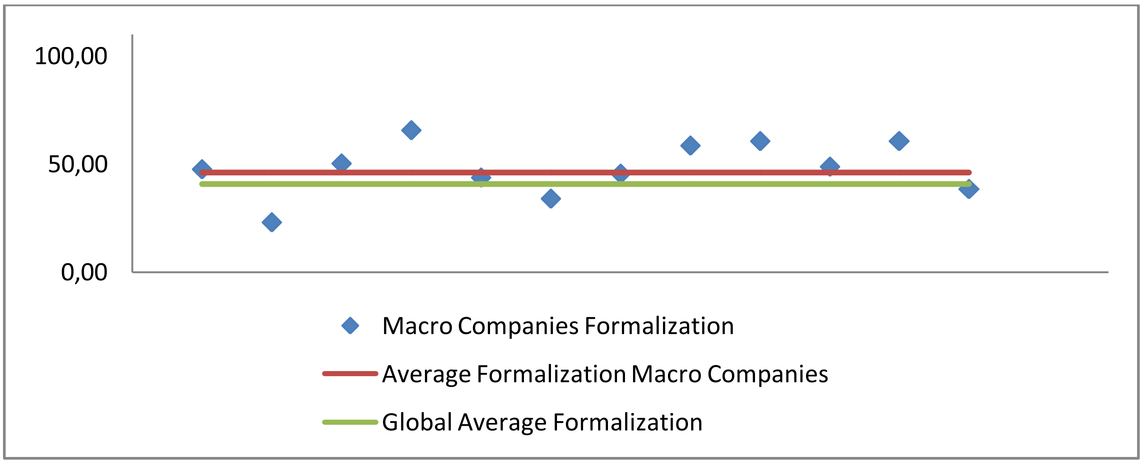

Figure 49. Formalization Area, whole sample with respect of the GA

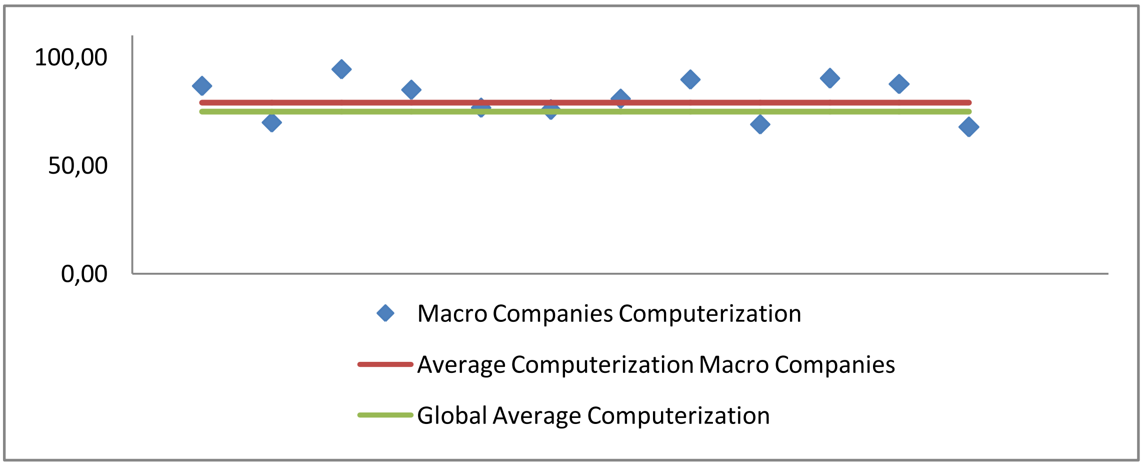

Figure 50. Computerization Area, whole sample with respect of the GA

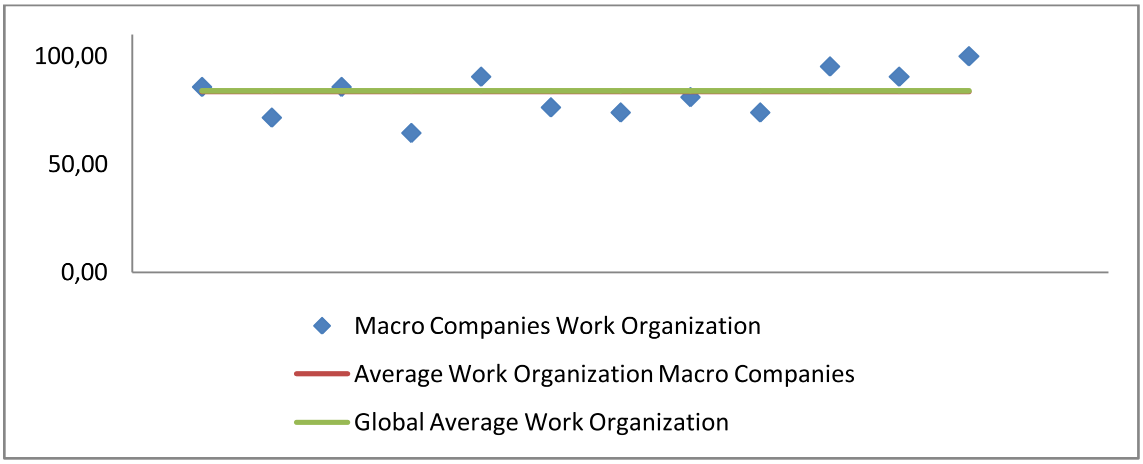

Figure 51. Work Organization Area, whole sample with respect of the GA

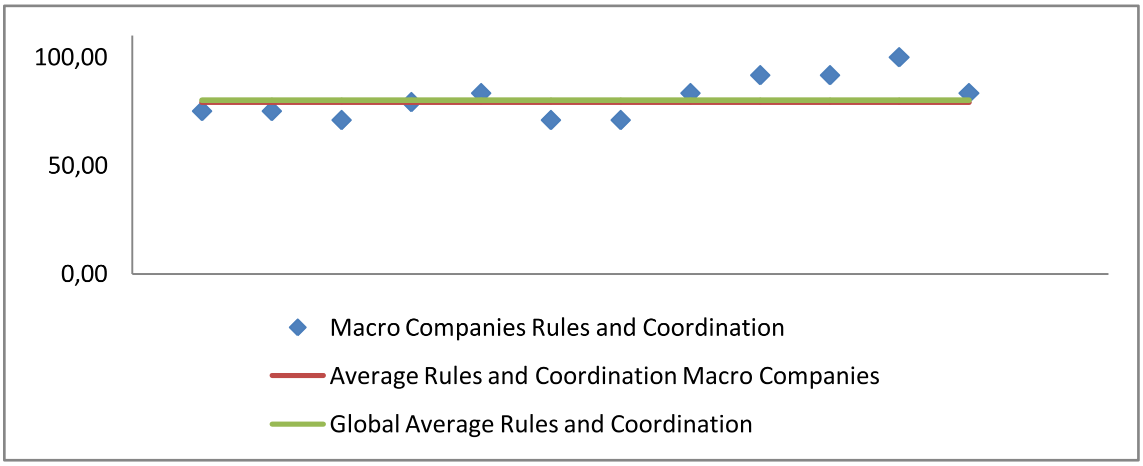

Figure 52. Rules and Coordination Area, whole sample with respect of the GA

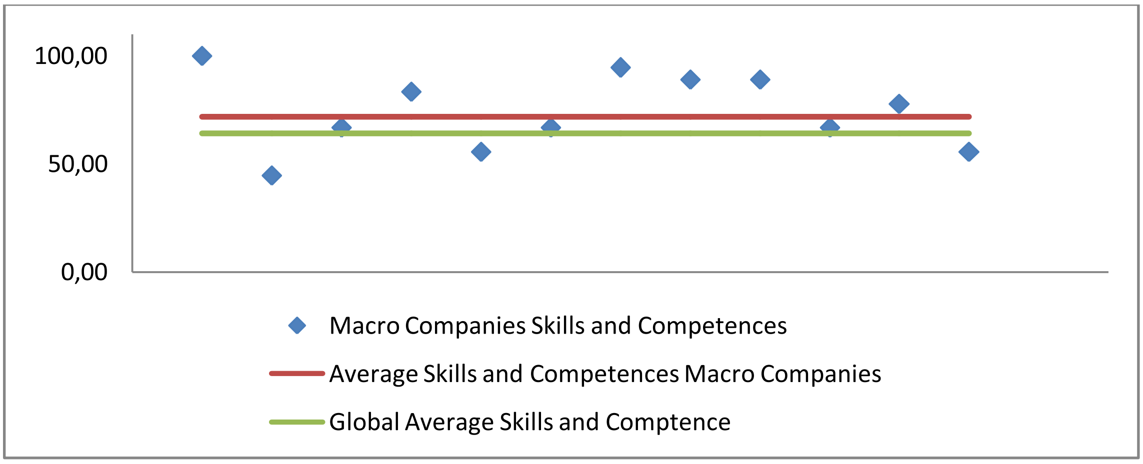

Figure 53. Skills and Competences Area, whole sample with respect of the GA

We can see in macro companies, most of case increase the average in spite of GA.

Standard Deviation

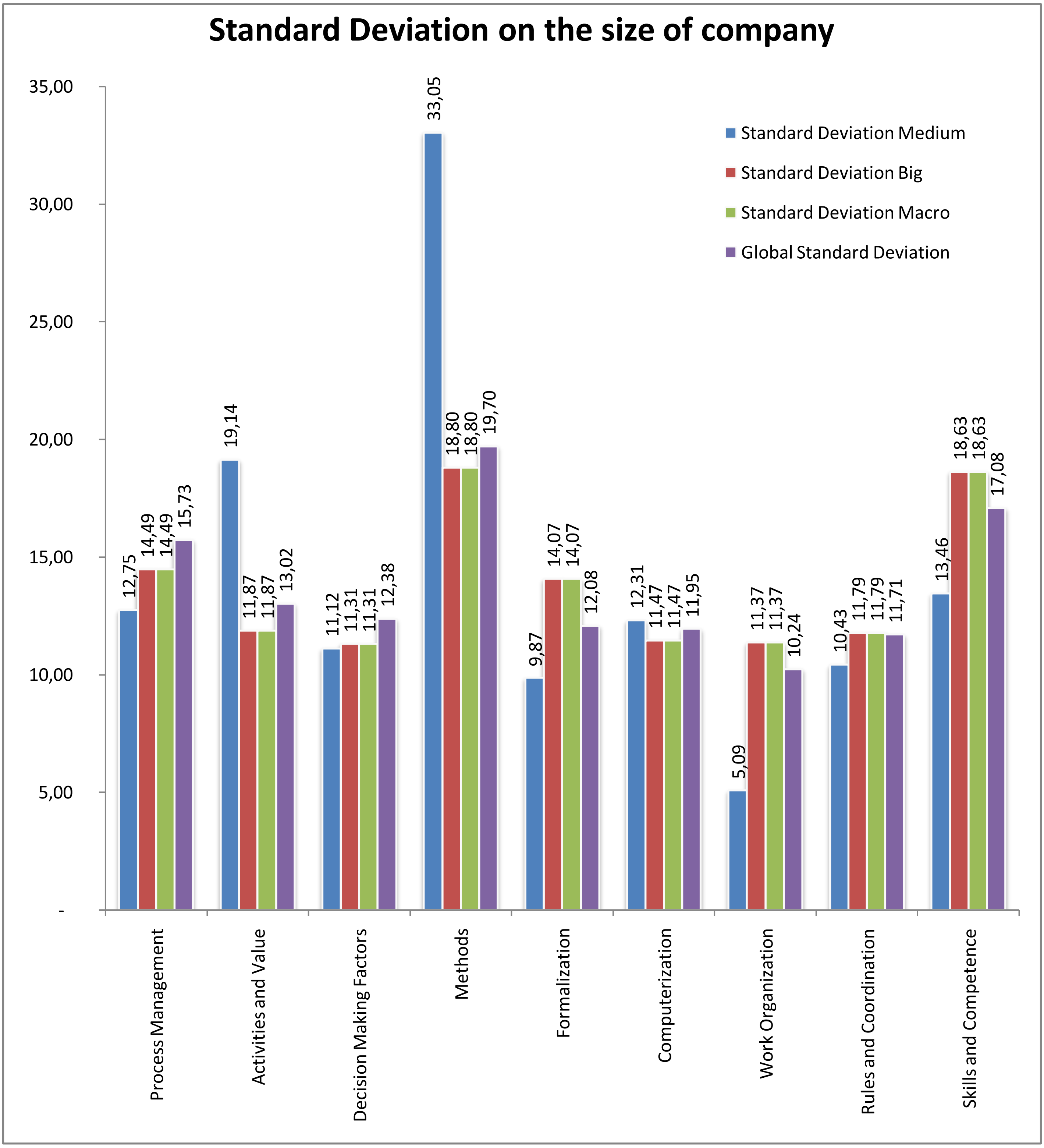

Figure 54. Standard deviation depending of the size of the companies

Summary Medium

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	59,03	62,96	54,17	53,28	33,11
SD	12,75	19,14	11,12	33,05	9,87

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	68,02	87,70	77,78	53,70
SD	12,31	5,09	10,43	13,46

Table 5. Summary of Standard Deviation for the medium company

Summary Big

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	79,70	76,26	63,47	56,63	38,80
SD	10,56	8,27	12,64	11,46	8,77

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	74,56	82,68	82,58	61,62
SD	11,17	11,22	13,28	14,58

Table 6. Summary of Standard Deviation for the big company

Summary Macro

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	81,73	76,07	68,45	64,33	45,90
SD	14,49	11,87	11,31	18,80	14,07

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	78,88	83,70	79,17	71,79
SD	11,47	11,37	11,79	18,63

Table 7. Summary of Standard Deviation for the macro company

The higher level of maturity is achieved in the Organization part, area of Work Organization, in which average scores overpass the 80% for medium, big and macro companies. The dispersion of values in this area is small, in fact we can see that in the case of medium companies standard deviation is around 5% (Table 5). The other two areas of Work Organization present different values. While Rules and Coordination has a good value (Figure 34, 43 and 52), near 80%, the other one Skills and Competence presents low values (Figure 35,44 and 53). The dispersion of values in this area is bigger than the other two areas, in fact many cases attest in Intermediate level of maturity (Table 5,6 and 7). It is important see that behavior is similar with the global report. Some macro companies are interested in the skills and competence of their employees and perhaps this value is a little bigger than medium and big companies, and for that it can be possible find some cases with mature practice. For the other case even if companies understand that enhancement of individual skills and competences results in a more agile and mature organization, with better product performance, the achieved maturity level is still Intermediate and, in few cases, Low.

If we study Process Management area we can see that there is a different between medium and big and macro companies. In the second case they attest around 80% in Process Management (Figure 36 and 45), even if some companies result to be less structured and still Intermediate. But in the first case (Figure 27), for medium companies the value is worst and we only obtain a result around 50%. Standard

Deviation is higher in this area for all type of companies, around 13%, showing this dispersion between different Italian companies.

Moreover, the NPD process requires a large number of decisions to be taken every day considering both internal (Decision Making Factors) and external (Activities and Value) factors. Interviewed medium companies are actually paying less attention on customer value (Figure 28). In the case of big and macro companies this value is bigger (Figure 37 and 46). In average, lower attention is paid on the whole product lifecycle, which not heavily affects companies' strategy and early decisions. This is demonstrated by the 54.17% in medium companies and 63, 47% and 68, 45 % in cases of big and macro companies (Figure 29, 38 and 47): most of the companies seem to be not aware of the simple fact that the decisions taken at the early NPD stages affect the whole product lifecycle. The lower level of maturity corresponds to the area Decision Making Factors. In spite of, standard deviation is very high.

The third investigated part is Knowledge Management, which results to be still precarious. In fact, even if the level of Computerization is in average Mature (Figure 32, 41 and 50), There is a different between medium and big and macro companies. The level of maturity in the first case is worst with the regard to big and macro companies. Formalization is still in the middle between Low and Intermediate (Figure 31, 40 and 49). Companies know that maintaining and protecting the internal know-how is crucial. For medium and big companies the behavior is similar between diverse companies. We can say that because SD is not very high. But in the case of macro companies, if the average is better than other type of companies, SD is higher, so the behavior is not representative for all the cases.

2.2.2.Average and SD depending on the sector

Manufacture of metal tools sector

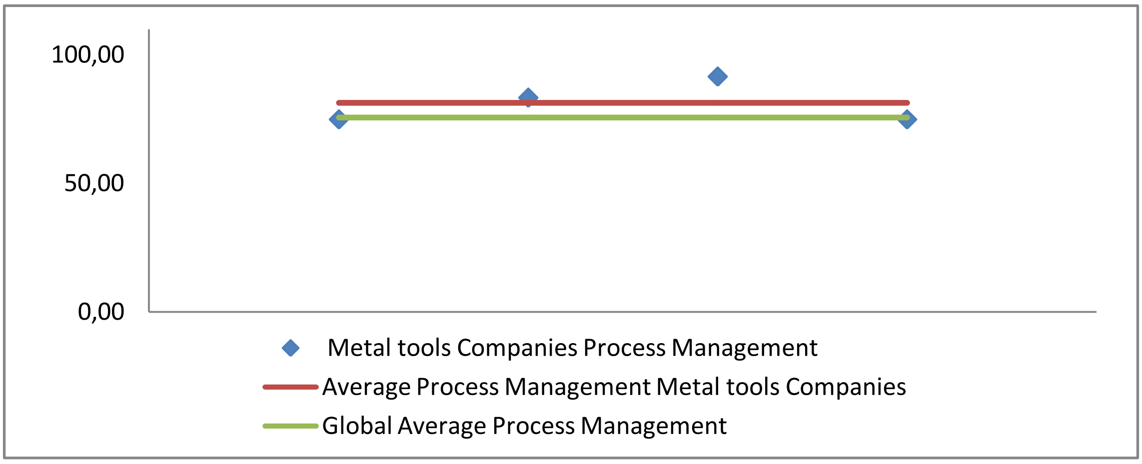

Figure 55. Process Management Area, whole sample with respect of the GA

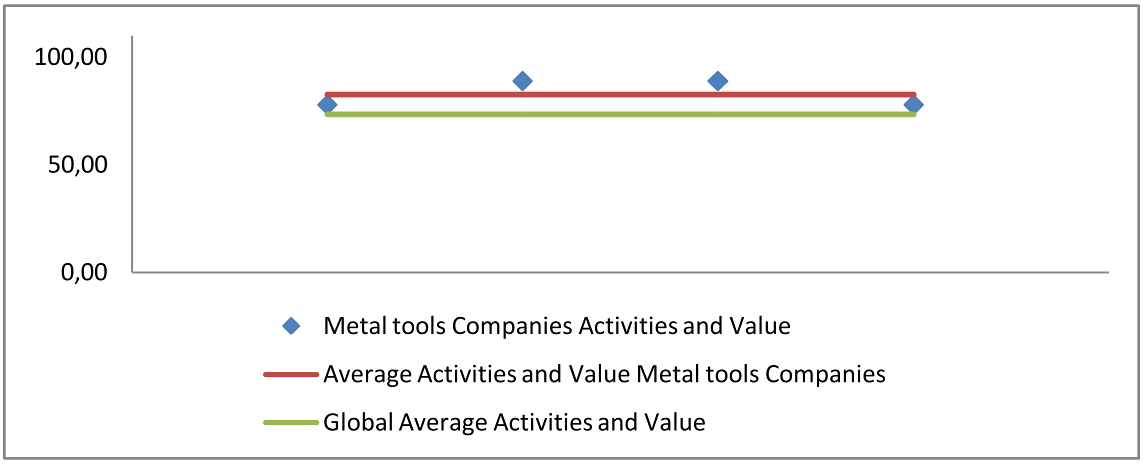

Figure 56. Activities and Value Area, whole sample with respect of the GA

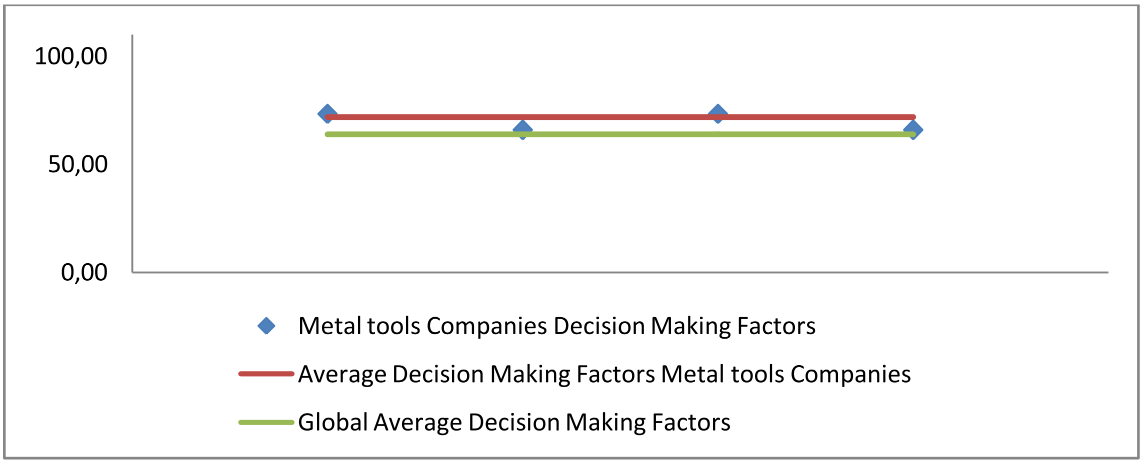

Figure 57. Decision Making Factors Area, whole sample with respect of the GA

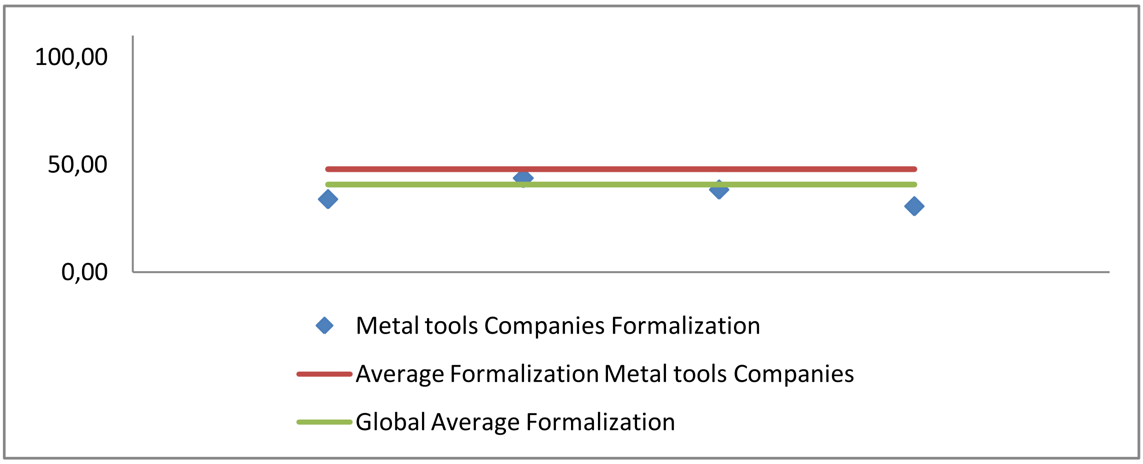

Figure 58. Formalization Area, whole sample with respect of the GA

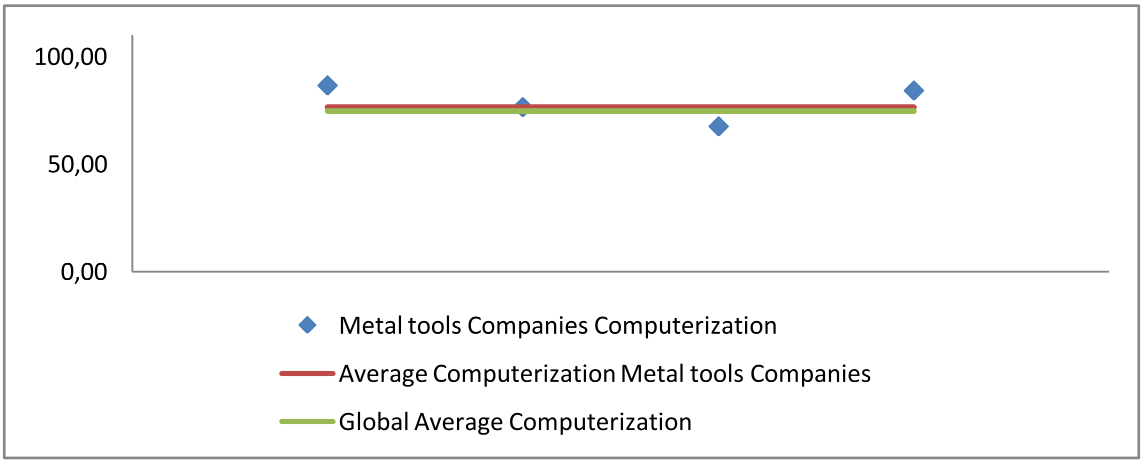

Figure 59. Computerization Area, whole sample with respect of the GA

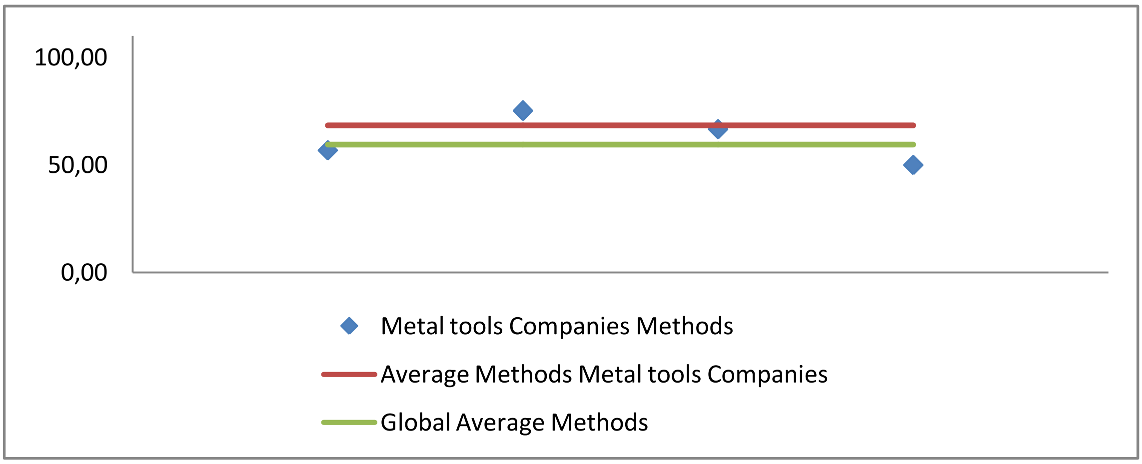

Figure 60. Methods Area, whole sample with respect of the GA

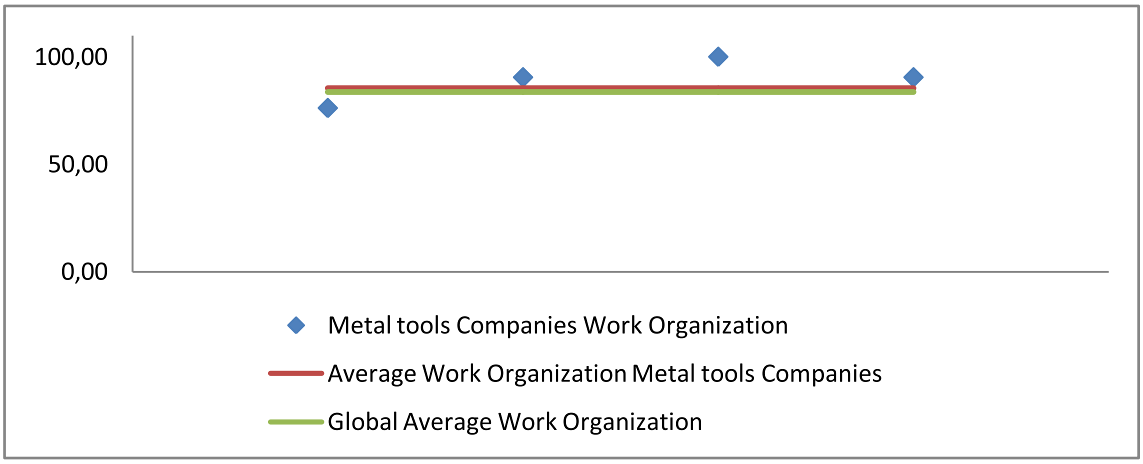

Figure 61. Work Organization Area, whole sample with respect of the GA

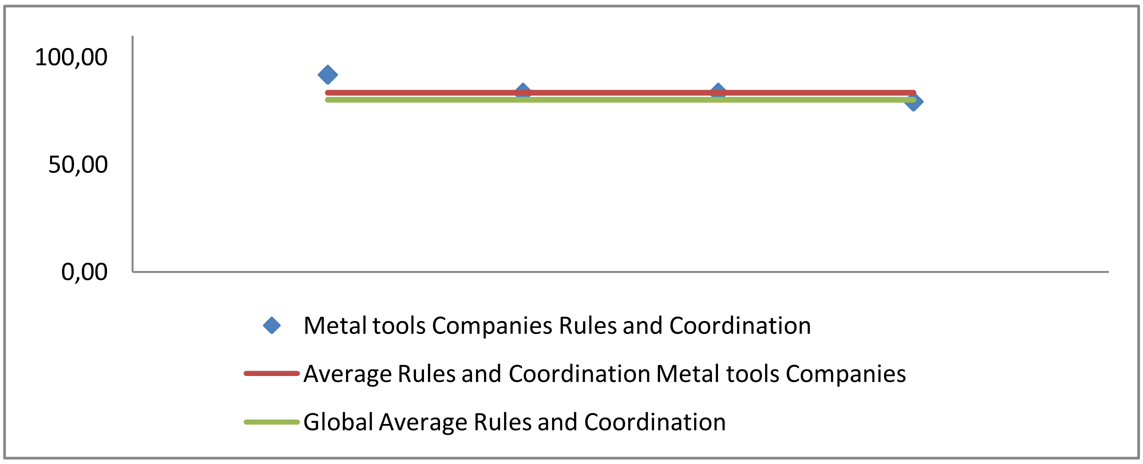

Figure 62. Rules and Coordination Area, whole sample with respect of the GA

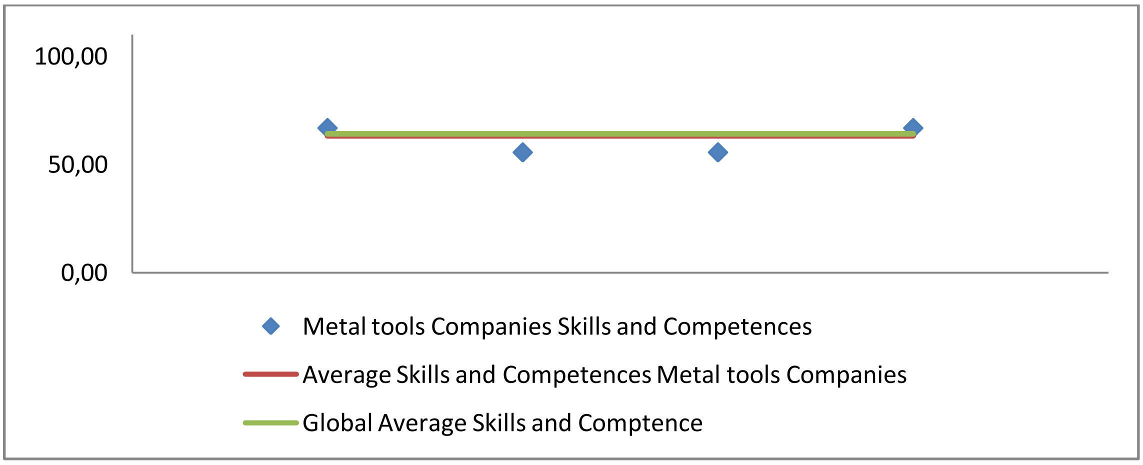

Figure 63. Skills and Competences Area, whole sample with respect of the GA

Manufacture of electronic components sector

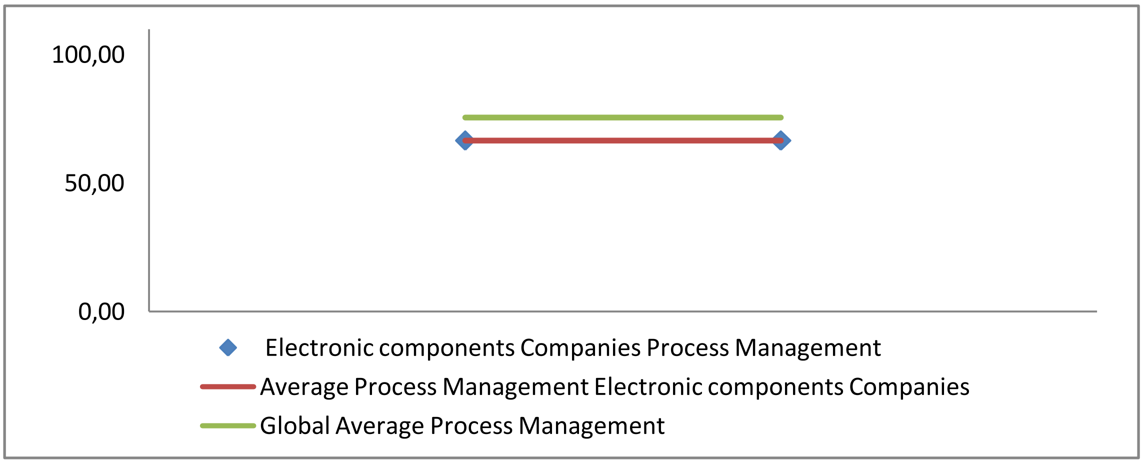

Figure 64.Process Management Area, whole sample with respect of the GA

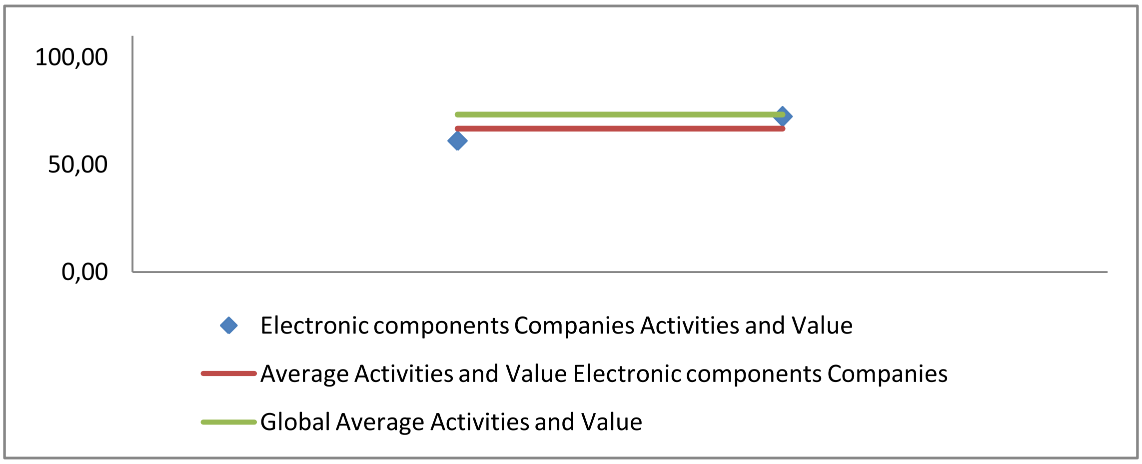

Figure 65. Activities and Value Area, whole sample with respect of the GA

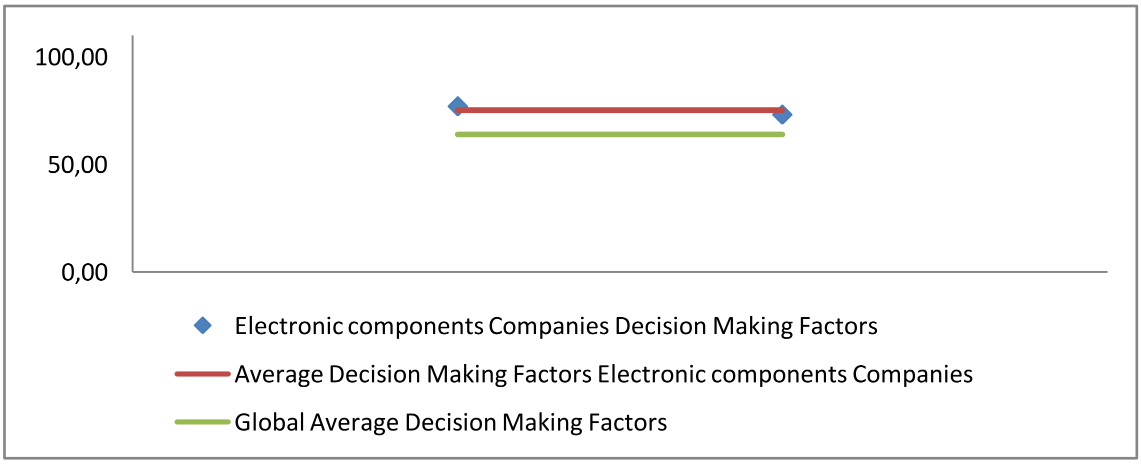

Figure 66. Decision Making Factors Area, whole sample with respect of the GA

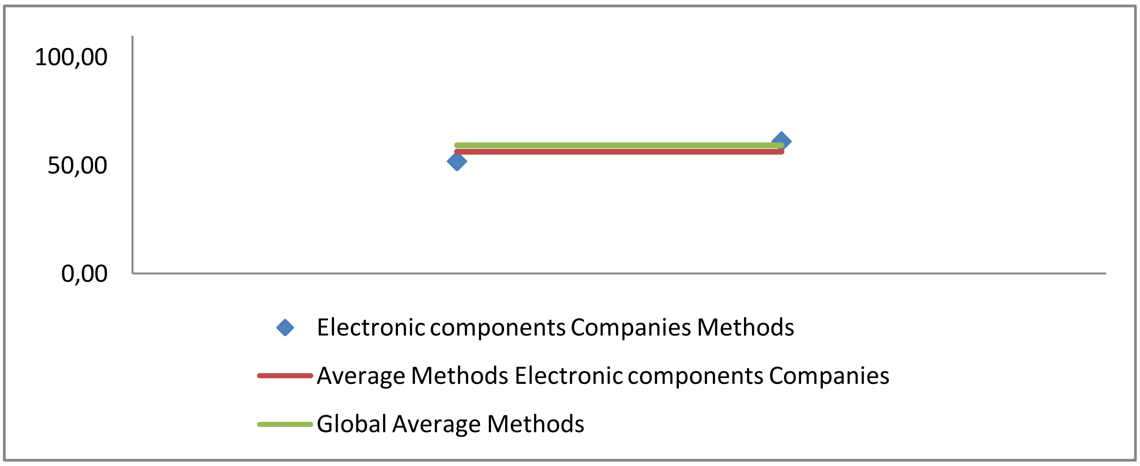

Figure 67. Methods Area, whole sample with respect of the GA

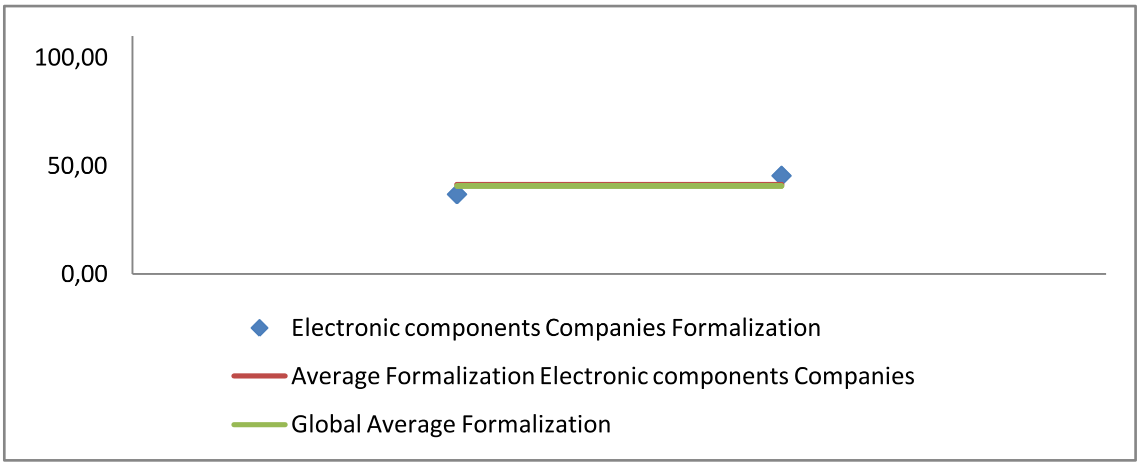

Figure 68. Formalization Area, whole sample with respect of the GA

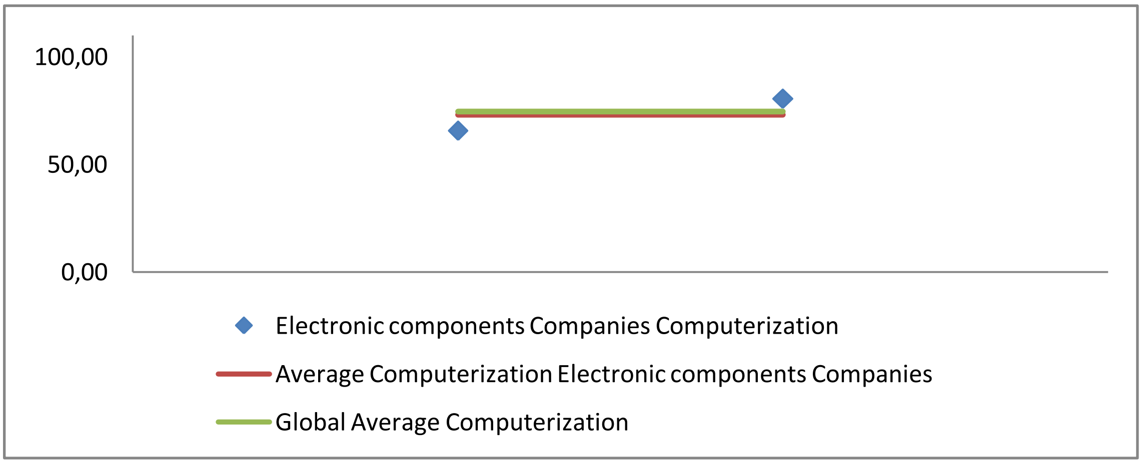

Figure 69. Computerization Area, whole sample with respect of the GA

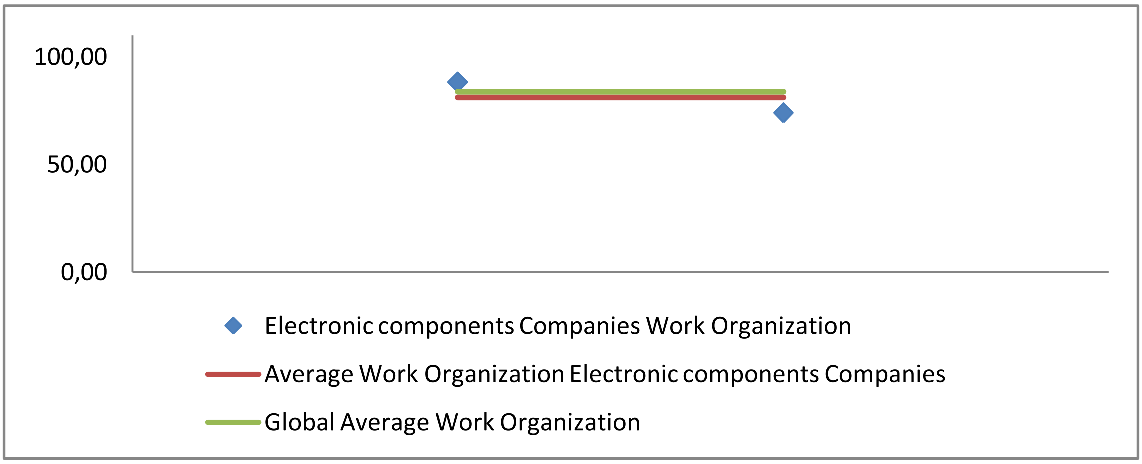

Figure 70. Work Organization Area, whole sample with respect of the GA

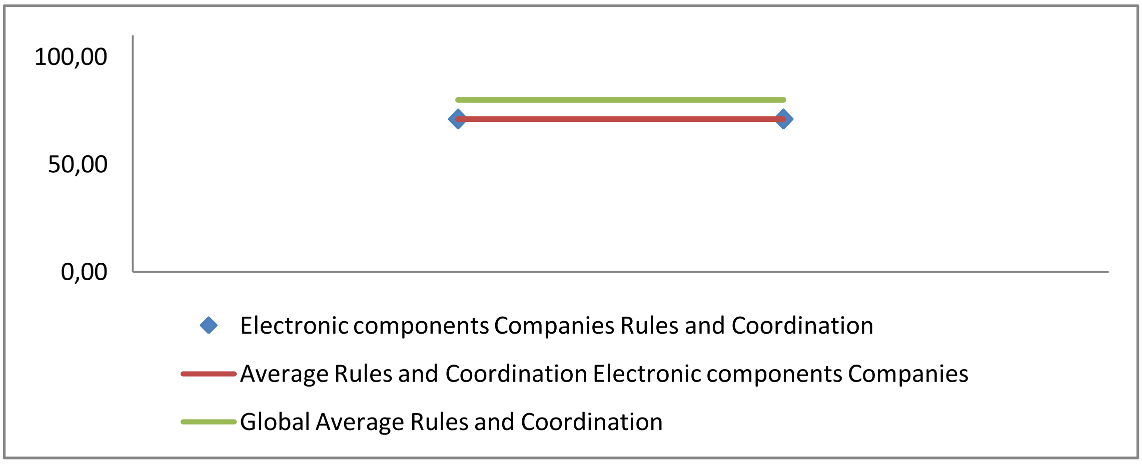

Figure 71. Rules and Competence Area, whole sample with respect of the GA

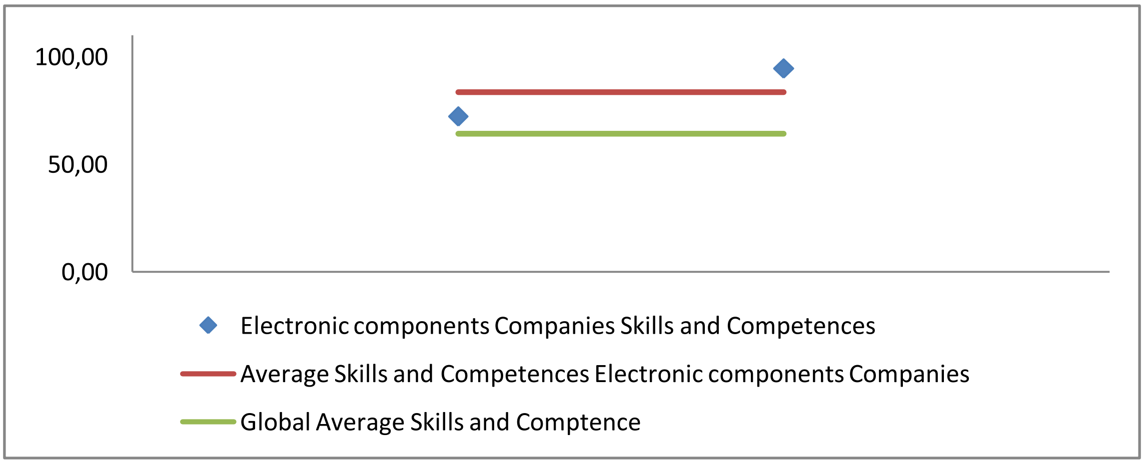

Figure 72. Skills and Competences Area, whole sample with respect of the GA

Manufacture of instruments and appliances for measuring, testing and navigation

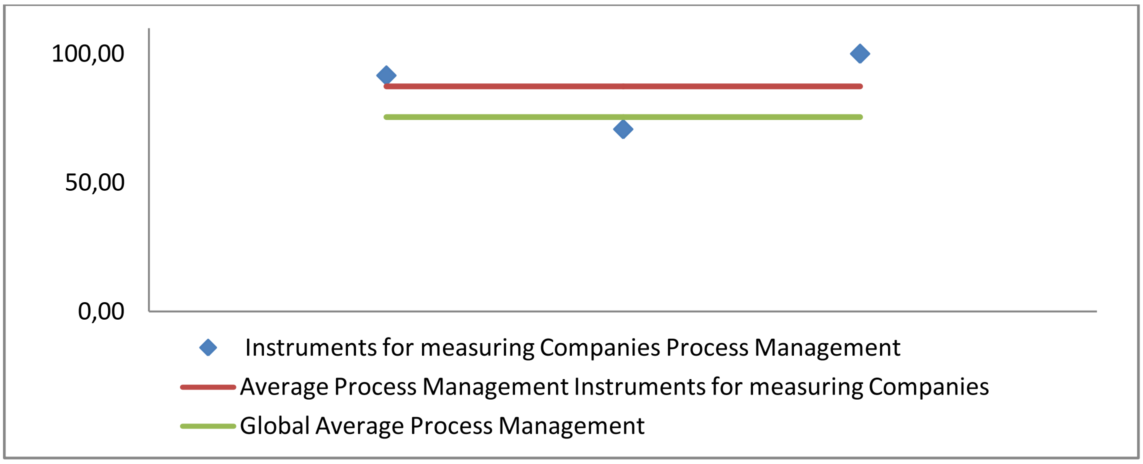

Figure 73. Process Management Area, whole sample with respect of the GA

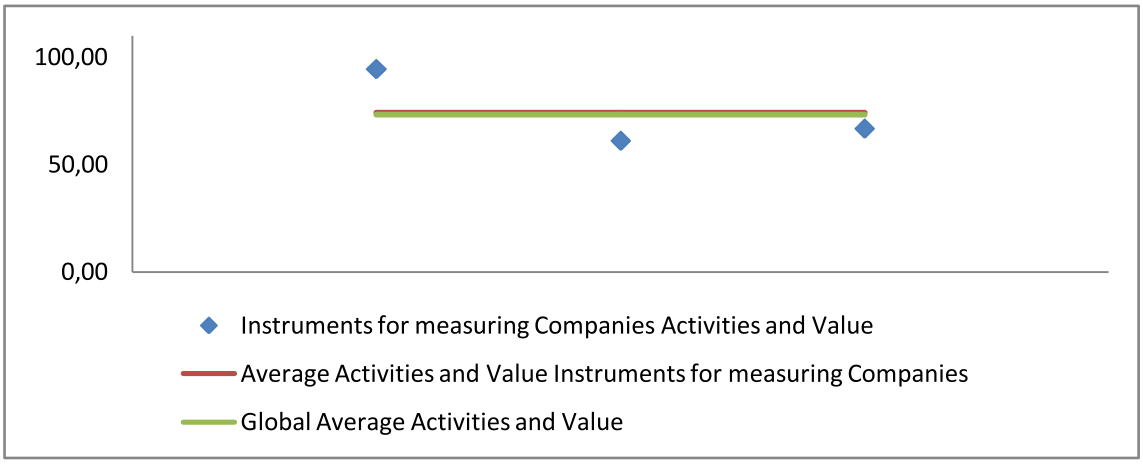

Figure74. Activities and Value Area, whole sample with respect of the GA

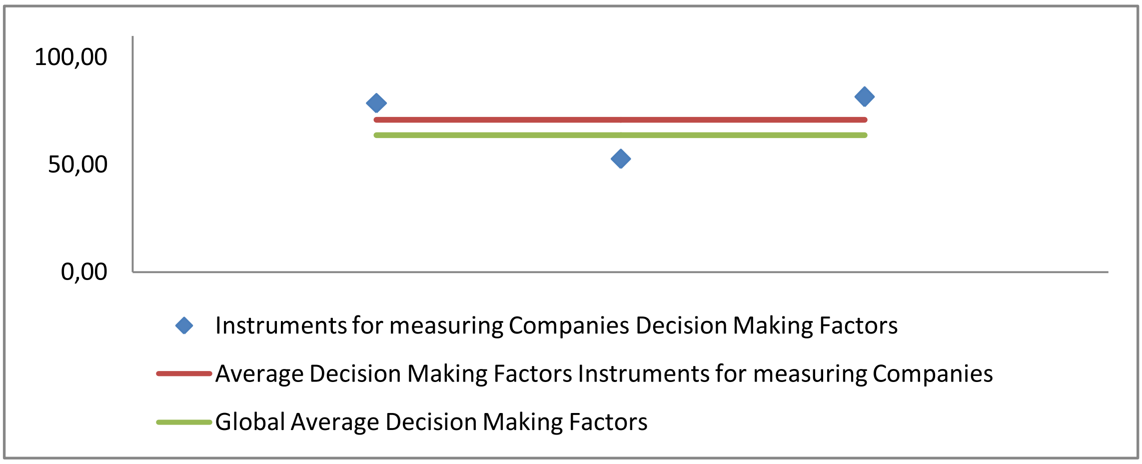

Figure 75. Decision Making Factors Area, whole sample with respect of the GA

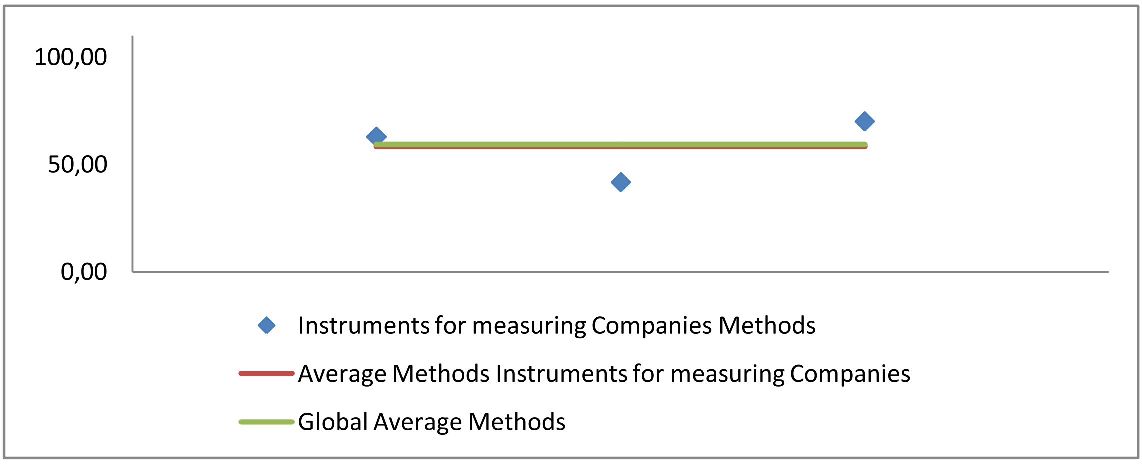

Figure 76. Methods Area, whole sample with respect of the GA

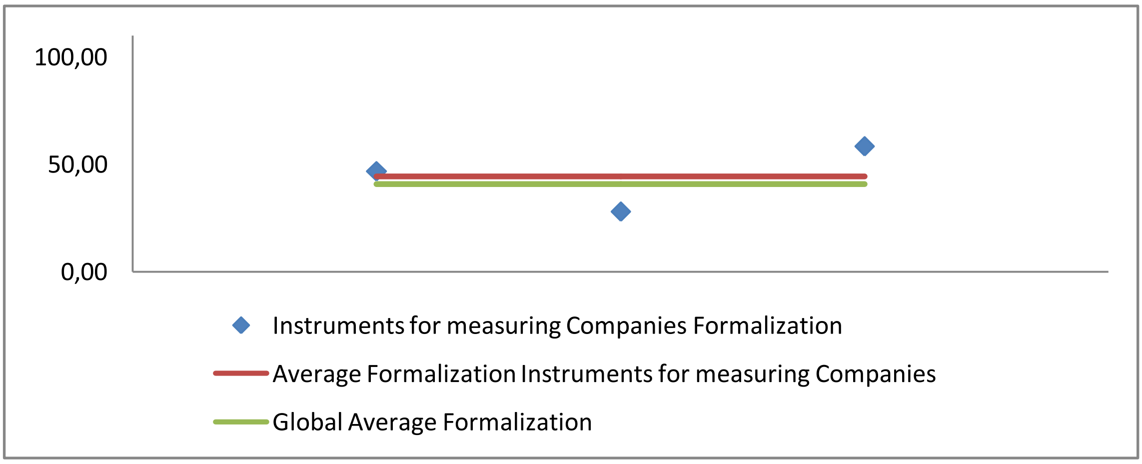

Figure 77. Formalization Area, whole sample with respect of the GA

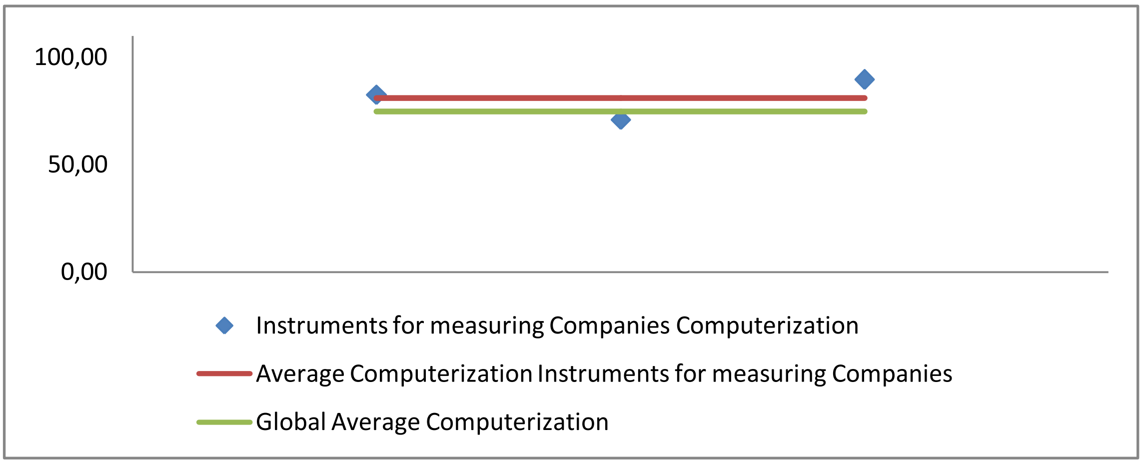

Figure 78. Computerization Area, whole sample with respect of the GA

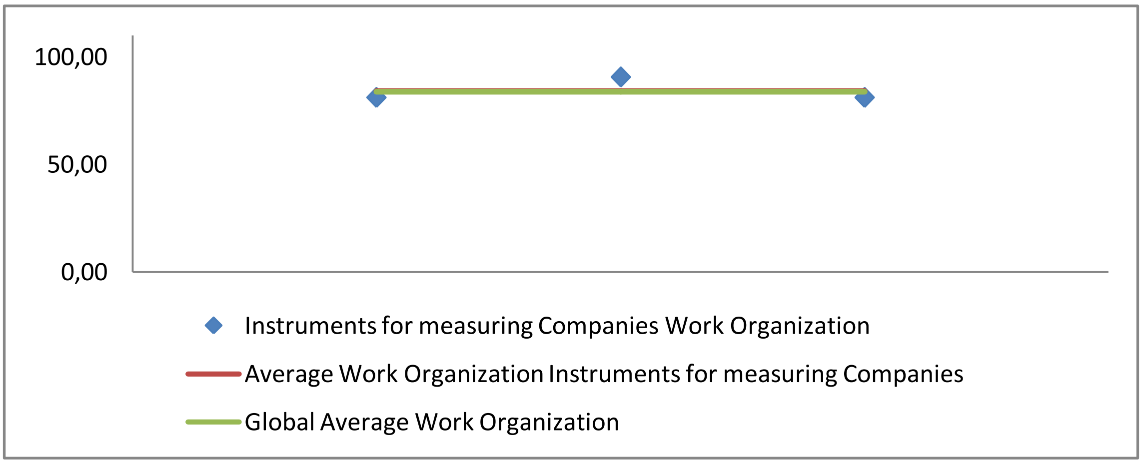

Figure 79. Work Organization Area, whole sample with respect of the GA

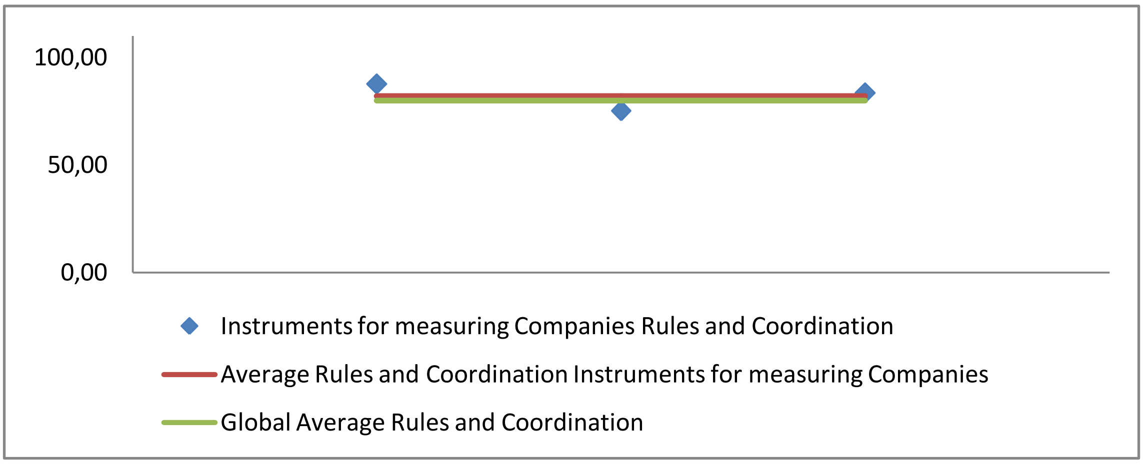

Figure 80. Rules and Coordination Area, whole sample with respect of the GA

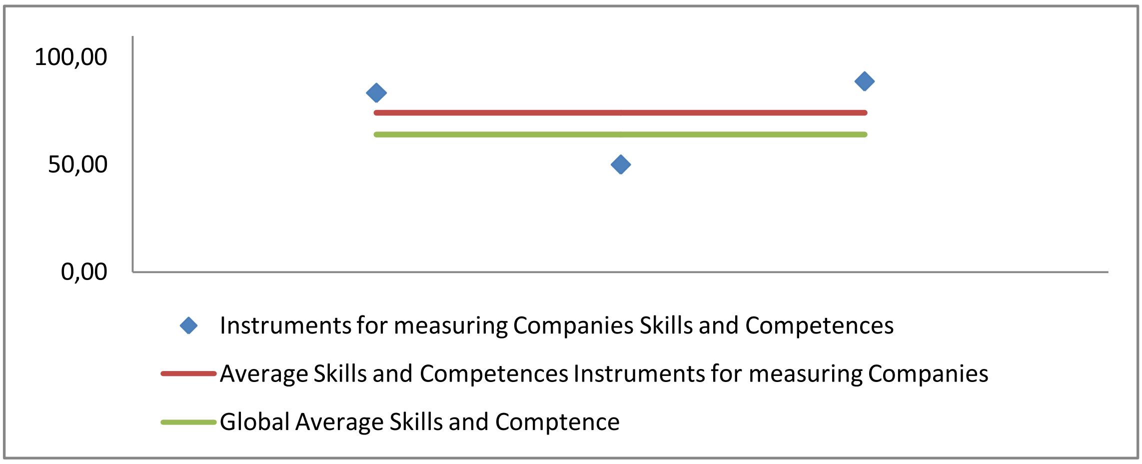

Figure 81. Skills and Competences Area, whole sample with respect of the GA

Manufacture of electric domestic appliances sector

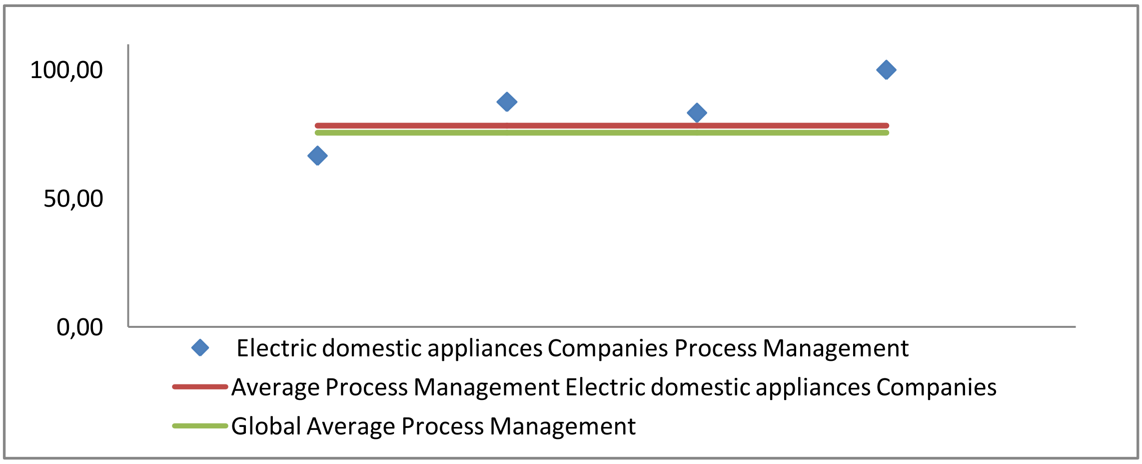

Figure 82. Process Management Area, whole sample with respect of the GA

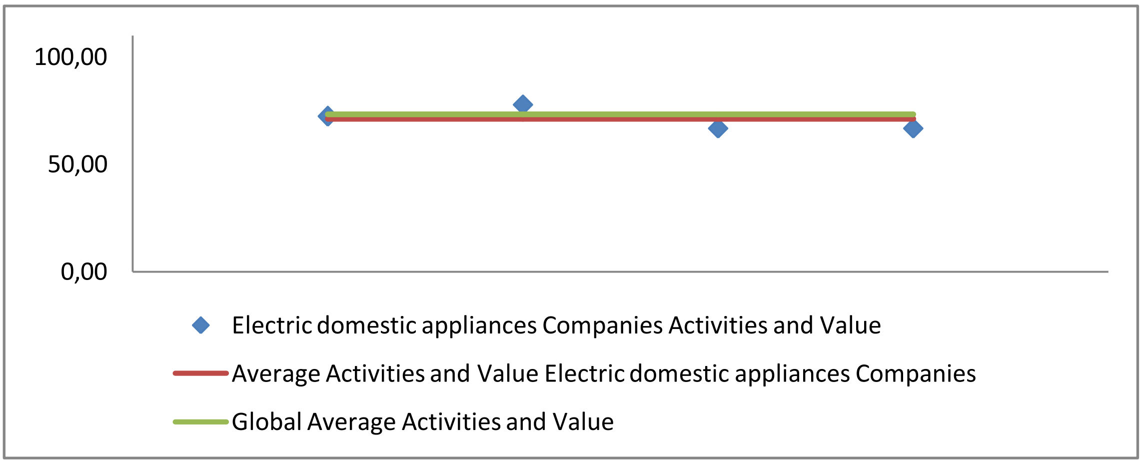

Figure 83. Activities and Value Area, whole sample with respect of the GA

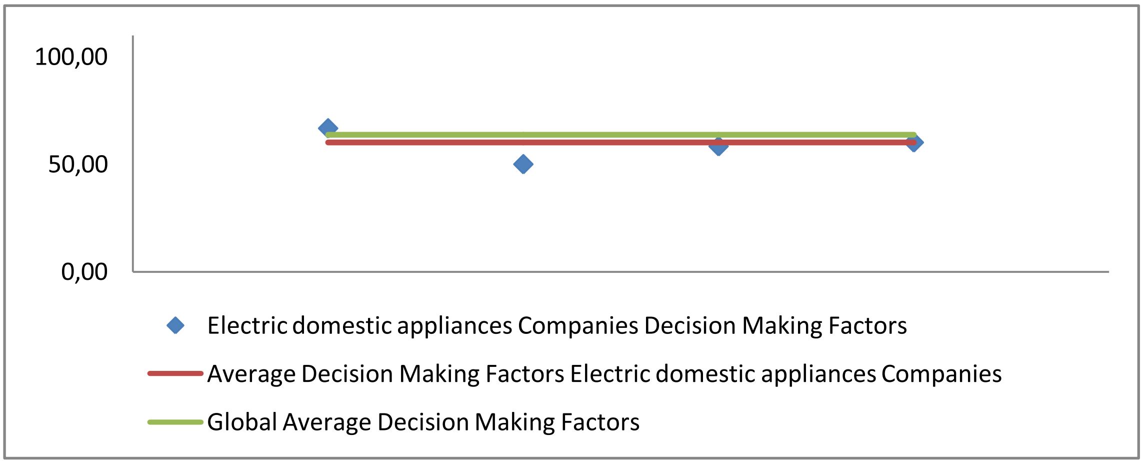

Figure 84. Decision Making Factors Area, whole sample with respect of the GA

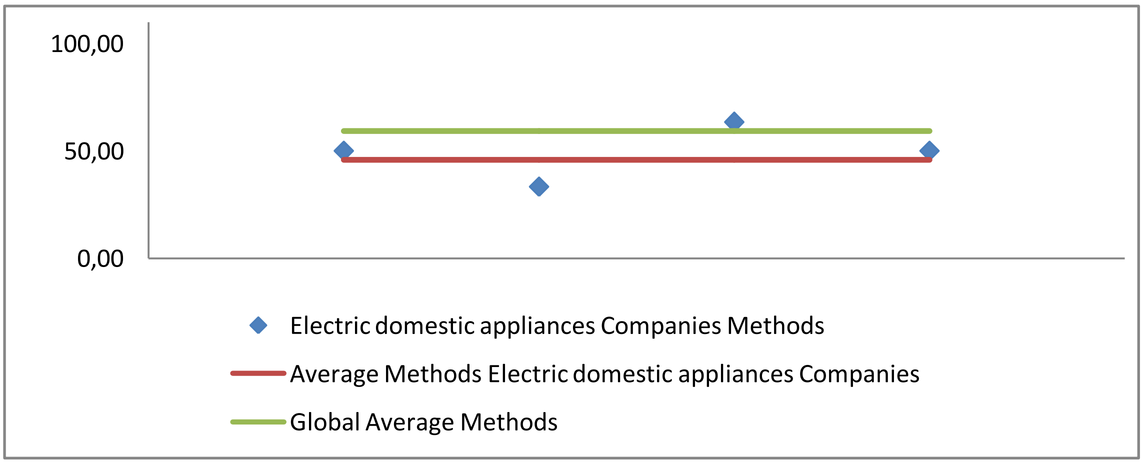

Figure 85. Methods Area, whole sample with respect of the GA

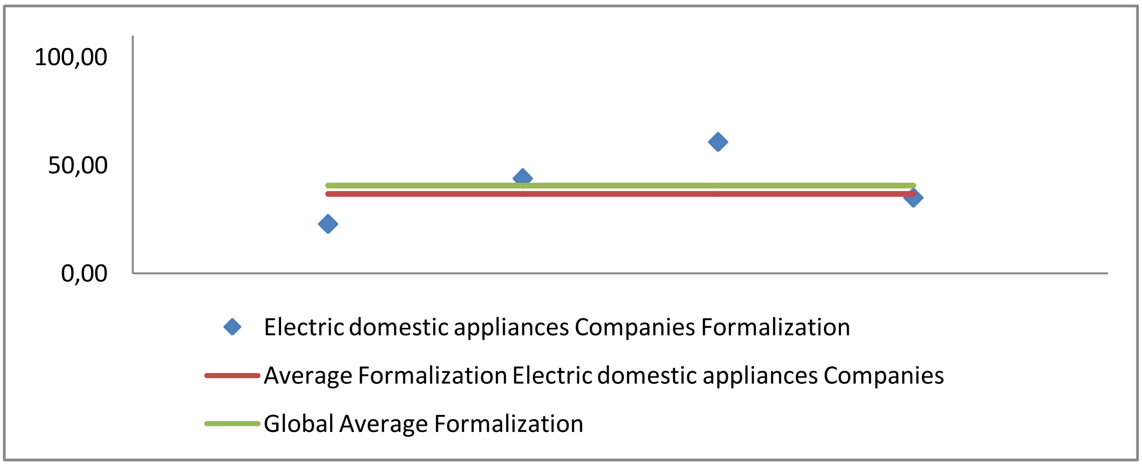

Figure 86. Formalization Area, whole sample with respect of the GA

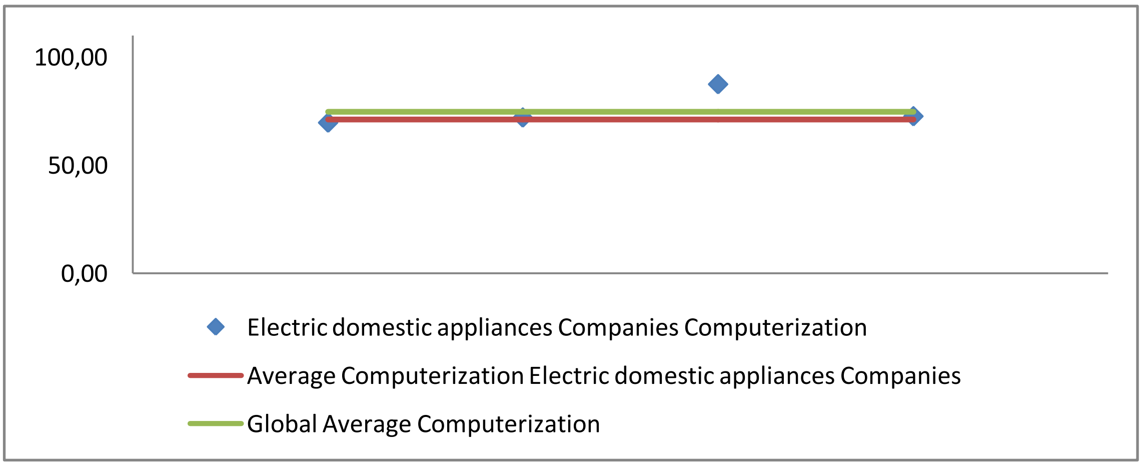

Figure 87. Computerization Area, whole sample with respect of the GA

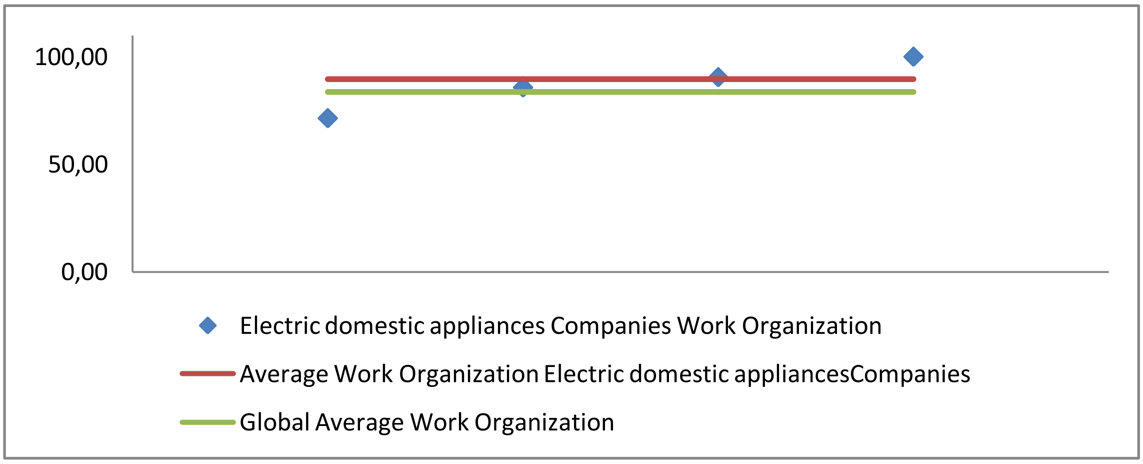

Figure 88. Work Organization Area, whole sample with respect of the GA

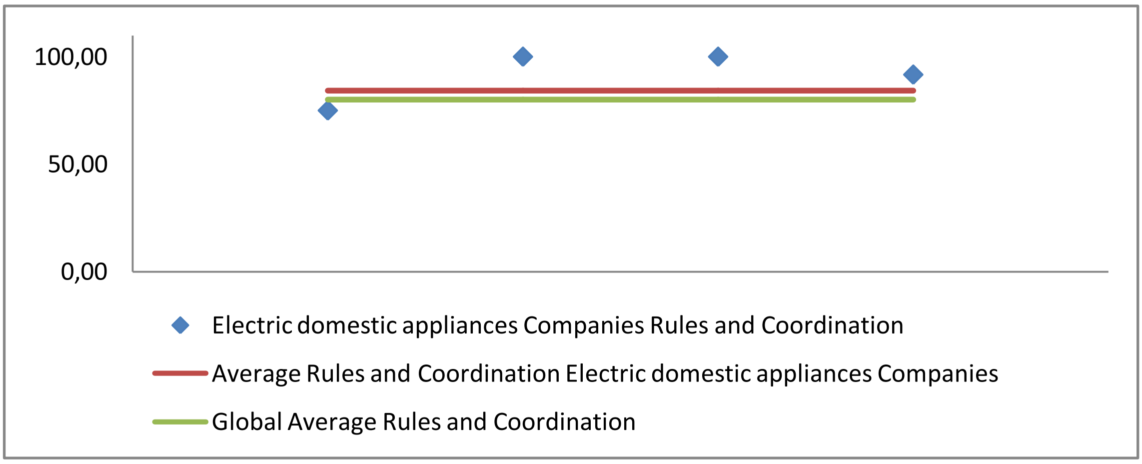

Figure 89. Rules and Coordination Area, whole sample with respect of the GA

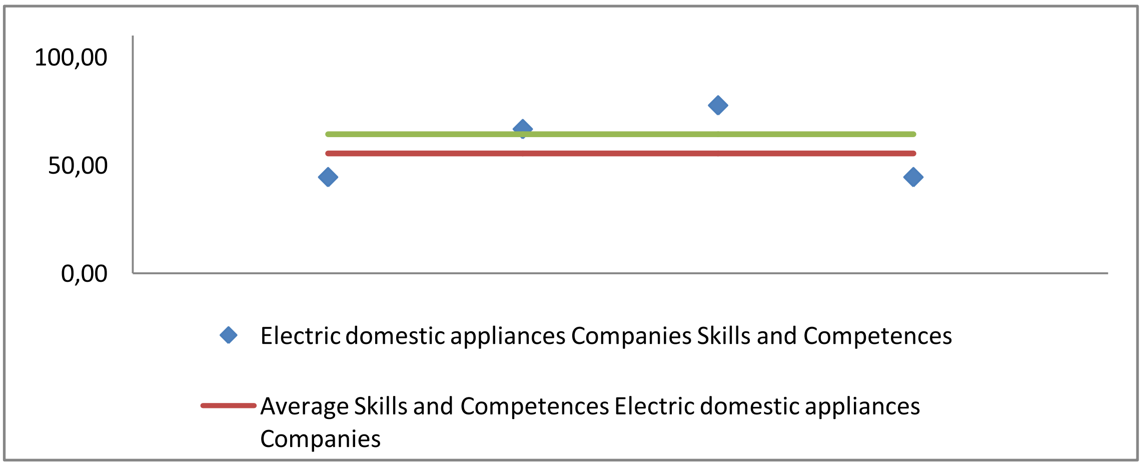

Figure 90. Skills and Competences Area, whole sample with respect of the GA

Manufacture of other electrical equipment sector

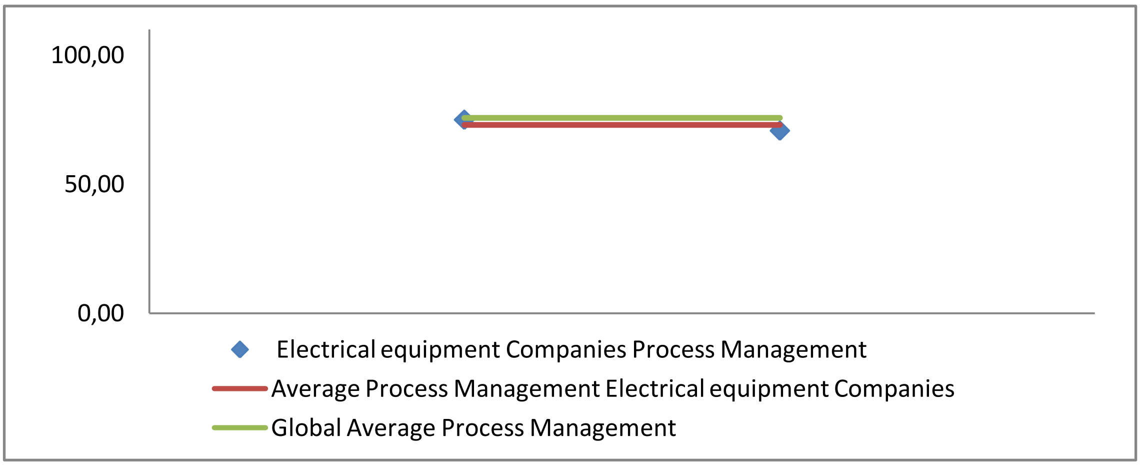

Figure 91. Process Management Area, whole sample with respect of the GA

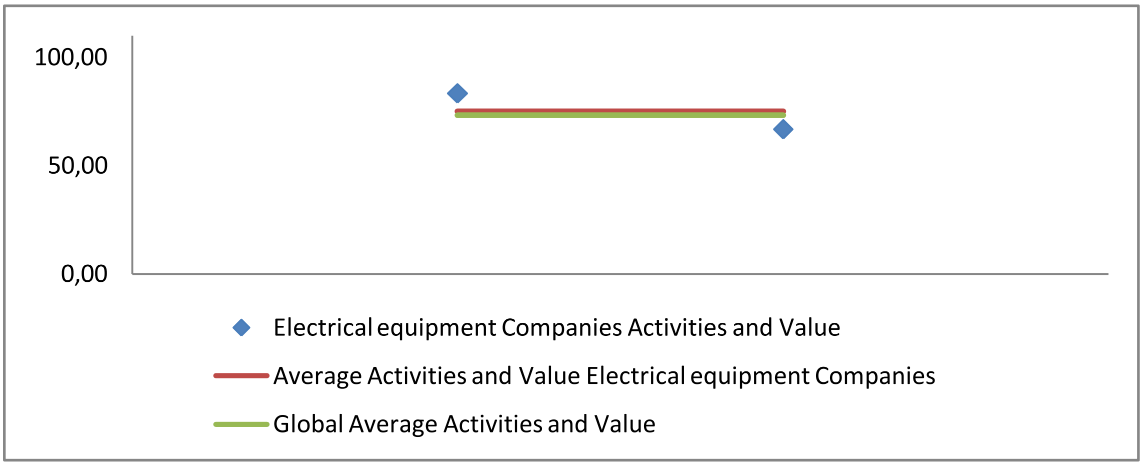

Figure 92. Activities and Value Area, whole sample with respect of the GA

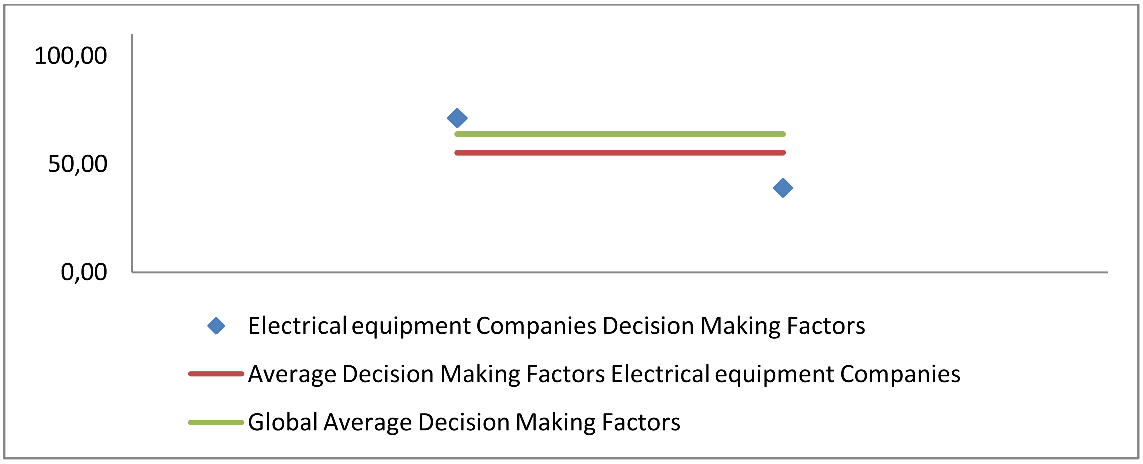

Figure 93. Decision Making Factors Area, whole sample with respect of the GA

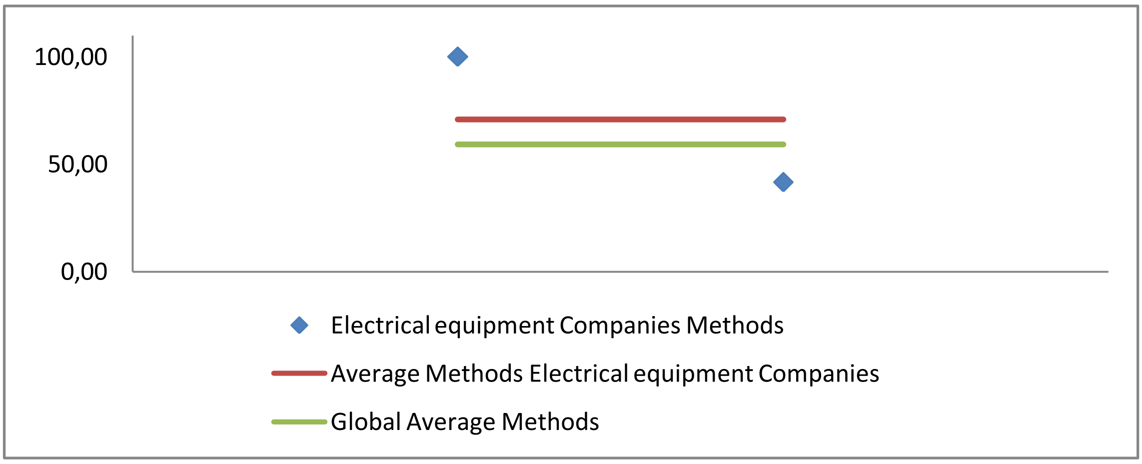

Figure 94. Methods Area, whole sample with respect of the GA

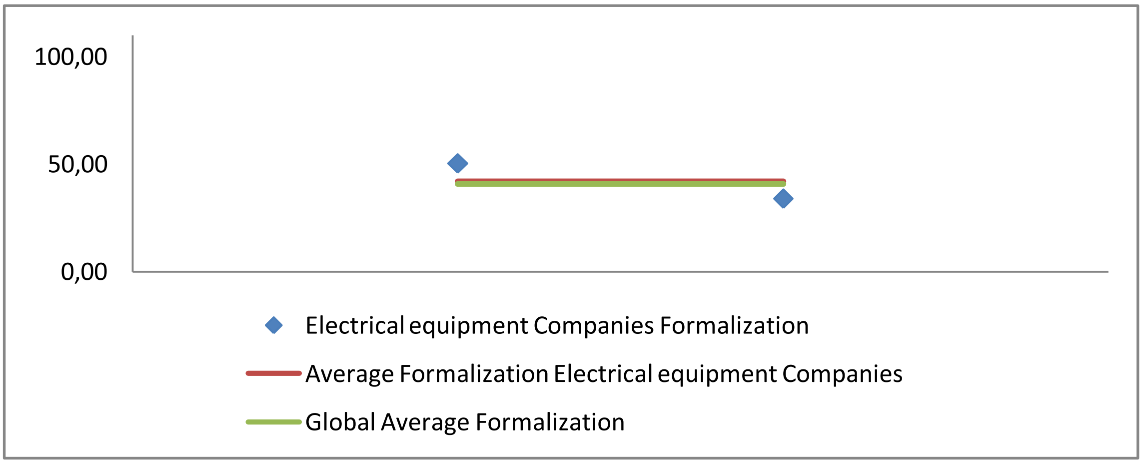

Figure 95. Formalization Area, whole sample with respect of the GA

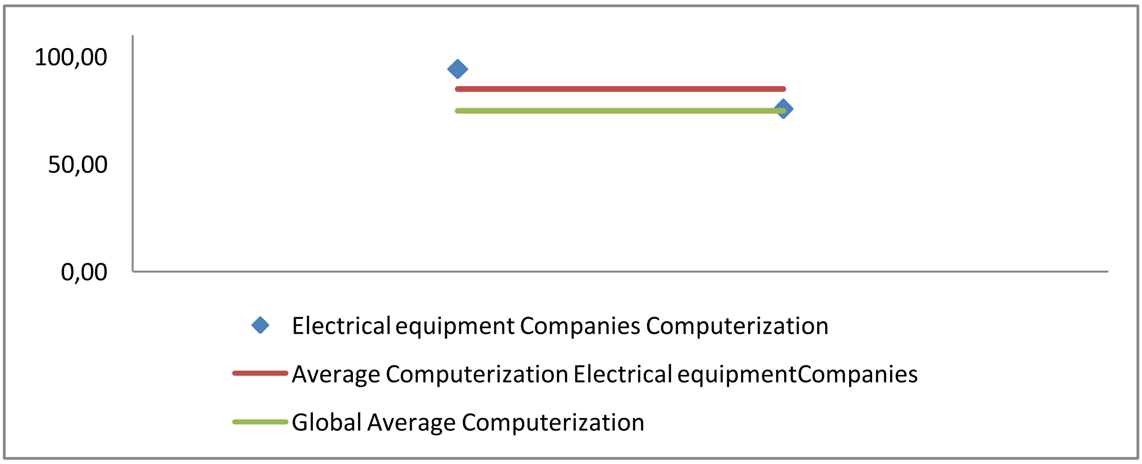

Figure 96. Computerization Area, whole sample with respect of the GA

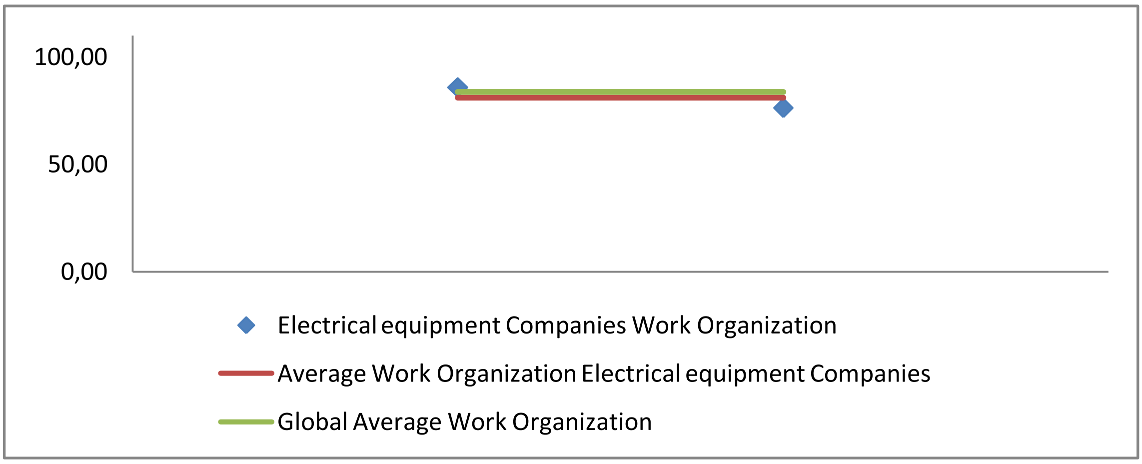

Figure 97. Work Organization Area, whole sample with respect of the GA

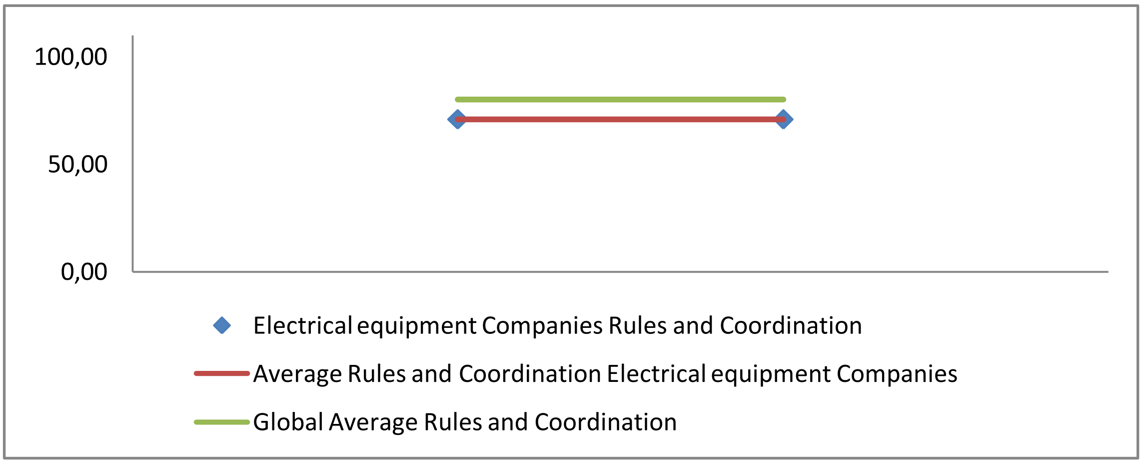

Figure 98. Rules and Coordination Area, whole sample with respect of the GA

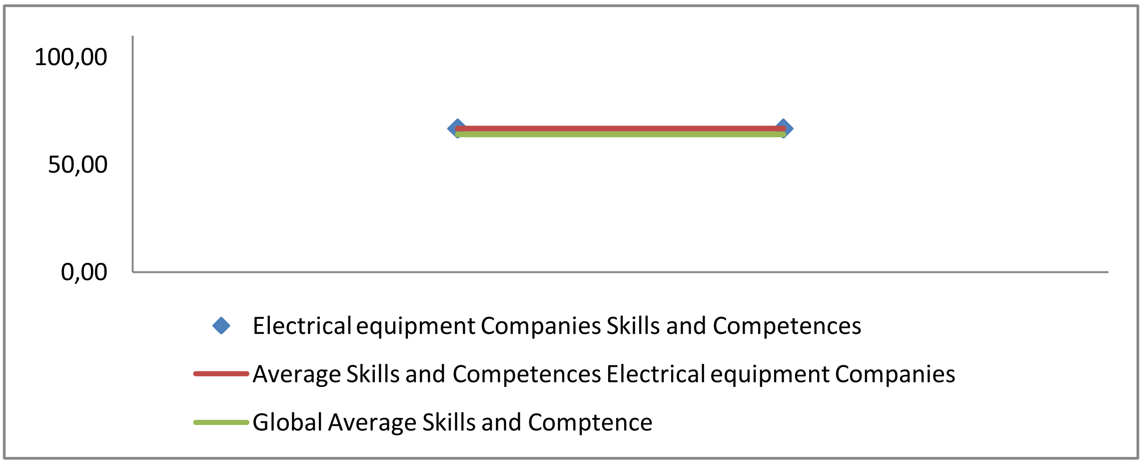

Figure 99. Skills and Competences Area, whole sample with respect of the GA

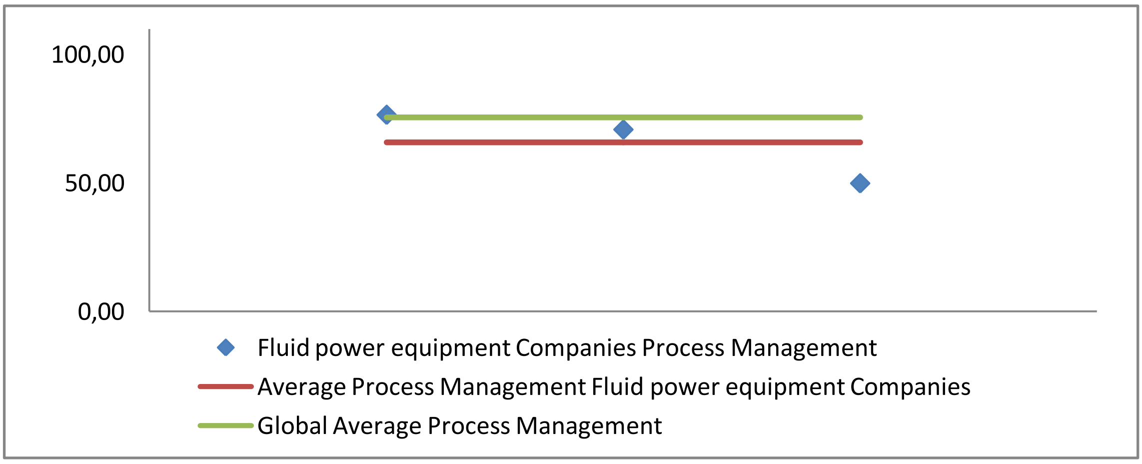

Figure 100. Process Management Area, whole sample with respect of the GA

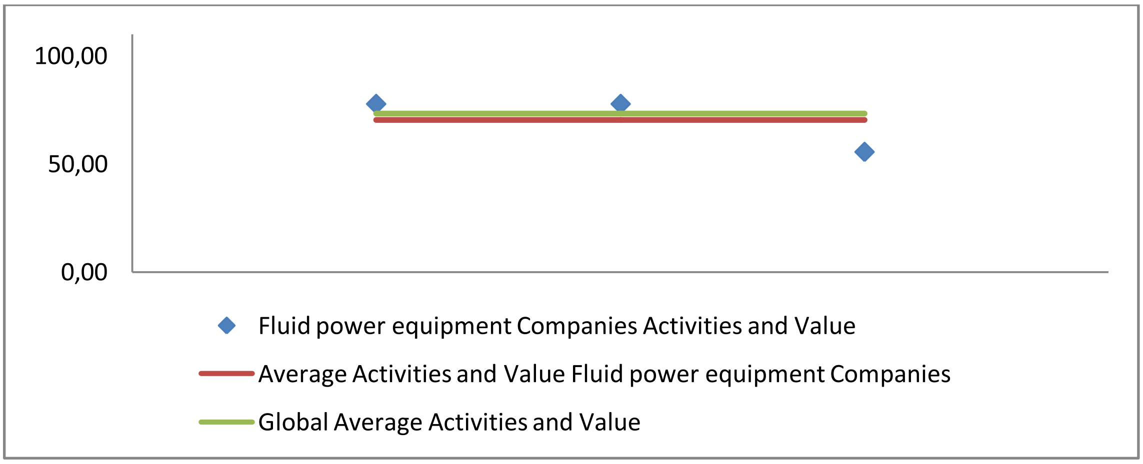

Figure 101. Activities and Value Area, whole sample with respect of the GA

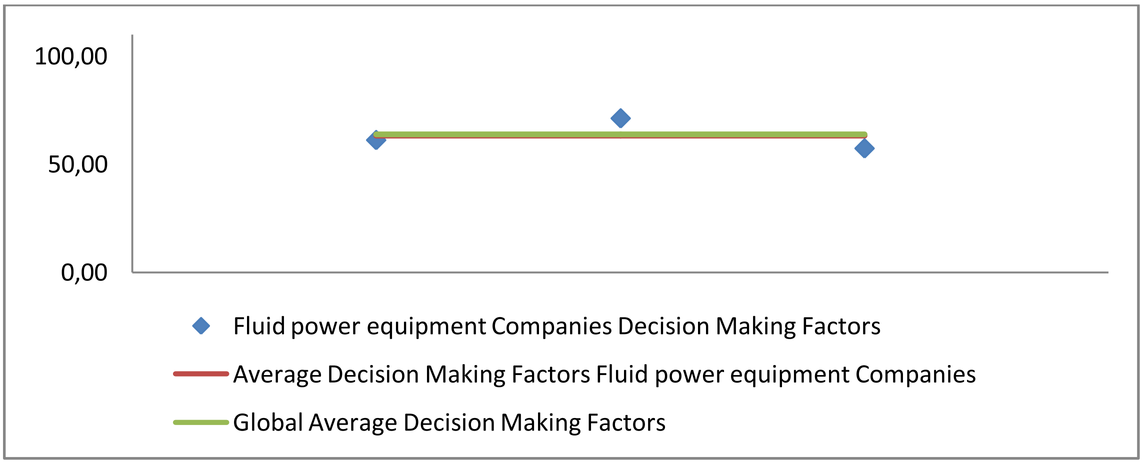

Figure 102. Decision Making Factors Area, whole sample with respect of the GA

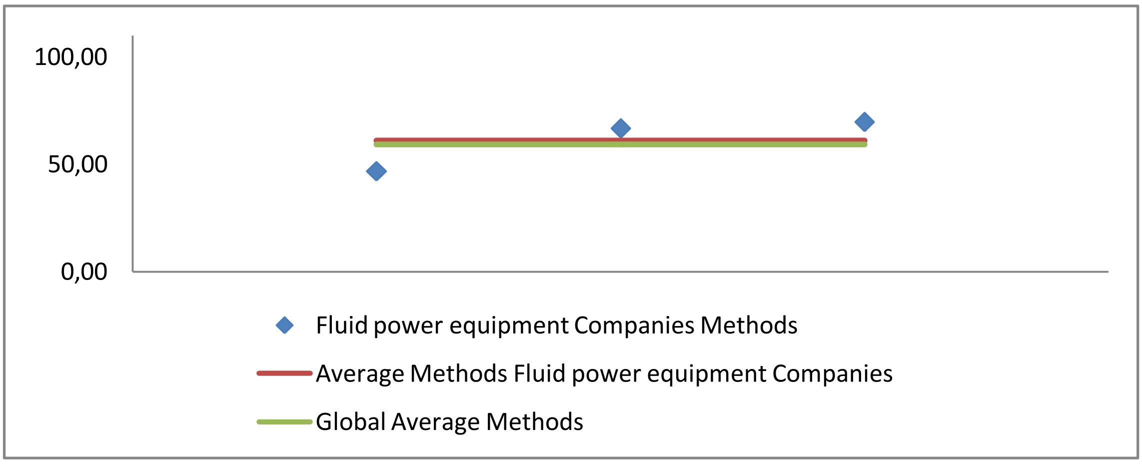

Figure 103. Methods Area, whole sample with respect of the GA

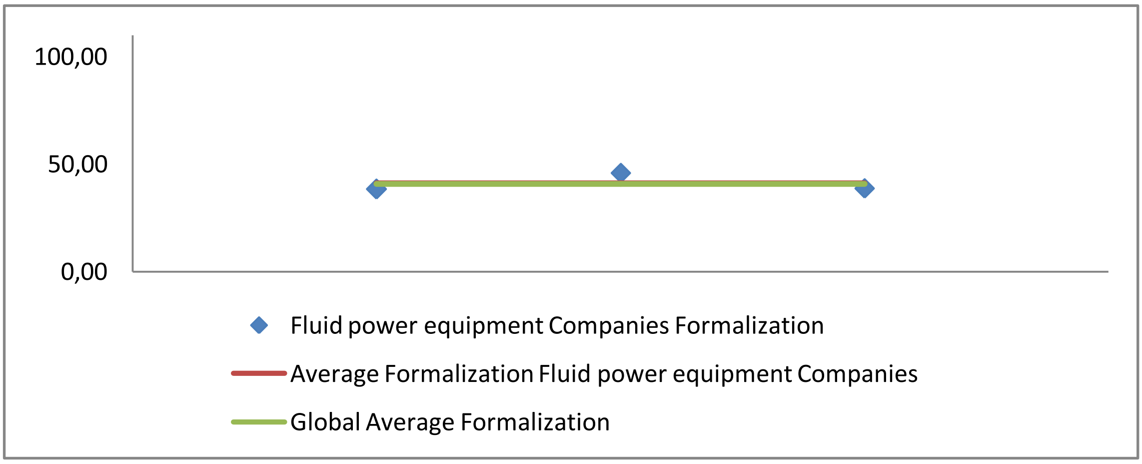

Figure 104. Formalization Area, whole sample with respect of the GA

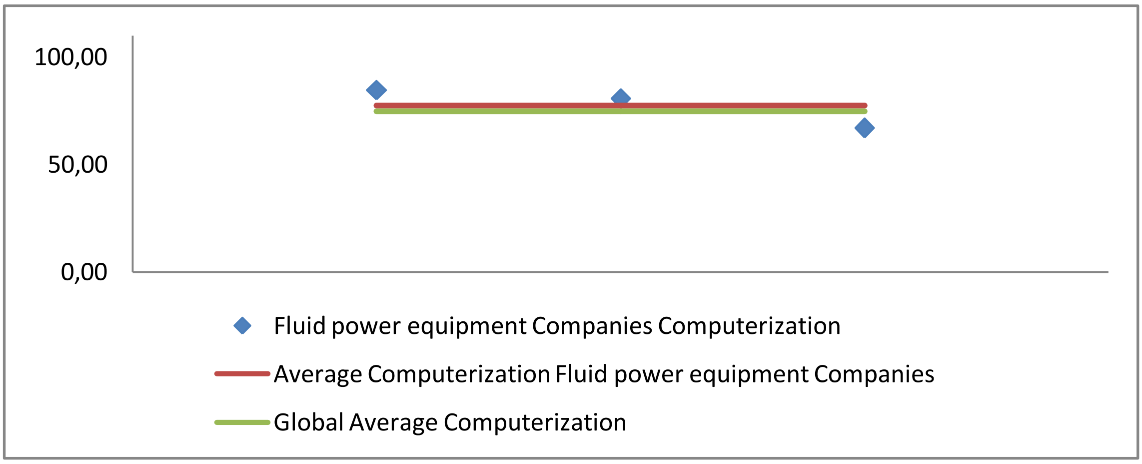

Figure 105. Computerization Area, whole sample with respect of the GA

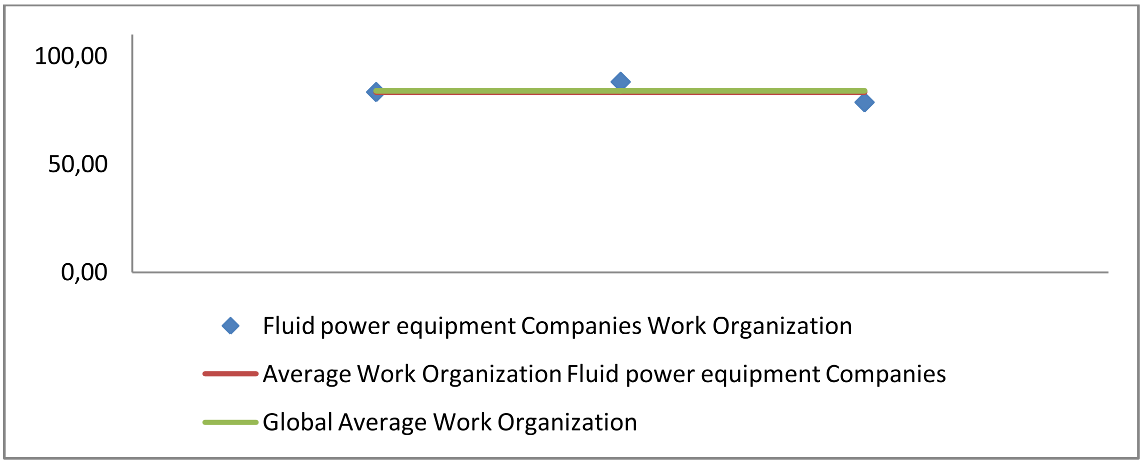

Figure 106. Work Organization Area, whole sample with respect of the GA

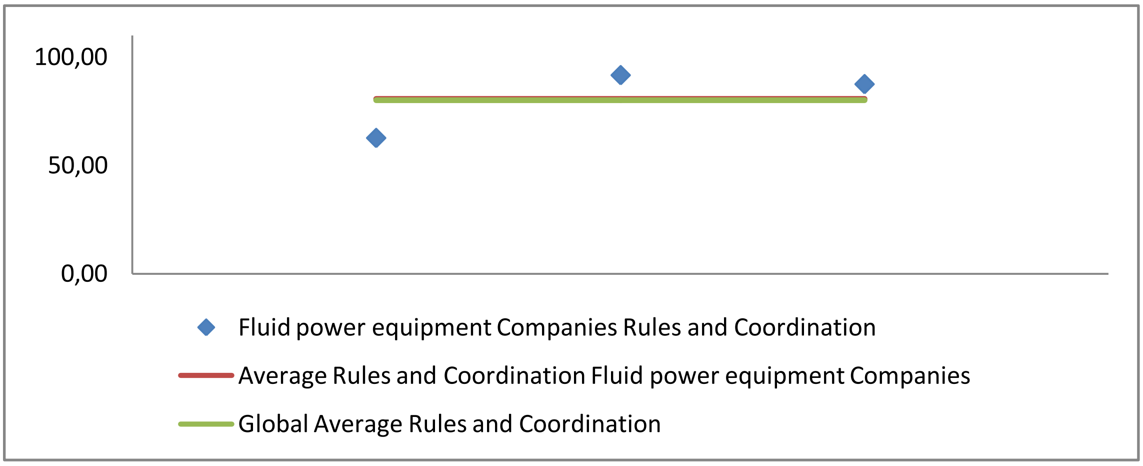

Figure 107. Skills and Competences Area, whole sample with respect of the GA

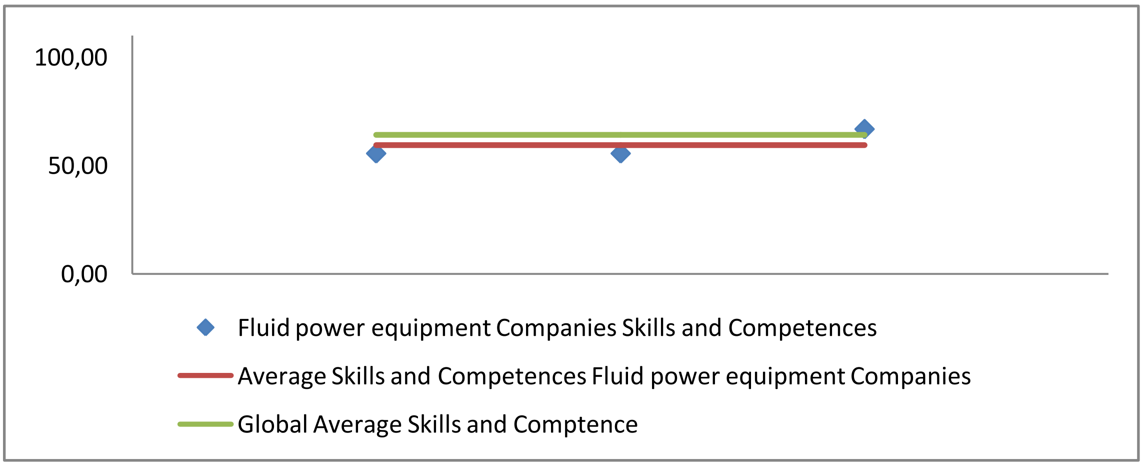

Figure 108. Skills and Competences Area, whole sample with respect of the GA

Manufacture of general-purpose machinery sector

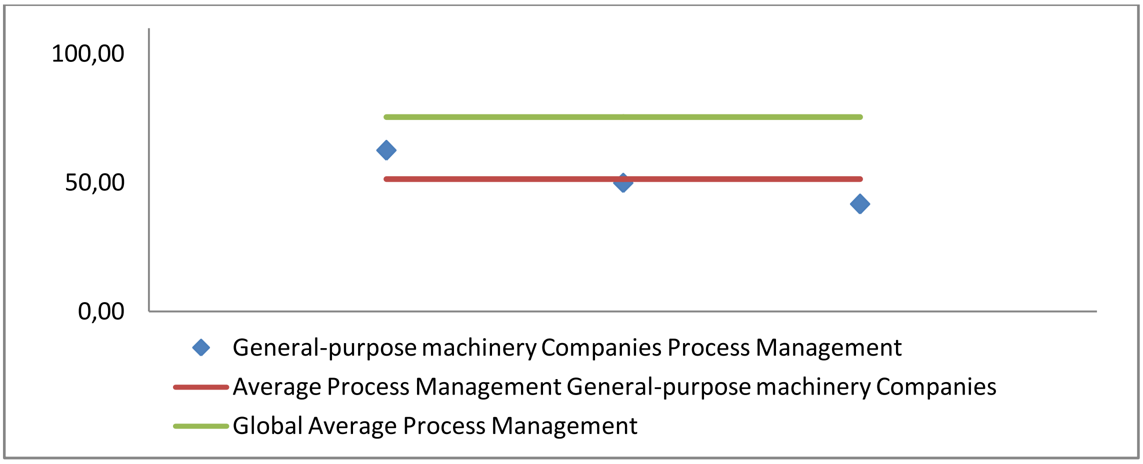

Figure 109. Process Management Area, whole sample with respect of the GA

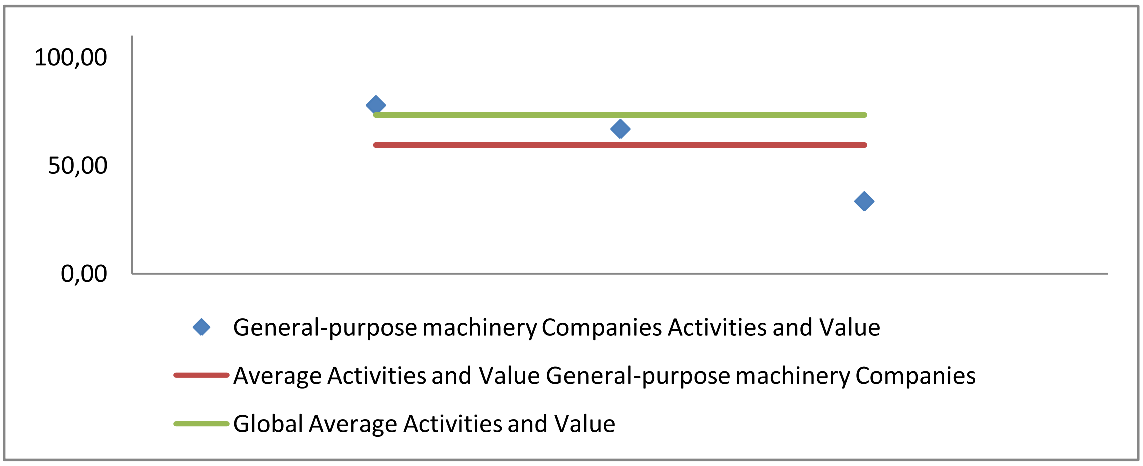

Figure 110. Activities and Value Area, whole sample with respect of the GA

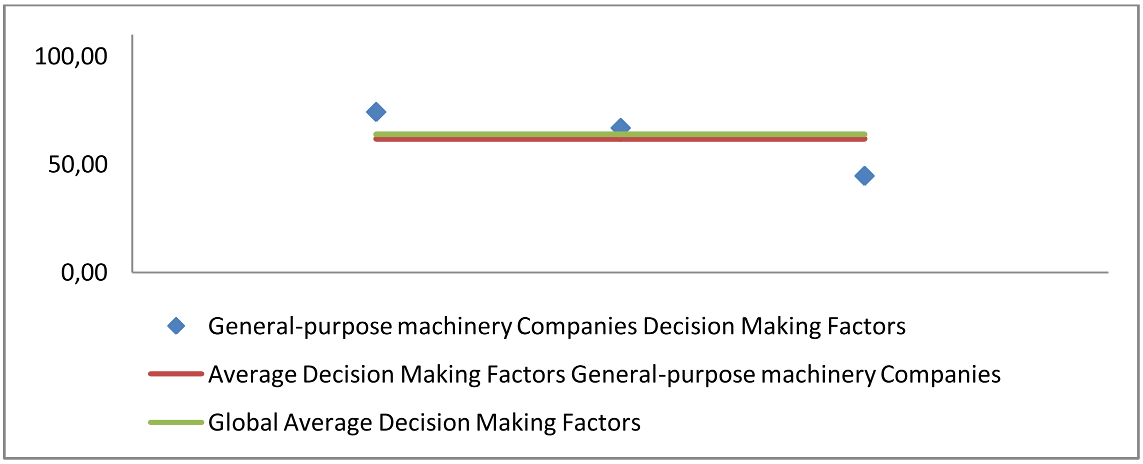

Figure 111. Decision Making Factors Area, whole sample with respect of the GA

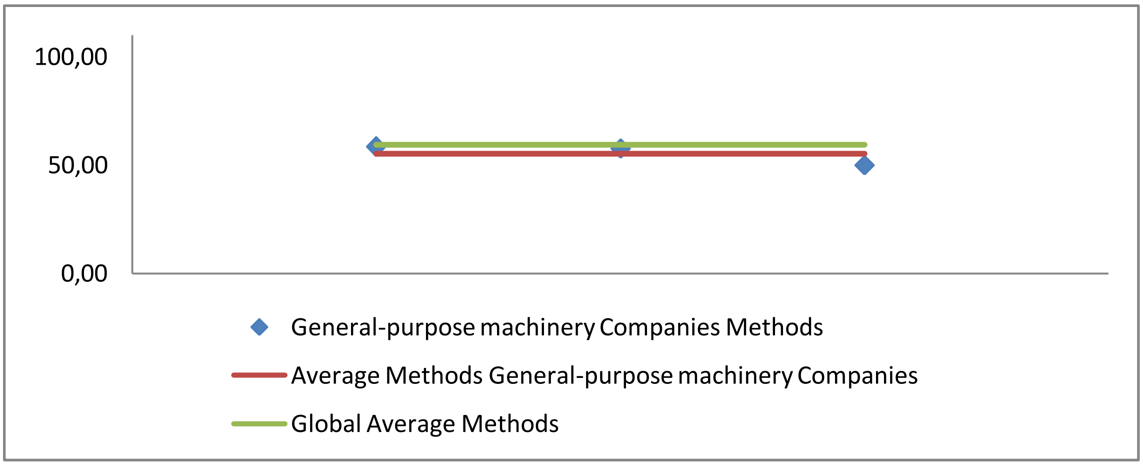

Figure 112. Methods Area, whole sample with respect of the GA

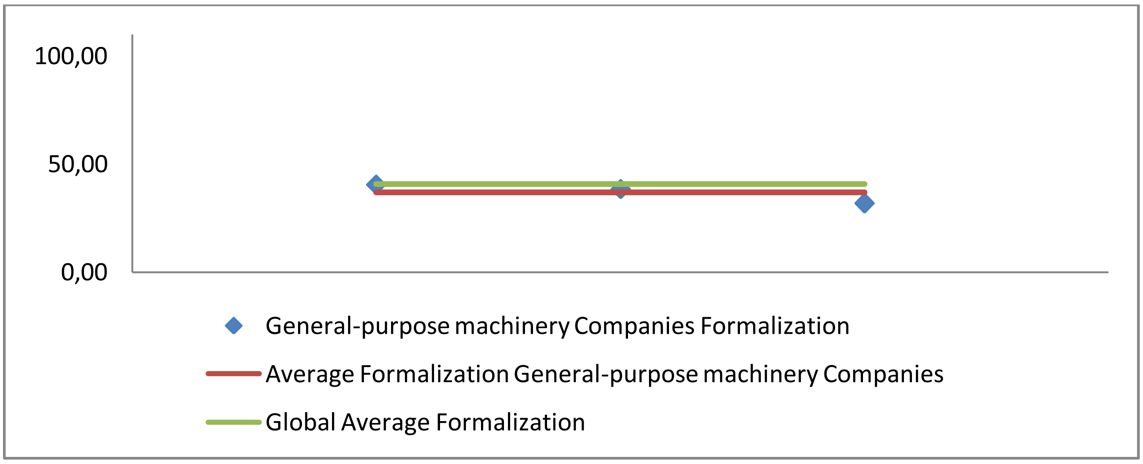

Figure 113. Formalization Area, whole sample with respect of the GA

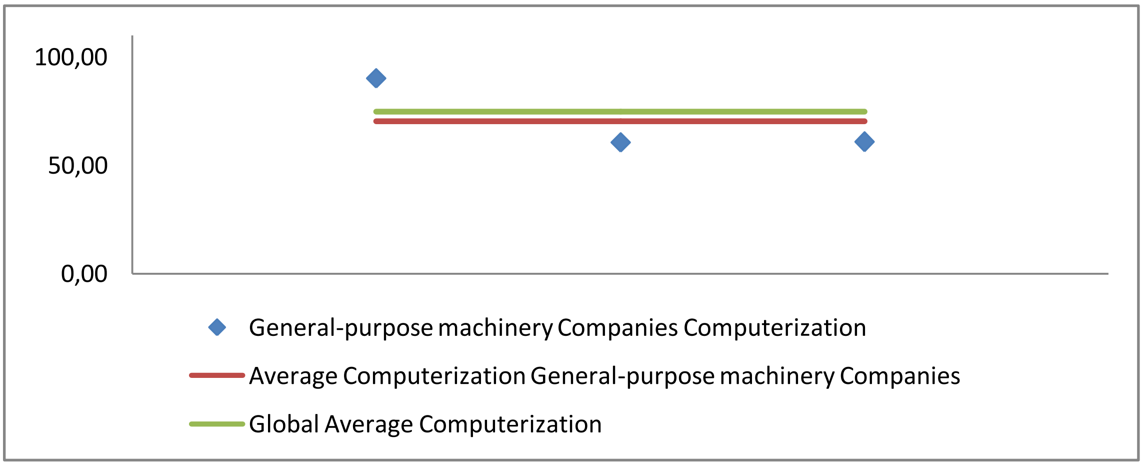

Figure 114. Computerization Area, whole sample with respect of the GA

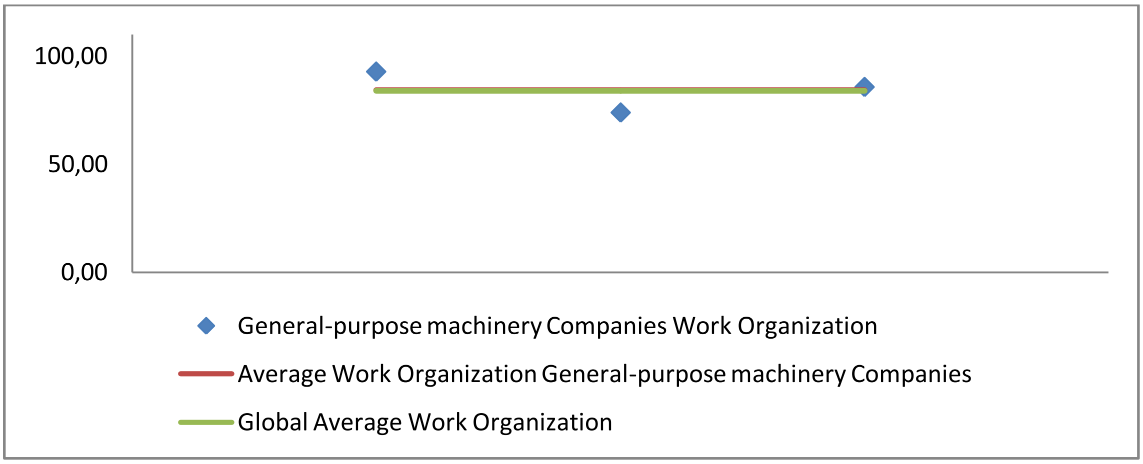

Figure 115. Work Organization Area, whole sample with respect of the GA

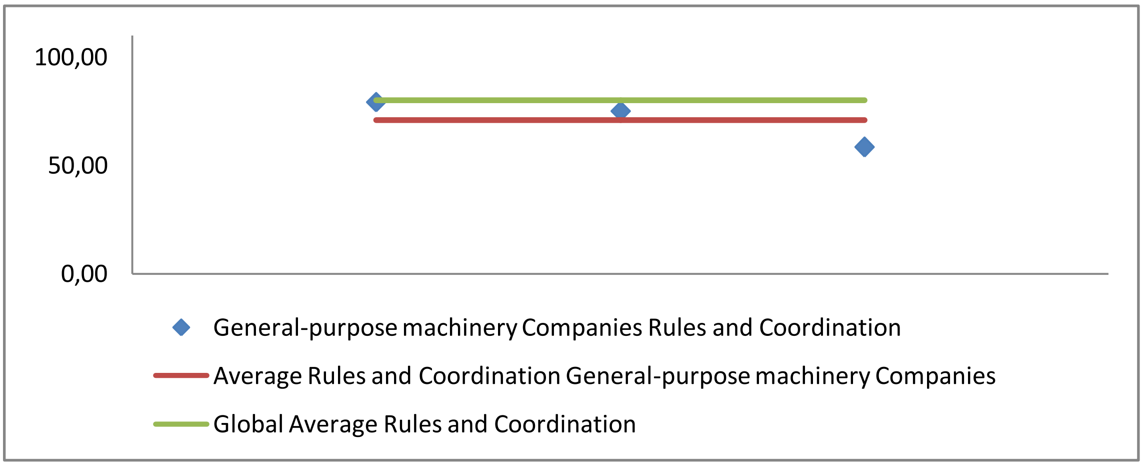

Figure 116. Rules and Coordination Area, whole sample with respect of the GA

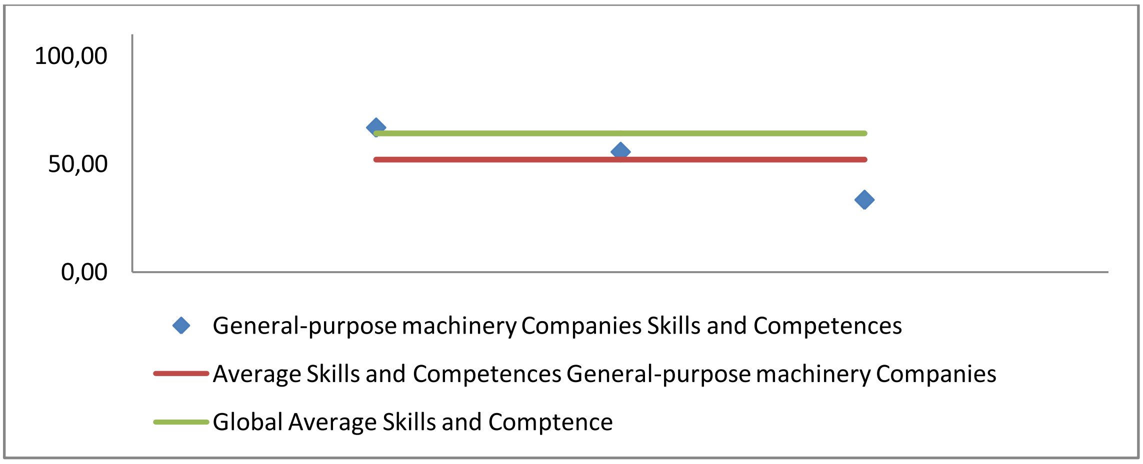

Figure 117. Skills and Competences Area, whole sample with respect of the GA

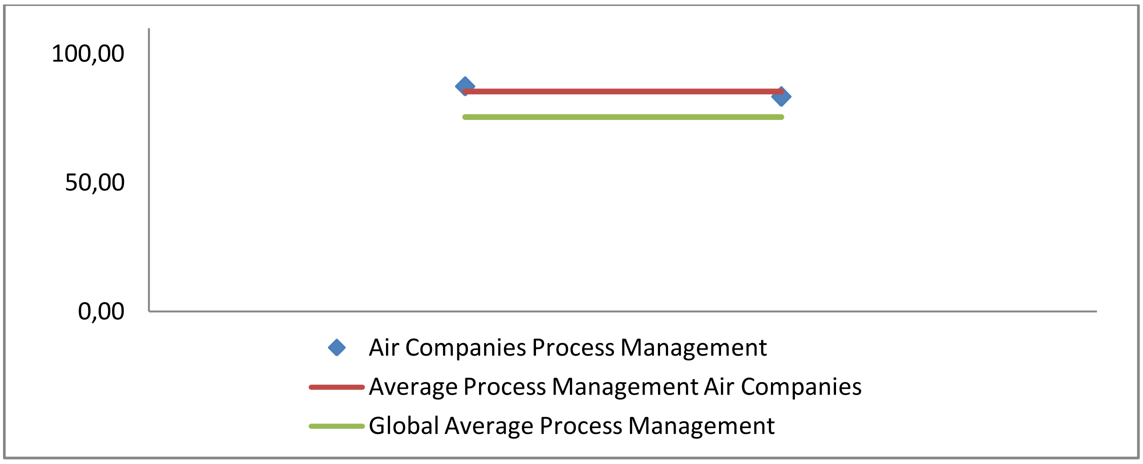

Figure 118. Process Management Area, whole sample with respect of the GA

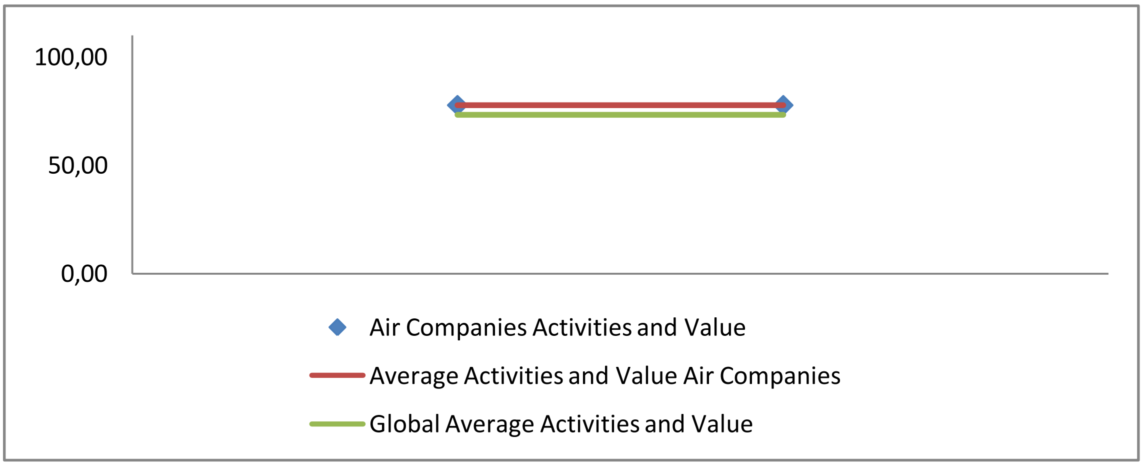

Figure 119. Activities and Value Area, whole sample with respect of the GA

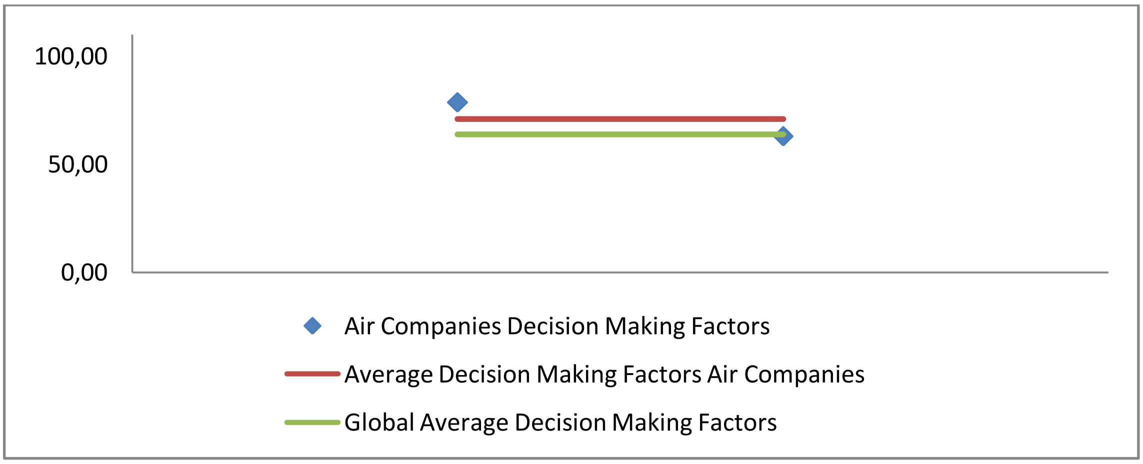

Figure 120. Decision Making Factors Area, whole sample with respect of the GA

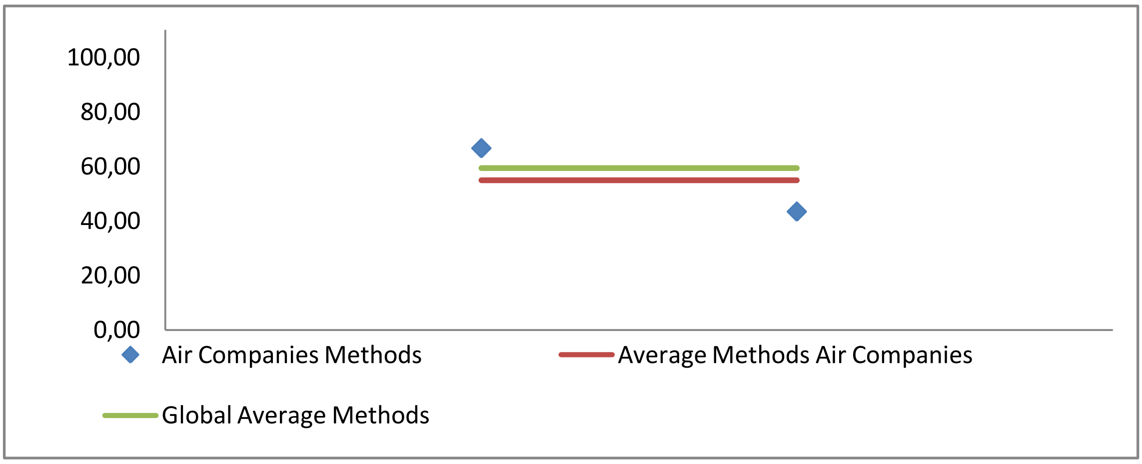

Figure 121. Methods Area, whole sample with respect of the GA

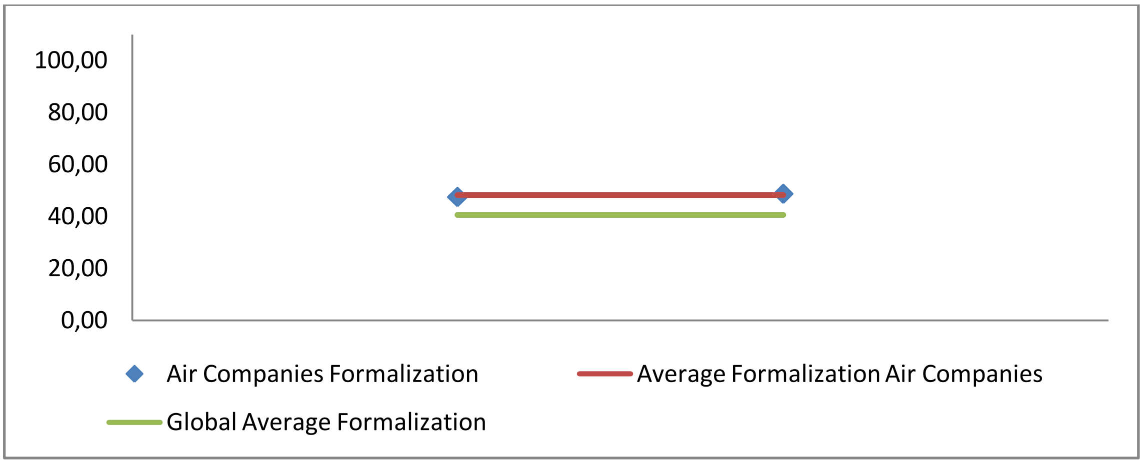

Figure 122. Formalization Area, whole sample with respect of the GA

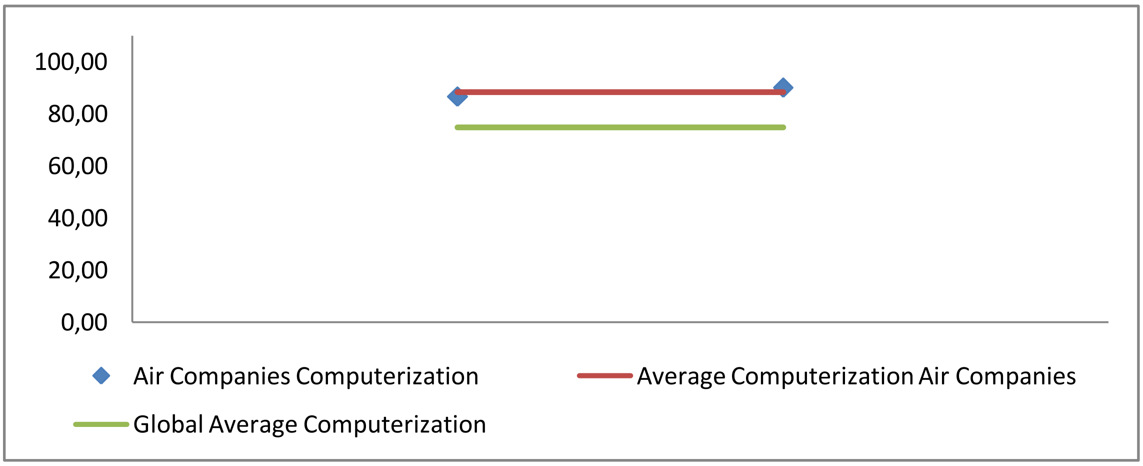

Figure 123. Computerization Area, whole sample with respect of the GA

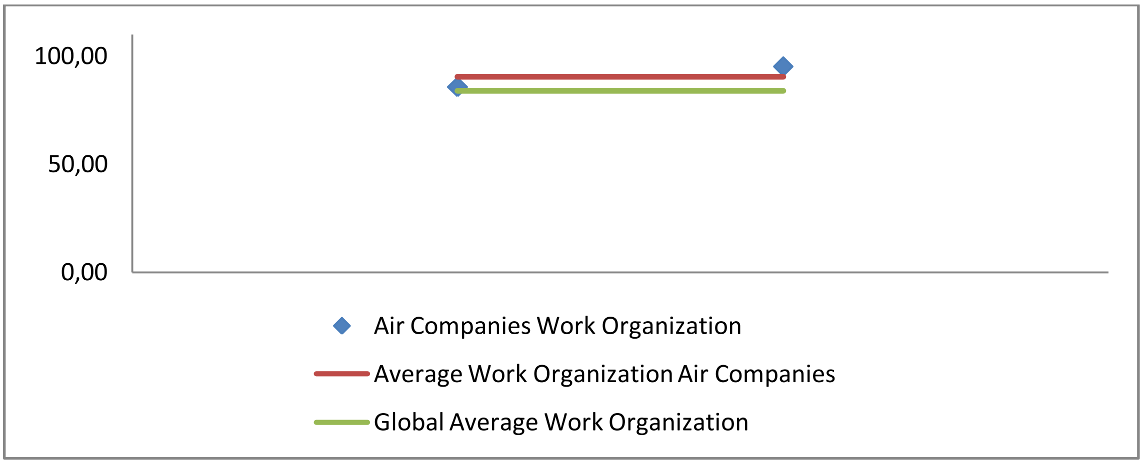

Figure 124. Work Organization Area, whole sample with respect of the GA

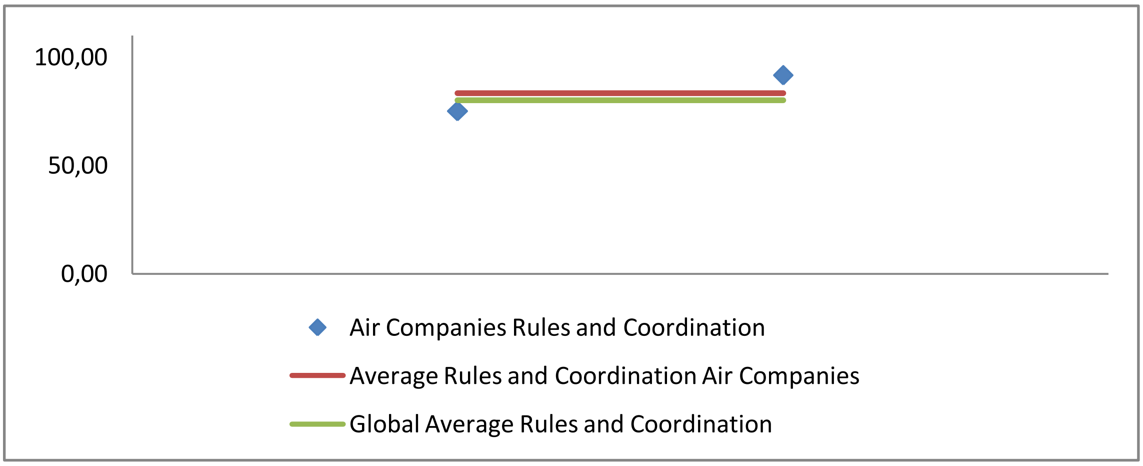

Figure 125. Rules and Coordination Area, whole sample with respect of the GA

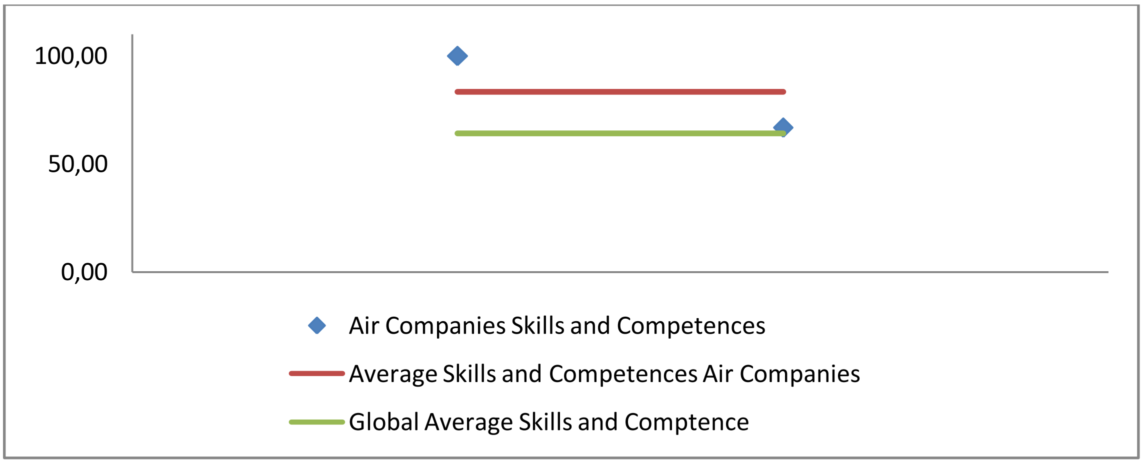

Figure 126. Skills and Competences Area, whole sample with respect of the GA

Average and standard deviation of other manufacturing n.e.c. sector

In this case of study we have only a company and it isn't notable execute a study. Nevertheless we show in the next table the average and the standard deviation.

Summary of other manufacturing n.e.c

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	54,17	61,11	52,78	100,00	16,67
Standard Deviation	0	0	0	0	0

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	53,79	88,10	83,33	44,44
Standard Deviation	0	0	0	0

Table 8. Summary of Standard Deviation for other manufacturing n.e.c. companies

Construction sector

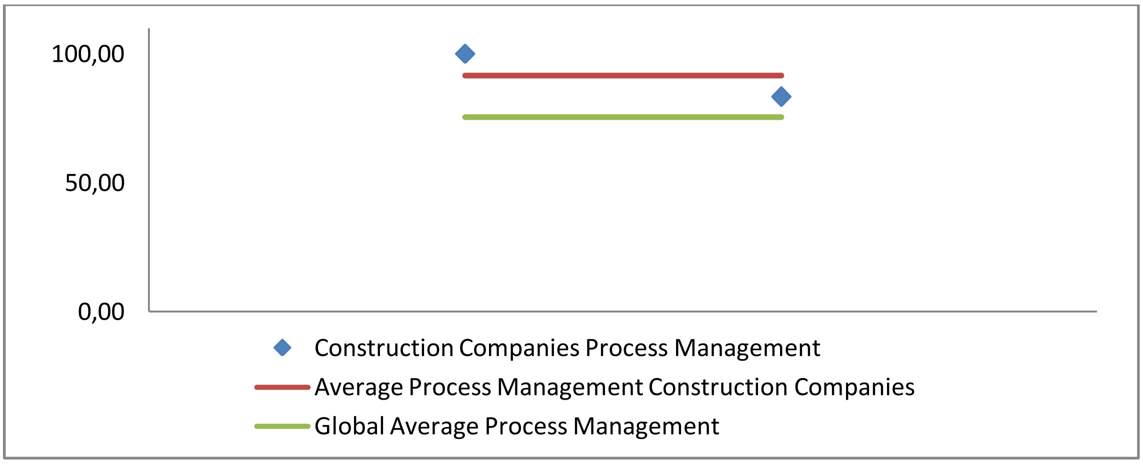

Figure 127. Process Management Area, whole sample with respect of the GA

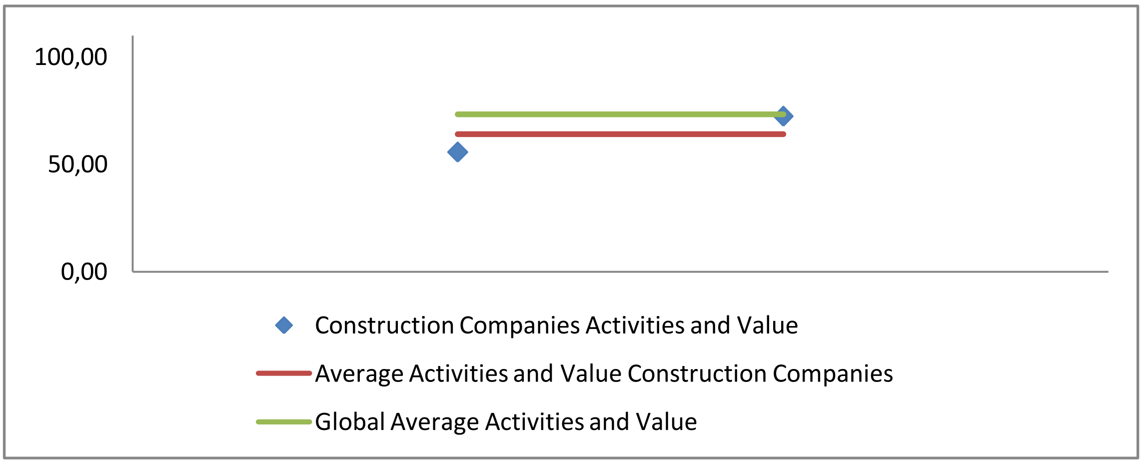

Figure 128. Activities and Value Area, whole sample with respect of the GA

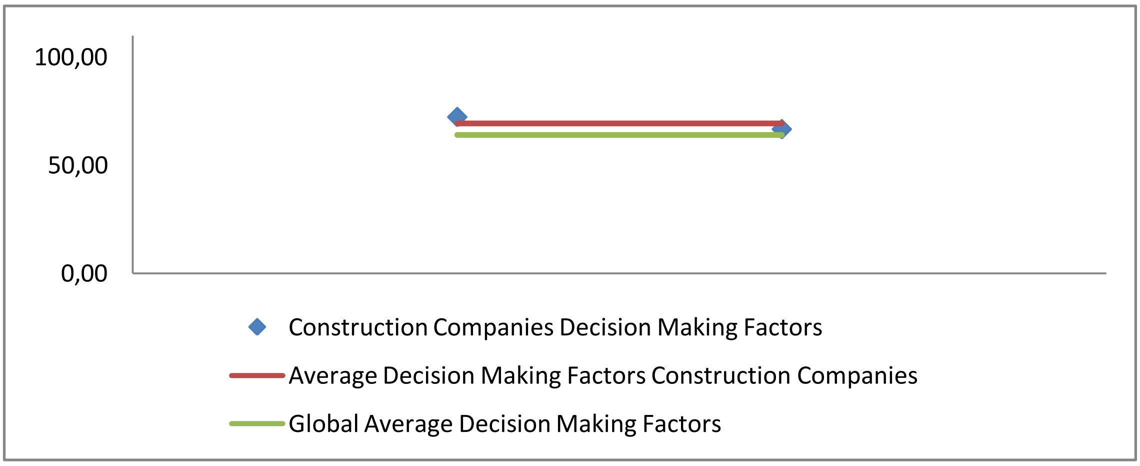

Figure 129. Decision Making Factors Area, whole sample with respect of the GA

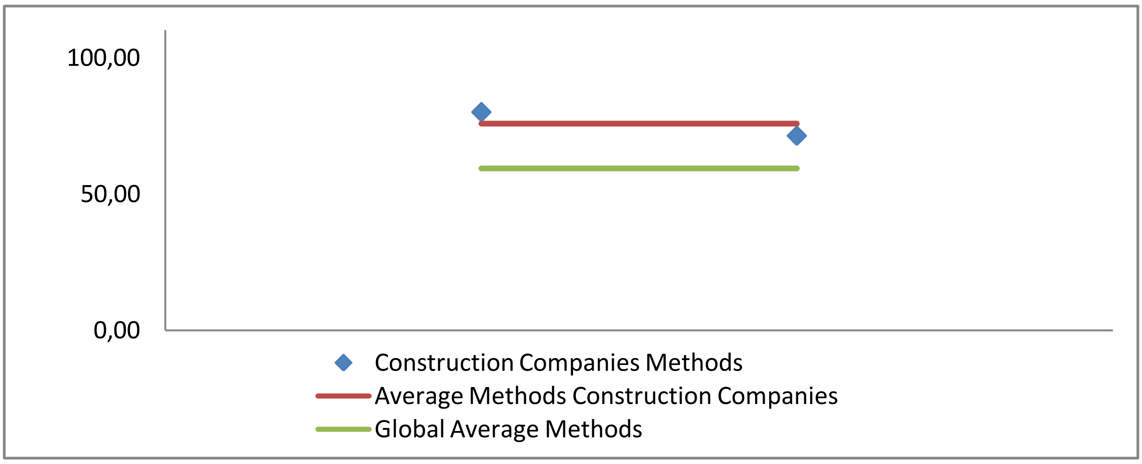

Figure 130. Methods Area, whole sample with respect of the GA

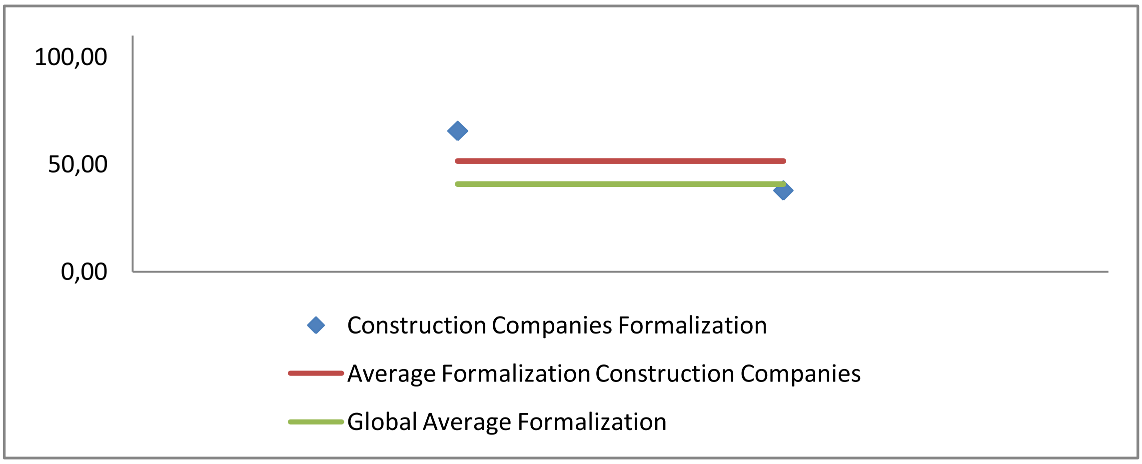

Figure 131. Formalization Area, whole sample with respect of the GA

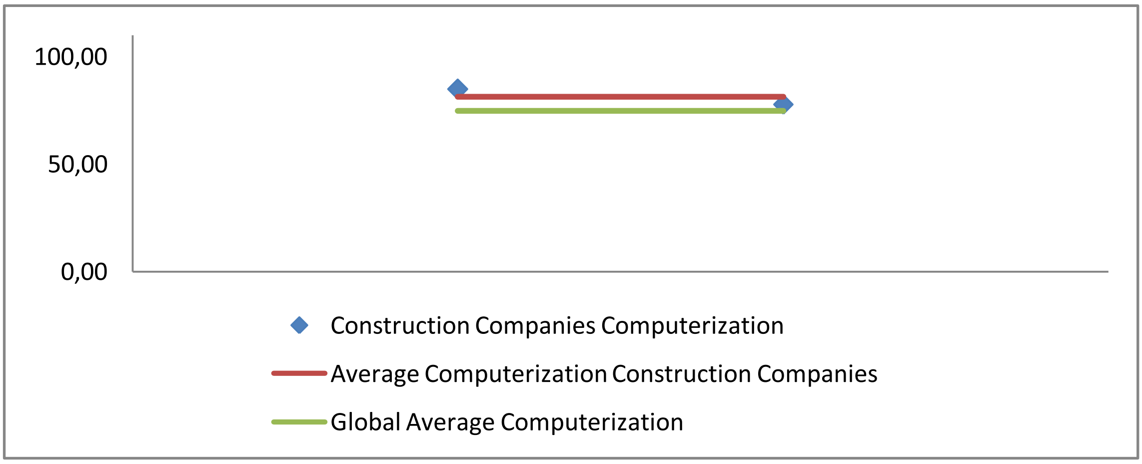

Figure 132. Computerization Area, whole sample with respect of the GA

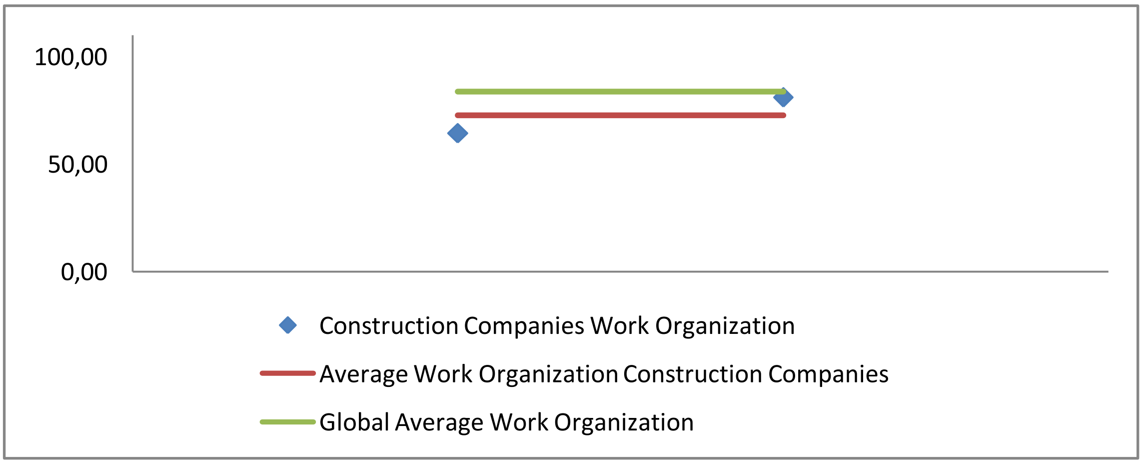

Figure 133. Work Organization Area, whole sample with respect of the GA

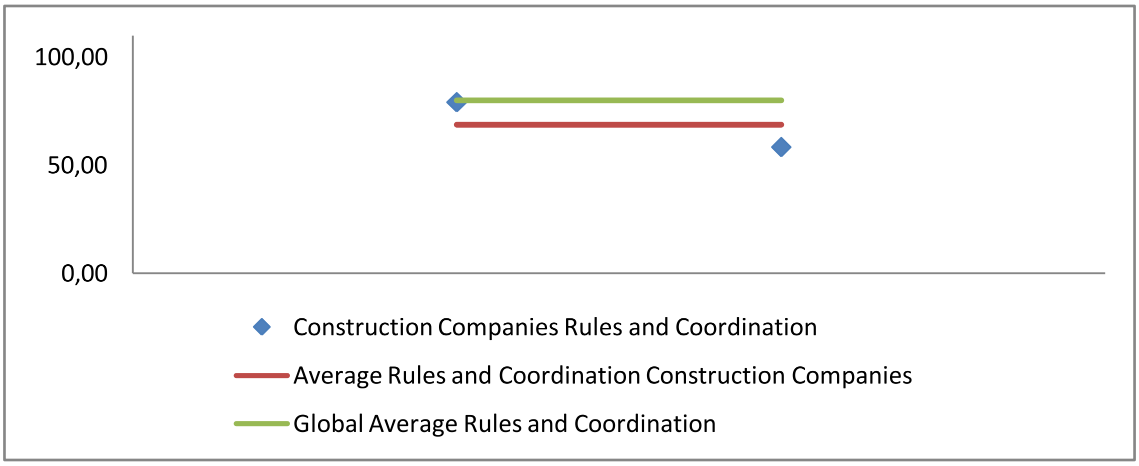

Figure 134. Rules and Coordination Area, whole sample with respect of the GA

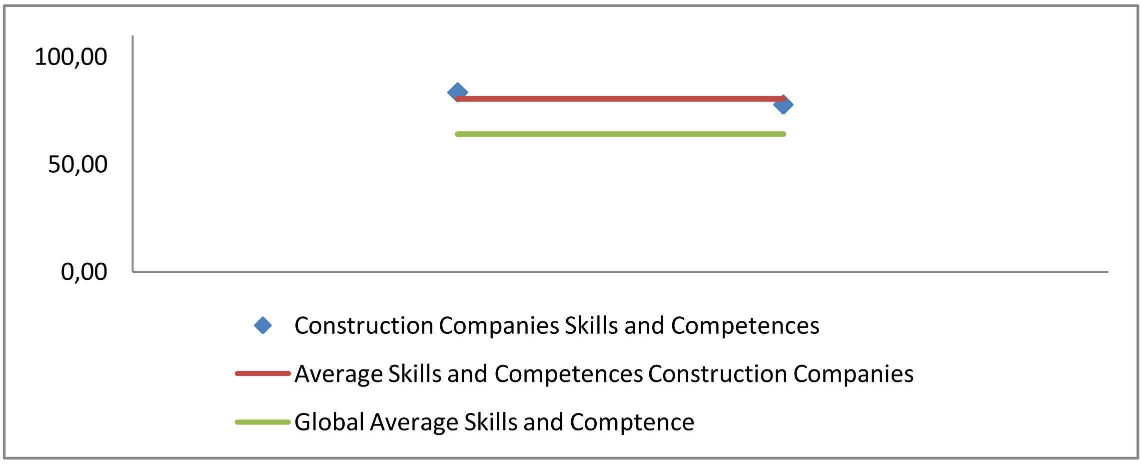

Figure 135. Skills and Competences Area, whole sample with respect of the GA

Specialized design activities sector

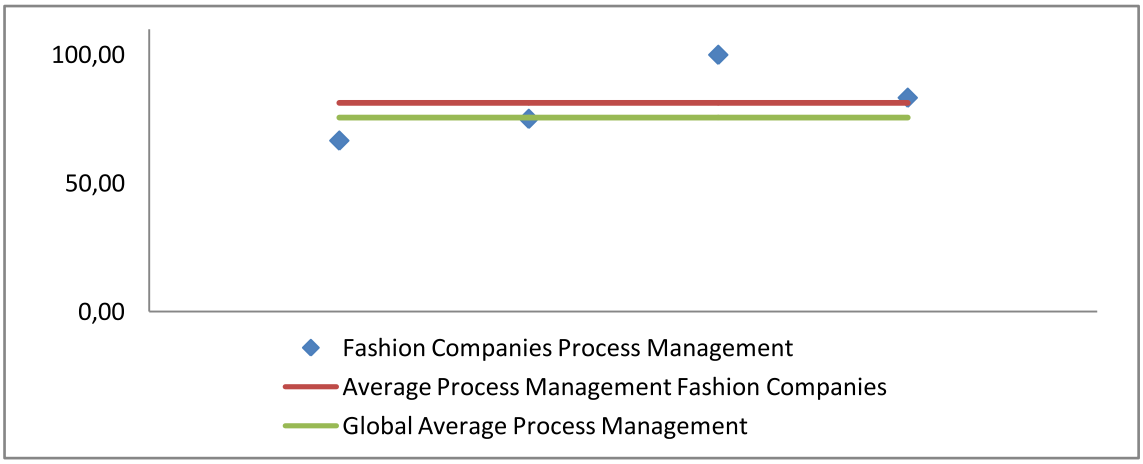

Figure 136. Process Management Area, whole sample with respect of the GA

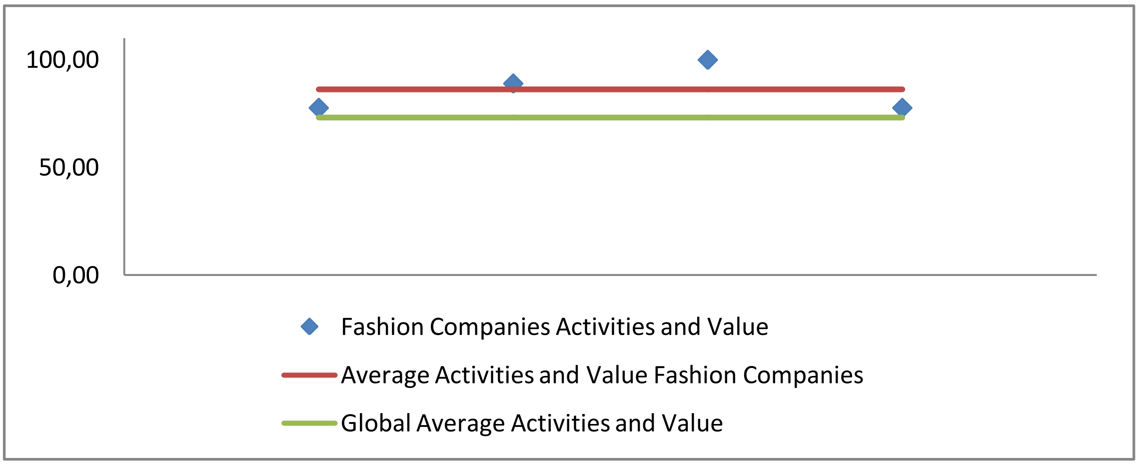

Figure 137. Activities and Value Area, whole sample with respect of the GA

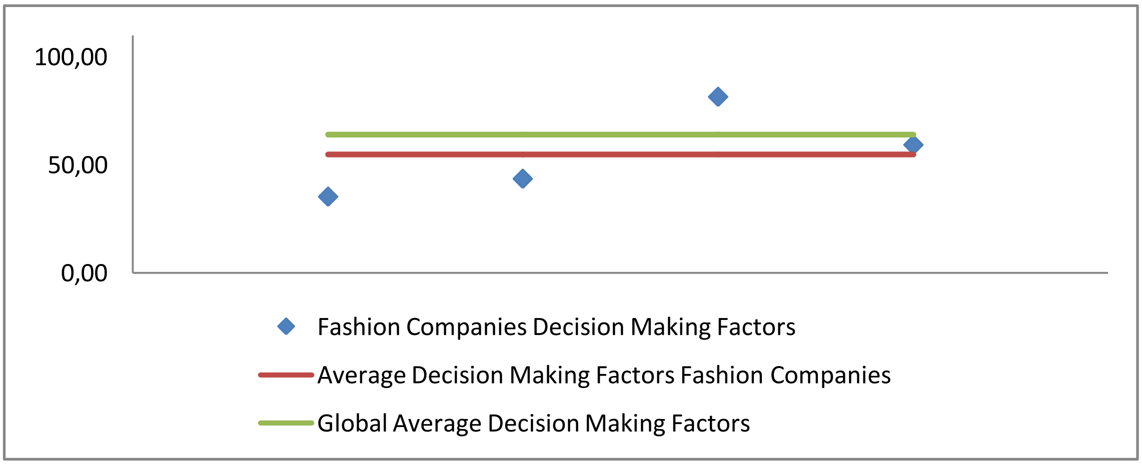

Figure 138. Decision Making Factors Area, whole sample with respect of the GA

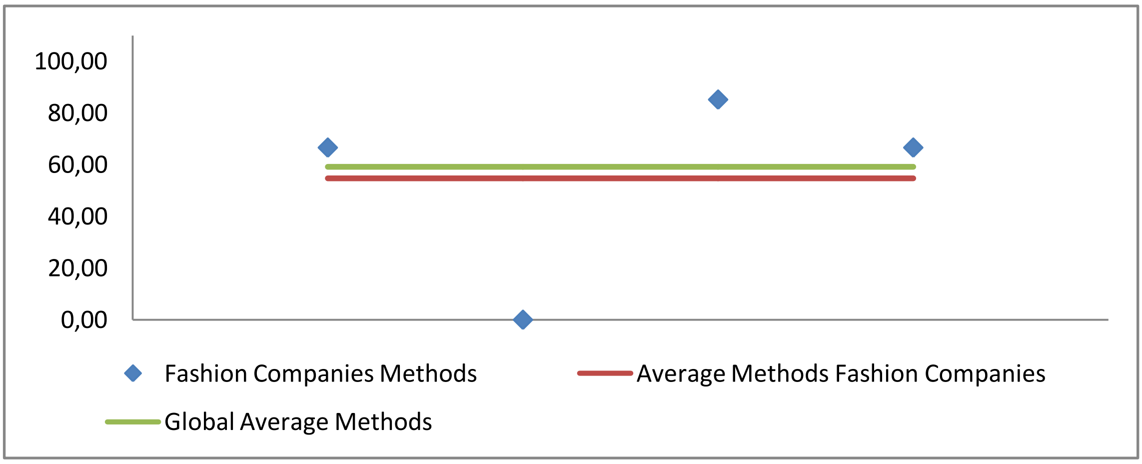

Figure 139. Methods Area, whole sample with respect of the GA

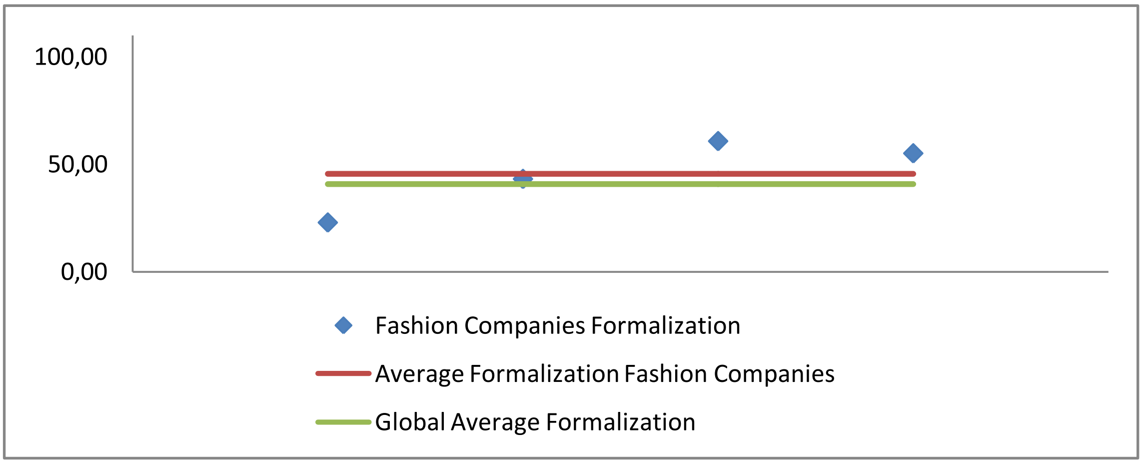

Figure 140. Formalization Area, whole sample with respect of the GA

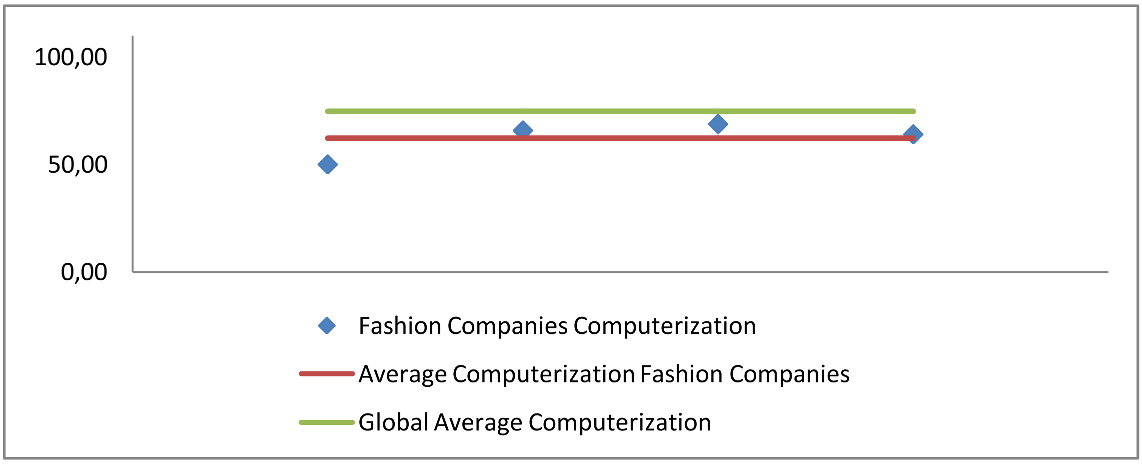

Figure 141. Computerization Area, whole sample with respect of the GA

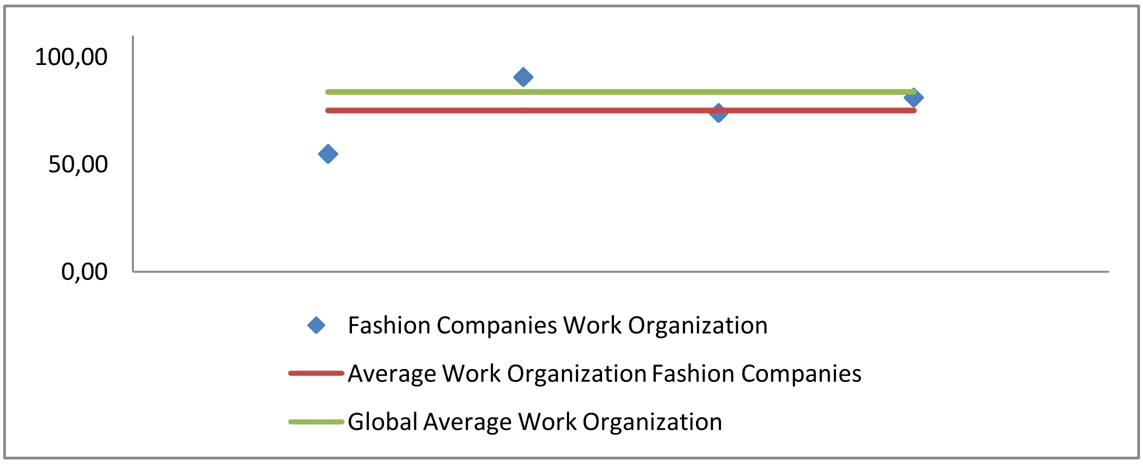

Figure 142. Work Organization Area, whole sample with respect of the GA

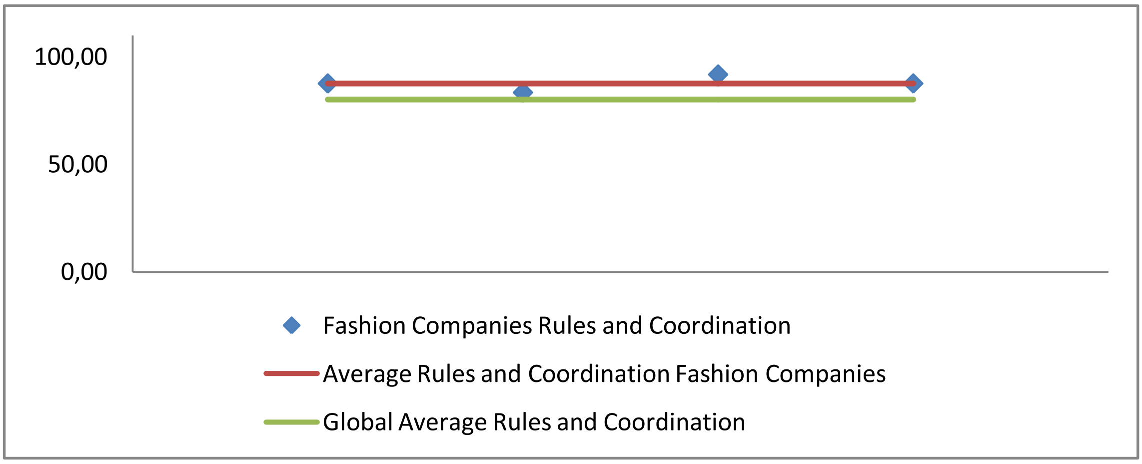

Figure 143. Rules and Coordination Area, whole sample with respect of the GA

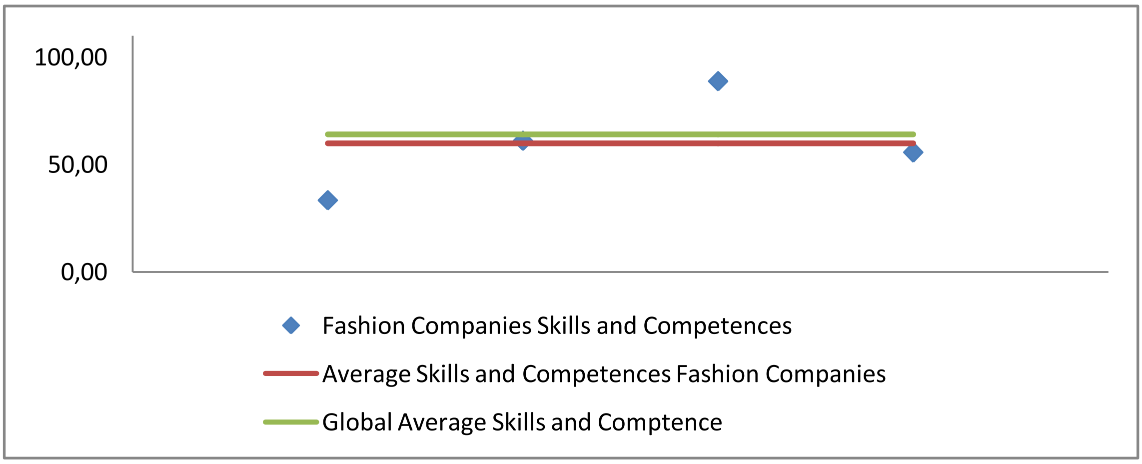

Figure 144. Skills and Competences Area, whole sample with respect of the GA

Standard Deviation

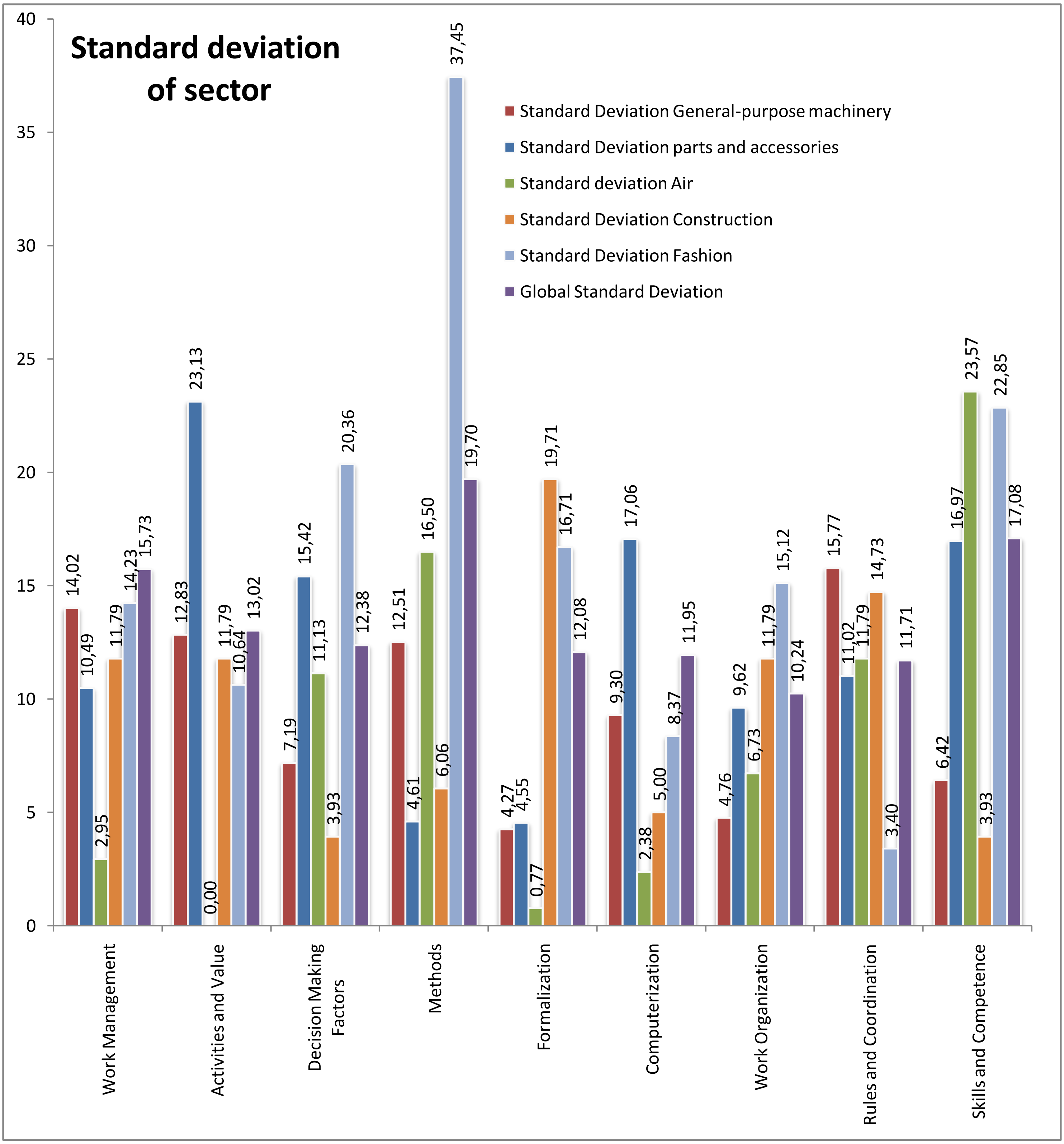

Figure 145. Standard deviation depending of the sector

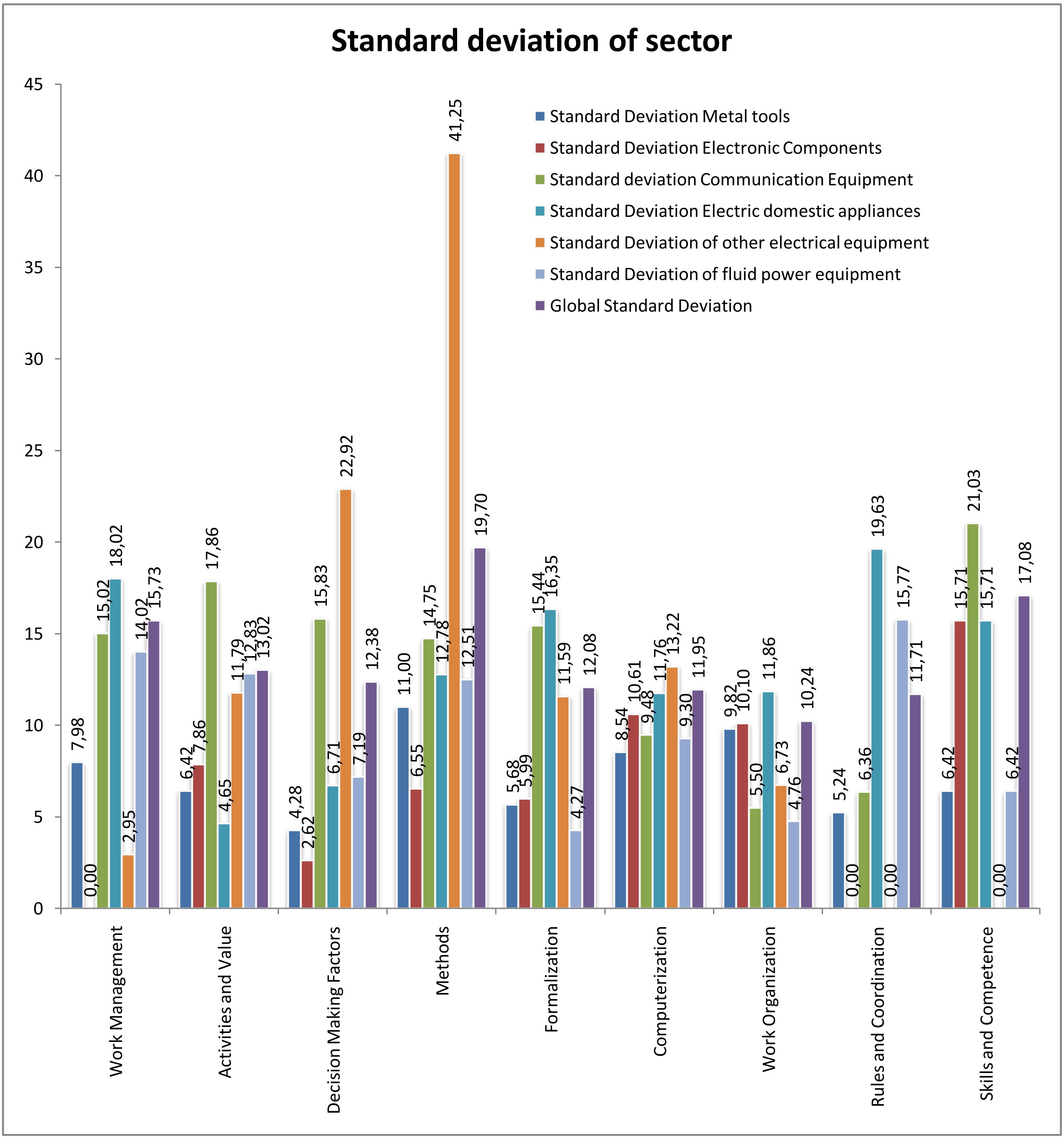

Figure 146. Standard deviation depending of the sector

Summary of manufacture of metal tools sector

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	81,25	83,33	69,44	62,08	36,61
Standard Deviation	7,98	6,42	4,28	11,00	5,68

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	78,59	89,29	84,38	61,11
Standard Deviation	8,54	9,82	5,24	6,42

Table 9. Summary of Standard Deviation for metal tools companies

Summary of manufacture of electronic components sector

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	66,67	66,67	75,00	56,48	41,12
Standard Deviation	0	7,86	2,62	6,55	5,99

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	73,07	80,95	70,83	83,33
Standard Deviation	10,61	10,10	0	15,71

Table 10. Summary of Standard Deviation for electronic components companies

Summary of manufacture of communication equipment sector

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	87,50	74,07	70,99	58,21	44,35
Standard Deviation	15,02	17,86	15,83	14,75	15,44

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	80,96	84,13	81,94	74,07
Standard Deviation	9,48	5,50	6,36	21,03

Table 11. Summary of Standard Deviation for communication equipment companies

Summary of manufacture of electric domestic appliances sector

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	78,33	71,11	60,19	46,00	36,61
Standard Deviation	18,02	4,65	6,71	12,78	16,35

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	71,21	89,52	84,17	55,56
Standard Deviation	11,76	11,86	19,63	15,71

Table 12. Summary of Standard Deviation for electric domestic appliances companies

Summary of manufacture of other electrical equipment sector

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	72,92	75,00	55,09	70,83	42,08
Standard Deviation	2,95	11,79	22,92	41,25	11,59

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	84,84	80,95	70,83	66,67
Standard Deviation	13,22	6,73	0	0

Table 13. Summary of Standard Deviation for other electrical equipment companies

Summary of manufacture of fluid power equipment sector

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	65,83	70,37	63,27	61,01	40,98
Standard Deviation	14,02	12,83	7,19	12,51	4,27

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	77,39	83,33	80,56	59,26
Standard Deviation	9,30	4,76	15,77	6,42

Table 14. Summary of Standard Deviation fluid power equipment companies

Summary of manufacture of other parts and accessories sector

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	51,39	59,26	61,73	55,30	36,79
Standard Deviation	10,49	23,13	15,42	4,61	4,55

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	70,41	84,13	70,83	51,85
Standard Deviation	17,06	9,62	11,02	16,97

Table 15. Summary of Standard Deviation for other parts and accessories companies

Summary of manufacture of air and spacecraft sector

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	85,42	77,78	70,83	55,00	48,09
Standard Deviation	2,95	0	11,13	16,50	0,77

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	88,32	90,48	83,33	83,33
Standard Deviation	2,38	6,73	11,79	23,57

Table 16. Summary of Standard Deviation for air and spacecraft companies

Summary of Construction sector

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	91,67	63,89	69,44	75,71	51,64
Standard Deviation	11,79	11,79	3,93	6,06	19,71

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	81,25	72,62	68,75	80,56
Standard Deviation	5,00	11,79	14,73	3,93

Table 17. Summary of Standard Deviation for construction companies

Summary of Specialized design activities sector

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	81,25	86,11	54,86	54,63	45,49
Standard Deviation	14,23	10,64	20,36	37,45	16,71

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	62,18	75,00	87,50	59,72
Standard Deviation	8,37	15,12	3,40	22,85

Table 18. Summary of Standard Deviation for specialized design companies

The higher level of maturity is achieved in the Organization part, area of Work Organization, for the majority number of the type of companies in which average scores overpass the 80%. We see that only the two last group show level of maturity under 80%. The dispersion of values in this area is small for all groups of companies. Only which works in construction and fashion sector have a big dispersion and for that we cannot say anything about their behavior. The other two areas of Work Organization present different values. While Rules and Coordination has a good value (around 80%) for all groups (only construction sector show a little value, 68,75%), the other one Skills and Competence presents low values and dispersion values. The dispersion of values in this area is bigger than the other two areas, in fact many cases arrives to 24%. It is important see that behavior is similar with the global report. A lot of sectors are interested in the skills and competence of their employees and for that it can be possible find some cases with mature practice. For the other case even if companies understand that enhancement of individual skills and competences results in a more agile and mature organization, with better product performance, the achieved maturity level is still Intermediate and, in few cases, Low.

If we study Process Management area we can see that there is a different between all the companies. We see that construction sector has the higher level of maturity (91,67%). But for the other case the values are very different. It is important see the companies in the air sector where have a elevate level of maturity (near 73%) and a SD very insignificant (near 3%). We understand with this value that companies have the same behavior. Standard Deviation is higher in the other areas and very dispersive, so we cannot create a relevant date of this value.

Moreover, the NPD process requires a large number of decisions to be taken every day considering both internal (Decision Making Factors) and external (Activities and Value) factors. Interviewed companies are actually paying less attention on customer value. In average, lower attention is paid on the whole product lifecycle, which not heavily affects companies' strategy and early decisions: most of the companies seem to be not aware of the simple fact that the decisions taken at the early NPD stages affect the whole product lifecycle. The lower level of maturity corresponds to the area Decision Making Factors. In spite of, standard deviation is very high.

The third investigated part is Knowledge Management, which results to be still precarious. In fact, even if the level of Computerization is in average Mature (only fashion companies present a lower value with the regard that others groups) , the

Formalization is still in the middle between Low and Intermediate. Companies know that maintaining and protecting the internal know-how is crucial.

2.2.3.Average and Standard Deviation depending on the type of market

Make to Order

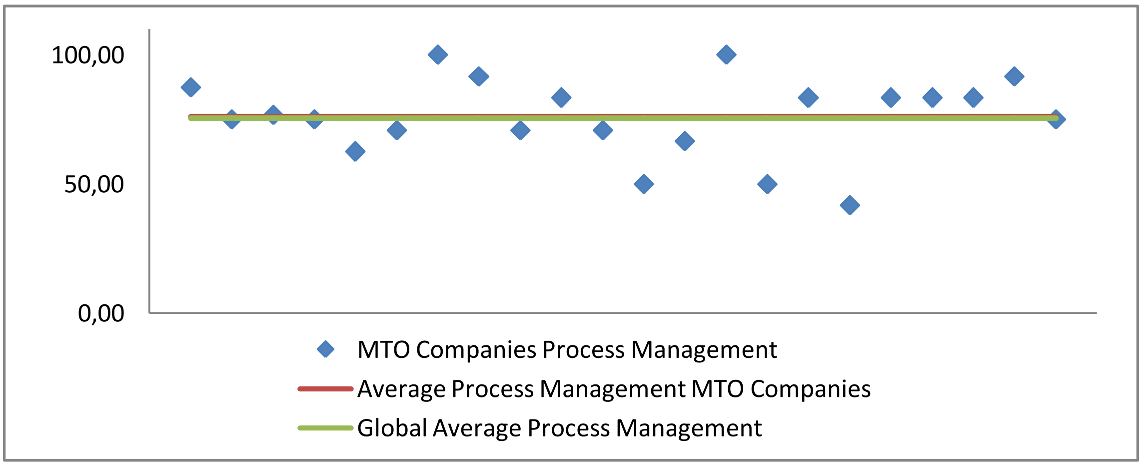

Figure 147. Process Management Area, whole sample with respect of the GA

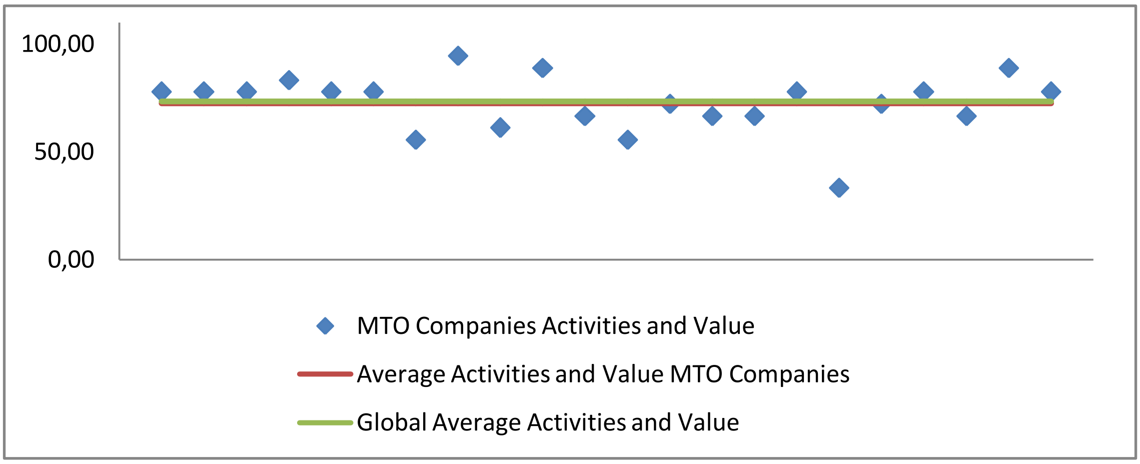

Figure 148. Activities and Value Area, whole sample with respect of the GA

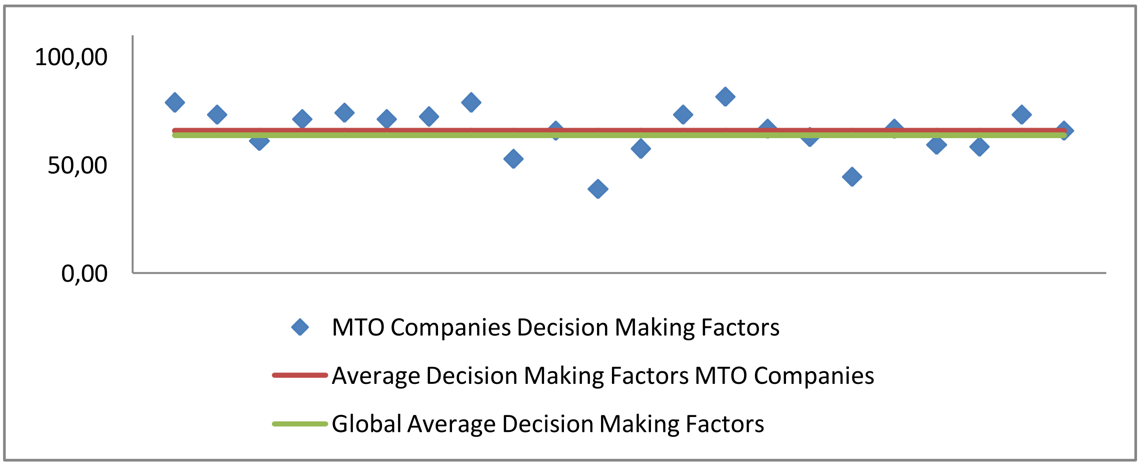

Figure 149. Decision Making Factors Area, whole sample with respect of the GA

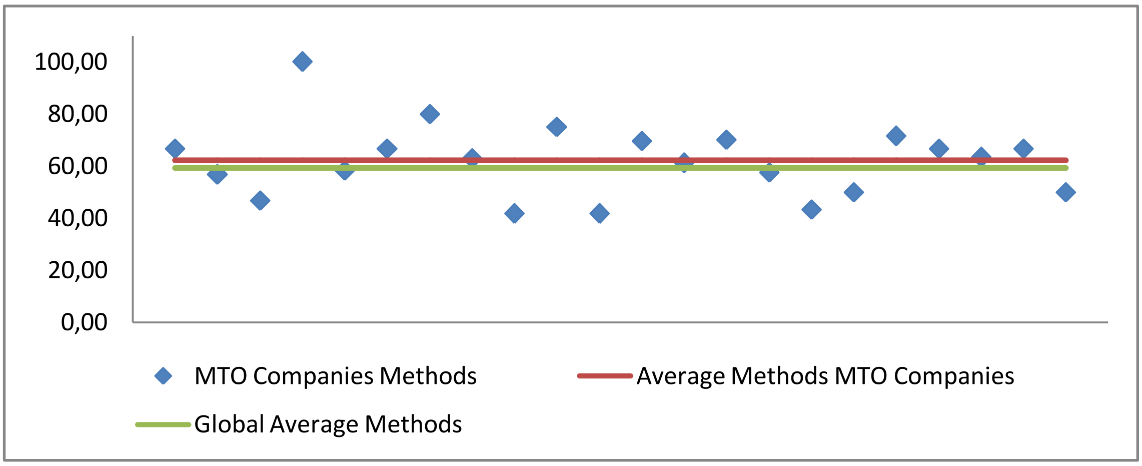

Figure 150. Methods Area, whole sample with respect of the GA

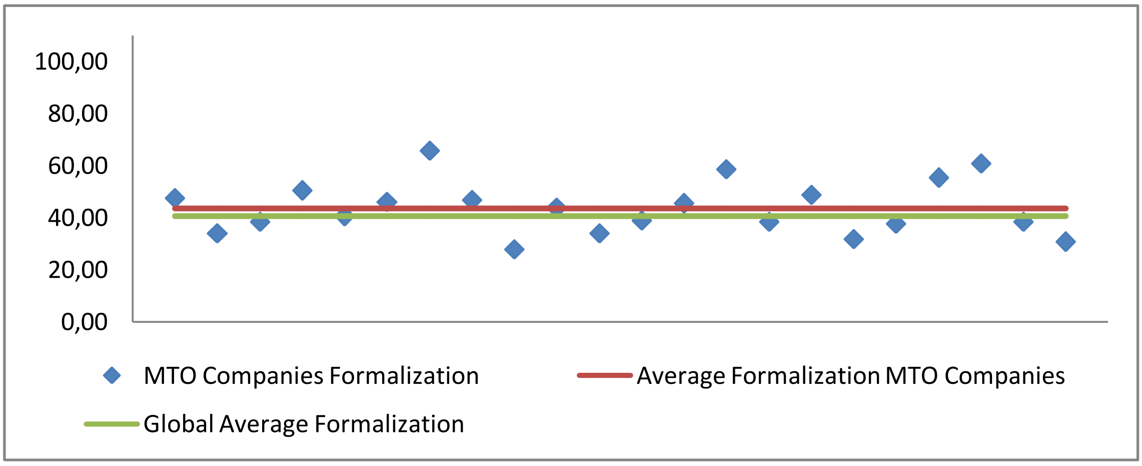

Figure 151. Formalization Area, whole sample with respect of the GA

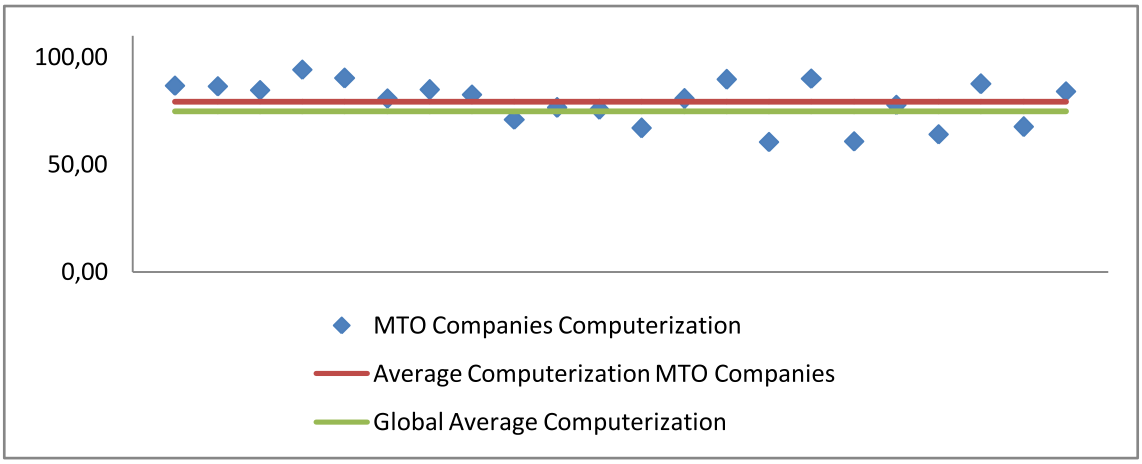

Figure 152. Computerization Area, whole sample with respect of the GA

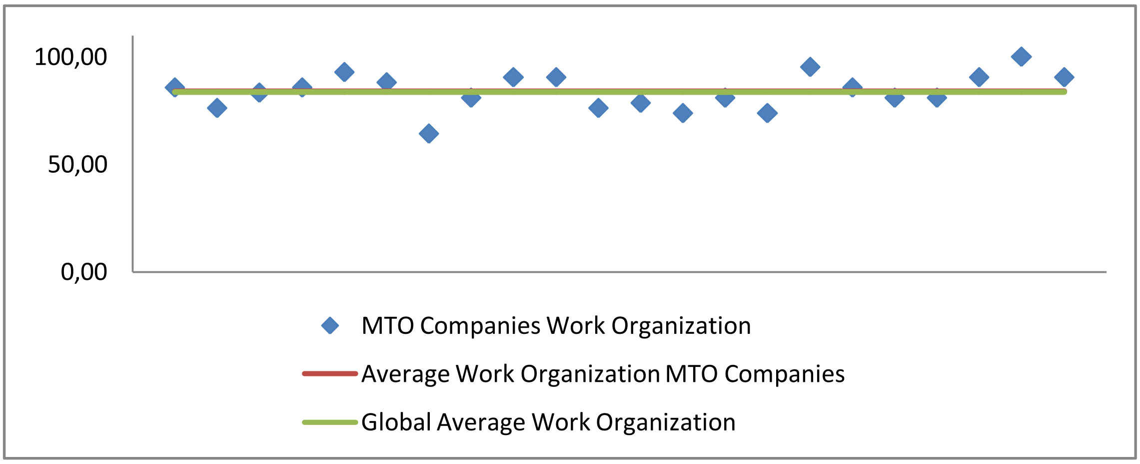

Figure 153. Work Organization Area, whole sample with respect of the GA

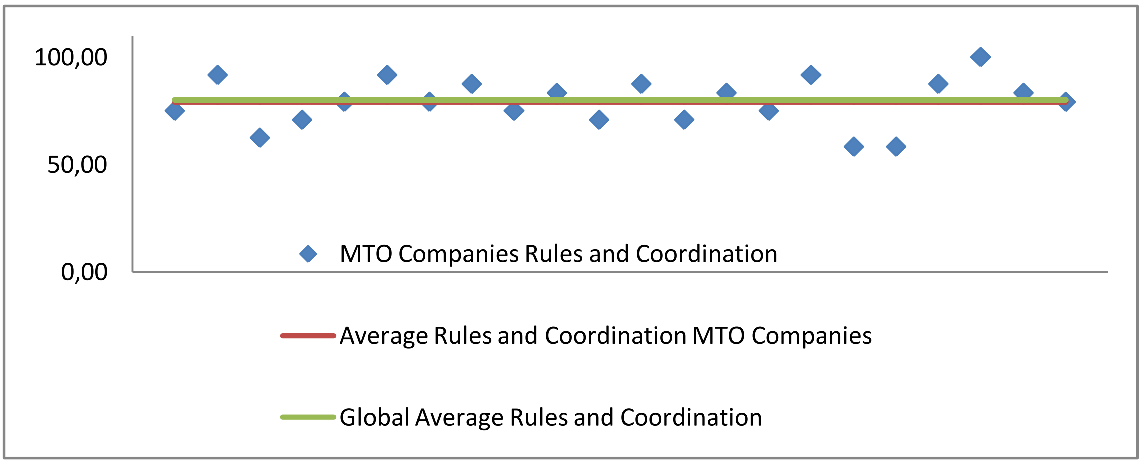

Figure 154. Rules and Coordination Area, whole sample with respect of the GA

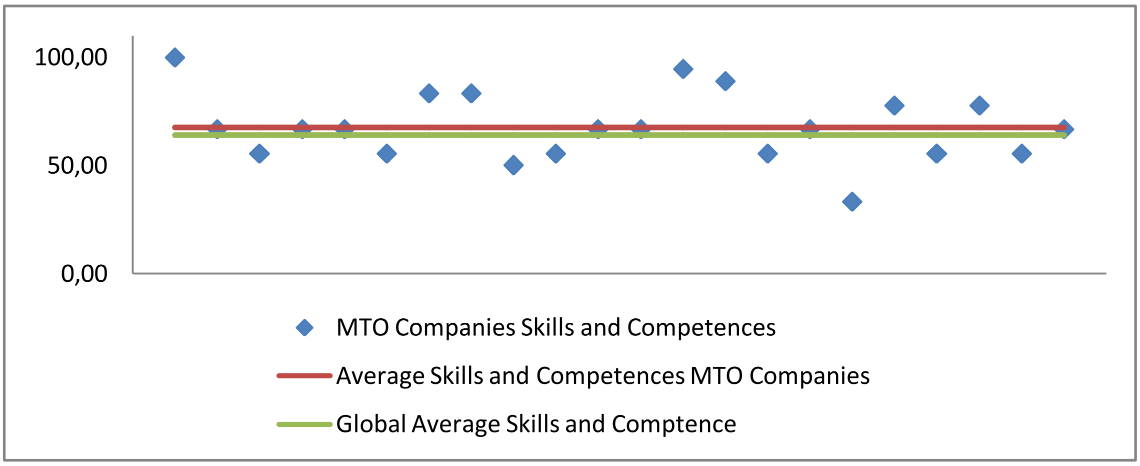

Figure 155. Skills and Competences Area, whole sample with respect of the GA

Summary of MTO companies

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	76,02	72,47	65,78	62,10	43,53
Standard Deviation	15,32	13,33	10,83	13,80	10,08

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	79,15	83,87	79,17	67,68
Standard Deviation	10,12	8,28	11,06	15,87

Table 19. Summary of Standard Deviation for MTO companies

MTS Companies

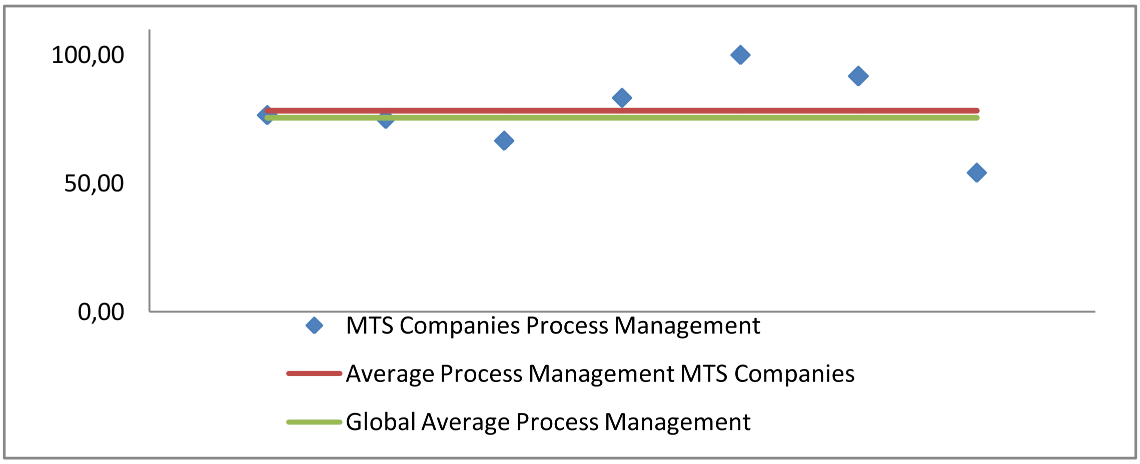

Figure 156. Process Management Area, whole sample with respect of the GA

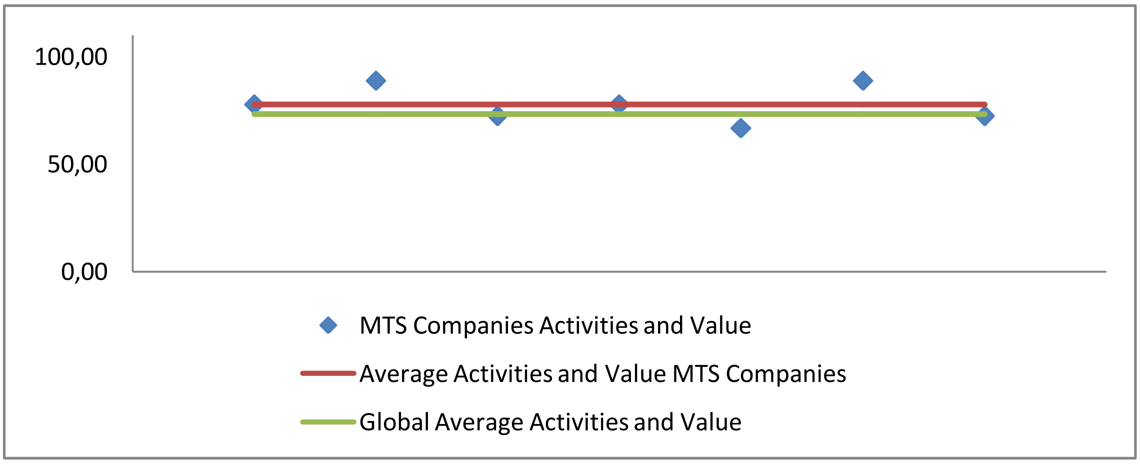

Figure 157. Activities and Value Area, whole sample with respect of the GA

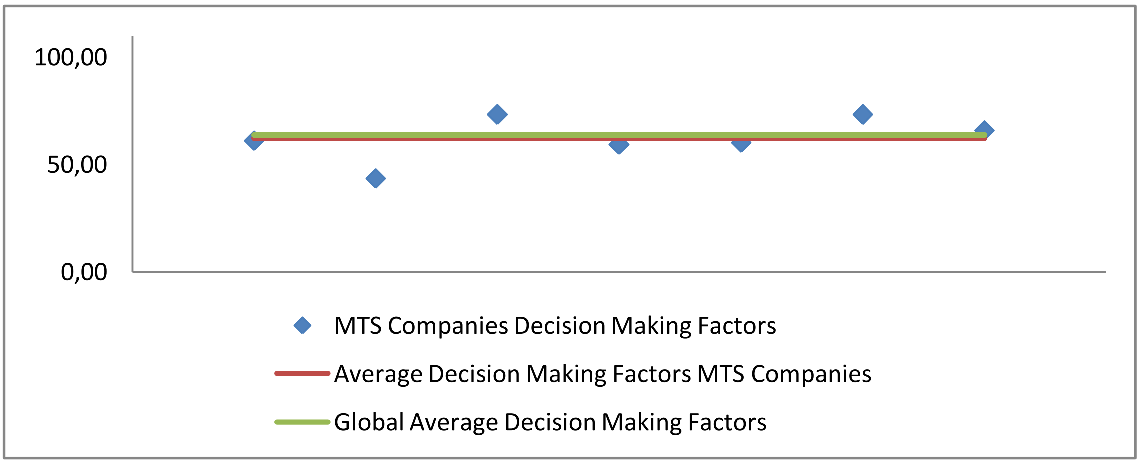

Figure 158. Decision Making Factors Area, whole sample with respect of the GA

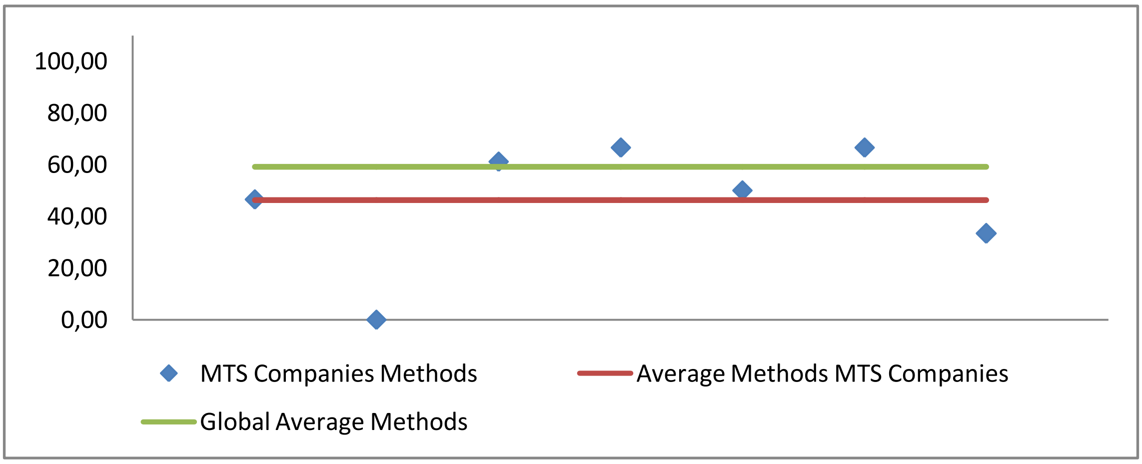

Figure 159. Methods Area, whole sample with respect of the GA

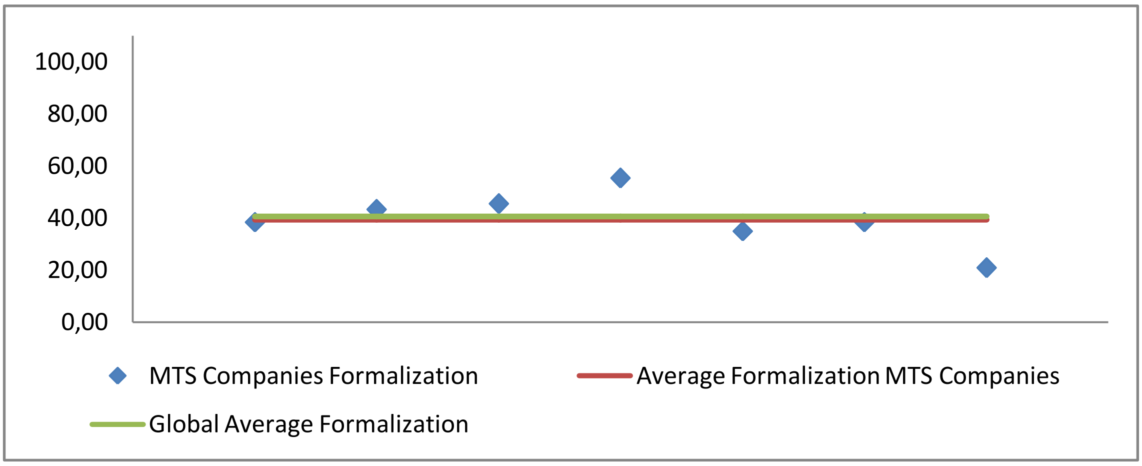

Figure 160. Formalization Area, whole sample with respect of the GA

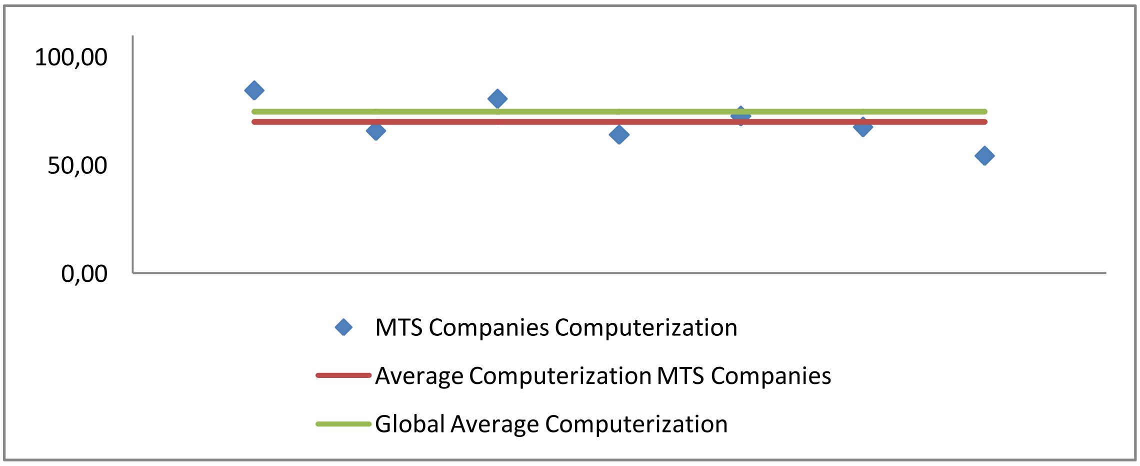

Figure 161. Computerization Area, whole sample with respect of the GA

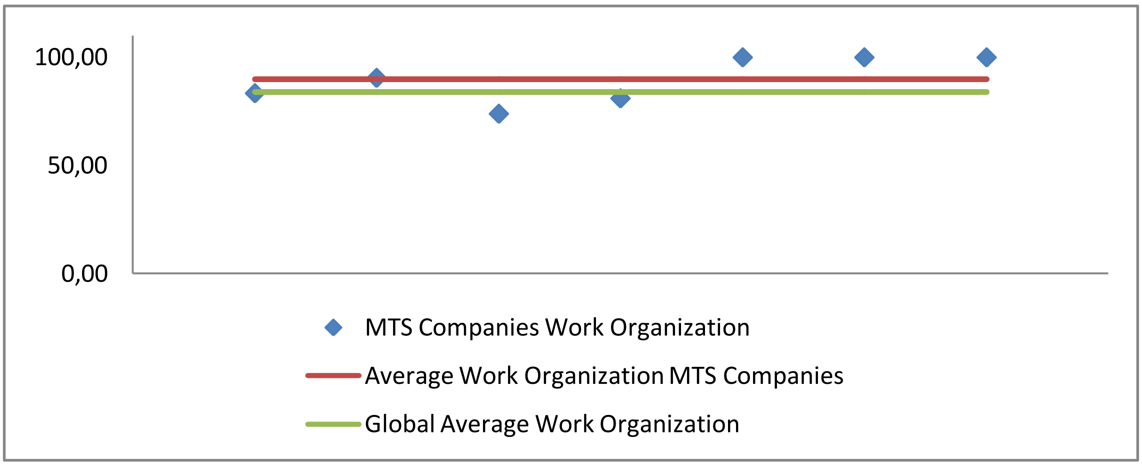

Figure 162.Work Organization Area, whole sample with respect of the GA

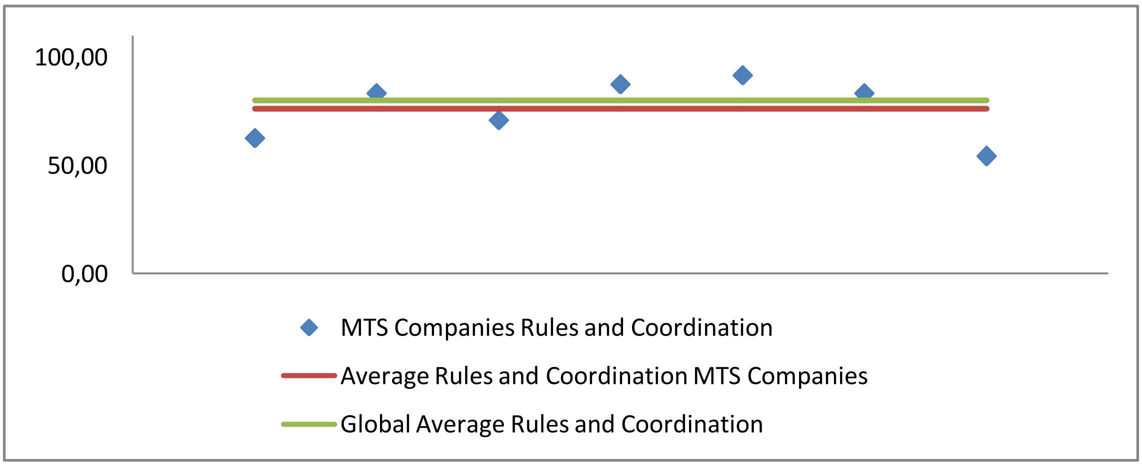

Figure 163. Rules and Competence Area, whole sample with respect of the GA

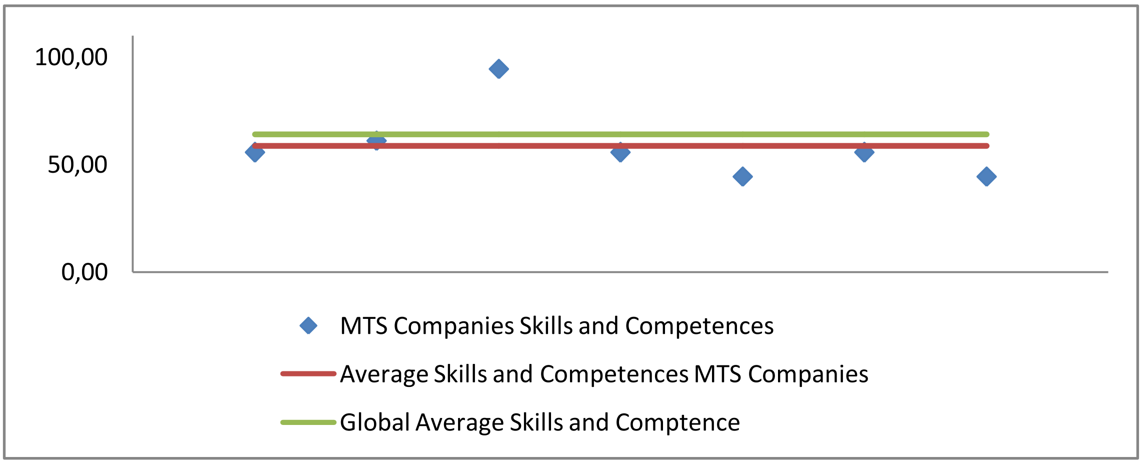

Figure 164. Skills and Competence Area, whole sample with respect of the GA

Summary of MTS companies

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	78,21	77,78	62,30	46,35	39,42
Standard Deviation	15,31	8,49	10,12	23,72	10,56

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	69,90	89,80	76,19	58,73
Standard Deviation	10,30	10,71	13,97	16,93

Table 20. Summary of Standard Deviation for MTS companies

Catalogue

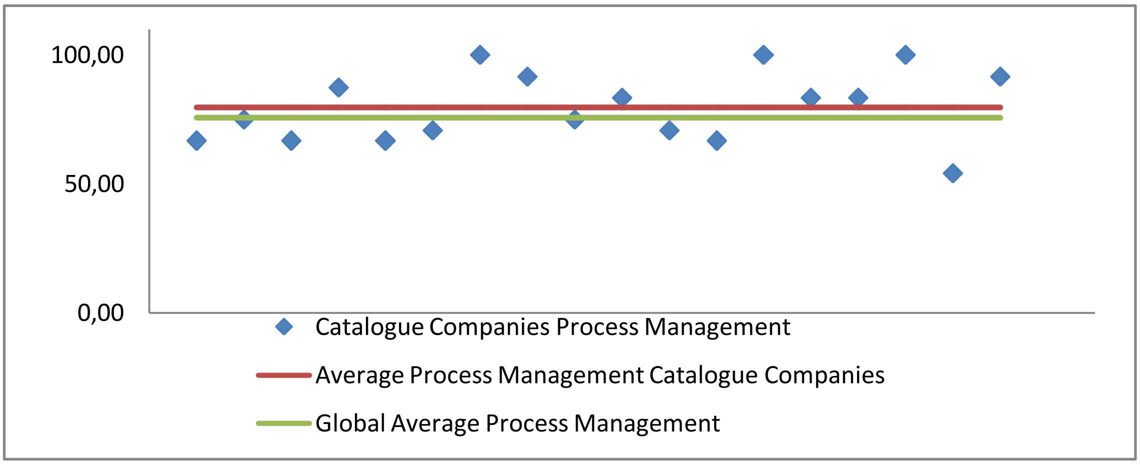

Figure 165. Process Management Area, whole sample with respect of the GA

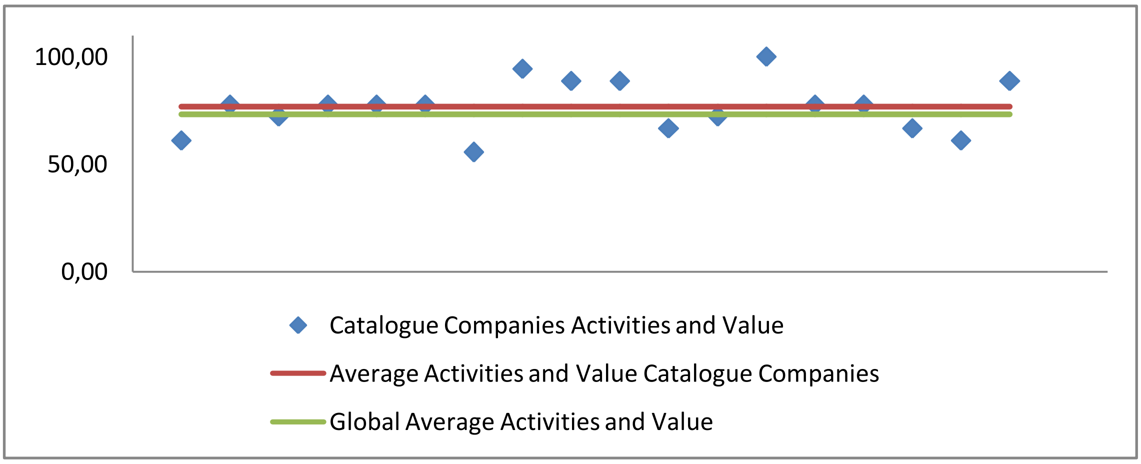

Figure 166. Activities and Values Area, whole sample with respect of the GA

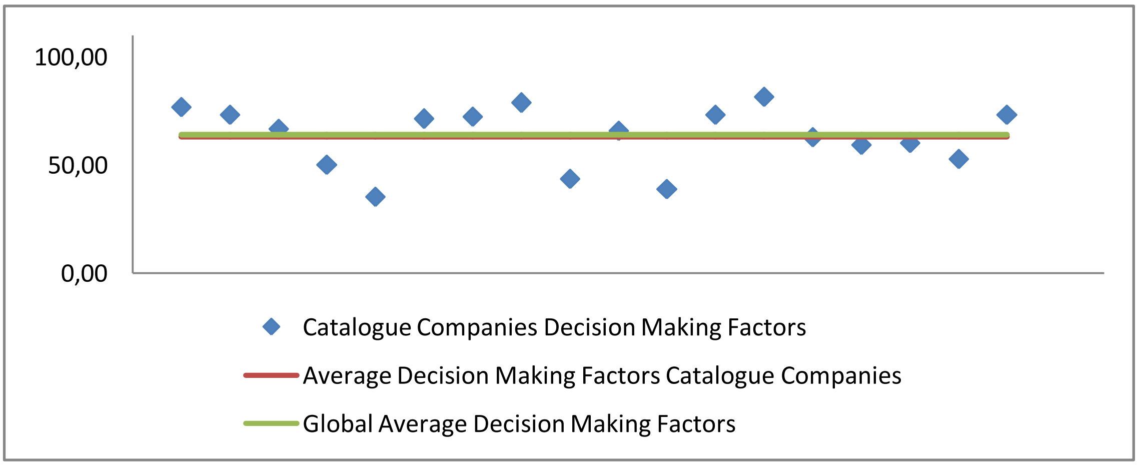

Figure 167. Decision Making Factors Area, whole sample with respect of the GA

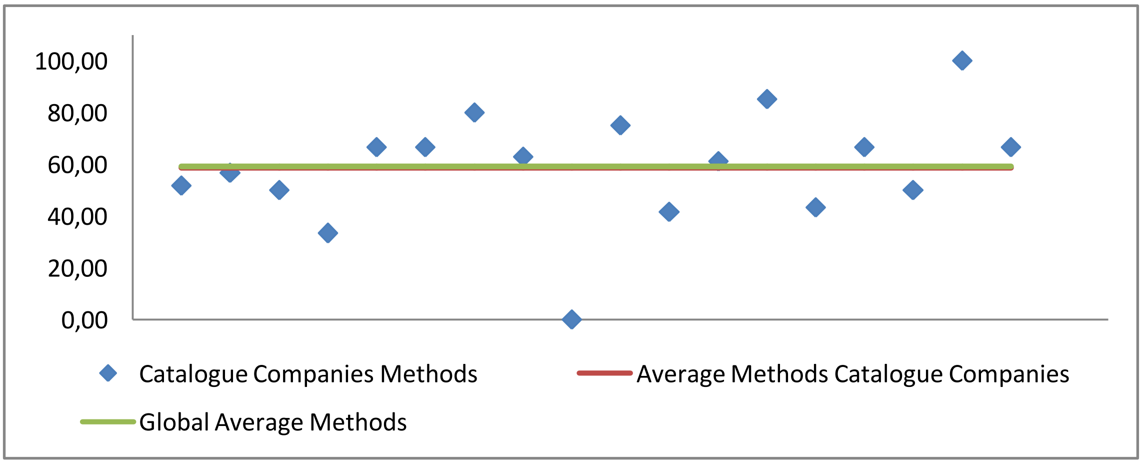

Figure 168. Methods Area, whole sample with respect of the GA

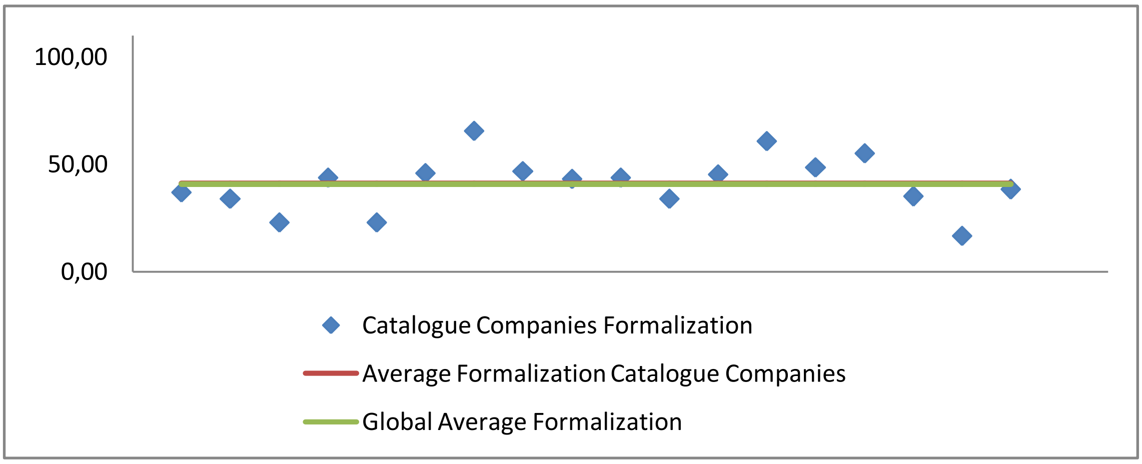

Figure 169. Formalization Area, whole sample with respect of the GA

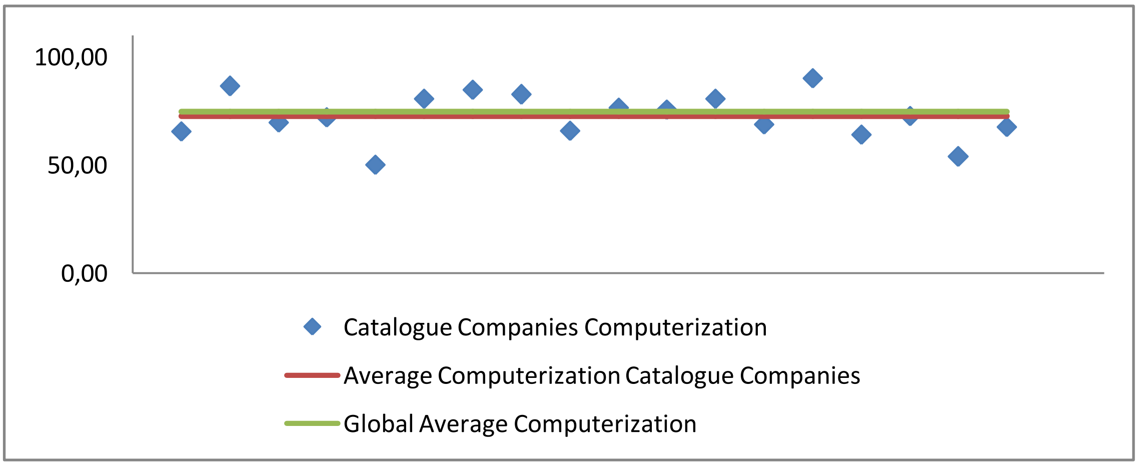

Figure 170. Computerization Area, whole sample with respect of the GA

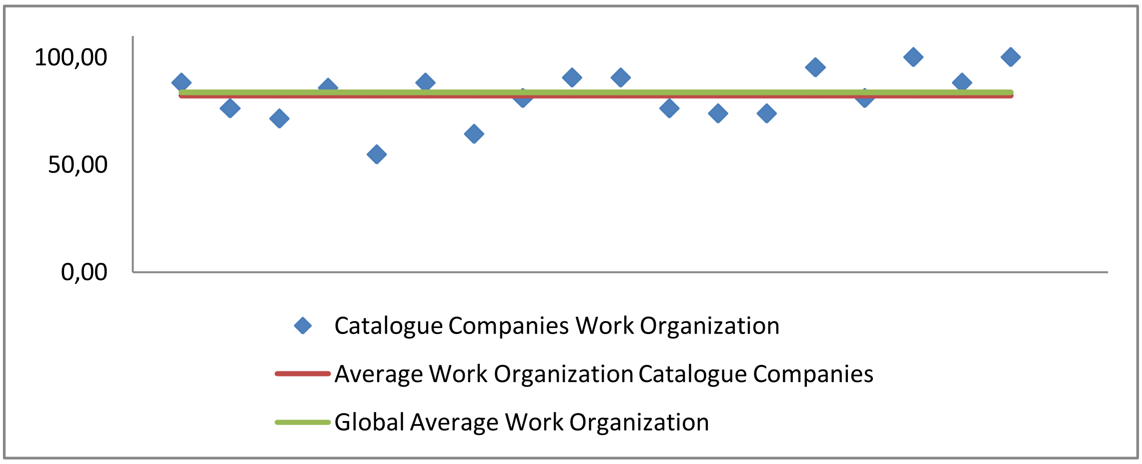

Figure 171. Work Organization Area, whole sample with respect of the GA

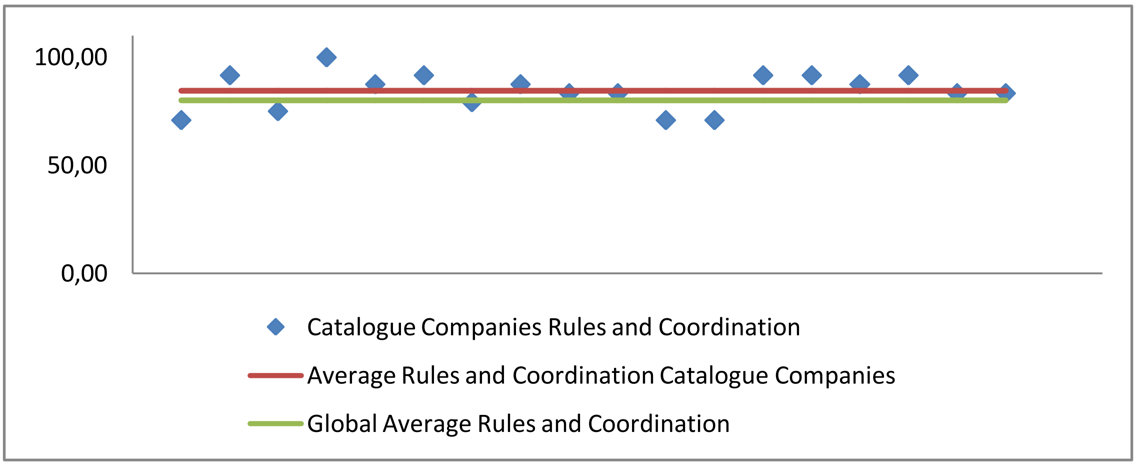

Figure 172. Rules and Coordination Area, whole sample with respect of the GA

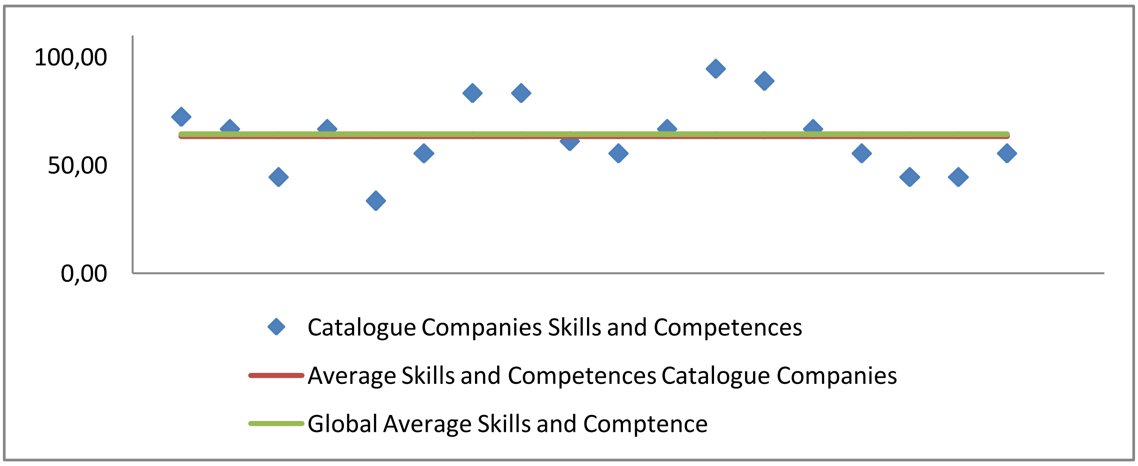

Figure 173. Skills and Competence Area, whole sample with respect of the GA

Summary of Catalogue companies

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	79,63	76,85	63,07	58,77	41,06
SD	13,62	12,09	13,92	21,99	12,72

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	72,60	82,14	84,49	63,27
SD	10,78	12,04	8,43	16,74

Table 21. Summary of Standard Deviation for Catalogue companies

Shelf

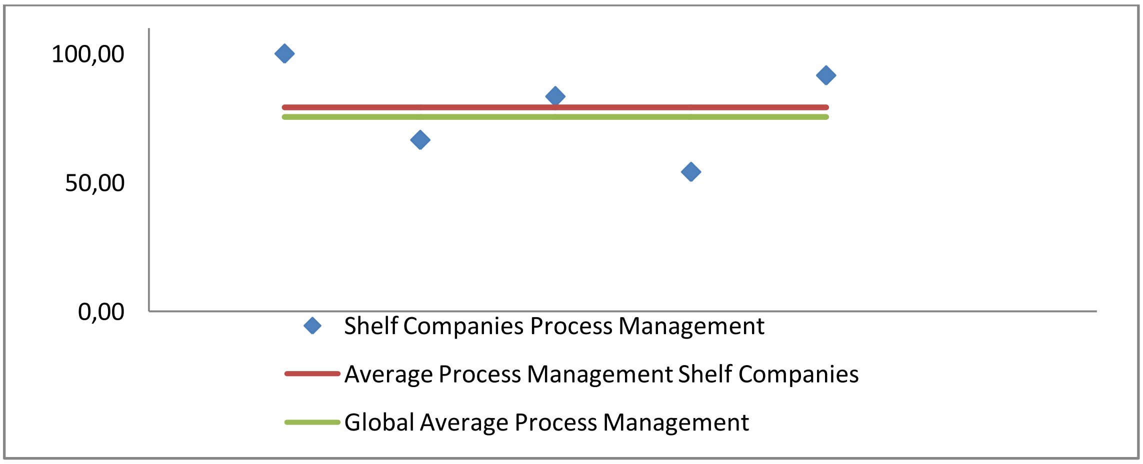

Figure 174. Process Management Area, whole sample with respect of the GA

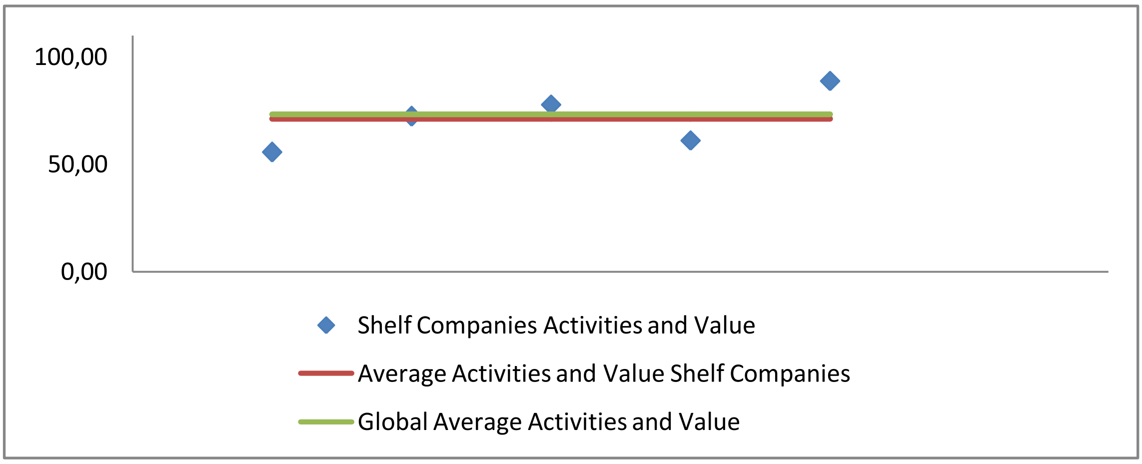

Figure 175. Activities and Value Area, whole sample with respect of the GA

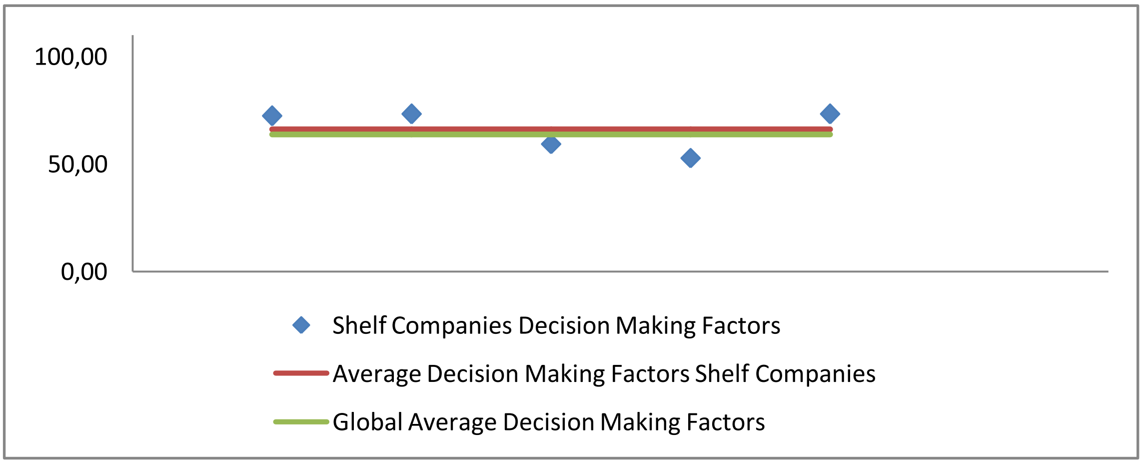

Figure 176. Decision Making Factors Area, whole sample with respect of the GA

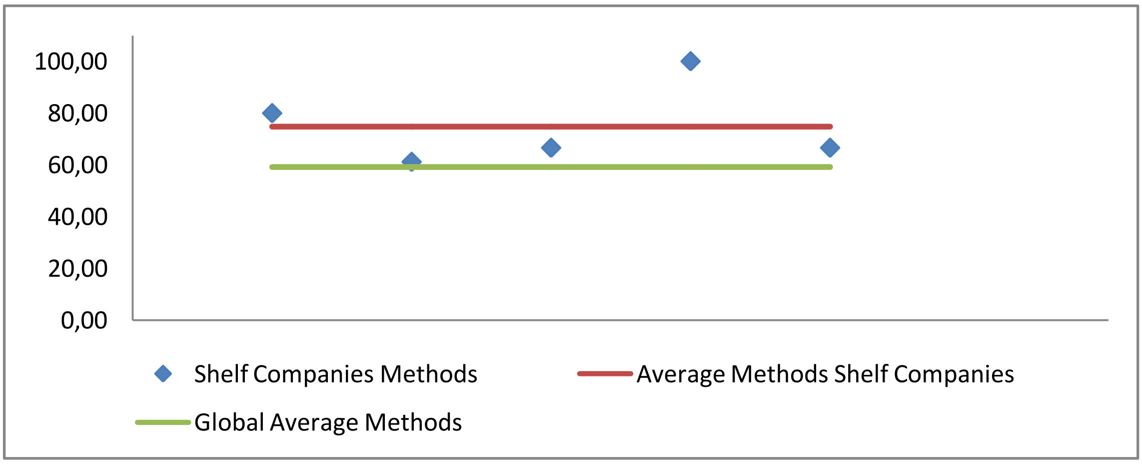

Figure 177. Methods Area, whole sample with respect of the GA

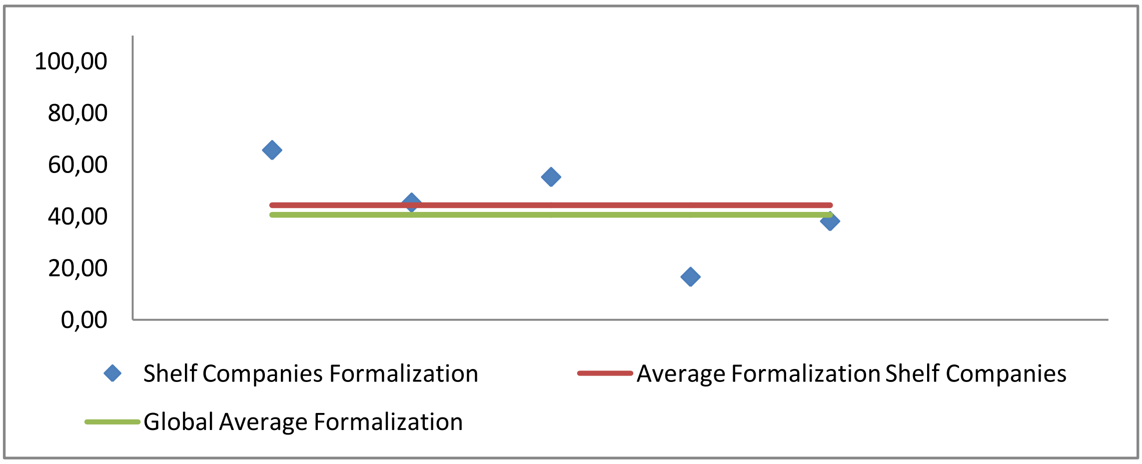

Figure 178. Formalization Area, whole sample with respect of the GA

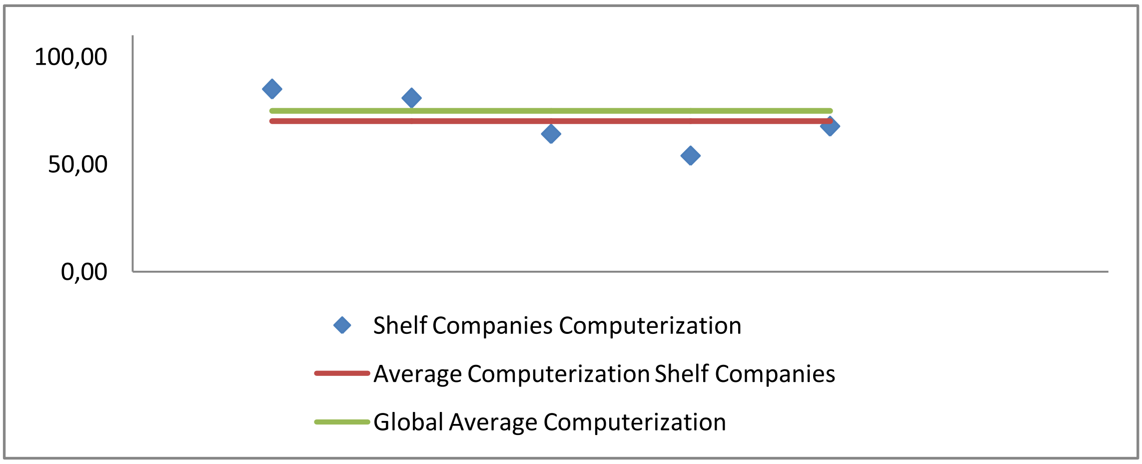

Figure 179. Computerization Area, whole sample with respect of the GA

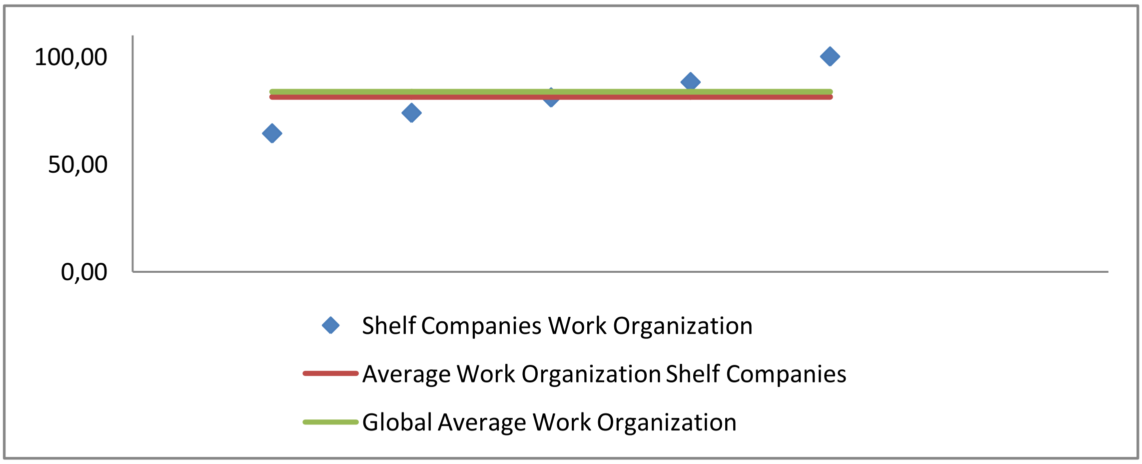

Figure 180. Work Organization Area, whole sample with respect of the GA

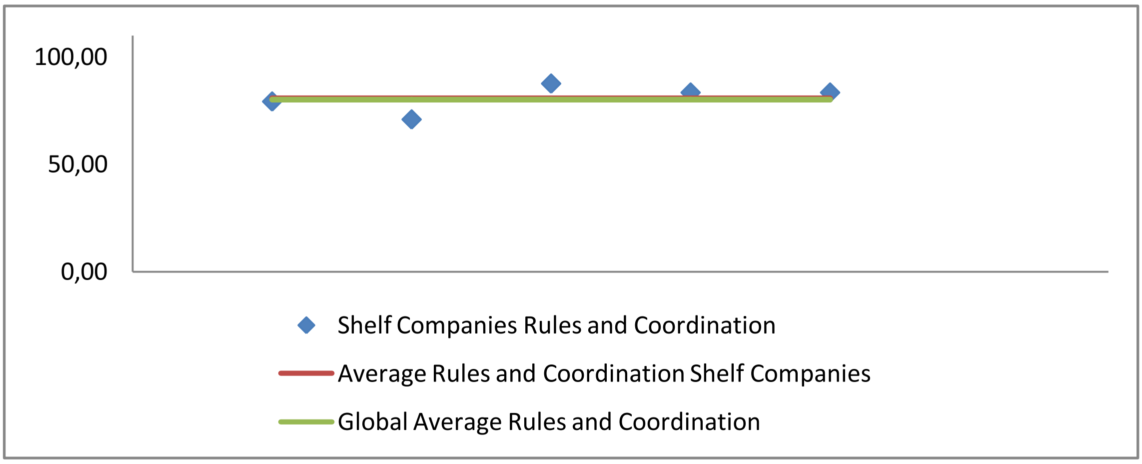

Figure 181. Rules and Coordination Area, whole sample with respect of the GA

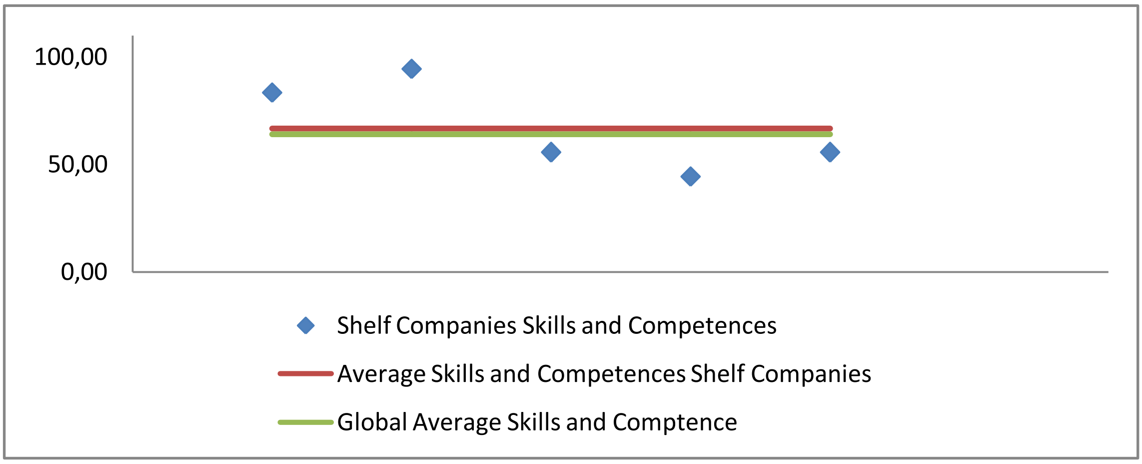

Figure 182. Skills and Competences Area, whole sample with respect of the GA

Summary of Shelf companies

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	79,17	71,11	66,11	74,89	44,21
Standard Deviation	18,63	13,26	9,50	15,67	18,52

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	70,12	81,43	80,83	66,67
Standard Deviation	12,61	13,62	6,32	21,15

Table 22. Summary of Standard Deviation for Shelf companies

Again the higher level of maturity is achieved in the Organization part, area of Work Organization; in which average scores overpass the 80% for MTO, MTS, Catalogue and Shelf companies. The dispersion of values in this area is around 10% and for that we cannot say anything relevant about significant behavior (Table 19, 20, 21 and 22). The other two areas of Work Organization present different values, but in the same tendency with the others classification that we have shown before. While Rules and Coordination has a good value, near 80%, the other one Skills and Competence presents low values. The dispersion of values Rules and Coordinated area is lower than the Skills and Competence. It is important see that behavior is similar with the global report.

If we study Process Management area we can see that there is not different between companies(Figure 144, 153, 162 and 171). In all case they attest around 80% in Process Management, even if some companies result to be less structured and still intermediate. Standard Deviation is higher in this area for all type of companies, around 15%, showing this dispersion between different Italian companies.

Moreover, the NPD process requires a large number of decisions to be taken every day considering both internal (Decision Making Factors) and external (Activities and Value) factors. Interviewed companies are actually paying less attention on customer value (near 75%) (Figure 145, 154, 163 and 172). In average, lower attention is paid on the whole product lifecycle, which not heavily affects companies' strategy and early decisions. Most of the companies seem to be not aware of the simple fact that the decisions taken at the early NPD stages affect the whole product lifecycle. The lower level of maturity corresponds to the area Decision Making Factors (Figure 146, 155, 164 and 173). In spite of, standard deviation is very high.

The third investigated part is Knowledge Management, which results to be still precarious. In fact, even if the level of Computerization is in average Mature, the Formalization is still in the middle between Low and Intermediate. Companies know that maintaining and protecting the internal know-how is crucial.

2.2.4.Average and Standard Deviation depending on the trend market

Leader market

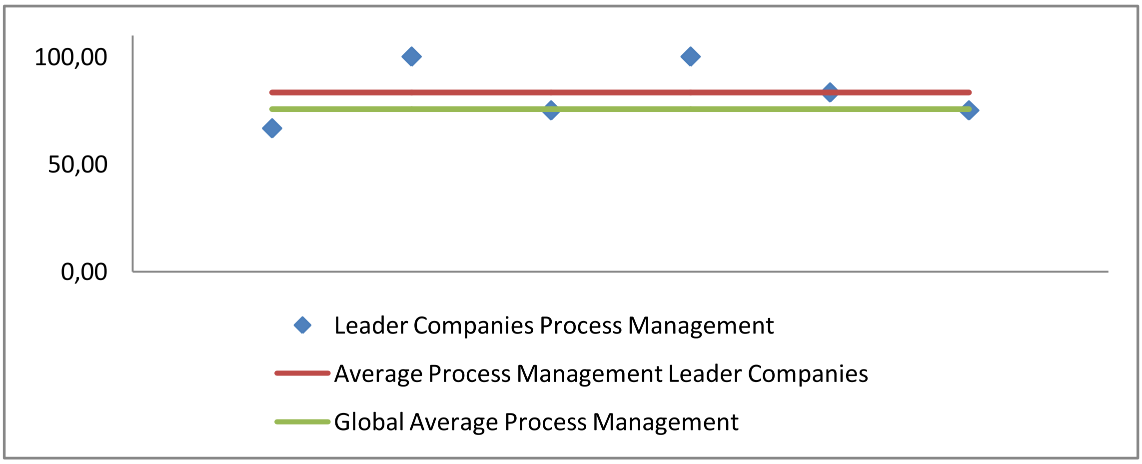

Figure 183. Process Management Area, whole sample with respect of the GA

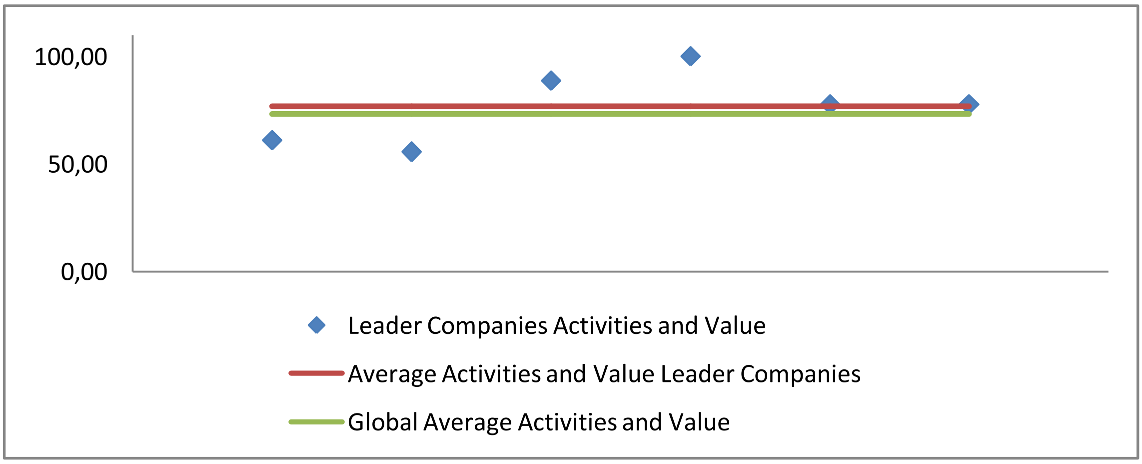

Figure 184. Activities and Value Area, whole sample with respect of the GA

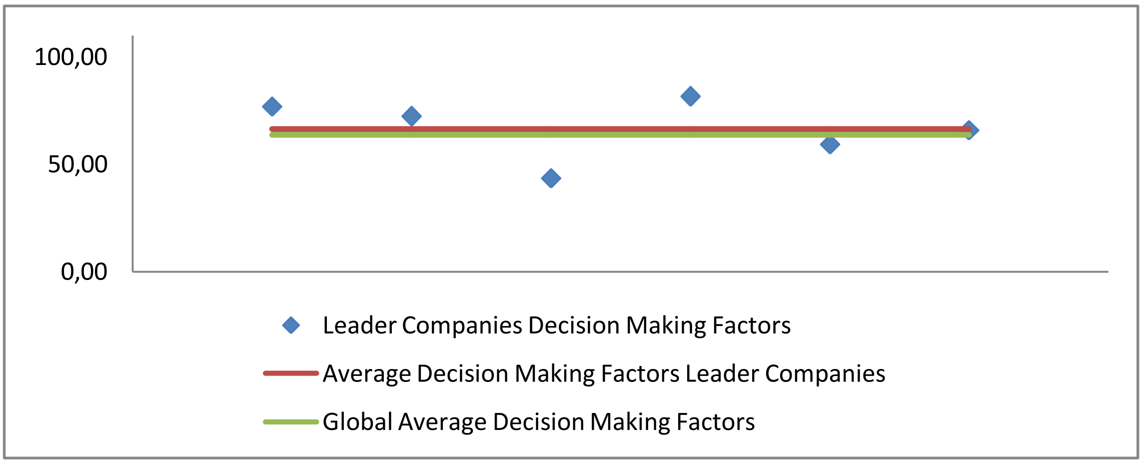

Figure 185. Decision Making Factors Area, whole sample with respect of the GA

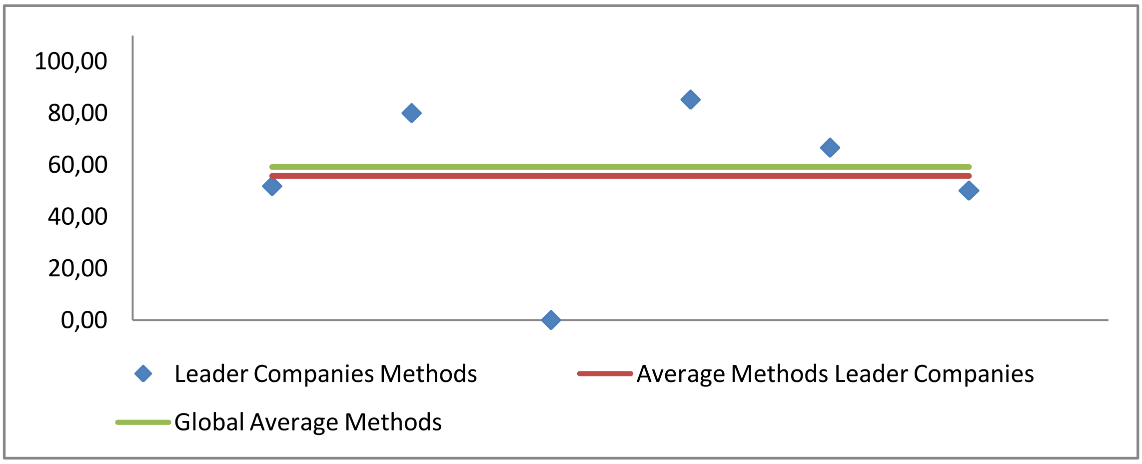

Figure 186. Methods Area, whole sample with respect of the GA

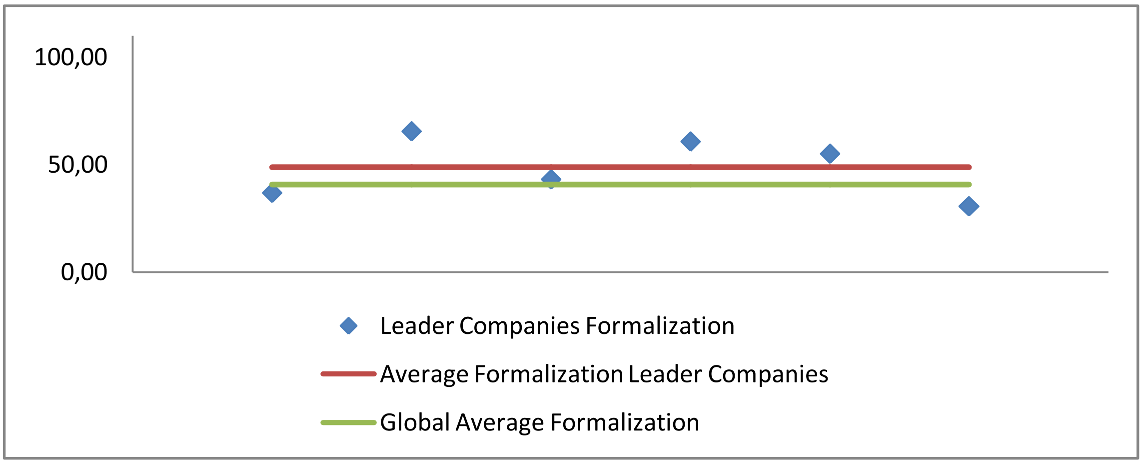

Figure 187. Formalization Area, whole sample with respect of the GA

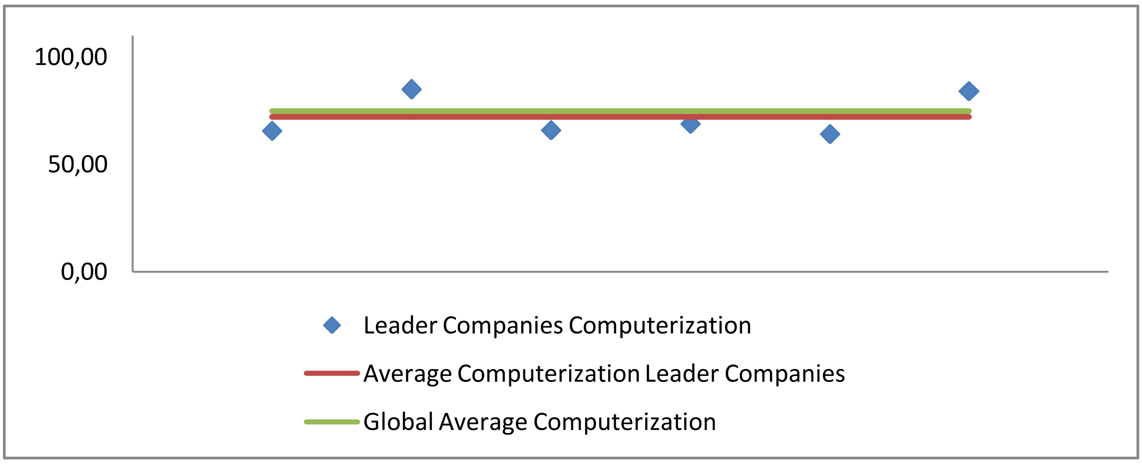

Figure 188. Computerization Area, whole sample with respect of the GA

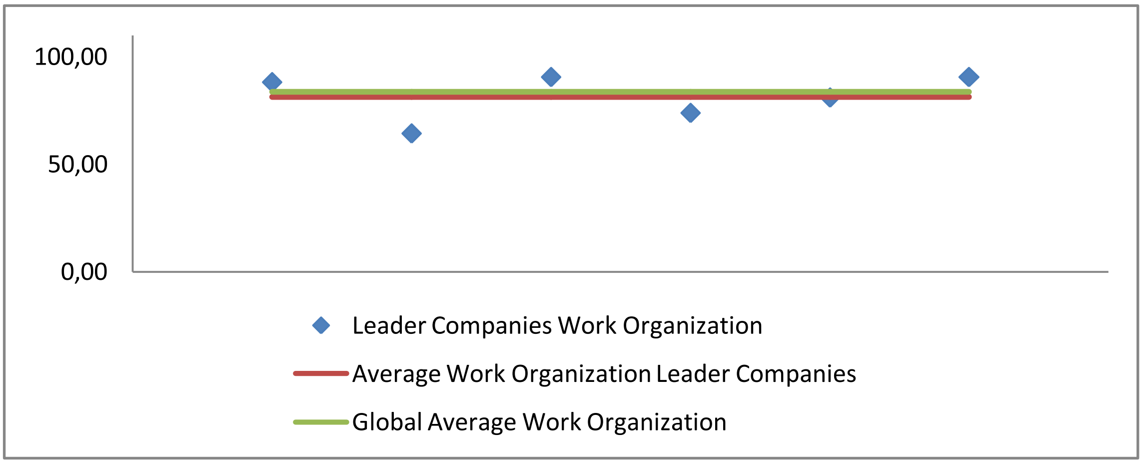

Figure 189. Work Organization Area, whole sample with respect of the GA

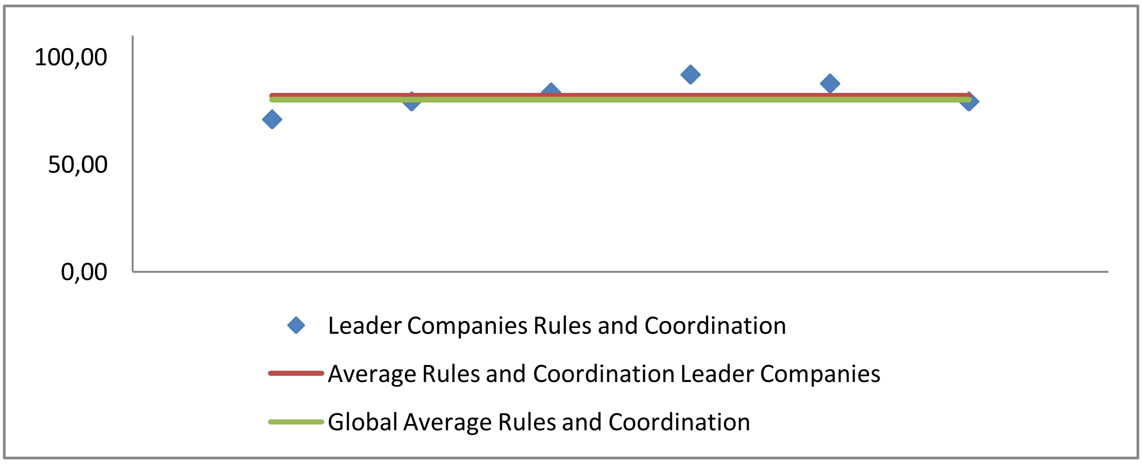

Figure 190. Rules and Coordination Area, whole sample with respect of the GA

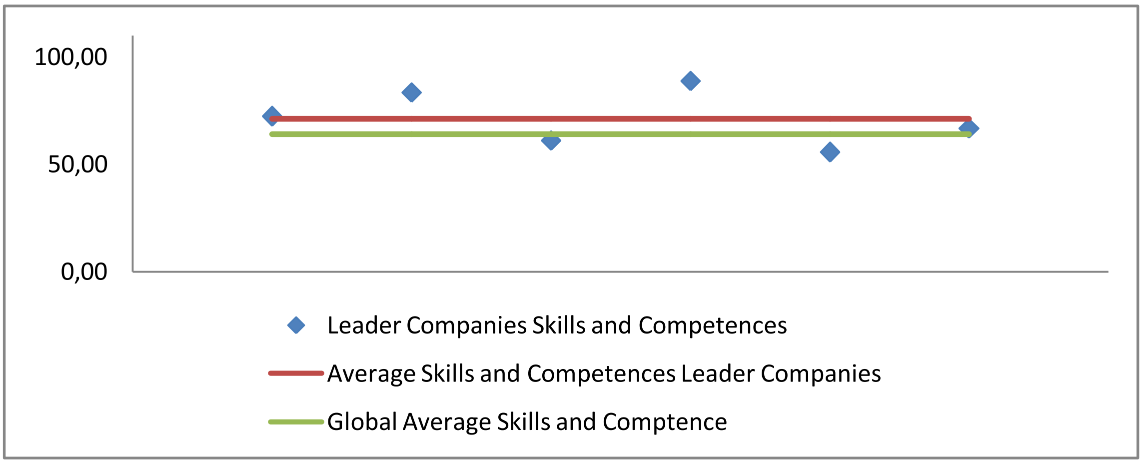

Figure 191. Skills and Competence Area, whole sample with respect of the GA

Summary of leader companies

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	83,33	76,85	66,51	55,62	48,68
Standard Deviation	13,94	16,64	13,75	30,76	13,91

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	72,18	81,35	81,94	71,30
Standard Deviation	9,61	10,59	7,30	12,87

Table 23. Summary of Standard Deviation for Leader companies

On trend companies

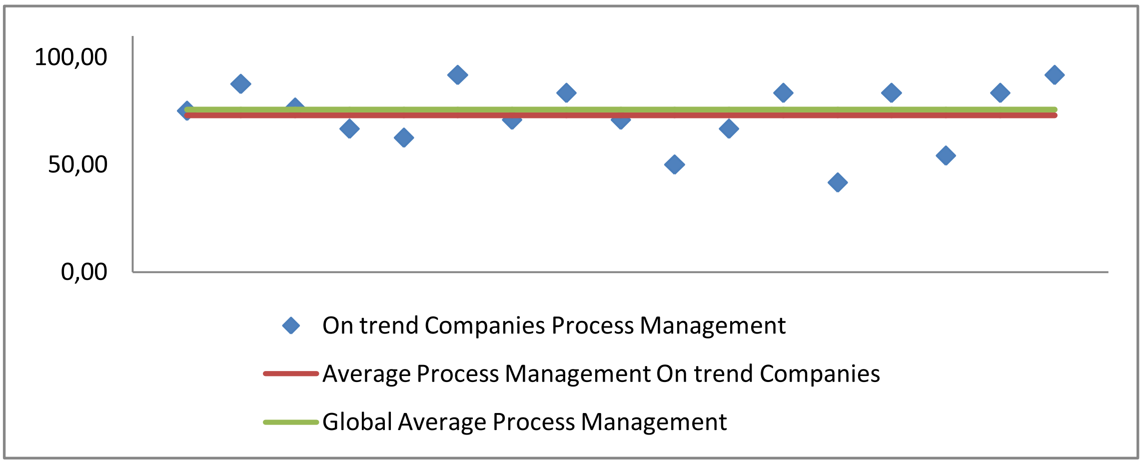

Figure 192. Process Management Area, whole sample with respect of the GA

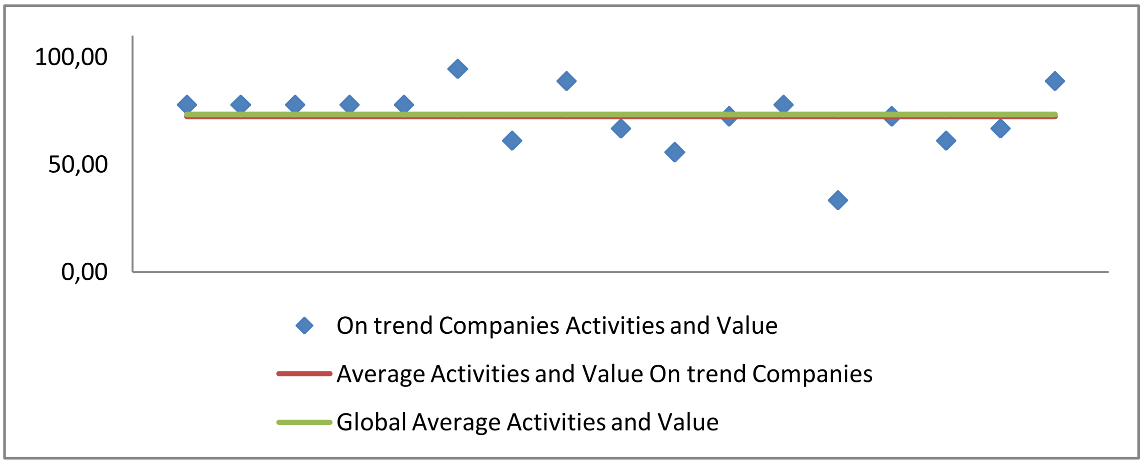

Figure 193. Activities and Value Area, whole sample with respect of the GA

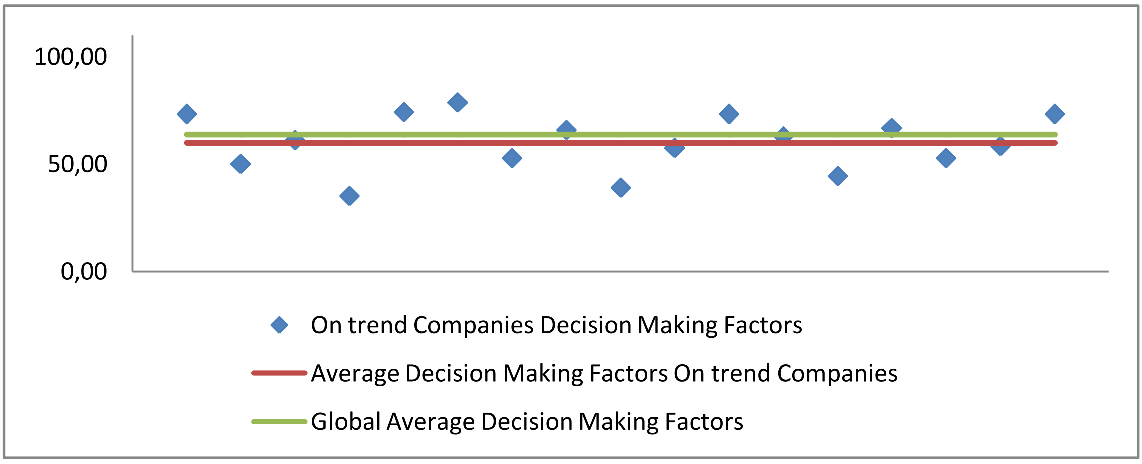

Figure 194. Decision Making Factors Area, whole sample with respect of the GA

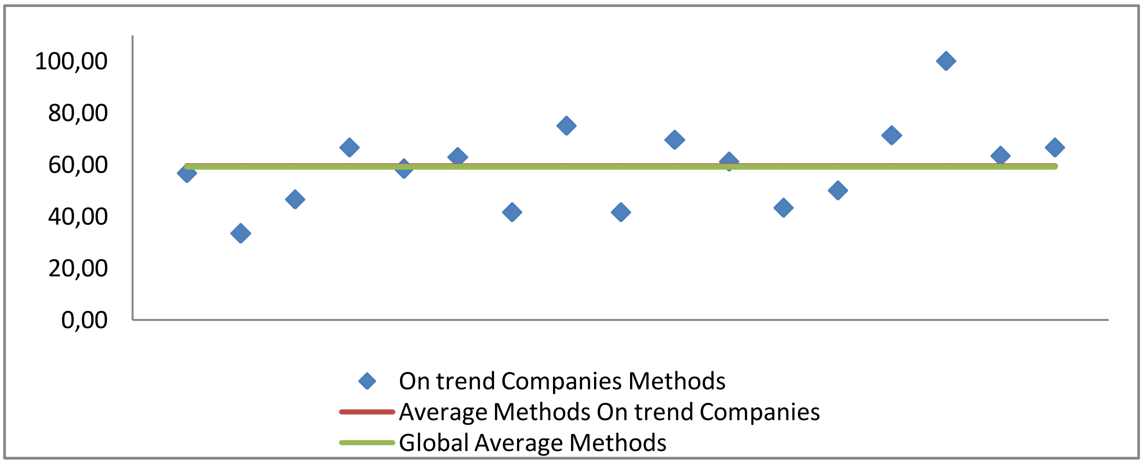

Figure 195. Methods Area, whole sample with respect of the GA

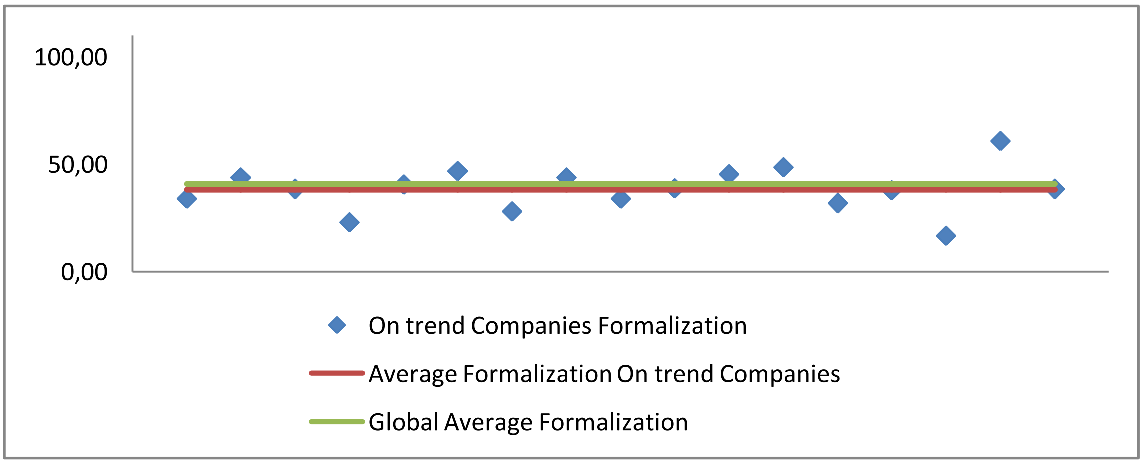

Figure 196. Formalization Area, whole sample with respect of the GA

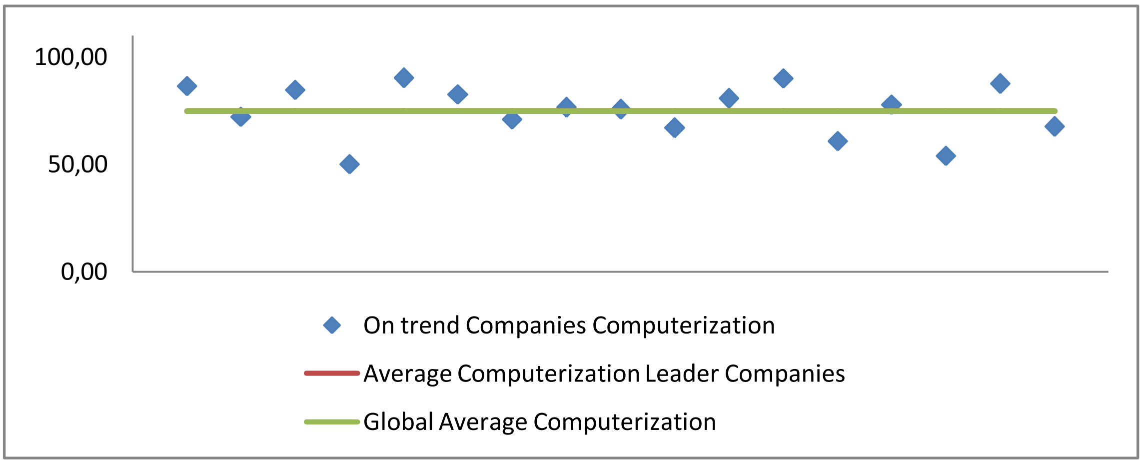

Figure 197. Computerization Area, whole sample with respect of the GA

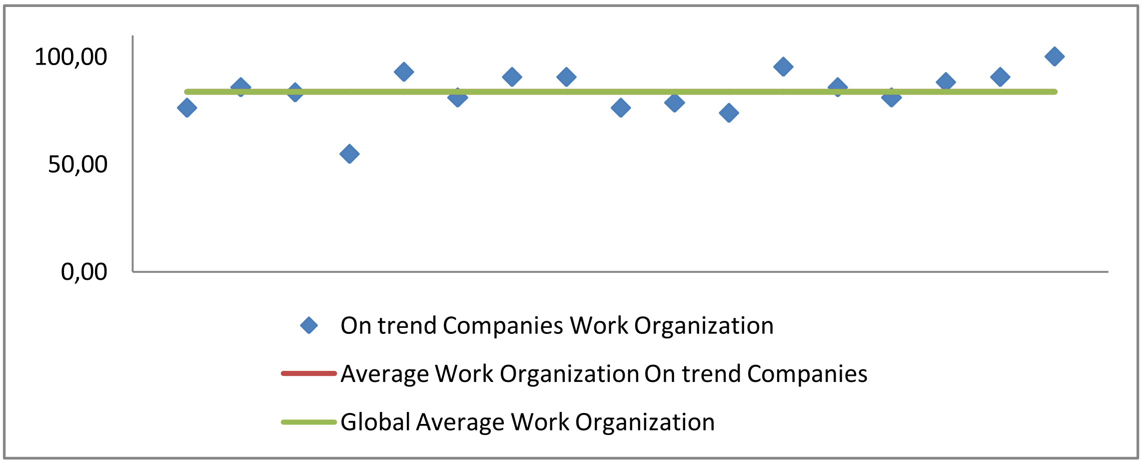

Figure 198. Work Organization Area, whole sample with respect of the GA

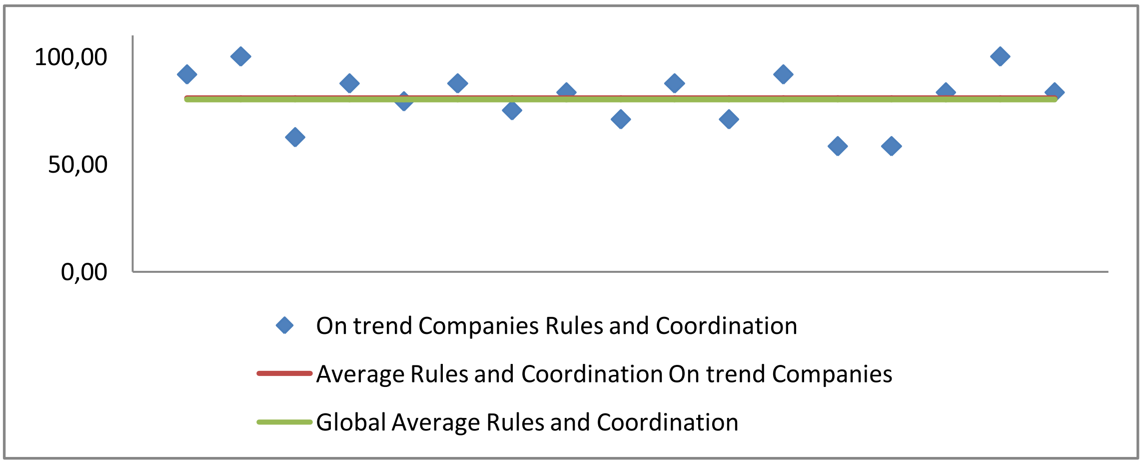

Figure 199. Rules and Coordination Area, whole sample with respect of the GA

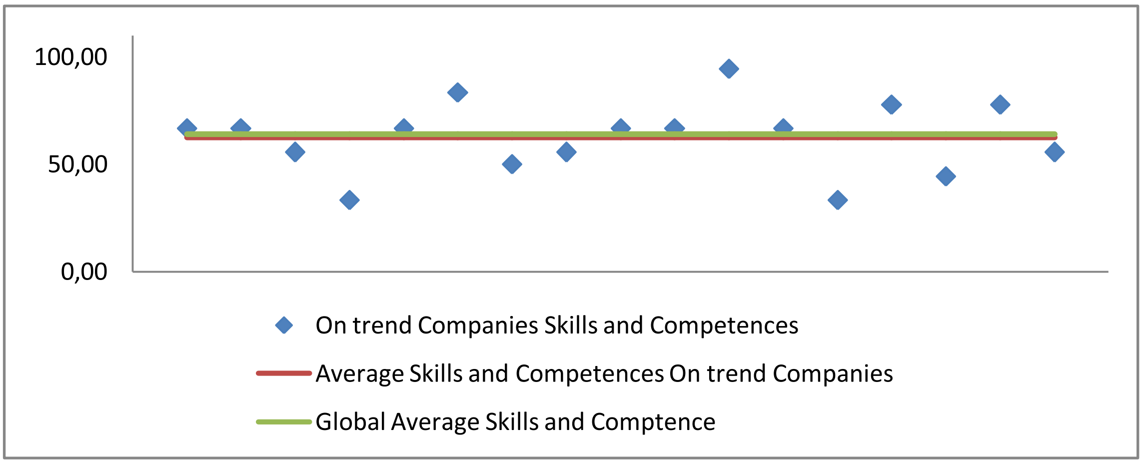

Figure 200. Skills and Competence Area, whole sample with respect of the GA

Summary of on trend companies

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	72,89	72,22	59,91	59,33	38,19
Standard Deviation	14,61	14,43	12,92	16,02	10,27

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	74,88	83,75	80,64	62,42
Standard Deviation	12,13	10,42	13,01	16,50

Table 24. Summary of Standard Deviation for On trend companies

2.2.2.2. Under trend companies

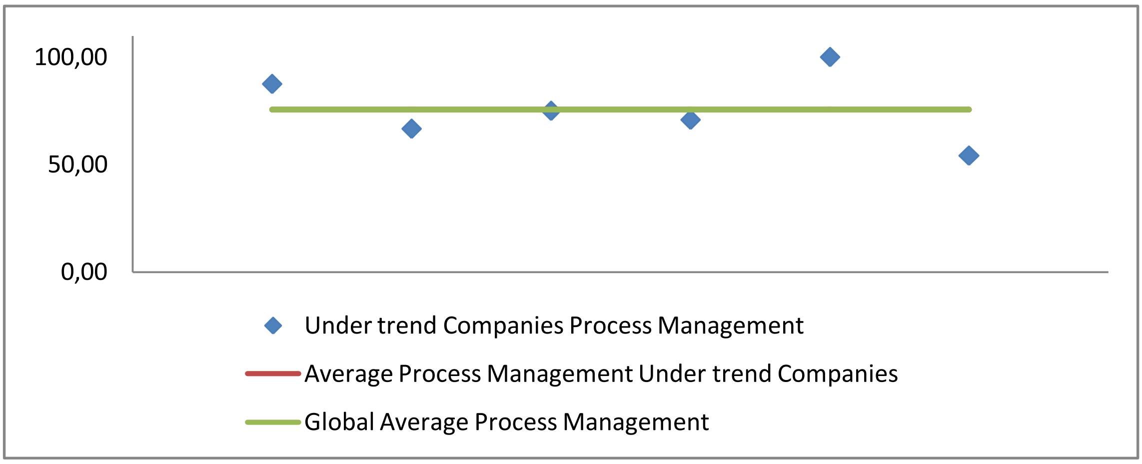

Figure 201. Process Management Area, whole sample with respect of the GA

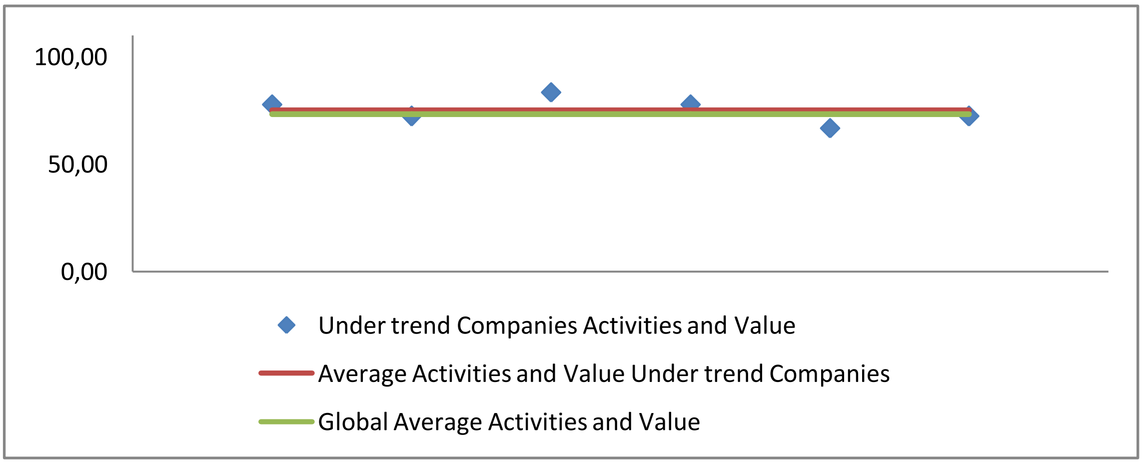

Figure 202. Activities and Value Area, whole sample with respect of the GA

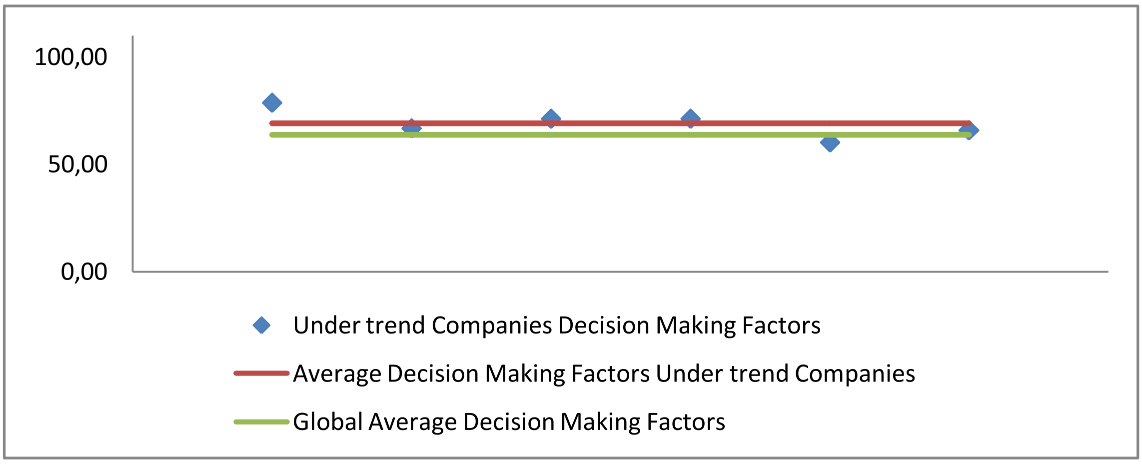

Figure 203. Decision Making Factors Area, whole sample with respect of the GA

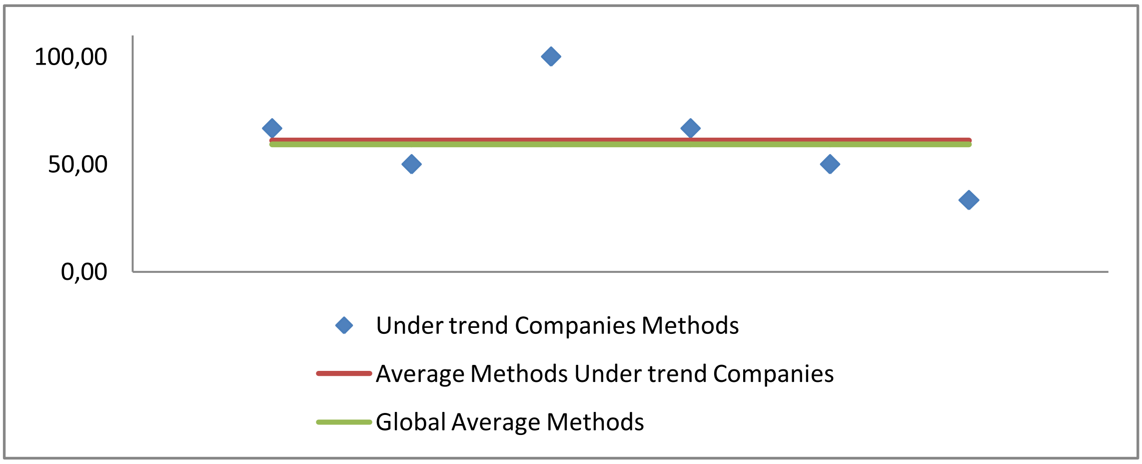

Figure 204. Methods Area, whole sample with respect of the GA

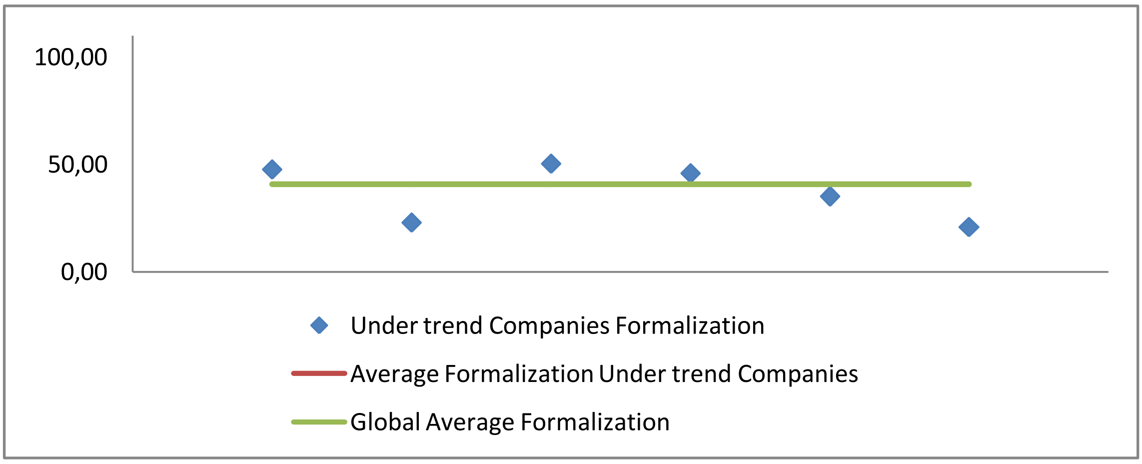

Figure 205. Formalization Area, whole sample with respect of the GA

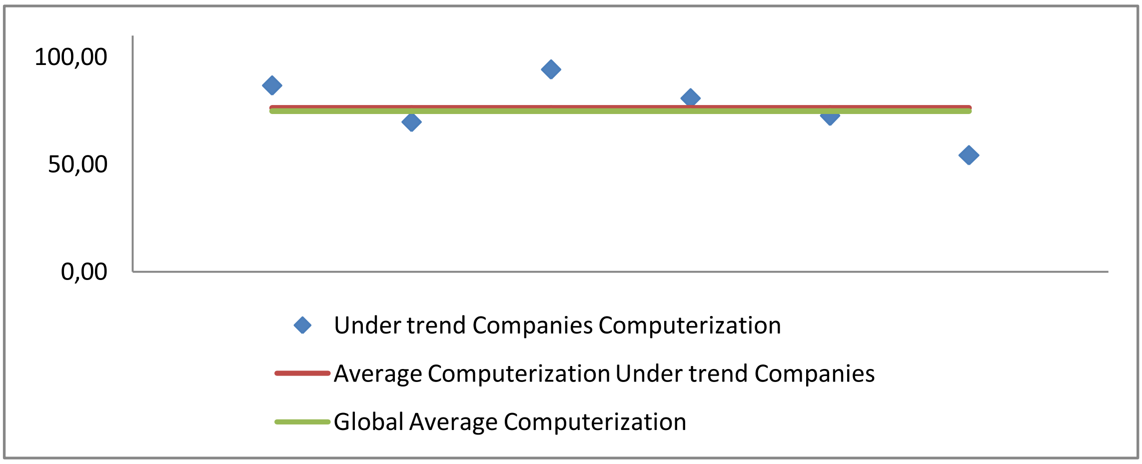

Figure 206. Computerization Area, whole sample with respect of the GA

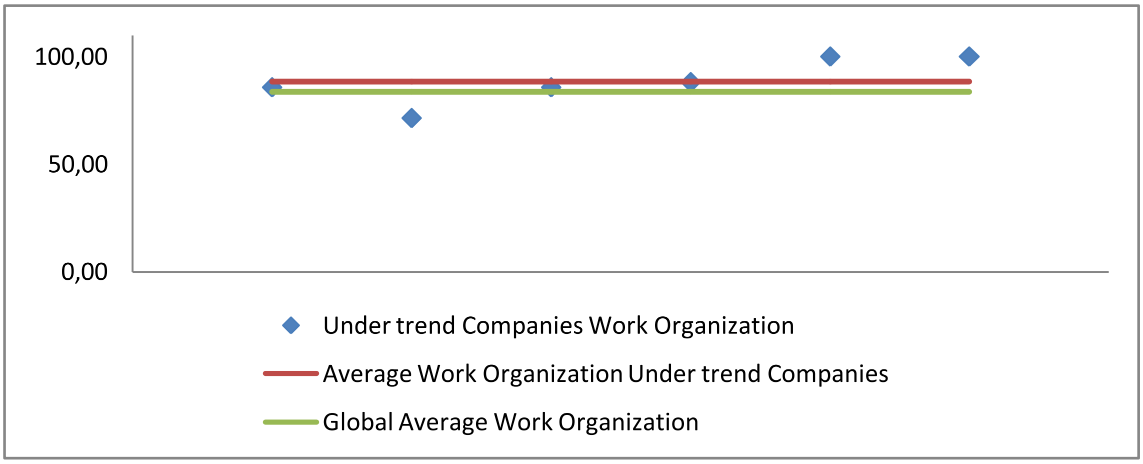

Figure 207. Work Organization Area, whole sample with respect of the GA

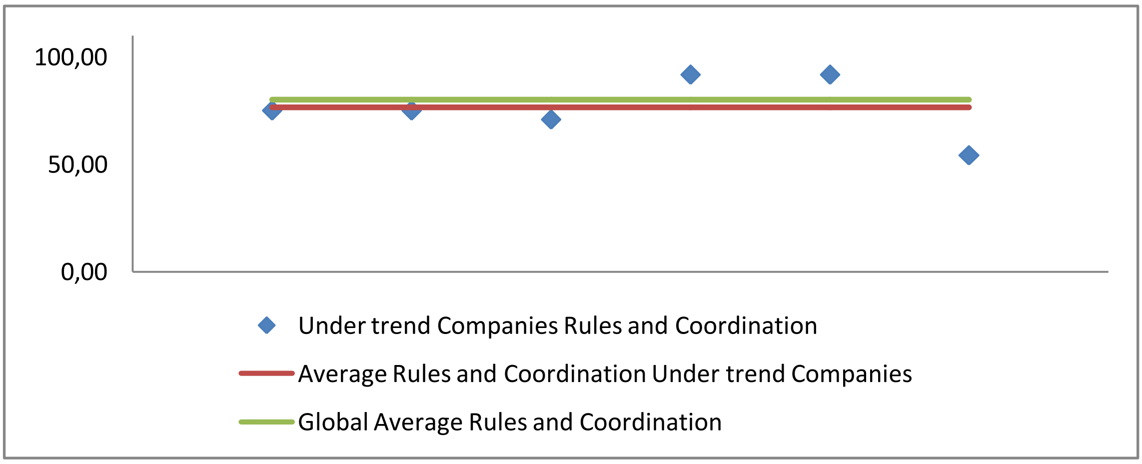

Figure 208. Rules and Coordination Area, whole sample with respect of the GA

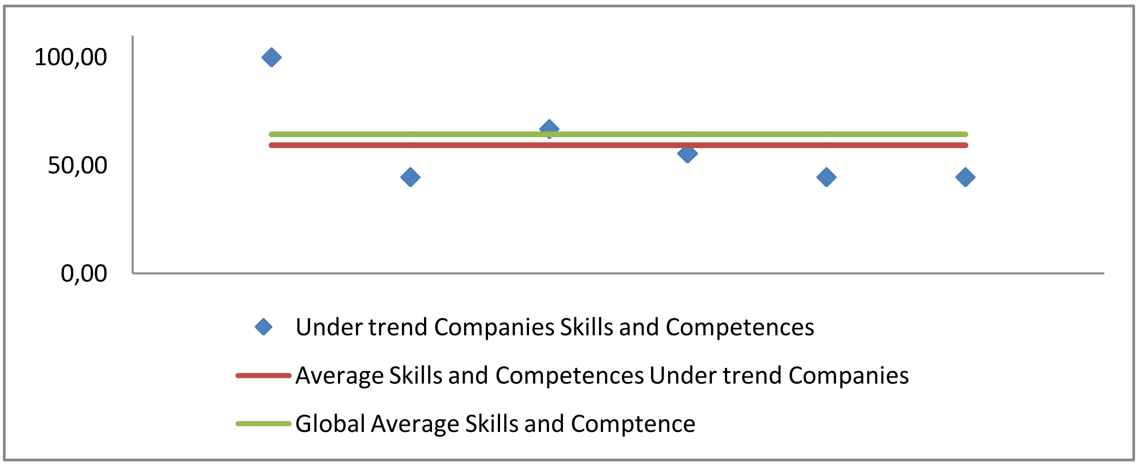

Figure 209. Skills and Competence Area, whole sample with respect of the GA

Summary of under trend companies

	Process Management	Activities and Value	Decision Making Factors	Methods	Formalization
Average	75,69	75,00	68,98	61,11	37,07
Standard Deviation	16,12	5,83	6,30	22,77	12,90

	Computerization	Work Organization	Rules and Coordination	Skills and Competence
Average	76,35	88,49	76,39	59,26
Standard Deviation	14,05	10,69	14,11	21,85

Table 25. Summary of Standard Deviation for On trend companies

The higher level of maturity is achieved in the Organization part, area of Work Organization, in which average scores overpass the 80% for all companies (Table 23,24,25). The other two areas of Work Organization present different values. While Rules and Coordination has a good value (Figure 187, 196 and 205), near 80%, the other one Skills and Competence presents low values (Figure 188, 197 and 206). The dispersion of values in this area is bigger than the other two areas, in fact many cases attest in Intermediate level of maturity. It is important see that behavior is similar with the global report.

If we study Process Management area we can see that there is a different between leader and on and under companies. In the second case they attest around 75% in Process Management (Figure 189 and 198), even if some companies result to be less structured and still Intermediate. But in the first case, for leader companies the value is best and we only obtain a result more than 850%. Standard Deviation is higher in this area for all type of companies, around 15%, showing this dispersion between different Italian companies.

Moreover, the NPD process requires a large number of decisions to be taken every day considering both internal (Decision Making Factors) and external (Activities and Value) factors. Interviewed companies are actually paying less attention on customer value (Figure 181, 190 and 199). In all the case this value is around 75%. In average, lower attention is paid on the whole product lifecycle, which not heavily affects companies' strategy and early decisions. Most of the companies seem to be not aware of the simple fact that the decisions taken at the early NPD stages affect the whole product lifecycle. The lower level of maturity corresponds to the area Decision Making Factors. But also we can see that the behavior is similar in the different type of companies that are under the market because SD is very similar.

The third investigated part is Knowledge Management, which results to be still precarious. In fact, even if the level of Computerization is in average Mature (Figure 185, 194 and 203), the Formalization is still in the middle between Low and Intermediate (Figure 184, 193 and 202). Companies know that maintaining and protecting the internal know-how is crucial.

At the end we can see that the behavior is similar for the different groups that we have created.

Chapter 3

Conclusion

3. The rating problem

As introduced in section 1, the aim of this work is to rate companies –defined as the objects to be rated- considering their performance in new product development process. Moreover we want to compare each companies with pre-defined levels of maturity, particularly 5 levels called CLIMB. Consequently we are in the class of ordered classification problem with predefined levels.

A series of steps will be followed in order to analyze the problem: definition of classes, choosing of indicators and weights, definition of the profile, and the rule of comparison. Then practical examples of real companies performance will be applied and companies will be rated. All these stages are described below.

	Not Predefined	Predefined
Ordered	RANKING	**RATING**
Not Ordered	CLUSTERING	ASSIGNMENT

Figure 210.Classification table

3.1. Steps for Rating Implementation

3.1.1. Definition of the Levels

In section 1 (Data Collection and Analysis) we introduced the maturity levels we are considering in the analysis and we defined them as follow in Table 1 and we represented them:

Level 1	**Chaos**	**0-19%**
Level 2	**L**ow	20-39%
Level 3	**I**ntermediate	40-59%
Level 4	**M**ature	60-79%
Level 5	**B**est	80-100%

Table 26. Maturity Levels

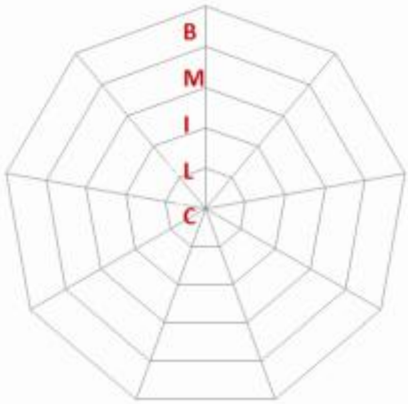

Figure 211. Maturity Levels Representation

3.1.2. Choose of Indicators

The indicators in our problem correspond to the 9 analyzed areas, evaluated as a % value.

A1. Work Organization

A2. Rules and Coordination

A3. Skills and Competencies

A4. Process Management

A5. Activities and Value

A6. Decision Making Factors

A7. Methods

A8. Formalization

A9. Computerization

As already explained, each area is composed of some questions (q_i), which can assume value from 1 –i.e. the company doesn't consider at all customer value during design process- to 3 –i.e. the company has clearly analyzed and defined the customer value and all the people involved in the development process are aware of this; the whole process is focused on continuous improvement and value creation.

The sum of the value of each question of the same area gives the total value of the area, which has to be normalized and transformed in %, in respect to the maximum possible value to be obtained (if all the answers assume value 3 for all the questions in the area). Formalizing, the value for generic area A_i, called a_i, can be calculated as follow.

$$a_i = \frac{\sum_1^{m_i} q_j}{3 * m_i}$$

Where:

a_i is the % value corresponding to area A_i

i=1…9, is the indicator for the areas

q_j is the value of the answer for area i and question j, which can assume value from 1 to 3

j=1…m_i, is the indicator for the questions, depending on the area the number of questions change

m_i, is the number of questions of the area A_i

3.1.3. Choose of the Weights

In our we didn't give different weights to indicators. Even if I thought about considering each area in different ways, basing for example on the context, the company strategy, or the sector, I didn't formalize this idea until now. I only had this trivial thought: maybe a company x is good in some areas and bad in other areas, because this is his strategy, maybe company x cares much more about formalizing and computerizing knowledge rather than improve skills and competencies. The problem is that as the model is now, we don't have a global evaluation for the companies, but we only have punctual analysis in each of the 9 areas. We have to delineate how to calculate the maturity level of the whole company, defining how different areas impact the whole evaluation. In this way we'll be able to assign a unique maturity level in the CLIMB scale to each company and rate them according.

If we consider that all the areas have the same importance, the company x will result to have lower performances rather than if we consider each area to have its own particular importance, which is maybe for company x major importance to knowledge management area and less importance to human resources management.

3.1.4. Definition of Profiles

The profiles which separate levels are predefined boarder lines. They are used as matter of comparison with the performance of the single company, which can be lower or upper than the profile itself.

3.1.5. Definition of Rules of Comparison

The comparison rules which can be used is the OUTRANKING. We can define a Treshold α at 85%. In fact, the sum of the weights in favor of the company should be > or at least equal to α. It means that knowledge management is crucial for the considered sector and is a necessary condition to satisfy in order to get the best practice level. Finally we don't have veto condition here.

We can now start the evaluation of the reason in favor of the sample. If we calculate maturity for each company we find the next graphic in function of the level of sales:

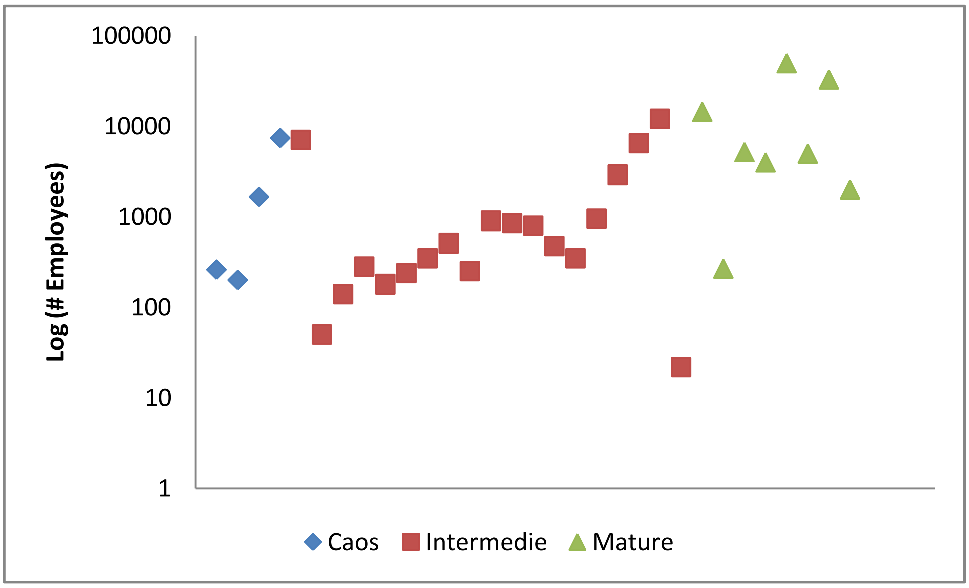

Figure 212. Sample distribution Log employees (y) Maturity (x)

Or this other one in function of the number of employees:

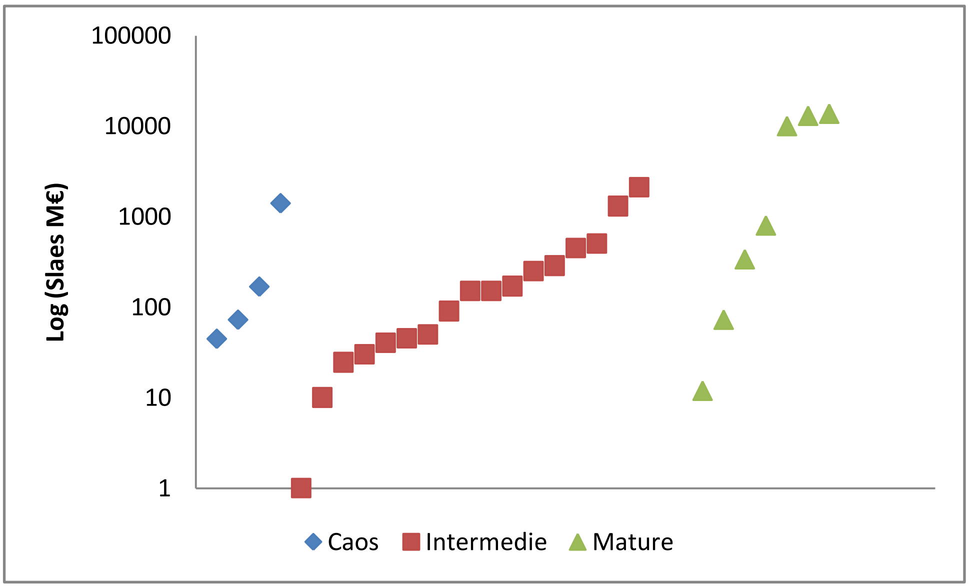

Figure 213. Sample distribution Log sales (y) Maturity (x)

Where we can see that the level of maturity is not correlated with the level of sales but there is a little the correlation with the number of employees.

Summarizing I adopted rating approach to rate companies basing on their measured performance in new product development process. The method results effective for company rating, benchmarking and possible identification of best practices. The rating framework was made of 5 maturity levels, 4 (levels-1) profiles, 9 criteria, 9 weights, and no veto. Future steps for validating the analysis could be done in terms of performing sensitivity analysis, trying to vary values for trash old to test the robustness of the rating, and perform rank reversal analysis on weights evaluation.

The weakest point found in this work is the evaluation of the weights. In fact this aspect has highlighted a very important gap in my research to be filled. The importance of having different significance of the criteria, or in this case, areas, is crucial. In fact, basing on the context different areas can assume different significance. A more accurate comparison is obtained when companies belonging to the same area are compared. Further and deeply investigation will surely be conducted about this relevant issue; this work represents an important and solid starting point of my long PhD path.

APPENDIX A

Questionnaire for the assessment for new product development process:

1 Part I - Introduction to the company and product offerings

1.1 Company's data

- Company name
- Number of employees, Average turnover
- Corporate group, Head Office
- Role in the corporate group, Role of the other offices
- Ownership, Management (family business, managerial business)

1.2 Interviewer's data

- Name, Surname, Telephone, E-mail
- Membership function, Role
- Number of years in the current role, Number of years in the company

1.3 Product and market position

- Which are the markets served? (national, international)
- How do you respond to market demand? (make to order, make to stock, catalogue, etc.)
- What is the trend in the current scenario? (growing, stable, declining)
- Which are the critical success factors for the company? (cost, quality, service, innovation, performance, timeliness, etc.)
- Which are your main competitors? What is your position compared to them?
- Which is your position in the market? (Leader, Follower, Niche)

1.4 Type of products

- Which is your core business? (Principal products produced and product evolution portfolio in the last years)
- Which are the technologies required to produce these products? (Mechanical, electronically, etc.)
- Which are the core components of your product?
- How many components/parts are there per product? (tens, hundreds, thousands, many thousands)

1.5 Value chain and company organization

- Which is your position in the Value Chain? (prime contractor / OEM, component supplier, raw material supplier, services supplier, etc.)
- How your company is organized? (Functional, matrix, divisional)
- What is the position of the technical direction and IT function in your organization chart?
- How many people are involved in the technical direction/design function? What are their main skills?
- How many people are involved in the IT function? What are their main skills?

1.6 How would you define the design/development process of your company?

1. Company designs one-of-a-kind solutions, receiving detailed requests from the customer (highly customized process).
2. Company designs dedicated solutions, attending to competitive tenders and making bids able to satisfy the main customer requests (bidding process).
3. Company designs different product solutions as catalogues, receiving requests and specifications from the Internal Marketing department, which makes analysis of the market and customer behaviors (market pull process).
4. Company designs different product solutions, pushed by the R&D and technical departments, to be sold in the market (technology push process).
5. Other (specify).

1.7 Is the design process geographically distributed? If it is, how?

1. No, the whole process is done in the same place. Partnership with design suppliers is possible, if necessary, according to own competences and suitability.
2. The design process is distributed in different national places, divided by type of product, technology skill, etc.
3. The design process is totally distributed, involving functions and company of different countries.
4. Other (specify).

2 Part II – DESIGN PROCESS AND KNOWLEDGE MANAGEMENT

2.1 Work Organization

2.1.1 Does the company have a formal and standard model for product development and innovation process?

a. Nowadays the company doesn't have a formal model. The process is based on the flexibility of the resources involved.
b. There is a formal model (ex. requirement for certification), but it is rarely followed.
c. There is a formal model, properly followed and documented by the various actors involved.
d. Other (specify).

2.1.2 Any development task consists of two key elements: (i) routine tasks and (ii) innovative tasks. The routine tasks are standard and done for all products/project, as most of the products/projects are not developed from scratch rather and they are successive from previous designs. Innovative tasks distinguish the new product from previous ones and have not been considered before. Estimate in percentage how much of your work is related to routine or innovative tasks.

100%Inno.	80%Inno.	60%Inno.	50%Inno.	40%Inno.	20%Inno.	0%Inno.
0%Rout.	20%Rout.	40%Rout.	50%Rout	60%Rout.	80%Rout.	100%Rout.

2.1.3 Considering a typical day of a designer/engineer, how do you assess the following tasks?

Attività	*5%*	*10%*	*20%*	*30%*	*40%*	*50%*
Time spent for value adding activities (design, test, development, etc.)						
Time spent for retrieving information from traditional resources (paper-based, chats, etc.)						
Time spent for retrieving information from digital sources (Values, Files, etc.)						
Time spent for elaborating specification and doing documents						
Time spent for coordination and collaboration with colleagues						
Other						

2.1.4 Which is the reference model for the design/development process of your company?

a. The typical development process is a sequential process: different functions work sequentially, exchanging specifications and review requests.
b. Tasks and activities are done in parallel (concurrent engineering) by a multidisciplinary and multifunctional team. People involved in quality, production and service function collaborate for the first stages of the development process.
c. Development process is strongly collaborative, involving technical functions and external resources (suppliers) from the first stages.
d. Others (specify).

2.1.5 If the company follows a concurrent engineering model, which of the following actors are involved in the project teams (core team and extended team) of design and where are they eventually localized?

Actor involved	*Involvement in team project*			*Localization*	
	No	*Core team*	*Extended team*	*Co-located*	*Scattered*
Concept designers					
Product engineer (ex. Mechanical, electronic, etc.)					
Production engineer					
Senior engineer/Technical leader					
Project manager (person responsible for project team).					
Product manager (responsible for a product family)					
Production function/logistic					
Maintenance function / Technical assistance					
Marketing function/ commercial					
Quality function					
Sustainability expert					
Analyst and cost control					
Client(beta-tester)					
Other (specify)					

2.2 Roles and coordination

2.2.1 How are the roles and responsibilities distributed throughout the design and development process?

a. There is no definition of roles and responsibilities for the different actors.
b. Roles and responsibilities throughout development process have not been clearly defined for different actors involved, but are defined for a project team.
c. There are clear rules and responsibilities throughout the development process for each individual technician and designer.
d. Other (specify).

2.2.2 How can designers and technicians execute their jobs? What is the level of flexibility and responsibility assigned to them in the development process?

a. Designers must complete defined tasks in order of predefined procedures.
b. Designers must complete defined tasks, but the order is flexible and they are free to administer their own workload.
c. Designers have freedom to experiment and they are responsible to manage their own time and tasks. Moreover, there is no mandate to prepare documentations of activities.
d. Other (specify).

2.2.3 Who is responsible for managing the overall innovation and development projects?

a. There no one responsible for the overall projects. The framework is strictly functional.
b. There is a Project Manager (no technical background) who controls development tasks (time, costs), while technical decisions are executed by the head/s of engineers and designers.
c. There is a Project Manager (technical background) who has full responsibilities (time, costs, quality and performance) for the whole development projects.
d. Other (specify).

2.2.4 How does the allocation of human resources work in the development process? Is there any formal and structural process to follow?

a. There is no formal process of allocating human resourcesin development projects, but designers are assigned as they become available during the various stages of design.
b. Early stages are conducted by ‘juniors’, but experienced experts are increasingly involved and fully committed to most critical design challenges.
c. Experienced designers mostly be involved from the earliest stages of a project, with the focus to reduce the overall uncertainty and mitigate potential risks of project failures.
d. Other (specify).

2.3 Training and competence

2.3.1 Product design is heavily based on skills and competence of the actors involved (technicians, designers, managers, etc.). On average, how does the company support training and skill development?

a. Any engineer/designer is personally responsible for developing and maintaining his/ her skills.
b. The company urges the development of strong technical skills, and give training on the job.
c. The company promotes multidisciplinary skills and supports knowledge management activities with formal programs (ex. training plans, rotation between project teams, etc,…).
d. Other (specify).

2.3.2 Inside the organization, is there a responsible trainer that supports training activities?

a. No, each technician/designer is expected to build his/her skills individually.
b. Yes, a technician/designer is encouraged to develop his/her own skill from his/her direct supervisor.
c. Yes, there is a one-to-one correspondence for tutoring (a junior designer is assigned with a more experienced designer, as a tutor, coach, or mentor).
d. Yes, other (specify).

2.3.3 How does the effectiveness of a training be evaluated in terms of the learning outcomes?

a. Using ‘visual’ evaluation where is verified improvements of individual behavior.
b. Using a test before and after the training session.
c. Using KPIs to assess the impact of training on business performances.
d. Other (specify).

2.4 Process Management

2.4.1 How do you manage the development process? i.e. How does the design process "flow"?

a. Projects are initiated solely based on customer requests.
b. Projects start at a regular interval, according to available plan, which considers the different types of projects (e.g. changes, radical redesign, innovation, etc.). The activities, however, often require reviews and delays are common.
c. The work plan between the different projects is regular on average, for times and types of projects. Delay is minimized.
d. Other (specify).

2.4.2 How do you measure the performances of the development process?

a. Performances are not defined and are not measured using any KPI (Key Performance Indicator) specific for the development process.
b. Time and cost indicators are used to evaluate the performance of the development process.
c. A complex set of indicators is used for measuring the performance of design/development processes (time, cost, quality, productivity etc.)
d. Other (specify).

2.4.3. H*ow open and flexible is your process?*

a. Each project follows a structured and a controlled plan. Specifications are strictly respected and changes in the advanced stages are not accepted.
b. Projects are subjected to continuous reviews in order to meet the requests of different actors (marketing, customers, suppliers, etc.). Changes often have significant impacts on development time and cost.
c. Considerable efforts are made at the early stages of the process for experimentation and analysis. Further changes are not accepted after a fixed time.
d. Other (specify).

2.4.4. How and in what way the company does improvement activities in development process?

a. There is no specific plan or strategy to follow. Improvement initiatives are activated according to emerging need and available resources in ad-hoc basis.
b. There is a strategic plan to improve the development process that we use to make the operational decisions.
c. The company follows a continues improvement initiative, and invest on important resources and new methods, tools, procedures, etc.
d. Other (specify).

2.5. Assets and values

2.5.3. How do you select between different designs alternatives? How does an optimal solution is identified?

a. A single solution is designed from the beginning to the end of the project
b. At the beginning different alternatives are identified but quickly the solution that best matches the design specifications at the lowest cost is selected and detailed.
c. Many solutions are designed and inferior solutions are progressively discarded when new information becomes available (e.g. thanks to prototyping, tests, etc.).
d. Other (specify).

2.5.4. A design process should target to create 'value' for the customer. Value is a product attribute the customers are willing to pay for it. How is this value defined and who is responsible for the definition?

a. There is no a formal way to define value.
b. The company has its own analytical method to define customer value (ex. marketing) but it is not clearly communicated (ex. through formulation of explicit requirements) to engineers and designers.
c. The development process is completely focused on offering value to customers. Main actors are involved in the definition of customer value and know the factors needed to achieve it
d. Other (specify).

2.5.5. Through the development process, to what extent the customers are taken into account, in order to satisfy their needs?

a. They are not taken into account. Customers can only purchase products in the catalogue.
b. Customers can customize their products choosing the standard modules to be assembled.
c. Customers are involved in the development process and the product is designed based on customer inputs.
d. Other (specify).

2.5.6. Nowadays in the industrial context there is a great competition in almost all sectors. Competitors are often very quick to create new solutions. What is the attitude your company maintains against competition in the innovation process?

a. The company does not consider competitive products because it works on its own ideas to satisfy its own customers.
b. In the early stages of new projects the company usually investigates competitors' products. Generally it is a task of marketing function with the target to have indications of market preferences.
c. There is an established process of analyzing competitors, where not only marketing function but also engineers/designers are heavily involved.
d. Other (specify).

2.6. Decision factors

2.6.3. Decisions that are made in the early stages of a design process (conceptual stages)affect throughout a product's lifecycle. From your experience, how would you rate the impact for the following lifecycle phases?

Lifecycle phases	*Low*	*Moderage*	*High*	*Very high*
Design and industrialization of parts and modules				
Components and semi-finished products manufacturing				
Components assemby				
Test and experimentation				
Packaging and warehousing				
Delivery and distribution				
Usage by the customer				
Maintenance and after-sale services				
Disassembly, recycling and disposal				
Check, reuse, updating, revamping				
Other (specify)				

2.6.4. Which are the main factors taken into account for design/development decisions? How would you rate the company's consideration of the following factors?

Factors	*Low*	*Moderage*	*High*	*Very high*
Final cost of the product/price of the product				
Development costs / ROI (Return On Investment)				
Costs related to usage and life cycle costs (e.g. maintenance, services, recycling, disposal etc.)				
Time To Market (TTM)/Complying with the date of delivery				
Number of different products				
Level of customization				
Level of innovation				
Brand image				
Functional performances (e.g. productivity, speed, accuracy, Usability, safety, maintainability, etc.)				

Factors				
	Low	*Moderage*	*High*	*Very high*
perceptible and not by the customer				
Quality performances (e.g. robustness, reliability, availability, durability, etc.) perceptible by the customer				
Performances compliance with laws and rules, provided to customers				
Other performances perceptible by customers (e.g. environmental sustainability, esthetics, social and moral aspects, etc.)				
Other (specify)				

2.6.5. How would you rate the importance of each stage of your development process, to retrieve knowledge form previous projects/product?

Design phase				
	Low	*Moderage*	*High*	*Very High*
Concept design				
Feasibility study				
Detailed design of product and its components				
Testing, prototyping, experimentation				
Final validation of the project				
Other (specify)				

2.6.6. Which of the following aspects are fundamental for the success of your company? (max 3)

Competitive area	
Value-added for the customer	
Segmentation	
Distribution channel	
Services	
Key resources (manufacturing system, other physical resources, human resource, financial resources)	
Key activities (production, problem solving, etc.)	
Strategic suppliers	

2.7. Methods

2.7.3. Which of the following formal methods for supporting design process are used in your company? How frequent are they used?

Design methods	*Frequency*			
	Never/Not known	*Sometimes*	*Often*	*Always*
Rules for parts modularization and standardization (modular, platform, cluster design, etc.)				
Design for X (DFX) for functional performances (e.g. Design for Manufacturing, for Assembly, Robust Design, etc.)				
Design for X (DFX) for quality performances (e.g. Design for Six Sigma, for Maintenance, etc.)				
Design for X (DFX) for other performances perceptible by the customer (design for Aesthetics, for Environment, Eco-Design, etc.)				
Design To Cost (DTC)/Target Cost Management (TCM)				
Lifecycle Cost (LCC)/Total Cost of Ownership (TCO)				
Lifecycle Analysis and Engineering (LCA&E)				
Value Analysis and Engineering (VA&E)				
QualityFunction Deployment (QFD)				
Risk and Fault Analysis, Failure Modes Effective Analysis (FMEA / FMECA)				
Methods for systematic innovation (e.g. TRIZ)				
Other (specify)				

2.8. Formalization

2.8.3. Do you have an existing or planned method to manage knowledge in design /development process?

a. The company does not have any formalexisting method nor has any plan in the future.
b. The company has introduced different initiatives to capture design knowledge, but there is no incentive that encourage employees.
c. There is a formal process to encourage and reward the sharing of knowledge. People can document and search information efficiently and they are very motivated to do it.
d. Other (specify).

2.8.4. What source of knowledge do you use to ensure the following factors/criteria are considered in your design/development process?

Factors	*Sources of Knowledge*						
	Written design rules defined by the company	*Written design rules in text books/standards*	*Written rules defined by external parts (e.g. customers, suppliers, etc.)*	*Previous projects*	*Personal experience*	*Personal intuition and ispiration*	*Collaborative work with colleagues*
Final cost of the product/Price of the product							
Development costs / ROI (Return on investment)							
Costs related to usage and life cycle costs (e.g. maintenance, services, recycling, disposal etc.)							
Short Time To Market (TTM)/Complying with the date of delivery							
Number of different products							
Level of customization							
Level of innovation							
Brand image							
Functional performances (e.g. productivity, speed, accuracy, usability, safety,							

Factors	*Sources of Knowledge*						
	Written design rules defined by the company	*Written design rules in text books/standards*	*Written rules defined by external parts (e.g. customers, suppliers, etc.)*	*Previous projects*	*Personal experience*	*Personal intuition and ispiration*	*Collaborative work with colleagues*
maintainability, etc.) perceptible and not by the customer							
Quality performances (e.g. robustness, reliability, availability, durability, etc.) perceptible by the customer							
Performances compliance with laws and rules, provided to customers							
Other performances perceptible by customers (e.g. environmental sustainability, aesthetics, social and moral aspects, etc.)							
Other (specify)							

2.8.5. How often are the 'sources' of knowledge updated and reviewed in the company?

Sources of Knowledge	*Updating*			
	Never	*Sometimes*	*Often*	*Always*
Written design rules defined by the company				
Written design rules text books/standards				
Written rules defined by external parts (e.g. customers, suppliers, etc.)				

2.8.6. How much, in percentage and on average, do you rely on knowledge from previous project when designing a new “product”?

100%	80%	70%	60%	50%	40%	20%	0%

2.8.7. Currently, which of the following techniques or/and tools are formally used to capture, share and reuse design knowledge in the company? How often are they used? Are they effective?

Initiative/ mode	*Use*				*Effectiveness*			
	Never	*Someti mes*	*Often*	*Always*	*Lo w*	*Moderat e*	*Hig h*	*Very High*
Verbal communication with colleagues								
Lessons learned documents								
Specification documents of the projects								
Questionnaire / Checklist								
Obeya rooms, poster and visual management								
Network shared folders								
Intranet								
Wiki								
Blogs, forum, noticeboards								
PDM/PLM Systems								
KBE software and design automation								
Other (specify)								

2.9. Computerization

2.9.3. Which software tools your company are using to support knowledge management in the design/development process?

Tool/Platform	*Installed solution*			
	Commercial Software	*Status of the implementation*		
		Just started	*In progress*	*Fully established*
Office Automation (Spread sheet, Word processing, etc.)				
CAD 2D				
CAD 3D				
Digital Mock-Up (DMU)				
Computer Aided Styling (CAS)				
Computer Aided Engineering (CAE)				
Finite Element Analysis Method (FEA/FEM)				
Computational Fluid Dynamics (CFD)				
Knowledge Based Engineering (KBE) and Design Automation				
Computer Aided Manufacturing (CAM)				
Computer Aided Process Planning (CAPP) / Digital Manufacturing & Factory				
Discrete Event Simulation (DES)				
Virtual / Augmented Reality (V/A R)				
Product Data Management (PDM/PLM)				
Document Management Systems (DMS)				
Workflow Management System (WMS)				
Enterprise Resource Planning (ERP)				
Supply Chain Management (SCM)				
CustomerRelationship Management (CRM)				
Supplier Relationship Management (SRM)				
Computerized Maintenance Management System (CMMS)				
Lifecycle Analysis Software (LCA)				
Software Project Management				
Other (specify)				

2.9.4. Knowledge management is carried out through a set of sub-processes implemented by different software functionality. Please, indicate which information system implements the following functionalities.

Software functionality		*State and implementation in information systems*			
Name	*Description*	*No*	*PDM/PLM*	*ERP*	*Other system*
Models and drawings management	It enables collaboration in the development process: synchronized access (check-in, check-out), user permissions and external user's access				
Change management	It supports change request				
Bill of materials management	It supports the generation of BOM, its different views, configurations and customizations				
Registers management	It enables product and component code management				
Product configurator	It supports marketing, budgeting and resources allocation activities				
Document management	It supports documents filing and sharing				
Projects management	It supports management automation and monitoring				
Products traceability	It supports product track and trace across its lifecycle				

2.9.5. How and which of the following product data are stored in your company for a specific product during the whole product life cycle?

Product Data	*Software support*						
	Paper form (not digital format)	*Personal Computer*	*Network Folder*	*Intranet, wiki, blog*	*Vault PDM / PLM*	*ERP Module*	*Other*
CAD Models (2D)							
Models and style sketches							
CAD Models							
CAE File							
CAM File							
Quality Function Deployment (QFD) Analysis							
Market Analysis / Business Plan							
Value Analysis and Engineering (VAE) Reports							
Specification Documents							
Process FMEA							
Product FMEA							
Test Reports							

Product Data	*Software support*						
	Paper form (not digital format)	*Personal Computer*	*Network Folder*	*Intranet, wiki, blog*	*Vault PDM / PLM*	*ERP Module*	*Other*
Design Validation Reports							
Customer Satisfaction Reports							
Production Part Approval Process (PPAP) Documents							
Engineering Change Request (ECR), Notification (ECN), Orders (ECO)							
Cost Analysis							
Feasibility and business plans							
Make or Buy Analysis							
Request for Quotations							
Request for Procurement							
Environmental Impact Analysis (LCA)							
Sustainability Reports							
Personal Data of created parts							
Personal Data normed codes							
Personal Data of finished products							
Safety / Compliance Reports and Certificates							
Sales catalogue							
Technical catalogue							
User manuals							
Maintenance, repair and technical manuals							
Product and service warranties							
Products and projects portfolio plan							
Bill of Materials (BOM)							
Manufacturing Bill of Materials (MBOM)							
Bill of Processes (BOP)							
Bill of Resources (BOR)							

Product Data	*Software support*						
	Paper form (not digital format)	*Personal Computer*	*Network Folder*	*Intranet, wiki, blog*	*Vault PDM / PLM*	*ERP Module*	*Other*
Other (specify)							

3. Part III – Critical factors and future improvements in product innovation and development process.

3.4.3. What are the main problems that affect the design process in your company? How often do these problems occur?

Problems	*Frequency*			
	Never	*Sometimes*	*Often*	*Always*
The development process involves too many signatures and bureaucracy is a norm				
In the development process the responsibilities are not well defined, as a result, the process is chaotic				
The projects are behind schedule because there are too many unnecessary and unneeded activities and tasks.				
Costs of projects are higher than pre-estimated budgets				
The projects are very complex to be adequately managed and designers get lost in the activities				
The designers are overloaded and cannot keep up with the overload				
Designers spend their considerable time for writing long documents, specifications and reports				
Designers often do many changes during the design process that frequently result in design reworks				
Engineers and designers have difficulties to extract knowledge from past projects				
The different systems used in the company(ex. CAD, PDM, ERP) have different formats (ex. file) which cause frequent manual work				
The products are not innovative enough to keep the existing market share or expand to new market segments				
Other (specify)				

3.5. In summary, what kind of improvement intervention has been or will be adopted in the design processes? When was it adopted?

Intervention	*Time and program of the intervention*				
	3 years ago	*In the last 3 years*	*Nowadays*	*Next future*	*Not expectie/Never*
Introduction of procedural changes and/or organization (ex. review of processes, change of responsibility, increased multidisciplinary teams, etc.)					
Outsourcing of activities and phases of the design and development processes (ex. outsourcing of design component, assembly design relocation, outsourcing of computing tasks and test, etc.)					
Adoption of formal methods and techniques to support the design (ex. use of QDF, Value Analysis, LCA studies, systematic innovation methods like TRIZ, etc.)					
Introduction of new Software for virtual prototyping (ex. CAD 3D, simulation tools, virtual reality, etc.)					
Introduction of new IT systems to support collaboration and data management (ex. PDM, PLM, etc.)					
Training plans and related activities					
Other (specify)					

3.6. Using the scheme of the previous question and analyzing the different conducted and/or planned actions to support the improvement of the design process, which have been and/or what will be the main expected benefits. Mark 'E' if the improvement is expected, and 'R' if the improvement is already realized (achieved)

Benefits	*Intervention*							
	Organization Intervention		*Outsourcing activities*		*SW virtual prototyping*		*Sw IT collaborat.*	
	E	*R*	*E*	*R*	*E*	*R*	*E*	*R*
Decrease development time								
Decrease development costs								
Decrease the number of necessaryresources								
Increase resource productivity								
Increase the speed of innovation								
Decrease production costs, storage, distribution, service, etc.								
Improve competitive position, flexibility to respond market demands								
Improve quality design (ex. greater reliability, better compliance, fewer errors, less waste, etc..)								
Improve control and management of the design and development activities								
Improve communication and collaboration between different designers and development actors								
Improve the knowledge management (ex. more capacity to retrieve and share experience from previous projects, avoid redoing work done in the past, etc.)								
Improve complexity products management (parts, components, variations, etc.)								
Improve the standardization and modularization of parts and components								
Other (specify)								

3.7. What problems the company was facing throughout the improvement interventions?

Criticality	*Intervention*			
	Organizat. Intervention	*Outsourcing design activities*	*SW Virtual protot.*	*SW IT collaborat.*
People has difficulty to accept changes				
The various actors involved have cultural and attitudinal differences.				
There is little support from management, absence of adequate sponsorship, low commitment from top management				
Technologic and information problems (ex. interoperability of systems, unreliability of applications, etc..)				
Other (specify)				

3.8. What are your future plans to improve the development and design process? What are the priorities?

Intervention	*Priority of future plans*				
	No interest	*Little interest*	*Moderage interest*	*High interest*	*Maxim priority*
Introduction of procedural and / or organizational changes					
Outsourcing of activities and stages of design and development process					
Adoption new methods and formal techniques to support the design					
Introduction new virtual prototyping SW					
Introduction new IT systems to support collaboration and data management					
Other (specify)					