UNIVERSIDADE CATÓLICA PORTUGUESA

Data Mining and Cluster Organisations

The case of PortugalFoods

Trabalho Final na modalidade de Dissertação
apresentado à Universidade Católica Portuguesa
para obtenção do grau de mestre em Gestão de Serviços

por

Rita Gomes Salgado Ferreira

sob orientação de
Professor Doutor Ricardo Morais
Professor Doutor António Andrade

Católica Porto Business School
Setembro 2016

Acknowledgements

This dissertation would not have been possible without the precious support received from several people.

I would like to start by expressing my deepest gratitude to my parents, who have constantly supported me throughout this process. My most sincere appreciation to them, for not only being an encouragement pillar but also for providing me the means to be fortunate enough to attend such a prestigious master.

I would also like to thank my thesis advisor, Professor Ricardo Morais for his dedicated guidance during this research, as well as to my thesis co-advisor, Professor António Andrade, who provided me with vital intuitions for the statistical treatment of this dissertation.

I am also extremely grateful to PortugalFoods' team, who from the very beginning have motivated and helped me during this project. I would like to particularly acknowledge the support given by Isabel Braga da Cruz, who was always available to provide me any help.

To all my close friends, I would like to express my sincere gratitude. Their support was vital for keeping me motivated, enthusiastic, and focused on the final goal. Among them I want to express a special thanks to Joana, Inês, Olga, Francisca, Diliana, and Adriana for all the patience and comprehension. I would also like to thank to Miguel Braga, who worked by my side for many hours, and Miguel Moreira for always making me laugh.

Finally, I would like to present my deepest thanks to my aunt, Natália, who provided me key insights for this dissertation.

Abstract

Even though the concept of clusters received a considerable amount of attention, the literature dedicated to cluster organisations is still very scarce.

On the other hand, the widely applicability of data mining to several industries, along with the benefits that it might bring to any organisation, have been the subject of various articles throughout the years.

This dissertation intends to assess how could cluster organisations benefit from the application of data mining on the type of services they provide.

Through the empirical study of a Portuguese cluster organisation – PortugalFoods – I analysed if data mining represents an opportunity for these governance bodies, particularly if applied as a new support tool on their market intelligence services. Supported by CRISP-DM methodology, and based on data provided by Mintel's databases, a prototype data mining project was developed. The findings of this study indicate that data mining could enhance PortugalFoods' market intelligence services, as well as their role as producers and disseminators of knowledge. Yet, challenges were also detected, due to the existence of several data's problems, which could jeopardize the future replication of this process.

Keywords: data mining; cluster organisations; market intelligence services; case study research; PortugalFoods; CRISP-DM

Table of Contents

List of Figures

List of Tables

1. Introduction

Over the last years, clusters organisations – intermediate organisations specifically set up to strength the competitiveness of selected clusters (Glaeser, 2013) – have proven to play a key role as important facilitators of entrepreneurship, cross-sectoral collaboration, and growth (Ketels et al., 2012). Aiming to successfully generate added value for the cluster participants, cluster organisations deliver specific, tailor-made services to their members, which can be categorized as "Cluster Identity and Attractiveness", "Innovation and R&D", and "Business development" (Sölvell & Williams, 2013).

Within "Business development", cluster organisations' activities include market intelligence collection. Key to successful anticipation of new business opportunities, market intelligence involves gathering accurate market information through surveys, interviews, roadmapping, foresight or market analyses, and/or trend scouting. This innovative service is particularly relevant for their SMEs members, whom may not have the resources to operate business development units for strategic market analysis and development (Lämmer-Gamp et al., 2014).

On the other hand, the past two decades have seen both a remarkable enlargement in data acquisition and data storage. The grand challenge of this new era of information sharing is to handle the massive volumes of high-dimensional data generated, collected, and stored on a daily basis, and turning them into knowledge. Responding to this challenge, the field of Knowledge Discovery on Databases (KDD) has emerged, as a reference to the "overall process of discovering useful knowledge from data" (Fayyad et al., 1996).

Data mining (DM), one of the crucial steps involved in this broader KDD process, is focused on extracting implicit and potentially valuable relationships, patterns, and interdependencies, established among data that has not yet been tapped (Han & Kamber, 2001). Whether considering insurance, direct-mail marketing, telecommunications, retail, or health care, DM has already proven to deliver major benefits to organisations - enhanced quality of service, improved profitability, and reduced cost of doing business (Apte et al., 2002) - by providing decision makers with the additional knowledge they need to make better informed decisions.

The necessity and importance of understanding cluster organisations' market intelligence services are first and foremost due to the confident decision-making that effective market intelligence provides on corporate strategy areas like market opportunity, market penetration strategy, and market development (Lämmer-Gamp et al., 2014). It is also due to the fact that literature devoted to these cluster governance bodies is still very scarce (Glaeser, 2013), and even less is dedicated to the type of services by them provided, which calls for further research in this area.

The main research question of this dissertation is therefore: How data mining influences the market intelligence services of cluster organisations?

This dissertation presents a five-chapter structure, including this introductory chapter.

Chapter 2 presents a review of literature on two of the keywords of this study – cluster organisations and data mining. Chapter 3 describes the empirical methodology that supported this research, aiming to address the research question in a more adequate manner.

Chapter 4 introduces the case study, particularly focusing on how PortugalFoods deliver their market intelligence services, but also describes the application of a prototype data mining project.

The fifth and final chapter presents the main advantages and challenges of applying data mining in cluster organisations' market intelligence services, as well as the study's limitations and suggestions for future research.

2. Literature Review

2.1 Clusters Organisations

The widespread interest in the economics of industrial location and, particularly, in the issue of industrial clusters as strategic entities in global industries (Tallman et al., 2004), follows as it became widely recognized that they can positively contribute to spur economic growth (Porter, 2003), regional industries' performance in terms of employment (Delgado et al., 2014) and entrepreneurship (Delgado et al., 2010). Besides, the global awareness of clusters as powerful engines to foster innovation (Porter, 1998; Baptista & Swann, 1998), enhance knowledge creation (Maskell, 2001; Tallman et al., 2004), and stimulate competitiveness (Lindqvist, 2009) has pushed governments and industry organisations to prominently introduce the cluster concept in several of their economic policy efforts (Coletti, 2010).

Accordingly, since Porter's (1990) seminal work, for whom "clusters are geographic concentrations of interconnected companies and institutions in a particular field" (Porter, 1998, p.78), a wide range of cluster development policies have been deployed, aiming to replicate the kind of synergies observed in spontaneous clusters[1] (Glaeser, 2013).

One frequent element of these cluster based strategies are cluster initiatives (CIs) (Lindqvist, 2009) – organised efforts carried out to launch, develop, and manage clusters, involving private industry, public authorities, and/or the research community (Coletti, 2010).

[1] Silicon Valley and Route 128, for example. See Saxenian (1996) for an overview.

CIs often imply the establishment of cluster organisations[2] (COs) - intermediate organisations that, by engaging in a wide variety of activities, attempt to strengthen the competitiveness of selected clusters (Glaeser, 2013).

COs are governed by a board in which private sector dominates (on average it represents 61% of the board) - with academia second, and public sector third – and are financially supported by a combination of public funding (54% of the revenues mainly come from regional and local public funding), consulting services, and membership fees (Lindqvist et al., 2013).

In Europe, the majority of these truly public-private partnerships[3] have been formed around 2007 - influenced by Michael Porter's (1990) book "The Competitive Advantages of Nations" - and are mostly common on technology intensive areas such as IT and Automotive (Ketels et al., 2012). Nevertheless, sectors including Food processing, Health care, Energy and Green Technology, are on the rise.

One way to analyse COs' mission is through the "gap model" (Lindqvist et al., 2013), according to which their key role consists on fostering value-enhancing interaction and cooperation between the different actors within a cluster (Sölvell & Williams, 2013). Figure 1 depicts the "gap model".

Hence, this model views clusters as a collective set of complementary actors of different types:

- Firms, the most relevant actors, and the ones who take innovations to markets and subject them to the test of competition;

[2] Even though both terms are frequently used interchangeably, these don't overlap completely. Whereas cluster initiatives refer to the process of cluster-related actions, cluster organisations refer to the organisational entities specifically set up to implement the strategies and tasks devised in it (Lindqvist, 2009).

[3] Regardless of the sector, geography, size, and age of clusters organisations, on average, they follow a 60/40 rule with 60% public financing (Lindqvist et al., 2013).

- Research organisations, which produce new and advanced knowledge;
- Educational institutions, such as school and polytechnics;
- Capital providers such as angel networks, venture capitalist, and commercial banking institutions, that provide the necessary capital to explore inventions and new business models;
- Government and public bodies, the ones responsible for implementing cluster policies, decide about infrastructure investment, and regulations, among other.

In an ideal cluster, these actors would collaborate perfectly, support each other, and form new ideas in both planned and unplanned meetings and interactions. However, in reality, communication between them tends to be flawed. Firstly, because these connections would hardly happen spontaneously, but also due to the existence of several types of obstacles to interaction like different norms and attitudes, weak networks, lack of trust, different vision, and limited knowledge across actor boundaries (Sölvell & Williams, 2013).

These barriers create gaps where should be paths, thereby restricting communication, collaboration, exchange of ideas, and diffusion of knowledge within the cluster, in short, preventing innovation processes. If persistent, as they usually are, such gaps can have major impact on clusters' competitiveness.

It is precisely to correct these knowledge, network, and collaboration failures that COs are formed. By building the cluster commons – the place where cluster actors meet and exchange ideas with cooperation as the main mechanism – the following seven innovation gaps can then be overcome (Sölvell & Williams, 2013):

- The research gap, limiting interaction between firms and research organisations;

- The education gap, limiting interaction between firms and education organisations;
- The capital gap, limiting interaction between firms and capital providers;
- The government gap, limiting interaction between firms and public bodies;
- The firm-to-firm gap, limiting interaction among firms in the cluster;
- The cross-cluster gap, limiting connections between firms in one cluster and another;
- The global market gap, limiting connections between cluster firms and global markets;

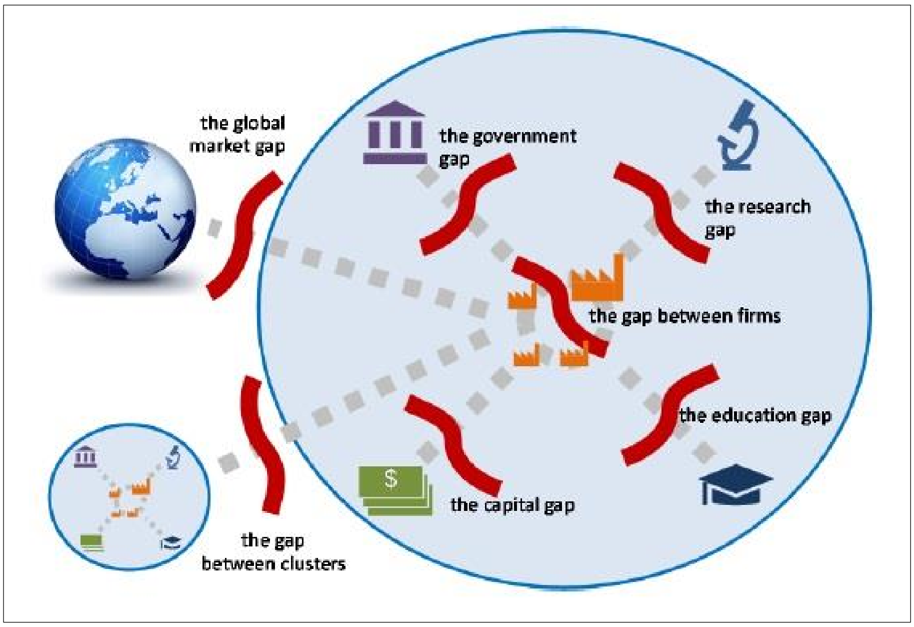


Figure 1 - The Gap Model (Source: Lindqvist et al., 2013)

COs are thus keen on creating proximity between the local actors (Glaeser, 2013). As "bridge builders", they connect businesses with academia, education with industry, and large firms with small firms. In order to do so, COs typically hold a variety of activities in parallel (Glaeser, 2013), equally considered as services provided by them to its members (Colleti, 2010).

Based on the comprehensive surveys by Lindqvist et al. (2013), and Sölvell & Williams (2013), such activities can be broadly categorized in three different types (see Figure 2): Cluster identity and attractiveness, innovation and R&D, and business development.

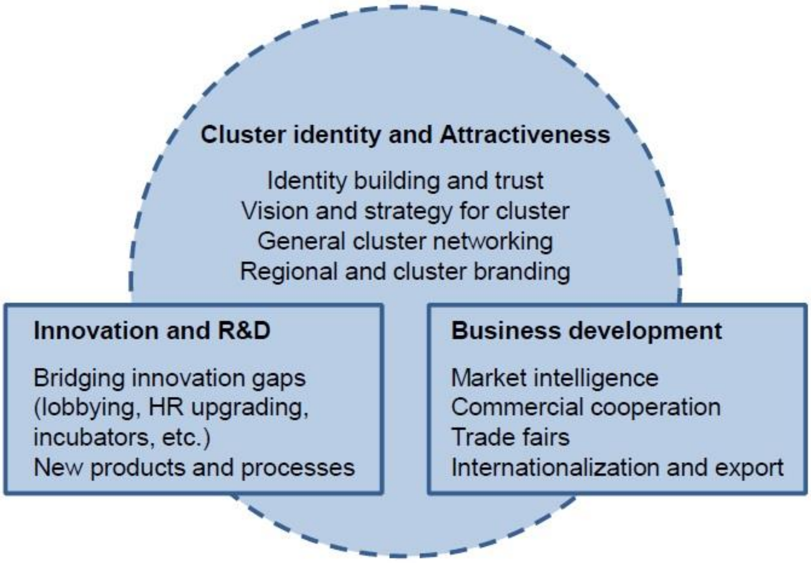

Cluster identity and Attractiveness
Identity building and trust
Vision and strategy for cluster
General cluster networking
Regional and cluster branding
Innovation and R&D
Bridging innovation gaps
(lobbying, HR upgrading,
incubators, etc.)
New products and processes
Business development
Market intelligence
Commercial cooperation
Trade fairs
Internationalization and export

Figure 2 - Types of activities performed by COs (Source: Sölvell & Williams, 2013)

The first type of activity, one of the most prioritised objectives pursued by COs (Ketels et al., 2012), is about overall cluster identity and attractiveness, that is, about building a brand, strategy, and vision for the cluster (Sölvell & Williams, 2013). Here the CO is mostly oriented towards building the cluster commons, which include building a sense of belonging, general trust, and networking (Lindqvist et al., 2013). By increasing the external visibility of the cluster through its brand awareness activities, this pillar then serves as an attractor for inflows of capital, investments, skilled labour, and new entrants (Lindqvist, 2009).

The second and third pillars, on the other hand, involve the type of activities aimed towards direct collaboration between firms and organisations (Lindqvist et al., 2013), i.e. focused on bringing different clusters' actors together to exploit to the fullest their (great) potential for dynamic interaction. Such collaboration can either be with an innovation and technology focus, which is the case of the second pillar, or with a business development focus, the case of the last pillar (Sölvell & Williams, 2013).

Therefore, in "Innovation and R&D", COs are dedicated to foster innovation and enhance cluster's R&D environment either through joint R&D projects or collaborations across the innovation gaps (Ketels et al., 2012): the research, education, and the government gap, more precisely. Accordingly, bridging to research might involve publishing cluster reports, sharing information through seminars, commercialization of research results, information gathering, incubator services, and inviting speakers (Sölvell et al., 2003). Bridging to education, on the other hand, covers education improvement and technical and management training, which improve and upgrade the HR supply within the cluster (Lindqvist, 2009). Lastly, bridging firms to public organisations might lead to reduced administrative obstacles, changes in regulation and policy, and redirection of public investments (Sölvell & Williams, 2013).

Finally, the third pillar comprises the type of activities that are mainly focused on business development among member firms. These include market intelligence collection, representing the cluster at trade fairs, export promotion/internalization, joint purchasing, and other commercial cooperation (Sölvell et al., 2003).

Market intelligence is about gathering accurate information related to a company's existing markets, customers, and competitors, as well as to its market growth potential for new products and services. Specifically collected through social media, surveys, interviews, roadmapping, foresight analyses, and/or trend scouting (Lämmer-Gamp et al., 2014) the analysis of this information provides confident decision-making on corporate strategy areas like market opportunity, market penetration strategy, and market development.

Key to successful anticipation of new business opportunities, gathering market intelligence may be done either in-house, through specialists agencies, or specialist sites companies. This innovative service often target SMEs, whom may not have the necessary resources to operate business development units for strategic market analysis and development (Lämmer-Gamp et al., 2014).

By delivering sophisticated market intelligence services to their members, COs can thus counterbalance such disadvantage.

Despite the evidences that COs can actually make an impact on their underlying clusters by enhancing innovation, growth, and competitiveness (Ketels et al., 2012), these cluster governance bodies have not been extensively studied (Glaeser, 2013). Therefore, and even though the EU-27 already counts 1400 COs and over 1600 other organisations playing critical roles within clusters[4], the vast majority of the academic literature devoted to cluster policies have treated COs merely as a secondary aspect of it, rather than addressing them as a sole research target (Lindqvist, 2009).

[4] Information accessed in January 2016 through http://www.clusterobservatory.eu.

2.2 Data Mining

The past two decades have seen an exponential growth in the amount of information and databases[5] (Santos & Azevedo, 2005), allied to a huge progress in information technology associated with them. However, as the flood of data swells (Witten et al., 2011), the need to have a technology that accesses, analyses, summarizes, and interprets information intelligently and automatically, became an evident necessity to most organisations (Chen & Liu, 2004).

Both the remarkable enlargement in data acquisition and data storage made it clear that, in order to better support the extraction of valuable information from these never seen streams of digital records, new manipulation techniques and special tools were being required (Fayyad et al., 1996). This need for automatic and effective approaches was even more augmented as the previous methods applied especially required direct hands-on data analysis (Hand, 1998). Based on that process constraint, but considering now the availability of huge volumes of data generated on a daily basis by institutions like hospitals, research laboratories, banks, insurance companies, retail stores, and by internet users (Pal & Jain, 2005), even accessing and sampling records could turn out to be a very complicated and time-consuming task. Additionally, as this accumulation of data takes place at an explosive rate, the proportion of it that people understand, decreases dramatically (Witten et al., 2005).

Rich sources of data, stored either in warehouses, databases, or other data repositories, were then readily available, but not easily analysable (Cios &

[5] It has been estimated that the amount of data stored in the world's datasets doubles every 20 months and that the size, and number, of databases are increasing even faster (Witten et al., 2005).

Kurgan, 2005). Lying hidden in all this data is a huge amount of information[6] , potentially important and useful, but that has not yet been discovered or made explicit (Han & Kamber, 2001). Handling this massive amount of generated, collected, and stored data and turning it into information, and then information into knowledge, became the grand challenge in this new era of digital information and information sharing (Witten et al., 2011).

As response to these major challenges and developments, the field of Knowledge Discovery in Databases (KDD) has emerged (Santos & Azevedo, 2005).

The term KDD was coined in 1989 as a reference to "the overall process of discovering useful knowledge from data" (Fayyad et al., 1996), based on the assumption that, regardless of the context and databases' size, the data itself (in raw form) is of little direct value. As so, being able to access to more quantity of data - whether related to business, medicine, science, or even government – does not necessarily means that the information contained on it is potentially useful to organisations (N. & Srivatsa, 2006). Instead, it is intelligently analysed data that is as a valuable resource to end users analysis, since it may help them gain the insights they need for improving business decision quality (Apte et al., 2002), as well as to lead, in commercial settings, to competitive advantages (Witten et al., 2005).

Data mining is one of the crucial steps involved in the broader KDD process (Figure 3), and is focused with the algorithmic means by which useful patterns can be extracted and enumerated from vast amounts of data (Fayyad et al., 1996). Even though the KDD and DM terms are often viewed as synonymous (Cios & Kurgan, 2005), Fayyad et al. (1996) emphasized the difference between them. According to these authors, KDD refers to the whole (interactive and iterative)

[6] If data is considered as recorded facts, then information is the set of patterns, or expectations, that underlie it (Wittten et al., 2005).

process of making sense of data - from the development and understanding of the application domain, to the action on the knowledge discovered. DM, on the other hand, represents the step within that same process that is concerned with the actual search and discovery of unsuspected and potentially valuable relationships, patterns, and interdependencies, established among data that has not yet been tapped (Han & Kamber, 2001).

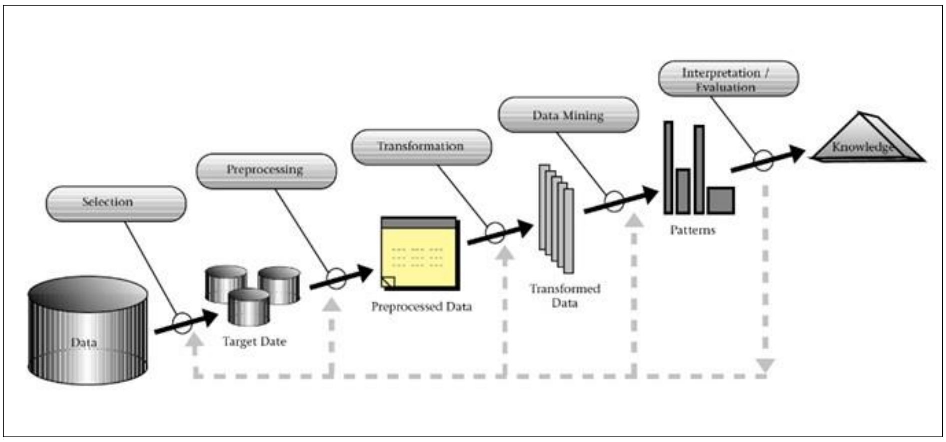


Figure 3 - The KDD Process (Source: Fayyad et al., 1996)

Aside from DM, all the additional steps comprised within the KDD process are equally important to ensure that useful knowledge is inferred from the data (Fayyad et al., 1996). Such steps include the data's selection, preprocessing, and transformation, as well as a proper interpretation/evaluation of the extracted (mining) results.

This is particularly evident when it is detected, in the available dataset, the existence of imperfect data. It commonly includes the presence of noise, inconsistency of the data, missing data, and ambiguous data, which ultimately may jeopardize the final goal of the overall KDD process. In such cases, the "preprocessing" and "transformation" steps will play a key role on both identifying and minimizing the occurrence of such data's problems (Gama et al., 2012).

As a truly multidisciplinary field, DM has incorporated several techniques from a wide range of other areas which include statistics, machine learning, pattern recognition, database technology, information retrieval, network science, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization (Han & Kamber, 2001; Fayyad et al., 1996).

Those fields are also concerned with inferring models from data, which leads to a natural question: how is DM different from these other data analysis' fields? The main difference relies both in the way each explore useful relationships among data (Santos & Azevedo, 2005), and in the dataset's size (Hand, 1998).

For example, while Pattern Recognition (PR) uses a method based on verification, i.e., the user builds its own hypothesis having apriori some goal in mind, the DM process is itself responsible for generating hypothesis, without any specific goal in mind, at the same time that ensures the improvement, autonomy, and reliability of results (Pal & Jain, 2005).

The first method is then almost exclusively dependent on analysist's ability to propose interesting hypothesis, in manipulating the attribute's complexity, and in refining the analysis based on the results of potentially complex databases (Santos & Azevedo, 2005). DM applications, by contrast, are an inductive exercise (Hand, 1998), which allows to go beyond data that has been explicitly stored. This will increase the opportunities to find patterns in data and subsequently to derive novel business knowledge (Witten et al., 2005).

This is where KDD or DM will deliver measurable benefits for any firm such as enhanced quality of service, improved profitability, and reduced cost of doing business (Apte et al., 2012). These have been visible demonstrated in a wide variety of industries like insurance, direct-mail marketing, telecommunications, retail, and health care. Typical applications include market segmentation,

customer profiling, fraud detection, evaluation of retail promotions, and credit risk analysis (Santos & Azevedo, 2005).
Additionally, DM applications deal with the analysis of large databases, whereas PR is typically concerned with datasets of moderate size (Pal & Jain, 2005).

Ultimately, the interdisciplinary nature of DM research and development, contributed to both its success and extensive applications, as well as to its reference in many different ways (Han & Kamber, 2001). Such capability to search for implicit and useful patterns is also known as knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing (Fayyad et al., 1996).

2.2.1 Data Mining Approaches, Tasks, and Techniques

Berry and Linoff (2000) have classified DM into two different approaches: undirected and directed data mining. These are also referred in the literature, respectively, as descriptive and predictive data mining (Fayyad et al., 1996; Han & Kamber, 2001; Gama et al., 2012), as well as unsupervised and supervised learning (Chen & Liu, 2004).

In undirected DM, the recognition of intrinsic structure, relationships, or affinities among data is accomplished without the imposition apriori of some source of restriction or initial guidance (Santos & Azevedo, 2005), i.e., no variable[7] is singled out as the target. Since it lets the data speak for itself, it is possible for this bottom-up approach to discover hidden patterns inside the data, which in turn, may provide very enlightening insights (Berry & Linoff, 2000). It is then up

[7] Also known as attributes or fields (Gama et al., 2012).

to the user to determine what significance, if any, these patterns have, and what they might mean. Often used during the data exploration steps, when the analyst has no idea what is looking for, undirected DM thus attempt to answer some of the following questions: What is in the data? What does it look like? Are there any unusual patterns in our customer base?

On the other hand, in directed DM, it is mandatory for data to be already pre-classified (Chen & Liu, 2004), which typically involves using known examples. Besides, this top-down approach implies that the analyst knows exactly what is looking for, i.e., has some direction for the search, or an idea of what might be predicted (Santos & Azevedo, 2005). For example: which customers are most likely to buy a specific type of car in the next year?

Thus, undirected DM focuses on exploring or describing the dataset to see what might be learned, whereas directed DM aims to build a model capable of predicting future trends (or values) of variables of interest to the user (Gama et al., 2012).

Both approaches can be addressed by using a variety of distinctive data mining tasks (Chen & Liu, 2004; Fayyad et al., 1996), being clustering, summarization, dependency analysis or affinity grouping, classification, and prediction the most commonly mentioned[8] in literature (Berry & Linoff, 2000; Santos & Azevedo, 2005; Han & Kamber, 2001; Fayyad et al., 1996).

Each of these tasks involve extracting novel and meaningful information from the data (Berry & Linoff, 2000), and are individually described below.

[8] Since there isn’t a standard terminology in the data mining community, several of these tasks are referred in the literature by different terms.

- **Clustering**[9] – Clustering consists on segmenting the data into interesting and meaningful subgroups, known as clusters (Chapman et al., 2000). However, as opposite to classification methods, clustering does not rely on predefined classes or examples. Instead, the groups are formed on the basis of self-similarity (Berry & Linoff, 2000): each cluster is characterized for being constituted only by objects that are, at the same time, similar between themselves and dissimilar to records of other clusters (Santos & Azevedo, 2005). This method can then be used in situations where there is not a training set of pre-classified records, which in turn, enhances the possibility to uncover unanticipated trends, correlations, or patterns (Chen & Liu, 2004). By presenting such advantages, clustering is often a step toward solving some other form of data mining (Han & Kamber, 2001). For example, in a marketing segmentation effort, clustering as the first step would allow to discover what kind of homogenous subpopulations of consumers (i.e., with similar buying habits) exist in those marketing databases, and only then analyse what type of promotion would better fit each cluster (Berry & Linoff, 2000). Appropriate techniques include clustering techniques, neural networks and visualization (Chapman et al., 2000).

- **Summarization** – Summarization aims to provide a concise description of characteristics for a subset of data (Fayyad et al., 1996), and is typically carried out at the early stages of a data mining project (Chapman et al., 2000). Summarization techniques are often applied in the exploratory data analysis and automated report generation (Santos & Azevedo, 2005), and

[9] In the literature, clustering can also be referred as segmentation or classification. Even though there is this ambiguity, Chapman et al., (2000) restricts the former term to the concept of creation of classes and the latter to the creation of models (that aim to predict known classes for previously unseen cases).

can range from simple descriptive statistical (the mean and the standard deviation for all fields, for example) to summary rules, multivariate visualization techniques, and functional relationships between variables (Fayyad et al., 1996). This gives the user an overview of the data's structure, a better understanding of its nature, and might also help him to form potential hypothesis for hidden information (Chapman et al., 2000). Even though summarization typically occurs in combination with other data mining methods (if possible as the first one being addressed) it may also be, by itself, an objective of a data mining project (Chapman et al., 2000). For example, in order to know which parts of a customer group call for further marketing strategies, a retailer might be interested in the distribution of customers by gender and age.

- **Dependency Analysis or Affinity Grouping** – Dependency analysis aims to find a model that describes significant dependencies among variables, i.e., that identifies groups of data which are typically associated to each other (Santos & Azevedo, 2005). Such models may exist at two levels: structured and quantitative. Whereas in the structural level the dependency model specifies which variables are locally dependent, in the quantitative level it specifies, according to some numerical scale, the strengths (or weights) of such dependencies (Fayyad et al., 1996). Associations are a special case of dependencies, and describe affinities of data items (Chapman et al., 2000), i.e., determine which data items frequently occur together in a given set of data (Han & kamber, 2001). Association analysis can then have its greatest impact on the sales area. For example, retail chains may use it to determine which items are usually purchased together, i.e., for market basket analysis (Berry & Linoff, 2000). A prototypical example for an association might be "in 40 percent of all purchases, chocolate and alcohol have been bought together". This type of analysis can also be used to identify cross-selling opportunities (Berry& Linoff, 2000). Association rules, Correlation and regression analysis, and Bayesian networks, are some of the most appropriate techniques for dependency analysis (Chapman et al., 2000). Both classification and prediction may need to be preceded by this task as dependencies are implicitly used for the formulation of predictive models (Chapman et al., 2000).

- **Classification** – Classification is the most frequently used data mining task and has a wide range of business applications - from financial areas to marketing and sales (Santos & Azevedo, 2005). The objective is to build classification models (also referred as classifiers) that, through the examination of the attributes or features of a newly presented data object, are able to assign it to one of a predefined set of class labels (Berry & Linoff, 2000). These are a discrete or symbolic value and can be given in advance, either defined by the user or derived from segmentation (Chapman et al., 2000). In order to decide how previously unseen and unlabeled data objects should be classified, and so, to predict the (correct) class label, these models base themselves on the analysis of a training set of pre-classified examples (i.e., data objects for which the class labels are known) (Han & Kamber, 2001). For example, in order to assess the credit risk of a new customer, banks may generate a classification model from existing customer data related to their credit behavior, sex, age, income, etc. By creating two classes (good and bad customers) such model could then be used to classify and assign new customers to one of these classes, and accordingly, to accept or reject them to get a loan (Chapman et al., 2000). Discriminant analysis, decision trees, case-based reasoning, rule induction methods, neural networks, K nearest neighbor and genetic algorithms are the most appropriated techniques for classification modelling (Chapman et al., 2000).

- **Prediction** [10] – Prediction is very similar to classification since both methods aim to assign (based on training examples) previously unseen and unlabeled data objects to a predefined class (Berry & Linoff, 2000). Nevertheless, prediction specifically refers to the case when the predicted values are numerical data, that is, to when the user wishes to predict some missing value or unavailable data values (rather than class labels) (Han & Kamber, 2001). Therefore, the only difference is that in prediction the target attribute (class) is not a discrete categorical attribute (as in classification), but instead a continuous one (Chapman et al., 2000). For example, based on the values (or reliable estimates) of attributes like advertisement, inflation and exchange rate, one company is able to predict its expected annual revenue for the next year. Prediction occurs in a wide range of applications, being regression analysis, regression trees, neural networks, k nearest neighbor and genetic algorithms the most appropriate and common techniques (Chapman et al., 2000).

In sum, undirected DM uses tools that support clustering, summarization, and affinity-grouping, whereas classification and prediction are examples of directed DM. DM efforts frequently involve a combination of both (Berry & Linoff, 2000), which together solve the business problem (Chapman et al., 2000).

[10] In the literature, prediction is also commonly referred as regression and forecasting (when dealing with time-series data) (Chapman et al., 2000).

Associated to each one of these tasks, there is a wide variety of data mining techniques, which are employed in order to extract information from the data (Berry and Linoff, 2000), being decision trees, neural networks, genetic algorithms, and bayesian networks the most commonly used (Santos & Azevedo, 2005). In turn, each technique comprises a myriad of different algorithms, defined by Berry & Linoff (2000) as "the step-by-step instructions" that lead to the actual technique's implementation.

Based on Chapman et al., (2000), a number of common DM techniques are summarized in Table 1, organized by their use for directed or undirected data mining.

Undirected Data Mining Techniques	**Directed Data Mining Techniques**
• Clustering Techniques • Neural Networks • Bayesian Networks • Association Rules • Correlation Analysis	• Discriminant Analysis • Decision Trees • Case-based Reasoning • K Nearest Neighbor • Genetic Algorithms • Regression Analysis • Neural Networks

Table 1 - Data Mining Techniques

2.2.2 Data Mining Methodologies

In order for a DM project to succeed, as well as to become easier to understand, plan, develop, and ultimately implement, it necessarily needs to be framed into the context of a methodology (Santos & Azevedo, 2005).

Currently, there are two well-known and properly developed methodologies for carrying out DM projects (Santos & Azevedo, 2005): CRISP-DM (Cross-Industry

Standard Process for Data Mining) and SEMMA (Sample, Explore, Modify, Model, Assessment). Nevertheless, CRISP-DM is by far the most widely used[11]. CRISP-DM methodology was conceived in late 1996, after several years of discussion, by a consortium of leading data mining specialists[12] - both users and suppliers: DaimlerChrysler AG, SPSS Inc., NCR, and OHRA (Chapman, et al., 2000).

Its development was driven by two reasons: on one hand, due to the increasing and generalized interest of the data mining market, and on the other hand, by the existing consensus that the industry quickly needed a standard process for data mining (Chapman, et al., 2000). At that time, the CRISP-DM project addressed part of these issues by proposing a comprehensive process model, which is independent of both the industry sector and the technology used, and that aims to make DM projects less costly, faster, as well as more reliable, repeatable, and manageable (Wirth & Hipp, 2000).

CRISP-DM methodology succeeds as it is both based on theory and on the real-world experience of how people conduct data mining projects (Santos & Azevedo, 2005). It is exactly by incorporating that practical knowledge that CRISP-DM provides the structure and the flexibility necessary to suit the needs either of experienced data mining people, as well as of people with lower technical skills and/or with little time to experiment different approaches (Wirth & Hipp, 2000).

The CRISP-DM methodology is described in terms of a hierarchical process model, which comprises an overview of the life cycle of a data mining project

[11] According to the latest KD nuggets' pool (http://www.kdnuggets.com), in 2014 CRISP-DM remains the most popular methodology for analytics, data mining, and data science projects, with essentially the same percentage as in 2007 (43% vs 42%). In turn, and considering the same period of time, SAS SEMMA methodology has suffered a big decline (from 13 to 8.5%).

[12] The project was also partly sponsored by the European Commission under the ESPRIT program (Chapman et.al, 2000; Wirth & Hipp, 2000).

(Chapman et al., 2000). This life cycle model consists of six phases[13] , depicted in Figure 4: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. Yet, this sequence is not strict (Wirth & Hipp, 2000).

Even though at the beginning of the DM project there is precedence between phases, as it evolves and more insights are gathered from data, often it is required to move back and forth between different phases (Chapman et al., 2000). In practice this will depend on the outcome and performance of each phase and/or of some specific task of a phase. The arrows in Figure 4 indicate the most important and frequent connections and dependencies that exist between the six phases along the cycle, as the outer circle symbolizes the cyclic nature of DM itself (Santos & Azevedo, 2005).

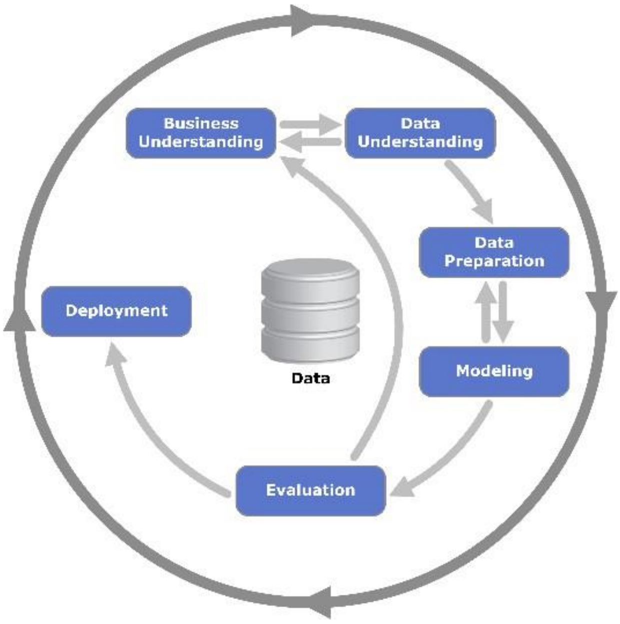


Figure 4 - Phases of the CRISP-DM process model (Source: Chapman et al., 2000)

[13] Each of these six phases is subdivided in (several) generic and specific tasks, in which are described how actions should be carried out according to certain specific (data mining) situations (Chapman et.al, 2000).

Business Understanding is considered to be the most crucial step in the whole DM process. Identifying and understanding the need to do DM, i.e. understanding the problem to be solved, both project's objectives and requirements - either functional, technical or temporal - from a business perspective (Chapman et al., 2000), along with uncovering factors that could influence the final outcomes are the main goals of this initial phase of the CRISP-DM methodology (Santos & Azevedo, 2005). This phase evolves 4 tasks, as shown in Figure 5, from which it is expected to obtain a DM problem definition, and a preliminary project plan[14] designed to achieve the objectives (Wirth & Hipp, 2000).

The *Data Understanding* phase starts with an initial data collection and proceeds with activities that enable the user to become familiar with the data (Chapman et al., 2000), discover first insights into it, identify interesting subsets and the most obvious associations among data to form hypotheses for hidden information, and/or to identify data quality problems (Santos & Azevedo, 2005). The four tasks included in this phase are presented in Figure 5.

The *Data Preparation* phase comprises all the activities evolved in the construction of the final dataset, i.e. the data that is going to be used into the modeling tool(s) and that inevitably has suffered several optimization tasks since the initial raw data (Santos & Azevedo, 2005). This phase includes five tasks (Figure 5), which are likely to be performed multiple times and without a prescribed order (Chapman et al., 2000), including table, record and attribute selection, data cleaning, construction of new attributes, and data formatting for modelling tools.

[14] This should include: business' objectives and success criteria, an inventory of the available resources, an evaluation of all the requirements, assumptions and constraints, and risks and contingencies, as well as an initial assessment of DM goals, tools and techniques (Santos & Azevedo, 2005).

In the *Modeling* phase numerous modelling techniques are selected (e.g., decision trees, neural networks, genetic algorithms) and applied, at the same time as their parameters are adjusted to optimal values (Chapman et al., 2000). Normally, there are several techniques for the same DM problem type (e.g., both decision trees and neural networks can be applied to classification problems), some of which require specific forms of data[15] (Wirth & Hipp, 2000). Even though both the DM problems and objectives were previously specified, it is only at this stage that the data (already prepared for modelling) is submitted, and the techniques that best fit the DM objectives chosen, i.e. is the model applied to the final dataset (Santos & Azevedo, 2005). This phase is structured in four tasks, as in Figure 5.

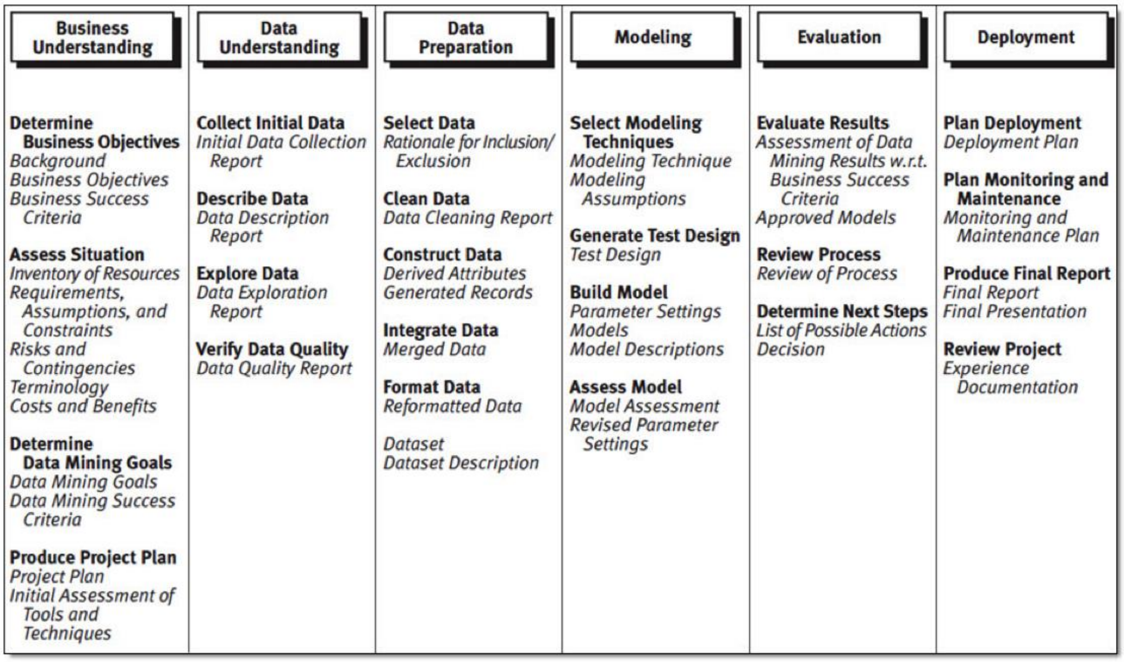
Before proceeding to the final deployment of the model(s) generated, the *Evaluation* stage aims to more thoroughly evaluate the utility of the model, review all the steps executed in its construction, and verify if it accomplishes the business objectives[16] (Santos & Azevedo, 2005). It includes three tasks (Figure 5) and it is expected that, at the end of this phase, a decision on the use of the DM results could be reached (Chapman et al., 2000). At this point, it is also frequent to return to the previous phases as new insights changes the understanding of what can be done, and new ideas for new problems appear.

The creation of the final model does not necessarily mean the end of a DM project. The knowledge extracted from the application of the model needs to be

[15] For this reason, the Modelling and the Data Preparation phases may be strongly interrelated: often, while modelling, the user gets ideas for constructing new data or realizes the existence of specific data problems (Wirth & Hipp, 2000) (e.g., the data available isn't enough and should be enriched). Hence, returning to the Data Preparation phase is frequently necessary.

[16] Regarding this last aspect, it is important to determine if any important business issue has not been (sufficiently) considered (Chapman et al., 2000).

organized and presented in a way that the customer can use it[17] (Santos & Azevedo, 2005), which is exactly what makes the DM project worthwhile. Depending on the requirements, the *Deployment* phase can either be very simple, as creating a report, or very complex, as implementing the overall DM process across the enterprise (Chapman et al., 2000). Often, it is the client, instead of the data analyst, the responsible for carrying out the four tasks comprised in this phase (Figure 5). In any case, it is important for the customer to understand up front what actions need to be executed in order to actually make use of the created models (Wirth & Hipp, 2000).



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

Figure 5 - Overview of the CRISP-DM tasks (Source: Chapman et al., 2000)

The CRISP-DM methodology is extremely complete and documented since each of its six phases is properly organized, structured, and defined (Santos & Azevedo, 2005), which allows for the project to be easily understood and

[17] Real-time personalization of web pages or repeated scoring of marketing databases, for example (Chapman et al., 2000).

reviewed, reliably and efficiently repeated by different people, and/or adapted to different situations (Wirth & Hipp, 2000). When applied, this methodology provides a set of different documents, as a support tool in the development of the DM process. Based on these aspects, which translate as advantages for the user, CRISP-DM was the methodology chosen to support the prototype data mining project developed in this study.

3. Research Methodology

3.1 Research Strategy and Design

Case study research was deemed the most appropriate research strategy to explore the potential of DM on the market intelligence services provided by COs, and thus answer this study's research question: "How data mining influences the market intelligence services of cluster organisations?"

Yin (1984, p. 23) defines the case study research strategy as "an empirical inquiry that investigates a contemporary phenomenon in depth and within its real-life context; when the boundaries between phenomenon and context are not clearly evident; and in which multiple sources of evidence are used."

According to Coutinho and Chaves (2002), it is exactly for involving an exhaustive and detailed study of a well-defined entity, "the case" - which can either be a single individual, a small group, an organisation, a community, or even a nation - that this research strategy is best recognized for.

Based on several authors, Benbasat et al. (1987) presented a summarized list of eleven key case studies' characteristics:

1. Phenomenon is examined in a natural setting;
2. Data are collected by multiple means;
3. One or few entities are examined;
4. The complexity of the unit is studied intensively;
5. The study focuses on exploration, classification and hypothesis development;
6. The set of independent and dependent variables is not specified in advance;

7. The results derived depend heavily on the integrative powers of the investigator;
8. Changes in site selection and data collection methods could take place as the investigator develops new hypothesis;
9. Focus on "how" and "why" questions because these deal with operational links to be traced over time rather than with frequency or incidence;
10. No experimental controls or manipulation are involved;
11. The focus is on contemporary events.

Yin (2009) states that these last three characteristics, which reflect 1) the type of research question, 2) the extent of control an investigator has over actual behavioral events, and 3) the degree of focus on contemporary as opposed to historical events, enables to distinguish case studies from other types of social science research. Accordingly, case study will then be the preferred method when 1) "how" and "why" questions are being posed, 2) the investigator has little control over events, and 3) the focus is on a contemporary phenomenon within a real-life context.

Additionally, the author highlights case study's unique ability to deal with a full variety of evidence. Therefore, in case study strategy the employed data collection techniques may be various – including documents, archival records, interviews, direct observation, participant observation, and physical artefacts (Yin, 1994) – and are likely to be used in combination.

Also according to Yin (2009), within the case study research strategy there are two approaches: single and multiple case studies. Often adopted where it represents a critical case or, alternatively, an extreme or unique case, a single case may also be selected as it provides the researcher with an opportunity to observe and analyse a phenomenon that few have considered before (Saunders et al., 2009).

Whereas for Yin (1994) the case study research strategy may be used to explore, describe and explain – whether it is an exploratory, descriptive, or explanatory case study - Ponte (1994) states that its main purpose is to describe and analyse. The investigator does not intend to change the situation, but rather to understand it as it is.

3.2 Case Study Selection

This dissertation's design includes a single case study of a Portuguese cluster organisation - PortugalFoods.

The reasons to choose PortugalFoods as the case were based on: i) the recognition of the organisation as the main intermediary and developer of the Portuguese agrofood sector, ii) the relevance of market intelligence services on their annual action plan, iii) the interest of PortugalFoods managers in the case study.

Considering the purpose of this investigation, this study is particularly focused on PortugalFoods' Knowledge Division department, the responsible for collecting market intelligence (mainly accomplish through PortugalFoods' access to two external sources of data – Innova Market Insights and Mintel GNP Analysis), and delivering it to their members (by annually sending three type of reports – periodicals, support, and tailored).

3.3 Data Collection

Considering the purpose of this investigation, multiple sources of evidence were used: direct observation, unstructured interviews, documents, and data analysis. Thus, in this dissertation primary data is mainly collected through direct observation and unstructured interviews, which is complemented with secondary data such as research articles, books, as well as organisation's documents (e.g., reports previously produced by Knowledge Division department).

Through direct observation - which allowed data to be often collected in a non-systematic manner thorough the process - and documental analysis it was possible to develop a deeper understanding of PortugalFoods' core activities, particularly focusing on how their market intelligence services were being provided to its members.

On the other hand, through data analysis it was possible to have an improved insight on how these particular type of services could potentially be enhanced through the use of data mining techniques.

To complement the research, unstructured interviews with the chief of PortugalFoods' Knowledge Division, Isabel Braga da Cruz, were also conducted. Crucial in the case study's development (Yin, 2009), interviews may be categorized according to their level of formality and structure into: unstructured or in-depth interviews, semi-structured interviews, and structured interviews.

Characterized for not having a predetermined list of questions, unstructured interviews give the opportunity to the interviewee to talk freely about events, behavior, and beliefs in relation to the topic area (Saunders et al., 2009).

In this study, the one-to-one interviews allowed to identify, from a business perspective, PortugalFoods' motivations and intended goals for the implementation of data mining on their market intelligence services, and therefore, for the realization of this study. Thus, as part of a case study strategy, these interviews aimed to understand the "what" and the "how", but particularly to explore in depth the "why" (Saunders et al., 2009).

4. Empirical Results

4.1 PortugalFoods

Considered as the "umbrella" brand of the Portuguese agrofood sector, PortugalFoods comprises several companies, entities from the national scientific and technological system, and regional and national entities that represent the various subsectors of the Portuguese agrofood sector. Currently, more than 130 companies (the majority of them SMEs and micro enterprises [18]), several Portuguese universities and institutes (University of Porto and Polytechnic Institute of Bragança, for example) and national entities such as DGS (the Portuguese General Direction of Health), are associated to PortugalFoods' brand[19].

Founded in 2009, this association is promoted by the AgroFood Competitiveness and Technology Centre, and it's recognized as the main intermediary and developer of the Portuguese agrofood sector throughout the industry, the Portuguese Ministry of Economy, and the Ministry of Agriculture and Fishing.

Based on two strategic pillars, innovation and internationalization, PortugalFoods aims to enhance the companies' competitiveness in the agrofood sector by:

- Increasing the agrofood companies' technological index;
- Promoting the production, transfer, and application of knowledge oriented towards innovation;

[18] According to the European classification of companies (Europeia, 2003).

[19] An updated list of all members can be found in their website (http://www.portugalfoods.org).

- Promoting the internationalization of the agrofood sector companies through an active support, either by identifying and capturing opportunities in priority markets, as well as by improving their capacities and qualifications for internationalization.

In order to successfully fulfill its mission, PortugalFoods' operational team organizes itself in two interconnected departments: Market and Knowledge Division.

Oriented towards innovation, Knowledge Division supports the active diffusion of knowledge, encourages innovative practices, strengthens synergies for strategic competitiveness, and promotes the interaction (networking) and cooperation between PortugalFoods' members.

The Knowledge Division's key activities then include: encouraging joint R&D projects, sharing information through seminars, inviting speakers, and producing reports as support tools for their members.

Aiming to provide a quick and easy access to relevant market information, product, and consumer, alluding to the activity area and the main requirements and needs of each member, such reports can be categorized as followed:

- Periodicals - sent to all PortugalFoods' members three times a year, which can either be categorized, that is, focused on a specific food category ("Dairy" or "Chocolate Confectionery", for example), or more generalized, that is, based on trends or problems that are having major impact on the agrofood sector as a whole;
- Support – sent to PortugalFoods' members as a back-up for their participation in international fairs or trade missions;
- Tailored – elaborated when a PortugalFoods' member specifically requests it; pretends to scan and prospect more detailed information: get

an overlook of some food category's performance in specific regions, for example.

Through these reports PortugalFoods' members can then have access to the major market trends as well as to the actual consumer "value added perception", relevant information for after exploit new and successful business opportunities, in addition to adapt to changes in customer behavior.

In order to provide such sophisticated market intelligence services, PortugalFoods mainly relies on two different external sources of data - Innova Market Insights [20] and Mintel GNP Analysis [21] - capable of continuously monitoring new product launches worldwide, within the several existing food categories.

Conscious of how this type of services generate added value to their members, the idea of employing data mining urged as the association wished to optimize to the fullest the privileged access that they have to the (already very enlightening) information provided by both data sources.

As a new support tool, data mining could then complement PortugalFoods' reports, by trying to find unsuspected and potentially useful information among data that have not yet been tapped.

This particular activity, along with the crucial efforts developed by Market Division's department, create competitive advantages for PortugalFoods' members, thereby ensuring their sustainability in a competitive environment.

In fact, as an acknowledgment for all the excellent work developed, PortugalFoods received the 'Gold Label of the European Cluster Excellence

[20] http://www.innovadatabase.com/

[21] http://www.mintel.com/

Initiative'. This certification was attributed in 2014 by ESCA (The European Secretariat for Cluster Analysis), being the association the first Portuguese entity to ever obtain it.

4.2 Data Analysis

This subchapter describes how data obtained for analysis was handled, processed, and analysed, through the application of a prototype DM project in PortugalFoods.

CRISP-DM was the methodology chosen to support the DM project due to its flexibility, widely applicability (standard process model independent of both the industry sector and the technology used), efficiency, reliability (Wirth & Hipp, 2000), and acceptance among the community as the most complete and well documented methodology (Santos & Azevedo, 2005). In that sense, the project followed (as illustrated in Figure 5) the sequence of phases, and respective tasks, comprised in CRISP-DM methodology. Yet, due to the small dimension of this DM project, prototype like, only the first three - "Business Understanding", "Data Understanding", and "Data Preparation" - were applied. Each will be individually described in the following subsections.

The presented DM project is then focused on exploring Mintel's databases, based on a sample of cheese product launches occurred between 2010 and 2014, in the European market. In order to accurately meet the purpose of this DM project, and due to its overall good quality and ease of use, the chosen DM tool[22] was IBM

[22] According to a survey conducted by Rexer Analytics, in 2013, the most used commercial data mining tools in 2013 were R (by 70% of data miners), IBM SPSS Statistics, Rapid Miner, SAS, and Weka. These results are available in the KDnuggets website (http://www.kdnuggets.com).

SPSS Statistics. Through this tool representations were built, which can later help PortugalFoods to draw conclusions related to the product's nutritional values.

4.2.1 Business Understanding

As illustrated in Figure 5, Business Understanding comprises 4 interconnected tasks: "Determine Business Objectives", "Assess Situation", "Determine Data Mining Goals", and "Produce Project Plan". This first phase is then focused on identifying both project's goals and requirements from a business perspective, i.e., understanding what the customer really wants to accomplish, for after convert this knowledge into a definition of the data mining problem and a preliminary project plan (Chapman et al., 2000).

In the case of this empirical study, this crucial step translated in an unstructured interview to Isabel Braga da Cruz, chief of PortugalFoods' Knowledge Division (the organisation's department most impacted by the prototype DM project), on the 20th of July 2015.

Knowledge Division's primary objective was to optimize to the fullest the privileged access that they have to both external data sources – Innova and Mintel.

In order to properly balance that objective with possible project's constraints and risks, the following aspects were also taken in consideration: the type of data sources available for this project (which included Innova and Mintel's databases, as well as reports by them provided to its members), the availability of enough data to reach out the intended objectives (often there is not enough observations, or more commonly, lack of key attributes), as well as the time schedule provided for the realization of this dissertation.

Based on the organisation's needs and expectations, the goal of this DM project consists in extracting the most "value-added" information possible from the databases that PortugalFoods has access to, that is, in refining the (already very enlightening) information possible to obtain with more accurate and statistically significant results.

4.2.2 Data Understanding

As illustrated in Figure 5, Data Understanding evolves 4 different tasks: "Collect Initial Data", "Describe Data", "Explore Data", and "Verify Data Quality".

Once more, in order to complete this step, an interview to Isabel Braga da Cruz was required. At this point, and before proceeding any further, it was crucial to determine which type of data was going to be the focus of this project, that is, which product's category, market, and time frame would be of most interest for PortugalFoods to get statically explored and analysed.

It was then decided, respectively, that the project's data would consist on launches of cheese products that occurred, in the European market, between 2010 and 2014. Having these parameters set, the data was after accessed (loaded in the excel format) through Mintel GNPD Analysis.

Therefore, and according to Mintel's terminology, "Dairy" was the (broader) product's category chosen for this project, along with the following five product's subcategories: "Curd & Quark", "Fresh Cheese & Cream Cheese", "Hard Cheese & Semi-Hard Cheese", "Processed Cheese" and"Fresh Cheese & Cream Cheese", "Hard Cheese & Semi-Hard Cheese", "Processed Cheese" and "Soft Cheese & Semi-soft Cheese". Each one is described, more or less accurately, via a number of different features (Appendix 1). Besides, according to Mintel's glossary, the European market includes the following 24 countries: Austria, Belgium, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Spain, Sweden, Switzerland, Turkey, UK, and Ukraine.

With all these criteria applied to the search, an initial dataset of 13274 records, each one representing a different launch of cheese products, was obtained, as well as the 41 attributes presented in Appendix 2.

In this dataset an interesting fact stir up most attention: all the product launches' nutritional information was compressed within just one text cellule named "Nutrition", i.e., presented as a whole to the analyst.
This was particularly relevant for the DM project since Mintel is already capable to perform simple statistical analysis in the majority of these attributes, present the results in graphs and plots, and ultimately extract lots of valuable information. Yet, the products launches' nutritional values is the exception. How Mintel operates, does not allow for the user to have access to that type of information, which not only represented a major opportunity for this project, but also turned out to be its main focus.

Given both the huge variety of attributes (many of them not particularly relevant to achieve the project's goal) and the high dimension of the available dataset, the two remaining tasks comprised within the "Data Understanding" phase, weren't strictly after followed. In fact, as mentioned in the literature review section by Wirth & Hipp (2000) the sequence of the six CRISP-DM model phases' isn't rigid, (i.e., moving back and forth between different phases is usually required). This particular project was no exception as the "Data Preparation" phase (and more precisely the "Select Data" task), were subsequently performed.

4.2.3 Data Preparation

Data Preparation phase includes 5 different tasks – “Select Data”, “Clean Data”, “Construct Data”, “Integrate Data”, and “Format Data” – which ultimately construct the final dataset(s) that will later be used for modelling or the major analysis work of the project (Chapman et al., 2000).

As suggested by Chapman et al. (2000), in order to select the data to be used for future analysis, three different criteria were applied: relevance to the data mining goal, quality, and technical constraints.

Consequently, “Product Variant”, “Category”, “Number of variants”, “Product Description”, “Bar code”, “Production code”, “Primary Image Thumbnail”, and “Record hyperlink”, were immediately removed from the initial dataset - neither of these attributes would be relevant for the data mining goal. Due to the same reasons, “Brand”, “Company”, “Ultimate Company”, “Package Material”, “Company county/state”, “Ingredient 1”, “Ingredient 2”, “Ingredient 3”, “Ingredient 4”, “Ingredient 5”, “Ingredient6”, “Ingredient 7”, “Ingredient 8”, “Ingredient 9”, and “Remaining Ingredients” were too removed.

Additionally, since Europe was the chosen market for this project, "Price in euros" was the attribute selected for the following analysis, and accordingly, "Price in US dollars", "Currency", and "Price in local currency" excluded from the dataset.

“Date published” was also selected, with the objective to deeper understand how some of the other attributes’ values evolved between 2010 and 2014 (the 4 years considered for this data mining project). However, this attribute was presented in the date format dd/mm/yyyy. Through excel formulas, it was extracted only the years in which the product launches occurred (the most relevant data for this project), which were subsequently assigned to “Year”, the newly created attribute.

Since this project is mainly focused on exploring the products' nutritional information, "Nutrition" was, by far, the attribute that deserved the most attention in this phase. In order to analyse this text cell[23], that hasn't yet been taped and might provide valuable information to the organization, several steps were followed as described below.

To begin with, seven new attributes (corresponding to the seven categories of nutrients mostly mentioned in "Nutrition") were created: "Energy", "Protein", "Fat", "Saturated Fat", "Salt", "Carbohydrate", and "Sugar". This division enable to better understand the nutritional constitution of each product (individually in terms of sugar and fat content, for example), as well as the impact that each one might have on other attributes considered for the analysis (the product's price, for example).

Since "Nutrition" also specifies the quantity of product to which the product's nutritional values are referring to ("per 100g", for example), the attribute "Quantity (g)" was also created.

Excel formulas were then used to split the values contained in "Nutrition", by the corresponding 8 newly created attributes.

Data quality analysis on the results of this procedure revealed major issues, which requested a considerable amount of time to solve. In this project, "Explore Data" and "Verify Data Quality" tasks were thus almost simultaneously performed with the ones included in Data Preparation step, which is perfectly acceptable with the CRISP-DM methodology - an "highly iterative, creative process with many parallel activities" (Chapman et al., 2000).

[23] A standard example of how the product's nutritional values are specified within the "Nutrition" attribute is: "per 100g: energy 1,370kJ/330kcal, protein 14g, carbohydrate 1g (of which sugar <0.5g), fat 30g (of which saturated fats 15g), Fibres 0g, sodium 1.11g, salt 2.8g"

Firstly, it was noticed that Mintel's database didn't had a standard terminology for some of these categories of nutrients since "Fat" was also regarded as "Lipid", as well as "Carbohydrate" by "Glucides". In was decided to adopt, in both cases, just one of the possible terms – "Fat" and "Carbohydrate" were the ones chosen.

Additionally, it also didn't had a standard level of measurement: it was detected, for all the seven nutritional attributes, that records could be introduced either in milligrams or grams. Again, decisions and actions have been taken, through the adoption of grams as the measure of reference, along with the manual correction of records from milligrams to grams.

Related problems, but particularly concerning the product's energy and salt values were also noticed: several records wouldn't present their values in terms of Kcal and content of salt, but instead in terms of KJ and sodium. Equally to the above procedure, it was adopted that "Energy" would always be presented in terms of Kcal, and "Salt" in terms of salt content.

In these particular cases, it was chosen to use the formulas below, and thus, to convert the values of records in KJ to Kcal, and its values in sodium in terms of table salt content:

- 1 Kcal = 4,184 KJ
- 1 gram of sodium = 2,5 grams of salt table

Further analysis on "Nutrition" displayed additional singularities: some of this attribute's cells, instead of having a number in front of each nutrients' category, had the text "trace" ("per 100g: energy 1,450kJ/350kcal, carbohydrate trace", for example). In such cases, it was opted to consider the value of zero for the corresponding nutritional attribute.

After having addressed these data quality problems, the values for “Energy”, “Protein”, “Fat”, “Saturated Fat”, “Salt”, “Carbohydrate”, and “Sugar” were all homogenized - in order to correctly compare these products’ nutritional values among different records, a standard measure was needed. Through basic statistics on the attribute “Quantity (g)”, it was noted that its values could range from 6 up to 220 grams. Faced with this dispersion it was opted to have all values of the nutritional attributes by 1 gram of product.

That was accomplished by dividing the values of those seven attributes by the corresponding “Quantity (g)” value. Each attribute was after renamed to “Energy/1gr”, “Protein/1gr”, “Fat/1gr”, and so on.

Due to identical reasons, the attribute “Price/1gr” was too constructed, this time derived through the attributes’ values of “Price in euros” and “Unit pack size (ml/g)” - in the original dataset there were product’s prices referent to 20 grams of (cheese) product, as well as to 2000 grams. Each product’s price became then presented by 1 gram of product.

When checking the quality of the data for this particular attribute, the existence of numerous errors of data in the “Unit pack size (ml/g)” attribute were revealed, i.e., incorrectly typed values that ultimately could lead to skewed results in “Price/1gr”. These were mostly evident with mozzarella products - for which it wasn’t shown the value of its net drained weight, but rather the total net weigh of the products - as well as products consisting of several small units of cheese – for which it was only present the weight of one unit, rather than the total weight of the product. When identified, such values were manually corrected.

Also in light of experience of data quality and data exploration, several records have also been removed from the analysis:

- Records without any nutritional values or, in turn that presented the text “Not indicated on pack” in the “Nutrition” attribute.
- Records that only had nutritional values for categories of nutrients that wouldn’t be considered for future analysis (e.g., per 100g: Calcium 570mg (71% RDA), Vitamin B12 1µg (100% RDA));
- Records that didn’t present, either in “Nutrition” or in Mintel’s database, specifically the quantity to which the nutritional values were referent to (e.g., per serving: energy 835kJ/201kcal);
- Records representing products with (at least) more than one type of cheese, which could lead to skewed results, both with regard to nutritional values of these type of cheese assortments (such as fat, salt content, or Kcal), as well as the values of the prices of these products. Note that this type of “combined” products could also include other elements rather than only a different type of cheese, as it’s the case of “Cheese Spread & Breadsticks” or “Cooked Ham Wrapped around Cheddar Cheese”.
- Records that presented all their nutritional values in terms of %RDA (e.g., per 100g: fat (33% RDA), protein (23% RDA), carbohydrate (2.5% RDA)) – Since these values of Recommended Dietary Allowance (RDA) vary according to different criteria such as age, gender, and physical activity, in such cases, it was opted to remove these records from the dataset. In turn, for the records that contained only some of their nutritional values in terms of %RDA (since not all), it was chosen to treat them as a missing value.

Additional transformations were made, which included removing “Date Published”, “Price in euros”, “Unit pack size (ml/g)”, “Packaging Units”, “Nutrition” and “Quantity (g)” from the dataset – these attributes no longer contribute to future analysis.

The final dataset, containing 9475 records (equivalent to 9475 cheese product launches) along with 19 attributes (Appendix 3), was then loaded into the IBM SPSS Statistics software.

However, the majority of these attributes hadn't yet been properly explored.

As so, in order to get familiar with that data, discover first insights, and/or detect interesting subsets to form hypotheses for hidden information, for each one of those attributes, descriptive statistics were used.

To simplify this analysis, the (nominal-level) attributes "Country", "Subcategory", "Storage", "Package Type", "Launch Type", and "Private Label" were firstly coded as presented in Appendix 4.

For these type of attributes, frequencies were used. From here it was possible to conclude that the attribute "Storage" wasn't sufficiently discriminative. For that reason, it was decided to not include this attribute in further analysis, and thereby to exclude it from the final dataset.

On the other hand, for the attributes of numeric type - all the nutritional ones "Price/1gr" – mean, maximum, minimum, range, standard deviation, variance, mode, and percentiles were computed. These results are summarized in

Table 2.

	Price/1gr	Energy/1gr	Protein/1gr	Fat/1gr	Saturated Fat/1gr	Salt/1gr	Carbohydrate/1gr	Sugar/1gr
Mean	0,012	2,931	0,187	0,232	0,152	0,015	0,021	0,016
Maximum	0,130	5,960	0,575	0,540	0,350	0,088	0,540	0,490
Minimum	0,001	0,320	0,000	0,000	0,000	0,000	0,000	0,000
Range	0,129	5,640	0,575	0,540	0,350	0,088	0,540	0,490
Standard Deviation	0,007	0,914	0,080	0,090	0,063	0,009	0,034	0,029
Variance	0,000	0,835	0,006	0,008	0,004	0,000	0,001	0,001
Mode	0,010	3,920	0,250	0,280	0,160	0,015	0,000	0,000
Percentiles 25	0,007	2,430	0,124	0,180	0,118	0,009	0,001	0,000
50	0,011	3,040	0,190	0,250	0,163	0,015	0,010	0,005
75	0,015	3,640	0,250	0,290	0,200	0,019	0,030	0,027

Table 2- Central Tendency for Numerical Attributes

Faced with the high dispersion of values of the nutritional attributes - from which relevant information and/or conclusions could not be extracted - it was after decided to divide the dataset by each one (of the five) product's subcategory. Histograms were created as presented in Appendix 5, 6, 7, 8, 9, 10 and 11, which allowed to particularly understand how the nutritional attributes were distributed, depending on the cheese's subcategory in which they were devised. Note that it is perfectly reasonable that certain type of cheeses present higher levels of fat or salt, for example, when comparing to others. Additionally, the repetition of this analysis allowed to assess if such dispersion of values was due to cheese's subcategory, or to the presence of noise (outliers).

5. Conclusions

This dissertation intends to assess how cluster organisations may benefit from the application of data mining on the type of market intelligence services they provide. In order to reach this goal, an empirical application was developed, which explored the particular case of a Portuguese cluster organisation – PortugalFoods. Through case study methodology, the following main research question was then addressed: "How data mining influences the market intelligence services of cluster organisations?".

This empirical study had its particular focus on PortugalFoods' Knowledge Division department, as well as on the market intelligence services by them provided. As producers and disseminators of knowledge, the delivery of such innovative services is mainly accomplished through the production of three different types of reports: periodicals, support, and tailored.
To collect the privileged information there contained, PortugalFoods relies on two different external sources of data – Innova Market Insights and Mintel GNPD Analysis. In this study, Mintel's databases were chosen for the implementation of a prototype data mining project (supported on CRISP-DM methodology), which enabled to analyse how could their market intelligence services be enhanced through the application of data mining.

On one hand, data mining would optimize the use that PortugalFoods has on Mintel's databases, and increase the amount of value-added information gathered within the cluster organisation. This was particularly evident when considering the product's nutritional values, data that have not yet been tapped, since it provides their members with new insights of how the market is reacting to the launch of certain type of products. Such information could then complement the already available on their reports, thereby enhancing PortugalFoods' market intelligence services.

On the other hand, considering how Mintel's databases present their data in the excel format, further replications of this prototype data mining project would turn out to be a major time-consuming task. This would happen due to the existence of the following data's problems: i) the incorrect introduction of values; ii) the inexistence of a standard terminology for all the nutrient constituents; iii) the inexistence of a standard level of measurement for those same attributes;
The sequence of steps and actions taken throughout this DM project provides PortugalFoods with a standard guideline to how to overcome such problems.

In sum, data mining could be an effective support tool in the market intelligence services of COs by providing a broader picture (to and of) their members' existing markets, customers, and competitors, as well as to their market growth potential for new products and services. Thus, by augmenting the amount of specific information available in these COs' tailor-made services, DM may improve their members' projections of the state of the industry, enhance their confident decision-making in corporate strategy areas like market opportunity, market penetration strategy, and market development, and ultimately strength their competitiveness - particularly relevant in the midst of rapidly globalizing competition.

5.1 Limitations and suggestions for future research

Given the extant literature devoted to DM in this dissertation, it would be expected for this project to evolve distinctive DM tasks to solve the business problem. Yet, considering the extension of data's problems and due to time constraints, this project only applied summarization techniques. Aiming to identify and minimize the occurrence of such problems, and therefore, to obtain the most accurate results possible, "data verify quality" task was often repeated along this project. Though, unless each observation was both manually and individually analysed it is not possible to ensure that all data is correctly entered, and therefore, to guarantee the results' reliability.

The application of other DM tasks was also restricted by the attributes of Mintel's databases: for example, since it does not include the sales values or quantities sold, prediction was not considered for analysis. Likewise, classification was not applied, as Mintel already assign records to a predefined set of class labels. In this study, this products' classification is reflected in Appendix 4.

The method found as the most appropriate to develop this dissertation was case study. Nevertheless, case study research strategy has always been criticized for its lack of rigor as a research tool as well as by the subjectivity and biased interpretation of the data that the researcher tends to have (Yin, 2009). Besides, the drawback of a single-case design is its inability to provide a generalizing conclusion, particularly if the events are rare. The question commonly raised is "How can you generalize from a single case?" (Yin, 1984, p. 21).

A suggestion for future research is to explore how COs' services are being perceived by their members. It would also be of great interest to analyse the impact that market intelligence services have on the definition of their strategical action plan, in terms of their internationalization process for example.

References

Apte, C., Liu, B., Ednault, E. P. & Smyth, P. 2002. Business Applications of Data Mining. ***Communications of the ACM***, 45: 49-53.

Baptista, R. & Swann, P. 1998. Do firms in clusters innovate more? ***Research Policy*,** 27: 525-540.

Benbasat, I., K.Goldstein, D. & Mead, M. 1987. The Case Research Strategy in Studies of Information Systems. ***MIS Quarterly***, 11: 369-386.

Berry, M. J. & Linoff, G. S. 2000. ***Mastering Data Mining***. New York: John Wiley & Sons, Inc.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. 2000. ***CRISP-DM 1.0: Step-by-step data mining guide***.

Chen, S. Y. & Liu, X. 2004. The Contribution of Data Mining in Information Science. ***Journal of Information Science***, 1-19.

Cios, K. J. & Kurgan, L. A. 2005. Trends in Data Mining and Knowledge Discovery. In N. R. Pal, & L. Jain, ***Advanced Techniques in Data Mining and Knowledge Discovery:*** 1-26. London: Springer.

Coletti, M. 2010. Technology and industrial clusters: how different are they to manage? ***Science and Public Policy*,** 37: 679–688.

Coutinho, C. P. & Chaves, J. H. 2002. O estudo de caso na investigação em Tecnologia Educativa em Portugal. ***Revista Portuguesa de Educação***, 221-243.

Delgado, M., Porter, M. E., & Stern, S. 2010. Clusters and Entrepreneurship. ***Journal of Economic Geography*,** 10: 495-518.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. 1996. The KDD Process for Extracting Useful Knowledge from Volumes of Data. ***Communications of the ACM,*** *39*: 27-34.

Gama, J., Carvalho, A. P., Faceli, K., Lorena, A. C. & Oliveira, M. 2012. ***Extração de Conhecimento de Dados - Data Mining***. Lisboa: Edições Sílabo, Lda.

Ganne, B. & Lecler, Y. 2009. From Industrial Districts to Poles of Competitiveness. In B. Ganne, & Y. Lecler, ***Asian Industrial Clusters, Global Competitiveness and New Policy Initiatives:*** 3-24. Singapore: World Scientific Publishing.

Glaeser, A. 2013. The Role of Cluster Organisations in the Construction of Collaborative R&D Projects: The Case of the ICT & Health Clusters in the Paris Region. *35th DRUID Celebration Conferecence*, (pp. 1-25). Barcelona.

Han, J. & Kamber, M. 2001. ***Data Mining: Concepts and Techniques***. San Francisco: Morgan Kaufmann Publishers.

Hand, D. J. 1998. Data Mining: Statistics and More? ***The American Statistician,*** *52*: 112-118.

Ingstrup, M. B. & Damgaard, T. 2011. Cluster facilitation in a cluster life cycle perspective. *IMP 2011 Conference at University of Strathclyde*, (pp. 1-20). UK.

Ketels, C., Lindqvist, G. & Sölvell, Ö. 2012. *Strengthening Clusters and Competitiveness in Europe: The Role of Cluster Organisations.* Stockholm: Center for Strategy and Competitiveness.

Lämmer-Gamp, T., Köcker, G. M. & Nerger, M. 2014. *Cluster Collaboration and Business Support Tools to Facilitate Entrepreneurship, Crosssectoral Collaboration and Growth.* European Comission, European Cluster Observatory.

Langen, P. W. 2002. Clustering and performance: the case of maritime clustering in The Netherlands. ***Maritime Policy & Management*** , 209-221.

Lindqvist, G. 2009. Disentangling Clusters: Agglomeration and Proximity Effects. ***Cluster organisations: activities and performance:*** 255-278. Stockholm: The Economic Research Institute.

Lindqvist, G., Ketels, C. & Sölvell, Ö. 2013. ***The Cluster Initiative Greenbook 2.0***. Stockholm: Ivory Tower Publishers.

Marshall, A. 1890. ***Principles of Economics***. London: Macmillan and Co., Ltd.

Maskell, P. 2001. Towards a Knowledge-based Theory of the Geographical Cluster. ***Oxford University Press***, 921-943.

N., G., & Srivatsa, S. 2006. A Research Study: Using Data Mining in Knowledge Base Business Strategies. ***Information Technology Journal***, 590-600.

Pal, N. R. & Jain, L. 2005. ***Advanced Techniques in Data Mining and Knowledge Discovery***. London: Springer.

Porter, M. E. 1990. The Competitive Advantage of Nations. ***Harvard Business Review***, 73-91.

Porter, M. E. 1998. Clusters and the New Economics of Competition. ***Harvard Business Review***, 77-90.

Porter, M. E. 2003. The Economic Performance of Regions. *Regional Studies Association*, 549-578.

Santos, M. F. & Azevedo, C. S. 2005. ***Data Mining - Descoberta de Conhecimento em Bases de Dados***. Portugal: FCA.

Saunders, M., Lewis, P. & Thornhill, A. 2009. ***Research methods for business students***. England: Pearson Education Limited.

Saxenian, A. 1996. Inside-Out: Regional Networks and Industrial Adaptation in Silicon Valley and Route 128. ***Cityscape: A Journal of Policy Development and Research***, 41-60.

Sölvell, Ö. & Williams, M. 2013. *Building the Cluster Commons - An Evaluation of 12 Cluster Organizations in Sweden 2005 - 2012*. Stockholm: Ivory Tower Publishers.

Sölvell, Ö., Lindqvist, G. & Ketels, C. 2003. ***The Cluster Initiative Greenbook .*** Stockholm: Ivory Tower Publishers.

Wirth, R. & Hipp, J. 2000. CRISP-DM: Towards a Standard Process Modell for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 29-39.

Witten, I. H., Frank, E. & Hall, M. A. 2011. ***Data Mining: Practical Machine Learning Tools and Techniques***. USA: Morgan Kaufmann Publishers.

Yin, R. K. 2009. ***Case Study Research: Design and Methods***. Thousand Oaks, California: SAGE Publications, Inc.

Appendix

Appendix 1: Subcategories' description according to Mintel's glossary

Subcategory	
Curd & Quark	Curd is obtained by curdling milk with rennet or an edible acidic substance, such as lemon juice or vinegar. Milk that is left to sour will naturally produce curds. Quark is soft, white and not aged. Cottage cheese and paneer are also categorized here. This sub-category includes all plain and flavoured curd and quark products, apart from sweet quark desserts, which should be categorized under the Soft Cheese Desserts sub-category.
Fresh Cheese & Cream Cheese	Fresh cheese refers to cheese that has not been ripened or aged but may be slightly cured. These cheeses have a high moisture content and a soft texture. Fromage frais, fromage blanc, queso blanco, mascarpone and cream cheese are all examples of fresh cheese. Also includes sour cream based spreads. This sub-category includes all plain and flavoured fresh cheese products, apart from sweet flavoured fresh cheese based desserts, which should be categorised under the *Soft Cheese Desserts* sub-category. Requeijão is considered processed and should be categorised under *Processed Cheese*.
Hard Cheese & Semi-Hard Cheese	This sub-category includes all plain and flavoured hard and semi-hard cheeses. These cheeses may also be referred to as firm. Gouda, cheddar, emmental and parmesan are all examples of hard and semi-hard cheeses.
Processed Cheese	Processed cheese consists of natural cheese(s) with additives such as salt, emulsifiers, stabilizers, flavour enhancers and food colourings. These products feature a consistent texture. This sub-category includes all plain and flavoured processed cheese. Many forms are available, including: cheese spread; cheese dip; cheese food; requeijão and spray cheese. Cheese alternatives, for example made from rice are categorized here. Processed cheese is often available in single serve slices or triangular portions, but is also available in blocks.
Soft Cheese & Semi-Soft Cheese	Soft and semi-soft cheeses are aged for a short time and feature a very soft texture. This sub-category includes all plain and flavoured soft and semi-soft cheeses. Brie and camembert are types of soft cheese. Blue cheeses are semi-soft. Also included in this sub-category are soft and semi-soft pasta filata cheeses, such as mozzarella. Pasta filata cheeses are cooked and kneaded, or "spun".

Appendix 2: Attributes initially obtained from Mintel's databases

Original Attributes

- Record ID
- Product
- Product variant
- Brand
- Company
- Ultimate Company
- Country
- Date Published
- Category
- Subcategory
- Price in US dollars
- Price in euros
- Positioning claims
- Storage
- Unit pack size (ml/g)
- Packaging units
- Package type
- Package material
- Nutrition
- Number of variants
- Product description
- Launch Type
- Company county/state
- Private label
- Currency
- Price in local currency
- Bar code
- Production code
- Flavours
- Primary Image Thumbnail
- Record hyperlink
- Ingredient 1
- Ingredient 2
- Ingredient 3
- Ingredient 4
- Ingredient 5
- Ingredient 6
- Ingredient 7
- Ingredient 8
- Ingredient 9
- Remaining Ingredients

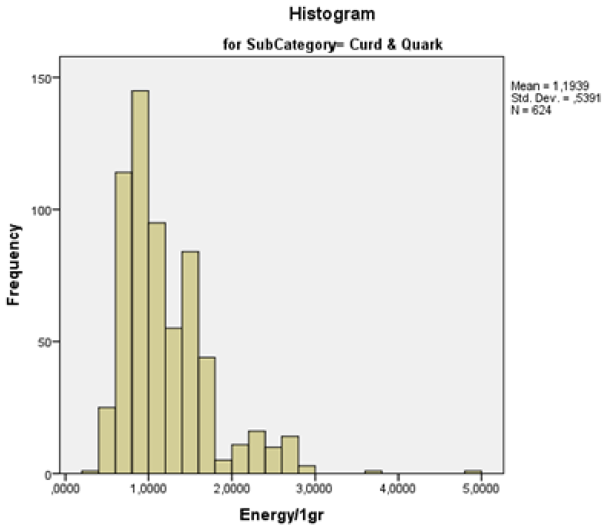
Appendix 3: Attributes in the final dataset

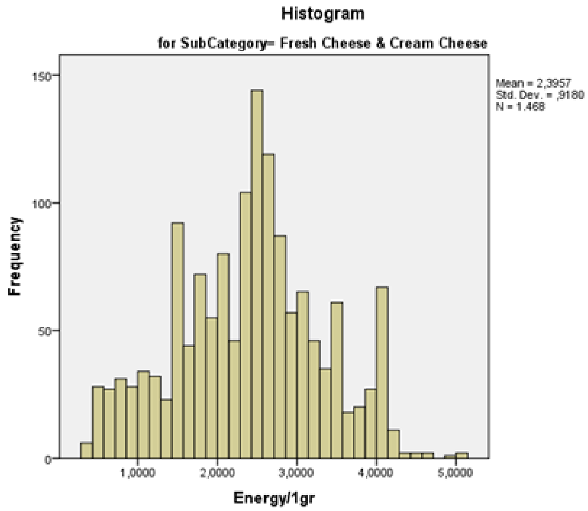
Final Attributes
• Record ID • Product • Country • Date Published • Subcategory • Price/1gr • Positioning claims • Storage • Package type • Energy/1gr • Protein/1gr • Fat/1gr • Saturated Fat/1gr • Salt/1gr • Carbohydrate/1gr • Sugar/1gr • Launch Type • Private label • Flavours

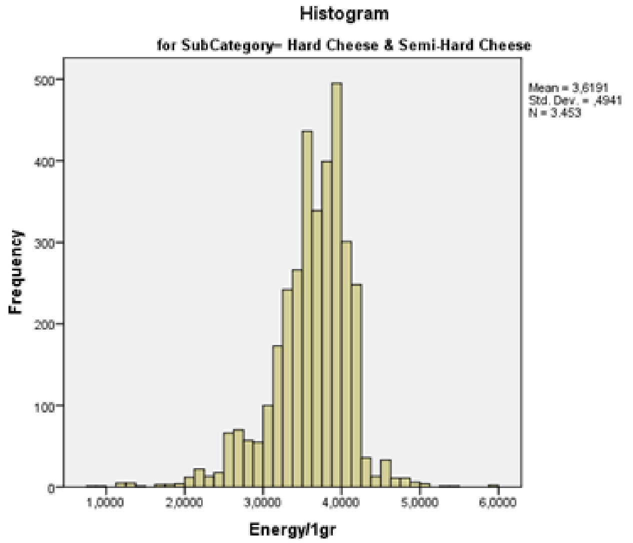
Appendix 4: Codebook for the nominal-level attributes

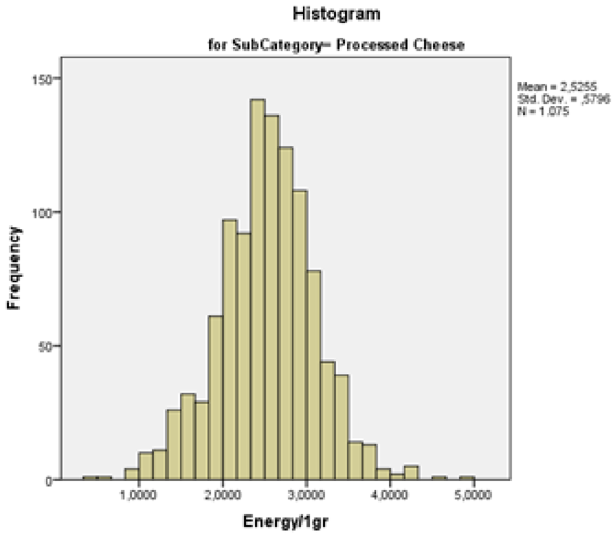
	Attributes					
	Country	**Sub-Category**	**Storage**	**Package Type**	**Launch Type**	**Private Label**
1	Austria	Curd & Quark	Chilled	Aerosol	New Formulation	Private Label
2	Belgium	Fresh Cheese & Cream Cheese	Frozen	Blister Pack	New Packaging	Branded
3	Czech Republic	Hard Cheese & Semi-Hard Cheese	Shelf stable	Bottle	New Product	
4	Denmark	Processed Cheese		Can	New Variety/Range Extension	
5	Finland	Soft Cheese & Semi-Soft Cheese		Carton	Relaunch	
6	France			Clam-pack		
7	Germany			Composite		
8	Greece			Flexible		
9	Hungary			Flexible sachet		
10	Ireland			Flexible stand-up pouch		
11	Italy			Jar		
12	Netherlands			Miscellaneous		
13	Norway			Rigid box		
14	Poland			Skinpack		
15	Portugal			Sleeve		
16	Spain			Tottle		
17	Sweden			Tray		
18	Switzerland			Tub		
19	Turkey			Tube		
20	UK					
21	Ukraine					

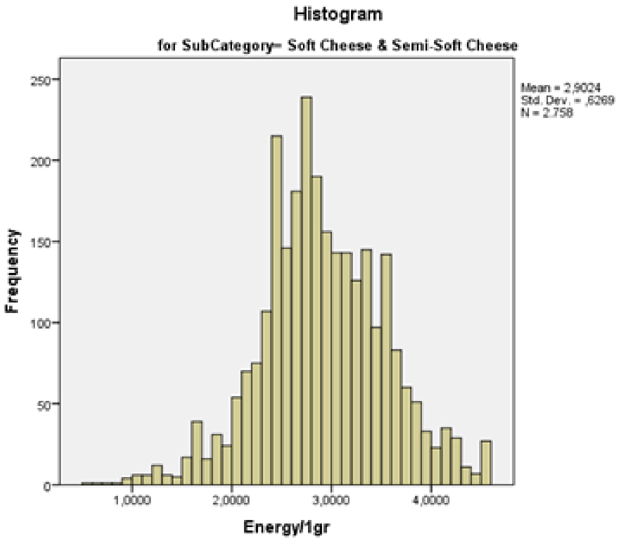
Appendix 5 - “Energy/1gr” histograms by cheese subcategory



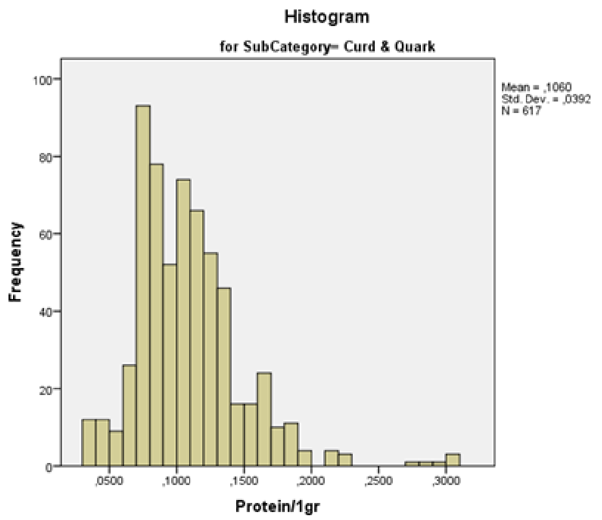


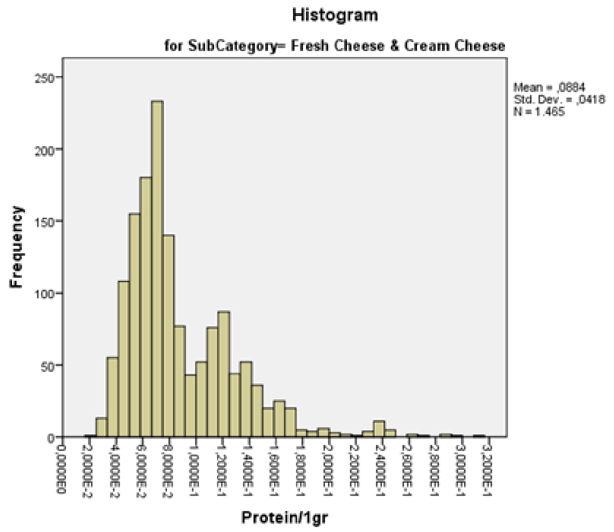


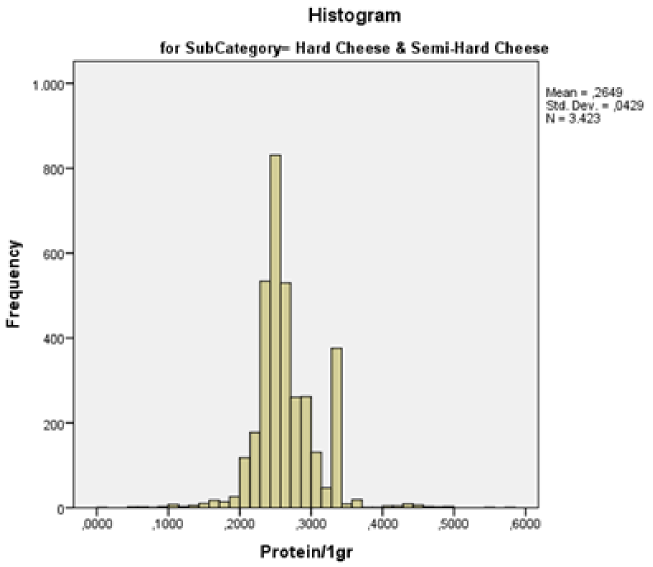


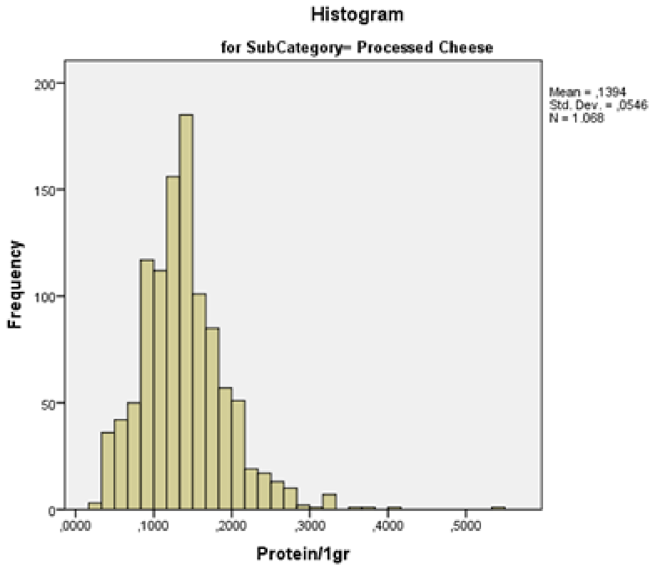


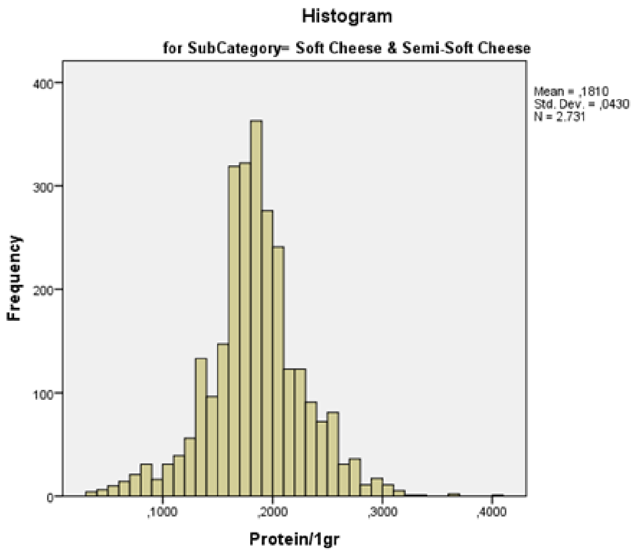
Appendix 6: "Protein/1gr" histograms by cheese subcategory



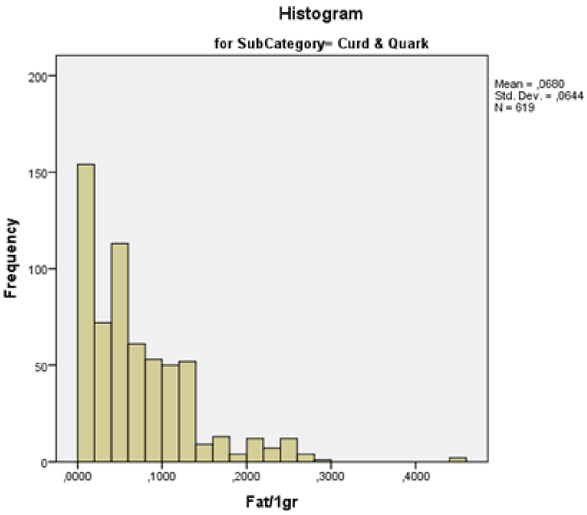


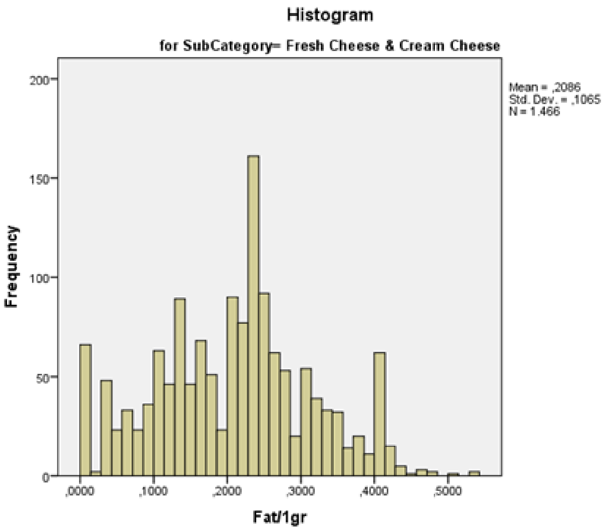


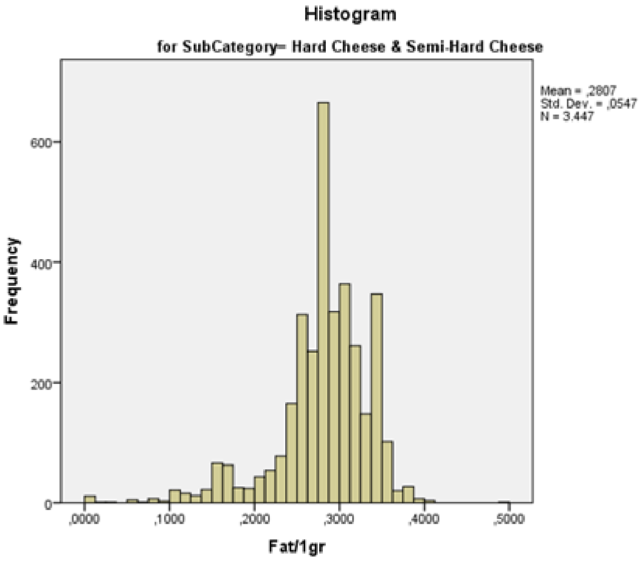


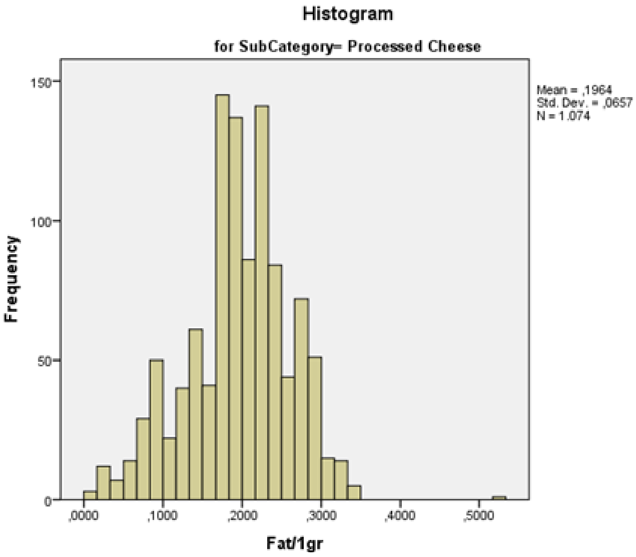


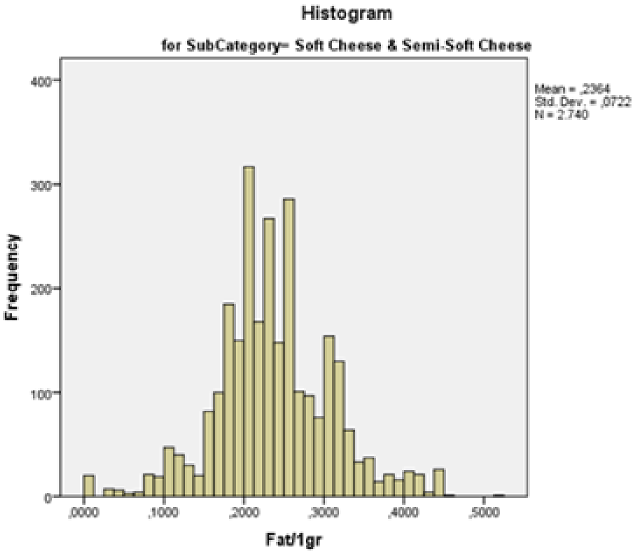
Appendix 7: "Fat/1gr" histograms by cheese subcategory



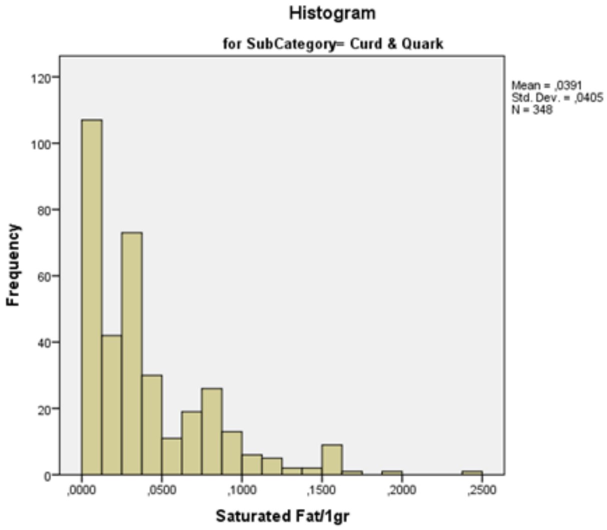


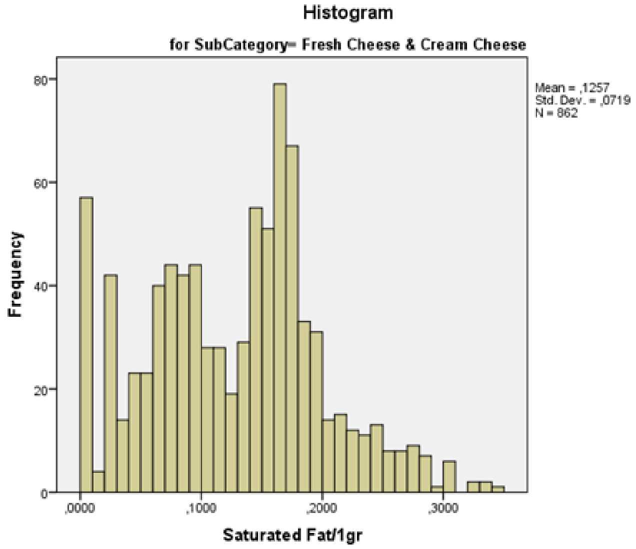


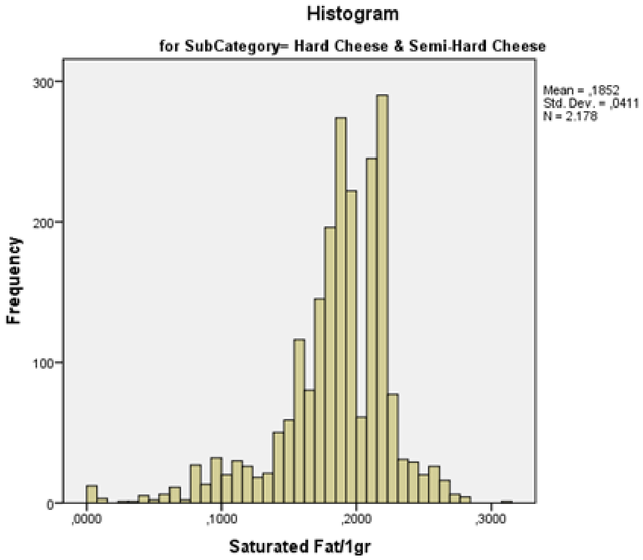


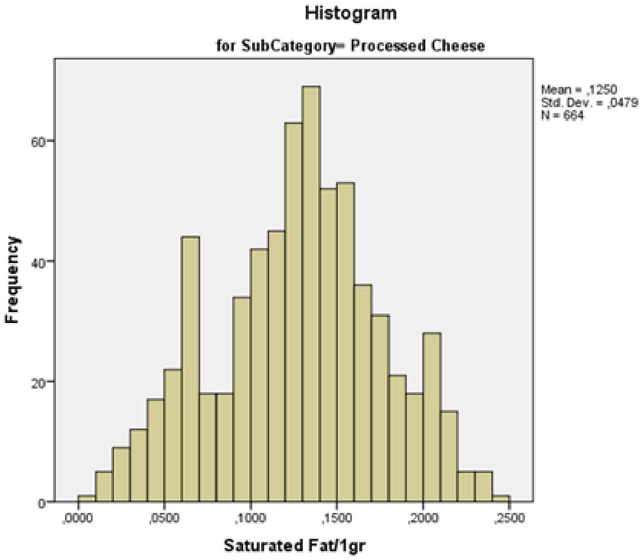


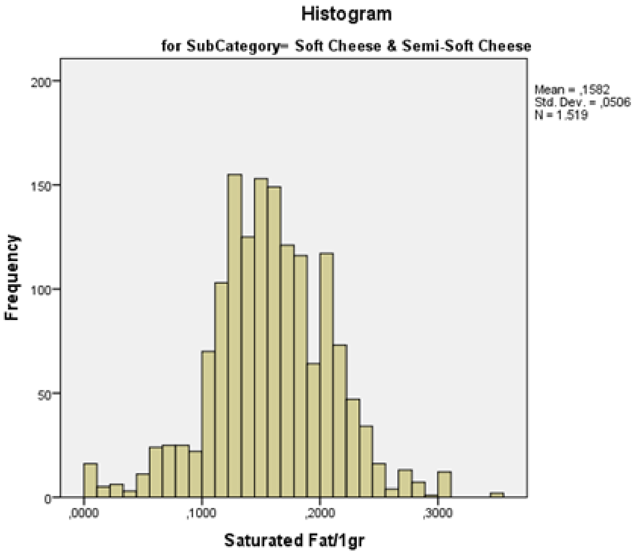
Appendix 8: "Saturated Fat/1gr" histograms by cheese subcategory



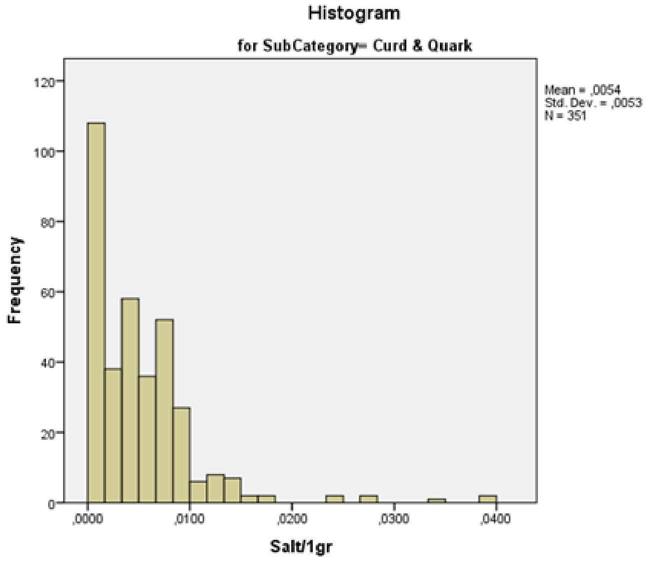


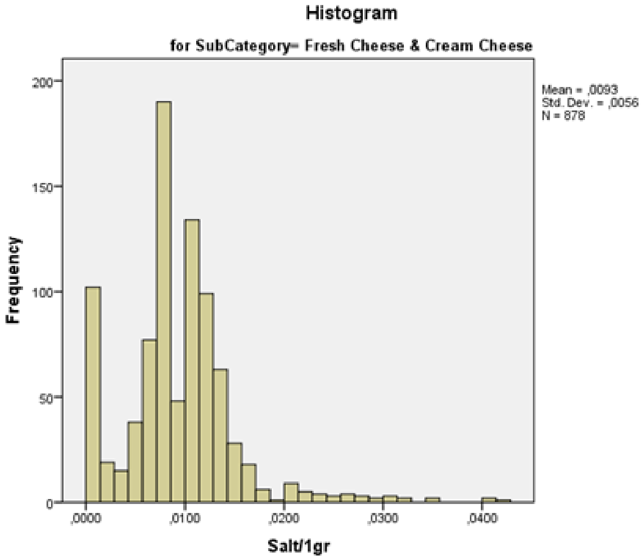


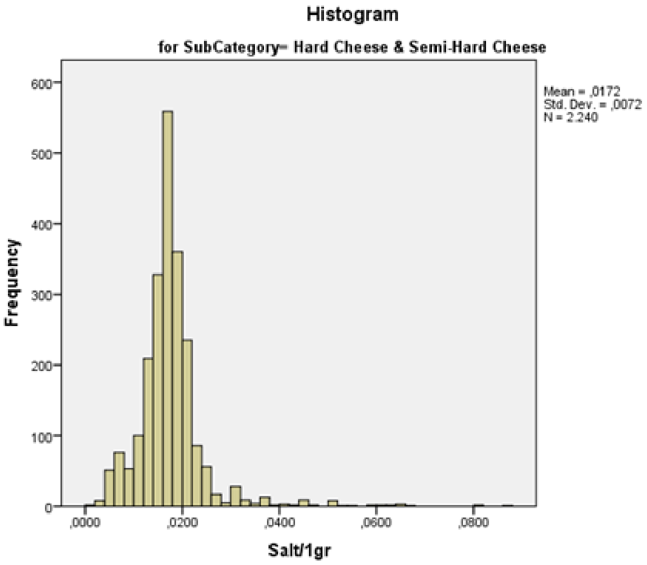


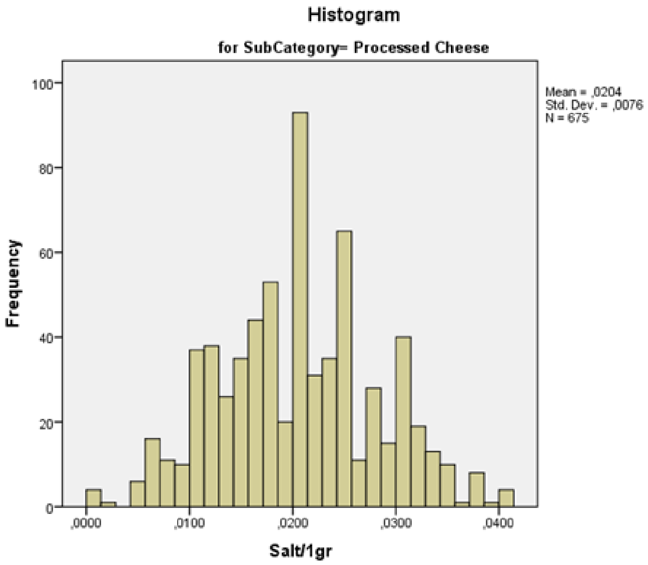


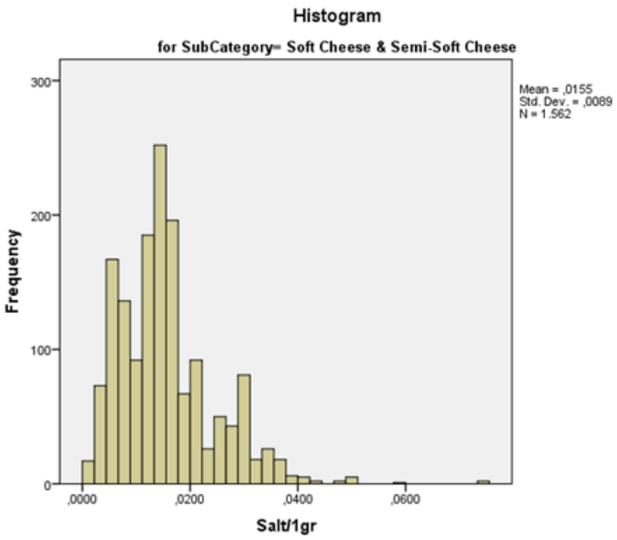
Appendix 9: “Salt/1gr” histograms by cheese subcategory



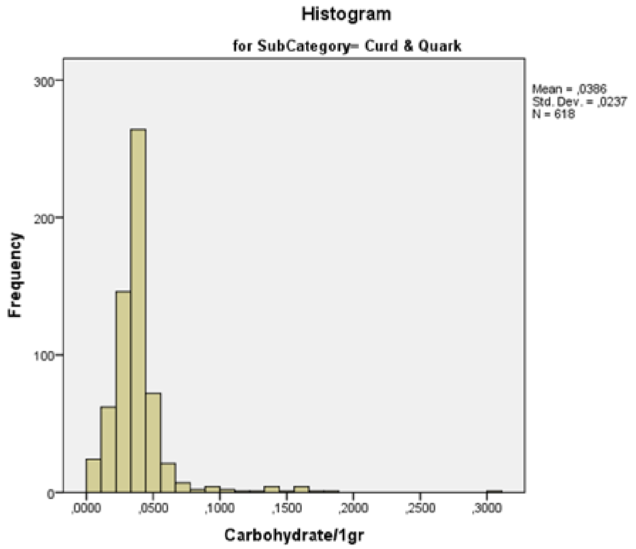


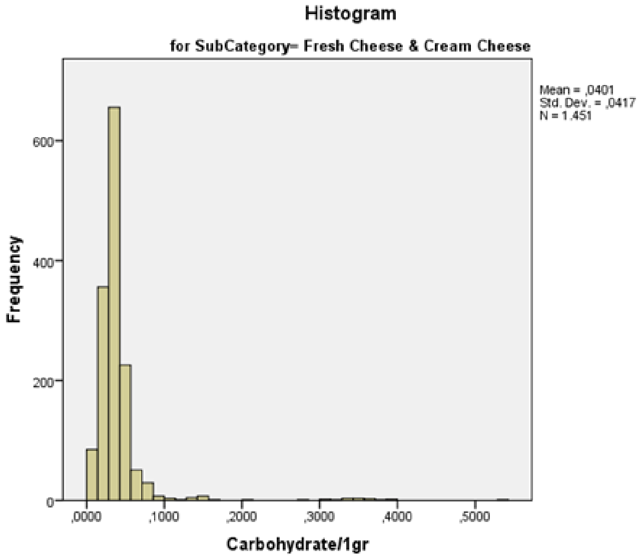


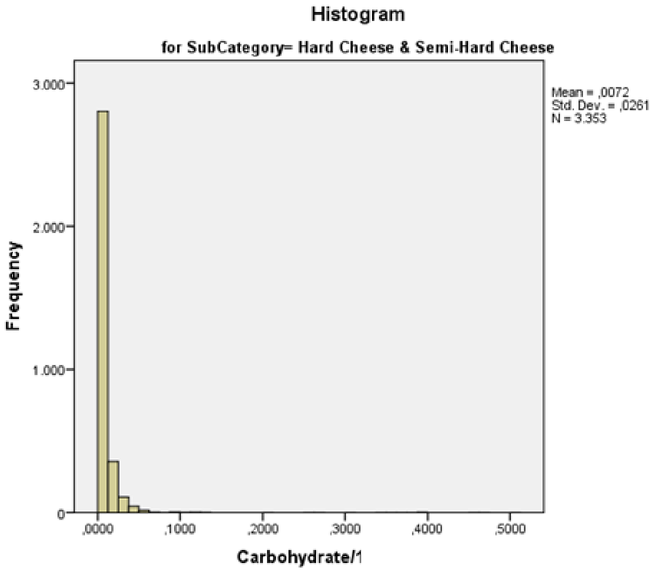


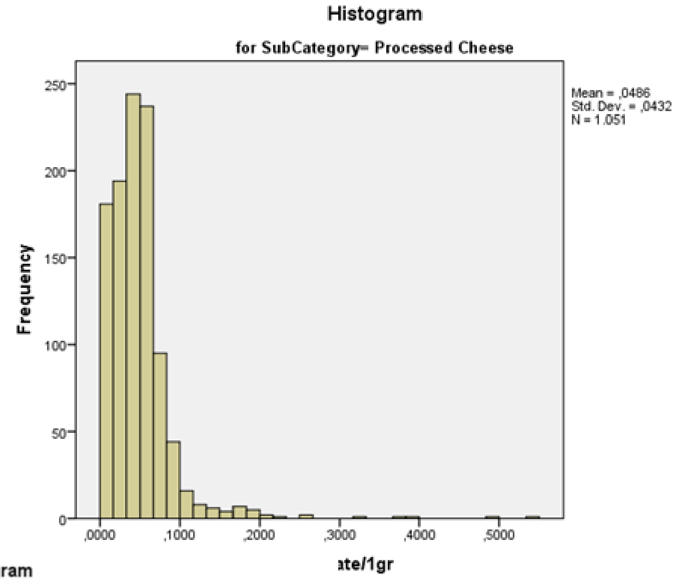


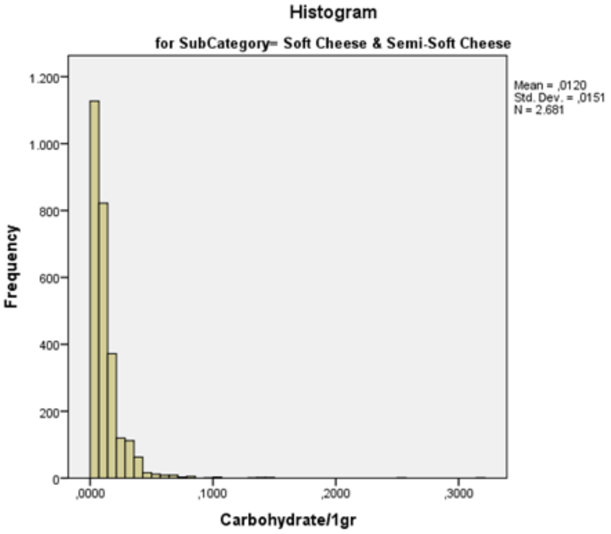
Appendix 10: “Carbohydrate/1gr” histograms by cheese subcategory



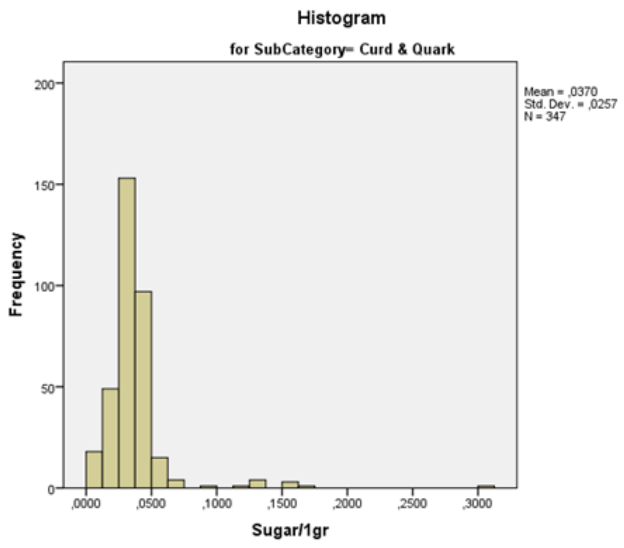


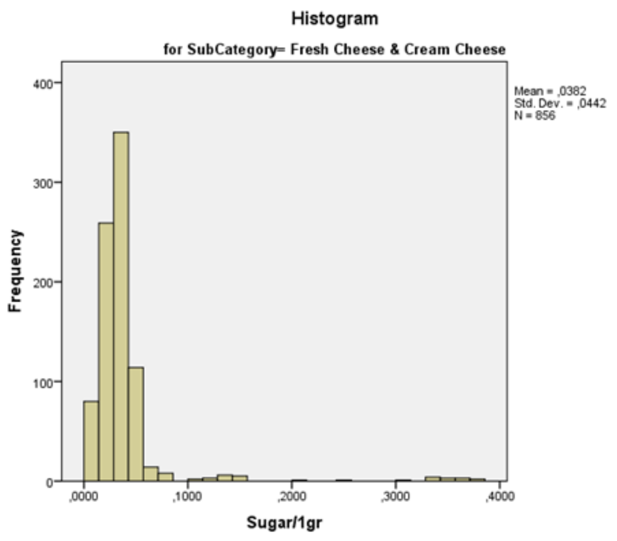


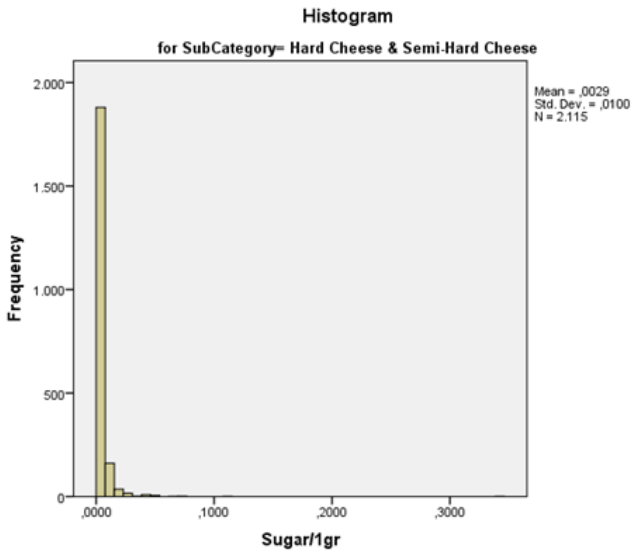


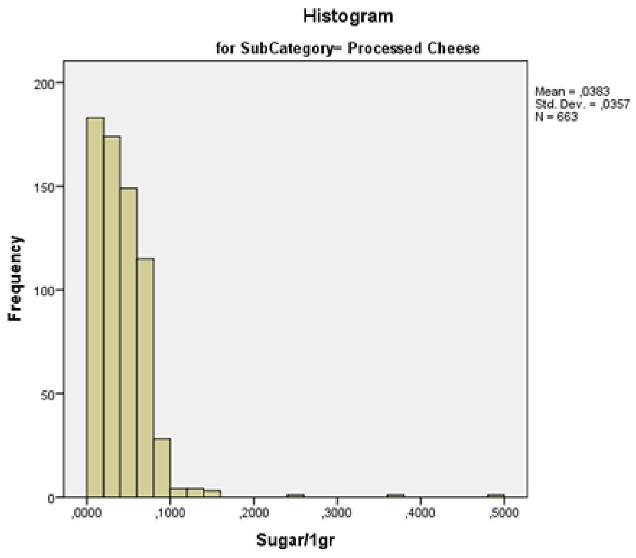


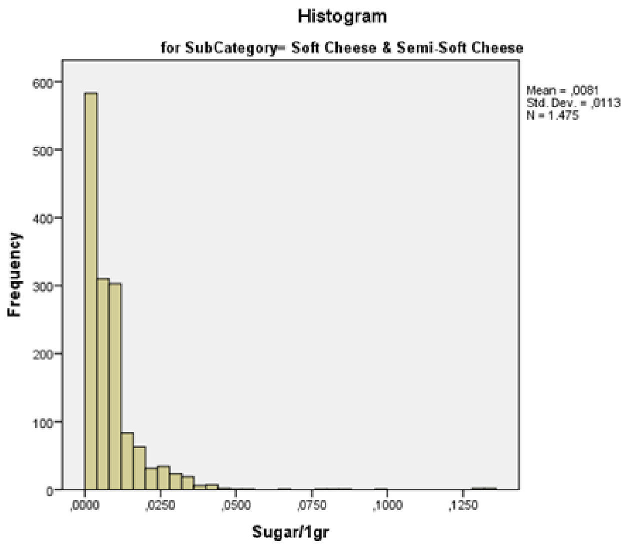
Appendix 11: “Sugar/1gr” histograms by cheese subcategory


Histogram
for SubCategory= Curd & Quark
Mean = ,0370
Std. Dev. = ,0257
N = 347
Frequency
Sugar/1gr


Histogram
for SubCategory= Fresh Cheese & Cream Cheese
Mean = ,0382
Std. Dev. = ,0442
N = 856
Frequency
Sugar/1gr


Histogram
for SubCategory= Hard Cheese & Semi-Hard Cheese
Mean = ,0029
Std. Dev. = ,0100
N = 2.115
Frequency
Sugar/1gr


Histogram
for SubCategory= Processed Cheese
Mean = ,0383
Std. Dev. = ,0357
N = 663
Frequency
Sugar/1gr


Histogram
for SubCategory= Soft Cheese & Semi-Soft Cheese
Mean = ,0081
Std. Dev. = ,0113
N = 1.475
Frequency
Sugar/1gr