• U C •

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

DEPARTAMENTO DE
ENGENHARIA MECÂNICA

Knowledge elicitation by merging heterogeneous data sources in a die-casting process

Submitted in Partial Fulfilment of the Requirements for the Degree of Master in Engineering and Industrial Management.

Autor

João Morgado

Orientador

Professor Doutor Pedro Mariano Simões Neto

Júri

Presidente	**Professor Doutor Cristóvão Silva** **Professor Auxiliar da Universidade de Coimbra**
Vogais	**Professor Doutor Nuno Alberto Marques Mendes** **Investigador Auxiliar da Universidade de Coimbra** **Professor Doutor Pedro Mariano Simões Neto** **Professor Auxiliar da Universidade de Coimbra**

Institutional Collaboration

University *of Ljubljana*
Faculty *of Mechanical Engineering*

University of Ljubljana

Coimbra, Setembro, 2015

Acknowledgements

The accomplishment of this master thesis was a great experience for me, not only for being able to work in a field of manufacturing process, but also because adapting to a new workplace in another country with a different culture and in another language represents a great challenge.

I would like to thank my mentor Prof. Dr. Peter Butala the opportunity he has given me to do this project and the activities in which I had the chance to participate which suppose an unexpected large learning about this field. I also want to express my gratitude to somentor Rok Vrabič for their interest and sacrifice for helping me and solving all the doubts i had during the achievement of this thesis.

Finally, I want to express my eternal gratitude to my parents, without them this experience would not have been possible. They have supported me to fulfil this project abroad and they have also covered most of the expenses here. I am also grateful to my partners from all parts of Slovenia for their help, support and patience. Because of this I want to dedicate this thesis to all of them.

Abstract

In order to establish adaptive control of a manufacturing process knowledge must be acquired about both, the process and its environment. This knowledge can be obtained by mining large amounts of data collected through the monitoring of the manufacturing process. This enables the study of process parameters and the correlations between the process parameters and with the parameters of the environment. Through this, knowledge about the process and its relation to the environment can be established and, in turn, used for adaptive process control. The aim of this thesis is to study real manufacturing data, obtained through monitoring of a die casting process. First, in order to better understand the problem at hand, a literature review of using Big data and merging data from heterogeneous sources is given. Second, using the real data, a robust algorithm to asses the quality rate was developed, due to data being incomplete and noisy. Merging the process and the environment data was done. In this way it is possible to visualize the influences of various parameters on quality rate and make suggestions for improvement of the die casting process.

Keywords Manufacturing systems, Adaptive process control, Data mining, Knowledge discovery, Big data analytics.

Contents

LIST OF FIGURES

LIST OF TABLES

Acronyms

KDD- Knowledge Discovery Database

DM- Data Mining

IBM- Institute for Business Value

XML- Extensible Markup Language

SME´s- Small and Medium-Sized Enterprises

KNN- K-Nearest Neighbour

SVM- Support Vector Machines

GDBT- Gradient Boosted Decision Trees

JDBC- Java Database Connectivity

SQL- Structured Query Language

QR- Quality Rate

NP- Hard Problems

MARS- Multivariate Adaptive Regression Splines

LOESS- Locally Estimated Scatter Plot Smoothing

LVQ- Learning Vector Quantization

SOM- Self-organizing Map

LASSO- Least Absolute Shrinkage and Selection Operator

CART- Classification and Regression Tree

ID3- Iterative Dichotomiser 3

CHAID- Chi-squared Automatic Interaction Detection

GBM- Gradient Boosting Machines

SVM- Support Vector Machines

RBF- Radial Basis Function

LDA- Linear Discriminant Analysis

SOM- Self-organizing Map

VLQ- Learning Vector Quantization

DBN- Deep Belief Networks

1. INTRODUCTION

1.1. Background

Since markets are shifting, economies are affected by crises, and competition is increasing at a global level, manufacturing companies are facing new challenges in dealing with the changes of the turbulent environment where new technologies are continuously emerging [1].

In today´s competitive business environment, the rapidly changing conditions, needs, and current opportunities in the global market are pushing manufacturing companies to adapt themselves to new situations. An important step for the adaptation is the ability to learn, a process which is based on knowledge discovery. By continuously learning, growing, and evolving over time, systems that are able to learn are the only ones that can adapt to the rapid changes happening right now in the manufacturing environment. In the manufacturing environment there are many different sources of knowledge and ways of learning that can be put into practice in order to improve productivity. In the past, traditional data analysis has been improving products and processes. Today, however, the emergence of concepts like Big data in the study of manufacturing data, allow us to extract knowledge which was previously impossible by using this new approach based on computer analysis methods, since humans can´t read large amounts of data and computers are better at analysing large amounts of data.

Knowledge discovery in databases (KDD) is an advanced learning technique. This technique enables learning by data mining (DM), i.e. extracting knowledge from the actual data collected during the operation of a business system and stored in large databases or data warehouses [1].

KDD enables detection of patterns of behaviour and causalities hidden in the data that might represent a valuable source of new specific knowledge, and thus may contribute to a better understanding of the manufacturing process and to enhance management and control of the manufacturing environment. To illustrate the utility of KDD, the concept is used in a high-pressure die-casting domain. A case study based on industry data collected during die-casting operations provides a demonstration of the concept.

1.2. Purpose of the master´s thesis

The aim of this master's thesis is to uncover knowledge by analyzing a large volume of real-time manufacturing data collected during manufacturing operations and to use the knowledge gained to support decision-making and adaptive process control.

This thesis also focuses on the idea of developing a system that enables us to detect patterns of behaviour and causalities in the data in the die-casting process in order to have a better understanding of the manufacturing process, as well as, to develop solutions that can improve management and control of manufacturing systems.

1.3. Problem Statement

The amount of data in our world has been exploding. Five years ago, most companies collected data that were part of their daily transactions with the primarily goal to keep track of operations. Nowadays, companies capture trillions of bytes of information about their customers, suppliers, and operations with the aim of creating and implement big data strategies that will able to give them a competitive advantage in the global market.

Data collected by companies is not actually analysed by anyone, at least not in a deep way, only in a superficial way. This means that there is a need to develop better methods, including engineering methods and engineering tools to provide quick assessments and interpretation of this data.

The main problem is to be able extract valuable knowledge from this large amount of data provided by the company. Correlations that could be found in this vast amount of data are also part of the main problem since it is not possible to know if there are any existing dependences between the parameters, and even if some kind of pattern could be found, it is not possible to infer what was the consequences or what is the causes of the problem only from one chart. So causality cannot be inferred on this, it can only helps in identyfying where to start to looking.

The other problem beginning to appear was that the quality rate was not part of the data set. In other words, no kind of information was available if whether pieces were well made or not, and if they were not well made if it is logged in the database or if there was any kind of alarms or if they were just discarded. It turns out that this is really difficult since large parts of the data can be incomplete or corrupt, thus complicating the analysis and search of valuable knowledge, as also the fact that the available data is almost infinite. That is why a manual approach is used to plot the data and discuss which kind of information is possible to get from it. Yet, generally, there is a lot of potential on other kinds of methods like machine learning, data mining and others in the industry.

1.4. Structure

To achieve the mentioned goals and go through them, this thesis is divided into four parts:

- The first part introduces the reader to the thesis and provides a necessary theoretical basis and understanding the concept of Big Data in manufacturing.
- The second part is an overview of the existing methods to analyse the big data.
- The third part shows the preparation and selection of the data and steps that are followed to discover more knowledge in the process, as well as, the algorithm developed.
- The fourth part focuses on the die casting process and the correlations between the parameters.
- The last part of the thesis gives attention to suggestions and improvements and also this study’s results and conclusions.

2. BIG DATA ON MANUFACTURING

This chapter focuses on the theoretical part based on general definitions of Big Data and the methods used to analyse the Big Data.

2.1. Introduction to Big Data

"Big data"- which admittedly means many things to many people is no longer confined to the realm of technology. Today, given its ability to profoundly affect commerce in the globally integrated economy, it is a business priority. In addition to providing solutions to long-standing business challenges, big data inspires new ways to transform processes, organizations, entire industries and even society itself [2].

While it is ever-present today, nevertheless, "big data" as a concept is being developed and taken to deeper level in manufacturing analysis. The first big data appearances were probably originated in lunch-table conversations at Silicon graphics Inc. in the mid-1990s. In spite of the references to the nineties, Figure 2.1 show that the term became widespread as recently as in 2011. The current hype can be attributed to the promotional initiatives by IBM and other leading technology companies who invested in building the niche analytics market[2].

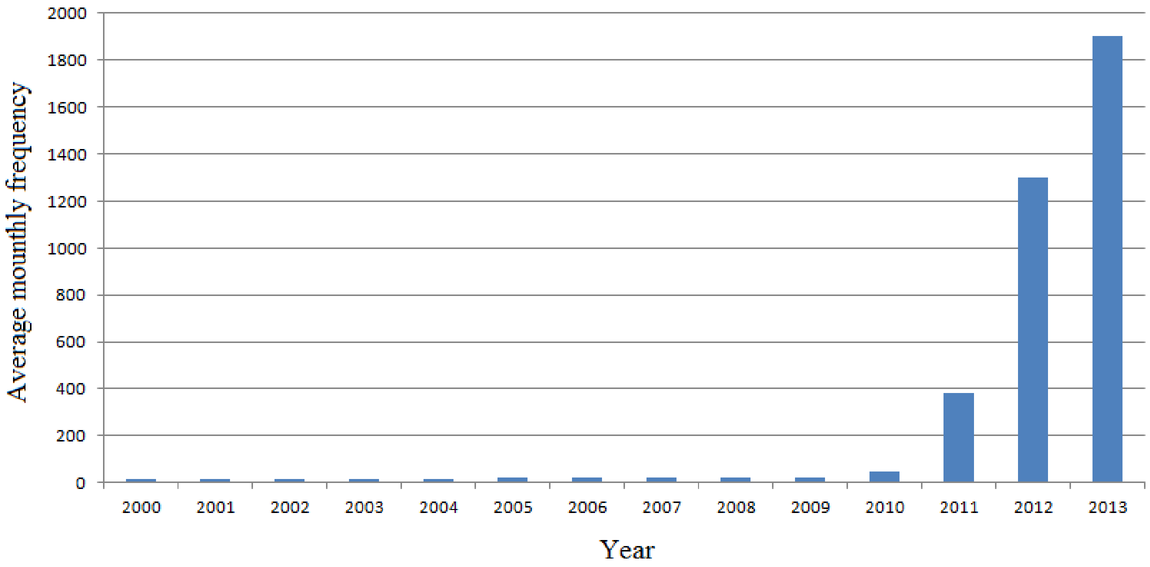


Figure. 2.1 Frequency distribution of documents containing the term "big data" in the ProQuest Research Library[3].

The term “big data” is pervasive, and yet still the notion engenders confusion. Big data has been used to convey all sorts of concepts, including: huge quantities of data, social media, analytics, next generation data management capabilities, real-time data, and much more. Whatever the label, organizations are starting to understand and explore how to process and analyse a vast array of information in new ways. Industries throughout the world and their executives, recognize the need to learn more about how to exploit big data. But despite what seems like unrelenting media attention, it can be hard to find in-depth information on what organizations are really doing[2].

Much of the confusion about big data begins with the definition itself. To understand its definition, this paper focus on a study conducted by the company IBM. This company asked each respondent to select up to two characteristics of big data. Rather than any single characteristic clearly dominating among the choices, respondents were divided in their views on whether big data is best described by today’s greater volume of data, the new types of data and analysis, or the emerging requirements for more real-time information analysis[2].

The Figure 2.2 shows the study made by IBM.

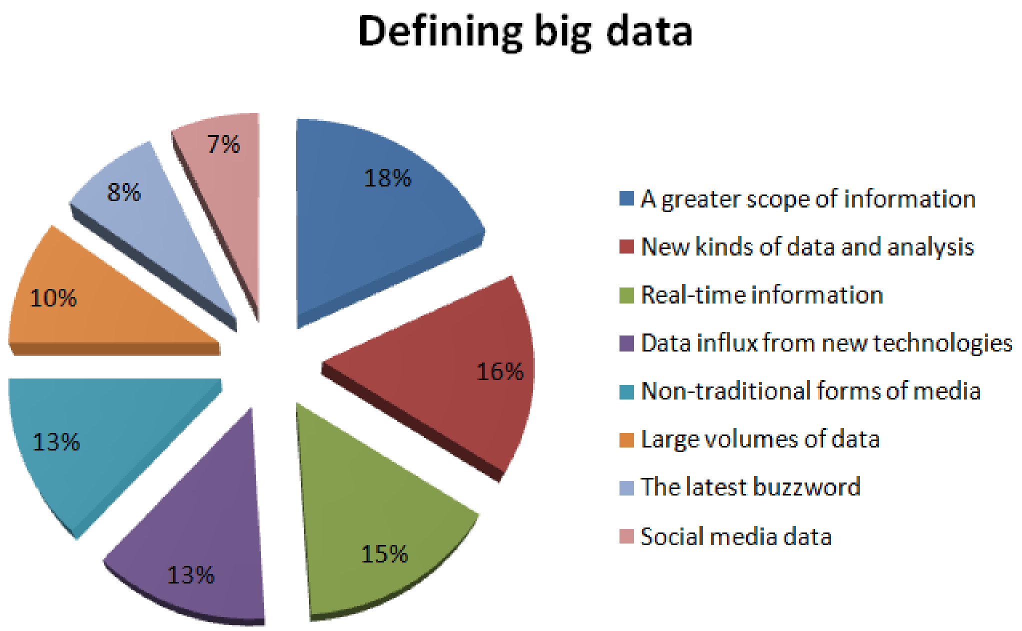


Figure. 2.2 View of the big data by the respondents[2]

Respondents were asked to choose up to two descriptions about how their organizations view big data from the choices above. Choices have been abbreviated, and selections have been normalized to equal 100%. Total respondents=1144 [2].

These results align with a valuable way of characterizing three dimensions of big data - "the three Vs: " volume, variety and velocity". Even thought these are the key attributes of big data, we believe organizations need to consider an important fourth dimension: veracity. Inclusion of veracity as the fourth big data attribute emphasizes the importance of addressing and managing for the uncertainty inherent within some types of data. This is shown in the Figure 2.3.

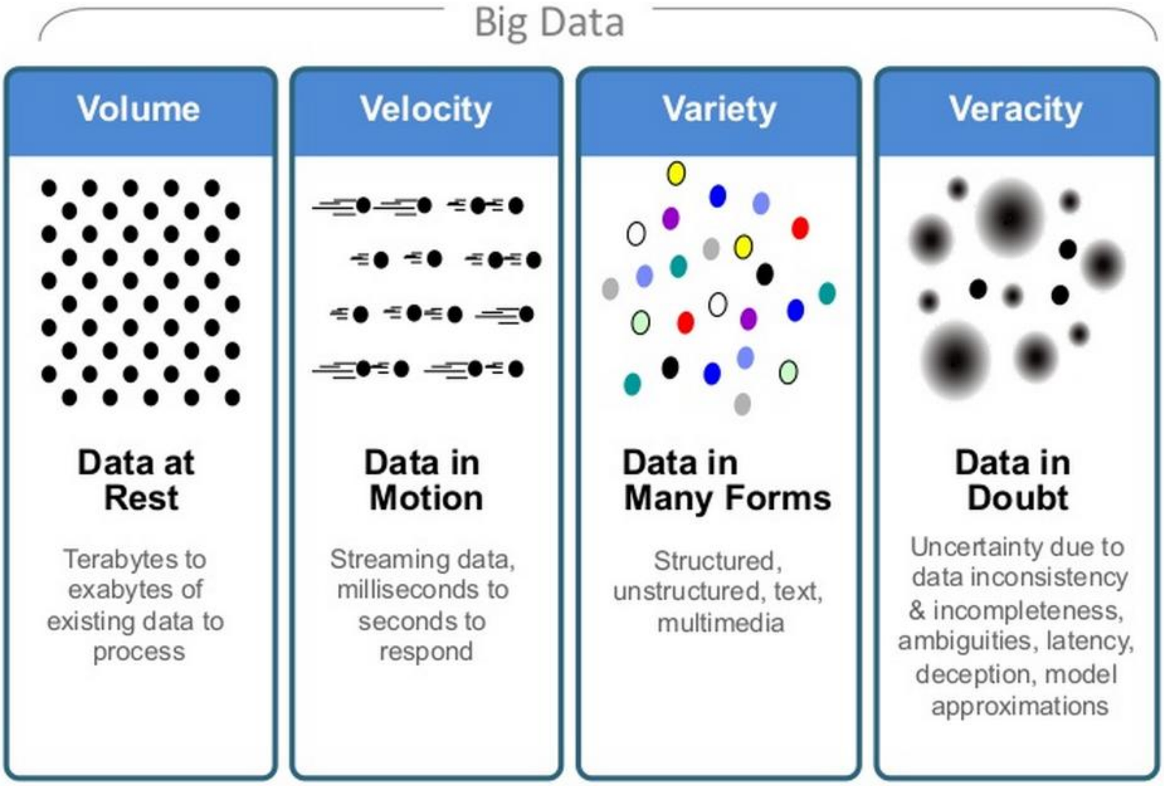

Big Data
Volume
Velocity
Variety
Veracity
Data at Rest
Terabytes to exabytes of existing data to process
Data in Motion
Streaming data, milliseconds to seconds to respond
Data in Many Forms
Structured, unstructured, text, multimedia
Data in Doubt
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Figure. 2.3 Four dimensions of big data[4].

The union of these four dimensions helps to both define and differentiate about big data:

Volume

This refers to the magnitude of data. It is the characteristic mostly associated with big data. Volume refers to the mass quantities of data that organizations are trying to harness to improve decision-making across the organization. Data volumes are reported in multiple terabytes and petabytes, and they continue to increase at an exceptional rate. The definitions of big data volumes are relative and vary by factors, such as time and the type of data. What may be deemed big data today may not meet the threshold in the future since storage capacities will be able to be increased, allowing even bigger data sets to be captured. In addition, the type of data, discussed under variety, defines what is meant by "big". Two datasets of the same size may require different data management technologies based on their type. In this manner, definitions of big data also depend upon the industry. These considerations therefore make it impractical to define a specific threshold for big data volumes[2].

Variety

This refers to the structural heterogeneity in a data set. Technological advances allow firms to use various types of structured, semi-structured and unstructured data. Structured data, which constitutes only 5% of all existing data, refer to the tabular data found in spreadsheets or relational databases (Cukier, 2010)[5]. Text, images, audio, and video are examples of unstructured data, which sometimes lack the structural organization required by machines for analysis. Spanning a continuum between fully structured and unstructured data, the format of semi-structured data does not conform to strict standards. Extensible Markup Language (XML), a textual language for exchanging data on the web, is a typical example of semi-structured data. XML documents contain user-defined data tags which make them machine-readable. A high level of variety, a defining characteristic of big data, is not necessarily new. Companies have been putting away unstructured data from internal sources (e.g., sensor data) and external sources (e.g., social media). Nevertheless, the emergence of new data management technologies and analytics, which enable organizations to leverage data in their business processes, is the innovative aspect. For instances, facial recognition technologies empower the brick-and-mortar retailers to acquire intelligence about store traffic, the age or gender composition of their customers, and their in-store movement patterns. This invaluable information is leveraged in decisions related to product promotions, placement, and staffing. Clickstream advises on the timing and sequence of pages viewed by a customer. By using big data analytics, even small and medium-sized enterprises (SMEs) can mine massive volumes of semi-structured data to improve website designs and implement effective cross-selling and personalized product recommendation systems [3].

Velocity

This refers to the rate at which data are generated and the speed at which it should be analysed and acted upon. The proliferation of digital devices such as smartphones and sensors has led to an unprecedented rate of data creation and is driving a growing need for real-time analytics and evidence-based planning. Even conventional retailers are generating high-frequency data. As an example, Wallmart processes more than one million transactions per hour (Cukier,2010)[5]. The data emanating from mobile devices and flowing through mobile apps produce torrents of information that can be used to generate real-time personalized offers for everyday customers. This data provides sound information about customers, such as geospatial location, demographics, and past buying patterns, which can be analysed in real time to create customer value.

Given the soaring popularity of smartphones, retailers will soon have to deal with hundreds of thousands of streaming data sources that demand real-time analytics. Traditional data management system are not capable of handling huge data feeds instantaneously. This is where big data technologies come into play. They enable firms to create real-time intelligence from high volumes of "perishable" data.

Veracity

This refers to data uncertainty. Veracity is the level of reliability associated with certain types of data. Striving for high data quality is an important big data requirement and a challenge, but even the best data cleansing methods cannot remove the inherent unpredictability of some data, like the weather, the economy, or a customer’s actual future buying decisions. The need to acknowledge and plan for uncertainty is a dimension of big data that has been introduced as executives seek to better understand the uncertain world around them. An example of this uncertainty is in energy production: the weather is uncertain, but a utility company must still forecast production. In many countries, regulators require a percentage of production to come from renewable sources, yet neither wind nor clouds can be forecasted with precision.

So, in order to manage uncertainty, analysts need to create context around the data. One way to achieve this is through data fusion, where combining multiple less reliable sources creates a more accurate and useful data point, such as social comments appended to the geospatial location information. Another way to manage uncertainty is through advanced mathematics that embraces it, such as robust optimization techniques and fuzzy logic approaches. Humans, by nature, dislike uncertainty, but just ignoring it can create even more problems than the uncertainty itself. In the era of big data, executives will need to approach the dimension of uncertainty differently. They will need to acknowledge it, embrace it and determine how to use it to their advantage; the one certainty about uncertainty is that it is not likely to go away.[3]

The relativity of big data volumes, as discussed earlier, applies to all dimensions. No universal benchmark exists for what volume, variety, and velocity defines big data. The defining limits depend upon the size, the sector, and location of the firm, and these limits evolve over time. Also important is the fact that these dimensions are not independent of each other. As one dimension changes, the others also change as a result. However, a "three-V tipping point" exists for every firm beyond which traditional data management and analysis technologies become inadequate for deriving timely intelligence. The Three-V tipping point is the threshold beyond which firms start dealing with big data. The firms should then trade-off the future value expected from big data technologies against their implementation costs.

The real value of big data has just been unlocked when leveraged to drive decision-making. To be able to have such evidence-based decision-making, organizations need efficient processes to turn high volumes of fast-moving and diverse data into meaningful insights. The overall process of extracting insights from big data can be broken down into five stages, as shown in Figure 2.4. [6]

These five stages form the two main sub-processes: data management and analytics. Data management involves processes and supporting technologies to acquire and store data and to prepare and retrieve it for analysis. Analytics, on the other hand, refers to techniques used to analyse and acquire intelligence from big data. Thus, big data analytics can be viewed as a sub-process in the overall process of ‘insight extraction’ from big data.

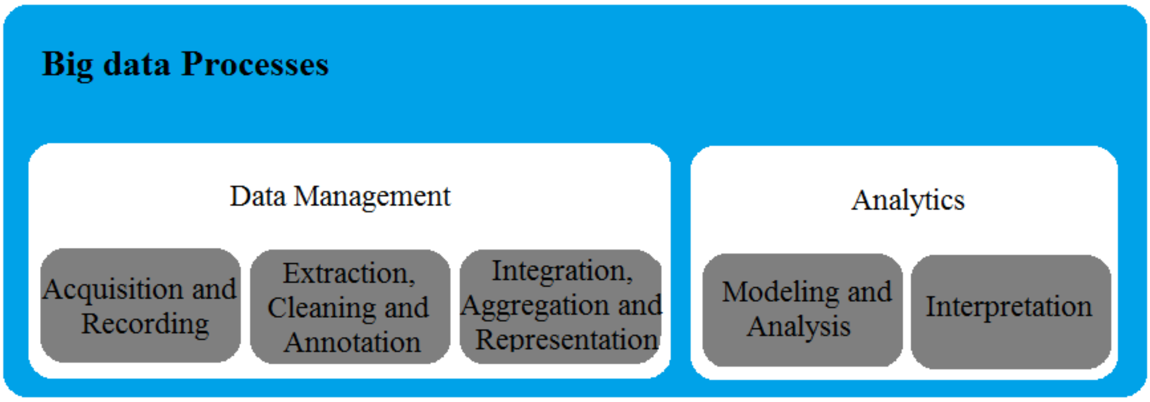

Big data Processes
Data Management
Acquisition and Recording
Extraction, Cleaning and Annotation
Integration, Aggregation and Representation
Analytics
Modeling and Analysis
Interpretation

Figure. 2.4 Processes for extracting insights from big data.[3].

There are many techniques that draw on disciplines such as statistics and computer science (particularly machine learning) that can be used to analyse datasets. In this section, we provide a list of some categories of techniques applicable across a range of industries. This list is by no means exhaustive. Indeed, researchers continue to develop new techniques and improve on existing ones, particularly in response to the need to analyze new combinations of data. We note that not all of these techniques strictly require the use of big data—some of them can be applied effectively to smaller datasets (e.g., A/B testing, regression analysis). However, all of the techniques listed here can be applied to big data and, in general, larger and more diverse datasets can be used to generate more numerous and insightful results than smaller, less diverse ones.

Techniques for analyzing Big Data[7]:

A/B testing: A technique in which a control group is compared with a variety of test groups in order to determine what treatments (i.e., changes) will improve a given objective variable, e.g., marketing response rate. This technique is also known as split testing or bucket testing. An example application is determining what copy text, layouts, images, or colours will improve conversion rates on an e-commerce website. Big data enables huge numbers of tests to be executed and analysed, ensuring that groups are of sufficient size to detect meaningful (i.e., statistically significant) differences between control and treatment groups (see statistics). When more than one variable is simultaneously manipulated in the treatment, the multivariate generalization of this technique, which applies statistical modeling, is often called "A/B/N" testing.

Association rule learning: A set of techniques for discovering interesting relationships, i.e., "association rules," among variables in large databases.27 These techniques consist of a variety of algorithms to generate and test possible rules. One application is market basket analysis, in which a retailer can determine which products are frequently bought together and use this information for marketing (a commonly cited example is the discovery that many supermarket shoppers who buy diapers also tend to buy beer). Used for data mining.

Classification: A set of techniques to identify the categories in which new data points belong, based on a training set containing data points that have already been categorized. One application is the prediction of segment-specific customer behavior (e.g., buying decisions, churn rate, consumption rate) where there is a clear hypothesis or objective outcome. These techniques are often described as supervised learning because of the existence of a training set; they stand in contrast to cluster analysis, a type of unsupervised learning. It is used for data mining.

Cluster analysis: A statistical method for classifying objects that splits a diverse group into smaller groups of similar objects, whose characteristics of similarity are not known in advance. An example of cluster analysis is segmenting consumers into self-similar groups for targeted marketing. This is a type of unsupervised learning because training data are not used. In contrast to classification, this technique is a type of supervised learning. It is used for data mining.

Data fusion and data integration: A set of techniques that integrate and analyze data from multiple sources in order to develop insights in ways that are more efficient and potentially more accurate than if they were developed by analyzing a single source of data. Signal processing techniques can be used to implement some types of data fusion. One example of an application is sensor data from the Internet of Things being combined to develop an integrated perspective on the performance of a complex distributed system such as an oil refinery. Data from social media, analyzed by natural language processing can be combined with real-time sales data in order to determine what effect a marketing campaign is having on customer sentiment and purchasing behaviour.

Data mining: A set of techniques to extract patterns from large datasets by combining methods from statistics and machine learning with database management. These techniques include association rule learning, cluster analysis, classification, and regression. Applications include mining customer data to determine segments most likely to respond to an offer, mining, human resources data to identify characteristics of the most successful employees, or market basket analysis to model the purchase behaviour of customers.

Machine Learning: A subspecialty of computer science (within a field historically called "artificial intelligence") concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data. Natural language processing is an example of machine learning.

Natural language processing (NLP): A set of techniques from a subspecialty of computer science (within a field historically called "artificial intelligence") and linguistics that uses computer algorithms to analyse human (natural) language. Many NLP techniques are types of machine learning. One application of NLP is using sentiment analysis on social media to determine how prospective customers are reacting to a branding campaign.

Neural networks: Computational models, inspired by the structure and workings of biological neural networks (i.e., the cells and connections within a brain), that find patterns in data. Neural networks are well-suited for finding nonlinear patterns. They can be used for pattern recognition and optimization. Some neural network applications involve supervised learning and others involve unsupervised learning. Examples of applications include identifying high-value customers that are at risk of leaving a particular company or identifying fraudulent insurance claims.

Network analysis: A set of techniques used to characterize relationships among discrete nodes in a graph or a network. In social network analysis, connections between individuals in a community or organization are analysed, e.g., how information travels, or who has the most influence over whom. Examples of applications include identifying key opinion leaders to target for marketing, and identifying bottlenecks in enterprise information flows.

Pattern recognition: A set of machine learning techniques that assign some sort of output value (or label) to a given input value (or instance) according to a specific algorithm. Classification techniques are an example.

2.2. Data Mining

2.2.1. Over view

Data mining represents the "core" step of the KDD process. The KDD was already explained in the point 1.1 in the beginning of the thesis. To have a better understanding about the KDD and how connected is with data mining, the Figure 2.5 show us the sequence of the main steps:

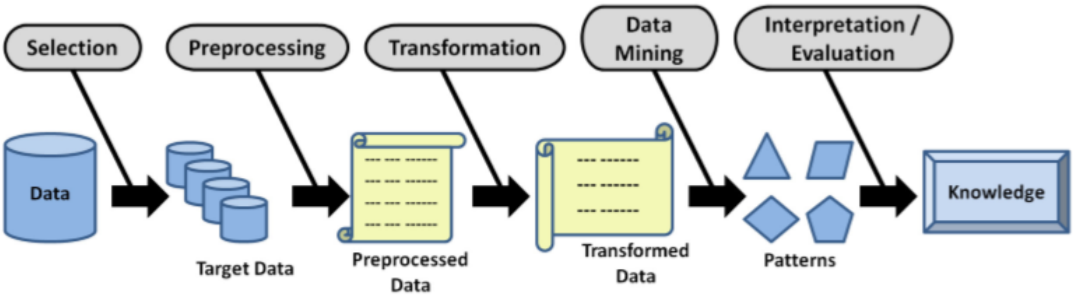


Figure. 2.5 The Knowledge Discovery in Databbases (KDD) process [8].

Description of the main steps [8]:

- Selection, whose main goal is to create a target data set from the original data, i.e., selecting a subset of variables or data samples, on which discovery has to be performed;

- Preprocessing, which aims to "clean" data by performing various operations, such as noise modeling and removal, defining proper strategies for handling missing data fields, accounting for time-sequence information;

- Transformation, which is in charge of reducing and projecting the data, in order to derive a representation suitable for the specific task to be performed; it is typically accomplished by involving transformation techniques or methods that are able to find invariant representations of the data;

- Data Mining, which deals with extracting interesting patterns by choosing a specific data-mining method or task (e.g., summarization, classification, clustering, regression, and so on), proper algorithm(s) for performing the task at hand, and an appropriate representation of the output results;

- Interpretation/evaluation, which is exploited by the user to interpret and extract knowledge from the mined patterns, by visualizing the patterns; this interpretation is typically carried out by visualizing the patterns, the models or the data given such models and, in case, iteratively looking back at the previous steps of the process.

Data mining is the fourth step of the KDD process. These are so much connected that "data mining” and "KDD" terms are often treated as synonyms. Several definitions of what data mining is used, e.g., “automated yet non-trivial extraction of implicit, previously unknown, and potentially useful information from data”, “automated exploration and analysis of large quantities of data in order to discover meaningful patterns”, and “computational process of automatically extracting useful knowledge from large amounts of data”. All definitions are roughly equivalent to each other. They all agree on the main aspects of data mining, which are: huge quantity of data that should be analysed so as to extract what is called “knowledge”, or “useful information”, or “patterns”, i.e., something that can be processed and profitably exploited by human beings. The current importance of data mining is mainly motivated by the lots of data that is collected and stored in a variety of today’s prominent applications. This data includes Web data, e-commerce data, purchases, bank transactions, and so on. Also, the number of applications dealing with data that needs to be processed at enormous speeds (GB/seconds or even more) is rapidly increasing. Examples include remote sensors on satellites, telescopes scanning the skies, microarray generating gene-expression data, and scientific simulations. Due to the peculiarity of the underlying data, it is apparent that data analysis in such challenging contexts cannot be performed with traditional data-analysis techniques, either manual or automated. Data mining aims at filling this gap, with its intrinsic interdisciplinary nature that poses it at the intersection of a number of more classical fields, such as artificial intelligence, statistics, database systems and machine learning.

2.2.2. Methods

Data mining incorporates a number of methods that can be used, even in combination, based on the requisites of the specific application context. Data mining methods are usually sorted into predictive and descriptive. Predictive methods refer to building a model useful for predicting future behaviour or values for certain features. Among others, these include classification and prediction, i.e., deriving some models (or functions) that describe data classes or concepts by a set of data objects whose class label is known (i.e., the training set), so as to predict the class of objects whose class label is unknown; deviation detection, i.e., dealing with deviations in data, which are defined as diferences between measured values and corresponding references such as previous values or normative values; evolution analysis, i.e., detecting and describing regular patterns in data whose behavior changes over time. On the other hand, in a descriptive data-mining method, the built-in model aims at describing the data in an understandable, effective, and efficient form. Relevant examples of descriptive tasks are data characterization, whose main goal is to summarize the general characteristics or features of a target class of data; data discrimination, i.e., a comparison of the general features of a target class of data objects with the general features of objects from a set of contrasting classes; association-rule discovery, i.e., discovering rules that show attribute-value conditions occurring frequently together in a given set of data; and clustering, which aims at forming highly cohesive and well separated groups of objects from the input set of data objects.

- **Classification**

The classification method takes, as input, a collection of records, called training sets, where each record is composed of a set of attributes and one of the attributes denotes the class of the record. The goal is to find a model for the class attribute as a function of the values of the other attributes. The model is then used to predict the class attribute of previously unobserved records. As an example, consider a collection of records describing the position held by the academic staff of any university. Assume that each record has the following attributes: name of the professor, position (i.e., assistant professor, associate professor, or full professor), number of years she has been affiliated to such a university, and the class attribute, that is a boolean attribute that indicates whether the professor holds a tenured position or not. Assume also that the input collection contains the following four records: (Mike, Assistant Prof, 3, no), (Mary, Assistant Prof, 7, yes), (Bill, Full Prof, 2, yes), and (Anne, Associate Prof, 7, yes). Based on this input, a classification algorithm would likely find a model expressed by the following (set of) rule (s): “IF position=Full Prof OR years > 3 THEN tenured=yes”. Thus, given a new record (Barbara, Full Prof, 4,?), the model would predict the missing class value as a yes. Classification is a long-standing area of research where a plethora of different approaches and algorithms have been defined, including k Nearest Neighbours (KNN), decision trees, Support Vector Machines (SVM), neural networks, and Gradient Boosted Decision Trees (GDBT; Kotsiantis, 2007).

- **Clustering**

Given a set of data objects, clustering aims at identifying a finite set of groups of objects, i.e., clusters, so that the objects within the same cluster are “similar” to each other, whereas the objects belonging to different clusters are “dissimilar”. The degrees of (dis)similarity between data objects are computed and evaluated according to a proximity measure that can be either specified by the user or inherently incorporated into the specific clustering algorithm. In a clustering task, there is no prior knowledge of the class labels associated with the objects to be grouped; for this reason, clustering is often also referred to as unsupervised classification, to emphasize the difference from the (supervised) classification task, in which the class labels of the objects in the training set are known.

Given a set of data objects, clustering aims at identifying a finite set of groups of objects, i.e., clusters, so that the objects within the same cluster are “similar” to each other, whereas the objects belonging to different clusters are “dissimilar”. The degrees of (dis)similarity between data objects are computed and evaluated according to a proximity measure that can be either specified by the user or inherently incorporated into the specific clustering algorithm. In a clustering task, there is no prior knowledge of the class labels associated with the objects to be grouped; for this reason, clustering is often also referred to as unsupervised classification, to emphasize the difference from the (supervised) classification task, in which the class labels of the objects in the training set are known. A clustering of the input set of objects is thus built in such a way that cluster cohesiveness and separation, measured in terms of the underlying proximity measure, are maximized. More precisely, clustering methods typically define a specific objective function to be optimized in order to formally define clusters that are compact and well-separated from each other. Since these formulations usually lead to computational problems too hard to be optimally solved for large-scale inputs (the so-called NP-hard problems), any specific clustering method should define the corresponding approximation/heuristic algorithm(s) to find good approximations of the optimal solution. The literature abounds with different clustering approaches and algorithms, which differ from each other in the optimization criterion, in the solution strategy, and in the computation of the distance between the input objects (Aggarwal and Reddy, 2014). These algorithms can be classified according to a lot of different taxonomies, which, however, usually all agree on the top-level division in two main categories, i.e., partitional (or partitioning) and hierarchical. Broadly, partitional-clustering algorithms compute a single part it on of the input data set. A considerable number of partitions algorithms exploit the relocation scheme, i.e., the objects are iteratively re-assigned to the clusters, until a stop criterion has met. As an example, such a scheme is at the basis of the popular K-Means algorithm MacQueen (1967). Rather than a single partition of the input dataset, hierarchical-clustering approaches output, instead, a hierarchy of clustering solutions that are organized into a tree like structures known as dendrograms (Aggarwal and Reddy, 2014).

- **Association-rule discovery**

Given a set of records (i.e., transactions), each of which containing a number of items from a given collection, the goal of the association-rule discovery is to produce dependency rules that can predict the occurrence of an item based on occurrences of other items. As an example, think about an electronic-device shop where, for marketing reasons, one is interested in understanding the best way to expose items to customers in order to increase purchases. In this case, one can analyse the past purchasing history in order to discover association rules like {camera, tripod} → {SD memory}, which informally states that, when customers buy a camera and a tripod, it is very likely that they buy an SD memory as well. Such a rule can be used in several ways. For instance, cameras and tripods can be used to boost the sales of SD memories by, for example, storing the cameras and tripods close to SD memories or putting cameras in bundle promotion with tripods. A preliminary step commonly required by association-rule-discovery algorithm corresponds to another classical data-mining task know as frequent pattern mining (Han et al., 2007), whose main goal is to find subsets of items that co-occur frequently in a set of transactions. For instance, the above example association rule {camera, tripod} → {SD memory} would derive from the preliminary discovery that cameras, tripods and SD memories frequently appear together in a purchasing data log.

Even being general enough to handle any type of data, data mining can, however, be “customized” to deal with a specific typology of data and to focus on specific data peculiarities. Therefore, data mining is a powerful tool that has been used for decades for advanced analysis of large quantities of data. It is defined as the automated, yet non trivial, extraction of implicit, previously unknown, and potentially useful information from data. This technical note provided a broad overview of the main data-mining principles and its interdisciplinary aspects.

2.3. Machine Learning

2.3.1. Overview

One of the most outstanding differences between how people and computers work is that humans, while executing any kind of activity, normally simultaneously expend effort to improve the way they perform it. This is to say that human performance of any task is inseparably intertwined with a learning process, while current computers are typically only executors of procedures supplied to them. They may perform very efficiently, but they do not self-improve with experience. Research in machine learning has been concerned with building computer programs able to construct new knowledge or to improve knowledge that is already present, by using input information. Machine Learning is a subfield of computer science that involves the study of pattern recognition and computational learning theory in artificial intelligence. It explores the construction and study algorithms that can learn from and make predictions on data. With the input information, it is possible to build models such algorithms in order to make data-driven predictions or decisions. Machine learning often overlaps with computational statistics, a discipline that also specializes in prediction making. It has strong ties to mathematical optimization, which deliver methods, theory and application domains in the field. When employed in industrial contexts, machine learning methods may be referred to as predictive analytics or predictive modelling. In the following pages we will shortly make an introduction to the most popular machine learning approaches. This machine learning approaches focus on the development of algorithms. The first is a grouping of algorithms by the learning style. The second is a grouping of algorithms by similarity in form or function (like grouping similar animals together). Both approaches are useful.

2.3.2. Methods

This chapter is all based on Literature [8].

One of the most striking differences between how people and computers work, is that humans, while performing any kind of activity, usually simultaneously expend effort to improve the way they perform it. This is to say that human performance of any task is inseparably intertwined with a learning process, while current computers are typically only executors of procedures supplied to them. They may execute very efficiently, but they do not self-improve with experience. Research in machine learning has been concerned with building computer programs able to construct new knowledge or to improve already possessed knowledge by using input information. Machine Learning is a subfield of computer science that involves the study of pattern recognition and computational learning theory in artificial intelligence. It explores the construction and study algorithms that can learn from and make predictions on data. With the input information, it is possible to build models such algorithms in order to make data-driven predictions or decisions. Machine learning often overlaps with computational statistics, a discipline that also specializes in prediction making. It has strong ties to mathematical optimization, which delivers methods, theory and application domains in the field. When employed in industrial contexts, machine learning methods may be referred to as predictive analytics or predictive modelling. In the following pages we will shortly make an introduction to the most popular machine learning approaches. This machine learning approaches focus on the development of algorithms. The first is a grouping of algorithms by learning style. The second is a grouping of algorithms by similarity in form or function (like grouping similar animals together). Both approaches are useful.

2.3.2.1. Learning Style

There are different ways an algorithm can model a problem, based on its interaction with the experience or environment or whatever we want to call the input data. It has become popular, in both machine learning and artificial intelligence text books, to first consider the learning styles that an algorithm can adopt.

There are only a few main learning styles or learning models that an algorithm can have and we'll go through them here with a few examples of algorithms and problem types that they suit. This taxonomy or way of organizing machine learning algorithms is useful because it forces one to think about the roles of input data and model preparation process, and select one that is the most appropriate for a problem in order to get the best result.

Supervised Learning: Input data are called training data and has a known label or result such as spam/not-spam or a stock price at a time. A model is prepared through a training process where it is required to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data. The example problems are classification and regression. Example algorithms are Logistic Regression and the Back Propagation Neural Network.

Unsupervised Learning: Input data are not labelled and do not have a known result. A model is prepared by deducing structures present in the input data. The example problems are association rule learning and clustering. Example algorithms are the Apriori algorithm and k-means.

Reinforcement Learning: Input data are provided as stimulus to a model from an environment to which the model must respond and react. Feedback is provided not from a teaching process as in supervised learning, but as punishments and rewards in the environment. The example problems are systems and robot control. Example algorithms are Q-learning and Temporal difference learning.

When crunching data to model business decisions, you are most typically using supervised and unsupervised learning methods. A hot topic at the moment is semi-supervised learning methods in areas such as image classification where there are large data sets with very few labelled examples. Reinforcement learning is more likely to turn up in robotic control and other control systems development.

2.3.3. Algorithm Similarity

Algorithms are universally present in groups by similarity in terms of function or form, as for example, tree based methods, and neural network inspired methods. This is a useful grouping method, but it is not perfect. There are still algorithms that could just as easily fit into multiple categories like Learning Vector Quantization that is both a neural network inspired method and an instance-based method. There are also categories that have the same name that describes the problem and the class of algorithm such as Regression and Clustering. As such, you will see variations in the way algorithms are grouped depending on the source you check. Like machine learning algorithms themselves, there is no perfect model, just a good enough model.

In this section I list many of the popular machine learning algorithms grouped the way I think is the most intuitive. It is not exhaustive in either the groups or the algorithms, but I think it is representative and will be useful for you to get an idea of the lay of the land. If you know of an algorithm or a group of algorithms not listed, put it in the comments and share it with us. Let’s dive in.

Regression

Regression is concerned with modelling the relationship between variables that is iteratively refined using a measure of error in the predictions made by the model. Regression methods are a work horse of statistics and have been copied into statistical machine learning. This may be confusing because we can use regression to refer to the class of problem and the class of algorithm. In all reality, regression is a process. Some example of algorithms are:

- Ordinary Least Squares
- Logistic Regression
- Stepwise Regression
- Multivariate Adaptive Regression Splines (MARS)
- Locally Estimated Scatter Plot Smoothing (LOESS)

Instance-based Methods

Instance based learning model is a decision problem with instances or examples of training data that are deemed important or required for the model. Such methods typically build up a database of example data and compare new data to the database using a similarity measure in order to find the best match and make a prediction. For this reason, instance-based methods are also called winner-take-all methods and memory-based learning. Focus is put on the representation of the stored instances and similarity measures used between instances.

- K-Nearest Neighbour (KNN)
- Learning Vector Quantization (LVQ)
- Self-organizing Map (SOM)

Decision Tree Learning

Decision tree methods construct a model of decisions made based on the actual values of attributes in the data. Decisions fork in tree structures until a prediction decision is made for a given record. Decision trees are trained on data for classification and regression problems.

- Classification and Regression Tree (CART)
- Iterative Dichotomiser 3 (ID3)
- C4.5
- Chi-squared Automatic Interaction Detection (CHAID)
- Multivariate Adaptive Regression Splines (MARS)
- Gradient Boosting Machines (GBM)

Kernel Methods

Kernel Methods are best known for the popular method Support Vector Machines which is really a constellation of methods in and of itself. Kernel Methods are concerned with mapping input data into a higher dimensional vector space where some classification or regression problems are easier to model.

- Support Vector Machines (SVM)
- Radial Basis Function (RBF)
- Linear Discriminant Analysis (LDA)

Clustering Methods

Clustering, like regression, describes the class of problem and the class of methods. Clustering methods are typically organized by the modelling approaches such as centroid-based and hierarchal. All methods are concerned with using the inherent structures in the data to best organize the data into groups of maximum commonality.

- k-Means
- Expectation Maximisation (EM)
- Association Rule Learning

Artificial Neural Networks

Artificial Neural Networks are models that are inspired by the structure and/or function of biological neural networks. They are a class of pattern matching that are commonly used for regression and classification problems, but are really an enormous subfield comprised of hundreds of algorithms and variations for all manner of problem types. Some of the classically popular methods include (I have separated Deep Learning from this category):

- Perceptron
- Back-Propagation
- Hopfield Network
- Self-organizing Map (SOM)
- Learning Vector Quantization (LVQ)

This explanation of machine learning algorithm was intended to give an overview of what is out there and some tools to relate algorithms that can be used.

3. KNOWLEDGE DISCOVERY IN DIE CASTING PROCESS

3.1. Preparation and Selection

It deals with gathering data from different sources. Data sources are firstly located and integrated. Target data are selected from all of the original data. Further discoveries are performed on the subset of the original data, samples and variables.

3.1.1. Database server

Data stored in a warehouse cannot be accessed directly from a computer. The access is usually provided by a database server. It is a computer program that defines a client - server model. In order to retrieve data from the warehouse we communicate with the server, which translates this into a query language (T-SQL and ANSI-SQL) readable by the database [9].

The client-server is a software architecture model that consisted of two parts. There is a server and a client. The client can request content from the server and does not share any of its resources. The server waits for an incoming request from the client to then share the content available in the warehouse. In Figure 3.1 we can see a schematic representation of the architecture of the client-server model.

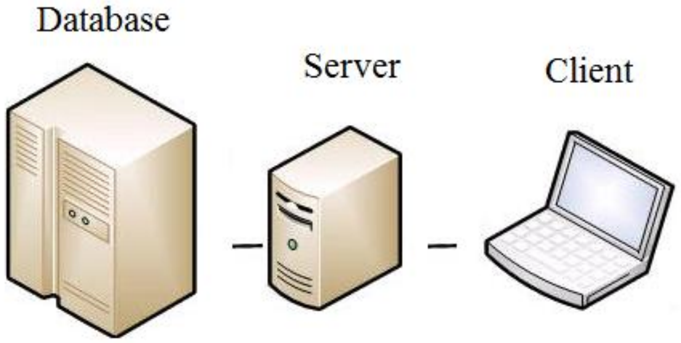


Figure. 3.1 Architecture of the client-server model.

3.1.2. Database structure

Data in SQL server Management Studio is organized in tables. The program provides a guide to the database server, which, in turn, provides access to the database. The Figure 3.2 shows an example of the diagram of a database.

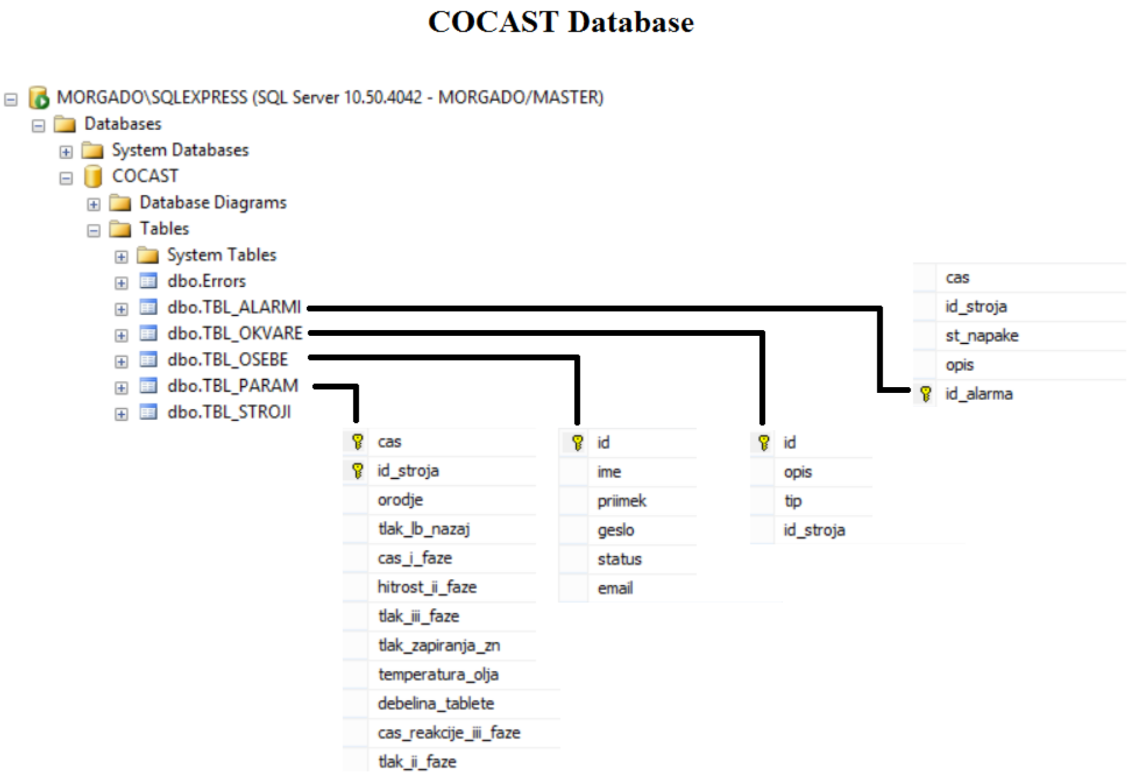
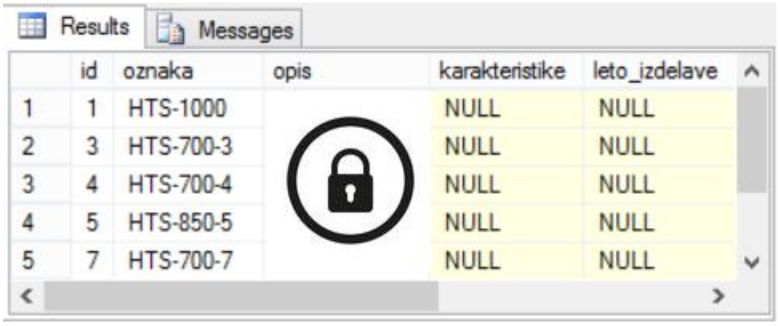


Figure. 3.2 COCAST Database Diagram.

The database we had access to is called COCAST database and consists in this specific example of 3 tables.(*the categories of information are in Croatian language, since this is the way they were collected). The Table 3.1 shows an example of a COCAST Database table



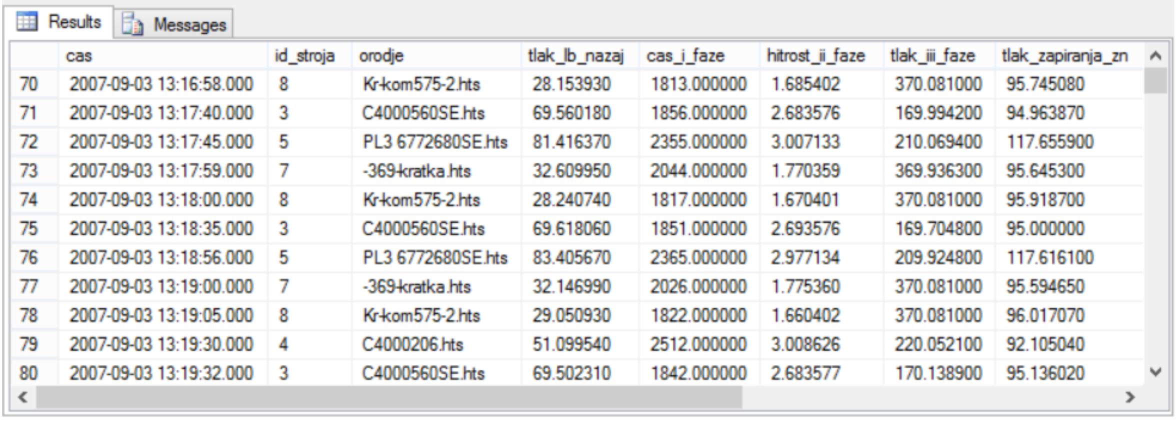
	id	oznaka	opis	karakteristike	leto_izdelave
1	1	HTS-1000		NULL	NULL
2	3	HTS-700-3		NULL	NULL
3	4	HTS-700-4		NULL	NULL
4	5	HTS-850-5		NULL	NULL
5	7	HTS-700-7		NULL	NULL

Table. 3.1 COCAST Database table.

The Table 3.1 shows us a table from the COCAST database where first, give us information about the machines, such as its number, produces, model, etc.

Table columns represents a specific category of information, while the rows represent information about specific objects.

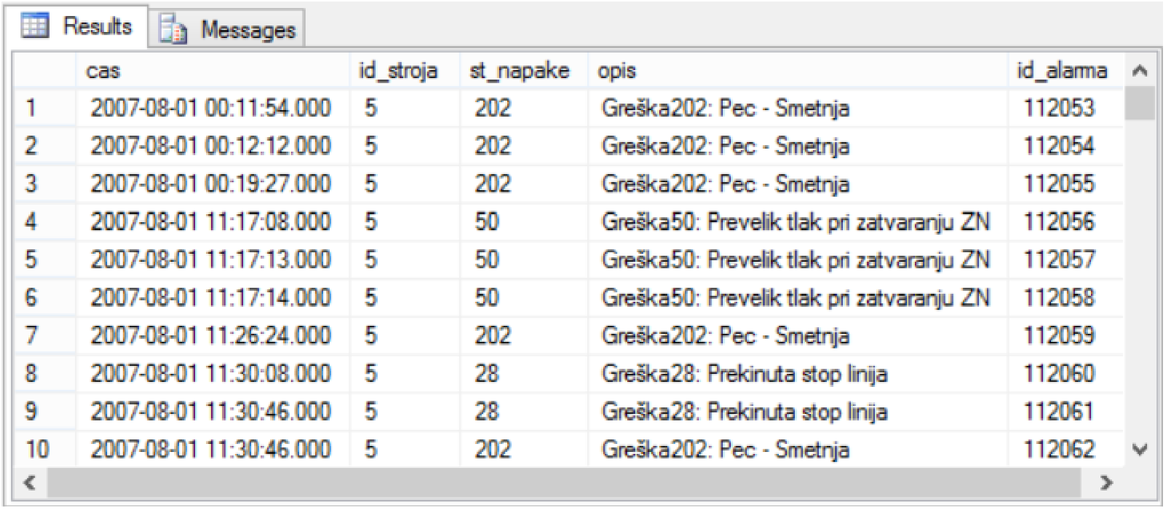
The Table 3.2 gives us information about the conditions at which the pieces were produced. It tells us the time of production, machine used and tool used, as well as, the parameters collected during the piece production (pressure at closing, temperature of oil, pressure at return, etc...).



	cas	id_stroja	orodje	tlak_lb_nazaj	cas_i_faze	hitrost_ii_faze	tlak_iii_faze	tlak_zapiranja_zn
70	2007-09-03 13:16:58.000	8	Kr-kom575-2.hts	28.153930	1813.000000	1.685402	370.081000	95.745080
71	2007-09-03 13:17:40.000	3	C4000560SE.hts	69.560180	1856.000000	2.683576	169.994200	94.963870
72	2007-09-03 13:17:45.000	5	PL3 6772680SE.hts	81.416370	2355.000000	3.007133	210.069400	117.655900
73	2007-09-03 13:17:59.000	7	-369-kratka.hts	32.609950	2044.000000	1.770359	369.936300	95.645300
74	2007-09-03 13:18:00.000	8	Kr-kom575-2.hts	28.240740	1817.000000	1.670401	370.081000	95.918700
75	2007-09-03 13:18:35.000	3	C4000560SE.hts	69.618060	1851.000000	2.693576	169.704800	95.000000
76	2007-09-03 13:18:56.000	5	PL3 6772680SE.hts	83.405670	2365.000000	2.977134	209.924800	117.616100
77	2007-09-03 13:19:00.000	7	-369-kratka.hts	32.146990	2026.000000	1.775360	370.081000	95.594650
78	2007-09-03 13:19:05.000	8	Kr-kom575-2.hts	29.050930	1822.000000	1.660402	370.081000	96.017070
79	2007-09-03 13:19:30.000	4	C4000206.hts	51.099540	2512.000000	3.008626	220.052100	92.105040
80	2007-09-03 13:19:32.000	3	C4000560SE.hts	69.502310	1842.000000	2.683577	170.138900	95.136020

Table. 3.2 Information about the conditions at which the pieces were produced.

The Table 3.3 gives us the information about triggered alarms during the manufacturing process. It tells us the time of alarm, number of the machine that triggered the alarm, the number of the alarm and brief description of the alarm.

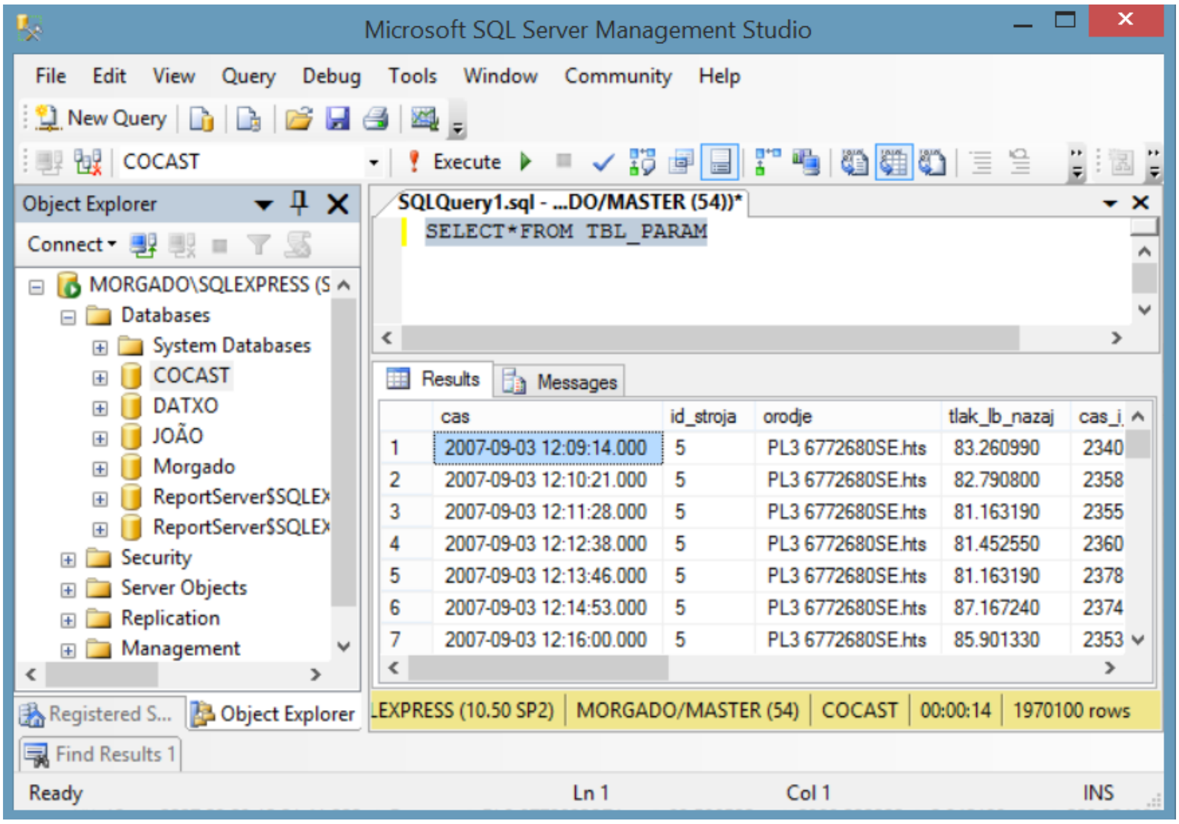


	cas	id_stroja	st_napake	opis	id_alarma
1	2007-08-01 00:11:54.000	5	202	Greška202: Pec - Smetnja	112053
2	2007-08-01 00:12:12.000	5	202	Greška202: Pec - Smetnja	112054
3	2007-08-01 00:19:27.000	5	202	Greška202: Pec - Smetnja	112055
4	2007-08-01 11:17:08.000	5	50	Greška50: Prevelik tlak pri zatvaranju ZN	112056
5	2007-08-01 11:17:13.000	5	50	Greška50: Prevelik tlak pri zatvaranju ZN	112057
6	2007-08-01 11:17:14.000	5	50	Greška50: Prevelik tlak pri zatvaranju ZN	112058
7	2007-08-01 11:26:24.000	5	202	Greška202: Pec - Smetnja	112059
8	2007-08-01 11:30:08.000	5	28	Greška28: Prekinuta stop linija	112060
9	2007-08-01 11:30:46.000	5	28	Greška28: Prekinuta stop linija	112061
10	2007-08-01 11:30:46.000	5	202	Greška202: Pec - Smetnja	112062

Table. 3.3 Information about triggered alarms during the manufacturing process.

3.1.3. SQL management studio

We used Microsoft SQL Server 2008 Management Studio, which enables us to interact with SQL Server trough user interface. It is an integrated environment that provides us with tools for configuring, managing and administrating data gathered on Microsoft SQL Server with graphics and visual design tools that simplify our work. Using the program, first we have to access the server. We do this by logging in to the server. After logging in we have access to all the data in the warehouse connected to the server. How the environment looks at this point is shown in Figure 3.3 (screenshot program) [10].



```
SELECT*FROM TBL_PARAM
```

	cas	id_stroja	orodje	tlak_lb_nazaj	cas_i
1	2007-09-03 12:09:14.000	5	PL3 6772680SE.hts	83.260990	2340
2	2007-09-03 12:10:21.000	5	PL3 6772680SE.hts	82.790800	2358
3	2007-09-03 12:11:28.000	5	PL3 6772680SE.hts	81.163190	2355
4	2007-09-03 12:12:38.000	5	PL3 6772680SE.hts	81.452550	2360
5	2007-09-03 12:13:46.000	5	PL3 6772680SE.hts	81.163190	2378
6	2007-09-03 12:14:53.000	5	PL3 6772680SE.hts	87.167240	2374
7	2007-09-03 12:16:00.000	5	PL3 6772680SE.hts	85.901330	2353

```
.EXPRESS (10.50 SP2) | MORGADO/MASTER (54) | COCAST | 00:00:14 | 1970100 rows
```

Figure. 3.3 Layout SQL management studio.

3.1.4. SQL language (How the language works)

SQL (Structured Query Language) is a language used to communicate with databases. It enables us to update, delete and request information from a certain database. It was standardized in 1986 by ANSI (American National Standards Institute). Many different database products use SQL, such as Microsoft SQL Server, Oracle, Sybase and others. These databases contain objects called tables, which are identified by their specific name. They store information in a form of columns and rows. We can manipulate this information using commands. To do that, client programs send SQL statements to the server. The server then processes these statements and returns result sets to the client program [11].

The SELECT statement is one of the many commands available in SQL language.

An SQL SELECT statement retrieves data from a database table according with the criteria that we specify. The Figure 3.4 shows the structure of statement to retrieve data [12]:

```
SELECT column1, column2 FROM table1, table2 WHERE column2='value';
```

Figure. 3.4 Mathematica statement to retrieve data.

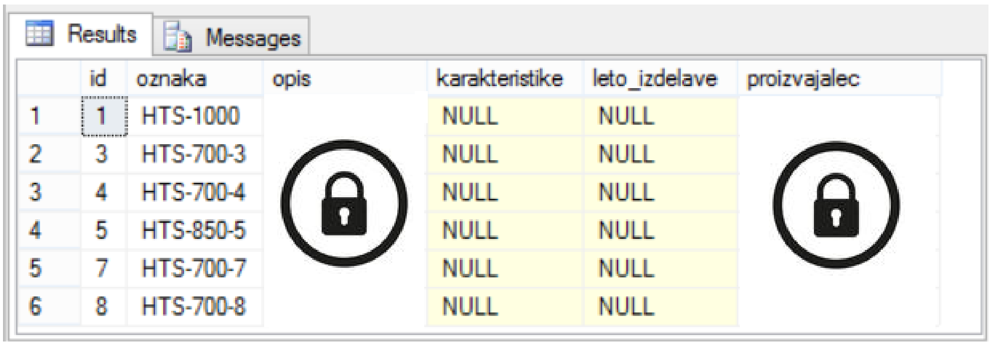
In the SQL statement above:

- The SELECT clause specifies one or more columns to be retrieved. To retrieve all columns, we use the symbol "*"(an asterisk).

- The FROM clause specifies one or more tables to be queried. We can use a comma and space between table names if we want to specify multiple tables.

- The WHERE clause selects only the rows in which the specified column contains the specified value. The value is enclosed in single quotes.(E.g., Where id tool='829.F1900').

Example of a SQL SELECT statement is shown in Table 3.4.



	id	oznaka	opis	karakteristike	leto_izdelave	proizvajalec
1	1	HTS-1000		NULL	NULL	
2	3	HTS-700-3		NULL	NULL	
3	4	HTS-700-4		NULL	NULL	
4	5	HTS-850-5		NULL	NULL	
5	7	HTS-700-7		NULL	NULL	
6	8	HTS-700-8		NULL	NULL	

Table. 3.4 Table of the machines existing in the COCAST database.

To select all columns from the table of the machines for rows where the machine id='5'(this means all the information from machine 5 that exist in the table of the machines), we must send this SELECT statement to the server back end. Figure 3.5 shows the statement that is send back to the server.

```
SELECT * FROM TBL_STROJI WHERE id='5'
```

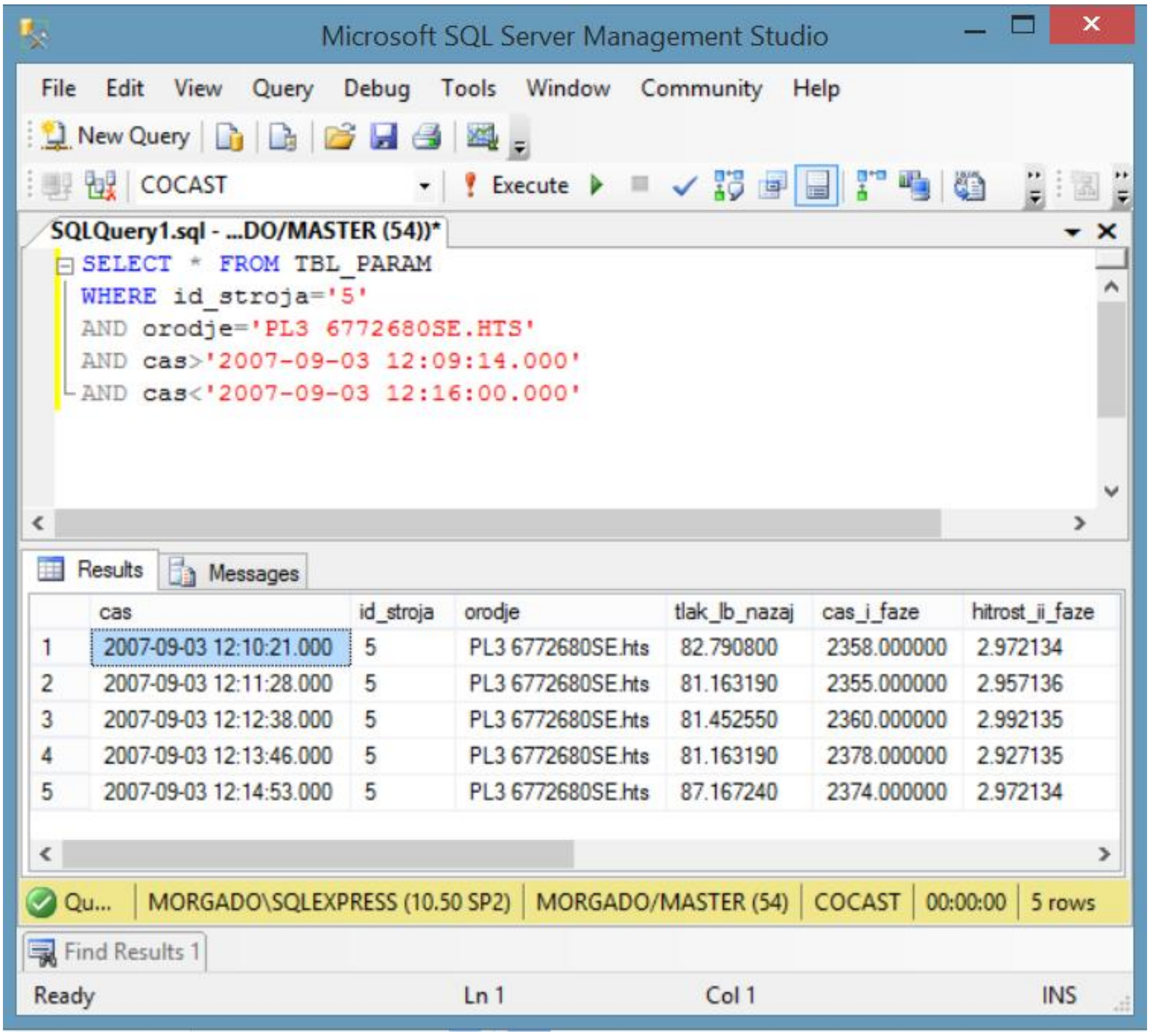
Figure. 3.5 Mathematica statement to retrieve machine 5 in the table of the machines.

The Figure 3.6 shows the selection commands that were used in the case study to select specific data sets.

The command SELECT * FROM enable us to select all the data from the table "TBL_PARAM"

The command WHERE enable us to specify the number of the machine.

The operator and enable us to specify the type of tool("orodje"), and also the start and end time when the pieces are being produced.



```
SELECT * FROM TBL_PARAM
WHERE id_stroja='5'
AND orodje='PL3 6772680SE.HTS'
AND cas>'2007-09-03 12:09:14.000'
AND cas<'2007-09-03 12:16:00.000'
```

	cas	id_stroja	orodje	tlak_lb_nazaj	cas_i_faze	hitrost_i_faze
1	2007-09-03 12:10:21.000	5	PL3 6772680SE.hts	82.790800	2358.000000	2.972134
2	2007-09-03 12:11:28.000	5	PL3 6772680SE.hts	81.163190	2355.000000	2.957136
3	2007-09-03 12:12:38.000	5	PL3 6772680SE.hts	81.452550	2360.000000	2.992135
4	2007-09-03 12:13:46.000	5	PL3 6772680SE.hts	81.163190	2378.000000	2.927135
5	2007-09-03 12:14:53.000	5	PL3 6772680SE.hts	87.167240	2374.000000	2.972134

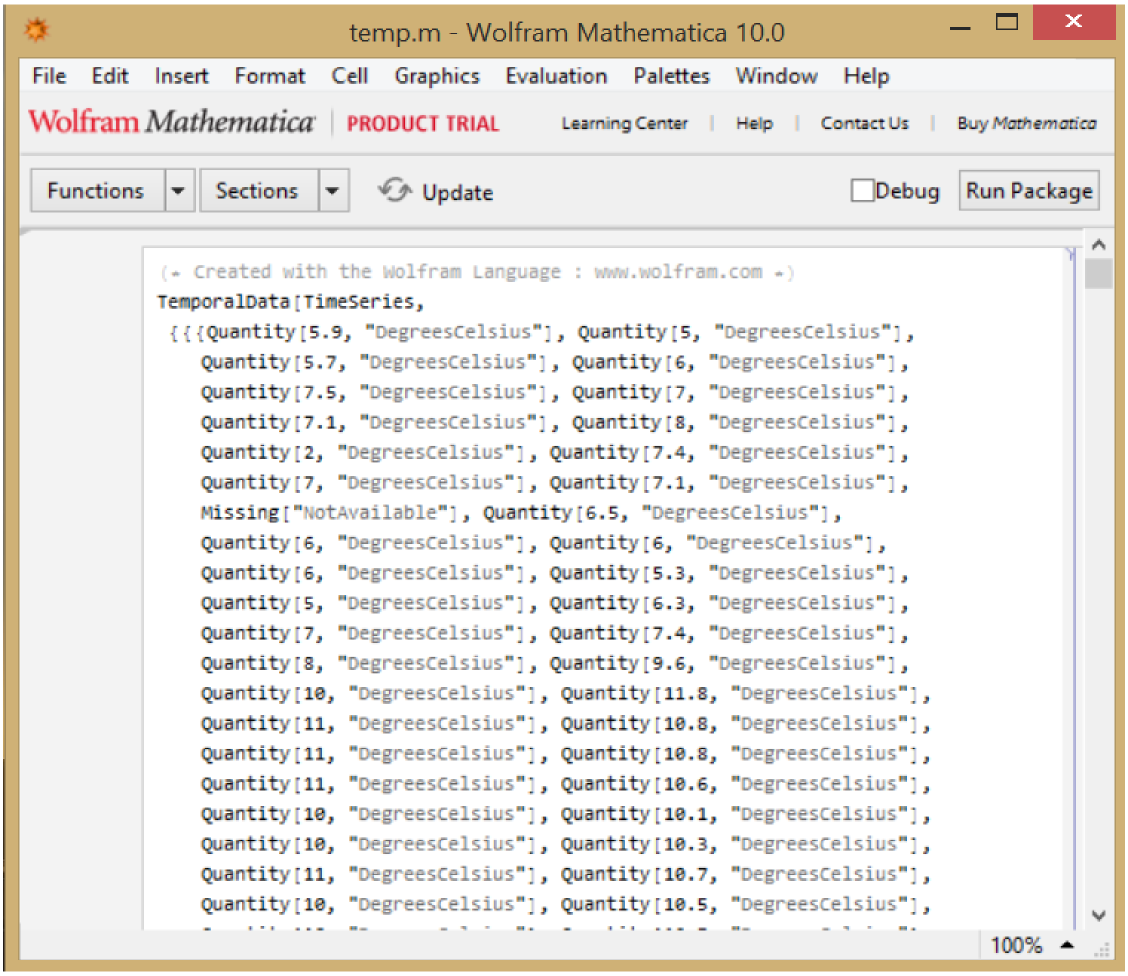
```
Qu... | MORGADO\SQLEXPRESS (10.50 SP2) | MORGADO/MASTER (54) | COCAST | 00:00:00 | 5 rows
```

Figure. 3.6 Mathematica statement to select specific data set.

3.1.5. Public Data

In terms of availability, we characterize data either as public or private. Private data is data collected for example, of a company and stored in its facilities (warehouses) and is therefore not available to people outside of the company [13]. Public data is data that people can usually access through internet or other ways without the credentials of the data owner. The amount of this sort of data is getting larger by the day. The reason for this is that there is a need for engineering tools to be developed in order to provide quick assessments and interpretations of data. Web sites like Amazon web services enable us to publish data free of charge as long as they can be available to everybody.

The weather information used in this thesis is characterized as public data, as anybody is able to access it. It gives us information about the humidity and temperature of the air at the location where the weather station was located. A part of the dataset is presented in the picture. The software Mathematica Wolfram enables the access to the public data through the internet. the Figure 3.7 shows the screen shot of Mathematica.



temp.m - Wolfram Mathematica 10.0
File Edit Insert Format Cell Graphics Evaluation Palettes Window Help
Wolfram Mathematica | PRODUCT TRIAL
Learning Center | Help | Contact Us | Buy Mathematica
Functions ▾ Sections ▾ Update
□Debug
Run Package
100%

```
(* Created with the Wolfram Language : www.wolfram.com *)
TemporalData[TimeSeries,
 {{{Quantity[5.9, "DegreesCelsius"], Quantity[5, "DegreesCelsius"],
    Quantity[5.7, "DegreesCelsius"], Quantity[6, "DegreesCelsius"],
    Quantity[7.5, "DegreesCelsius"], Quantity[7, "DegreesCelsius"],
    Quantity[7.1, "DegreesCelsius"], Quantity[8, "DegreesCelsius"],
    Quantity[2, "DegreesCelsius"], Quantity[7.4, "DegreesCelsius"],
    Quantity[7, "DegreesCelsius"], Quantity[7.1, "DegreesCelsius"],
    Missing["NotAvailable"], Quantity[6.5, "DegreesCelsius"],
    Quantity[6, "DegreesCelsius"], Quantity[6, "DegreesCelsius"],
    Quantity[6, "DegreesCelsius"], Quantity[5.3, "DegreesCelsius"],
    Quantity[5, "DegreesCelsius"], Quantity[6.3, "DegreesCelsius"],
    Quantity[7, "DegreesCelsius"], Quantity[7.4, "DegreesCelsius"],
    Quantity[8, "DegreesCelsius"], Quantity[9.6, "DegreesCelsius"],
    Quantity[10, "DegreesCelsius"], Quantity[11.8, "DegreesCelsius"],
    Quantity[11, "DegreesCelsius"], Quantity[10.8, "DegreesCelsius"],
    Quantity[11, "DegreesCelsius"], Quantity[10.8, "DegreesCelsius"],
    Quantity[11, "DegreesCelsius"], Quantity[10.6, "DegreesCelsius"],
    Quantity[10, "DegreesCelsius"], Quantity[10.1, "DegreesCelsius"],
    Quantity[10, "DegreesCelsius"], Quantity[10.3, "DegreesCelsius"],
    Quantity[11, "DegreesCelsius"], Quantity[10.7, "DegreesCelsius"],
    Quantity[10, "DegreesCelsius"], Quantity[10.5, "DegreesCelsius"],
```

Figure. 3.7 Layout Mathematica Wolfram.

3.2. Pre-processing

3.2.1. Wolfram Mathematica Software

Mathematica is a computer program, originally designed by Stephen Wolfram, and continuously being developed by the company Wolfram Research, that implements a computer algebra system [14]. In addition to a programming language, the program contains a number of programming libraries ready to be used for various purposes in various fields of exact sciences. The program comes in various areas of engineering, biological, chemical, image processing, finance, statistics, mathematics, among others, and also serves as an environment for rapid development programs. Newer versions allow the exchange of information with programs in Java, C++, among others, using libraries for communication between applications. Mathematica is based on Wolfram language which becomes its the main interface language. The language is really extensive, has contact on numerous domains and has a built-in high-level interface to all standard SQL databases that allows immediate searching, reading, and writing of arbitrary data and expressions, as well as, supporting general SQL database features, including discovery, result sets, and transactions [15]. The program also includes built-in functions that can perform mathematical computations, data analysis 3D models analysis, visualizations, numerical computations, and many others. Mathematica is not a free software. It is a proprietary software licenser so we can have access depending on the intended use (commercial, educational, public administration, among others).

Mathematica software need a database connectivity to connect SQL server to be able to have access to all the data provided by the company that we specify in SQL software Management Studio. This connection is made through Java Database Connectivity (JDBC). The application program interface allow us to encrypt access request statements in Structure Query Language(SQL) that later is passed to the program that manages the database. This means that, when we access a database on a PC, the Mathematica software will have access to the database through the JDBC statements that make the bridge between Mathematica and SQL server management studio [16].

The Figure 3.8 shows the JDBC statement that was used to connect Mathematica to SQL Server Management Studio.

```
Needs["JLink`"]
Needs["DatabaseLink`"]
LoadJavaClass["java.lang.Runtime"];
java`lang`Runtime`getRuntime[]@maxMemory[];
ReinstallJava[JVMArguments -> "-Xmx64g"];
LoadJavaClass["java.lang.Runtime"];
java`lang`Runtime`getRuntime[]@maxMemory[];
conn = OpenSQLConnection[JDBC["Microsoft SQL Server(jTDS)", "127.0.0.1"],
"Catalog" -> "COCAST", "Username" -> "MORGADO/MASTER", "Password" -> "benfica"]
```

Figure. 3.8 JDBC statement.

After the connection is made between those two programs we already able to upload data into Mathematica. To do that we need to use various functions provided by Mathematica software. The Figure 3.9 shows the statement to upload the dataset when the pieces are being produced.

```
query = "Select * from TBL_PARAM
where id_stroja=3
";
data = SQLExecute[conn, query];
times = #[[1, 1]] & /@ data
```

Figure. 3.9 Mathematica statement.

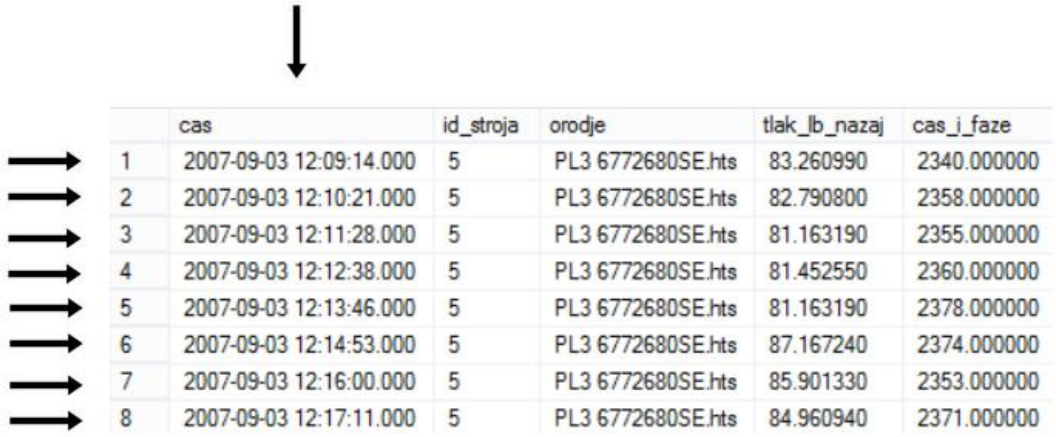
The command query enable us to retrieve the dataset from the program SQL server management studio. The command data ask the connection between Mathematica and SQL server management and the request dataset.

The command times shown in Figure 3.10 able us to specify the dataset , in this case:

The "@" means we want all the elements from the data and #[1,1] specify this elements for only the first element of the first column like is show in the Table 3.5.

```
times = #[[1, 1]] & /@ data
```

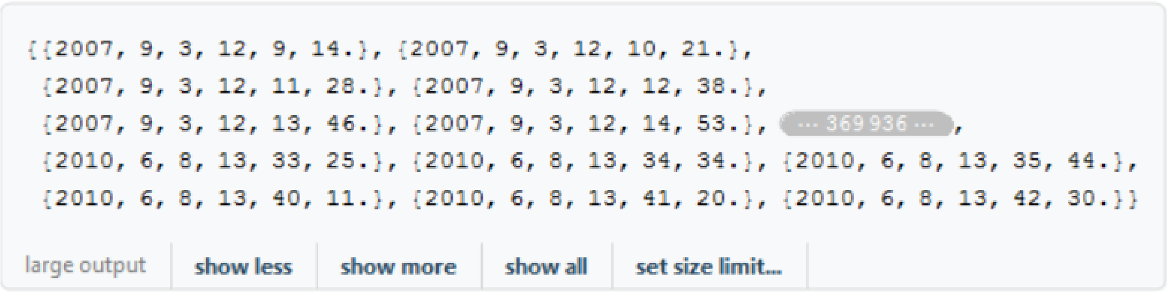
Figure. 3.10 Mathematica statement.



	cas	id_stroja	orodje	tlak_lb_nazaj	cas_i_faze
1	2007-09-03 12:09:14.000	5	PL3 6772680SE.hts	83.260990	2340.000000
2	2007-09-03 12:10:21.000	5	PL3 6772680SE.hts	82.790800	2358.000000
3	2007-09-03 12:11:28.000	5	PL3 6772680SE.hts	81.163190	2355.000000
4	2007-09-03 12:12:38.000	5	PL3 6772680SE.hts	81.452550	2360.000000
5	2007-09-03 12:13:46.000	5	PL3 6772680SE.hts	81.163190	2378.000000
6	2007-09-03 12:14:53.000	5	PL3 6772680SE.hts	87.167240	2374.000000
7	2007-09-03 12:16:00.000	5	PL3 6772680SE.hts	85.901330	2353.000000
8	2007-09-03 12:17:11.000	5	PL3 6772680SE.hts	84.960940	2371.000000

Table. 3.5 Selecting data in SQL server management studio.

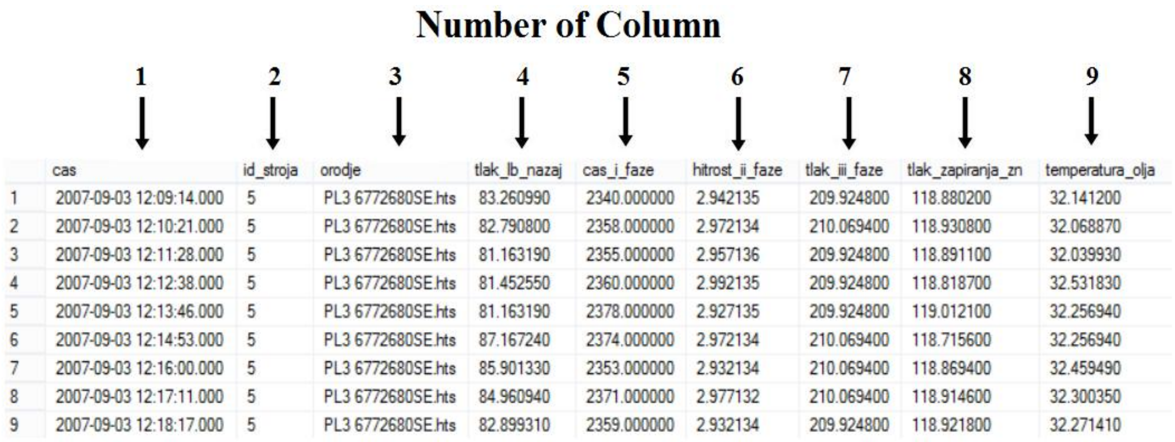
The output of the previous statement is shown in the Figure 3.11.



```
{{2007, 9, 3, 12, 9, 14.}, {2007, 9, 3, 12, 10, 21.},
 {2007, 9, 3, 12, 11, 28.}, {2007, 9, 3, 12, 12, 38.},
 {2007, 9, 3, 12, 13, 46.}, {2007, 9, 3, 12, 14, 53.}, ··· 369 936 ···,
 {2010, 6, 8, 13, 33, 25.}, {2010, 6, 8, 13, 34, 34.}, {2010, 6, 8, 13, 35, 44.},
 {2010, 6, 8, 13, 40, 11.}, {2010, 6, 8, 13, 41, 20.}, {2010, 6, 8, 13, 42, 30.}}
```

Figure. 3.11 Output of the previous Mathematica statement.

From the picture, we can see that now we have a dataset only with all the elements of the first column. In the data analysis, we select the data from various parameters of the manufacturing process. To select the data and specify the parameters we first needed to check which column was the request parameter. The following picture shows how we were doing that:



	cas	id_stroja	orodje	tlak_lb_nazaj	cas_i_faze	hitrost_ii_faze	tlak_iii_faze	tlak_zapiranja_zn	temperatura_olja
1	2007-09-03 12:09:14.000	5	PL3 6772680SE.hts	83.260990	2340.000000	2.942135	209.924800	118.880200	32.141200
2	2007-09-03 12:10:21.000	5	PL3 6772680SE.hts	82.790800	2358.000000	2.972134	210.069400	118.930800	32.068870
3	2007-09-03 12:11:28.000	5	PL3 6772680SE.hts	81.163190	2355.000000	2.957136	209.924800	118.891100	32.039930
4	2007-09-03 12:12:38.000	5	PL3 6772680SE.hts	81.452550	2360.000000	2.992135	209.924800	118.818700	32.531830
5	2007-09-03 12:13:46.000	5	PL3 6772680SE.hts	81.163190	2378.000000	2.927135	209.924800	119.012100	32.256940
6	2007-09-03 12:14:53.000	5	PL3 6772680SE.hts	87.167240	2374.000000	2.972134	210.069400	118.715600	32.256940
7	2007-09-03 12:16:00.000	5	PL3 6772680SE.hts	85.901330	2353.000000	2.932134	210.069400	118.869400	32.459490
8	2007-09-03 12:17:11.000	5	PL3 6772680SE.hts	84.960940	2371.000000	2.977132	210.069400	118.914600	32.300350
9	2007-09-03 12:18:17.000	5	PL3 6772680SE.hts	82.899310	2359.000000	2.932134	209.924800	118.921800	32.271410

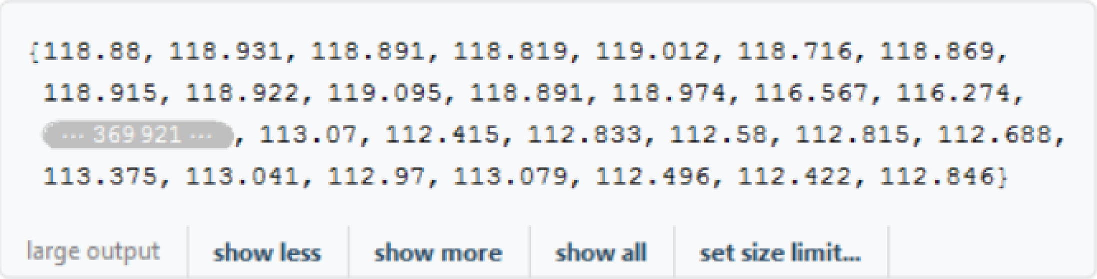
Table. 3.6 Selection of the parameters.

After we knew the number of the column of the parameter we just use the following statement(In this case we specified for the parameter closing pressure), Is shown on the FIgure 3.12.

```
closingpressure = #[[8]] & /@ data
```

Figure. 3.12 Mathematica statement.

The output of the statement is shown in Figure. 3.13, which give us the Closing pressure of each piece that was produced.

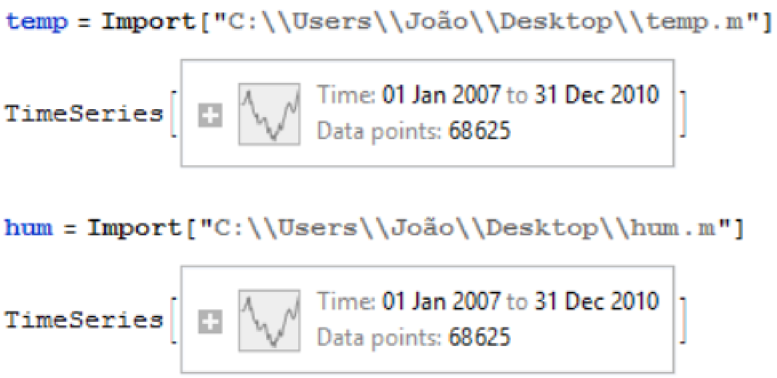


```
{118.88, 118.931, 118.891, 118.819, 119.012, 118.716, 118.869,
 118.915, 118.922, 119.095, 118.891, 118.974, 116.567, 116.274,
 ··· 369 921 ···, 113.07, 112.415, 112.833, 112.58, 112.815, 112.688,
 113.375, 113.041, 112.97, 113.079, 112.496, 112.422, 112.846}
```

Figure. 3.13 Output of the previous Mathematica statement.

To be able to select the data of the other parameters we just need to change the number inside the brackets according the column or, in this case, the parameter desired in the previous table.

To get the data from public sources we use the command Import. This command handles with a large number of formats, each typically with many different possible elements. The public data in our study are the temperature and humidity outside of the manufacturing local during a certain period of time. The data were uploaded in Mathematica with the following statement shown in the Figure 3.14.



```
temp = Import["C:\\Users\\João\\Desktop\\temp.m"]

TimeSeries[ Time: 01 Jan 2007 to 31 Dec 2010
            Data points: 68625 ]

hum = Import["C:\\Users\\João\\Desktop\\hum.m"]

TimeSeries[ Time: 01 Jan 2007 to 31 Dec 2010
            Data points: 68625 ]
```

Figure. 3.14 Mathematica statement.

The temperature and humidity data are uploaded in Mathematica in a form of Time series. Time Series represents a series of time-value, which in our case represents the temperature or humidity over a period of time depending on what we want. The following image represents the visualization of the Time Series of the temperature for a period of 4 years. Each Data point represents a temperature for a specific time.

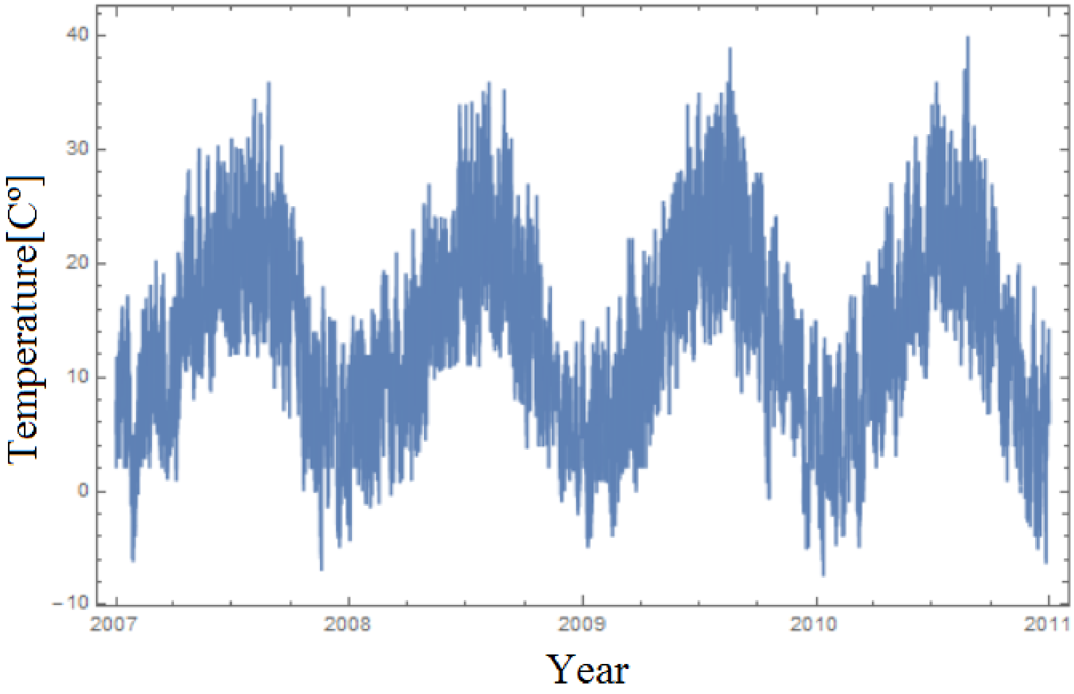


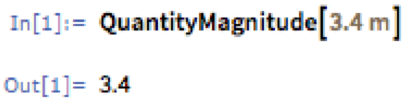
Figure. 3.15 Time Series of the temperature for a period of 4 years.

3.2.2. Mash up

In this chapter we will explain how we combine information to visualize and detect the hidden patterns in the 3D plots. After we upload all the data from different sources into the Mathematica program, the next step was to combine all this data. In order to do this we need to have all data points of the parameters in study to do the plots and to be able to do that all the parameters needed to be in the same format. The parameters did not have the same format because they did not have the same unit. The parameters that we combine were: outside temperature, outside humidity, closing pressure, back pressure, and oil temperature. To be able to change the parameters for all the same format we use the function Quantity Magnitude in Mathematica program.

This function able us to have all the data from any parameter in the same format, in this case it gives us the magnitude of a specific quantity [17].

The Figure 3.16 shows one example where the unit is in meters:



```
In[1]:= QuantityMagnitude[3.4 m]

Out[1]= 3.4
```

Figure. 3.16 Mathematica function.

After we have all the parameters in the same format we use the statement shown in Figure 3.17 to cross the information about the parameters:

```
dataset =
 Cases[
{QuantityMagnitude[temp[#[[1, 1]]]], #[[8]]} -> qrtime[rates, #[[1, 1]]] & /@ data, x_ /; x[[2]] != 0]
```

Figure. 3.17 Mathematica statement with information between the temperature, the closing pressure, and the quality rate.

Since the quality rate was not part of the data set but we use to combine the parameters we will explain how we get that information in the next chapter with a more in-depth explanation (Algorithm chapter). In the Mathematica chapter we explain the following statement that was the time of when the pieces were being produced. Shown in Figure 3.18.

```
times = #[[1, 1]] & /@ data
```

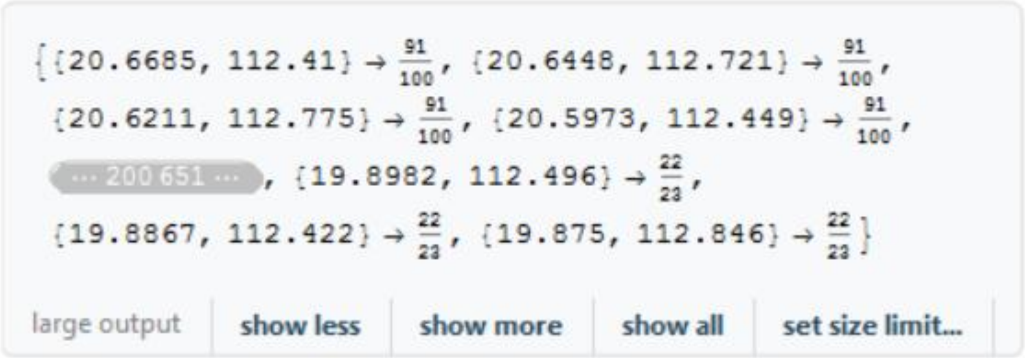
Figure. 3.18 Mathematica statement.

So to be able to have the outside temperature of when a piece was produced we use the statement shown in Figure 3.19.

```
temp[#[[1, 1]]]
```

Figure. 3.19 Mathematica statement.

The statement take the values of the outside temperature taking account the variable times, this means the outside temperature for when each piece is being produced and not for a random time. The other part of the statement that refers the number #[8] that is the closing pressure is explained in the previous chapter. The following picture show us the output of this combination of parameters:



```
{{20.6685, 112.41} → 91/100, {20.6448, 112.721} → 91/100,
 {20.6211, 112.775} → 91/100, {20.5973, 112.449} → 91/100,
 ⋯ 200651 ⋯, {19.8982, 112.496} → 22/23,
 {19.8867, 112.422} → 22/23, {19.875, 112.846} → 22/23}
```

Figure. 3.20 Output of combination of parameters.

So we can see in the Figure 3.20 the values of the outside temperature, closing pressure and quality rate for a specific time.

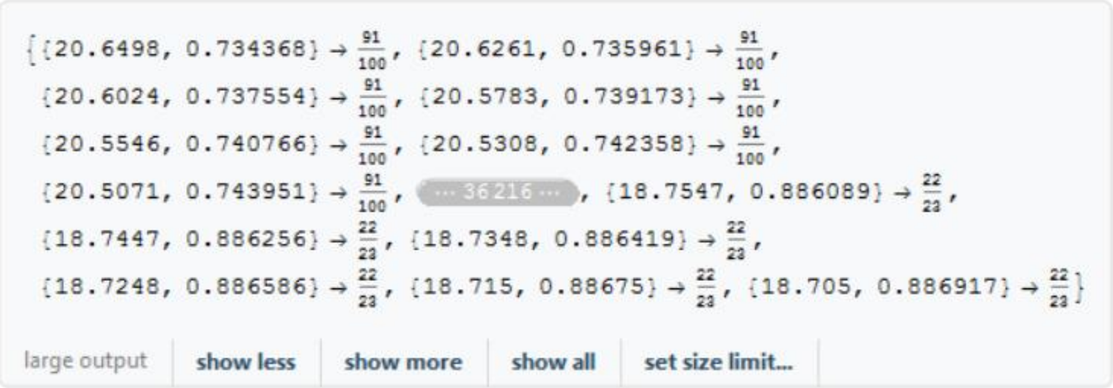
In the Figure 3.21 we can see another kind of combination between the data, in this case it is temperature, humidity and quality rate.

```
dataset =

 Cases[{QuantityMagnitude[temp[#]], hum[#]} -> qrtime[rates, #] & /@
 times, x_ /; x[[2]] != 0]
```

Figure. 3.21 Mathematica statement.

The Figure 3.22 shows the Output of the previous Mathematica statement.



```
{{20.6498, 0.734368} → 91/100, {20.6261, 0.735961} → 91/100,
 {20.6024, 0.737554} → 91/100, {20.5783, 0.739173} → 91/100,
 {20.5546, 0.740766} → 91/100, {20.5308, 0.742358} → 91/100,
 {20.5071, 0.743951} → 91/100, ··· 36216 ···, {18.7547, 0.886089} → 22/23,
 {18.7447, 0.886256} → 22/23, {18.7348, 0.886419} → 22/23,
 {18.7248, 0.886586} → 22/23, {18.715, 0.88675} → 22/23, {18.705, 0.886917} → 22/23}
```

Figure. 3.22 Output of the previous Mathematica statement.

3.3. Inferring the quality rate from incomplete data

The quality rate was not part of the data set, this means we didn´t have any information, whether a certain piece was well made or not. To work on that problem, we first looked at the timeline when the pieces were being produced, as shown in the Figure 3.23.

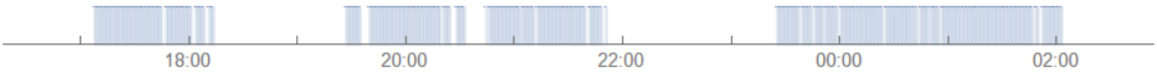


Figure. 3.23 Timeline of when the pieces were being produced.

After looking to the timeline of the pieces that are being produce we realize that quality rate is not a property of a specific piece, but a property of multiple pieces. This happens because the unit of the quality rate comes in percent of pieces so it could not be a property of a specific piece. To get the quality rate we first draw a point in the middle of each of the interval of pieces produced and we said the quality rate was the same for the pieces that were being produced in that specific interval. The Figure 3.24 illustrates this:

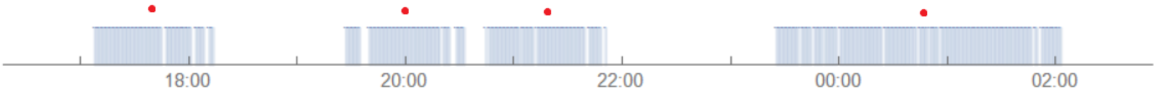


Figure. 3.24 Timeline of when the pieces were being produced.

The problem ocurred once we realized that the actual quality rate changed during the interval. So we said, that if we draw a line between two points it means that not all the pieces have the same quality rate, even within the interval. The Figure 3.25 illustrates this:

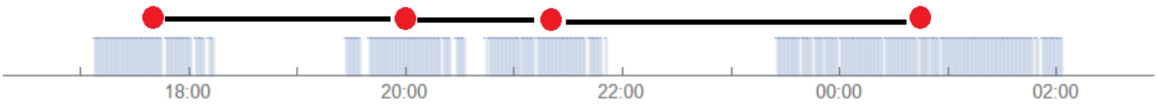


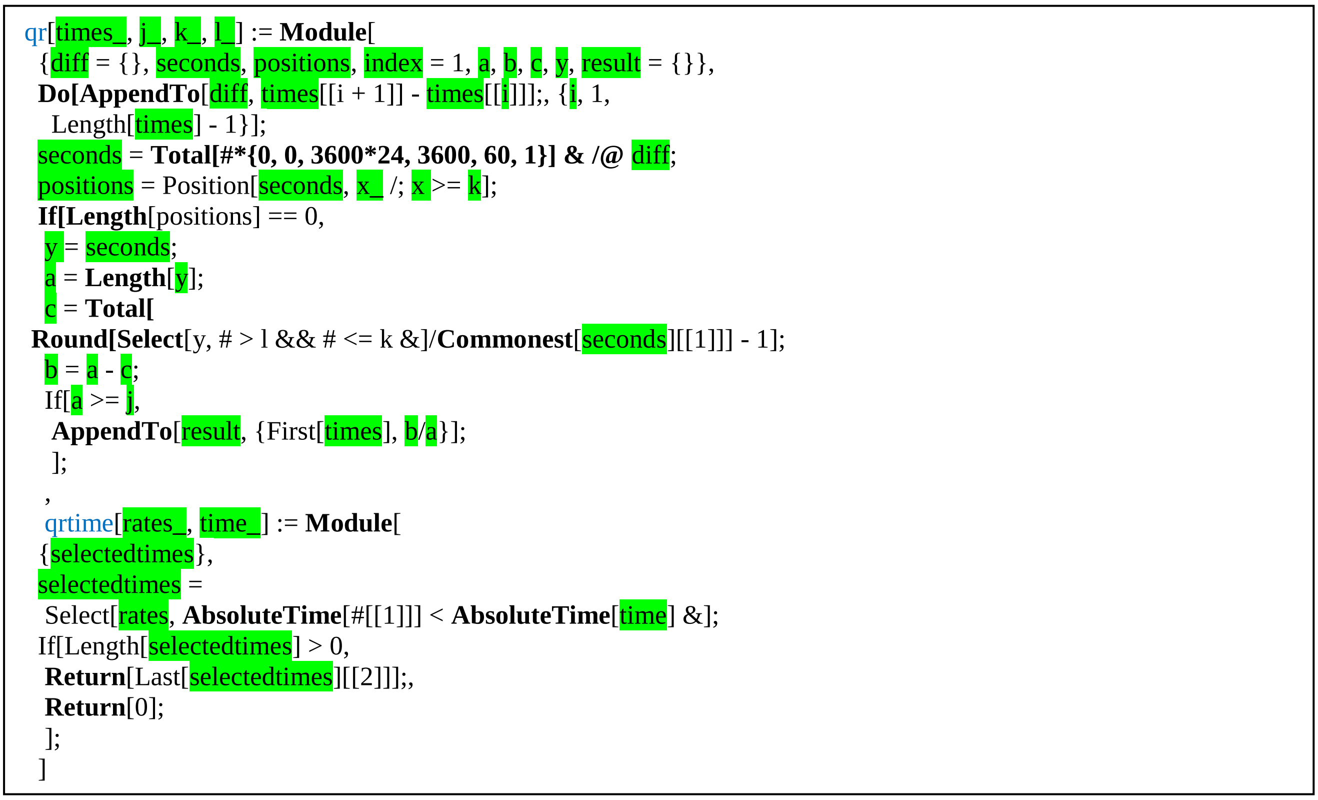
Figure. 3.25 Timeline of when the pieces were being produced.

We did that because it was easier to program, but in the final, it did not make much sense. Thus, the solution was that if we write a time of a piece the code will return the quality rate of the interval since we assume that the quality rate of a certain piece is the quality rate of the interval of each piece was produced. It is actually difficult to talk about the quality rate of one piece so we decide that the quality rate of a piece is the quality rate of the interval which was produced.

3.4. Algorithm

The algorithm is not necessarily a computer program, but the steps needed to perform a task. A properly executed algorithm will not solve a problem if is implemented in the incorrect way or if it is not suitable for the problem. Its implementation can be performed by a computer, another type of robot or even a human. Different algorithms can perform the same task using a different set of instructions in more or less time, space, or effort than others [18]. This different set of instructions may reflect the computational complexity applied, which depends on appropriate data structures to the algorithm. In this paper, we develop an algorithm that was able to give us the information about the quality rate of a certain amount of pieces that were being produced. In Figure 3.23 we will explain how we achieve this.

The following pictures show us part of the developed algorithm and the main loop flowchart and the QR calculation flowchart to better understanding the algorithm for the reader.



```
qr[times_, j_, k_, l_] := Module[
  {diff = {}, seconds, positions, index = 1, a, b, c, y, result = {}},
  Do[AppendTo[diff, times[[i + 1]] - times[[i]]];, {i, 1,
    Length[times] - 1}];
  seconds = Total[#*{0, 0, 3600*24, 3600, 60, 1}] & /@ diff;
  positions = Position[seconds, x_ /; x >= k];
  If[Length[positions] == 0,
   y = seconds;
   a = Length[y];
   c = Total[
 Round[Select[y, # > l && # <= k &]/Commonest[seconds][[1]]] - 1];
   b = a - c;
   If[a >= j,
    AppendTo[result, {First[times], b/a}];
    ];
   ,
   qrtime[rates_, time_] := Module[
  {selectedtimes},
  selectedtimes =
   Select[rates, AbsoluteTime[#[[1]]] < AbsoluteTime[time] &];
  If[Length[selectedtimes] > 0,
   Return[Last[selectedtimes][[2]]];,
   Return[0];
   ];
  ]
```

Figure. 3.26 Part of the developed algorithm.

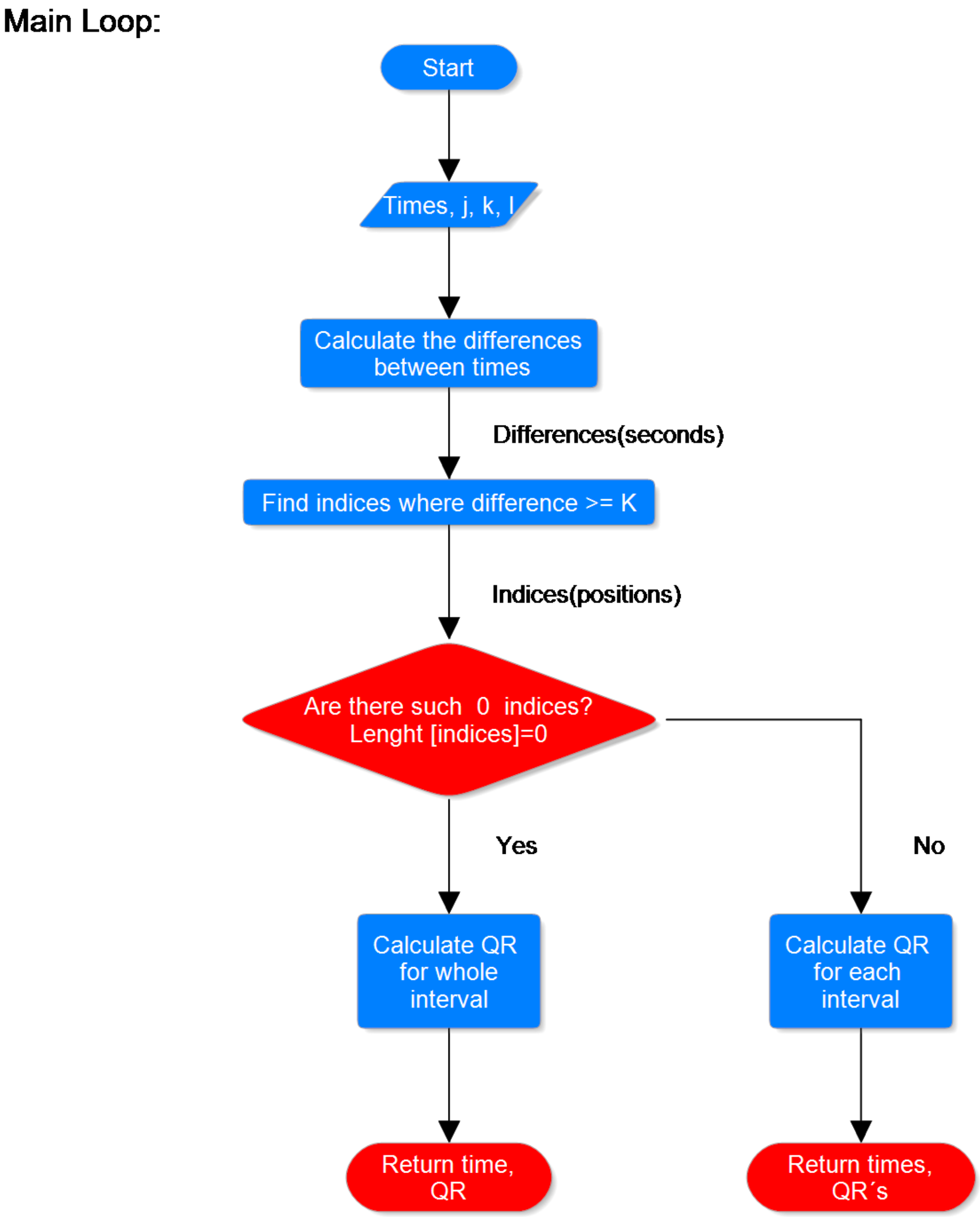

Main Loop:
Start
Times, j, k, l
Calculate the differences between times
Differences(seconds)
Find indices where difference >= K
Indices(positions)
Are there such 0 indices? Lenght [indices]=0
Yes
No
Calculate QR for whole interval
Calculate QR for each interval
Return time, QR
Return times, QR´s

Figure. 3.27 Flow chart of the main loop of the algorithm [19].

The main loop represents the outer program. When it starts, it gets the 4 inputs (Times, j, k, l) and converts the differences between the times to seconds. The next step is to find indices or evidences? where the differences are equal or bigger than the input k. The number of indices depends on how many intervals exist. If the number of intervals is equal to zero, the calculation of the QR function is made for the whole interval and it just returns one quality rate. If the number of intervals is different than zero the calculation of the QR function is made. Yet, in this case the calculation is for each interval. After that, it return the times and QR for each interval.

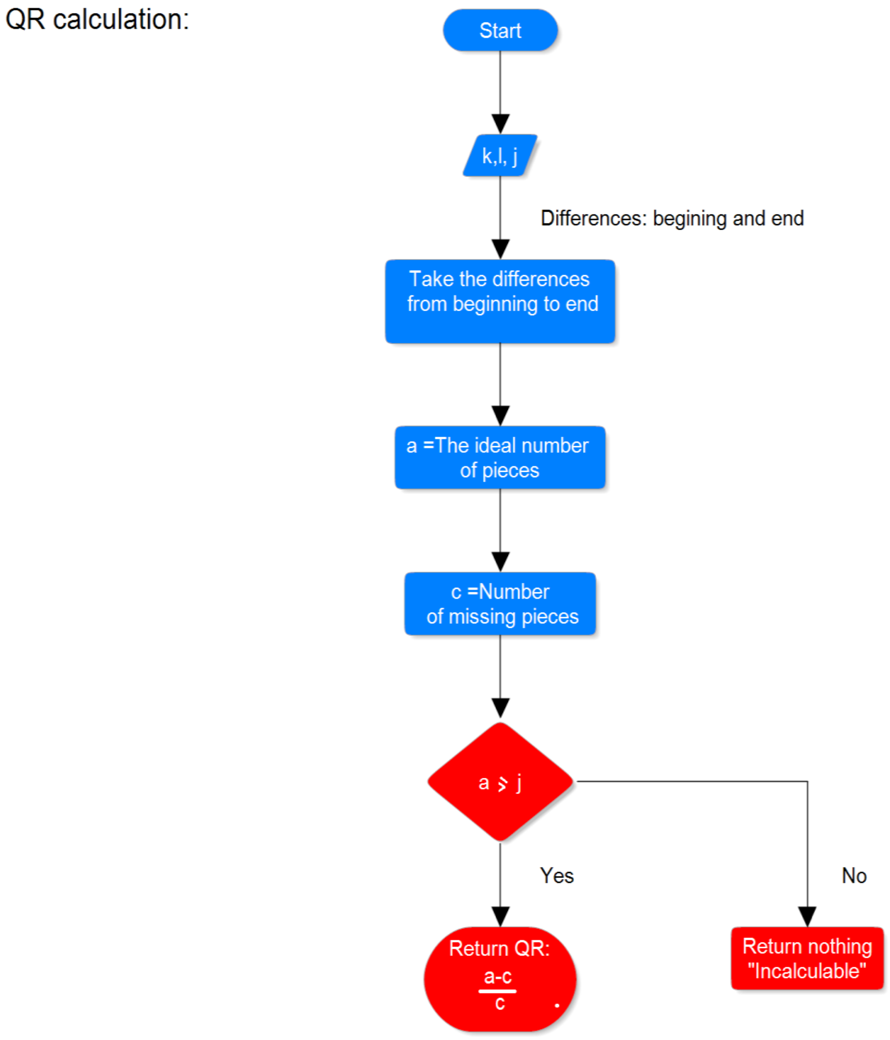

QR calculation:
Start
k,l, j
Differences: begining and end
Take the differences from beginning to end
a =The ideal number of pieces
c =Number of missing pieces
a ⩾ j
Yes
No
Return QR: a-c / c
Return nothing "Incalculable"

Figure. 3.28 Flow chart of the quality rate calculation of the algorithm [19].

In the QR (quality rate) flowchart calculation of the entries are k, j l. Differences are calculated between the beginning and the end of each interval and then calculated the optimum number of pieces followed by the calculation of the number of missing pieces.

If there are not enough pieces this means that "a" is not bigger or equal than "j" then the quality rate is incalculable, so return nothing. If "a" is bigger or equal than "j" the calculation is made for the ideal number of pieces less the number of missing pieces and after divided by the number of ideal pieces where it will give us the QR.

3.5. Evaluation and Interpretation

The point of the work was not only try to calculate the correlations that we can discover in the data, since there are plenty of methods that can do that. The important goal was to see which kind of information is left to the observer when there is no correlation calculation of the variables since part of that correlation was made directly from the plots and not using the proper correlation method. To evaluate a plot we need to take into account the correlation and dependence on that plot or graph.

Dependence is any mathematical relationship between two random variables or two sets of data. It refers to any kind of situation in which random variables do not satisfy a mathematical condition of a probabilistic independence. This probabilistic independence happens when two events are independent, this means that two random variables are independent if the realization of one does not affect the probability distribution of the other [20].

Correlation mention to any of a broad class of statistical relationships involving dependence. It can indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. In this example, there is a casual relationship, because for extreme weather people may use more electricity for heating or cooling, nevertheless, statistical dependence is not sufficient to demonstrate the presence of such causal relationship. This means that correlation cannot be used to infer a casual relationship between the variables. This previous sentence should not be taken that correlations cannot indicate the potential existence of causal relations. However, the causes underlying the correlation, if any, may be indirect and unknown, and high correlations also overlap with identity relations, where no causal process exists. There are several correlation coefficients that are able to measure the degree of correlation. The most common of these is the Pearson correlation coefficient shown in the following picture (Several sets of (x, y) points, with the correlation coefficient of x and y for each set).

The Pearson correlation coefficient is a measure of the linear correlation (dependence) between two variables X and Y, giving a value between +1 and -1 inclusive, where 1 is total positive correlation, 0 is no correlation, and -1 is a total negative correlation. This is shown in the Figure 3.29.

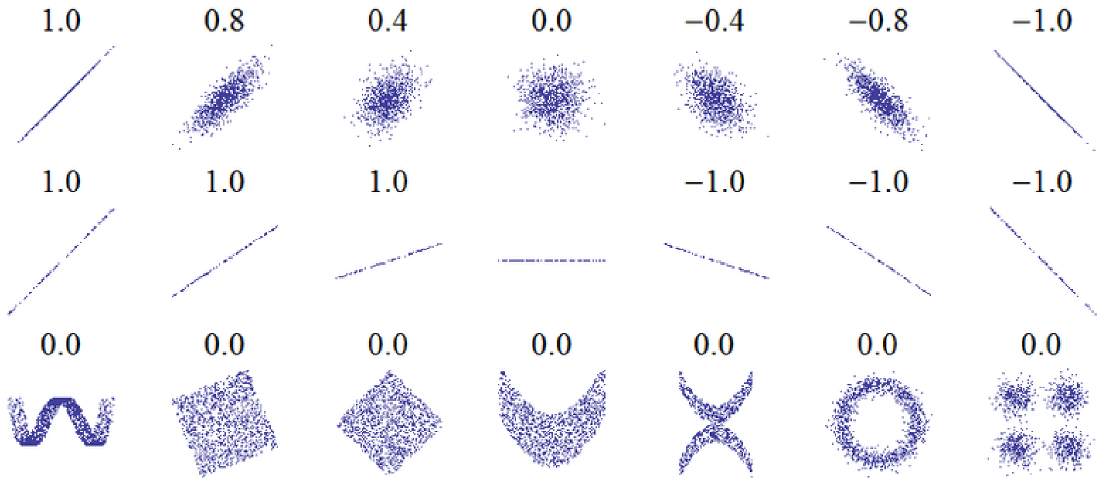


Figure. 3.29 Different values for the Pearson Correlation Coefficient [21].

The correlation methods are in a way good to detect inner correlations but not good enough to detect more complicate different kinds of connections between variables.

Therefore, if the data is linear the correlation can be calculated very easy, but if is less correlated like the noisy sign wave, the correlation coefficient is zero so it´s impossible to detect the nonlinear relation between the variables

3.6. Data Visualization

Data visualization is seen by many fields as a modern equivalent of visual communication. Is not directed to one specific field, but rather finds interpretation across many of them. It evolves the development and study of the visual representation of data, meaning "information that has been abstracted in some schematic form, including attributes or variables for the units of information" [22].

The main goal of data visualization is to communicate information clearly and easily to users via the statistical graphics, plots, information graphics, tables and charts selected. Effective visualization helps users in analysing and reasoning about data and evidence. It makes the people seeing the things in a way that before were not obvious for them. Even when data volumes are very large, patterns can be spotted quickly and easily. The main thing of visualizations is that can transport information in a universal manner and make it simple to share ideas with others. It lets people ask other, "Do you see what I see?" And it can even answer questions like "What would happen if we made an adjustment to that area?". In our case we plot the different combination of parameters to be able to visualize the data in a 3D plot. The Figure 3.30 shows us the statement that was used to plot the data:

```
ListPointPlot3D[{#[[1, 1]], #[[1, 2]], #[[2]]} & /@ dataset,
 PlotRange -> {All, All, {0.5, 1}}, ColorFunction -> "RedGreenSplit"]
```

Figure. 3.30 Mathematica statement.

The object "#[1,1]" represents the temperature in the data set that we specified.

The object "#[1,2]" represents the humidity in the data set and the object"#[2]" represents the quality rate. To be able to understand the statement, we also need to show the statement that combine the information of the parameters before being plotted. Is showed in the Figure 3.31.

```
dataset =

Cases[{QuantityMagnitude[temp[#]], hum[#]} -> qrtime[rates, #] & /@
times, x_ /; x[[2]] != 0]
```

Figure. 3.31 Mathematica statement.

3.6.1. 3D Plot

The 3D plot is used to plot data on three axes in the attempt to show the relationship between three variables. The relationship between different variables is called correlation. If the points in the 3D plot are close to make a straight line in any direction in the three-dimensional space of the 3D plot, the correlation between the corresponding variables is high. If the point or lines in the plot are equally distributed in the 3D plot, the correlation is low, or zero. However, even thought a correlation may seem to be present, this might not always be the case. We can change how the 3D plot is viewed by zooming in and out as well as rotating it by using the Mathematica controls.

Example: In the 3D plot of Figure 3.23, time (xxx), temperature (yyy) and quality rate(zzz) are plotted.

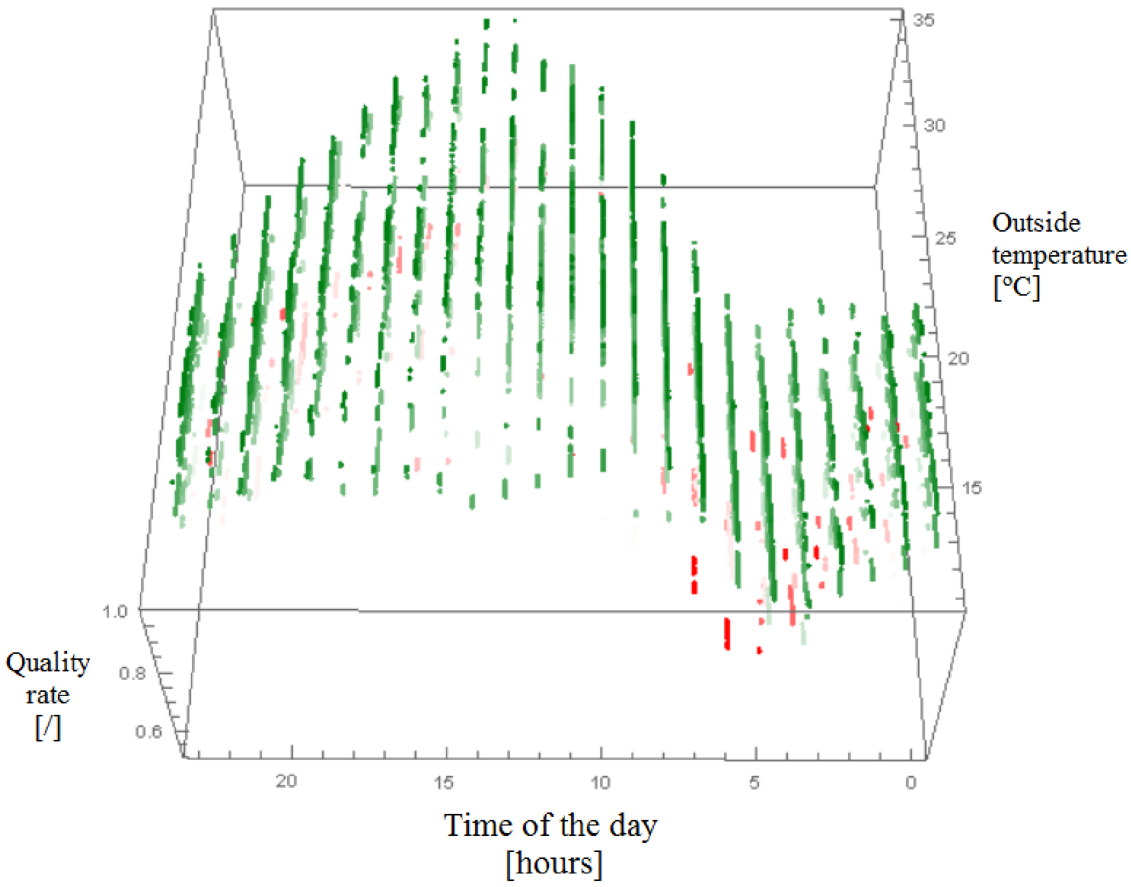


Figure. 3.32 3D plot.

We can see from the plot that the data is highly correlated due to all the lines. In the plot we use a Mathematica command called ColorFunction. ColorFunction is an option for graphics functions that specifies a function to apply to determine colours of elements [23]. In other words, this command able us to color the lines on the plot. This means higher values of the quality rate will be close to green, and the lower values of the quality rate will be close to red. In this way we can have a different kind of visualization and also able us to have a better analysis of the plot.

4. CASE STUDY

4.1. Brief description of the factory

The large amounts of data used in this research were collected by a company dealing with a process called die casting. For the purpose of this paper the company shall remain unnamed.

The company has worked for a long time in this kind of business and also in the development of knowledge of engineering staff and up-date information technology are highly valuable for the design and manufacturing of products for all the fields in which the company works.

The majority of the products they produce are used in the automotive industry and the largest business unit is Turbo, where still remain one of the leading manufacturers in Europe.

This product is introduced in the car engines of all leading automotive manufacturers, including Volkswagen, GM, Opel, Ford, Daimler, Fiat, PSA, Audi, BMW. The product key of the Turbo business unit is turbo compressor housings, back plates, inserts, turbine housings, central housings and nozzles. The other part of the business unit is on the chassis and car-body that represents almost one third of the total sales of the automotive division.

The company has a fully integrated level on quality. Each employee represents a key factor in the performance of activities within a particular operating process. There are few machines from a Slovenian company that are all located in the same large space. The level of satisfaction is expressed through the operation process and taking account the customer needs.

4.2. Die Casting process

Die casting is a process that was developed in the 1900's. This process is often used in the production of automotive parts, machine components, industrial telecommunications equipment and tools. It enables us to produce parts with good surfaces, dimensional accuracy and very thin walls. Because of the high cost of the tool, it is usually used for large production quantities (over 50 000 parts). High-pressure die casting technology is really complex in the practise area and that is the crucial point to set-up the parameters of the HPDC process to acquire good mechanical properties and the right performance for the manufactured aluminium-alloy components. According to ref , the quality of a die pressure casting is the result of a great number of parameters. Some of these parameters are controllable, and others are disturbance factors. A simplified scheme of the machine is shown in Figure 4.1. Before the die casting process can begin the tool for a specific part must be produced. It consists of two halves, which can open and close to enable us removal of the produced part. Die casting machines are generally large and strong, due to the large forces used to hold the two halves of the tool together during the process. The process is composed of five main phases. First molten metal is loaded into the chamber and pushed towards the gate with low speed. The second phase is forcing molten metal into the casting cavity with high speed. Thirdly, high pressure is ensured and maintained. The fourth is the solidifying phase. And lastly the fifth phase is ejection from the mould and spraying the mould with lubricant [24].

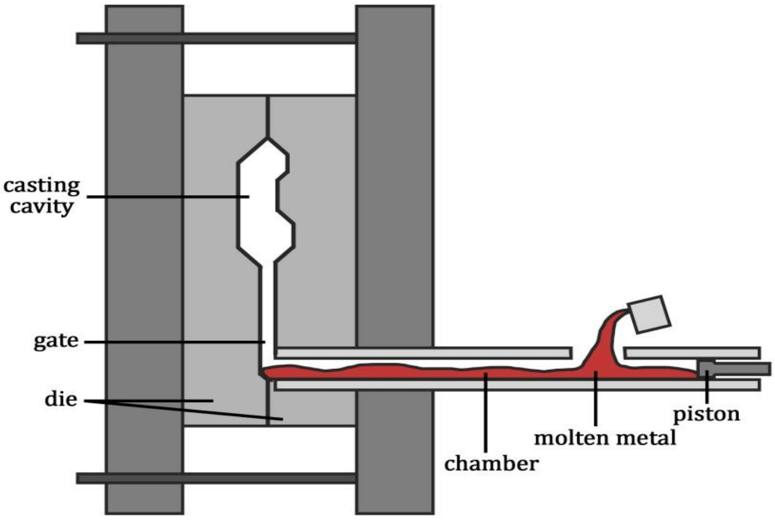


Figure. 4.1 Die casting scheme [24].

4.2.1. Parameters of Die casting process

The quality of a die pressure casting is the result of a great number of parameters. Some of these parameters are controllable and other are disturbance factors. During the operations, each die casting process cycle is monitored and recorded on the cells. Each cycle record consists on the values related to input process and the output process parameters [1].

The input process parameters are:

- Time of phase 1
- Velocity of phase 2
- Pressure at phase 2
- Pressure at phase 3
- Reaction time of phase 3

The output process parameters are:

- Pressure at return
- Pressure at closing
- Temperature of the oil
- Thickness of tablet

The scheme in Figure 4.2 shows the measuring points in a die casting cell where the input and output process parameters are acquired. The subset of parameters we focused on consisted of pressure at closing, pressure at return and the temperature of the oil. The other set of data we used were information on weather, where the main parameters were temperature and humidity. Pressure at closing consists in the pressure needed to maintain the two die halves attached inside the die to keep it securely closed and securely clamped together. Sufficient force must be applied to the die to keep it securely closed while the metal is injected. The back pressure is the pressure that injects the molten metal into the die. This pressure holds the molten metal in the dies during solidification. The temperature of the oil provides very high cooling capacities to maintain the temperature of the die at the optimum level.

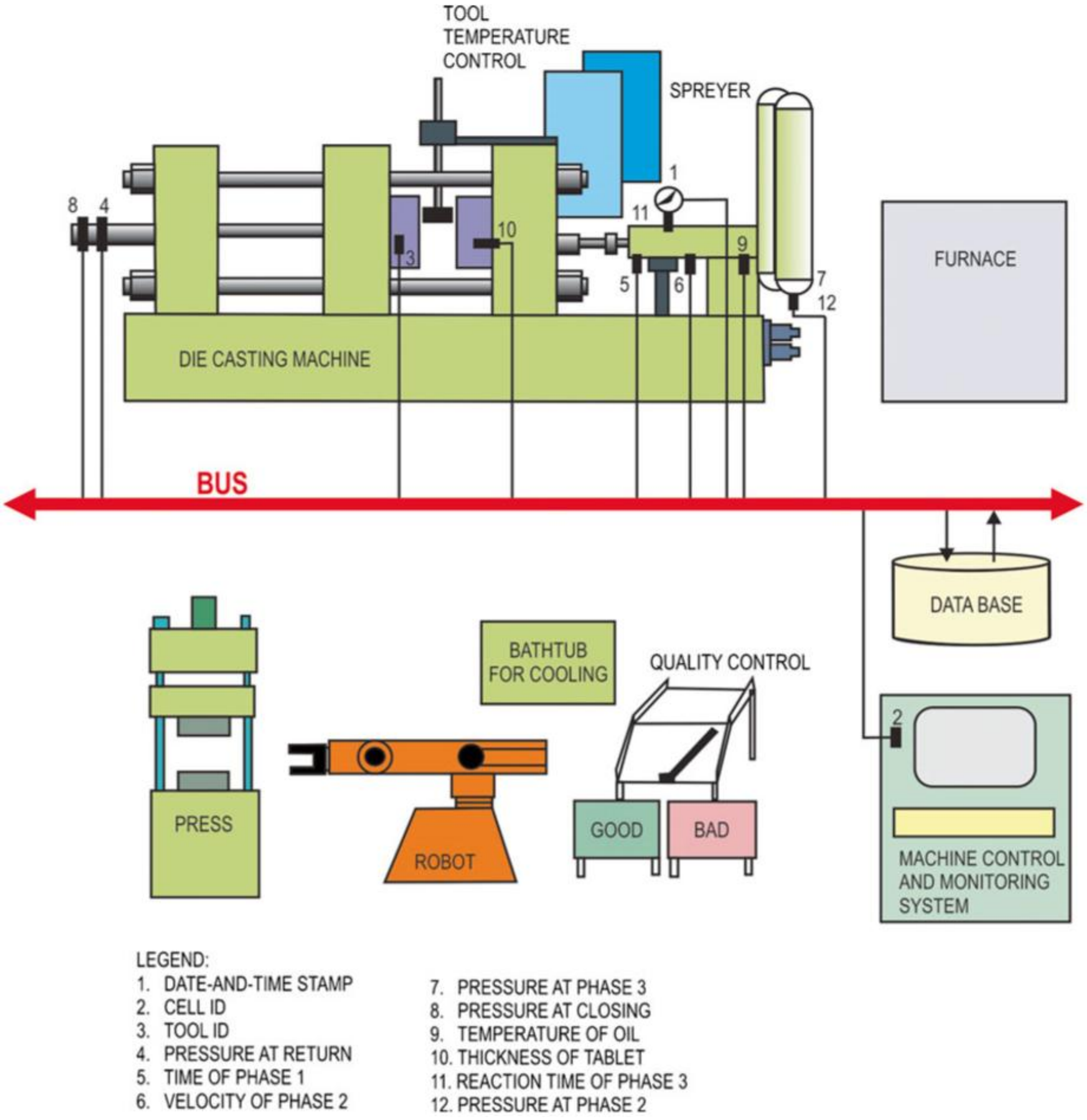


Figure. 4.2 Measuring points for input and output process parameter in a die casting cell [1].

4.3. Finding correlations between the parameters

Since we believed that the data is highly correlated, not just within the manufacturing data set, but also with the weather data, we selected the appropriate data taking into account the colder and warmer months of the year.

Each point on the 3D plot called Figure 4.3 represents a produced piece in the manufacturing process. The bright red points represent a low quality rate and the dark green points represent high quality rate. Looking at the plot we can see a big cluster of red points at around 96 MPa of closing pressure and 21°C outside temperature. The cluster of red points might mean that that the combination between closing pressure and outside temperature in that specific area is somehow affecting the quality rate in a negative way. As we can see, the quality rate dropped when the closing pressure was above 93 MPa. This tells us that the closing pressure should be held within tighter limits. Also, when the temperatures were above 16°C the quality rate drops, meaning the quality rate will not be as adequate in summer months as in winter months.

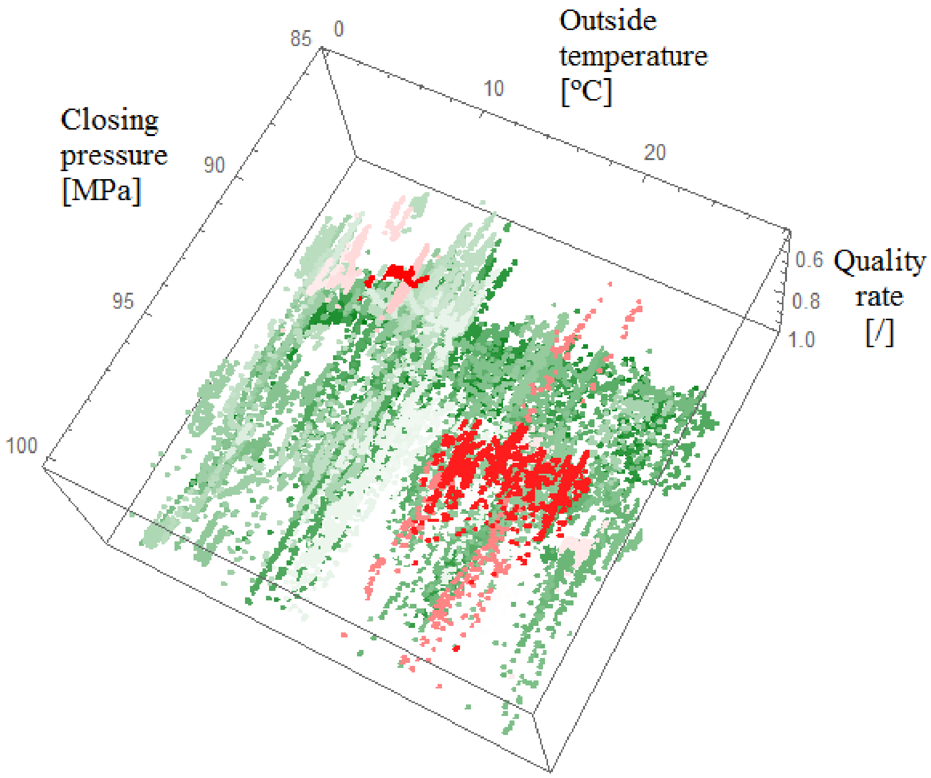


Figure. 4.3 Machine 3, Tool 829 Fi100, between 1.6.2008 and 1.1.2009.

Figure. 4.4 shows the second visualization, however, these time different parameters are used to display the correlations within the data. Along the axis we can see the temperature of the oil, pressure at closing and quality rate for a specific machine for the period of one month. From this 3D plot we can see that the optimal combination of the parameters is achieved when pressure at closing is around 90 MPa and temperature of the oil is around 45°C.

We can see that the temperature of the oil is oscillating at around 20°C which is too much for this kind of process. Since the temperature of the oil is a parameter that is easier to control than the pressure at closing they should make sure that the interval of the temperature is narrower. To ensure that the quality rate is higher, the temperature of the oil should be between 40°C and 45°C. The visualizations above provide us with information about the optimal combination of parameters. However it is impossible to say where the causes of low or high quality rate are just by looking from this perspective. Therefore, we have a holistic view of the data and can only try to manually interpret what could be the possible causes and consequences. The latter can be a starting point for other research, as this will not be discussed in this thesis.

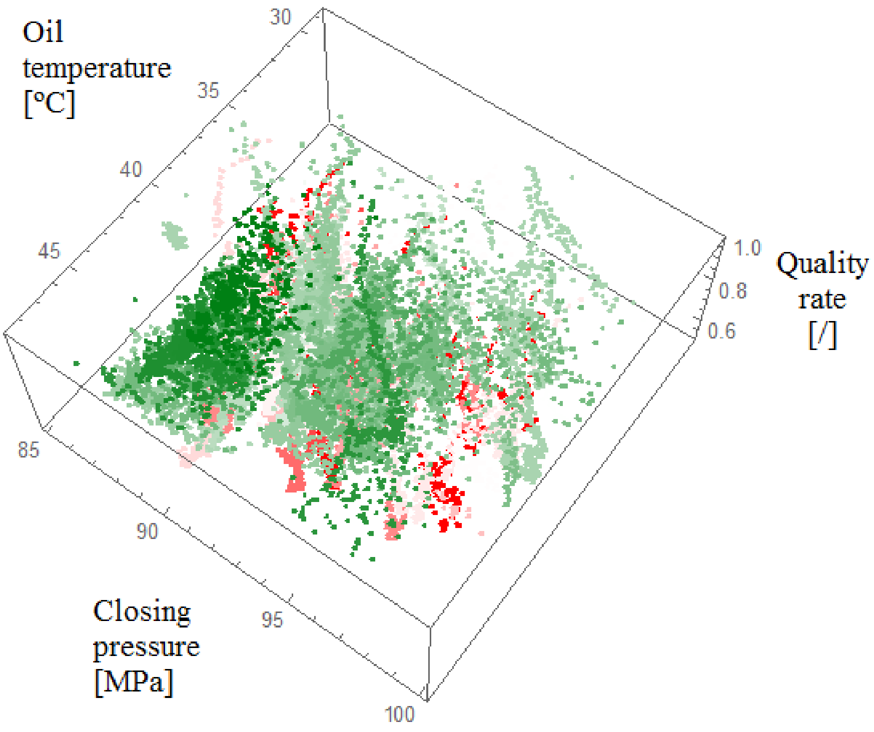


Figure. 4.4 Machine 3, Tool 829 Fi100, between 1.6.2008 and 1.1.2009.

5. CONCLUSIONS

The main goal of this master thesis was to prove that it is possible to find certain correlations between manufacturing data and publicly available data. Using described methods we were able to identify some correlations by manually interpreting a visualization of the plots created with Mathematica software. From the analysis and results we are also able to suggest certain improvements that could make the difference in the near future of the company. One of the suggestions to the company would be to make sure the temperature in its facilities are constantly as "high" as it in the winter months. The variance should be controlled in tight limits. Likewise, all the materials used in the die casting process should have the same temperature as they do in the colder months of the year. They could also increase the quality rate by trying to reduce the interval of the closing pressure. Closing pressure should never exceed 93 Mpa. With this kind of study, we are able to better understand the complex relationships among process variables and other relevant factors of the work system, as well as, the environment that influence process stability, product quality, and system productivity. Therefore, the learning feedback is established, which enables the work system to learn continuously from experiences that can make a difference in the future of the productivity of the company. An important necessity for industrial applications is speed in learning, and developing the data mining models and their solutions, due to continually changing customer and technical requirements in manufacturing systems. Data mining software with comparable qualities needs to be developed. In turn, this demands the availability of more robust, easy to learn and implement data handling and data mining approaches for solving manufacturing problems. Such software should also have the capability to help users select the most appropriate methods for quality improvement problems, and to interpret the results obtained from the applications. This study can help the researchers to developing or further improving upon their methods and tools by providing them with critical information on the typical characteristics of quality improvement data collected, necessary data mining functions and methods, and expected results.

6. BIBLIOGRAPHY

[1] S. Žapčević and P. Butala, “Adaptive process control based on a self-learning mechanism in autonomous manufacturing systems,” *Int. J. Adv. Manuf. Technol.*, vol. 66, no. 9–12, pp. 1725–1743, 2013.

[2] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, “Analytics: The real-world use of big data,” *IBM Glob. Bus. Serv. Saïd Bus. Sch. Univ. Oxford*, pp. 1–20, 2012.

[3] A. Gandomi and M. Haider, “Beyond the hype: Big data concepts, methods, and analytics,” *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, Apr. 2015.

[4] L V (Venkat) Subramaniam, “Big data veracity challenges.” [Online]. Available: http://www.slideshare.net/prayukth1/big-data-veracity-challenges. [Accessed: 17-Aug-2015].

[5] Kenneth Cukier, “Data, data everywhere | The Economist.” [Online]. Available: http://www.economist.com/node/15557443. [Accessed: 17-Aug-2015].

[6] A. Labrinidis and H. V. Jagadish, “Challenges and opportunities with big data,” *Proc. VLDB Endow.*, vol. 5, no. 12, pp. 2032–2033, Aug. 2012.

[7] McKinsey & Company, “Big data: The next frontier for innovation, competition, and productivity,” *McKinsey Glob. Inst.*, no. June, p. 156, 2011.

[8] F. Gullo, “From Patterns in Data to Knowledge Discovery: What Data Mining Can Do,” *Phys. Procedia*, vol. 62, pp. 18–22, 2015.

[9] Margaret Rouse, “What is database?” [Online]. Available: http://searchsqlserver.techtarget.com/definition/database. [Accessed: 17-Aug-2015].

[10] Microsoft, “SQL Server Management Studio.” [Online]. Available: https://msdn.microsoft.com/en-us/library/hh213248.aspx. [Accessed: 17-Aug-2015].

[11] Indiana University, “What is SQL, and what are some example statements for retrieving data from a table?” [Online]. Available: https://kb.iu.edu/d/ahux. [Accessed: 17-Aug-2015].

[12] W3School, “SQL AND & OR Operators.” [Online]. Available: http://www.w3schools.com/sql/sql_and_or.asp. [Accessed: 17-Aug-2015].

[13] Margaret Rouse, “What is public data? - Definition from WhatIs.com.” [Online]. Available: http://searchcio.techtarget.com/definition/public-data. [Accessed: 17-Aug-2015].

[14] Wolfram, “Wolfram Mathematica: Definitive System for Modern Technical Computing.” [Online]. Available: http://www.wolfram.com/mathematica/. [Accessed: 17-Aug-2015].

[15] Len Gallagher, “Database Language SQL.” [Online]. Available: http://www.itl.nist.gov/div897/ctg/dm/sql_info.html. [Accessed: 17-Aug-2015].

[16] “What is Java Database Connectivity (JDBC)? - Definition from WhatIs.com.” [Online]. Available: http://searchoracle.techtarget.com/definition/Java-Database-Connectivity. [Accessed: 17-Aug-2015].

[17] “QuantityMagnitude—Wolfram Language Documentation.” [Online]. Available: https://reference.wolfram.com/language/ref/QuantityMagnitude.html. [Accessed: 17-Aug-2015].

[18] Peter JB King, “Describing an Algorithm.” [Online]. Available: https://www.macs.hw.ac.uk/~pjbk/pathways/cpp1/node33.html. [Accessed: 17-Aug-2015].

[19] Flowchart Software, “RFFlow Flowchart Software.” [Online]. Available: https://www.rff.com/. [Accessed: 17-Aug-2015].

[20] J. L. Rodgers and W. A. Nicewander, “Thirteen Ways to Look at the Correlation Coefficient,” *Am. Stat.*, vol. 42, no. 1, pp. 59 – 66, Feb. 1988.

[21] DenisBoigelot, “File:Correlation examples2.svg - Wikimedia Commons.” [Online]. Available: https://commons.wikimedia.org/wiki/File:Correlation_examples2.svg. [Accessed: 18-Aug-2015].

[22] Wikipedia, “Data_visualization @ en.wikipedia.org.” .

[23] Wolfram, “ColorFunction—Wolfram Language Documentation.” [Online]. Available: https://reference.wolfram.com/language/ref/ColorFunction.html. [Accessed: 18-Aug-2015].

[24] S. Žapčević, “Model of a Self-Learning Manufacturing System,” PhD thesis, University of Ljubljana, 2013.