Big Data Analytics – A Review of Data Mining Models for SMEs in the Transportation Sector

[1]Siti Aishah Mohd Selamat, [1]Simant Prakoonwit, [2]Reza Sahandi, [1]Wajid Khan, [2]Manoharan Ramachandran

[1]Department of Creative Technology
[2]Department of Computing
Bournemouth University
Bournemouth, United Kingdom

{aishah,sprakoonwit,rsahandi,wkhan,mramchandran}@bournemouth.ac.uk

Abstract

The need for Small and Medium Enterprises (SMEs) to adopt data analytics has reached a critical point, given the surge of data collected from the advancement of technology. Despite data mining being widely used in the transportation sector, it is staggering to note that there are minimal research case studies being done on the application of data mining by SMEs specifically in the transportation sector. From the extensive review conducted, the three most common data mining models used by large enterprises in the transportation sector are "Knowledge Discovery in Database" (KDD), "Sample, Explore, Modify, Model and Assess" (SEMMA) and "CRoss Industry Standard Process for Data Mining" (CRISP-DM). The same finding was revealed in the SMEs' context across the various industries. It was also uncovered that amongst the three models, CRISP-DM had been widely applied commercially. However, despite CRISP-DM being the *de-facto* data-mining model in practice, a study carried out to assess the strengths and weakness of the models reveal that they have several limitations in respect of SMEs. This paper concludes that there is a critical need for a novel model to be developed in order to cater to the SMEs' prerequisite, especially so in the transportation sector context.

Key Words

Data Mining, KDD, CRISP-DM, SEMMA, SMEs and Transportation

Introduction

Intelligent Transportation Systems (ITS) utilise advanced technologies and systems to provide efficient and safe transportation services while minimising the operational cost and environmental impacts [1]. The ITS evolution has seen a dramatic development in the last two decades – whereby, from the 1970s to 1980s the primary area of` development was concentrated in curbing Traffic Congestion [2]. From the 1980s to 1990s, the building of Intelligent Infrastructure and Vehicles was the core focus of development [2]. With the advancement of technology in the 21st Century, data is increasingly collected every hour, every minute and every second causing a data explosion era. The International Data Corporation (IDC) forecast that the volume of data is expected to grow up to 50 Zettabytes globally (equivalent to fifty billion Terabytes) by the year 2020 [3]. These can revolutionise

the development of ITS, by shaping a traditional technology- driven system [4] into a more robust ITS ecosystem [3]. The influx of data can only become an asset to the organisation if they are implicitly intelligible to translate useful knowledge for small and medium enterprise (SMEs) organisations [5, 6]. As shared by Lyons in a thought-provoking editorial piece in *Transport Reviews* – in a progressive environment load with robustness, interconnectedness, it is yet 'uncertain'. Therefore, there is a crucial need to evaluate the relevancy of the transport analysis purposes [7].

In a conclusive report by European Commission, the key economic drivers of growth in the European continent is the SMEs – contributed 3.9 trillion euros to the economy in 2015 [8]. This is twice as much in comparison to the large enterprises [9]. The transportation and the storage enterprise made up of 5% of the 22.3 million of the non-financial business economy in 2012 [10]. SMEs can further reap two to three times growth rate through the exploitation of advanced technologies (such as social media, big data, cloud computing and mobile). Eurostat identified that less than 7% of the European SMEs have employed data analytics in it's their business; making a need for digital transformation as a high priority for the EU [11] in this data explosion era. In an in-depth report by IDC on IDC European Vertical Market Survey 2012, it was ascertained that only 3% (an estimation of 1,500 out of 50,000 organisations) of the SMEs in the transportation and storage sector have deployed data analytics in their business [12]. The transport and storage SMEs have been relentlessly labelled as laggards in the adoption of Big Data technologies. SMEs can become 5-6% more productive through the utilisation of data analytics in the business – as evident in the larger transport companies [11]. Despite the momentous potential benefits of utilisation of big data analytics, the transport and storage SMEs are still dawdling in their adoption efforts. In 2012, the adoption rate of big data analytics of SMEs in the UK stood at 0.2% compared to the large enterprise, with an uptake of 25% [8]. This is indeed an alarming figure, as the fast adoption rate by the large enterprise may eventually implicate SMEs to become irrelevant and absolute. Therefore, there is an urgent need for SMEs to begin exploring the implementation of big data analytic and data mining (DM).

To ensure the relevance of this study, articles published with the last 10 years were only included. The selections of the literature were divided into three categories. The first category includes papers relating to big data analytics for SMEs. The second and third category encompasses papers relating to DM models in the transportation sector from the SMEs and the large enterprises subsequently. The full text of each article was screened in order to validate the relevancy and applicability of the articles. Upon screening, only suitable articles were included for this study. This paper presents the outcome of a critical study of the big data analytics literature, in respect of data mining models for SMEs in the transportation sector in particular. This information was extracted from online databases such as ACM Digital Library, Science Direct, Springer, EBSCOhost EJS, Semantic Scholar, Google Scholar (search engine) and IGI Global. The research aims to provide researchers, transportation business leaders and policy makers' eminent findings of the big data analytics research studies. It is anticipated that this paper will magnify the emergence of big data technologies aiding SME's understanding and capitalisation to facilitate and spur business growth.

In the next section, the paper will provide background information on the big data analytics challenges and problems faced by SMEs. The third and fourth sections will cover the DM application case studies in both the transportation sector and SMEs context. The fifth section will present a comparative study on the most commonly used DM models. This is then followed by a discussion on

the model's strengths and deficiencies. Lastly, the paper will conclude with several key points and a discourse of future research work to be undertaken.

1. Background: SMEs Adoption Barriers in Big Data Analytics

Aside from the transport and storage sector, the SMEs group are straggling with the implementation of big data analytics in their businesses. This raises an alarm as to what is the hindering factor(s) that is curbing the SMEs from advancing with the evolution of big data technology. In a recent in-depth study by Coleman et al., they uncover several core factors contributing to the slow acceptance of big data analytics by SMEs in the European continent [13]. The factors are listed below:

(1) Minimal cognizance in the big data analytics domain.

(2) Little or no interest in new management trends.

(3) Insufficient in-house data analytics experts.

(4) Increasing shortage of competent data analyst in the labour market.

(5) Lack of exemplary successful case studies for SMEs to refer to.

(6) Lack of effective analytics consulting services.

(7) Highly complex analytics solutions in the software market.

(8) Data security concerns.

(9) Data protection and privacy concerns.

(10) Lack of financial access to invest in new technologies.

1.1 Classifications of Area of Concerns

To further comprehend the nature of the identified barriers, the SMEs' areas of concerns are classified into three groups – resources, knowledge and data management (as outlined in Table 1). Building on the research findings of Coleman et al., several the three classified group will constitute examples from other research studies to support and validate these key barriers as identified.

Areas of Concerns	List of Challenges and Barriers
Resources	• Insufficient in-house data analytics experts. • Increasing shortage of competent data analyst in the labour market. • Lack of effective and value for money analytics consulting services available. • Highly complex analytics solution in the software market. • Lack of financial access to invest in new technologies.
Knowledge	• Low understanding of big data analytics domain. • Little or no interest in new management trends. • Lack of exemplary successful case studies for the SMEs to refer to.
Data Management	• Data security concerns. • Data protection and privacy concerns.

Table 1. Classification of SMEs Big Data Analytic Barriers

1.1.1 Resources

In respect of grouping classification, the SMEs nucleus area of concern is mostly in relation to the subject of resources. This is followed by knowledge and data management concerns. It can be deduced that the lack of an in-house data analytics specialist is an implication caused by the insufficient number of available qualified data analytics talent. For instance, in the United States, it is envisaged that by 2018, there will be a shortage of close to 190,000 skilled analytical talent and also a shortfall of 1.5 million analysts and managers with the relevant competency to derive strategic decision(s) from big data analysis [14]. A survey carried out amongst recruiters in the UK revealed that up to 57% of the recruiters are facing obscurity in filling up the big data analytics gaps – this is inclusive of the large companies [15]. The scarcity of qualified data scientists would deter the analytics development scene in the European market [16]. Given the shortfall of talent supply, it is expected that the existing analytics software, readily available in the market would aid in curbing the impending expertise gap. There are plenty of analytics solutions available in the market. Nonetheless, to find a solution that is both user-friendly and embedded with robust analytical capabilities is scarce. The need for an instinctive user interface is critical in shortening the user learning curve [16] – allowing faster implementation for the SMEs. As evaluation platforms may tend to be vendor biased; an end user with minimal or zero proficiency in analytics may find it difficult to select a solution with a decent price-performance ratio. With an innumerable number of studies highlighting that financial limitations are the SMEs' major hindrance block [17, 18], a lower price-performance ratio solution would be more desirable by the SMEs.

1.1.2 Knowledge

In the area of knowledge concerns, a survey conducted in the UK reveals that SMEs' personnel has an exceptionally low comprehension of the big data analytics domain [15]. A similar survey conducted in Germany shared an identical result [13]. To a great extent, the SMEs are uncertain of their datasets potentiality, in turn drawing hesitation on the need to invest in data science capabilities to reap the intended benefits as affirmed by the various analytics *connoisseur*. In spite of the fact that there are guidelines available for the SMEs to make reference to, there is still a lack of exemplary research case studies that propagates the successful implementation of analytics in the SMEs sphere [13]. The existing big data use cases generated in the European Union (EU) often do not correlate with the SMEs' points of interests [19]. More case studies are needed to possibly fill the knowledge gap and jumpstart the SMEs enthusiasm to take more interest in the big data analytics domain.

1.1.3 Data Management

Finally, yet importantly, data security, protection and privacy are the SMEs' key concerns in the area of data management. Close to 50% of SMEs identified data security and protection as the key barrier to big data analytics – in a worldwide survey of more than 82 SMEs companies [13]. In comparison to larger companies, SMEs lack the competency to scale up their IT security level [20]. The use of obsolete and non-updated database management system raised a critical IT security gap for SMEs. In consequence, making SMEs less resilient against cyber-attacks and intrusions. The processing and analysis of customer's data by SMEs, on the other hand, has to abide by European Union (EU) legality on data protection and privacy. The lengthy EU data protection law [21], mooted in 2012,

creates an added constraint for SMEs. Predominantly, in view of the lack of financial resources, SMEs could not meet the expenses to engage a legal expert in order to fully grasp the EU's data protection legislation requirements.

1.2 Supplemental Intangible Barriers

In addition to the discussions in respect of adoption barriers in Sections 1 and 1.1, it is worth considering the supplemental intangible barriers that may also hinder the adoption of analytics for SMEs. These intangible barriers relate to SME's organisational culture, organisation structure and decision-making. First and foremost, in terms of organisational culture – given that SMEs are highly domain-specialised, they have little or no interest in new management trends that might be beneficial for the organisation [22]. This culture of intrinsic conservatism is leading the SME's attitude in taking big data analytics as a management hype instead of an opportunistic viewpoint. The second aspect of organisational structure implies the need to have a fitting management concept within an organisation, in order to create an economic success on the adoption of analytics[23]. Unlike the large enterprise, the organisational structures of most SMEs are flat with few or no levels of middle management between the executives and staffs [24]. The flat organisational structure of SMEs would in turn impact; on the way the organisation makes it decisions. The decision makers in SMEs are often the business owners, which are in a way usually tied up with the owner's identity and life[22]. The decision making within the large enterprise tends to be more rational because of the complexity of the organisation structure and decision making units [25]. In view of the scarcity of resources and expertise, SMEs would be in a limiting position to make a complex decision, as it is often reliant on the business owner's intuition [25]. In other words, if the business owner is not personally attuned with the latest business trends of analytic adoption, it will be an intricate barrier to overcome.

1.3 What Data Mining Can Mean for SMEs?

The explosion of data is deemed critical for SMEs because, during the DM process, organisations can radically learn more about their business and translate the new knowledge into better decision making and performance [23]. In other words, DM has the potential to transform traditional SMEs to organisations with a competitive advantage. For instance, suppose an SME aims to mine its customer data, the potential benefit would entail creating cross-selling avenues at a higher margin, improving its customer retention and satisfaction rates, identifying the most profitable customer group and last but not least, enhancing the SME's marketing and sales strategy [26]. In the mining of inventory data, on the other hand, SMEs can gain an advantage by forecasting of the inventory that will help to reduce the total value of stock held. This would create a positive implication in allowing timely inventory purchase from the supplier, creating a supplier lock-in, leading the SME to a better trading agreement [27]. It is therefore evident that DM has the capability to create business opportunities to enable SMEs to stay ahead of their competition and leverage on the possibilities.

2. Data Mining in Transportation Sector

In the area of transportation, there have been several pieces of research developing novel approaches for traffic management, motorist and commuter safety, transport mobility, road accident management and much more – with the application of DM. Table 2 is a compilation of applications of DM in the transportation sector.

The table illustrates that DM is widely applied in all three transportation modes – land, air, and sea. An example of a DM application in land transportation is research carried out by Giovanni et al., in which DM is used to the predict the railroad demands to facilitate operational and manpower planning for Malha Regional Sudeste (MRS) Logistica [31]. Cristobal et al.'s research denote a similar area of interest in the effective management of resource planning in predicting passenger demands for Gran Canaria Island Public Transport [32]. An example of sea transportation DM application is Greis et al.'s research, which the study involves in applying DM to identify high-risk shipments reaching the United States of America (U.S.A.) ports [33]. Finally, Lukacova et al.'s research adopt DM to assist the Federal Aviation Administration [34] [34] to predict potential incidents and implications [35].

From the compilation of DM application in the transportation sector, Table 2 reflects a distinctive commonality whereby DM was applied to extract new information/knowledge for prediction capabilities. Secondly, the three recurring DM models adopted by the enterprises were Knowledge Discovery in Database [36], Sample, Explore, Modify, Model and Assess [36] and CRoss Industry Standard Process for Data Mining (CRISP-DM). Out of the 10 industrial examples quoted, six enterprises had adopted the CRISP-DM model and the remaining four enterprises had used the SEMMA and KDD model equally. Of the table, it is evident that CRISP-DM marks as the most commonly used model. And these findings correlate with the industrial polls conducted by KdNuggets.Com [37-39]. On the types of data used, the majority of the enterprises are leveraging on their historical data for data processing and analysing. The data is in the form of structured data – referring to data that are organised in a relational database that is structured in columns (fields) and rows (record) [40]. On a different note, one key prominent finding derived from Table 2 indicates that little is known of research studies on SMEs in the transportation sector.

3. Data Mining in the SMEs Context

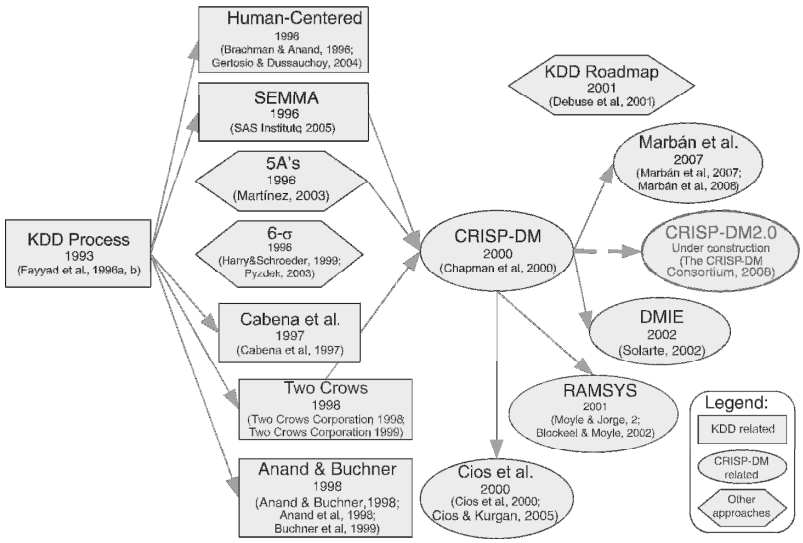
Having understood that there is a minimal study on SMEs in the transportation sector, this section aims to encapsulate the use of DM in the SMEs context across various industries. In all, ten recent case studies have been tabulated, as seen in Table 3. The case studies reflect the application of DM in the food & beverage, tourism, Information Technology (IT), financial, aviation, trading and manufacturing industries. From the ten case studies, DM is primarily used for prediction and improvement of the decision-making process. For instance, under the finance industry, Mandala et al.'s research are involved in assessing the credit risk of loan lenders [41]. On the other hand, Koyuncugil et al. in their research, employ DM to predict to detect financial risks for SMEs [42]. In respect of IT, SMEs are also utilising DM for various usage. In a research conducted by Bozdogand and Zincir-Heywood, DM is used to facilitate the SME's in collecting its public resources automatically to create a knowledge base to support it's IT management [43]. In another study conducted by Ibukun et al., the objective for applying DM is to identify customer segmentation in order to carry out target marketing [44]. A recent study by Packianather et al. for SMEs in the manufacturing industry, they applied DM to generate unique and new knowledge for forecasting and strategic decision-making [45]. The case studies for application of DM by SMEs as depicted in Table 3 insinuate that DM can be applied across many industries. The CRISP-DM is a widely used model. Our in-depth review, (see Tables 2 and 3), exhibits that CRISP-DM is the most frequently used DM model. Out of the ten case studies considered, six studies utilised the CRISP-DM model, three used the KDD model and one adopted the SEMMA model. Another synonymous result reflects that the SMEs are in most cases using their historical structured data for DM applications – as reflected in

Table 2. Details of the elements and functions of the most frequently used DM models will be discussed in the next section.

4. The Evolution of Data Mining

Data mining (DM) refers to the science of identifying valuable and unique information from a substantial size of data sets or databases. The mining procedures involve intensive computing of data analysis [28]. The core goal of DM is processing a large amount of data to generate new knowledge [29]. DM serves two primary goals – (1) uncover new insights and (2) generate predictions [30]. DM is a process for Knowledge Discovery in Database [36] [30]. The term KDD refers to a set of broad processes used to discover valuable knowledge from a set of data collection [46]. The emergence of KDD was sparked with the rising establishment of big databases in the varying number of organisations in the early 90s [47]. This created a paradigm shift for the need to develop data mining algorithms with the capabilities to unearth gainful insights from the big volume of data that are residing in companies' databases. Figure 1 depicts the overall evolution of the data mining process models with KDD as its foundation and CRISP-DM as the core focal point of the evolution. This section will only discuss the three most applied models, KDD, SEMMA and CRISP-DM in depth. A critical comparison of these models has been made which is discussed in the next section.

Figure 1. Evolution of Data Mining Methodologies
Source: Mariscal et al. [47]



4.1 KDD

The KDD process can be defined as an un-superficial way of distinguishing potentially useful, valid and conclusively understandable patterns from the data [48]. The term process refers to the many stages that are involved in the KDD process. KDD can also be described as the overall approach of uncovering valuable knowledge from data [30]. It also entails the evaluation and (perhaps) the interpretation of the new insights and knowledge for decision-making. Outlined in Figure 2 is the overall overview of the KDD process from the data viewpoint – interactive, iterative and with many feedback loop points. In all, the KDD process encompasses nine steps [30].

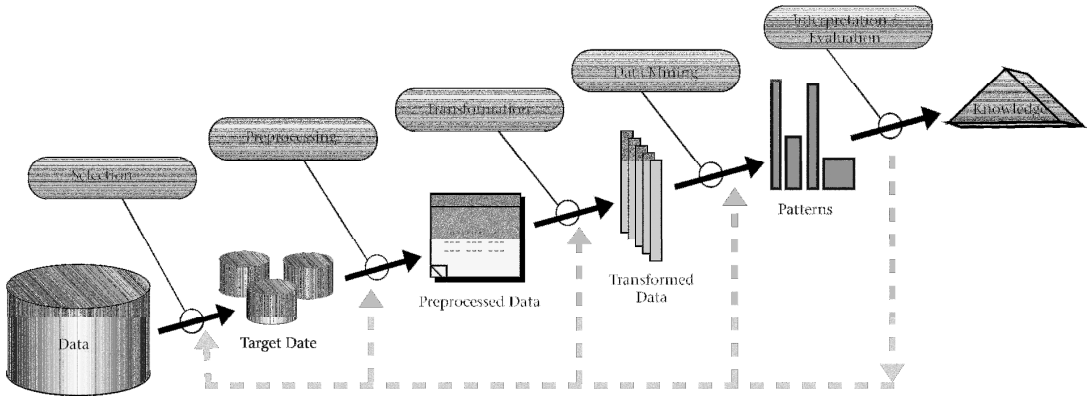
Studies	Sector Type	Company	Transportation Mode / Type	Data Type	Model Adopted	Application
Viglioni, Cury [31]	Private	MRS Logistica	Land / Train	• Historical Data • Structured Data	CRISP-DM	Prediction of railroad demands to facilitate operation and manpower planning.
Wong and Chung [49]	Private	Taiwanese Domestic Airline	Air / Airflight	• Historical Data • Structured Data	KDD	Mining passengers' demographic, travel behaviour and core service quality information for customer retention initiatives.
Haluzová [50]	Public	Prague Public Transit Company	Land / Bus	• Historical Data • Structured Data	CRISP-DM	Identification of the accident influences between car and tram on the electric tramway net.
Shin, Park [51]	Private	Jeju Taxi Service	Land / Taxi	• Historical Data • Unstructured Data	SEMMA	Analysing passenger pick-up location patterns to proposed potential pick-up locations for empty taxis.
Mirabadi and Sharifian [52]	Public	Iranian Railways	Land / Train	• Historical Data • Structured Data	CRISP-DM	Analysing historical accident data to discover the unsafe condition contributing factors.
Zhang, Huang [53]	Public	China Railways	Land / Train	• Historical Data • Structured Data	KDD	Deriving intelligent based decision-making in accident treatments.
Greis and Nogueira [33]	Public	U.S. Seaport (Department of Homeland Security)	Sea / Shipping Cargo	• Real-time Data • Structured Data • Unstructured Data	CRISP-DM	Identification of high-risk shipments reaching the U.S.A. ports.
de Almeida and Ferreira [54]	Public	BUS Public Transport	Land / Bus	• Historical Data • Structured Data	SEMMA	Identification of the most fuel-efficient resources in route operation and areas of resources for improvements.
Lukáčová, Babič [35]	Public	Federal Aviation Administration	Air / Airflight	• Historical Data • Structured Data	CRISP-DM	Analysing the aviation historical incident data to predict potential incidents and implications.
Moreno-Díaz, Pichler [32]	Public	Gran Canaria Island Public Transport	Land / Bus	• Historical Data • Structured Data	CRISP-DM	Predicting passenger demand for efficient resource planning and deployment.

Table 2. Case Studies on DM Applications in the Transportation Sector

Studies	Research Context	Industry	Data Type	Model Adopted	Application
Raju, Kang [55]	UK based SME wholesaler	Food & Beverage	• Historical Data • Structured Data	CRISP-DM	Forecasting freshly produced (short shelf life) product demands.
Rebón, Castander [56]	Tourism SMEs	Tourism	• Real-time Data • Structured Data	KDD	Enhancing the analysis technique to improve decision-making process of credit fraud transaction detection.
Pytel, Britos [57]	Project Planning for SME	Information Technology	• Historical Data • Structured Data	CRISP-DM	To predict the cost and effort estimation for small-sized software projects.
Mandala, Nawangpalupi [41]	Rural Bank	Financial	• Historical Data • Structured Data	CRISP-DM	To develop credit assessment to in order to classify lenders as performing or non-performing loans risk.
Koyuncugil and Ozgulbas [42]	SMEs	Financial	• Historical Data • Structured Data • Unstructured Data	KDD	To predict financial risk detection for SMEs.
Bozdogand and Zincir-Heywood[43]	IT Management for SMEs	Information Technology	• Historical Data • Structured Data	SEMMA	To automatically generate an IT management support knowledge base from public resources.
Cheung and Li [58]	SMEs	Trading	• Historical Data • Structured Data	CRISP-DM	To uncover hidden patterns in the sales and market domain.
Packianather et al. [45]	SMEs	Manufacturing	• Historical Data • Structured Data	KDD	To generate unique and new knowledge for forecasting and strategic decision making.
Young et al. [59]	SMEs	Aviation	• Historical Data • Structured Data	CRISP-DM	To decide how and where aircraft maintenance process can be enhanced or amended.
Ibukun et al. [44]	SMEs	Information Technology	• Historical Data • Structured Data	CRISP-DM	To identify customer segmentation in order to carry out target marketing.

Table 3. Case Studies on DM Applications in the SMEs Context

Figure 2. Knowledge Discovery in Database [36]
Source: Fayyad et al. [30]



The outline of the KDD steps is as follows:

(1) Understanding the application domain.

(2) Constructing a target data set.

(3) Data cleaning and pre-processing.

(4) Data transformation.

(5) Data mining function selection.

(6) Data mining algorithm selection.

(7) Data mining.

(8) Examination and evaluation of mined data.

(9) Employing newly discovered knowledge.

4.2 SEMMA

Developed by the Suite of Analytics (SAS) Institute, SEMMA refers to the systematic tool set of SAS enterprise miner for delivering the data mining core tasks [60]. The SEMMA model can only function with the enterprise miner tool, which had been developed by the SAS Institute. The KDD process, on the other hand, is an open source process that can be administered in various environments. The SEMMA model's principle focus is on its model development point of data mining [61]. Figure 3 illustrates the five SEMMA steps. The steps consist of Sample, Explore, Modify, Model and Assess.

Figure 3. Sample, Explore, Modify, Model, Assess (SEMMA) Methodology
Source: Mariscal et al. [47]

4.3 CRISP-DM

The CRISP-DM model was mooted together by highly acclaimed organisations such Teradata, Daimler-Chrysler, SPSS and OHRA in the mid-1990s [62]. It is considered as a *de facto* standard for

establishing data mining projects. The popularity of CRISP-DM was contributed to the fact that the model is applicable across all industries [47]. Unlike the KDD and SEMMA models, the CRISP-DM process model renders a continuous life cycle modus operandi. In addition, in each phase of the project, it corresponds to the designated tasks and interrelation between each task. As depicted in Figure 4, the overall cycle of the CRISP-DM data mining project comprises of six stages. The chain of cycle in CRISP-DM is flexible, allowing the end user to move back and forth freely. The chain of sequence is really dependant on the result of the specific task of the concerning phase.

The six phases of CRISP-DM are:

(1) Business Understanding

(2) Data Understanding

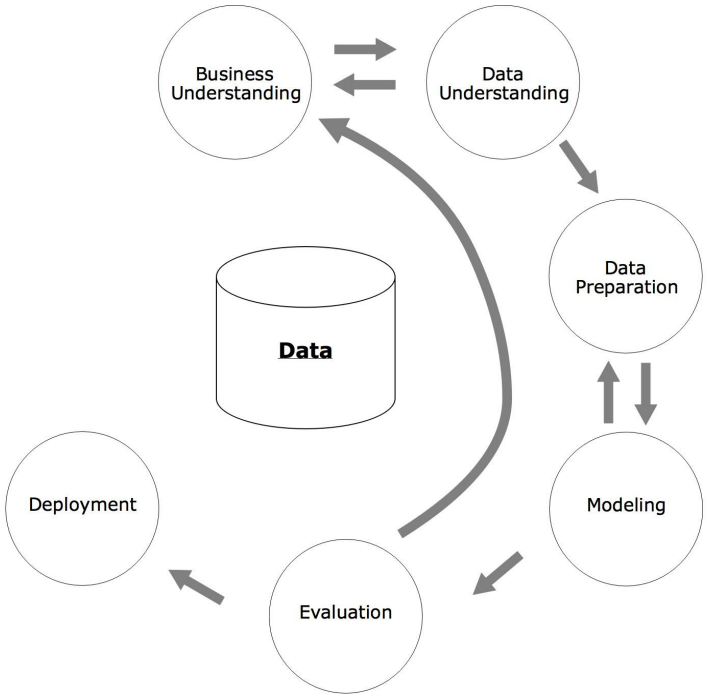
(3) Data Preparation

(4) Modelling

(5) Evaluation

(6) Deployment

Figure 4. CRISP-DM Methodology
Adapted from: Chapman, Clinton [62]



5. Data Mining Models Detailed Discussions

In this section, the DM models will be discussed in two parts. The first part will consist of a critical comparison of the KDD, CRISP-DM and SEMMA model from various aspects. The strengths and limitations of the DM model in the SMEs context will be discussed. The objective of this discussion is

to synthesise the findings, which would be key in the proposal for a new DM model, especially for SMEs in the transportation sector, in which further elaboration will be provided.

5.1 Models Comparison

Quantitative and qualitative comparisons of the three models are shown in Table 4. For illustration purposes, the first and second rows outline the key facts of the models, namely the creator(s) and the year when each model was first introduced. The third to fifth rows indicate the models' functions, the total number of implementation steps and a brief description of each phase. Subsequently the sixth to the tenth row specify the industry involvement of each model, the requirement for background knowledge in DM, the status of the software tool supported, availability of model documentation for users' reference and lastly, the status whether the model can support an open source tool. The eleventh to the final row consist of industry related components like total case studies in the transportation sector, total case studies in the context of SMEs, overall case studies across all sectors, application areas and last but not least, the KDnuggets poll results for 2007 and 2014. The core motivation for carrying out the model comparisons is shown in Table 4. It is worth pointing out that despite the variation of phases entailed by each DM model, the three models are entrusted with the same core functions. The eventual intended outcome of the three DM models is to uncover new insights and to generate predictions. One prominent difference amongst the three models is the need for an initial understanding (in the domain or business) of the first phase of data-mining projects. Unlike SEMMA, both KDD and CRISP-DM require this particular phase. For SEMMA, the DM methodology focal point is in its technical characteristic that is involved during the development process, starting with data sampling. Another key difference that differentiates SEMMA from the other two models is its shortfall in applying and deploying the new knowledge, which is uncovered. In terms of the development of the models, only KDD had not had any involvement from industry and has no supporting documentation for the users to make reference to. A common component that all the three models share is the need to have prior knowledge in DM in order to be able to apply the model in practice. With reference to software support, unlike KDD and CRISP-DM, SEMMA does not support open source tools. As indicated in table 3, CRISP-DM is the most applied DM model by SMEs in the transport sector. This finding is also affirmed by a poll conducted by KDnuggets – a leading online resource on DM. The first poll conducted on 200 respondents in 2007 indicate that 42% of the respondents used the CRISP-DM model, 19% created their own model, 13% used SEMMA, 4% used other non-domain specific models, 7.3% used the KDD process and the remaining 5.3% used other models. A second poll conducted in 2014, on the same number of respondents shows that 43% of the respondents used the CRISP-DM model, 27.5% created their own model, 8.5% used SEMMA, 8% used other non-domain specific models, 7.5% used the KDD process and the remaining 5.5% used other models. The two polls observed an increase in the usage of the CRISP-DM model. It is worth noting that almost one-quarter of the respondents were using their own model in their own individual domain. This pre-eminent finding may suggest that CRISP-DM is not ultimately the *de-facto* DM model for all domains. Nonetheless, as shown in Table 4, CRISP-DM outweighs the KDD and SEMMA model in view of it being industrially attuned. This is credited to its development on real-world knowledge discovery experience [62].

5.2 Strengths and Limitations of Models in the SMEs Context

Following on from the model's comparison, the strengths and limitations of each model will be discussed in this section. This will be in the context of SMEs in the transportation sector. Table 5 shows strengths and limitations of each model according to the case studies listed. As Table 5 shows CRISP-DM. The consistent list of strengths of CRISP-DM illustrates that the model is applicable to the industry, addressing business objectives and issues, providing structured approaches and processes, as well as having the flexibility to use any DM tool. This holistically means that CRISP-DM is practically applicable to the business environment. The compiled list of limitations, on the other hand, is as follows:

1. Long and arduous process with detail steps to be undertaken in each process
2. Requires DM knowledge
3. Explicit need for detailed DM requirements
4. Challenge faced in deriving how and when the selection of data is necessary or irrelevant
5. The late selection of DM technique affects the data format compatibility causing to return to the data analysis stage
6. Inadequate knowledge on domain expert's terminology

Based on the CRISP-DM limitations perhaps the outcome of the case studies is suggesting that the user is experiencing an exhaustive process when applying the model in view of the detail steps that each phase contain. Further, with insufficient knowledge on DM, the user may face challenges in grasping the CRISP-DM concept and mode of application. For instance, in the study by Cheung and Li, the limitation encountered during the identification of the necessary or irrelevant data for analysis[58]. In the study by Ibukun et al., the issue faced was the inadequacy knowledge of DM's terminologies [44]. The technical aspect as encountered by Young et al. was the delay in the selection of the DM technique[59]. When the selected DM techniques do not correspond with the selected data format, incompatibility will occur, causing the user to return again to the data analysis stage. The strengths of the second most applied model, KDD were accredited to its highly interactive process, containing feedback loops in each process. Like CRISP-DM, the KDD model is endorsed for being industrially applicable in the business environment. The limitations of KDD on the contrary, share similar aspects to that of CRISP-DM. Whereby, the incompatibility of tool and data sets would require the KDD users to return to the KDD process again. Lastly, the strengths of the least applied model, SEMMA are attributed to the availability of its DM robust user support. Its drawback is in the case of the model's highly technical centric process. This may pose a great challenge if an organisation has assigned a user who is not technically equipped in the SEMMA domain. In addition, in comparison to CRISP-DM and KDD, the SEMMA model falls short on the knowledge application phase.

MODEL	KDD	CRISP-DM	SEMMA
Developed by	Fayyad et al.	CRISP-DM Consortium	SAS Institute
Year of model introduced	1996	1996 (Officially released in 2000)	1997
Functions	1. Uncover new and unique insights 2. Generate predictions		
Total Steps	9	6	5
Phase	1. Application Domain Understanding	1. Business Understanding	-
	2. Creating a Target Data Set	2. Data Understanding	1. Sample
	3. Data Cleaning and		2. Explore
	4. Data Transformation	3. Data Preparation	3. Modify
	5. Data Mining Method Selection	4. Modelling	4. Model
	6. Data Mining Algorithm Selection		
	7. Data Mining Application		
	8. Discovered Patterns Interpretation	5. Evaluation	5. Assessment
	9. Using Discovered Knowledge	6. Deployment	-
Industry Involvement	No	Yes Consortium of companies involving Teradata, Daimler-Chrysler, SPSS and OHRA	Yes Individually by SAS Institute
Requires background knowledge in DM	Yes	Yes	Yes
Software Tool Support	Yes Mineset™	Yes SPSS Clementine™	Yes SAS™
Documentation	No	Yes	Yes
Open Source Tool Support	Yes	Yes	No
Total Case Studies in The Transportation Sector Count	2	6	2
Total Case Studies in The SME Context Count	3	6	1
Overall Case Studies Count	5	12	3
Application Areas	Aviation, rail, tourism, financial, manufacturing	Logistic, cargo, aviation, rail, public transport, software, financial, marketing and sales, trading	Software, public transport, street taxis
Kdnuggets poll results for 2007 (200 votes total) [63]	7.3%	42%	13%
Kdnuggets poll results for 2014 (200 votes total) [63]	7.5%	43%	8.5%

Table 4. Models Comparison - KDD, CRISP-DM and SEMMA

MODEL	Case Studies	SMEs Industry	Strengths	Limitations
CRISP-DM	Raju, Kang [55]	Food & Beverage	• Applicable to industry context • Addresses business objective and issues • Structured approach and process	• Long and ardous process with detail steps to be undertaken in each process
	Pytel, Britos [57]	Information Technology	• Applicable to industry context • Having the flexibility of using any DM tool	• Requires DM knowledge • Explicit need for detailed DM requirements
	Mandala, Nawangpalupi [41]	Financial	• Applicable to industry context • Addresses business objective and issues	• Requires DM knowledge
	Cheung and Li [58]	Trading	• Applicable to industry context • Structured approach and process • Having the flexibility of using any DM tool	• Challenge faced in deriving how and when the selection of data is necessary or irrelevant • Long and arduous process with detail steps to be undertaken in each process
	Young et al. [59]	Aviation	• Applicable to industry context • Having the flexibility of using any DM tool • Structured approach and process	• Requires DM knowledge • The late selection of DM technique affects the data format compatibility causing to return to the data analysis stage
	Ibukun et al. [44]	Information Technology	• Applicable to industry context • Having the flexibility of using any DM tool	• Inadequate knowledge on domain expert's terminology
KDD	Rebón, Castander [56]	Tourism	• Interactive and iterative process • Can be applied to industry context	• Requires background knowledge in DM
	Koyuncugil and Ozgulbas [42]	Financial	• User-friendly process • Contains feedback loops in each process	• Requires background knowledge in DM • Insuitability of tool with the prepared data cause to make unnecessary loop back to the earlier process
	Packianather et al. [45]	Manufacturing	• Interactive and iterative process • Can be applied to industry context	• Wrong selection of data resulted to the wrong results
SEMMA	Bozdogand and Zincir-Heywood [43]	Information Technology	• Applicable to industry context • Robust support by DM provider	• Highly technical centric process • No clear indication on how to apply the new knowledge

Table 5. DM Models Strengths and Limitations - SMEs Context

Conclusion and Future Work

The need for SMEs to deploy data analytics has reached a point of criticality; with the immense surge of data collected via the advancement of technologies. In accordance to Eurostat, SMEs will reap a higher productivity level of up to 6% through the utilisation of data analytics in the business. Ignoring the call for technology advancement can risk SMEs falling behind large enterprises that are more forthcoming towards the adoption of the new technology. As identified in Section 2, resource, knowledge and data management are the key areas of concerns that are hindering SMEs from adopting analytics – all of which needs to be addressed. In this paper, the applications of data mining models are examined in the transportation sector in the context of both SMEs and large organisations. The paper reveals a compelling finding (see Tables 2 and 3) that CRISP-DM is the most prominent DM model that is widely used by SMEs in the transportation sector. The model is mainly used for prediction and facilitation of decision-making processes. Another commonality in the findings is the types of data being used to run the CRISP-DM model. The majority of the businesses are leveraging on their historical structured data. Lastly, the paper highlights that there is limited case study research on DM application by SMEs in the transportation sector in particular. Three common data mining processes (KDD, SEMMA and CRISP-DM) were critically compared. The comparison was made against the overall implementation processes, the model's strengths and limitations. In all, the findings show the reason why CRISP-DM has been commercially adopted. In addition, our research shows that the model is flexibility to suit any business using any data-mining tool.

Despite CRISP-DM being the *de facto* data-mining model for businesses to adopt – as examined in the study of the model's strengths and limitations in the SME context, there are several shortfalls that require addressing. The core limitation is the principal expectation of the need to have background knowledge on DM in order to fully grasp the terms, concept, and application of DM for the organisation. The second limitation relates to the intense and exhaustive process that the CRISP-DM entails for applying the model in practice. Last but not least, the delay due to the selection of a DM technique that may implicate on the data format compatibility affecting the DM overall process. Following up from this paper, the future research work aims to develop a novel DM model to suit SMEs in the transportation sector. Taking CRISP-DM as the foundation model, an Intelligent Transportation Analytical Model (ITAM) is to be developed. The ITAM aims to conduct an intelligent analysis with the objectives of churning out new insights, showing hidden patterns and relationship within the existing datasets to aid business decision-making. This would undertake the impending limitations of the CRISP-DM model and at the same time taking, into consideration the impending SMEs' constraints as learned – primarily in terms of time and human capacity constraints. Transportation SMEs sector will be identified. The companies' datasets will be collected for evaluation that is to understand the datasets characteristics. Following that, the ITAM will be proposed, tested in a real-life application and undergoes evaluations.

References

1. An S-h, Lee B-H, Shin D-R. A Survey of Intelligent Transportation Systems. 2011:332-337, doi:10.1109/CICSyN.2011.76.

2. Neil Taylor IS, Jon Parker, Jim Bradley – Integrated Transport Planning Ltd., Consulting AGWW, Sustainability CTA, Institute JMHDER. The Transport Data Revolution. *Catapult* 2015.
3. Zeng D, Lusch R. Big data analytics Perspective shifting from transactions to ecosystems. *IEEE Access* 2013.
4. Zhang J, Wang F-Y, Wang K, Lin W-H, Xu X, Chen C. Data-Driven Intelligent Transportation Systems: A Survey. *IEEE Transactions on Intelligent Transportation Systems* 2011, 12:1624-1639, doi:10.1109/tits.2011.2158001.
5. Iansiti M, Lakhani KR. Digital ubiquity: How connections, sensors, and data are revolutionizing business. *Harvard Business Review* 2014:11p.
6. Derrick H. The C-level is coming around on big data [infographic]. 2012. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=edsnbk&AN=146FD1536214CEC8&site=eds-live&scope=site.
7. Lyons G. Transport analysis in an uncertain world. *Transport Reviews* 2016:1-5.
8. Commission E. Annual Report on European SMEs 2015/2016 – SME recovery continues. 2017.
9. Eurostat. SMEs were the main drivers of economic growth between 2004 and 2006. Available at: http://ec.europa.eu/eurostat/en/web/products-statistics-in-focus/-/KS-SF-09-071.
10. Eurostat. Structural business statistics overview - Statistics Explained. Available at: http://ec.europa.eu/eurostat/statistics-explained/index.php/Structural_business_statistics_overview - Main_statistical_findings.
11. Commission E. STRATEGIC POLICY FORUM ON DIGITAL ENTREPRENEURSHIP Fuelling Digital Entrepreneurship in Europe Background paper. 2013.
12. Commission E. Business Opportunities: Big Data. 2013.
13. Coleman S, Göb R, Manco G, Pievatolo A, Tort-Martorell X, Reis MS. How Can SMEs Benefit from Big Data? Challenges and a Path Forward. *Quality and Reliability Engineering International* 2016, doi:10.1002/qre.2008.
14. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. Big data: The next frontier for innovation, competition, and productivity. 2011.
15. UK E-s. Big data analytics: adoption and employment trends, 2012-2017. 2013.
16. Probst L, Frideres L, Demetri D, Vomhof B, Lonkeu O-K, Luxembourg. P. Business Innovation Observatory - Customer Experience. *European Union* 2014.
17. Bartlett W, Bukvič V. Barriers to SME growth in Slovenia. *MOST: Economic Policy in Transitional Economies* 2001, 11:177-195.
18. Fuller□Love N. Management development in small firms. *International Journal of Management Reviews* 2006, 8:175-190.
19. Wadhwa K. BYTE: Big data roadmap and cross-disciplinary community for addressin…. 2014.
20. Lacey D, James BE. Review of availability of advice on security for small/medium sized organisations. *Retrieved* 2010, 2:2013.
21. Rights EUAfF. Handbook on European data protection law. Available at: http://fra.europa.eu/en/publication/2014/handbook-european-data-protection-law.
22. Goebel R, Norman A, Karanasios S. Exploring the Value of Business Analytics Solutions for SMEs. *Association for Information Systems AIS Electronic Library (AISeL)* 2015.
23. McAfee A, Brynjolfsson E. Big data: the management revolution. *Harvard business review* 2012, 90:60-68.

24. Capgemini. Measuring Organizational Maturity in Predictive Analytics: the First Step to Enabling the Vision | Resource. 2012.
25. Culkin N, Smith D. An emotional business: a guide to understanding the motivations of small business decision takers. *Qualitative Market Research: An International Journal* 2000, 3:145-157.
26. Tan D-W, Yeoh W, Boo YL, Liew S-Y. The Impact of Feature Selection: A Data-Mining Application in Direct Marketing. *Intelligent Systems in Accounting, Finance and Management* 2013, 20:23-38, doi:10.1002/isaf.1335.
27. Waller MA, Fawcett SE. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics* 2013, 34:77-84.
28. Xindong W, Xingquan Z, Gong-Qing W, Wei D. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering* 2014, 26:97-107, doi:10.1109/tkde.2013.109.
29. Fan W, Bifet A. Mining Big Data: Current Status, and Forecast to the Future. *SIGKDD Explorations* 2013, 14.
30. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI magazine* 1996, 17:37.
31. Viglioni GMC, Cury MVQ, Silva PALd. Methodology for Railway Demand Forecasting Using Data Mining. 2007.
32. Moreno-Díaz R, Pichler F, Quesada-Arencibia A. Using Data Mining to Improve the Public Transport in Gran Canaria Island. *Springer International Publishing Switzerland* 2015, 9520, doi:10.1007/978-3-319-27340-2.
33. Greis NP, Nogueira ML. Use of Data Mining for Validation and Verification of Maritime Cargo Movement. 2011.
34. Abdul Rahman F, Shamsuddin SM, Hassan S, Abu Haris N. A Review of KDD-Data Mining Framework and Its Application in Logistics and Transportation. *International Journal of Supply Chain Management* 2016, 5:77-84.
35. Lukáčová A, Babič F, Paralič J. Building the prediction model from the aviation incident data. In: *Applied Machine Intelligence and Informatics (SAMI), 2014 IEEE 12th International Symposium on*: IEEE; 2014.
36. CRISP-DM: A PARALLEL OVERVIEW–.
37. Poll: What main methodology are you using for data mining? Available at: http://www.kdnuggets.com/polls/2002/methodology.htm.
38. Poll: Data Mining Methodology. Available at: http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm.
39. Poll: Data Mining Methodology. Available at: http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm.
40. Duan L, Xiong Y. Big data analytics and business analytics. *Journal of Management Analytics* 2015, 2:1-21, doi:10.1080/23270012.2015.1020891.
41. Mandala IGNN, Nawangpalupi CB, Praktikto FR. Assessing Credit Risk: An Application of Data Mining in a Rural Bank. *Procedia Economics and Finance* 2012, 4:406-412, doi:http://dx.doi.org/10.1016/S2212-5671(12)00355-3.
42. Koyuncugil AS, Ozgulbas N. Financial early warning system model and data mining application for risk detection. *Expert Systems with Applications* 2012, 39:6238-6253, doi:http://dx.doi.org/10.1016/j.eswa.2011.12.021.
43. Bozdogan C, Zincir-Heywood N. Data mining for supporting it management. In: *Network Operations and Management Symposium (NOMS), 2012 IEEE*: IEEE; 2012.

44. Afolabi IT, Worlu RE, Uwadia OC. Data Mining Approach for Target Marketing SMEs in Nigeria. *Covenant Journal of Informatics and Communication Technology* 2016, 4.
45. Packianather MS, Davies A, Harraden S, Soman S, White J. Data Mining Techniques Applied to a Manufacturing SME. *Procedia CIRP* 2017, 62:123-128.
46. Piatetsky-Shapiro G. Discovery, analysis, and presentation of strong rules. *Knowledge discovery in databases* 1991:229-238.
47. Mariscal G, Marbán Ó, Fernández C. A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review* 2010, 25:137-166, doi:10.1017/s0269888910000032.
48. Marbán Ó, Mariscal G, Segovia J. A data mining & knowledge discovery process model. *Data Mining and Knowledge Discovery in Real Life Applications* 2009, 2009:8.
49. Wong J-Y, Chung P-H. Managing valuable Taiwanese airline passengers using knowledge discovery in database techniques. *Journal of Air Transport Management* 2007, 13:362-370, doi:10.1016/j.jairtraman.2007.07.001.
50. Haluzová P. Effective data mining for a transportation information system. *Acta Polytechnica* 2008, 48.
51. Shin I-H, Park G-L, Saha A, Kwak H-y, Kim H. Analysis of Moving Patterns of Moving Objects with the Proposed Framework. 2009, 5593:443-452, doi:10.1007/978-3-642-02457-3_38.
52. Mirabadi A, Sharifian S. Application of association rules in Iranian Railways (RAI) accident data analysis. *Safety Science* 2010, 48:1427-1435, doi:10.1016/j.ssci.2010.06.006.
53. Zhang C, Huang Y, Zong G. Study on the Application of Knowledge Discovery in Data Bases to the Decision Making of Railway Traffic Safety in China. In: *Management and Service Science (MASS), 2010 International Conference on*; 2010, doi:10.1109/ICMSS.2010.5577012.
54. de Almeida J, Ferreira JC. BUS public transportation system fuel efficiency patterns. In: *Int. Conf. on Machine Learning and Computer Science, Kuala Lumpur, Malaysia*; 2013.
55. Raju Y, Kang PS, Moroz A, Clement R, Hopwell A, Duffy A. Investigating the Demand for Short-shelf Life Food Products for SME Wholesalers. 2016.
56. Rebón F, Castander Ii, Argandoña J, Gerrikagoitia JK, Alzua-Sorzabal A. An Antifraud System for Tourism SMEs in the Context of Electronic Operations with Credit Cards. *American Journal of Intelligent Systems* 2015, 5:27-33, doi:10.5923/j.ajis.20150501.03.
57. Pytel P, Britos P, García-Martínez R. A Proposal of Effort Estimation Method for Information Mining Projects Oriented to SMEs. In: Poels G, ed. *Enterprise Information Systems of the Future: 6th IFIP WG 8.9 Working Conference, CONFENIS 2012, Ghent, Belgium, September 19-21, 2012, Revised Selected Papers*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013, 58-74.
58. Cheung C, Li F. A quantitative correlation coefficient mining method for business intelligence in small and medium enterprises of trading business. *Expert systems with applications* 2012, 39:6279-6291.
59. Young T, Fehskens M, Pujara P, Burger M, Edwards G. Utilizing data mining to influence maintenance actions. In: *AUTOTESTCON, 2010 IEEE*: IEEE; 2010.
60. Inc SI. SAS version 9.1. 2005.
61. Shmueli G, Patel NR, Bruce PC. Data Mining for Business Intelligence. 2011.

62. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R. CRISP-DM 1.0 Step-by-step data mining guide. 2000.
63. Kdnuggets. What main methodology are you using for your analytics, data mining, or data science projects? Poll. Available at: http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html.

Figure Legend

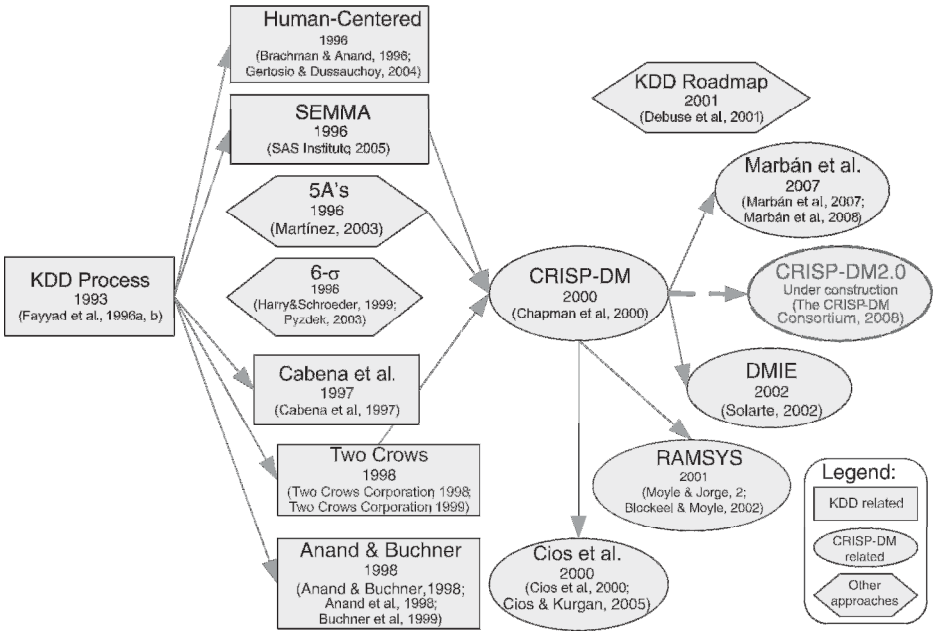


Figure 1. Evolution of Data Mining Methodologies
The figure represents the evolution of data mining methodologies from the 1990s to 2000s.

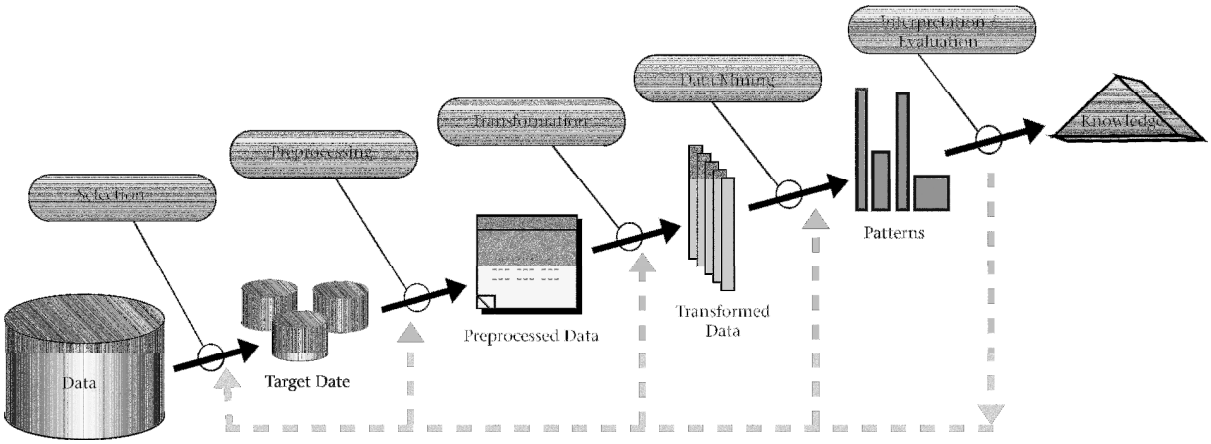


Figure 2. Knowledge Discovery in Database [36]
The figure represents the overall KDD methodology process.

Figure 3. Sample, Explore, Modify, Model, Assess [36] Methodology
The figure represents the overall SEMMA methodology process.

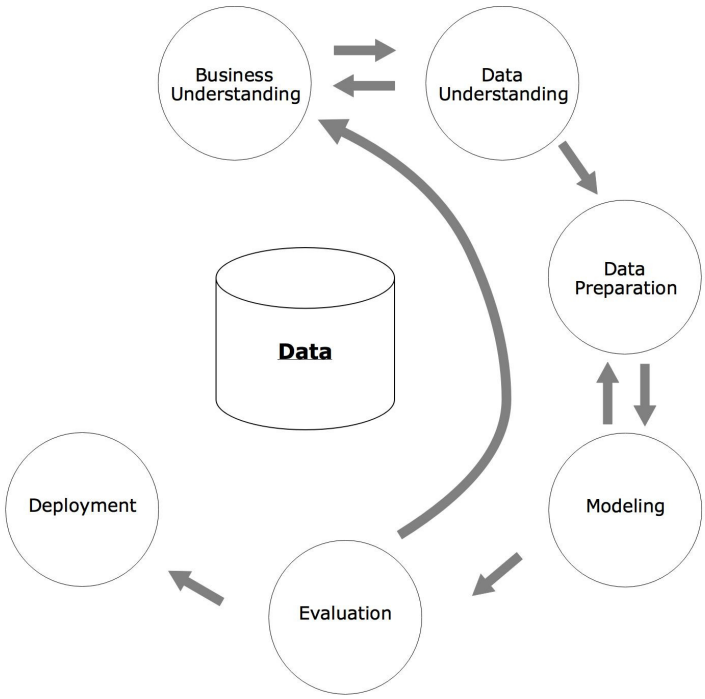


Figure 4. CRISP-DM Methodology
The figure represents the overall CRISP-DM methodology process.