Problemy Zarządzania – Management Issues, vol. 17, no. 2(82): 233–251
ISSN 1644-9584, © Wydział Zarządzania UW
DOI 10.7172/1644-9584.82.13

# Data Mining Process Maturity – Result of Empirical Research

Submitted: 07.01.19 | Accepted: 28.03.19

**Piotr Sliż***

The main goal of the article is to present the results of the study relating to the assessment of data mining process maturity on the example of Polish organizations. Several partial objectives were added to the main goal. CT1: To diagnose the current state of knowledge regarding the data-mining process in the discipline of management sciences. Attempts at attaining this objective served to identify the knowledge gap. CT2: To adopt an appropriate theoretical perspective in the form of a theoretical model, enabling the implementation of future research challenges. The first section of the article describes the results of quantitative and qualitative bibliometric analysis. The second section presents the parameters and the definition of the data mining process. Then, the theoretical model used for measuring the maturity of the data mining process is discussed. In the fourth section, the structure of the empirical research conducted and its partial results are outlined. It transpired that the vast majority of the surveyed organizations qualified at the first level of process maturity, defined as a state in which organizations are not aware of the need to identify activities aimed at data mining. Research objectives formulated in the article have been implemented using such research methods as quantitative and qualitative bibliometric analysis, opinion polls and statistical methods.

**Keywords:** data mining, data mining process, process management, process maturity.

# Dojrzałość procesu eksploracji danych – wynik badania empirycznego

Nadesłany: 07.01.19 | Zaakceptowany do druku: 28.03.19

Celem głównym artykułu było przedstawienie wyników badania oceny dojrzałości procesu eksploracji danych na przykładzie polskich organizacji. Realizacji celu głównego przyporządkowano cele cząstkowe. CT1: Określenie istniejącego stanu wiedzy dotyczącego data-mining process w dyscyplinie nauk o zarządzaniu. Podjęta próba realizacji tego celu służyła identyfikacji luki poznawczej. CT2: Przyjęcie odpowiedniej perspektywy teoretycznej w postaci modelu teoretycznego, umożliwiającego realizację przyszłych wyzwań badawczych. W pierwszej sekcji artykułu opisano wyniki ilościowej i jakościowej analizy bibliometrycznej. Następnie, w sekcji drugiej przedstawiono parametry i definicję procesu eksploracji danych. W sekcji następnej przedstawiono model teoretyczny, wykorzystany do pomiaru dojrzałości procesu eksploracji danych. W sekcji czwartej, w wyniku zrealizowanego postępowania empirycznego scharakteryzowano strukturę badania oraz cząstkowe wyniki. W jego rezultacie stwierdzono, że zdecydowana większość

* **Piotr Sliż** – dr inż., Institute of Organisation and Management, Faculty of Management, University of Gdańsk. https://orcid.org/0000-0001-6776-3369.

Correspondence address: Faculty of Management, University of Gdańsk. 101 Armii Krajowej Street, 81-824 Sopot.

Ministry of Science and Higher Education
Republic of Poland

The creation of the English-language version of these publications is financed in the framework of contract No. 607/P-DUN/2018 by the Ministry of Science and Higher Education committed to activities aimed at the promotion of education.

badanych organizacji została zakwalifikowana do pierwszego poziomu dojrzałości procesu, definiowanego jako stan, w którym organizacje nie wykazują świadomości potrzeby identyfikacji działań zmierzających do eksploracji danych. Sformułowane w artykule cele badawcze zostały zrealizowane z wykorzystaniem takich metod badawczych, jak: ilościowa i jakościowa analiza bibliometryczna, sondażowe badanie opinii oraz metody statystyczne.

**Słowa kluczowe:** eksploracja danych, proces eksploracji danych, zarządzanie procesami, dojrzałość procesów.

**JEL:** O310, O330, M210

## 1. Introduction

When presenting modern concepts and management methods, J. Brilman, formulated a thesis that the voice of customers penetrates deep into the enterprise through processes (2002, p. 287). In contemporary market realities, the development of process management is determined by the turbulent market environment, the derivative of which is a dynamic growth in the implementation of information technologies. This has a positive effect on communication with the client, but also enables precise measurement of processes, implementation of intelligent processes and designing new generation processes, identified as editable data and, in the distant future, defined as the executed code (Morzy, 2005, p. 258). Examples of such solutions include: artificial intelligence (AI), machine learning (ML), virtual reality, Internet of Things and robotization of processes (RPA). The progressive digitalization of contacts with clients implies the need to transfer organization's relations with stakeholders from the verbal to the digital level. At this point, it must be emphasized that the described technical factor, conducive to communication with the environment and the implementation of elements of the process approach in management, should be expanded with systemic, cultural and competence factors (Grajewski, 2016, pp. 197–199). With further specification, for the organization to reach a state in which it discounts the benefits resulting from the implementation of the process approach, in addition to increased computerization in the organization, it must afford to process participants a greater freedom of influencing its course and, as a result, reconfigure laminar processes into turbulent ones. It also requires understanding the growing phenomenon of presumption, redefining the concept of the customer, thus far understood as the recipient of manufactured products and services, towards a prosumer, defined as a co-creator of process modelling in the organization (Czubasiewicz, Grajewski & Sliż, 2018, p. 244). As a result, through the intelligent architecture of processes, the client's voice can reach the organization and have an even deeper effect. For this to happen, it must be identified as data collected and explored using modern methods and tools. The increasing supply of huge volumes of data, generated

 DOI 10.7172/1644-9584.82.13

both in organizations and in their market environment, is a derivative of this state. The solution enabling their processing is data mining, referred to as a well-established field of science (Fayyad, Piatetsky-Shapiro & Smyth, 1996, pp. 7–34; Chen & Liu, 2005; Wang, 2005; Pechenizkiy, Puuronen & Tsymbal, 2005, pp. 67–71). Data mining is defined in literature as a business process in which activities related to the exploration and analysis of large amounts of data are carried out in order to discover theses and rules (Linof & Berry, 2011, p. 6). Synonyms of data mining are data fusion, data analysis and decision support (Suzhen, 2018, p. 1).

Following an analysis of the Web of Science database, it can be concluded that the issue of data mining is interdisciplinary in nature and usually encompasses such categories as computer science, engineering and mathematics; in this article, it is explored in the context of business management. Certain themes related to data mining techniques and tools shall be examined, in an attempt to look at this process holistically from the perspective or two research questions (RQ). RQ1: Do activities carried out in the course of data mining bear some semblance to a business process? RQ2: What is the level of maturity of the data mining process in Polish organizations?

The main goal of the article is to present the results of the study relating to the maturity assessment of the data mining process on the example of Polish organizations. Partial epistemological objectives have been formulated in the article (CT). CT1: To diagnose the current state of knowledge regarding the data mining process in selected organizations. CT2: To construct a theoretical model and to adopt an appropriate theoretical perspective for empirical research.

The theoretical and empirical study was carried out within the framework of a project financed from funds for scientific research and development and related tasks aimed at the development of young scientists (project No. 538-2200-B128-18).

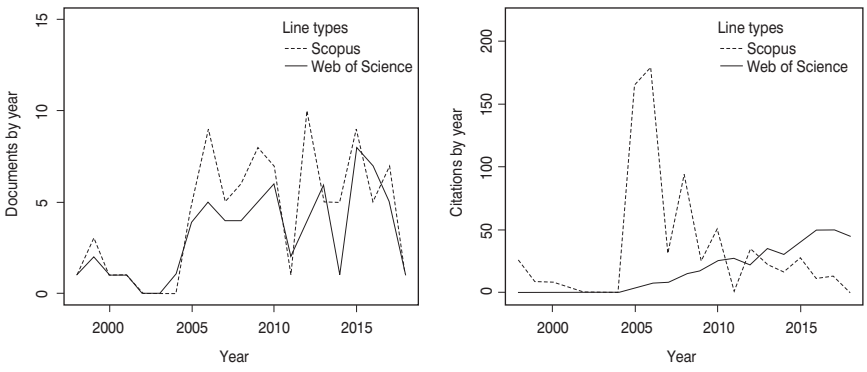
## 2. Literature review

The examination of issues explored in this article was induced by a theoretical study. For this purpose, quantitative and qualitative bibliometric analysis was carried out. Based on its results, a knowledge gap was identified: the absence of publications regarding the maturity assessment of the data mining process. Next, a theoretical model was constructed, as a result of which a research tool was designed, which forms the basis for the design and implementation of the empirical procedure.

### 2.1. Quantitative bibliometric analysis

As a result of the quantitative bibliometric analysis, the following parameters were verified: the number of publications and the number of

citations. The analysis was based on two scientific databases: Web of Science and Scopus. For the *article title* category in the Web of Science database, a total of 68 publications were identified with respect to entries: *data mining process* (57) and *data mining processes* (11).




1a 1b

*Fig. 1. The number of publications (1a) and the number of citations (1b) based on Scopus and Web of Science databases. Source: own study based on data from Scopus, Web of Science and Google Trends using the R programming language.*

In turn, 89 publications were identified for both entries in the Scopus database using the same search criterion (Figure 1a and 1b). Only 11.23% were classified in business, management and accounting categories.

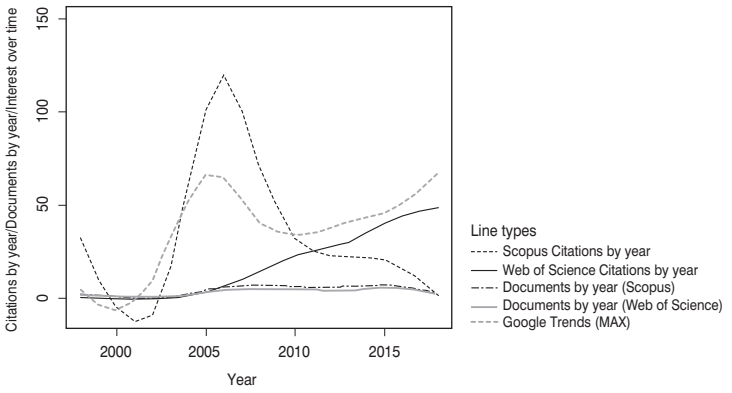



*Fig. 2. The number of publications, citations and popularity of research entries using LOESS regression with the 50 basis. Source: own study based on data from Scopus, Web of Science and Google Trends using the R programming language.*

 DOI 10.7172/1644-9584.82.13

The results of the quantitative bibliometric analysis using LOESS regression at the 50% basis are presented in Figure 2 (Cleveland, 1979; Cleveland, 1981; Cleveland & Devlin, 1988; Cleveland, Devlin & Grosse, 1988). In addition, the results of the data mining process popularity analysis using the Google Trends tool have been included in lines exemplifying the number of citations and publications. Presented results testify to a stable number of publications within the examined time series and a growing interest in the matter (Figure 2).

### 2.2. Qualitative bibliometric analysis

As a result of the qualitative analysis for the *data mining process* entry, the following types of studies were identified using Scopus and Web of Science databases: proceedings papers, articles, book chapters, reviews and editorials materials. It was observed that in the vast majority of cases, they refer to examples of the application of data mining in a wide spectrum of impact on such areas as distributions patterns of energy consumption (Chen & Wu, 2018, pp. 240−251), medicine (Gironi, Saresella, Rovaris, Vaghi, Nemni, Clerici & Grossi, 2013; Gironi, 2014; Becerra-García, García-Bermúdez, Joya-Caparrós, Fernández-Higuera, Velázquez-Rodríguez, Velázquez-Mariño & Rodraguez-Labrada, 2017, pp. 28−36; Mansingh, Osei-Bryson & Asnani, 2016; Tavares, Paredes, Rocha, Carvalho, Ramos, Mendes & Morais, 2015, p. 52−65; Gironi, Borgiani, Farina, Mariani, Cursano, Alberoni & Grossi E., 2015), soil testing (Belabed, Alaoui, Belabed & Alaoui, 2017, pp. 234−244), spatial development (Besheli, Zare, Umali &, Nakhaeezadeh, 2015, pp. 821−835), behaviour profiling of bank customers (Mansingh, Rao, Osei-Bryson & Mills, 2015, pp. 193–215), and hydrology (Keskin, Taylan & Kucuksille, 2013). At this point, it must be emphasized that the range of data mining impact is much broader and concerns the vast majority of areas of economy and social life.

In the data mining process, the following methods are used: association discovery, classification and prediction, grouping (cluster analysis, clustering), sequence and time sequence analysis, discovery of characteristics, text exploration, website exploration, graph exploration, social network exploration, multimedia data and spatial data exploration, and the detection of singularities (Hinneburg & Keim, 1998; Tan, Steinbach & Kumar, 2006). In addition, as a result of a theoretical study, entries similar to *data mining process* were identified. They were: *big-data mining process* (Song, Jung & Chung, 2017, pp. 1–10) and *visual data mining process* (VDMP) (Ltifi, Benmohamed, Kolski & Ayed, 2016).

## 3. Data mining as a process

### 3.1. Overview of definitions of the data mining process

As the result of the theoretical study, Table 1 presents definitions of the data mining process.

| Author/Authors | Year | Definition |
|---|---|---|
| M. Kantardzic (2011, p. 6) | 2011 | "Data mining is a process of discovering various models, summaries, and derived values from a given collection of data". |
| D. Jannach, S. Fischer (2014, p. 337) | 2014 | "Practical data mining processes typically consist of several sequential steps including, for example, data-pre-processing, feature selection, model learning or performance evaluation". |
| K. Atanassov (2015, p. 5) | 2015 | "Data Mining is a process of finding reasonable correlations, repeating patterns and trends in large DBs". |
| K.M. Osei-Bryson, C. Barclay (2015, p. 11) | 2015 | "Data mining is the process of extracting useful, relevant knowledge from any data repository". |
| G. Mansingh, L. Rao, K.M. Osei-Bryson, A. Mills (2015, p. 193) | 2015 | "Data mining is the process of extracting interesting patterns and trends from large datasets". |
| Q. Suzhen (2018) | 2018 | "Data Mining is the extraction of information that is hidden in the prior, but potentially useful, from a large number of incomplete, noisy, fuzzy, and random application data". |

*Tab. 1. Selected definitions of the data mining process. Source: own study based on the selected literature.*

In the majority of studies under analysis, no terminological discussions on the characterization of the data mining process were found. In addition, proposals are superficial and, most of all, describe the type of activities carried out in the process.

Table 2 presents the characteristics of activities in the data mining process.

| Data mining process | |
|---|---|
| Data cleaning | Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data (Rahm E., Do H.H., 2000, p. 3). |
| Data integration | Data integration is the problem of combining data residing at different sources, and providing the user with a unified view of these data (Lenzerini M., 2002, p. 233). |
| Data selection | Selecting data or variables, which are useful in further activities of the data mining process. |
| Data transformation | Performing data transformation using data mining tools, methods and techniques, to a form useful in further stages of the data mining process. |
| Data visualization | Data visualization is very important for humans to understand structural relations among variables within a system. It is also a critical step to eliminate some unrealistic models (Yang H.H., Moody J., 2000, p. 668). |

*Tab. 2. General characteristics of activities carried out in the data mining process. Source: own study.*

 DOI 10.7172/1644-9584.82.13

As a result of the theoretical study, the stage referred to as the pre-processing of data was also observed, including activities related to cleaning the noise information, missing values and scaling outliers by normalizing data (Belabed, Alaoui, Belabed & Alaoui, 2017).

### 3.2. Data mining process and the mega-process of knowledge discovery and diffusion

In this section of the article, we shall attempt to identify the effect of the data mining process. First, it was necessary to specify whether the identified and formalized set of actions, presented in Table 2, should be regarded as a separate process in process architecture, or as a set of activities in the process of knowledge discovery and diffusion. This required us to clarify the effect of the set of data mining activities (Table 2) and to answer the following question: is knowledge or information the result of data mining? At this point, it must be emphasized that knowledge in this article is defined as "information combined with experience, context, interpretation, and reflection" (Davenport, De Long & Beers, 1998), while information as "data endowed with relevance and purpose" (Drucker, 1998). According to P. Beynon-Davies, information is data equipped with semantics (2003, p. 15).

As a result of the theoretical study, arguments for identifying information as the effect of the data mining process were presented.

First of all, links between data mining and knowledge management are described in the literature, taking into account their independence within the organization (Wang H., Wang S., 2008). Secondly, assuming that the effect (result) of the process should be consistent with customers' expectations (Davenport, 1993, p. 5; Rummler & Brache, 2000, p. 75), the interactive nature of the data mining process implies limitations resulting from the inability to determine what can be discovered in databases prior to their exploration (Han & Kamber, 2011, p. 36). In turn, the identification of expectations with respect to database exploration activities in terms of obtaining information, e.g. regarding product segmentation based on sale transaction data and using cart analysis, can be specified by the external or internal client. This thesis can be confirmed by the words of T.H. Davenport and C.P. Seely, according to whom knowledge management "is concerned with human subjective knowledge, not data or objective information" (2006). Moreover, H. Wang and S. Wang., describing the Data Mining model developed by M.J.A. Berry and G.S. Linoff (2000) and its individual stages state that it ignores the essential element – knowledge. At the same time, they point to the limitations of the practical application of the model, emphasizing two aspects: "first, people often find that 'knowledge' gained from DM does not always lead to an action in all situations, particularly when the piece of 'knowledge' is hard to apply. In fact, this model overstates the role of DM in action, and in turn fails to recognize the roles of business insiders in developing their knowledge for coordination of actions for

business. Second, this model mixes non-sequential processes into a single cycle, and de-emphasizes distinctive roles of different people involved in DM for BI" (2008, pp. 624–625).

It is claimed that the discovery of knowledge is identified as the entire process; data mining is one of its stages (Morzy, 1999, pp. 3–4). V. Medvedev, O. Kurasova, J. Bernatavičienė, P. Treigys, P., V. Marcinkevičius and G. Dzemyda share this opinion. They conside data mining as an important part of knowledge discovery processes in areas such as medicine, economics, finance, telecommunication, and various scientific fields (2015, p. 11).

The study assumes that activities implemented in the data mining process are defined as supporting the described process of knowledge discovery and diffusion. This means that information is regarded as the effect of the data mining process.

### 3.3. Parameters of the data mining process

In this section of the article, based on the key parameters of the business process (Rummler & Brache, 2000, p. 75; Kaplan & Norton, 2001, p. 43; Irani & Hlupic, 2002, pp. 5–10; Grajewski, 2003, p. 104; Sliż, 2018, pp. 12–17), parameters of the data mining process have been outlined.

- **Sequence of activities** – activities related to data mining have a structured course described in a simplified model according to M. Kantardzic: state the problem, collect the data, pre-process the data, estimate the model (mine the data), interpret the model and draw conclusions (2011, p. 9). In the literature, the sequence of activities is also divided into process stages, such as purification, integration, selection, transformation and presentation of data that are identified as phases or stages of the process (Azevedo & Santos, 2008). These are:
- **the client** – defined in external and internal terms (Brillman, 2002, p. 286).
- **the supplier** – identified on the basis of the specification of the environment being the source of data. It should be understood that feeds for the data mining process may come from the environment, but also from within the organization.
- **inputs** – defined as structured or unstructured data of various types, which "defines a collection of data values and a set of predefined operations on those values" (Sebesta, 2016, p. 260).
- **added value** – it is generated by actions in subsequent stages of processes and can be calculated on the basis of the final effect (Cyfert, 2014, pp. 232–233), assuming that each activity should add new value to the effect of an earlier activity (Porter, 1985, p. 3).
- **outputs** – messages, information, patterns, graphical representation of data, enabling the generation of new knowledge in basic processes of the organization.

 DOI 10.7172/1644-9584.82.13

In summary, **the data mining process is defined as a highly flexible set of actions transforming with the use of information technology and statistical or mathematical methods intangible inputs of a structured or unstructured nature into structured or intangible outputs in order to generate added value consistent with the client's expectations in external or internal terms**.

In addition, the data mining process should be qualified as a group of intelligent processes, i.e. those that have a system of using their own knowledge in the structure to optimize the flow of individual operations from the perspective of results that are estimated each time (Grajewski, 2012, p. 59).

Creating an environment that enables flexible and effective management of the data mining process requires:

- ensuring that process participants have a greater freedom in influencing its structure and course;
- designing the flow of activities performed by humans and robots (RPA) at the same time;
- using modern methods and tools that enable the generation of the process effect (Grajewski, 2003, p. 104);
- reorientation of the manager's role towards the position of a leader managing the diffusion of knowledge within the organization;
- appointing process owners to design processes in different versions depending on the final effect, as well as ensuring a high level of self-organization capacity of the process;
- building highly interdisciplinary teams;
- creating solutions that enable the changing and overlapping of roles instead of their precise description, which contributes to postponing their petrification; – creating an organizational culture based on the cult of knowledge (Grajewski, 2016, p. 29; Czubasiewicz, Grajewski & Sliż, 2018, p. 63).

More precisely, the adoption of these parameters limits the possibilities of a detailed design of the process. This means that process boundaries are determined by the parameters presented earlier, as well as the type of project or transaction implemented.

## 4. A model for assessing the maturity of the data mining process

A five-degree maturity model was used to assess the level of maturity of the data mining process. The model presented in Table 3 has a descriptive function (Becker, Knackstedt & Pöppelbuß, 2009).

| Level | Level description | Level characteristics |
|---|---|---|
| L1 | Lack of awareness of the need to identify processes | The organization is unaware of the need to identify a set of activities related to data mining in the process. Therefore, it does not make decisions that allow its formalization, measurement and management. |
| L2 | Identified and formalized process | The organization is aware of the need to identify data mining activities. Data is stored in the organization's databases. The use of IT tools for data processing is fragmentary, and their implementation is determined by legal regulations (e.g. protection of personal data) or by external decisions (e.g. by an external planning centre). The organization collects available data generated through customer inquiries, sales transactions or traffic on the website. The organization is aware of the need to expand knowledge, but training in the field of data mining relates to the implementation of new IT tools. Despite the organisation's awareness of the data mining process, it is not measured. |
| L3 | Measured process | The organization uses basic measures of the data mining process, such as the cost, length or revenues generated by the process. Superficial symptoms indicating management decisions based on the results of measuring the data mining process are noticeable in the organization. |
| L4 | Managed process | The data mining process is managed in the organization. This means that, on the basis of information obtained in the data mining process, decisions are made in the space of all processes in the organization. The organization manages the intellectual potential of its employees. This means that cycles of training are carried out by specialist companies or employees undertake post-graduate studies, e.g. in the field of data science. |
| L5 | Improved process | The organization deliberately discounts the benefits of measuring and managing the data mining process. This means that it looks for new areas in which data will be generated to feed the data mining process. |

*Tab. 3. A five-degree model for assessing the maturity of the data mining process. Source: own study.*

In the presented model it is assumed that an organization at a higher level has fulfilled the requirements of all levels below.

## 5. Assessment of the maturity of the data mining process – results of the empirical study

### 5.1. Structure of empirical proceedings

Empirical proceedings were carried out in 2018 on a sample of 97 objects selected using a non-probabilistic method with a targeted choice, out of 350 Polish organizations that participated in the study relating to the assessment of process maturity (Sliż, 2018). The study was carried out

 DOI 10.7172/1644-9584.82.13

in organizations whose genotypic activity was not related to data mining. The structure organizations classified according to their size is presented in Table 4. The study was carried out using the CAWI technique. The questionnaire was addressed to representatives of the management in the surveyed organizations. The structure of respondents was as follows: president/owner (59.79%), director/manager (11.34%), manager/team leader (12.37%) and expert/specialist (16.49%).

The questionnaire used in empirical proceedings contained questions related to the following areas:

- **v1** – identification of a set of planned activities related to data mining as a process;
- **v2** – types of databases in which organization's data is stored;
- **v3** – identification of data sources to be explored;
- **v4** – types of training implemented in the field of data mining;
- **v5** – types of tools used (software) in the data mining process;
- **v6** – types of data mining methods used;
- **v7** – the data mining system used for measuring the data mining process;
- **v8** – types of activities carried out in the data mining process.

Moreover, on the basis of the obtained answers, the reliability of the research tool was verified using the Alfa Cronbach test. The result was 0.79, and standardized for Alfa = 0.84 (with the average correlation between items = 0.39), which proves a high degree of reliability of the questionnaire.

### 5.2. Results of empirical proceedings

First, respondents were asked to identify activities related to data mining. Their answers are presented in Table 5.

| Class | Size of the organization [number of employees] | | | |
|---|---|---|---|---|
| | **Up to 9** | **From 10 to 49** | **From 50 to 249** | **More than 250** |
| No, but the organization plans to appoint a person to implement the data mining process. | 2,06% | 2,06% | – | – |
| No. | 30,93% | 17,53% | 2,06% | 2,06% |
| Yes, but activities related to data mining are carried out by an external company. | 1,03% | 2,06% | – | – |
| Yes, the organization carries out data mining activities. | 17,53% | 10,31% | 8,25% | 4,12% |
| Total | 51,55% | 31,96% | 10,31% | 6,19% |

*Tab. 4. Identification of activities related to data mining. Source: own study using Tableau.*

Based on the results obtained, presented in Table 4, it was concluded that the vast majority of entities are unaware of the need to identify data mining activities.

Then, the variability of groups of organizations was verified taking into account their size. The variability index v was verified, as expressed by the formula (1). The results are shown in Table 5.

$$v = \frac{s}{\bar{x}} \cdot 100\%. \tag{1}$$

Where:
$s$ – standard deviation
$\bar{x}$ – arithmetic mean

The analysis of the variability index demonstrated that as the size of the organization (micro, small and medium-size) increases, the value of the index decreases (table 5).

| Measure | Number of employees | | | |
|---|---|---|---|---|
| | **<1,9>** | **<10;49>** | **<50;249>** | **<249** |
| Average | 2,879 | 3,35 | 6,87 | 7,241667 |
| Standard deviation | 2,110713 | 1,788575 | 2,786794 | 3,325119 |
| Value of the V coefficient | 0,733141 | 0,533903 | 0,405647 | 0,459165 |

*Tab. 5. Results of the variability index V. Source: own study using Statistica13.*

Then, the Chi-square ($\chi 2$) independence test by Pearson was used. The analysis verified the variables, identified as the size of the enterprise. To this end, the following statistical hypotheses were formulated:

**H0**: the studied variables are dependent ($p>0.05$)
**HA**: the variables are not independent ($p<0.05$)

As a result, the $p < 0.05$ result was obtained. Therefore, the null hypothesis was rejected. This means that the result of the study of the maturity of the data mining process depends on the size of the organization.

In turn, Table 6 presents the distribution of answers to the question about the types of databases in which data in the organization is stored.

 DOI 10.7172/1644-9584.82.13

| Class | Company's transactional databases | Data warehouses | Company's websites (e.g. store, order site) | Relational databases | Data cloud | Data sets on individual computers depending on the person's position in the organisation | Other: |
|---|---|---|---|---|---|---|---|
| Yes, but the activities related to data mining are carried out by an external company. | 2.06% | 0.00% | 0.00% | 1.03% | 0.00% | 0.00% | 0.00% |
| Yes, the organization carries out activities related to data mining. | 14.43% | 6.19% | 7.22% | 7.22% | 5.15% | 21.65% | 1.03% |
| No, but the organization plans to appoint a person to implement the data mining process. | 0.00% | 0.00% | 0.00% | 0.00% | 2.06% | 2.06% | 1.03% |
| No. | 13.40% | 0.00% | 4.12% | 3.09% | 2.06% | 29.90% | 1.03% |

*Tab. 6. Types of databases in which organization's data is stored (variable v2). Source: own study using Tableau.*

Table 7 presents the answers to the question about sources of data generation in the surveyed organization.

| Class | Yes, by an external company | | Yes, in the organization | |
|---|---|---|---|---|
| | Number | Percentage share in the group | Number | Percentage share in the group |
| Traffic analysis on the company's website | 0.00 | 0.00% | 5.00 | 4.95% |
| Data from the measurement of activities in the organization (e.g. sales, delivery, production, quality control processes, etc.) | 0.00 | 0.00% | 13.00 | 12.87% |
| Correspondence with clients (forms on the website, e-mail) | 2.00 | 33.33% | 23.00 | 22.77% |
| The organization obtains data from external sources | 0.00 | 0.00% | 11.00 | 10.89% |
| Sale transactions | 3.00 | 50.00% | 27.00 | 26.73% |
| Customer inquiries | 1.00 | 16.67% | 19.00 | 18.81% |
| Own answer (other) | 0.00 | 0.00% | 3.00 | 2.97% |
| Total | 6.00 | 100.00% | 101.00 | 100.00% |

*Tab. 7. Sources of data acquisition (variable v3). Source: own study using Tableau.*

In the surveyed group of organizations, an attempt was made to verify the management of employees' intellectual potential related to data mining. The following results were obtained: in 41.24% of organizations, employees themselves broaden their knowledge, in 14.43% trainings are organised by software providers, in 18.56% additional trainings are carried out by external companies, in 12.37% e-learning is provided, and only in 3.09% employees embark on post-graduate studies. In 30.93% of the surveyed group of organizations, training is not organised.

The assessment of tools used for data processing shows that the vast majority of organizations used Microsoft Excel (81.44%) and Access (12.37%). Individual answers were obtained for the Tableau program (3.09%) and the Python programming language (2.06%). The use of such tools as R, SAS, SPSS and Statistica language has not been recorded. Respondents also indicated other tools (8.25%): MS Word, Inelo 4Trans, Marcosbis Nawigator, PostGresq, MiniTab, PHP and SAP BO).

| Class | Size of the organization [number of employees] | v6.1. | v6.2. | v6.3. | v6.4. | v6.5. | v.6.6. |
|---|---|---|---|---|---|---|---|
| Yes, but the activities related to data mining are carried out by an external company. | up to 9 | 0.00% | 0.00% | 1.03% | 1.03% | 1.03% | 0.00% |
| | from 10 to 49 | 1.03% | 0.00% | 2.06% | 1.03% | 0.00% | 1.03% |
| Yes, the organization carries out activities related to data mining. | up to 9 | 8.25% | 0.00% | 5.15% | 2.06% | 4.12% | 10.31% |
| | from 10 to 49 | 7.22% | 0.00% | 2.06% | 2.06% | 6.19% | 8.25% |
| | from 50 to 249 | 7.22% | 1.03% | 5.15% | 4.12% | 6.19% | 7.22% |
| | more than 250 | 3.09% | 1.03% | 2.06% | 2.06% | 3.09% | 3.09% |
| No, but the organization plans to appoint a person to implement the data mining process. | up to 9 | 1.03% | 0.00% | 2.06% | 0.00% | 1.03% | 1.03% |
| | from 10 to 49 | 1.03% | 0.00% | 0.00% | 0.00% | 2.06% | 2.06% |
| No. | up to 9 | 10.31% | 0.00% | 14.43% | 0.00% | 2.06% | 8.25% |
| | from 10 to 49 | 2.06% | 0.00% | 9.28% | 0.00% | 1.03% | 4.12% |
| | from 50 to 249 | 2.06% | 0.00% | 2.06% | 0.00% | 1.03% | 0.00% |
| | more than 250 | 1.03% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |

*Tab. 8. Purpose of data mining activities. Source: own study using Tableau.*

 DOI 10.7172/1644-9584.82.13

Next, we verified the measures taken as part of the data mining process in the organization. The following answer variants are presented in the research questionnaire: v6.1. – adjusting the rules to the provisions of the law on protection and data processing, v6.2. – designing new positions and departments related to data mining, v6.3. – regular reviews and corrections of backups, v6.4. – development and implementation of modern information technologies to increase the efficiency of the data mining process, v6.5. – setting rules for data processing, v6.6. – ensuring the security of sensitive data (Table 8).

Then, respondents were asked about the measures used for the data mining process. Among organizations that identified the tested process, 14.43% measure the cost of the process implementation, 11.34% the level of the internal client's satisfaction (recipient of the data mining process), 10.31% the length of process implementation, while 8.25% the income generated by the process. Moreover, 20.62% of the respondents indicated that process measurements are not being carried out.

| Class | Percentage share |
|---|---|
| Planning and forecasting sales. | 25.77% |
| Defining profiles of clients. | 17.53% |
| Grouping customers and/or identifying market segmentation. | 16.49% |
| Searching for fraud or activities that do not fit the data trend. | 8.25% |
| Analysis of the use of the Web to find typical patterns of user behaviour on the website. | 7.22% |
| Analysis of customers' reactions to an advertising campaign or a social-media campaign. | 7.22% |
| Searching for natural patterns of consumer behaviour, e.g. through analysing which products/services are bought together | 7.22% |
| Analysis of content posted on social networks allowing the search for new information or evaluation and verification | 4.12% |

*Table 9. Objectives of the data mining process in the surveyed organizations (variable v8). Source: Own study.*

For the v8 variable, the objectives of activities carried out in the data mining process have been verified, i.e. those that have an impact on improving not only the data mining process, but above all the entire architecture of processes in the organization (Table 9).

## 6. Summary and conclusions

The article presents a new look at the issues of data mining from a management perspective. As a result of the theoretical and empirical

study carried out, four generalizing conclusions were formulated. At this point, it should be emphasized that the non-structuralist sampling technique used determines the presentation of the following conclusions only to the group of the surveyed organizations.

Firstly, as a result of the bibliometric analysis, it was found that the discussed issue is interdisciplinary in nature, and the growing share of studies in the management category may be a challenge for further discussions and scientific research.

The next conclusion, as a result of the empirical proceeding carried out, the classification of the organizations was made due to the level of maturity of the data mining process. The results were as follows: 58.76% qualified to level 1, 21.65% to level 2, 11.35%, to level 3, 5.15% to level 4 and 3.09% to level 5, the highest one.

Thirdly, as a result of the performed correlation analysis of the variables tested, positive correlations were observed between the following variables (for $p < 0.05$ and $N = 97$): *v6* and *v8* (0.5837), *v1* and *v7* (0.4992), *v3* and *v8* (0.4701), *v2* and *v6* (0.4414), *v6* and *v7* (0.4379), and *v7* and *v8* (0.4334).

The last conclusions encourages to look at the limitations and barriers of the use of data mining. According to H. Wang and S. Wang, according to whim "DM is considered to be useful for business decision making, especially when the problem is well defined. Because of this, DM often gives people an illusion that one can acquire knowledge from computers through pushing buttons. The danger of this misperception lies in the over-emphasis on "knowledge discovery" in the DM field and de-emphasis on the role of user interaction with DM technologies in developing knowledge through learning." (2018, p. 624–625).

This study should be considered as an introduction to a broader discussion regarding the mega-process of knowledge discovery and diffusion in the organization, with particular emphasis on the data mining process. This means that the next objective of the author is to explore the established empirical facts of data mining processes in organizations, ways of their measurement, management and improvement.

**References**

Atanassov, K. (2015). Intuitionistic fuzzy logics as tools for evaluation of Data Mining processes. *Knowledge-Based Systems*, *80*, pp. 122–130.

Azevedo, A.I.R.L., & Santos, M.F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. IADS-DM.

Becerra-García, R.A., García-Bermúdez, R.V., Joya-Caparrós, G., Fernández-Higuera, A., Velázquez-Rodríguez, C., Velázquez-Mariño, M., & Rodraguez-Labrada, R. (2017). Data mining process for identification of non-spontaneous saccadic movements in clinical electrooculography. *Neurocomputing*, 250, 28–36.

Becker, J., Knackstedt, R., & Pöppelbuß, J. (2009). Developing maturity models for IT management – a procedure model and its application. *Business & Information Systems Engineering*, *1*(3), 213–222.

 DOI 10.7172/1644-9584.82.13

Belabed, I., Alaoui, M.T., Belabed, A., & Alaoui, Y.T. (2017, April). *Analysis of Soil Data from Eastern of Morocco Based on Data Mining Process*. International Conference on Bioinformatics and Biomedical Engineering (pp. 234-244). Cham: Springer.

Berry, M.J.A., & Linoff, G.S. (2000). *Mastering Data Mining*. New York, NY: Wiley.

Besheli, P.R., Zare, M., Umali, R.R., & Nakhaeezadeh, G. (2015). Zoning Iran based on earthquake precursor importance and introducing a main zone using a data-mining process. *Natural Hazards*, *78*(2), pp. 821–835.

Beynon-Davies, P. (2003). *Systemy baz danych [Database systems]*. Warszawa: Wydawnictwa Naukowo-Techniczne.

Brilman, J. (2002). *Nowoczesne koncepcje i metody zarządzania [Modern Management Concepts and Methods]*. Warszawa: PWE.

Chen, S.Y., & Liu, X. (2005). Data mining from 1994 to 2004: an application-oriented review. *International Journal of Business Intelligence and Data Mining*, 1(1).

Chen, Y., & Wu, J. (2018). Distribution patterns of energy consumed in classified public buildings through the data mining process. *Applied Energy*, *226*, 240–251.

Cleveland, W.S., & Devlin, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, *83*(403).

Cleveland, W.S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *American Statistician*, *35*(1).

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, *74*(368).

Cleveland, W.S., Devlin, S. J., & Grosse, E. (1988). Regression by local fitting: methods, properties, and computational algorithms. *Journal of econometrics*, *37*(1).

Cyfert, S. (2014). *Organizacja i kierowanie [Organization and management]*. Warszawa: Komitet Nauk Organizacji i Zarządzania Polskiej Akademii Nauk i SGH w Warszawie.

Czubasiewicz, H., Grajewski, P., & Sliż, P. (2018). Dojrzałość procesowa hoteli i obiektów noclegowych w Polsce [Business Process Maturity of Hotels and Accommodation]. *Zeszyty Naukowe Politechniki Poznańskiej*, *76*, 243–257.

Davenport, T.H. (1993). *Process innovation: reengineering work through information technology*. Harvard Business Press.

Davenport, T.H., & Seely, C.P. (2006). KM meets business intelligence: merging knowledge and information at Intel. *Knowledge Management Review*, January/February, 10–15.

Davenport, T.H., De Long, D.W., & Beers, M.C. (1998). Successful knowledge management projects. *Sloan Management Review*, *39*(2), 43–57.

Drucker, P.F. (1998). The coming of the new organization. *Harvard Business Review on Knowledge Management*. Boston, MA: Harvard Business School Press.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11), 7–34.

Gironi, M. (2014). A global immune deficit in Alzheimer and mild cognitive impairment disclosed by a novel data mining process. *Journal of Neuroimmunology*, *275*(1), 60.

Gironi, M., Borgiani, B., Farina, E., Mariani, E., Cursano, C., Alberoni, M., & Grossi, E. (2015). A global immune deficit in Alzheimer's disease and mild cognitive impairment disclosed by a novel data mining process. *Journal of Alzheimer's Disease*, *43*(4), 1199–1213.

Gironi, M., Saresella, M., Rovaris, M., Vaghi, M., Nemni, R., Clerici, M., & Grossi, E. (2013). A novel data mining system points out hidden relationships between immunological markers in multiple sclerosis. *Immunity & Ageing*, *10*(1), 1.

Grajewski, P. (2003). *Koncepcja struktury organizacji procesowej [The concept of the structure of the process organization]*. Toruń: TNOiK "Dom Organizatora".

Grajewski, P. (2012*). Procesowe zarządzanie organizacją [Process management of the organization]*. Warszawa: PWE.

Grajewski, P. (2016). *Organizacja procesowa [Process organization]*. Warszawa: PWE.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Han, J. & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
Hinneburg, A., & Keim, D.A. (1998). An efficient approach to clustering in large multimedia databases with noise. Proceeding: 4th International Conference on Knowledge Discovery and Data Mining, pp. 58–65.
Irani, Z., Hlupic, V., Baldwin, L.P., & Love, P.E. (2000). Re-engineering manufacturing processes through simulation modelling. *Logistics Information Management*, *13*(1), 7–13.
Jannach, D., & Fischer, S. (2014, October). *Recommendation-based modelling support for data mining processes*. Proceedings of the 8th ACM Conference on Recommender systems (pp. 337–340). ACM.
Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.
Kaplan, R.S., Norton, D.P., Pniewski, K., Jarugowa, A., Polakowski, M., & Kabalski, P. (2001). *Strategiczna karta wyników: jak przełożyć strategię na działanie [Strategic scorecard: how to translate strategy into action]*. Warszawa: PWN.
Keskin, M.E., Taylan, D., & Kucuksille, E.U. (2013). Data mining process for modeling hydrological time series. *Hydrology Research*, 44(1), 78–88.
Lenzerini, M. (2002, June). *Data integration: A theoretical perspective*. Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 233–246). ACM.
Linoff, G.S., & Berry, M.J. (2011). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons.
Ltifi, H., Benmohamed, E., Kolski, C., & Ayed, M.B. (2016). Enhanced visual data mining process for dynamic decision-making. *Knowledge-Based Systems*, *112,* 166–181.
Mansingh, G., Osei-Bryson, K.M., & Asnani, M. (2016). Exploring the antecedents of the quality of life of patients with sickle cell disease: using a knowledge discovery and data mining process model-based framework. *Health Systems*, *5*(1), 52–65.
Mansingh, G., Rao, L., Osei-Bryson, K.M., & Mills, A. (2015). Profiling internet banking users: A knowledge discovery in data mining process model based approach. *Information Systems Frontiers*, *17*(1), 193–215.
Medvedev, V., Kurasova, O., Bernatavičienė, J., Treigys, P., Marcinkevičius, V., & Dzemyda, G. (2017). A new web-based solution for modelling data mining processes. *Simulation Modelling Practice and Theory*, 76, 34–46.
Morzy, T. (1999). *Eksploracja danych: problemy i rozwiązania [Data mining: problems and solutions]*. Proceedings of the PLOUG Conference.
Osei-Bryson, K.M., & Barclay, C. (eds). (2015). *Knowledge Discovery Process and Methods to Enhance Organizational Performance*. CRC Press.
Pechenizkiy, M., Puuronen, S., & Tsymbal, A. (2005). Why data mining research does not contribute to business? In: C. Soares et al. (eds), *Proc. of Data Mining for Business Workshop DMBiz (ECML/PKDD'05)* (pp. 67-71). Portugal: Porto.
Porter, M. (1985). *Competitive Advantage*. *Creating and Sustaining Superior Performance*. New York: Free Press.
Rahm, E., & Do, H.H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull*., *23*(4), 3–13.
Rummler, G.A., & Brache, A.P. (2000). *Podnoszenie efektywności organizacji [Improving the effectiveness of the organization]*. Warszawa: PWE.
Sebesta, R.W. (2016). *Concepts of programming languages*. Pearson.
Sliż, P. (2018). *Dojrzałość procesowa współczesnych organizacji w Polsce [Process maturity of contemporary organizations in Poland]*. Sopot: Wydawnictwo Uniwersytetu Gdańskiego.
Song, C.W., Jung, H., & Chung, K. (2017). Development of a medical big-data mining process using topic modeling. *Cluster Computing*, 1–10.

 DOI 10.7172/1644-9584.82.13

Suzhen, Q. (2018, November). Data Mining and Business Process Management of Apriori Algorithm. IOP Conference Series: Materials Science and Engineering (Vol. 439, No. 3, p. 032012). IOP Publishing.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Pearson Education.

Tavares, M., Paredes, S., Rocha, T., Carvalho, P., Ramos, J., Mendes, D., & Morais, J. (2015, August). *Expert knowledge integration in the data mining process with application to cardiovascular risk assessment*. In Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE (pp. 2538–2542). IEEE.

The Google Trends tool. Retrieved from: https://trends.google.com/trends.

Wang, H., & Wang, S. (2008). A knowledge management approach to data mining process for business intelligence. *Industrial Management & Data Systems*, *108*(5), 622–634.

Wang, J. (ed.). (2005). *Encyclopedia of Data Warehousing and Mining*. Hershey, PA: Idea Group Inc.

Yang, H.H., & Moody, J. (2000). Data visualization and feature selection: New algorithms for nongaussian data. *Advances in Neural Information Processing Systems*, *12*, 687–693.