IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 29, 2018, accepted December 6, 2018, date of publication December 17, 2018,
date of current version January 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2887057

Understanding Time-Based Trends in Stakeholders' Choice of Learning Activity Type Using Predictive Models

MARTIN DRLIK AND MICHAL MUNK
Department of Computer Science, Constantine the Philosopher University in Nitra, 949 74 Nitra, Slovakia

Corresponding author: Martin Drlik (mdrlik@ukf.sk)

This work was supported in part by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and of Slovak Academy of Sciences under Contract VEGA-1/0776/18, and in part by the scientific research project of the Czech Sciences Foundation under Grant GA16-19590S.

ABSTRACT The learning analytics communities, as well as most learning analytics research, have not frequently focused on time-based trends in the same virtual learning environment over different years of deployment, or on temporal trends in the selection of different activity types over a typical day. This paper contributes to this debate and provides a novel approach to learning analytics using a multinomial logit model for modeling the probabilities of students' choice of learning activities during the hours of the day over several academic years. An abstraction called activity is introduced, which categorizes individual student's log accesses to the virtual learning environment into more semantically meaningful categories. Consequently, the activity represents a sequence of semantically meaningful Web accesses related to a particular activity or task that a student of the virtual learning environment performs. This paper includes a comprehensive explanation of the model and an evaluation of the model. This paper introduces a case study, which shows that the multinomial logit model can give useful insight into the course schedule, as it shows what the peak times are for different types of activities. This paper also discusses the possible implications of the results in the context of virtual learning environment management and content improvement at the institutional level.

INDEX TERMS Computational and artificial intelligence, learning management systems, predictive models, Web mining.

I. INTRODUCTION

The analysis of data collected from the interaction of users in the virtual learning environment (VLE) has attracted much attention as a promising approach for advancing the current understanding of the learning process as well as students' behavior. This promise motivated the emergence of the new learning analytics (LA) research field.

LA is a research field that aspires to use data analysis to support decisions made at every level of the educational institution [1]. LA deals with the gathering, measuring, and analysis of available data about stakeholders for understanding and optimizing the learning process and the whole environment, where the learning process is realized [2].

The review of the current literature indicates, that educational data could be successfully used in areas such as user behavioral pattern modelling, user knowledge and experience modelling, user profiling, personalization and adaptive personalized learning, adaptive technologies and tools, identification of learning problems, study program measurement and evaluation, as well as improvement of learning and teaching experiences [3]. Predictive methods, structure discovery, relation mining, a distillation of data for human judgment and discovery with models belong to the most frequent categories of advanced LA methods [2], [4], [5].

This paper belongs to the category of papers focused on temporal analysis of educational data, which deal with the study of common and consequential sequences of events and how these events are associated with learning outcomes, how they relate to the changes in students' behaviour over time, as well as ways in which knowledge and skills evolve over time. This area still provides both technical and theoretical challenges in appropriating suitable techniques and interpreting results in the context of learning [6].

This paper diverges from the majority of existing work in this subfield of LA primarily in the type of question it examines. An abstraction called activity is introduced [7], which categorizes VLE stakeholders' accesses (logs) into individual parts of the e-learning courses into more semantically meaningful categories. Consequently, the activity represents a sequence of semantically meaningful accesses to the parts of the e-learning courses, which relate to a particular activity or task that a VLE stakeholder executes.

A probability modelling of the students' accesses to different types of activities in the selected VLE depending on time is the main aim of the paper. In other words, the paper attempts to understand time-based trends in students' choice of activity type during the day in a VLE. These time-based trends over several years are not frequently researched. Moreover, the analysis of related research papers showed that only several LA papers focused on trends in the same VLE over different consecutive years of deployment. Simultaneously, the paper provides an example of the probability modelling of activities in the VLE during the hours of the day. As a result, both trends together can give useful insight into effective e-learning course scheduling and its lifecycle over the academic year. The teacher can use this information, for example, to determine when he/she should be most active in the VLE. On the other hand, the manager of the VLE can better estimate the periods suitable for VLE improvements or learning processes optimization in general.

The paper has the following structure. The second section summarizes the outcomes of scientific resources, which closely relate to time-based trends from the LA point of view. The necessary tasks of data wrangling and pre-processing are summarized in the third section. A multinomial logit model (MLM) is described in detail in the fourth section. The results of the MLM application are available in the fifth section. The sixth section is focused on the MLM evaluation. Finally, the utilization of the MLM at the various levels of the educational institution as well as of the e-learning course's development life cycle is discussed in the last section.

II. RELATED WORK

A large set of scholar publications were published in the last decade, which deal with many different facets of learning analytics. As a result, comprehensive resources of papers related to the LA research discipline, which provide a meta-analysis of this research field, as well as which try to identify current trends, are quite frequent [3], [8]–[11]. The analysis of these resources leads to the finding that most LA research papers are predominantly focused on educational tasks, methods, and algorithms, often on the level of the whole course.

Predictive modelling, the method used in this paper, can use several types of data, for instance, demographic/static data or stakeholders' interactions with the educational systems [12]. LA researchers tend to focus on classical approaches to prediction modelling, like classification and regression [5], [13]. According to Papamitsiou and Economides [14], the prediction of dropout and retention represent key issues for data mining in education in general.

Even though the concept of time is fuzzily defined in education, it influences many key teaching and learning aspects of the educational situation [15]. While temporal aspects of teaching and learning are extremely important, the time-factor has not received much attention in LA research. Thus, educational practice lacks adequate research on an important aspect that can improve learning in general and e-learning in particular [16]. This statement is in accordance with the findings of Knight [17], who stated that although it might be expected that a study of the temporal nature of learning will be central in learning analytics research and applications, the current situation is rather the opposite and the temporal nature of learning takes on importance only slowly. Temporality has typically been underexplored in both basic and applied educational research. Moreover, the details of how processes are expected to unfold over time are rarely well conceptualized.

This paper deals with a special kind of predictive modelling, which is focused on modelling the behaviour of the stakeholders over time. In other words, it tries to estimate the probability, with which a VLE's stakeholder will deal with a particular activity. The observed period could be different, for example, a day, a part of a week, a period of the term, a year, etc. Considering the assertions mentioned in the previous paragraph, this kind of research does not occur very frequently in LA research, although the predictive value of data about students' activity may be enhanced when considering either the order in which students engage in activities or the timing of engaging in activities [18].

Barbera *et al.* [16] assume, that the reason for limited LA research in temporal data is that most LA methods and approaches require aggregation across educational data due to a collection of conceptual, methodological, and operational challenges. Consequently, applying such a ''coding and counting'' approach often leads to the loss of temporal information related to students' interactions. As counting data are aggregated over time, the data gets ''flattened out'' in the temporal dimension and information about temporal variation is abandoned. Moreover, Chen *et al.* [19] stated, that despite the abundance of available temporal educational data, there has been a paucity of research on the temporal features of learning, with a tendency to minimize, if not totally ignore, the temporal dimension by using a ''coding and counting'' approach that aggregates over time. Subsequently, such operating in a ''snapshot'' mode shows stakeholders only a current picture of the data. It can lead to overlooking or misrepresenting patterns that change over time.

Ceddia *et al.* [20] and Ceddia and Sheard [21] developed a web-based educational system WIER. They defined different levels of log abstraction depending on time and researched the students' behavior. However, they focused mainly on analyzing a students' behaviour in a particular activity; e.g. the attempts at quizzes. They did not analyze their behaviour using the modelling of the probabilities of accesses.

The paper introduces a case study, in which a multinomial logit model (MLM) is adapted for modelling the probability of accesses of VLE's stakeholders depending on time. More precisely, the stakeholders' behaviour over several academic years is analyzed, which is covered by LA researchers only partially. While an increasing body of literature has become available regarding how a VLE stakeholders' behaviour (meaning their performance, knowledge, final grades) can be estimated, to the best of the authors' knowledge there is no available research dealing with a predictive modelling of the VLE stakeholders' behaviour in time using MLM.

The educational data used in the case study are stored in the form of logs in the VLE Moodle. The VLE Moodle belongs to the most used VLEs for several years. Therefore, many researchers focused their research on the implementation of data mining methods on the educational data recorded in this system [22], [23]. Dimopoulos *et al.* [24] evaluated available data mining tools, which can be used with the VLE Moodle.

Logs about the VLE stakeholders' activity stored in a structured form, can be considered time-oriented data. The number of papers, which take into account the time of accesses, is low. In general, they concluded that the patterns of student behaviour over time could be identified and suitably interpreted [25]–[28].

Młynarska *et al.* [29] examined a large VLE and performed a time series clustering of students achieving low and high grades, to observe the different behavioural patterns characteristic of these different groups. The clustering of activity data revealed several distinct behavioural patterns that highlight the relationship between VLE activity in relation to assignments and their final grades. Considering the aim of this paper, the studies mentioned earlier mainly analyzed and visualized the educational data and tried to bridge different didactical theories with VLE stakeholders' requirements [30].

The application of MLM extends the previous research, in which the probability of the students' accesses to the different activities and educational resources for the selected e-learning course was modeled [31]. The different behaviour of teachers and students using the MLM methodology was also analyzed in [32]. These studies confirmed statistically significant results of modelling the probability of accesses to the different parts of the e-learning course during different periods. They provided evidence as well as examples, how an application of the MLM could improve the management of a particular e-learning course and its participants [33].

This paper extends the methodology of MLM application at the higher institutional level. It describes how it is possible to estimate the probability, that the stakeholders will do a particular activity in a given time. Subsequently, it also discusses the opportunities for predictive modelling of VLE stakeholders using MLM in the management of a set of e-learning courses covered by a particular study program or managed by an organizational unit of a university.

III. DATA WRANGLING

Data wrangling consists of data cleaning, user identification, session identification, path completion steps [34]–[36].

A. DATA UNDERSTANDING

The educational data used in this case study describes stakeholders' accesses to the activities (modules, resources, interactive and collaborative activities) of e-learning courses during eight academic years. The e-learning courses were opened in the e-learning system of the university. The teachers and students used the VLE predominantly in the blended learning form. About 1000 of unique stakeholders' logins daily on average were recorded during the observed period. The stakeholders' logs from 302 e-learning courses used at the Faculty of Natural Sciences during 2010-2017 were selected. 5033 unique participants enrolled and actively participated in these e-learning courses during the mentioned period.

B. DATA GATHERING

The multi-tier software architecture of the VLE assumes that the users' data are not stored only in server log files but also in the more structured form of the relational database system. This system contains the data and metadata about the e-learning content, the structure of the courses, as well preserving the state of the VLE and its parts. It also stores data about the VLE stakeholders' activities [37].

A relational database system provides an integrated and structured form of data, which minimalizes the preprocessing phase of the data analysis [38].

Subsequently, a set of SQL queries can be prepared, which returns all required information about the stakeholders' activity in the VLE. For this case study, an SQL script was written. It selected the attributes *ID*, *userID*, *IP*, *course*, *time*, *module*, *action* from the tables *mdl_log*, *mdl_logstore_standard_log* and *mdl_log_display*. These tables created the main resource of logs of the VLE Moodle. Additionally, other tables were used where they were needed to improve human understanding of the attributes. The structure of the log tables of the VLE Moodle changed in 2015. For that reason, the records about the users' activity in the e-learning courses were mapped into the previous proprietary structure.

Usually, the stakeholders' logs should be cleaned of unnecessary items. Since the VLE stakeholders' had a unique ID, this step of data pre-processing is not necessary. The entries about users with the role other than the student or teacher were removed. Finally, 4,435,175 entries were accepted to be used in the next steps of data preparation.

C. REDUCTION OF CATEGORIES OF MODULES AND ACTIONS

Subsequently, special attention was paid to the attribute *activity*. Romero *et al.* define the abstractions as groupings of related records [3]. Page views, sessions, tasks or activities are typical examples of these abstractions. Sherd [39] defines an *activity* as an abstraction of discrete behavioural activity of the stakeholder stored in the VLE Moodle log file. It categorizes individual e-learning course parts into particular activities. As a result, the *activity*, in this case study, represents a sequence of semantically meaningful stakeholders' accesses related to a particular activity that a stakeholder of the VLE accomplished.

The usefulness of the abstraction of e-learning course parts into semantically more meaningful activities grows if the accesses to the individual parts were low [40]. The original variables *action* and *module* took many values (28 modules, 76 actions) in this case. Since many values led to lower counts of accesses, an abstraction, called *activity*, was also introduced. It reduced the number of different types of stakeholders' actions. A group of VLE and e-learning experts was asked to divide the combination of available modules and actions into the four *activities*:

- *learn* – contains all modules and actions which are connected with the learning process (view discussion, read educational resources), and require only small interactivity from the stakeholder,
- *browse* – contains mainly the module *course* and its actions, which enable browsing between the e-learning course resources and modules (course view), covers the passive presence of the stakeholder in the course predominantly,
- *manage* – includes all modules and actions related to the interactivity and managing of the learning process tasks (assign grade submission, assignment view all, grade, report, calendar, messaging),
- *develop* – covers all activities, which are related to the create, update and delete operations realized by the course participants (feedback, surveys about the educational resources, creating and editing quiz questions, glossary items, creating wiki pages, participating in the module database).

The reason for dividing the log records into these more general activities based on the values in columns *module* and *action* is that the aim of the research was to verify if there were any changes in generalized activities of the VLE stakeholders' behaviour over several academic years and what kind of the activity is prevalent at a particular time. The higher the number of categories of activities would cause a worse interpretation of the results.

D. USER AND SESSION IDENTIFICATION

The user identification is a less demanding problem in the VLE domain compared with other data mining application domains because the stakeholders can be identified by their unique ID [41]. Anonymous logins to the university VLE have not been allowed. Hence, only the activities of identified VLE stakeholders were logged.

A user session is defined as a sequence of requests made by a particular user over a certain navigation period. A user may have one or multiple sessions during this period. The session identification is a process of segmenting the logs of each stakeholder into disjoint sequences of individual sessions [42]. A more detailed discussion about the different session identification techniques can be found in [37] and [43].

User sessions can be identified in several ways. Finally, a reactive time-oriented heuristic method based on time threshold for identifying the users' sessions was selected based on the previous research [43], [44]. A 100-minutes timeout threshold was adopted for starting a new session regarding the results of the previous study [45], the preferred blended learning form, as well as settings of the university VLE. Therefore, this threshold was calculated considering a typical duration of a lesson at the university (90 minutes of lessons + 10 additional minutes to upload exercise or close another task after the lesson). As a result, the variable *session* was added into the final dataset as an output of this preprocessing step (Table 1).

E. PATH RECONSTRUCTION

The path reconstruction represents the last step of the data wrangling phase. Different path reconstruction techniques are summarized in [46] and [47]. The outcomes of the previous research indicate that the data wrangling phase can be reduced to a reconstruction of the activities of VLE users in the educational domain [43]. In other words, it has been proven that the identification of the transactions/sequences of visitors' actions in the VLE has a significant impact on the quality and quantity of identification of the useful sequence patterns. Considering this finding, if the VLE provides a sophisticated form of navigation and a firmly defined structure of the e-learning courses, the completion of the path is not an inevitable step of data preparation in the process of discovering patterns of VLE stakeholder's behaviour [45].

F. CALCULATION OF DERIVED VARIABLES

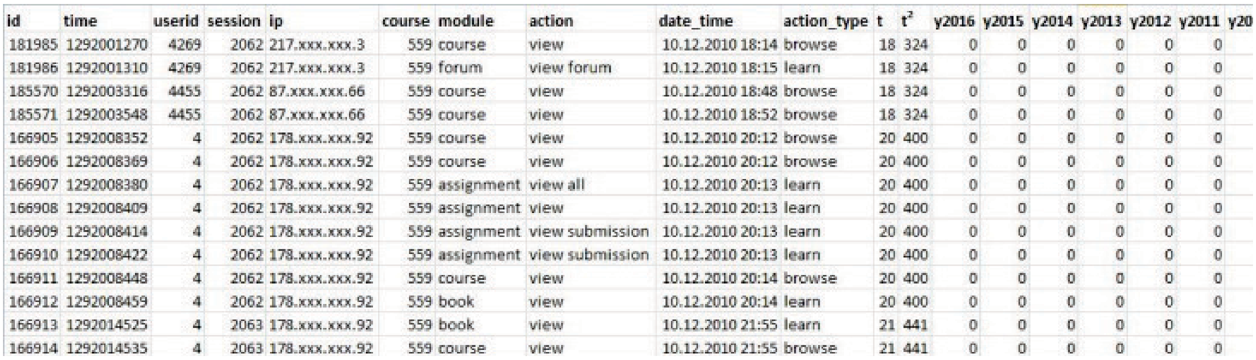
The construction of some derived variables was the last step of the data preparation stage. The original log file contained only the variable *datetime*. An independent variable was a time variable t with values from the interval 0-23 in this case study (hours in a day). It was calculated from the variable *datetime*. The type of dependence on time was identified considering the results of calculation and visualization of empirical logits. The previous research proved that logits compose a quadratic function of time [40] (Figure 9). Therefore, a new variable t^2 was defined.

It should be mentioned for completeness, that other independent variables based on time could be considered and used in MLM. For example, different week days could be distinguished [40], eventually working days and weekends [31] or different periods of the academic year [33]. In some other cases, it is desirable to create several models for different groups of users or types of accesses to the defined parts of the observed system [40]. Finally, several dummy variables *y2010*, *y2011*, *y2012*, *y2013*, *y2014*, *y2015*, *y2016* were calculated for the presented case study.

IV. MODEL DESCRIPTION

The multinomial logit model (MLM) introduced in this paper was described in the book [48] in detail. This MLM is a special case of the Generalized Linear Model. The basis for this theory was received from the books [4], [49]. In the next

TABLE 1. An example of the attributes of the final dataset.



id	time	userid	session	ip	course	module	action	date_time	action_type	t	t^2	y2016	y2015	y2014	y2013	y2012	y2011	y20
181985	1292001270	4269	2062	217.xxx.xxx.3	559	course	view	10.12.2010 18:14	browse	18	324	0	0	0	0	0	0	
181986	1292001310	4269	2062	217.xxx.xxx.3	559	forum	view forum	10.12.2010 18:15	learn	18	324	0	0	0	0	0	0	
185570	1292003316	4455	2062	87.xxx.xxx.66	559	course	view	10.12.2010 18:48	browse	18	324	0	0	0	0	0	0	
185571	1292003548	4455	2062	87.xxx.xxx.66	559	course	view	10.12.2010 18:52	browse	18	324	0	0	0	0	0	0	
166905	1292008352	4	2062	178.xxx.xxx.92	559	course	view	10.12.2010 20:12	browse	20	400	0	0	0	0	0	0	
166906	1292008369	4	2062	178.xxx.xxx.92	559	course	view	10.12.2010 20:12	browse	20	400	0	0	0	0	0	0	
166907	1292008380	4	2062	178.xxx.xxx.92	559	assignment	view all	10.12.2010 20:13	learn	20	400	0	0	0	0	0	0	
166908	1292008409	4	2062	178.xxx.xxx.92	559	assignment	view	10.12.2010 20:13	learn	20	400	0	0	0	0	0	0	
166909	1292008414	4	2062	178.xxx.xxx.92	559	assignment	view submission	10.12.2010 20:13	learn	20	400	0	0	0	0	0	0	
166910	1292008422	4	2062	178.xxx.xxx.92	559	assignment	view submission	10.12.2010 20:13	learn	20	400	0	0	0	0	0	0	
166911	1292008448	4	2062	178.xxx.xxx.92	559	course	view	10.12.2010 20:14	browse	20	400	0	0	0	0	0	0	
166912	1292008459	4	2062	178.xxx.xxx.92	559	book	view	10.12.2010 20:14	learn	20	400	0	0	0	0	0	0	
166913	1292014525	4	2063	178.xxx.xxx.92	559	book	view	10.12.2010 21:55	learn	21	441	0	0	0	0	0	0	
166914	1292014535	4	2063	178.xxx.xxx.92	559	course	view	10.12.2010 21:55	browse	21	441	0	0	0	0	0	0	

part, the used multinomial logit model will be described in detail.

The variables included in the model were described in the previous section. The investigated categorical dependent variable was a variable *activity* with categories: *learn*, *browse*, *develop* and *manage* for a model. The time t with values 0-23 represented an independent variable. Moreover, the variable square of time t^2 as well as several dummy variables mentioned earlier, which MLM required, were calculated.

Let π_{ij} be the probability that the user will choose the activity j, in hour i, while $j = 1, 2, \ldots, J$, where J is a number of *activities* and $i = 0, 1, \ldots, 23$.

Since $\sum_{j=1}^{J} \pi_{ij} = 1$ is true, there are only J–1 parameters. Let Y_{ij} be the number of accesses to the activity j with observations y_{ij} in hour i, then $\sum_{j=1}^{J} y_{ij} = n_i$ is the number of accesses in hour i.

The probability distribution of the vector $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})^T$, if the sum n_i is given, is multinomial

$$f_i\left(y_{i1}, y_{i2}, \ldots, y_{iJ}\right) = P\left[Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iJ} = y_{iJ}\right] = \frac{n_i!}{y_{i1}!y_{i2}!\ldots y_{iJ}!}\pi_{i1}^{y_{i1}}\pi_{i2}^{y_{i2}}\ldots\pi_{iJ}^{y_{iJ}}.$$

Consequently, it can be found by applying of the logarithm, that

$$ln\, f_i\left(y_{i1}, y_{i2}, \ldots, y_{iJ}\right) = \sum_{j=1}^{J} y_{ij} ln\, \pi_{ij} + ln\left(\frac{n_i!}{y_{i1}!y_{i2}!\ldots y_{iJ}!}\right).$$

Since $\sum_{j=1}^{J} \pi_{ij} = 1$, it is possible to put $\pi_{iJ} = 1 - \sum_{j=1}^{J-1} \pi_{ij}$ and get

$$\begin{aligned}
&ln\, f_i\left(y_{i1}, y_{i2}, \ldots, y_{iJ}\right)\\
&= \sum_{j=1}^{J-1} y_{ij} ln\pi_{ij} + y_{iJ}\, ln\left(1 - \sum_{j=1}^{J-1}\pi_{ij}\right) + ln\left(\frac{n_i!}{y_{i1}!y_{i2}!\ldots y_{iJ}!}\right)\\
&= \sum_{j=1}^{J-1} y_{ij} ln\pi_{ij} + \left(n_i - \sum_{j=1}^{J-1} y_{ij}\right) ln\left(1 - \sum_{j=1}^{J-1}\pi_{ij}\right) + ln\left(\frac{n_i!}{y_{i1}!y_{i2}!\ldots y_{iJ}!}\right)\\
&= \sum_{j=1}^{J-1} y_{ij} ln\pi_{ij} + n_i ln\left(1 - \sum_{j=1}^{J-1}\pi_{ij}\right) - \sum_{j=1}^{J-1} y_{ij} ln\left(1 - \sum_{j=1}^{J-1}\pi_{ij}\right) + ln\left(\frac{n_i!}{y_{i1}!y_{i2}!\ldots y_{iJ}!}\right)\\
&= \sum_{j=1}^{J-1} y_{ij}\left(ln\pi_{ij} - ln\left(1 - \sum_{j=1}^{J-1}\pi_{ij}\right)\right) + n_i ln\left(1 - \sum_{j=1}^{J-1}\pi_{ij}\right) + ln\left(\frac{n_i!}{y_{i1}!y_{i2}!\ldots y_{iJ}!}\right)\\
&= \sum_{j=1}^{J-1} y_{ij} ln\frac{\pi_{ij}}{1 - \sum_{j=1}^{J-1}\pi_{ij}} + n_i ln\left(1 - \sum_{j=1}^{J-1}\pi_{ij}\right) + ln\left(\frac{n_i!}{y_{i1}!y_{i2}!\ldots y_{iJ}!}\right).
\end{aligned}$$

From the last modification, the formula is obtained

$$lnf_i\left(y_{i1}, y_{i2}, \ldots, y_{iJ}\right) = \sum_{j=1}^{J-1} y_{ij} ln\frac{\pi_{ij}}{\pi_{iJ}} + ln\left(\pi_{iJ}\right)^{n_i} + ln\left(\frac{n_i!}{y_{i1}!y_{i2}!\ldots y_{iJ}!}\right).$$

Then, the log is taken

$$f_i(y_{i1}, y_{i2}, \ldots, y_{iJ}) = \exp\left(\sum_{j=1}^{J-1} y_{ij} ln\frac{\pi_{ij}}{\pi_{iJ}}\right)\pi_{iJ}^{n_i}\frac{n_i!}{y_{i1}!y_{i2}!\ldots y_{iJ}!} \tag{1}$$

and put

$$\eta_{ij} = ln\frac{\pi_{ij}}{\pi_{iJ}}, \quad \eta_{iJ} = 0,$$

$$\eta_i = \left(\eta_{i1}, \eta_{i2}, \ldots, \eta_{iJ-1}\right)^T, \quad y_i = \left(y_{i1}, y_{i2}, \ldots, y_{iJ}\right)^T.$$

Then

$$\pi_{ij} = \pi_{iJ} e^{\eta_{ij}}, \quad j = 1, 2, \ldots, J - 1,$$

$$1=\sum_{j=1}^{J}\pi_{ij}=\pi_{iJ}\sum_{j=1}^{J}e^{\eta_{ij}}=\pi_{iJ}\left(1+\sum_{j=1}^{J-1}e^{\eta_{ij}}\right),$$

of which the formula is obtained

$$\pi_{iJ}=\frac{1}{1+\sum_{j=1}^{J-1}e^{\eta_{ij}}}.$$

Now the probability distribution function has a general exponential form

$$f_i\left(\mathbf{y}_i,\boldsymbol{\eta}_i\right)=C\left(\boldsymbol{\eta}_i\right)\exp\left(\sum_{j=1}^{J-1}Q_j\left(\boldsymbol{\eta}_i\right)T_j\left(\mathbf{y}_i\right)\right)u\left(\mathbf{y}_i\right),$$

where

$$C\left(\boldsymbol{\eta}_i\right)=\left(1+\sum_{j=1}^{J-1}e^{\eta_{ij}}\right)^{-n_i},\quad Q_j\left(\boldsymbol{\eta}_i\right)=\eta_{ij},$$

$$T_j\left(\mathbf{y}_i\right)=y_{ij},\quad u\left(\mathbf{y}_i\right)=\frac{n_i!}{y_{i1}!y_{i2}!\ldots y_{iJ}!}.$$

Hence, a generalized linear model with link function logit can be applied to estimate the probabilities π_{ij} of selecting activity j with respect to the hour i. From (Eq.1) is obtained

$$ln\,f_i\left(\mathbf{y}_i,\boldsymbol{\eta}_i\right)=ln\frac{e^{\sum_{j=1}^{J-1}y_{ij}\eta_{ij}}}{\left(1+\sum_{j=1}^{J-1}e^{\eta_{ij}}\right)^{n_i}}\frac{n_i!}{y_{i1}!y_{i2}!\ldots y_{iJ}!}.$$

The log-likelihood function has a form

$$ln\prod_i f_i\left(\mathbf{y}_i,\boldsymbol{\eta}_i\right)=\sum_i\left(\sum_{j=1}^{J-1}y_{ij}\eta_{ij}-n_i ln\left(1+\sum_{j=1}^{J-1}e^{\eta_{ij}}\right)\right)+\sum_i ln\frac{n_i!}{y_{i1}!y_{i2}!\ldots y_{iJ}!}.$$

It is assumed that the following model is valid

$$\eta_{ij}=ln\frac{\pi_{ij}}{\pi_{iJ}}=\mathbf{x}_i^T\boldsymbol{\beta}_j,\quad j=1,2,\ldots,J-1,\quad i\in\{0,1,\ldots,23\},\tag{2}$$

where π_{iJ} is the probability of the last activity which will be chosen as a referential, $\mathbf{x}_i^T$ is a line vector, $\boldsymbol{\beta}_j=\left(\beta_{j1},\beta_{j2},\ldots,\beta_{jk}\right)^T$ is a vector of regression coefficients for $j=1,2,\ldots,J-1$.

There are $J-1$ equations which describe the contrasts between the activity j, for $j=1,2,\ldots,J-1$ and the last activity J (any other category could be chosen as a reference category). Then there is a log-likelihood function without the constants in the form

$$ln\,L\left(\mathbf{y},\mathbf{x},\boldsymbol{\beta}\right)=ln\prod_i f_i\left(\mathbf{y}_i,\mathbf{x}_i,\boldsymbol{\beta}_j\right)=\sum_i\sum_{j=1}^{J-1}y_{ij}\left(\mathbf{x}_i^T\boldsymbol{\beta}_j\right)-\sum_i n_i ln\left(1+\sum_{j=1}^{J-1}e^{\left(x_i^T\beta_j\right)}\right).\tag{3}$$

TABLE 2. Test of all effects.

	Degree of Freedom	Wald Stat.	p
Intercept	3	48368.90	0.0000
t	3	1461.90	0.0000
t2	3	752.23	0.0000
y2010	3	10648.68	0.0000
y2011	3	19754.06	0.0000
y2012	3	31185.01	0.0000
y2013	3	46135.42	0.0000
y2014	3	81205.33	0.0000
y2015	3	31632.16	0.0000
y2016	3	809.06	0.0000

Maximum likelihood estimation of the parameters of the model (Eq. 2) proceeds by maximization of the log of the multinomial likelihood function (without the constants) (Eq. 3).

The estimation of the parameters can be done by using an iteratively re-weighted least squares method as the Newton-Raphson technique or Fisher scoring.

The starting values for estimations β_{j0} are computed from empirical logits

$$\eta_{ij0}=ln\frac{p_{ij}}{p_{iJ}}=\mathbf{x}_i^T\boldsymbol{\beta}_j,\quad p_{ij}=\frac{y_{ij}}{n_i},\quad j=1,2,\ldots,J-1,\quad i\in\{0,1,\ldots,23\}$$

by linear regression.

The maximum likelihood estimation $\hat{\beta}_j$ has approximately in large samples a multivariate normal distribution with mean equals to the true parameter value and with variance-covariance matrix given by the inverse of the information matrix.

Information matrix is a mean value of the matrix of the second partial derivatives of the log-likelihood function concerning its parameters. Standard errors of parameters' estimations are the square roots of diagonal elements of the variance-covariance matrix divided by $\sqrt{n}$.

The hypothesis H0 :$\boldsymbol{\beta}_\mathrm{j}=\mathbf{0}$ can be tested by the Wald test.

V. RESULTS

The parameters $\alpha_j,\boldsymbol{\beta}_j$ of the model were estimated by maximizing of the logarithm of the multinomial likelihood function in the next step. The *STATISTICA Generalized Linear/Nonlinear Models* was used for parameter estimation of individual values. The significance of parameters was tested by the Wald test (Table 2).

It could be confirmed that the parameters of the model were statistically significant considering the results of the test of all effects (Table 2). The academic years implemented as dummy variables (*y2010* - *y2016*) in the model, represented statistically significant characteristics of the created logit model.

TABLE 3. Estimation of model parameters.

			Stand.		Lower	Upper	
	Level of Response	Estimate	Error	Wald Stat.	CL 95%	CL 95%	p
Intercept 1	*browse*	0.3174	0.0116	746.02	0.2947	0.3402	0.0000
t	*browse*	0.0207	0.0016	160.67	0.0175	0.0239	0.0000
t2	*browse*	-0.0012	0.0001	406.00	-0.0013	-0.0010	0.0000
y2010	*browse*	0.7622	0.0105	5274.10	0.7417	0.7828	0.0000
y2011	*browse*	0.7961	0.0078	10323.96	0.7807	0.8114	0.0000
y2012	*browse*	0.8622	0.0084	10486.80	0.8457	0.8787	0.0000
y2013	*browse*	0.9359	0.0079	14131.25	0.9205	0.9514	0.0000
y2014	*browse*	1.2548	0.0082	23166.61	1.2387	1.2710	0.0000
y2015	*browse*	0.3719	0.0062	3573.36	0.3597	0.3841	0.0000
y2016	*browse*	-0.0335	0.0053	40.11	-0.0439	-0.0231	0.0000
Intercept 2	*learn*	1.3322	0.0103	16634.10	1.3119	1.3524	0.0000
t	*learn*	-0.0185	0.0015	162.07	-0.0214	-0.0157	0.0000
t2	*learn*	-0.0002	0.0001	17.69	-0.0003	-0.0001	0.0000
y2010	*learn*	0.9451	0.0097	9567.60	0.9261	0.9640	0.0000
y2011	*learn*	0.9418	0.0072	17112.33	0.9277	0.9559	0.0000
y2012	*learn*	1.2872	0.0077	28026.15	1.2721	1.3023	0.0000
y2013	*learn*	1.4303	0.0072	39529.94	1.4162	1.4444	0.0000
y2014	*learn*	1.9800	0.0076	67645.15	1.9651	1.9950	0.0000
y2015	*learn*	0.8622	0.0055	24409.63	0.8514	0.8731	0.0000
y2016	*learn*	-0.1188	0.0047	628.72	-0.1281	-0.1095	0.0000
Intercept 3	*develop*	-1.9388	0.0219	7870.64	-1.9817	-1.8960	0.0000
t	*develop*	-0.0403	0.0029	195.84	-0.0460	-0.0347	0.0000
t2	*develop*	0.0004	0.0001	14.95	0.0002	0.0006	0.0001
y2010	*develop*	1.3775	0.0200	4755.94	1.3384	1.4167	0.0000
y2011	*develop*	1.5318	0.0157	9501.29	1.5010	1.5626	0.0000
y2012	*develop*	1.4883	0.0167	7970.55	1.4556	1.5209	0.0000
y2013	*develop*	1.8711	0.0151	15335.64	1.8415	1.9007	0.0000
y2014	*develop*	2.0973	0.0153	18887.85	2.0674	2.1272	0.0000
y2015	*develop*	1.1903	0.0140	7279.39	1.1630	1.2177	0.0000
y2016	*develop*	-0.1017	0.0143	50.81	-0.1296	-0.0737	0.0000

At the same time, hours of the day represented by variables t and t^2 were statistically significant.

The STATISTICA *Generalized Linear/Nonlinear Models* was also used for estimating the model's parameters. Again, the Wald test was used to test the significance of the parameters. Significant parameters are highlighted in Table 3. The significant dependence of logits of the activities *browse, learn* and *develop* on hours of the day and its square can be seen there. The activity *manage* was taken as a reference in this case. The academic years significantly influenced the values of these logits. For instance, such impact of academic years 2010-2016 can be seen in the case of the activity *browse*.

These estimated parameters were also used for calculation of the logits estimations. Subsequently, the probabilities of selection of individual activities in a given hour of the day were also calculated. These results were in accordance with the calculated probabilities.

The estimation of the probabilities represents the outputs of the logit model. However, the knowledge of the model's parameters has the same importance. Their absolute values inform about which predictor mostly influences the observed variable. A higher absolute value of the parameter means higher dependence. A positive value means proportional dependence and vice versa.

The estimation of logits η_{ij} for all values of independent variables is

$$\hat{\eta}_{ij} = a_j + \mathbf{x}_i^T \mathbf{b}_j, \quad j = 1, 2, \ldots, J-1.$$

A multinomial logit model was used for modelling the distribution of a categorical variable. The observed variable *activity* represented the categorical variable of the stakeholders' behavior analysis. This variable has four categories (*browse*, *learn* and *develop, manage*).

Finally, dummy variables of academic years, as well as time variables, represented by an hour of the day and its square, were used as predictive variables

$$\begin{aligned}\eta_{ij} &= \alpha_j + \beta_{1j} t_i + \beta_{2j} t_i^2 + \gamma_{1j} y2016_i + \gamma_{2j} y2015_i \\ &\quad + \gamma_{3j} y2014_i + \gamma_{4j} y2013_i + \gamma_{5j} y2012_i + \gamma_{6j} y2011_i \\ &\quad + \gamma_{7j} y2010_i.\end{aligned}$$

This step of the modelling phase contains two tasks:

1. Probability estimation of accesses π_{iJ} in time i for reference activity J

$$\hat{\pi}_{iJ} = \frac{1}{1 + \sum_{j=1}^{J-1} e^{\hat{\eta}_{ij}}}.$$

2. Probability estimation of accesses π_{ij} in time i for activity j,

$$\hat{\pi}_{ij} = e^{\hat{\eta}_{ij}} \hat{\pi}_{iJ}, \quad j = 1, 2, \ldots, J-1.$$

The most interesting visualizations are summarized in the following subsections. The probability of individual activities j at time $i, j = 1, 2, \ldots, J, i \in \{0, 1, \ldots, 23\}$ could be also visualized. For that reason, a set of charts (Figures 1 - 9) was created, which visualized the probability of selection of observed activities by the stakeholders depending on time in the observed academic years. It can be seen that the probability of accesses to the particular *activity* depends on time.

A. VISUALIZATION OF STAKEHOLDERS' BEHAVIOUR IN DIFFERENT ACADEMIC YEARS

The probability of stakeholders' accesses to the activities *browse*, *develop* and *manage* in the academic year 2010 are shown in Figure 1. The probability of the activity *learn* was much higher in comparison with others, with the maximum late at night (0.68) and with a minimum 0.57. Therefore, it is omitted from the figure. The probability of the activity *browse* was in the interval 0.20 – 0.27. Other activities were very low, less than 0.10.

The curve of the probability of the activity *develop* reflected the fact, that the structure and educational content of the e-learning course changed only slightly because course creators, as well as teachers, restored their e-learning courses from previous instances of VLEs, which had been used at the university before the launching the centralized e-learning system.

The activity *browse* had its maximum in the afternoon and evening. It indicates that the stakeholders used to browse last changes, posts or other information available in the e-learning course in this period.

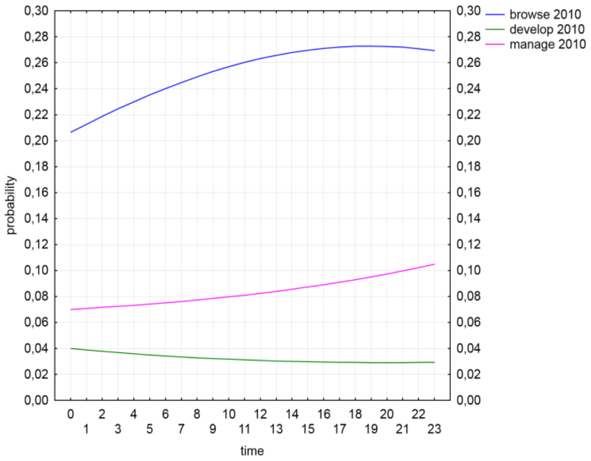


FIGURE 1. **Probability visualization of accesses of the stakeholders to the observed types of activities in 2010.**

Considering other academic years, it is worth mentioning the academic year 2014 (Figure 2a), in which the probability of activity *learn* increased (from 0.57 – 0.68 to 0.73 – 0.79), and the probability of activity *browse* slightly decreased (0.14 – 0.19) (Figure 2b).

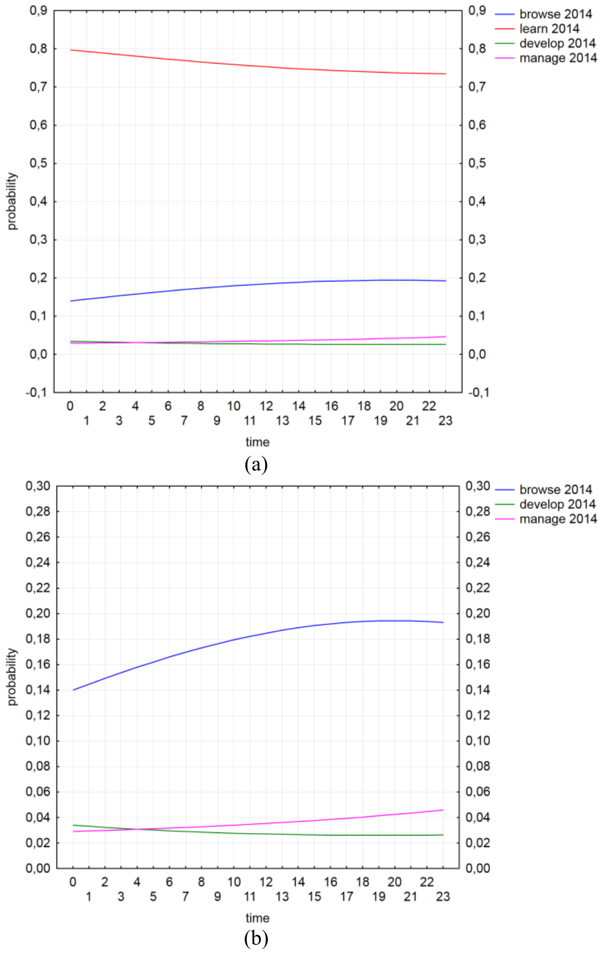
Considering the small values of activities *manage* and *develop*, it can be assumed, that the overall activity of the stakeholders focused solely on passive visiting of the e-learning courses and available study resources. Figure 3b provides a zoomed view on the activities with small values of calculated probabilities.

During the next years, the probabilities of the activity *manage* and *browse* have slowly increased with the maximum ($\pi = 0.22$) around 11 p.m. It was at the expense of the probability of the activity *learn*. This situation was probably caused by the running of several educational projects at the university focused on the effective utilization of e-learning courses in blended learning form.

Figure 3 visualizes the probability of stakeholder's accesses to the different types of *activities* in the last academic year of the observed period as an example.

A decrease in the probability of the activity *learn*, and an increase of probabilities of the activities *manage* and *browse* had several reasons. The overall consolidation of e-learning educational study resources (Figure 4a) and mainly the portfolio of interactive activities provided in the e-courses represent the most probable reason (Figure 4b). Whereas stakeholders' activity in the meaning of the number of logs continually was increasing, the portfolio of module and resources types remained almost the same.

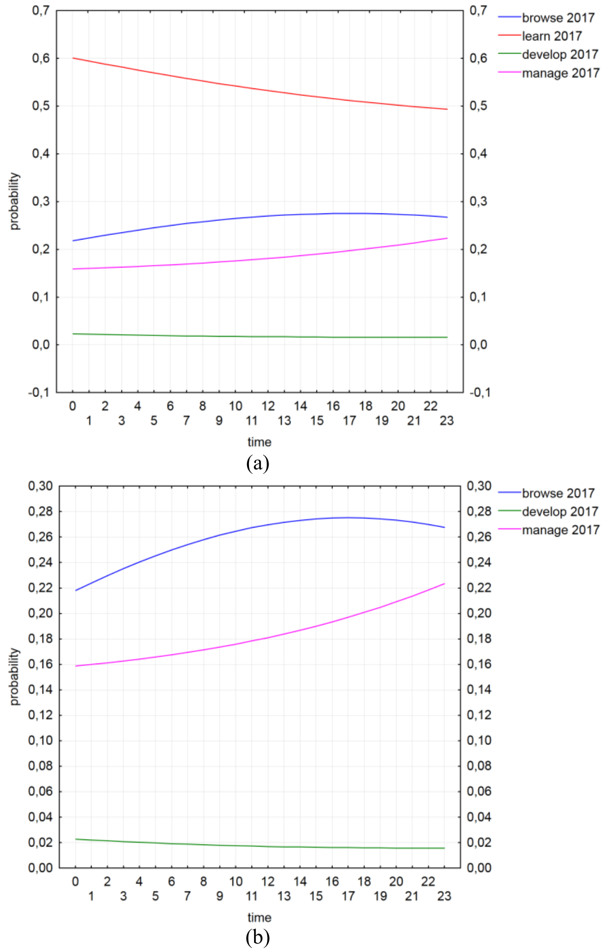
The changes in stakeholders' behaviour could be considered the second reason. The improvement of teachers' training, who participated in the realized e-learning projects, meant that teachers began using the e-learning courses not only as a repository of the educational resources but also



(a)

(b)

FIGURE 2. a) Probability visualization of accesses of stakeholders to the observed activities in 2014. b) Zoomed view on activities with a smaller probability.



(a)

(b)

FIGURE 3. a) Probability visualization of accesses of stakeholders to different types of activities in 2017. b) Zoomed view on activities with a smaller probability.

as a platform for more active interaction with the students (Figure 4b). At the same time, it can be noticed that the teachers graded the students, wrote posts and did other managerial tasks predominantly after their working hours even though the e-learning courses have been used mainly in blended form.

B. VISUALIZATION OF CHANGES IN INDIVIDUAL STAKEHOLDERS' ACTIVITIES DURING THE OBSERVED PERIOD OF ACADEMIC YEARS

Visualization of particular stakeholders' activity during the observed period of academic years uncovers another perspective on the changes in stakeholders' activity. Furthermore, this view confirms the previous conclusions.

Figure 5 emphasizes the changes in the preferred activities of the stakeholders between the academic years 2010 - 2017. An increase in the probability of the activity *manage* (Figure 5a) as well as a related decrease in the probability of the activity *learn* can be seen there (Figure 5b).

The course of the curves is in line with the subjective opinions of the teachers, who used to claim that students' activity used to increase during the night hours. The increasing in the probability of the activity *manage* can indicate, that the teachers partially improved e-learning courses towards greater interactivity in the last academic years of the observed period and students have been forced to be more active.

The visualization of the probability of the activity *learn* during the academic years 2010-2017 (Figure 5b) indicates that if the stakeholders visited the VLE in the evening or the night, they probably read or downloaded available materials and actively participated in other course activities.

The probability visualization of the activity *browse* is the highest in the afternoon (Figure 6a) during all observed years. It can be assumed that the stakeholders visited the e-learning courses more often with the aim only to check new contributions from other stakeholders, news or messages.

The probability visualization of the activity *develop* (Figure 6b) should also be explained for completeness. Its probability seemed almost unchanged during the observed period and very low compared with other activities depicted in the previous figures. The reason is that particular actions grouped in this activity were not frequent enough.

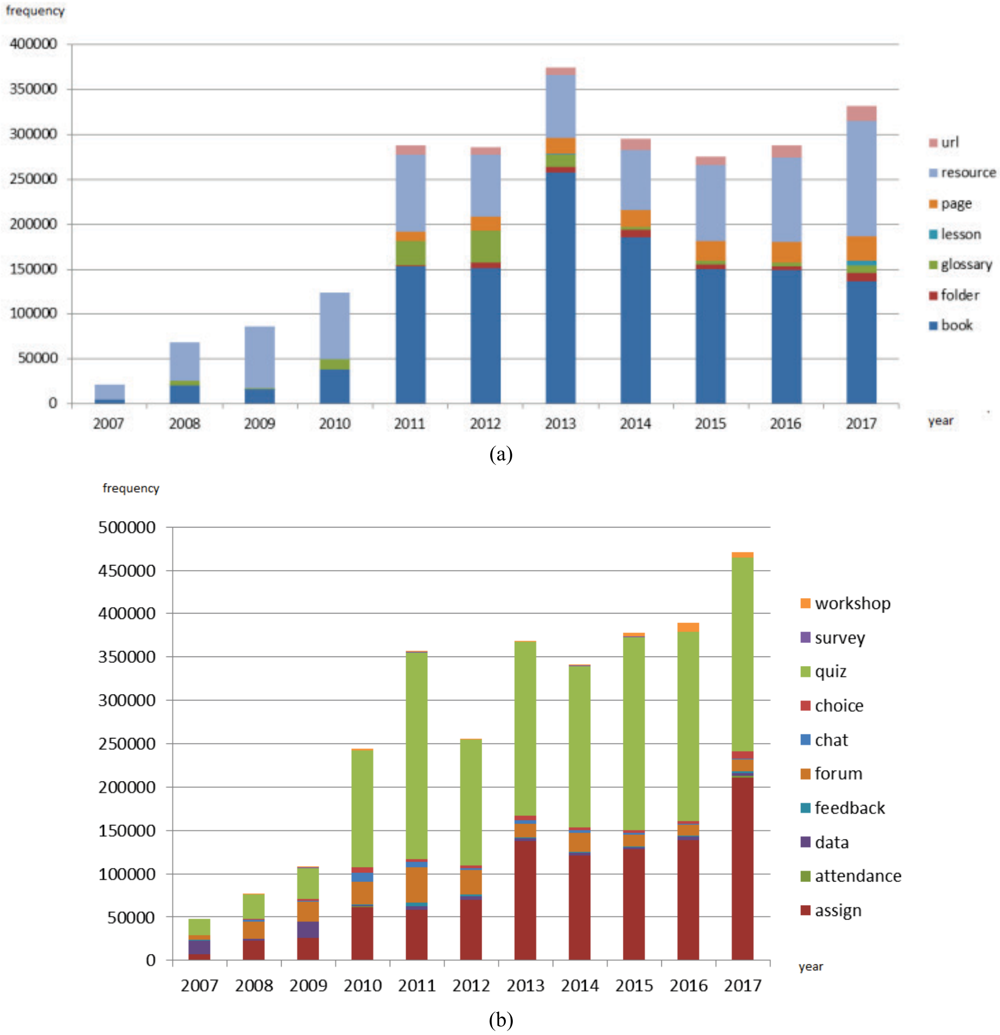


FIGURE 4. Comparison of real usage of different resources and activity modules in academic years 2007-2017 and their relative frequencies. a) The total count of logs in different academic years for different learning resources. b) The total count of logs in different academic years for different interactive activities.

The students of the university VLE are not used to participate in educational content development actively. The frequency of usage of collaborative activities like wikis, workshops, blogs, has been seldom, as can also be seen in figure 5. The number of logs related to these modules was very low comparing other types of modules and actions.

VI. MODEL EVALUATION

Only the results of the model evaluation of the year 2017 will be shown in this section as an example of the realized evaluation of the model. The reason is that similar results were also obtained for other academic years (2010-2017). Several steps should be realized during the model evaluation phase:

1. empirical counts determination,
2. theoretical counts estimation,
3. visualization of differences in the empirical and theoretical counts of accesses (Figure 7),
4. extremes identification,
5. calculation of relative empirical counts of accesses,
6. comparison of the distribution of the relative empirical counts of accesses with the estimated probabilities of selecting the activity j in hour i (Table 4),
7. calculation of empirical logits,
8. visualization of empirical and theoretical logits for individual *activity* except the referential (Figure 8) [33].

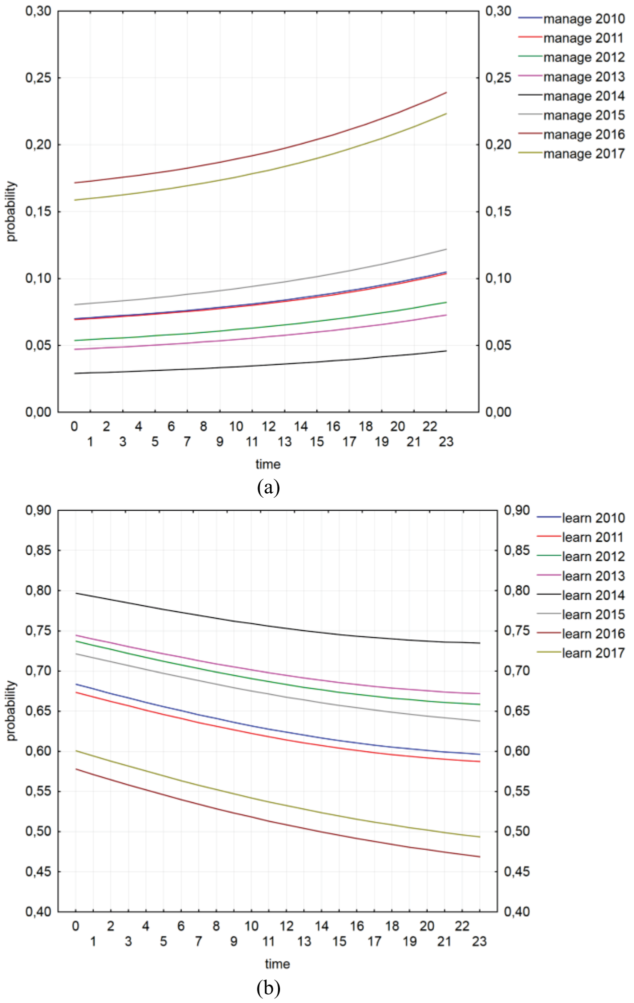
Several important aspects of this case study should be emphasized. Empirical counts of accesses y_{ij} were specified in the first step. Theoretical counts estimation were determined

$$\hat{y}_{ij} = \hat{\pi}_{ij} \sum\nolimits_{j} y_{ij}.$$

Subsequently, differences in the empirical and theoretical counts of accesses were visualized

$$d_{ij} = y_{ij} - \hat{y}_{ij}.$$

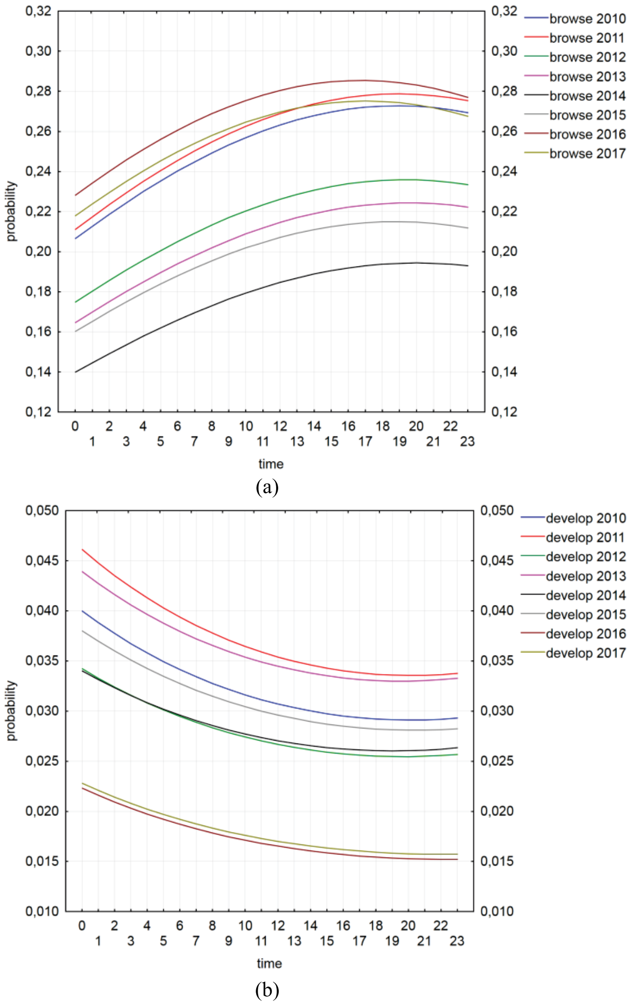
In the next step, a visualization of the differences between empirical and theoretical counts of accesses to the given type



(a)

(b)

FIGURE 5. **a) Probability visualization of activity "manage" during the academic years 2010-2017. b) Probability visualization of activity "learn" during the academic years 2010-2017.**



(a)

(b)

FIGURE 6. **a) Probability visualization of activity "browse" during the academic years 2010-2017. Probability visualization of activity "develop" during the academic years 2010-2017.**

of activities in the e-learning courses (Figure 7) was applied to identify the extremes. It means that hours, in which the forecast was overestimated ($d_{ij} < \bar{d}_j - 2s$) or underestimated ($d_{ij} > \bar{d}_j + 2s$), were identified [31], [32].

In the presented case, the suitability of the model was confirmed by means of differences, which was approximately equal to zero. Extremes were identified using the application of the 2sigma rule. The biggest difference between theoretical and empirical counts occurred at 8 a.m. in the case of all activities. Finally, four extreme cases were identified. The prediction for the activities *browse*, *develop* as well as *manage* was overestimated at this hour. Contrary, the prediction for the activity *learn* was underestimated. Another two extremes were identified in the case of the activity *develop*. The prediction was underestimated in both cases (10 a.m. and 12 a.m.).

Subsequently, the relative empirical counts of accesses were calculated

$$\mathrm{p_{ij}} = \frac{\mathrm{y_{ij}}}{\sum_{\mathrm{j}} \mathrm{y_{ij}}}.$$

Finally, the distribution of the relative empirical counts of accesses with the estimated probabilities of selecting a particular activities j in hour i were compared

$$r_{ij} = p_{ij} - \pi_{ij}, \quad H0 : F(-r) = 1 - F(r).$$

Considering the fact that the distribution of pairs is symmetrical around the zero value, the Wilcoxon matched pairs test was used for testing the zero hypothesis (H0). A statistically significant difference between empirical and theoretical probabilities was found only in the case of the activity *develop* (Table 4). The zero hypothesis for other activities was not rejected. The distribution of pairs was symmetrical around the zero value.

In the next step, empirical logits were calculated

$$h_{ij} = \ln\left(\frac{p_{ij}}{p_{iJ}}\right), \quad j = 1, 2, \ldots, J-1, \; i \in \{0, 1, \ldots, 23\}.$$

Subsequently, the empirical and theoretical logits of examined activities were visualized whether the theoretical logits fit the empirical logits [31], [32]. Figure 9 visualizes the

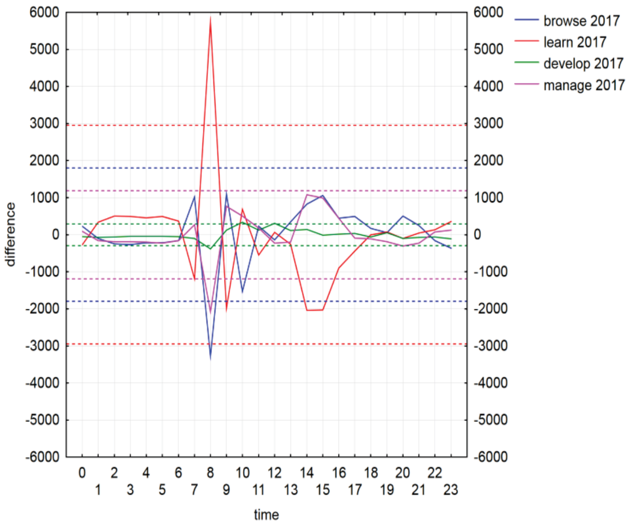


FIGURE 7. Visualization of differences in the empirical and theoretical counts of accesses for particular activities for the academic year 2017.

TABLE 4. Wilcoxon matched pairs test.

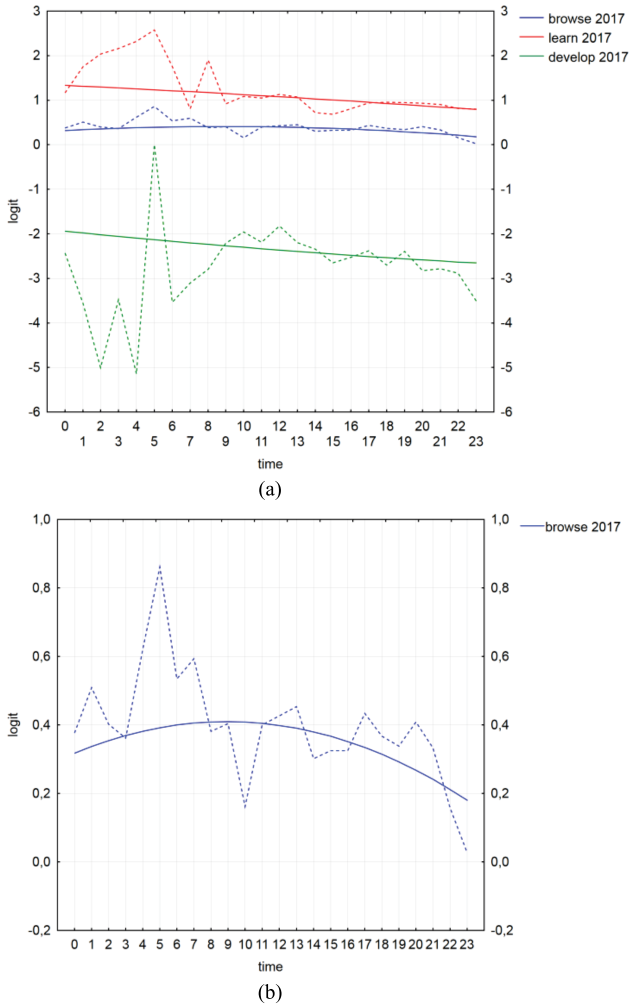
	Valid N	T	Z	p-value
theoretical_browse 2017 & empirical_browse 2017	24	126	0.685714	0.492894
theoretical _learn 2017 & empirical _learn 2017	24	118	0.914286	0.360567
theoretical _develop 2017 & empirical _ develop 2017	24	78	2.057143	0.039673
theoretical _manage 2017 & empirical _manage 2017	24	111	1.114286	0.265157

empirical and theoretical logits for the activities *browse*, *learn* and *develop*. It could be confirmed considering this visualization, that the estimations of the theoretical logits fitted to the empirical logits except for logits for the activity develop. In this last case, the empirical logits fitted to the theoretical logits only partially.

An evaluation of the proposed multinomial logit model can also be done using other evaluation techniques. The likelihood-ratio test (LR test, Deviance) was used followed by the Pearson statistics chi-square.

The statistics G^2 (Deviance) can also be applied in case the expected counts are large enough (it means that none is below 1 and no more than 20 % of the expected counts are below 5) for comparing the current model to the saturated model that estimates the probabilities independently for $i = 0, 1, \ldots, 23$.

$$\begin{aligned} G^2 &= LR(\hat{\pi}) = 2\left(L(p) - L(\hat{\pi})\right) \\ &= 2\sum_{i=0}^{23}\sum_{j=1}^{J} y_{ij}\left(\ln p_{ij} - \ln \hat{\pi}_{ij}\right) \\ &= 2\sum_{i=0}^{23}\sum_{j=1}^{J} y_{ij} \ln \frac{p_{ij}}{\hat{\pi}_{ij}} \\ &= 2\sum_{i=0}^{23}\sum_{j=1}^{J} y_{ij} \ln \frac{y_{ij}}{n_i \hat{\pi}_{ij}}. \end{aligned}$$



(a)

(b)

FIGURE 8. a) Visualization of the empirical and theoretical logits of observed activities. b) Zoomed view, which confirmed that logits compose a quadratic function of time.

TABLE 5. Statistics for the goodness of fit.

	Df	Stat.	Stat/df
Deviance (LR test)	1289344	8158320	0.6327
Pearson statistics chi-square	1289344	12896817	1.0003

The next formula is obtained from the last modification

$$G^2 = 2\sum_{i=0}^{23}\sum_{j=1}^{J} y_{ij} \ln \frac{y_{ij}}{\hat{y}_{ij}}.$$

Estimations $\hat{y}_{ij}$ to y_{ij} through the LR test could be compared. The model can be considered useful because the value of the LR test was low in this case (Table 5).

The saturated model has $24(J - 1)$ free parameters. The current model has $k(J - 1)$, and then the degrees of freedom *df* are equal to $(24 - k)(J - 1)$. The statistics G^2 has approximately χ^2 (df) distribution. The Pearson statistics can also be

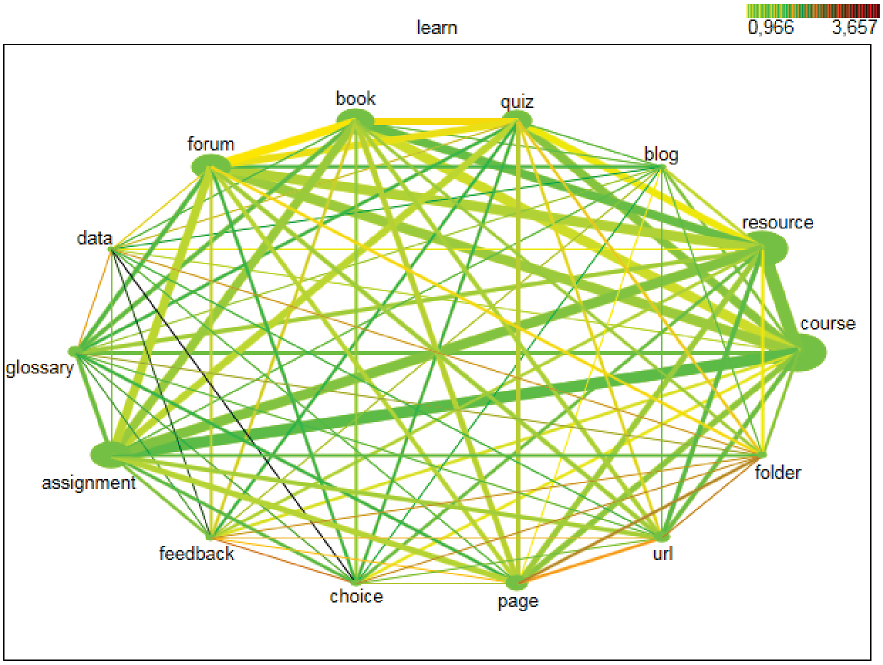


FIGURE 9. Web graph of the association rules related to the modules used in the e-learning courses with minimal support > 5%.

applied to compare the estimations $\hat{y}_{ij}$ to y_{ij}

$$\chi^2 = \sum_{i=0}^{23} \sum_{j=1}^{J} r_{ij}^2,$$

where $r_{ij} = \frac{y_{ij} - \hat{y}_{ij}}{\sqrt{\hat{y}_{ij}}}$ is the Pearson residual having $\chi^2(df)$ distribution. The model can be considered appropriate, if Pearson statistics equals approximately to one.

In the given area of LA, the use of the LR test is often interrupted. The examined variable usually has a considerable number of levels presenting the parts of the observed system (e-learning parts, pages, contents categories, activities, etc.). It has an impact on interruption of the usage term of the LR test. Consequently, the expected counts are not large enough. Therefore, alternative techniques for the model evaluation were used such as a visualization of differences between the empirical and theoretical counts, extremes identification, comparison of the distribution of the empirical relative counts of accesses with the estimated probabilities of selecting activity j in hour i as well as a visualization of empirical and theoretical logits for individual activities except referential [32].

The results of the LR test and Pearson statistics chi-square, as well as other alternative model evaluation techniques, confirmed that the model could be deemed suitable in general.

On the other hand, alternative techniques uncovered worse results only for the activity *develop* on the level of frequency, probability as well as logits. The markedly lower frequency of the accesses to the activity *develop* in the last academic year is probably caused by stakeholders' inactivity in this kind of activities, as well as by less accurate estimation of the probabilities of the activity *develop* in the academic year 2017.

VII. RESULTS VERIFICATION USING BEHAVIOURAL PATTERNS ANALYSIS OF ACTIVITIES

The visualization of stakeholders' behaviour in the individual academic years, as well as visualization of the possible changes in their activities during the observed period, uncover that the changes occurred mostly in the activities *learn* and *manage*. Whereas these changes were explained in the previous sections, the low probability of the activity *develop* remains partially unanswered.

The selection of the modules and corresponding actions to the activities is considered the main reason, why the probability of the activity *develop* has changed at least. This statement can be verified using the results of an association rules analysis shortly described in this section.

The modules can be found in different activities based on their accompanied actions. However, their probabilities of occurrence will be different in different activities (*manage*, *develop*, *browse*, *learn*). An association rules analysis can identify the most frequent modules and their actions in the meaning of their occurrence in the identified sessions.

Consequently, it can be expected that the results of this analysis could uncover the most frequently visited modules and their actions, which belong to a particular *activity*.

Association rules analysis is one of the most well studied data mining tasks in LA. It discovers relationships between attributes, producing if-then statements concerning attribute-values. It has been applied to help teachers obtain detailed feedback how students learn on the web, to evaluate students based on their navigation patterns, to classify students into groups, and to restructure e-learning course contents [3].

The ''interest'' of an association rule is evaluated by the *support*, *confidence* and *lift* metrics. The support for an itemset is given by a proportion of records in the transactions dataset that have the itemset. That means that for an itemset (A) the support can be calculated

$$support\ (A) = \frac{frequency\ of\,(A)}{number\ \ of\ \ transactions\ in\ the\ dataset} * 100.$$

The lift of rules can be similarly calculated. Based on support and confidence a lift for a rule can be defined and computed

$$lift\ (if\ \ A\ then\ C) = \frac{confidence\ (if\ \ A\ then\ C)}{support(C)},$$

where

$$confidence\ (if\ \ A\ then\ C) = \frac{support(if\ \ A\ then\ C)}{support(A)} * 100.$$

The *support* of the association rule is the same as the support of the itemset. It represents the probability of the given module occurrence in the identified stakeholders' sessions. The *confidence* represents the probability of the given combination of the modules occurrence in the identified stakeholders' sessions. Finally, *lift* is defined as the correlation. In other words, it means, how many times the combination of the modules occurs more frequently than in case if the modules would be statistically independent.

A transaction represents a set of visited types of modules by the stakeholder during a session [31], [40]. Given a set of transactions, and user-specified thresholds minimum support (represented by the variable *minsup*) and minimum confidence (represented by *minconf*), the problem of mining association rules is to generate all association rules that have the value of metrics *support* and *confidence* greater than *minsup* and *minconf*. In this case, only modules with *minsup* greater than 5% were taken into account. The results were processed by association rule analysis using *STATISTICA Sequence, Association, & Link Analysis*, which is an implementation of the algorithm using a-priori algorithm together with a tree-structured procedure that requires only one pass through the data [45].

Figure 9 depicts frequently found itemsets where the size of each node represents the *support* of particular type of modules, which were used in e-learning courses of the university VLE - 1-itemset (set of only one item). The thickness of the line between two types of modules represents the level of *support* for the 2-itemset or a combination of two types of modules. The brightness of the line represents the *lift* of a pair of module types [50]. The final visualization of the given metrics (*support*, *lift*) of association rule analysis confirms the findings of the MLM application.

The visualization confirms the previous assumption, that the modules in the activity *develop* were not frequently used. 1-itemsets of modules wiki, the workshop did not fulfil the condition of minimal *support*. Other metrics of the modules in this type of activity are also less important compared to modules in other types of activities.

At the same time, the results of the association rules analysis are in line with the findings obtained by the MLM application. The highest values of *support* in the activity *learn* belong to the modules Resource, Book, Page, URL. On the other hand, Assignments, Forum, Quiz are characterized by the highest values of *support* in the activity *manage*. The metrics *confidence* as well as *lift* of the 2-itemsets created from these modules confirm the findings that these modules are often used together in e-learning courses.

VIII. DISCUSSION AND CONCLUSIONS

As was mentioned earlier, the paper is focused on the temporal analysis of educational data. This area still provides both technical and theoretical challenges in finding suitable techniques and interpreting results in the context of education. The analysis of related research papers showed that only several LA papers focused on time-based trends in the same VLE over different consecutive years of deployment. However, methods based on the theory of time series were mainly used in these papers.

This paper tries to contribute to the research in this subtopic of LA by modelling the probabilities of the stakeholders' accesses to the particular activities of the e-learning courses in the VLE at different time intervals using MLM. The paper attempts to understand time-based trends in students' choice of activity type in a VLE. These time-based trends over several years are not frequently researched.

The efficacy of the proposed MLM can be assessed using comparative analysis with several other state-of-the-art approaches to the analysis of time-based trends in the stakeholders' behavior.

Data stored in the VLE can be considered as time-based data, which represents the accesses to the individual activities available in the system. Nevertheless, predictive tasks, which would be considered as time variables, is missing not only in the LA research field, but also in web usage mining in general. Sequence rules (patterns) and Markov chains are the only exceptions [3].

However, the time variable is used only in a limited form in the case of sequence rules. It can help to determine the order of the visited web parts (activities) in the identified sessions [15].

Similarly, the time variable used in Markov chains can again determine only the order of the observed events.

Considering this argument and the limitations of similar approaches to the analysis of time-based trends, the MLM was applied in the described research. As a result, the stakeholders' behaviour in the VLE depending on time was analyzed and a new approach to how to solve a given predictive task was proposed.

The proposed approach has several implications in practice. The estimation of probabilities, with which stakeholders access the individual activities of the e-learning course at different times, can be utilized in different phases of the e-learning course management, and VLE administration.

The probability of selecting an individual activity in a particular period by the stakeholder is given by the rigid structure of the e-learning course. Therefore, the most valuable asset of the MLM application does not lie in the process of developing new courses, but mainly in the process of restructuring the e-learning courses between individual cycles. The creator of an e-learning course can modify a portfolio of activities in the course according to the students' behaviour considering the probabilities estimated by the MLM. If any activity seems to be preferred by the students only with a little probability, the creator should replace it by an alternative activity with comparable outcomes.

The MLM application can improve a teacher's approach to the students considering their behaviour. For example, the teacher can modify the course schedule according to the students' learning preferences. Eventually, the teacher can research the impact of changes in the structure of the interactive activities on the students' behavior over several cycles of the same e-learning course.

The application of the MLM can also be useful for the university management. As a result, its application at different levels (an e-learning course, several cycles of the e-learning course, the courses of a particular study program) can improve stakeholders' feedback. For example, the students' answers in a survey focused on their study preferences can be compared with the found probabilities of accesses to the individual kinds of activities.

In addition, the managers can identify changes in students' learning preferences and behaviour over time. The coordinators of the online study programs can research the changes in the behaviour of teachers in e-learning courses over several academic years. Moreover, the application of the MLM can improve the visibility of teachers in the e-learning course, as well as contribute to the creation of institutional recommendations, which reflect the students' behaviour.

The proposed approach has several limitations. The distribution of the data based on the attributes module and action to the abstract activities, which was recommended by the experts, did not fully correspond to the real distribution of these activities at the university. As a result, the probability of accesses to the modules belonging to this activity was very small. It also caused some problems with the identification of possible changes in the stakeholders' behaviour over the observed academic years. Therefore, the activities could be defined using other principles or results of other exploratory analytical methods. However, this fact has not affected the usefulness of the MLM itself.

Future work should be focused on further improvement of the temporal educational data analysis. A common application of MLM and other data mining methods could lead to a better understanding of the behaviour of the VLE stakeholders over longer periods, identifying the impact of seasonality and time in general on the stakeholders' behaviour and learning outcomes. It could also be interesting to research how the VLE stakeholders' behaviour changes during the study and what is the impact of the learning design on their behavior.

REFERENCES

[1] L. Johnson, S. A. Becker, M. Cummins, V. Estrada, A. Freeman, and H. Ludgate. (2013). *NMC Horizon Report: 2013 Higher Education Edition*. [Online]. Available: http://net.educause.edu/ir/library/pdf/HR2013.pdf

[2] R. Baker and G. Siemens, "Educational data mining and learning analytics," in *The Cambridge Handbook of the Learning Sciences*, R. K. Sawyer, Ed. Cambridge, U.K.: Cambridge Univ. Press, 2014. [Online]. Available: https://www.cambridge.org/core/books/cambridge-handbook-of-the-learning-sciences/2E4224681267E61DBCE9B27630ED17BA

[3] C. Romero, S. Ventura, M. Pechenizkiy, and R. S. J. D. Baker, *Handbook of Educational Data Mining*. London, U.K.: Chapman & Hall, 2010.

[4] G. Rodríguez, *Generalized Linear Models*. Princeton, NJ, USA: Princeton Univ., 2011.

[5] A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," *IEEE Access*, vol. 5, pp. 15991–16005, 2017.

[6] S. Knight, A. F. Wise, B. Chen, and B. H. Cheng, "It's about time: 4th international workshop on temporal analyses of learning data," presented at the Proc. 5th Int. Conf. Learn. Anal. Knowl., Poughkeepsie, NY, USA, 2015.

[7] J. I. Sheard, "Analysis of log data from a Web-based learning environment," in *Handbook of Educational Data Mining*, C. Romero, S. Ventura, M. Pechnizkiy, and R. S. J. D. Baker, Eds. Boca Raton, FL, USA: CRC Press, 2011, pp. 311–322.

[8] C. Lang, G. Siemens, A. Wise, and D. Gašević, *The Handbook of Learning Analytics: Society for Learning Analytics Research*. Canada: Society for Learning Analytics Research, 2017.

[9] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Syst. Appl.*, vol. 41, pp. 1432–1462, Mar. 2014.

[10] A. Peña-Ayala, *Learning Analytics: Fundaments, Applications, and Trends: A View of the Current State of the Art to Enhance e-Learning*. Berlin, Germany: Springer, 2017.

[11] A. Peña-Ayala, "Learning analytics: A glance of evolution, status, and trends according to a proposed taxonomy," *Wiley Interdiscipl. Rev.-Data Mining Knowl. Discovery*, vol. 8, no. 3, p. e1243, May/Jun. 2018.

[12] B. Rienties, A. Boroowa, S. Cross, C. Kubiak, K. Mayles, and S. Murphy, "Analytics4Action evaluation framework: A review of evidence-based learning analytics interventions at the Open University UK," *J. Interact. Media Edu.*, vol. 2, no. 1, pp. 1–11, 2016.

[13] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Learning Analytics: From Research to Practice*, A. J. Larusson and B. White, Eds. New York, NY, USA: Springer, 2014, pp. 61–75.

[14] Z. Papamitsiou and A. Economides, "Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence," *Edu. Technol. Soc.*, vol. 17, no. 4, pp. 49–64, 2014.

[15] B. Chen, M. Resendes, C. S. Chai, and H.-Y. Hong, "Two tales of time: Uncovering the significance of sequential patterns among contribution types in knowledge-building discourse," *Interact. Learn. Environ.*, vol. 25, pp. 162–175, Feb. 2017.

[16] E. Barbera, B. Gros, and P. A. Kirschner, "Paradox of time in research on educational technology," *Time & Soc.*, vol. 24, no. 1, pp. 96–108, 2015.

[17] S. Knight, A. F. Wise, X. Ochoa, and A. Hershkovitz, "Learning analytics: Looking to the future," *J. Learn. Anal.*, vol. 4, no. 2, p. 5, 2017.

[18] A. van Leeuwen, N. Bos, H. van Ravenswaaij, and J. van Oostenrijk, "The role of temporal patterns in students' behavior for predicting course performance: A comparison of two blended learning courses," *Brit. J. Educ. Technol.*, vol. 50, no. 1, pp. 1–13, 2019, doi: 10.1111/bjet.12616.

[19] B. Chen, A. F. Wise, S. Knight, and B. H. Cheng, "Putting temporal analytics into practice: the 5th international workshop on temporality in learning data," presented at the Proc. 6th Int. Conf. Learn. Anal. Knowl., Edinburgh, U.K., 2016.

[20] J. Ceddia, J. Sheard, and G. Tibbey, "WAT: a tool for classifying learning activities from a log file," presented at the Proc. 9th Australas. Conf. Comput. Educ., Ballarat, VIC, Australia, 2007.

[21] J. Ceddia and J. Sheard, "Log Files for Educational Applications," presented at the World Conf. Educ. Media Technol. (EdMedia), Montreal, QC, Canada, 2005.

[22] C. Romero, S. Ventura, and E. García, "Data mining in course management systems: Moodle case study and tutorial," *Comput. Edu.*, vol. 51, pp. 368–384, Aug. 2008.

[23] C. G. Marquardt, K. Becker, and D. D. Ruiz, ''A pre-processing tool for Web usage mining in the distance education domain,'' in *Proc. Int. Database Eng. Appl. Symp. (IDEAS)*, 2004, pp. 78–87.

[24] I. Dimopoulos, O. Petropoulou, and S. Retalis, ''Assessing students' performance using the learning analytics enriched rubrics,'' presented at the Proc. 3rd Int. Conf. Learn. Anal. Knowl., Leuven, Belgium, 2013.

[25] W.-Y. Hwang and C.-C. Li, ''What the user log shows based on learning time distribution,'' *J. Comput. Assist. Learn.*, vol. 18, pp. 232–233, Jun. 2002.

[26] L. Tobarra, A. Robles-Gómez, S. Ros, R. Hernández, and A. C. Caminero, ''Analyzing the students' behavior and relevant topics in virtual learning communities,'' *Comput. Hum. Behav.*, vol. 31, pp. 659–669, Feb. 2014.

[27] M. Fakir and K. Touya, ''Mining students' learning behavior in moodle system,'' *J. Inf. Technol. Res.*, vol. 7, no. 4, pp. 12–26, 2014.

[28] T. Haig, K. Falkner, and N. Falkner, ''Visualisation of learning management system usage for detecting student behaviour patterns,'' presented at the Proc. 15th Australas. Comput. Educ. Conf., Adelaide, SA, Australia, 2013.

[29] E. Młynarska, D. Greene, and P. Cunningham, ''Time series clustering of moodle activity data,'' in *Proc. 24th Irish Conf. Artif. Intell. Cogn. Sci. (AICS)*, University College Dublin, Dublin, Ireland, Sep. 2016.

[30] R. Mazza, M. Bettoni, M. Faré, and L. Mazzola, ''MOCLog—Monitoring Online Courses with log data,'' presented at the 1st Moodle Res. Conf., Heraklion, Crete-Greece, 2014.

[31] M. Munk, M. Drlik, and M. Vrábelová, ''Probability modelling of accesses to the course activities in the Web-based educational system,'' in *Computational Science and Its Applications*, vol. 6786, B. Murgante, O. Gervasi, A. Iglesias, D. Taniar, and B. Apduhan, Eds. Berlin, Germany: Springer, 2011, pp. 485–499.

[32] M. Munk and M. Drlík, ''Analysis of stakeholders' behaviour depending on time in virtual learning environment,'' *Appl. Math. Inf. Sci.*, vol. 8, pp. 773–785, Mar. 2014.

[33] M. Munk and M. Drlík, ''Methodology of predictive modeling of students' behavior in virtual learning environment,'' in *Formative Assessment, Learning Data Analytics and Gamification*, S. Caballé and R. Clarisó, Eds. Boston, MA, USA: Academic, 2016, ch. 10, pp. 187–216.

[34] B. Liu, *Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data*, 2nd ed. New York, NY, USA: Springer, 2011.

[35] R. Cooley, B. Mobasher, and J. Srivastava, ''Data preparation for mining World Wide Web browsing patterns,'' *Knowl. Inf. Syst.*, vol. 1, pp. 5–32, Feb. 1999.

[36] V. Chitraa and A. S. Davamani, ''A survey on preprocessing methods for Web usage data,'' *Int. J. Comput. Sci. Inf. Secur.*, vol. 7, no. 3, pp. 78–83, 2010.

[37] C. Romero, J. R. Romero, and S. Ventura, ''A survey on pre-processing educational data,'' in *Educational Data Mining*, vol. 524, A. Peña-Ayala, Ed. Cham, Switzerland: Springer, 2014, pp. 29–64.

[38] E. Gaudioso and L. Talavera, ''Data mining to support tutoring in virtual learning communities: Experiences and challenges,'' in *Data Mining E-Learning*. Southampton, U.K.: WIT Press, 2005, pp. 207–226.

[39] J. Sheard, ''Basics of statistical analysis of interactions data from Web-based learning environments,'' in *Handbook of Educational Data Mining*, C. Romero, S. Ventura, M. Pechnizkiy, and R. S. J. D. Baker, Eds. Boca Raton, FL, USA: CRC Press, 2011.

[40] M. Munk, M. Vrábelová, and J. Kapusta, ''Probability modeling of accesses to the Web parts of portal,'' *Procedia Comput. Sci.*, vol. 3, pp. 677–683, Jan. 2011.

[41] H. Ba-Omar, I. Petrounias, and F. Anwar, ''A framework for using Web usage mining to personalise e-learning,'' in *Proc. 7th IEEE Int. Conf. Adv. Learn. Technol. (ICALT)*, Jul. 2007, pp. 937–938.

[42] C. Romero and S. Ventura, ''Educational data mining: A survey from 1995 to 2005,'' *Expert Syst. Appl.*, vol. 33, pp. 135–146, Jul. 2007.

[43] M. Munk, M. Drlík, J. Kapusta, and D. Munková, ''Methodology design for data preparation in the process of discovering patterns of Web users behaviour,'' *Appl. Math. Inf. Sci.*, vol. 7, no. 1L, pp. 27–36, 2013.

[44] M. Munk and M. Drlík, ''Impact of different pre-processing tasks on effective identification of users' behavioral patterns in Web-based educational system,'' *Procedia Comput. Sci.*, vol. 4, pp. 1640–1649, Jan. 2011.

[45] M. Munk, M. Drlík, L. Benko, and J. Reichel, ''Quantitative and qualitative evaluation of sequence patterns found by application of different educational data preprocessing techniques,'' *IEEE Access*, vol. 5, pp. 8989–9004, 2017.

[46] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa, ''A framework for the evaluation of session reconstruction heuristics in Web-usage analysis,'' *INFORMS J. Comput.*, vol. 15, no. 2, pp. 171–190, 2003.

[47] L. Yan and F. Bo-Qin, ''The construction of transactions for Web usage mining,'' in *Proc. Int. Conf. Comput. Intell. Natural Comput. (CINC)*, 2009, pp. 121–124.

[48] D. W. Hosmer and S. Lemeshow, ''Multiple logistic regression,'' in *Applied Logistic Regression*. Hoboken, NJ, USA: Wiley, 2005, pp. 31–46.

[49] B. H. Baltagi, *Econometrics*. Berlin, Germany: Springer, 2007.

[50] M. Munk, A. Pilkova, L. Benko, and P. Blažeková, ''Pillar 3: market discipline of the key stakeholders in CEE commercial bank and turbulent times,'' *J. Bus. Econ. Manage.*, vol. 18, no. 5, pp. 954–973, 2017.

MARTIN DRLIK received the M.S. degree in biophysics from the Faculty of Mathematics, Physics and Computer Science, Comenius University, Bratislava, Slovak, in 2001, and the Ph.D. degree in theory of computer science education from the Constantine the Philosopher University in Nitra, Nitra, Slovak, in 2009.

Since 2002, he has been an Assistant Professor with the Computer Science Department, Constantine the Philosopher University in Nitra. His research interests include learning analytics, educational data mining, software engineering, and database systems.

Dr. Drlik has been a member of the ACM, since 2007. He was a recipient of the Green Group Award (Best paper) of the International Conference on Computational Science, Workshop on Computational Finance and Business Intelligence, Barcelona, in 2013.

MICHAL MUNK was born in Piešt'any, Slovakia, in 1979. He received the M.S. degree in mathematics and informatics and the Ph.D. degree in mathematics from the University of Constantine the Philosopher, Nitra, Slovakia, in 2003 and 2007, respectively. In 2018, he was appointed as a Professor in system engineering and informatics with the Faculty of Informatics and Management, University of Hradec Králové, Czech Republic.

He is currently a Professor with the Computer Science Department, Constantine the Philosopher University in Nitra. His research interests include data analysis, Web mining, and natural language processing.

Dr. Munk has been a member of the Slovak Statistical and Demographic Society, since 2005. He was a recipient of the Green Group Award (Best paper) of the International Conference on Computational Science, Workshop on Computational Finance and Business Intelligence Barcelona, in 2013.