Impact of multimedia in learning profiles

Ariel Zambrano*

CAETI Research Lab,
UAI University,
Buenos Aires, Argentina
Email: arielzambrano@gmail.com
*Corresponding author

Daniela López De Luise

CI2S Labs,
Francisco Acuña de Figueroa 146,
2nd Fl, C1180AAB,
Ciudad Autónoma de Buenos Aires, Argentina
Email: daniela_ldl@ieee.org

Abstract: The present paper has as original contribution the definition of an automated model of the behaviour of a user against a certain type of images in a context of playful learning. Therefore, the entropy is used to classify profiles, starting from temporary information, which is mixed with certain characteristics previously extracted from the images. The aim of all this is to determine to what extent visual images trigger functions of comprehension and abstraction on topics of high degree complexity. Part of the obtained model is intended to generate learning profiles, which will enrich in the future with other non-invasive device, and to observe the behaviour of the user. For example: cameras, monitory keyboard, mouse and among others. The profiles are discovered and described with the minimum information needed. The collected information is processed with bio inspired techniques, which are essentially bases on 'deep learning' concept.

Keywords: audiovisual techniques; engineering teaching; video games; learning model; deep learning; multimedia; data mining.

Reference to this paper should be made as follows: Zambrano, A. and López De Luise, D. (2023) 'Impact of multimedia in learning profiles', *Int. J. Advanced Intelligence Paradigms*, Vol. 24, Nos. 1/2, pp.12–37.

Biographical notes: Ariel Zambrano completed the Information Systems Engineer at Universidad Abierta Interamericana (UAI). He has a wide experience in software development, databases analysis and design. He is an Assistant Researcher at CAETI Investigation Center and Special mention for project of postgraduate thesis in Congress CIITI 2016. He is a member of the MIDA-Learnitron project developed by the researcher PhD Daniela López de Luise.

Daniela López De Luise is a PhD in Computer Science (UNLP). She is a member of the Advisor Committee of the Swiss Innovation Valley and the Director of CI2S and IDTI research Labs. She is a member of CAETI and a Local Founder and first Chair of Local IEEE Computational Intelligence Society (and current Vice-Chair), a Local IEEE SIGHT and WCI. She was the

President of IEEE Argentina and declared 'Outstanding Engineer' by IEEE. She wrote many books, articles and scientific papers in the field of computational intelligence. She is the author of theories like *Mowphosyntactic: Linguistic Wavelets and Harmonic Systems*.

1 Introduction

Modelling the learning profiles before a new concept requires the use of different strategies. For example, de Luise et al. (2016) explain the architecture and functioning of the LEARNITRON learning model and the MIDA virtual museum. This project has as one of its main objectives, to promote the teaching of engineering through video games as a tool. In this way, an attempt is made to tackle the problem of university desertion in engineering careers (Pérez et al., 2013).

To get a notion of this problem, Table 1 is shown (Pérez et al., 2013). It can be clearly seen that the percentage of students dropping out of engineering degrees is very high, reaching approximately 70%.

Table 1 Abandonment in engineering careers

Engineering career		*Current student situation*			*Total*
		Abandonment	*Regular*	*Graduated*	
201 computer engineering	N	8,946	2,731	1,039	12,716
	%	70.4%	21.5%	8.2%	100%
202 electronic engineering	N	1,556	541	211	2,308
	%	67.4%	23.4%	9.1%	100%
203 industrial engineering	N	924	591	33	1,548
	%	59.7%	38.2%	2.1%	100%
207 civil engineering	N	94	169	0	263
	%	35.7%	64.3%	0.0%	100%
TOTAL	*N*	*11,520*	*4,032*	*1,283*	*16,835*
	%	*68.4%*	*24.0%*	*7.6%*	*100%*

Although this study focuses only on one university in Argentina, university dropout is a problem that affects many countries, as can be seen in the data obtained in González (2006).

The MIDA prototype, through the collection of information to obtain usage statistics, can help students to improve their understanding of engineering subjects, all through a playful aspect (de Luise et al., 2016).

1.1 Video games in teaching

Since the late 1980s, the importance of having different audio-visual media in classrooms has been mentioned as a support for teaching (Peris, 2008; Mitchell and Savill-Smith, 2004; Tomás, 2009). At present, video games are incorporated. We can consider them a didactic element (Charsky, 2010) and they have the advantages of having a playful

component, massification, versatility and relative easy access (Domínguez and Antequera, 2012).

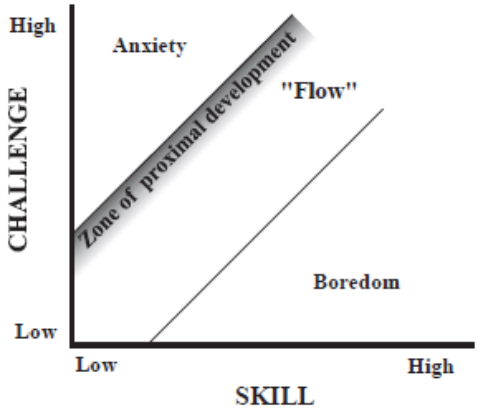
Video games will be effective in the teaching-learning process, if they have certain specific characteristics. Some of these are explained in Kiili (2005). In this paper, the importance of using a positive psychology model is mentioned. It also details the three fundamental elements that must be present in video games and which are the most important in learning:

- the story
- balance (difficulty levels, rewards, awards)
- optimise the cognitive load (videos, sounds, images, animations).

In this work is also mentioned a concept taken from the psychology: the flow model (Kiili, 2005). In this model, the existence of an optimal learning experience is discussed. This particular experience is due to a personal cognitive-emotional state, originated by the balance between the challenge and the skills in the presented tasks.

If that balance does not exist, two situations can occur. In the first one, if the challenges exceed the individual competences, an anxiety state is created due to an excess of difficulty. The other situation is the opposite. If skills far outweigh the challenges, the individual will be close to boredom and therefore less motivated. The flow process is shown in Figure 1.

Figure 1 Three channel model of flow

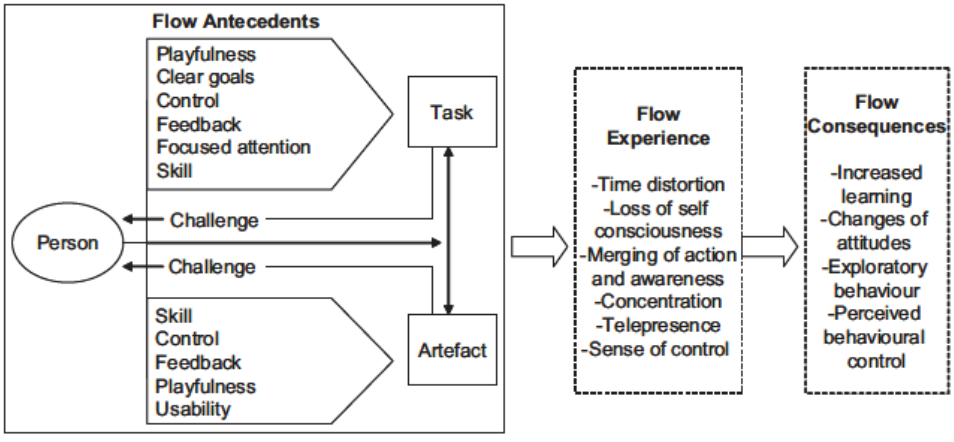


Source: Adopted from Csíkszentmihályi (1975)

In McCauley (2011) it is explained how the flow model is applied in computational environments and what is the impact on the person and the consequences (see Figure 2).

Another term that is mentioned is that of gamification. It is the use of game mechanics in non-play environments in order to enhance motivation, concentration, effort, loyalty and other positive values common to all games. From this point of view can positively impact learning processes (McCauley, 2011).

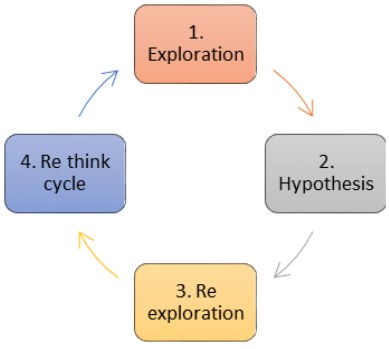
Figure 2 Blueprint of flow in computer-mediated environments



It is also mentioned that there are four processes in video games, similar to implementing a scientific method in real life. The four processes are: exploration, hypothesis, re exploration and rethinking the cycle.

The four processes are shown in Figure 3.

Figure 3 Four steps of learning in video games (see online version for colours)



The relevance of these processes in video-games is clearly demonstrated in Steinkuehler and Duncan (2008). In this work the behaviour of players of a massively multi-player online video game is being analysed. One of the main difficulties facing players in such games is to defeat the final level bosses.

It is interesting to analyse the approach used by these players. A group of them use spreadsheets in Excel in which they enter all the information collected about the behaviour of each boss, including:

- type of attacks that affect them
- attacks to perform
- damage it causes and when.

The mathematical model explains how the boss works, predicts how to defeat him. The first models obtained did not work well, then, the group discussed how to improve it trying to defeat the level's boss using the strategies of the updated model. This way the players were practicing science, applying a scientific method. They established a hypothesis and then collected evidence to see if that hypothesis is correct. If it is not, they improve it until it corresponds to the observed data.

1.2 Video-games and emotions

The technological advances of both software and hardware have made video-games create an increasingly realistic virtual environment and allow players to be immersed in them. This way it is possible to make the players experience emotions, condition necessary to stimulate the learning (Marcano, 2006).

This experience of immersion is created by video games through multiple sensory stimulations. That affects on:

- cognitive structures
- information processing
- experiences
- short and long term memories.

Thus, they provoke an increase of learning rate through the experience of the game (Marcano, 2006). The effects of video games on the body, short and long-term memory were verified in Shilling et al.(2002) and Ulate (2002), where a study was carried out on two groups of players. Both teams play with a combat simulator with very high level of realism. After playing, users must fill in a questionnaire to assess the number of details of the game that are remembered.

A test with 'high stimulation' is performed, placing headphones at high volume, which makes the game more realistic. They perform also a second test with low volume and no headphones. Many body sensors collect information about physical reactions during both tests. Among other sensors it can be mentioned cardiac, thermal and another for the galvanic response of the skin. Players with high stimulation experienced remarkable changes in heart rate and skin resistance, showing the effects on the body for those players. That is a clear indicator for those that feel 'immersed' in the game.

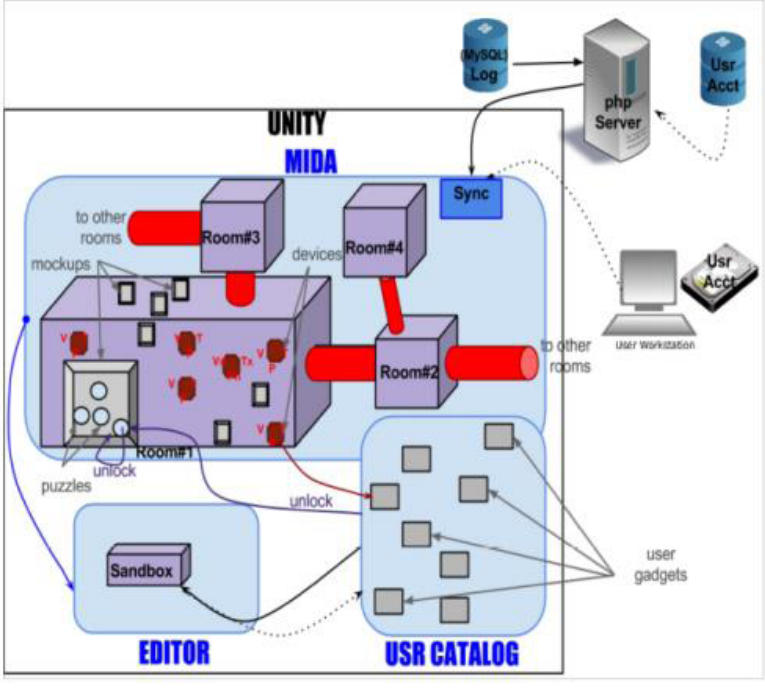
Results for players with high stimulation show a better performance recalling more details of the game experience even 24 hours after it. Therefore, it can be said that video games can be a good educational tool (Steinkuehler and Duncan, 2008).

The rest of this paper is organised as follows: Section 2 briefly describes the architecture of MIDA project. Section 3 describes the specific learning goals. Section 4 describes the data analysis performed. Section 5 applies principal component analysis to reduce dimensionality and find relevant information descriptors. Section 6 compares hierarchical clustering and k-means to derive main user behaviours. Section 7 explains the profiles obtained in the previous section. Section 8 shows prediction model and how it works. Section 9 has conclusions and future work.

2 MIDA architecture and features

In order to provide solutions to the problem of students' desertion, MIDA provides a platform that introduces gamification tips in engineering careers. It is video game, enhanced with a small kernel that implements an intelligent learning model. Such combination is intended to help teaching-learning engineering concepts. It has a dynamic and playful interface able to evaluate how the learning process evolves during the user interaction.

Figure 4 MIDA global architecture (see online version for colours)



Among others, MIDA has the following characteristics (de Luise et al., 2016):

1 Traces behavioural information to be able to discriminate junior from senior visitors and their abilities.

2 Shows ancient devices as a way to introduce complex concepts from the physics, mathematics, chemical, hydraulics, mechanics, etc.

3 Provides a confident and efficient tool to assess the impact of certain multimedia artefacts to learning and conceptualisation processes. Among other usages of such model, could be custom advice for a specific classroom about which are the best suited sequence of concepts, multimedia approaches, learning speed and reinforcement requirements.

4 Shows and explains real world applications of abstract concepts that are usually hard to visualise.

5 Promotes engineering careers, as it can be visited by anyone that takes time to play with devices, read/listen information and play in the sandbox.

6 Tracks and evaluates certain factors in the dramatic decrease of the engineering students and high desertion in first levels. This is done by indirect tracking the persistence, motivation and other clues.

Figure 4 shows the MIDA architecture.

The figure shows hot the video-game (enclosing black rectangle) is connected to other components that are mounted out of the UNITY (c) platform, mainly a data server or a terminal of user (de Luise et al., 2016).

2.1 How it works

MIDA is organised in several rooms according to the type of concepts exposed in the showroom. But all of them have the following two components (de Luise et al., 2016):

1 Mock-ups: with many puzzles that may or may not be unlocked according to the learning stage of the user. The more puzzles are successfully completed; the more ones will be unlocked. This is because a puzzle is a mini test of certain tip or concept within a topic.

2 Devices: ancient objects selected by its characteristics (most of them simple machines that use simple physics, mathematics, mechanics, etc.). Every object in the showroom has videos (introductory or function's demo), texts (introductory or technical) and images or photos. Besides, it is possible to assemble and disassemble the device to understand its architecture.

In despite of the fact that the museum shows antiques and simple devices, it also serves to teach basic engineering principles whenever the user must learn how the exposed objects operate. Every room associates objects to challenges that evaluate the level of understanding of the user. As the user plays within the Museum a specific intelligent activity is performed to derive the user learning preferences. This would be used to bias the Museum behaviour, improving the visitors' experience (de Luise et al., 2016).

3 Learning profiles

This work is part of the project named LEARNITRON, a branch of the research project MIDA (de Luise et al., 2016). It aims to add a set of tools based on machine learning and other approaches from the Computational Intelligence. All those modules intend to provide self-adaptation of MIDA's behaviour and to enhance the experience of learning certain very abstract concepts in engineering. As a spring off, there are also a set of metrics and models to describe users and their performance.

3.1 Why learning profiles

As the student population trend to change due to many factors, the main goal of this work is to automatically derive a model describing typical learning profiles through the usage of a video game (MIDA museum). The model is able to process real-time information

from the MIDA repository in order to detect the current user profile. Data are collected, pre-processed and feed to the model proposed here, obtained mainly with data mining. That way it is possible to determine the impact of the stimuli generated by the museum during the learning process of the players.

3.1.1 Specific goals

But the global goal requires a set of more specific goals, in order to be accomplished. Those are:

- Evaluate multimedia aspects that serve to specify parameters concerning expressiveness, efficiency and flexibility of the learning process for every individual, considering the interference of non-tangible aspects of the personality.
- Generate a theoretical model based on the application of soft computing technologies to detect, analyse and define the best algorithm for the alternatives found in the previous point.
- Build a model with statistically evaluated tests that validate the model.
- Perform a reasonable good prototype implementing the model, under the restrictions of applicability and reproducibility of the solution. It should be able to process at industrial level on-line services for entities and individuals.
- Expand the current knowledge of the area covering multidisciplinary aspects partially studied to the present and extending the scope of application in terms of age range and teaching topics on which it could be applied.

4 Data analysis and testing

In order to derive the current learning profiles, the visitor activity is registered in log files such as the one shown in Figure 5.

Figure 5 Log file example (see online version for colours)

localhost » metricas » actions

Examinar | Estructura | SQL | Buscar | Insertar | Exportar | Importar

+ Opciones

	id	userId	date	actionId	objectId	value
Editar Copiar Borrar	1	10	2014-03-17 00:36:30	0	10000	1
Editar Copiar Borrar	2	10	2014-03-17 00:36:31	0	20001	1
Editar Copiar Borrar	3	10	2014-03-17 00:36:32	0	40001	1
Editar Copiar Borrar	4	10	2014-03-17 00:36:37	0	50001	1
Editar Copiar Borrar	5	10	2014-03-17 00:36:39	0	50002	1
Editar Copiar Borrar	6	10	2014-03-17 00:36:40	0	50001	1
Editar Copiar Borrar	7	10	2014-03-17 00:36:41	0	60001	1
Editar Copiar Borrar	8	10	2014-03-17 00:36:44	0	50001	1
Editar Copiar Borrar	9	10	2014-03-17 00:36:51	0	40001	1
Editar Copiar Borrar	10	10	2014-03-17 00:36:55	0	40002	1

The figure corresponds to less than a minute of playing, with a visitor identified as ID 10. This section will show what type of information is initially collected, as well as the name of the module that is recording the information.

But keeping track of the activity not only includes the exact timing of each action, but also the component (objectID) and the action (value).

The record in Figure 5 corresponds to a user who enters the room (objectID 20001), interacts with an invention (objectID 40001), reads their texts (objectID 50001 and 50002), assembles and disassembles the device but does not look at the Video but the photo (objectID 60001). The collected data shows the way in which the user prefers to access the knowledge implicit in the invention's manipulation.

Table 2 shows a histogram of frequencies of the use of the objects during the test session.

Table 2 Usage frequency of the objects

ObjectID	*FA*	*FR*	*MV*
[1-26001]	7	0.21	13,001
[26001-52001]	7	0.21	39,001
[52001-78001]	1	0.03	65,001
[78001-104001]	10	0.3	91,001
[104001-1320001]	8	0.24	117,001

It shows that the player used very little objectID 52001-78001 (text) meaning that the player is not interested in obtaining additional information.

4.1 Pre-processing data

In order to manage the log data, they must be prepared to be processed in the statistical software Infostat © (Di Rienzo et al., 2015). This task includes:

- convert the records into a comma-separated format (csv)
- remove irrelevant data columns
- convert date format to Julian
- detect and remove outliers
- add hour difference column between record and record of the same ID.

It can be seen (Table 3) that the VALUE column does not provide information and can be removed.

The following is the list of actions performed on the log file in order to get the data correctly formatted for processing:

1 the names are added to the columns in a text file and saved in.csv format

2 the CSV file is imported into Microsoft Excel software (González, 2006)

3 the DATE column must be reformatted from date (dd/mm/yyyy) and time (hh: mm: sss) format

4 the DATE should be displayed with hours, minutes and seconds.

The data are reorganised as shown in Table 4.

Table 3 Log file

ActionsID	*UserID*	*ActionID*	*ObjectTypeID*	*ObjectID*	*Date*	*Value*
1	1	9	1	1	08/062014 12:48:40	0
2	1	2	1	1	08/062014 12:48:40	0
3	1	1	2	1	08/062014 12:48:42	0
4	1	1	4	1	08/062014 12:48:47	0
5	1	1	2	1	08/062014 12:48:51	0
6	1	1	3	1	08/062014 12:48:52	0
7	1	5	3	1	08/062014 12:48:53	0
8	1	1	2	1	08/062014 12:48:54	0
9	1	1	4	1	08/062014 12:48:56	0

Table 4 Log file imported

ActionsID	*UserID*	*ActionID*	*ObjectTypeID*	*ObjectID*	*Date*
1	1	9	1	1	08/062014 12:48:40
2	1	2	1	1	08/062014 12:48:40
3	1	1	2	1	08/062014 12:48:42
4	1	1	4	1	08/062014 12:48:47
5	1	1	2	1	08/062014 12:48:51
6	1	1	3	1	08/062014 12:48:52
7	1	5	3	1	08/062014 12:48:53
8	1	1	2	1	08/062014 12:48:54
9	1	1	4	1	08/062014 12:48:56
10	1	1	8	4	08/062014 12:48:59
11	1	2	8	4	08/062014 12:49:01
12	1	3	8	4	08/062014 12:49:02
13	1	1	2	1	08/062014 12:49:03
14	1	3	1	1	08/062014 12:49:04

As one further step column DATE must be adapted to a number format, in order to be able to measure the times in which a player occupies in each invention in the game, that date must be converted to Julian format.

The dates have an information component that disproportionately increases the entropy with respect to the rest of the variables, so if should not be included in the statistical analysis or it must be modified to avoid that.

The tasks performed for the date and time were:

1 create a new column named TIME, which only has the time of the date column

2 time is removed from DATE column, leaving only in DD/MM/YYYY format

3 apply the formula to pass Julian to the date column

4 apply the formula to convert Julian to the TIME column. To do this, the chosen cell format must be ‘Custom – [ss]’.

Formula in Excel for date in Julian is in equation (1):

$$\text{TEXT}((\text{date}),"yy")\&\text{TEXT}(((\text{date})-\text{DATEVALUE}("1/1/"\&\text{TEXT}((\text{date}),"yy"))+1),"000") \quad (1)$$

The columns are organised then as shown in Table 5.

Table 5 Log file modified

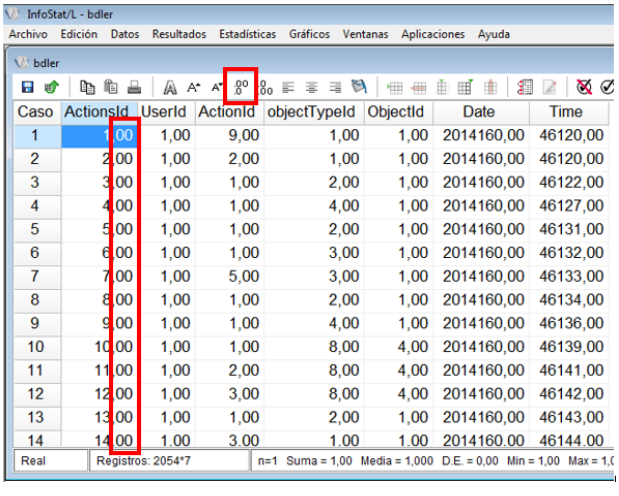
ActionsID	*UserID*	*ActionID*	*ObjectTypeID*	*ObjectID*	*Date*	*Time*
1	1	9	1	1	2014160	46120
2	1	2	1	1	2014160	46120
3	1	1	2	1	2014160	46122
4	1	1	4	1	2014160	46127
5	1	1	2	1	2014160	46131
6	1	1	3	1	2014160	46132
7	1	5	3	1	2014160	46133
8	1	1	2	1	2014160	46134
9	1	1	4	1	2014160	46136
10	1	1	8	4	2014160	46139
11	1	2	8	4	2014160	46141
12	1	3	8	4	2014160	46142
13	1	1	2	1	2014160	46143
14	1	3	1	1	2014160	46144
15	1	2	1	1	2014160	46153

Finally, the file should be saved as CSV format.

4.2 Import to statistical software

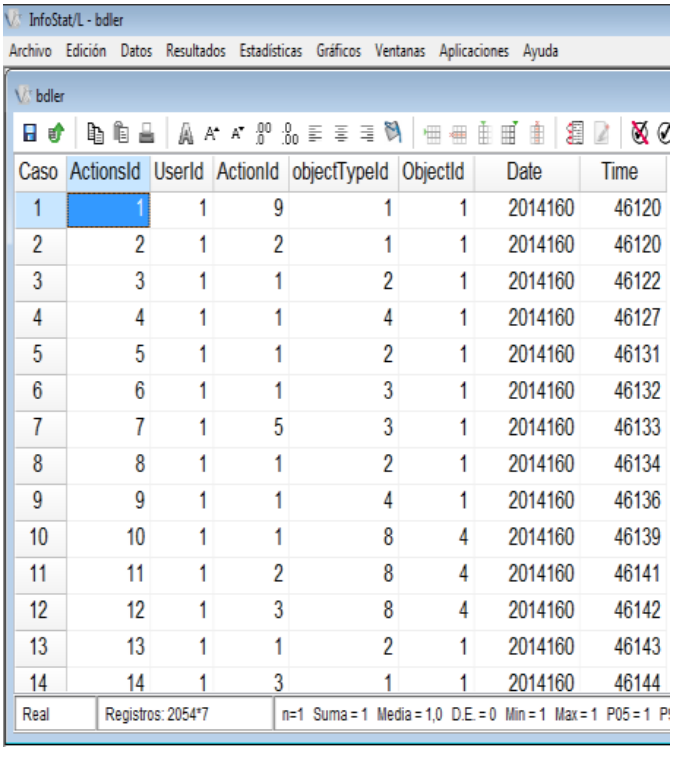
Once the csv file is created, it is imported into the statistical software. Once the file is imported, it is noted that the numeric data has two additional decimal places (, 00). These decimals should be removed using the corresponding option on the top button bar. See the formatting in Figure 6.

Figure 6 Log file modification (see online version for colours)



The file is successfully imported and formatted is in Figure 7.

Figure 7 Log file after modifications (see online version for colours)



InfoStat/L - bdler

Archivo Edición Datos Resultados Estadísticas Gráficos Ventanas Aplicaciones Ayuda

bdler

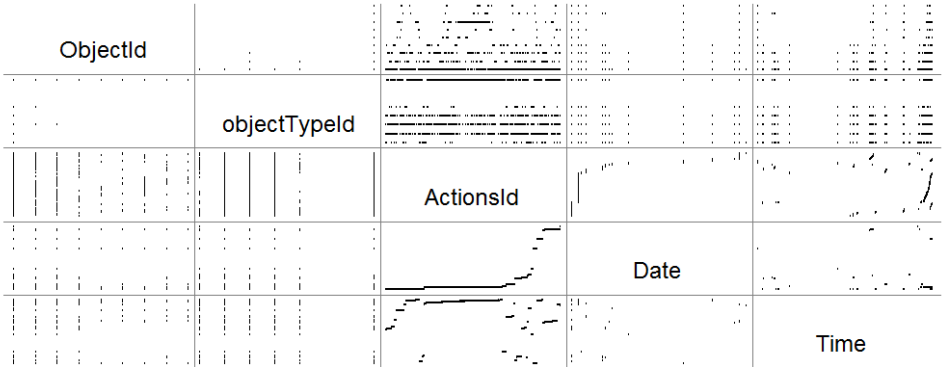
Caso	ActionsId	UserId	ActionId	objectTypeId	ObjectId	Date	Time
1	1	1	9	1	1	2014160	46120
2	2	1	2	1	1	2014160	46120
3	3	1	1	2	1	2014160	46122
4	4	1	1	4	1	2014160	46127
5	5	1	1	2	1	2014160	46131
6	6	1	1	3	1	2014160	46132
7	7	1	5	3	1	2014160	46133
8	8	1	1	2	1	2014160	46134
9	9	1	1	4	1	2014160	46136
10	10	1	1	8	4	2014160	46139
11	11	1	2	8	4	2014160	46141
12	12	1	3	8	4	2014160	46142
13	13	1	1	2	1	2014160	46143
14	14	1	3	1	1	2014160	46144

Real | Registros: 2054*7 | n=1 Suma = 1 Media = 1,0 D.E. = 0 Min = 1 Max = 1 P05 = 1 P

4.3 Data visualisation

To get an overview of the data in the log it is useful to use the scatter-plot matrix. This graph is practical for cases where more than one variable is measured, but not so many as to prevent all relations from being paired.

Figure 8 SPlotM obtained



To elaborate this graph in InfoStat, in the menu graphs select the sub-menu matrix of dispersion diagrams (SplotM). The selected variables are ObjectId, ObjectTypeId, ActionsId, date, time.

The matrix obtained is shown in Figure 8.

The advantage of this type of matrix is that they allow seeing in a relatively simple way if there are inconsistent data in the relations between the variables. These data are known as outliers.

4.4 Outliers detection

An outlier is an observation or set of observations that appear to be inconsistent with the rest of the data set.

The characteristic in the observation of outliers is the impact that it produces in the statistic when the data will be analysed. The presence of outliers in a dataset can lead to errors in our attempt to make inferences about the population from which they come, hence the presence of these data raises a fundamental problem in data analysis.

Analysing the generated matrix, possible outliers could be analysed in the relation of the following variables.

1 objectTypeID vs. time

2 actionsID vs. time

3 actionsID vs. date.

In order to continue, an operation known as principal component analysis is performed through the bi-plot graph.

5 Principal component analysis (PCA)

It is a statistical technique of information synthesis, or reduction of the dimension (number of variables). That is, before a set of data with many variables, its goal is to reduce them to a smaller number by losing the least amount of information possible.

The bi-plot graph of principal component analysis is used for multivariate observations where all variables are quantitative in nature.

A principal components analysis is performed to combine the variables into indexes and then dispersion diagrams are constructed using these indexes to define the axis.

Synthetic indexes or variables are called principal components (CP). The most widespread graphic is based on the first two main components (CP1 and CP2) because these combinations are the ones that best explain the differences between units of analysis. The graph is called bi-plot, because in the same space (which make up the CP1 and CP2) represent the units of analysis and variables, i.e., the two dimensions of the data table.

The bi-plot obtained from the variables ObjectTypeId and time is shown in Figure 9.

To identify the possible outlier, select the point farthest from the others and click on it. The number that appears to us is the number of ActionId that we can consult to analyse, in this is the case 2,052 (see Figures 10 and 11). The PCA generated are stored for future analysis.

Figure 9 ObjectTypeID vs. time (see online version for colours)

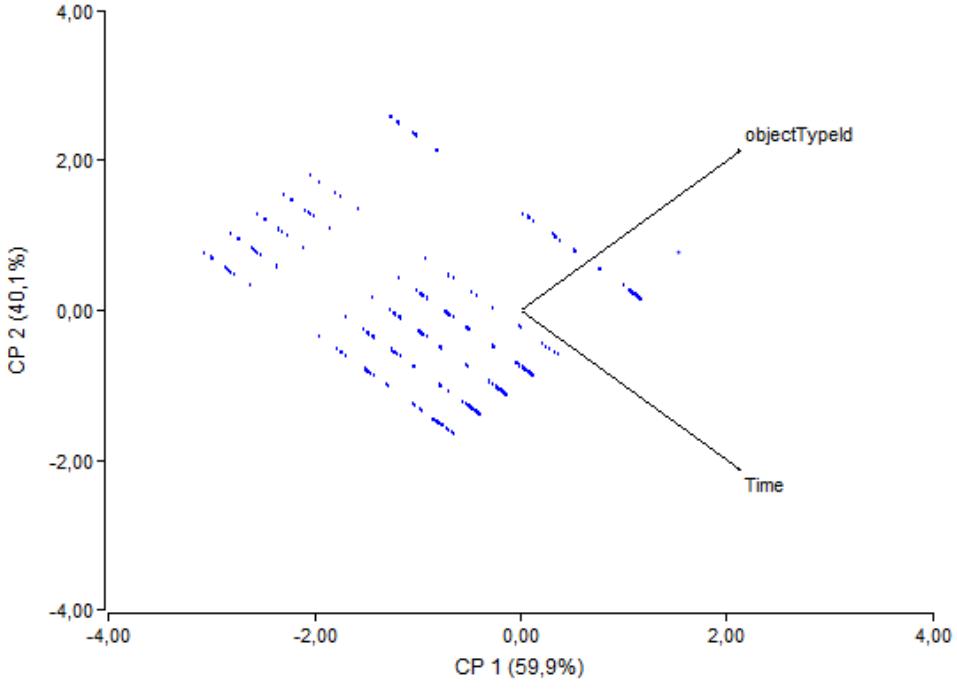


Table 6 PCA analysis – ObjectTypeID vs. time

Principal component analysis			
Read cases 2,054			
Omitted cases 0			
Classification variables			
		Eigenvalues	
Lambda	*Value*	*Proportion*	*Acum. prop.*
1	1.2	0.6	0.6
2	0.8	0.4	1
		Eigenvectors	
Variables	*e1*	*e2*	
ObjectTypeID	0.71	0.71	
Time	0.71	-0.71	

A similar analysis is performed for ActionsID and time. The bi-plot of rotated axis (CP1 and CP2) and original ActionsID, time shows the relationship between them (Figure 12).

Figure 10 Outlier detected (see online version for colours)

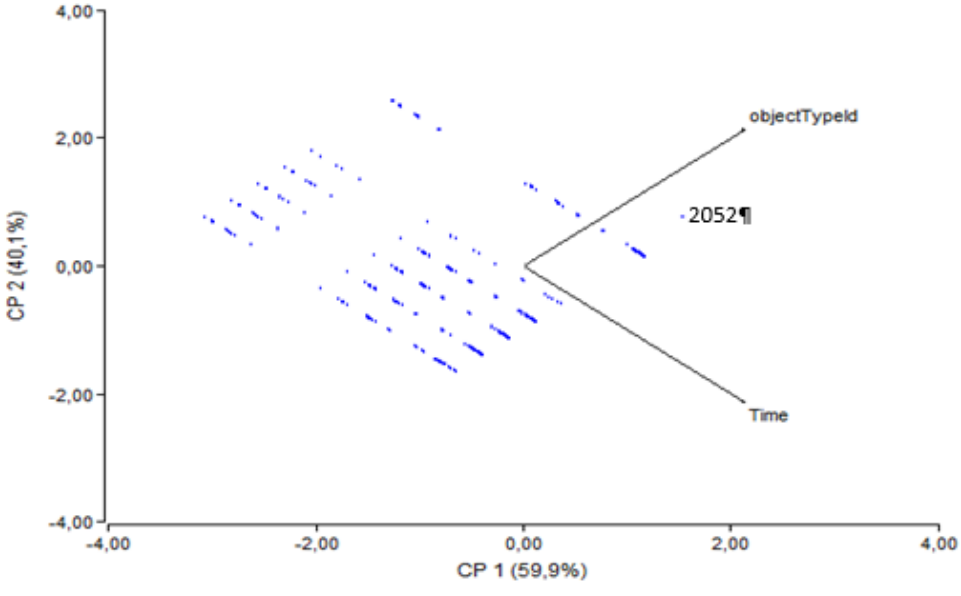


Figure 11 Outlier detected in table (see online version for colours)

ActionsId	UserId	ActionId	objectTypeId	ObjectId	Date	Time
2052	23	2	10	1	2014233	78241

Figure 12 ActionsID vs. time (see online version for colours)

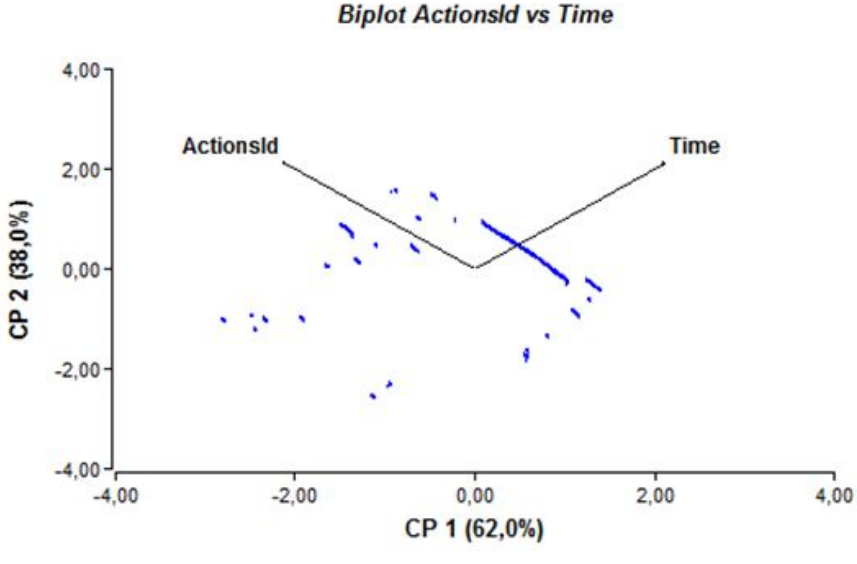


Table 7 PCA analysis – actionsID vs. time

Principal component analysis			
Read cases 2,054			
Omitted cases 0			
Classification variables			
		Eigenvalues	
Lambda	*Value*	*Proportion*	*Acum. prop.*
1	1.24	0.62	0.62
2	0.76	0.38	1
		Eigenvectors	
Variables	*e1*	*e2*	
ActionsID	–0.71	0.71	
Time	0.71	0.71	

And the remaining PCA for ActionsID vs. time appear as shown in Figures 13 and 14.

Figure 13 Outliers detected (see online version for colours)

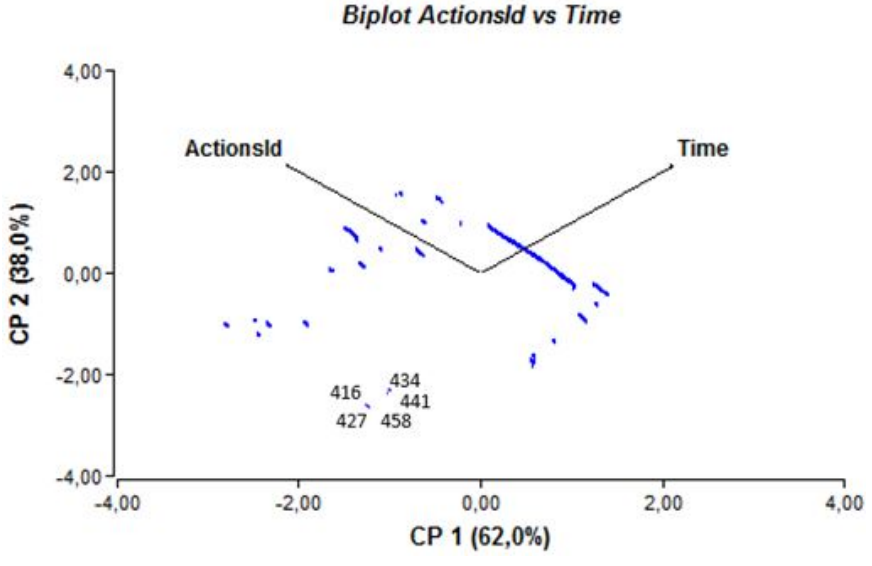


Figure 14 Outlier detected shown in table (see online version for colours)

ActionsId	UserId	ActionId	objectTypeId	ObjectId	Date	Time
416	8	4	3	1	2014160	4311
427	8	2	8	1	2014160	4430
434	8	3	1	1	2014160	4480
441	8	1	4	1	2014160	10897
458	8	3	1	1	2014160	12045

The records correspond to an expert user who has taken less than normal, which is part of his behaviour. Finally, PCA for ActionsID and time are in Figure 15.

Figure 15 ActionsID vs. date (see online version for colours)

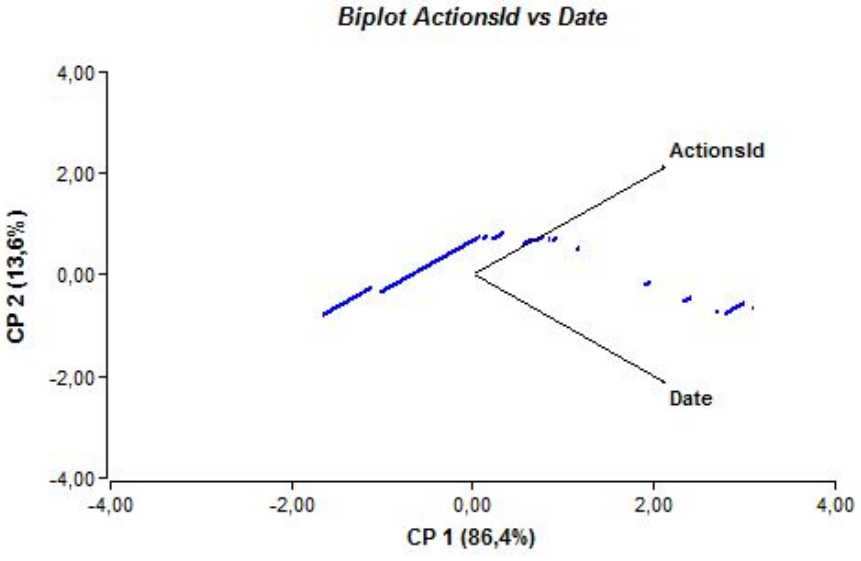


Table 8 PCA analysis – actionsID vs. date

Principal component analysis			
Read cases 2,054			
Omitted cases 0			
Classification variables			
		Eigenvalues	
Lambda	*Value*	*Proportion*	*Acum. prop.*
1	1.73	0.86	0.86
2	0.27	0.14	1
		Eigenvectors	
Variables	*e1*	*e2*	
ActionsID	0.71	–0.71	
Time	0.71	0.71	

5.1 Actions on outliers

There are two users with 'outlier' candidates, eight and 23. The record number 23 is clearly an outlier, shows a different behaviour. In contrast, the record number eight shows a different behaviour since it seems to be a minimal deviation but shares general characteristics of data. Consequently, record number 23 must be discarded but not record eight. The next step is proceeded to delete that record before continuing.

6 Clustering analysis

It is a multivariate technique that seeks to group elements trying to achieve maximum homogeneity in each group and the largest difference between groups. It is being applied to detect natural similarities in data. In this case, it used to derive the common characteristics of main natural data groupings.

6.1 Hierarchical clustering

An analysis of hierarchical clusters is performed. The dendogram is the graphical representation. It helps interpret the result of the clustering. This activity is performed with Minitab software (Minitab, 2014). Results are described below. It is important to note that clusters ID are automatically determined by the platform.

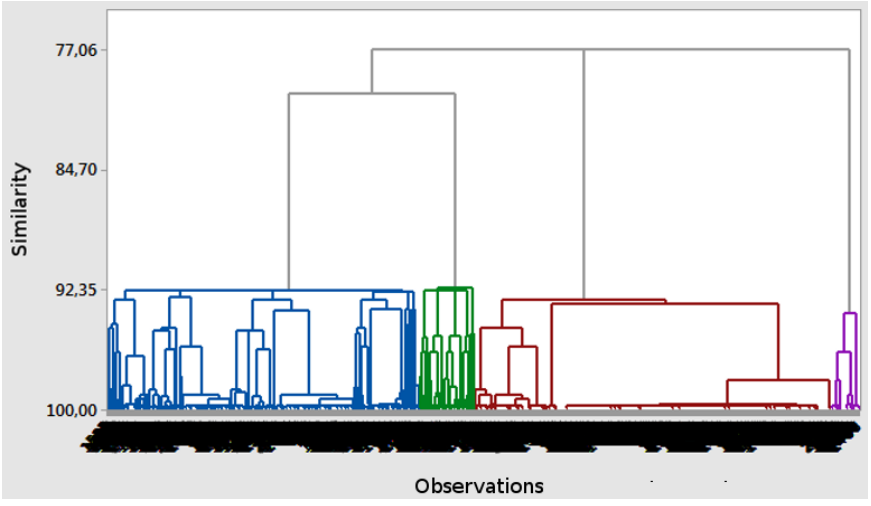
6.1.1 Number of clusters: 4

The dendogram obtained shows that there are four identifiable clusters (Figure 16).

Table 9 Clustering analysis

Number of clusters: 4				
Distance evaluation				
ClusterID	*# cases*	*Accum. quadratic dist. to centroid*	*Average distance to centroid*	*Max. distance to centroid*
Cluster 1	853	446.447	0.672288	1.70421
Cluster 2	975	230.382	0.396410	1.31871
Cluster 3	149	32.858	0.408025	0.93411
Cluster 4	76	4.598	0.218349	0.37846
Cluster centroids				
Principal component variable	*Cluster1*	*Cluster2*	*Cluster3*	*Cluster4*
CP1	–0.525458	0.870847	–2.19786	–0.96551
CP2	–0.917070	0.419556	1.14546	2.66473
Distance between cluster centroids				
	Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*
Cluster 1	0.00000	1.93294	2.65536	3.60874
Cluster 2	1.93294	0.00000	3.15340	2.90052
Cluster 3	2.65536	3.15340	0.00000	1.95624
Cluster 4	3.60874	2.90052	1.95624	0.00000

Figure 16 Dendogram (see online version for colours)



6.2 K-means

It is a method designed to assign cases to a fixed number of groups, whose characteristics are not known, but are based on a set of variables that must be quantitative. It is a method of grouping cases that is based on the distances between them in a set of quantitative variables. When a K-means is performed, the results are as follows. It is important to note that clusters ID are automatically determined by the platform.

Table 10 Analysis of K-means

Analysis of K-means				
Clusters: CP1; CP2				
Number of clusters: 4				
In the distance				
Observation	*#clusters*	*Accum. quadratic distance*	*Average distance to centroid*	*Max. distance to centroid*
Cluster 1	561	90.818	0.343	0.891
Cluster 2	225	23.058	0.929	1.434
Cluster 3	292	73.635	0.447	1.061
Cluster 4	975	230.382	0.396	1.319
		Centroids		
Centroid variable	*Centroid cluster 1*	*Centroid cluster 2*	*Centroid cluster 3*	*Centroid cluster 4*
CP1	–0.2730	–1.7816	–1.0105	0.8708
CP2	–1.2462	1.6586	–0.2847	0.4196

Table 10 Analysis of K-means (continued)

Analysis of K-means *Clusters: CP1; CP2* *Number of clusters: 4*				
	The distance between cluster centroids are:			
	Centroids cluster 1	*Centroids cluster 2*	*Centroids cluster 3*	*Centroids cluster 4*
Cent of cluster 1	0.0000	3.2732	1.2118	2.0207
Cent of cluster 2	3.2732	0.0000	2.0907	2.9276
Cent of cluster 3	1.2118	2.0907	0.0000	2.0088
Cent of cluster 4	2.0207	2.9276	2.0088	0.0000

6.3 Statistical evaluation and analysis

This section performs a comparison and analysis of the results from the two previous approaches. A summary of the results from k-means and dendograms are in Table 11.

Table 11 Dendogram and K-means comparison

		k-means		
Variable	*Cluster 1*	*Cluster 2*	*Cluster 3*	*Cluster 4*
CP1	–0.2730	–1.7816	–1.0105	0.8708
CP2	–1.2462	1.6586	–0.2847	0.4196
		Dendogram		
Variable	*Cluster 1*	*Cluster 2*	*Cluster 3*	*Cluster 4*
CP1	–0.525458	0.870847	–2.19786	–0.96551
CP2	–0.917070	0.419556	1.14546	2.66473

Considering that cluster's numbers are determined every time the algorithm is started and that each algorithm numerates them independently, then the first step is to find if there is a correspondence between clusters in k-means and dendogram.

Observing summaries in both cases, it is possible to find a remarkable analogy between them:

- rotation axis CP2 in dendogram corresponds to CP2 in k-means, since both rotate 3 variables to the negative values and the rest to positive ones
- CP1 in dendogram may correspond to CP1 in k-means, since both show negative rotation for two variables.

To confirm that hypothesis, the behaviour of each cluster is analysed.

- CP1 in k-means vs. dendogram:
  a cluster 2 (KMeans) is related to cluster 3 (dendogram) due to both have minimal rotation

b cluster 1 (KMeans) is related to cluster 1 (dendogram) since both present maximal rotation in negative values

c cluster 3 (KMeans) corresponds to cluster 4 (dendogram) since both cases are intermediate values for rotation

d cluster 4 (KMeans) corresponds to cluster 2 (dendogram) since it is the unique positive rotation.

- CP2 in k-means vs. dendogram:
  a cluster 1 (KMeans) corresponds to cluster 1 (dendogram) since both present highest rotation in negative value
  b cluster 3 (KMeans) corresponds to cluster 2 (dendogram) since both present minimal rotation (the difference may be due to statistical errors)
  c cluster 2 (KMeans) corresponds to cluster 3 (dendogram) since both presents minimal total value
  d cluster 4 (KMeans) corresponds to cluster 4 (dendogram) since both presents reduced positive rotation.

Results show that clusters are the same in case of cluster 1. For cluster 3 in CP2, it corresponds to cluster 2 (instead of cluster 4). But if cluster 2 is taken (its values are reduced) then both clustering are identical. To do so, it is mandatory to hypothesise that cumulative errors have inversed the rotation.

Reorganising according the precious criteria, then CP2 has the following correspondence:

- cluster 1 (KMeans) corresponds to cluster 1 (dendogram)
- cluster 3 (KMeans) corresponds to cluster 4 (dendogram)
- cluster 2 (KMeans) corresponds to cluster 3 (dendogram)
- cluster 4 (KMeans) corresponds to cluster 2 (dendogram).

Summarising, it can be said that there is a correspondence between clustering and results are compatible in both cases and models are similar.

7 Profiles model

Infostat generates a dispersion diagram with the main components and where it is shown to which conglomerate obtained belongs.

The dispersion is in Figure 17, where each colour represents a cluster:

Once the diagram is constructed, the form of the obtained point clouds is analysed to determine the relationships between the four data types.

As conclusion, taking the model derived with k-means, it can be said that there are four profiles that indicate that it is possible to manage global preferences of the users in a systematic way.

From here, k-means is taken as model since it describes better the total and partial error, while dendogram is being used here just to verify results from k-means.

Analysing the centroids of the clusters (from dendogram in Section 6), it can be said that profiles corresponds to:

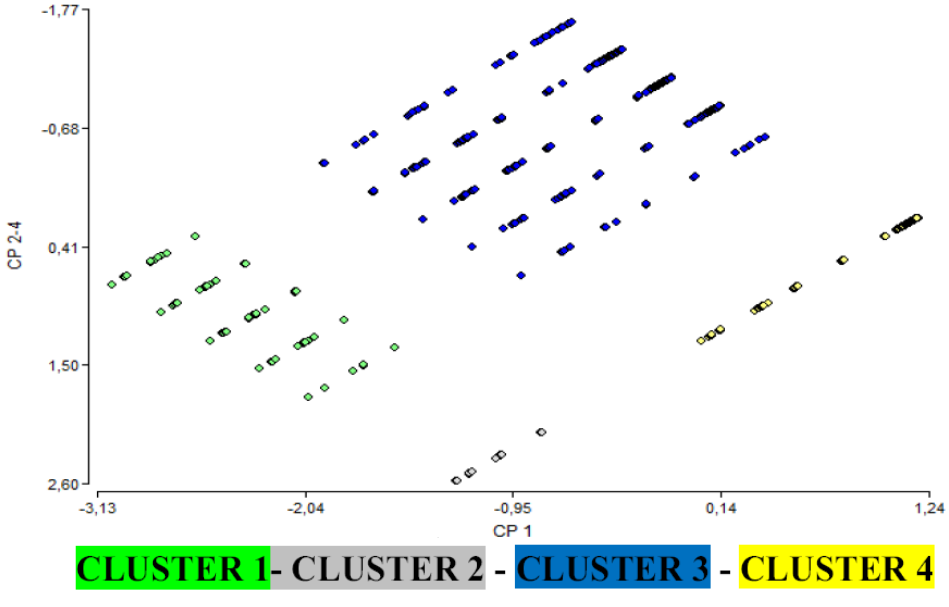
Cluster 1 CP1 y CP2 negatives, profiles with small variation of activities and short timings. It is a precise (no exploration of the solution) and efficient solver.

Cluster 2 Both positive, profile with big activity variation and longer timings. It is an explorer and less systemic solver.

Cluster 3 CP1 negative with high variations. CP2 positive, profile of an explorer and systemic with high efficiency.

Cluster 4 CP1 negative with less variation. CP2 positive and much higher value. Profile of an explorer, systemic but with low efficiency.

Figure 17 Dispersion diagram with clusters (see online version for colours)



It is highly probable that efficiency may be due to background knowledge that would indicate that cluster 1 differs from cluster 3 by the degree of previous knowledge. But they have a similar approach for learning.

Cluster 2 differs from cluster 4 in that the approach for learning is less systemic, that would indicate the trend to not organise knowledge and presumably not to detect in advance generalities.

The topic immersion is more intuitive, with less previous structuration of concepts: it is learning from details to generality. It is the opposite of cluster 4.

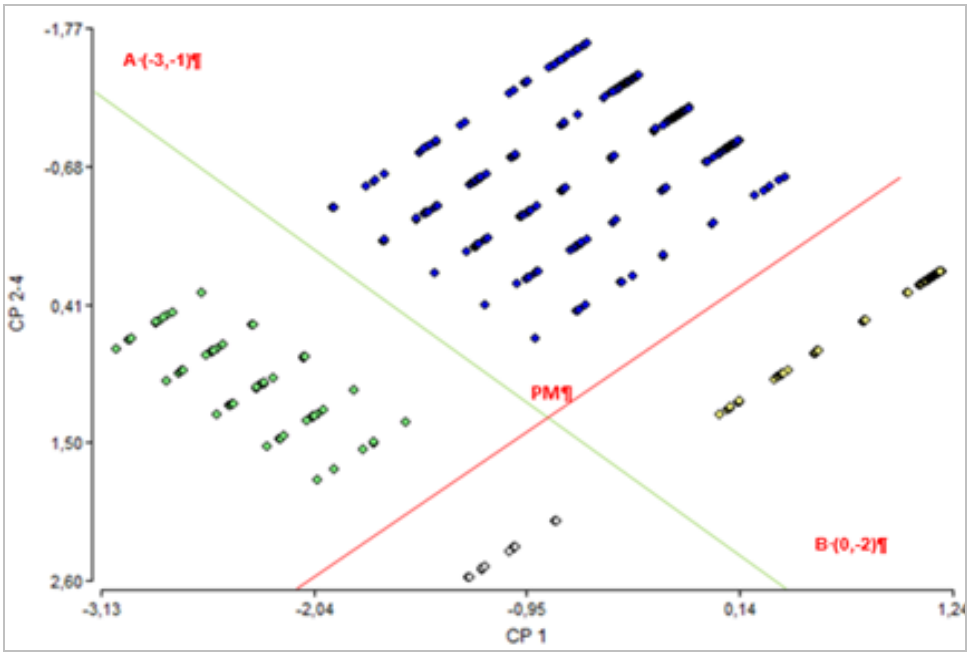
It is expected that a user from cluster 2 does not like to read definitions and global concepts before testing them by itself. The opposite happens with clusters 3 and 4. Figure 18 shows a simple linear separation for the main profiles found.

Lines can be draw for separating the four groups: Let A (x1, y1) and B (x2, y2) be two points on a line. Based on these two known points on a line, it is possible to determine its equation.

For that, a third point R(x, y), also belonging to the line, is taken.

Since A, B, R belongs to the same straight line, we have that PA and PR must have the same slope.

Figure 18 Straight lines separating clusters (see online version for colours)



8 Predictions

8.1 Bayes net

It is a probabilistic model that represents a set of variables and its conditional dependencies as a cyclic oriented graph (Witten et al., 2006). Weka Software (Witten et al., 2006) has been used to perform a Bayes network analysis over the log file. Data must contain non-numeric values. Values of the column CLUSTER have been changed replacing the cluster number with the corresponding descriptive text.

Once the file has been imported in Weka, Bayes classifiers must be selected.

Weka generates a Bayes graph. Selecting any node, it displays a probability distribution table, as shown in Figure 21.

To evaluate all the alternatives, different Bayes classifiers have been selected and analysed. The result of the comparison is shown in Table 12.

As a summary, Naïve-Bayes-multinomial is the most suitable Bayes classifier for the log, since it is the one with highest percentage of positive classifications, with the lowest absolute average error from the three classifiers.

Figure 19 Log imported in Weka (see online version for colours)

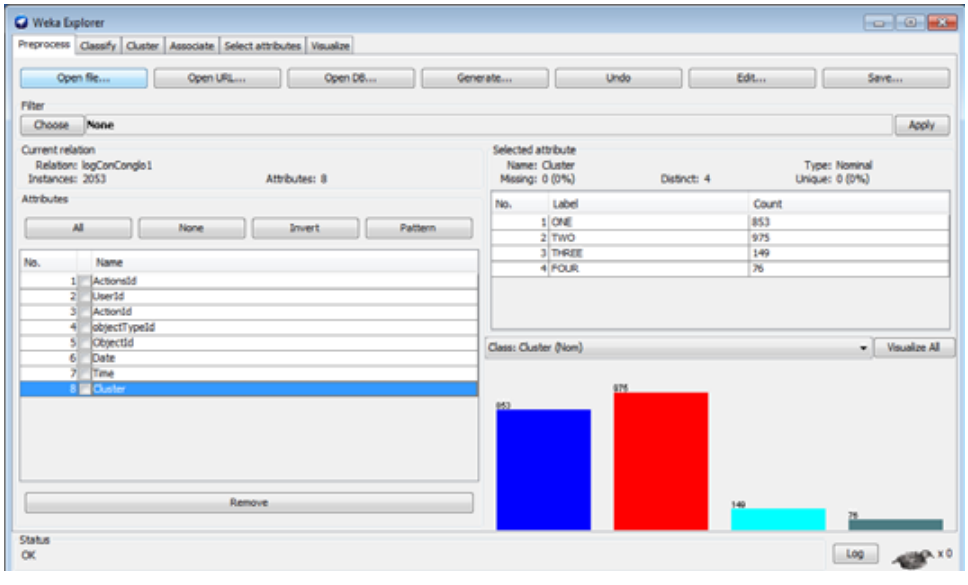


Figure 20 Different Bayes classifiers (see online version for colours)

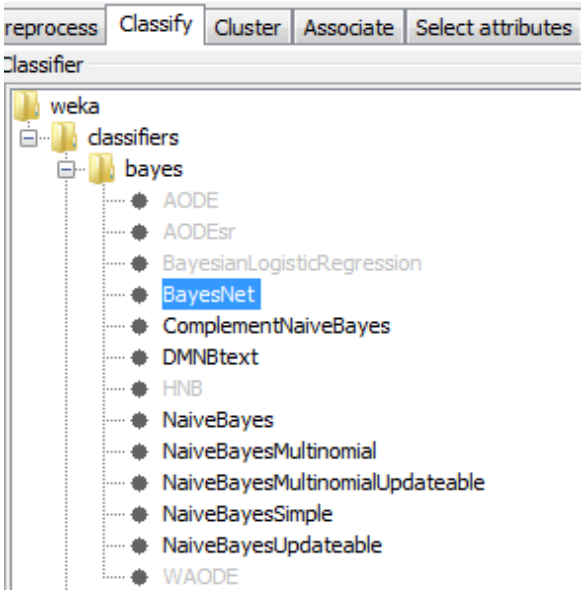


Figure 21 Graph with probability distribution generated with Weka (see online version for colours)

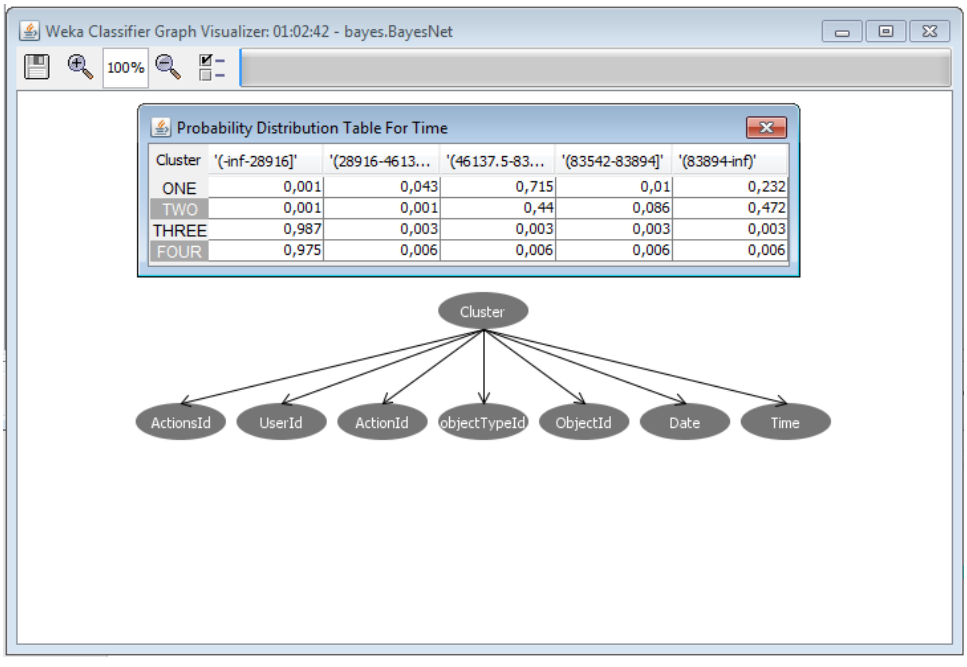


Table 12 Comparison of different techniques of Bayes networks analysis on the log

	Complement naïve Bayes	*Naïve Bayes multinomial*	*DMNBTEXT*
Correctly classified instances	44.9586	59.2304 %	47.4915 %
Incorrectly classified instances	55.0414 %	40.7696 %	52.5085 %
Kappa statistic	0.1728	0.3074	0
Mean absolute error	0.2752	0.2038	0.3244
Root mean squared error	0.5246	0.4515	0.3917
Relative absolute error	92.4308 %	68.4606 %	108.9465 %
Root relative squared error	135.9981 %	117.0378 %	101.5571 %

9 Conclusions and future work

The MIDA project was presented as a basis for the study of learning profiles. While they can be dynamic, it is possible to generate a classification model based on overall behaviour within the virtual museum. Four profiles were found, reflecting main trends to behave when learning. The analysis performed shows a clear difference between them. Taking that result as an initial model, it is possible to perform a lightweight classification based on logs. Thus the museum's organisation and contents can be dynamically adapted to the user.

As a future work, the statistical analysis should be implemented by software automatically and continue with this line of research to a greater level of players.

References

Charsky, D. (2010) 'From edutainment to serious games: a change in the use of game characteristics', *Games and Culture*, Vol. 5, No. 2, pp.177–198.

de Luise, D.L., Gelvez, J., Borromeo, N., Maguet, L. and Dima, L. (2016) 'Multimedia as a tool for leaning engineering', in Balas, V., Jain, L.C. and Kovačević, B. (Eds.): *Soft Computing Applications: Advances in Intelligent Systems and Computing*, Vol. 356, Springer, Cham

Di Rienzo, J.A., Casanoves, F., Balzarini, M.G., Gonzalez, L., Tablada, M. and Robledo, C.W. (2015) *InfoStat versión*, Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina.

Domínguez, F.I.R. and Antequera, J.G. (2012) 'Qué aprendo con videojuegos? Una perspectiva de meta-aprendizaje del video jugador', *RED, Revista de Educación a Distancia*, Vol. 33, No. 15, p.10.

González, L.E. (2006) *Repitencia y deserción en América Latina, World Wide Web* [online] http://www.intercontacto.com/gente.

Kiili, K. (2005) 'Digital game-based learning: towards an experiential gaming model', *The Internet and Higher Education*, Vol. 8, No. 1, pp.13–24.

Marcano, B. (2006) 'Estimulación emocional de los videojuegos: efectos en el aprendizaje', *Education in the Knowledge Society (EKS)*, Vol. 7, No. 2, p.8.

McCauley, K. (2011) *Gaming Impacts Student Learning*, Doctoral Dissertation, University of Central Missouri.

Minitab, I. (2014) *MINITAB Release 17: Statistical Software for Windows*, MinitabInc, USA.

Mitchell, A. and Savill-Smith, C. (2004) *The Use of Computer and Videogames for Learning*, Learning and Skills Development Agency.

Pérez, S.N., Giuliano, M., Sacerdoti, A. and Spositto, O. (2013) 'Abandono y egresos en las carreras de ingeniería de la Universidad nacional de la matanza', in *Tercera Conferencia Latinoamericana sobre el Abandono en la Educación Superior*, México, DF Obtenido el, Vol. 1, No. 6, p.15.

Peris, F.J.S. (2008) 'Videjuegos: Una Herramienta En El Proceso Educativo Del" Homo Digitalis', *Educación y Cultura en la Sociedad de la Información*, Vol. 9, No. 3, pp.4–10.

Shilling, R., Zyda, M. and Wardynski, E.C. (2002) 'Introducing emotion into military simulation and video game design America's army: operations and VIRTE', in *GAME-ON*, November.

Steinkuehler, C. and Duncan, S. (2008) 'Scientific habits of mind in virtual worlds', *Journal of Science Education and Technology*, Vol. 17, No. 6, pp.530–543.

Tomás, A.A. (2009) *Medios audiovisuales en el aula: Pedagogía de los medios audiovisuales*, No. 19.

Ulate, S.O. (2002) *The Impact of Emotional Arousal on Learning in Virtual Environments*, Naval Postgraduate School, Monterey, CA.

Witten, I., Frank, E., Trigg, L., Holmes, M. and Cunningham, S. (2006) *Weka*, Version 3.6, [software].