Assessing data analysis performance in research contexts: an experiment on accuracy, efficiency, productivity and researchers' satisfaction

Patricia Martin-Rodilla[1], Jose Ignacio Panach[2], Cesar Gonzalez-Perez[1] and Oscar Pastor[3]

[1]*Institute of Heritage Sciences. Spanish National Research Council. Santiago de Compostela, Spain.*
[patricia.martin-rodilla, cesar.gonzalez-perez]@incipit.csic.es

[2]*Escola Tècnica Superior d'Enginyeria, Departament d'Informàtica, Universitat de València. Valencia, Spain*
joigpana@uv.es

[3]*Centro de Investigación en Métodos de Producción de software, Universitat Politècnica de València. Valencia, Spain*
opastor@pros.upv.es

ABSTRACT

Any knowledge generation process involves raw data comprehension, evaluation and inferential reasoning. These practices, common to different disciplines, are known as data analysis, and represent the most important set of activities in research contexts. Researchers use software methods and tools for generating new knowledge in their daily data analysis. In recent years, data analysis software has been incorporating explicit references in modelling of cognitive processes, in order to improve the assistance offered in data analysis tasks. However, data analysis software commercial suites are still resisting this inclusion, and there is little empirical work done in knowing more about how cognitive aspects inclusion in software helps researchers in analyzing data.

In this paper, we evaluate the impact produced by the explicit inclusion of cognitive processes in the assistance logic of software tools design and development. We conducted an empirical experiment comparing data analysis performance using traditional software versus data analysis performance using software-assistance tools which incorporate cognitive processes in their design. The experiment is designed in terms of accuracy, efficiency, productivity and user satisfaction during the data analysis made by researchers. It allowed us to find some clear benefits of the cognitive inclusion in the software designed for research contexts, with statistically significant differences in terms of accuracy, productivity and researcher's satisfaction in support of this explicit inclusion, although some efficiency weaknesses are detected. We also discuss the implications of these results for the priority of cognitive inclusion in the software tools design for research contexts data analysis.

Keywords: data-analysis; software-assistance; data-analysis measurement; data-analysis performance, cognitive processes;

1. INTRODUCTION

The set of practices performed by humans in order to comprehend, interiorize and interpret raw data and generate new knowledge based on this data is commonly known as data analysis. Due to the amplitude of the term, there is no a specific consensus about the definition of data analysis, although all definitions include references to the data analysis goals and the set of processes included in a data analysis performance. For instance, we could explain data analysis as "*the process of evaluating data using analytical and logical reasoning to examine each component of the data provided. (...) Data from various sources is gathered, reviewed, and then analyzed to form some sort of finding or conclusion*." [1]

It is also common to refer the entire corpus of the data analysis studies as Data Analytics (hereinafter DA), as "the science of examining raw data with the purpose of drawing conclusions about that information." [2] Note that, for the scope of this paper and as other authors proposed [2], data mining techniques are not included as DA practices due to the purpose and goal of the analysis. While data mining contains techniques to sort through huge data sets using software to discover data patterns and to establish hidden relationships, data analytics focuses on inference. Inference is a derivation process from raw data to conclusions, based entirely on what is already known by the researcher.

The data analysis process has been studied as part of DA from a great variety of disciplines, from cognitive psychology [3, 4] to decision support systems [5]. In some of these last studies, DA is divided into exploratory data analysis (EDA), where new features in the data are discovered, and confirmatory data analysis (CDA), where existing hypotheses are proven true or false [2]. Thus, a data analysis brings together physical and cognitive human practices. Physical tasks are those performed by the human in a tangible way to analyze the data in first place. These tasks could include or not the use of software tools (e.g. sketching in a paper, creating a chart, searching specific concepts inside the data, etc.). Cognitive tasks refer to cognitive processes performed by the human reasoning in order to achieve some conclusions based on the data, such as causal or spatio-temporal reasoning.

Due to the generalization of software tools in DA practices, some sub-disciplines of DA software have emerged in order to apply the information systems corpus to DA contexts. Some examples are grid computing (using several software infrastructures for helping in data analysis processes) or information visualization (improving the ways of visualizing the data for a better analysis of them). Inside these areas, the previously called "physical" practices are extensively studied from the information systems and software engineering point of view. These studies includes tasks characterization [6], empirical studies with users to observe data analysis workflows [7-9] and analysis of usability and performance measurement of software tools created for researchers [10-12]. However, the inclusion of the cognitive dimension in these studies is still residual despite its relevance in research contexts, whose main target is to generate new knowledge based on raw data and evidences through inferences. In recent years, several trends have called for the formal inclusion of the cognitive dimension in software engineering [13-15]. Following these trends, previous studies have included cognitive processes to elicit, design, develop and evaluate software to assist in cognitive practices of data analysis in research contexts. These studies focus mainly on the biomedical domain [16, 17], although there are examples in industrial [18], legal [19] or humanities research environments [20]. Does this inclusion represent a real improvement in the performance of the data analysis through software in research contexts? Is this kind of software improving the data analysis results in comparison with standard software not designed under cognitive processes parameters?

In this paper, we try to give an initial response to these questions. Hence, the main contribution is the empirical evaluation carried out in order to compare the data analysis performance obtained in two different situations. On the one hand, we measure the data analysis performance obtained using DA standard software tools (without cognitive processes assistance). On the other hand, we compare it with the data analysis performed using software-assistance tools (following a data analysis software method with user assistance governed by cognitive processes). The data analysis performance is measured in terms of four different variables: accuracy, efficiency, productivity and user's satisfaction.

As we explained before, the aim of the experiment is to measure the real impact of including cognitive processes in the logic of assistance of software tools design and development in research contexts. Thus, we carried out an experiment with a repeated measures design with 16 subjects, all of them researchers. As a result, we obtained statistically significant differences in terms of accuracy, productivity and researcher's satisfaction in support of data analysis performance using tools with the cognitive explicit modelling inclusion. However, efficiency results show an increase of the researcher's time consumption in data analysis tasks using DA tools with cognitive inclusion.

These results have allowed us to assess the need and priority level of this cognitive modelling inclusion in the design of software tools for data analysis. The design of these tools often involves lengthy, costly and complex processes due to the need of specific information about scientist's practices [9], among other reasons.

This paper is organized as follows. Section 2 presents related work on the evaluation of data analysis performance on research contexts. Section 3 describes the traditional software data analysis practices in research context, which we have called here "autonomous seeking". This section also includes our approach of assisting researchers in data-analysis tasks by including cognitive processes in the software definition and design. Section 4 presents details of the definition, planning and execution of the experiment conducted in order to compare both approaches. In section 4 we also discuss threats to validity of the experiment. Section 5 explains all the analysis performed on evaluated subjects, and finally Section 6 presents conclusions and future work.

2. RELATED WORK

We have previously reviewed in the introduction section what is data analysis, what kind of software is designed and used for data analysis performance in research contexts, and recent trends that has included the cognitive dimension in the design of DA software for this research context. Then, we analyze how the data analysis performance is evaluated in the existing literature, as well as what are the most common metrics used for this purpose.

As we briefly described in the introduction section, DA is a broad and vast concept, and there are not general frameworks or universal metrics that are able to measure the data analysis results and performance with software in all scenarios, tools and domains. Regarding research contexts, it is common to find specific aspects of the domain that serves as an ad hoc metric for examining the data analysis performance. Some examples are proteomics analysis results [21], in the biomedical domain, sensor-based workflows result in terrestrial ecology and oceanography [22] or tourism performance and user satisfaction indicators in cultural heritage studies [23].

It is also common to find metrics referring the specific software tool used in the data analysis performance, such as mistakes evaluation in creating, programming and reasoning using spreadsheets [8, 24] or empirical studies about how users analyze data using spreadsheet and databases format [7]. We can also find these tool-based evaluations using more sophisticated commercial suites for DA, as programming environments [25, 26] (e.g. R [27] or Strata [26]) or statistical suites (e.g. SPSS [28]). However, all these evaluations are mainly based on usability metrics of the software tool, ignoring the data analysis methods presented in the background and without any assistance to the user in the cognitive processes that they daily performed.

In order to consider cognitive factors in measuring the data analysis performance, another approach is currently including studies not only about the software tool employed and the domain involved (as in previous works presented) but also about the technological environment (materialized as the DA method or the technological platform that the researchers used in their data analysis). Thus, DA analysis performance could be materialize defining metrics in order to compare results in desktop applications (e.g. case tools [29]), tablet and mobile solutions [30] or cloud environments [31]. Most of these DA analysis performance studies measure only productivity aspects regarding the software environment as representation metric of the cognitive factors involved, and using qualitative usability aspects for functional analysis [32].

In summary, data analysis performance presents an amount of heterogeneous approaches that offer us valuable results about software tool choices in terms of usability or physical workflows performance, and also fragmented results in terms of DA methods and differences in domains of application. However, most of the evaluations found do not consider the cognitive dimension of the data analysis method employed: neither in the DA software tools employed nor in the evaluation experiment design and the interpretation of the DA results. As a result, there is not an evaluation about the impact of the cognitive dimension inclusion in these software tools. For this reason, we design an initial experiment that compares the data analysis performance in research context using a DA traditional method —using commercial software without cognitive dimension design— and a proposed method that includes this cognitive design in the software tool. The experiment considers common metrics in software evaluation (accuracy, efficiency and productivity) and also the level of researchers' satisfaction. In next section we explained in detail the comparison of both methods.

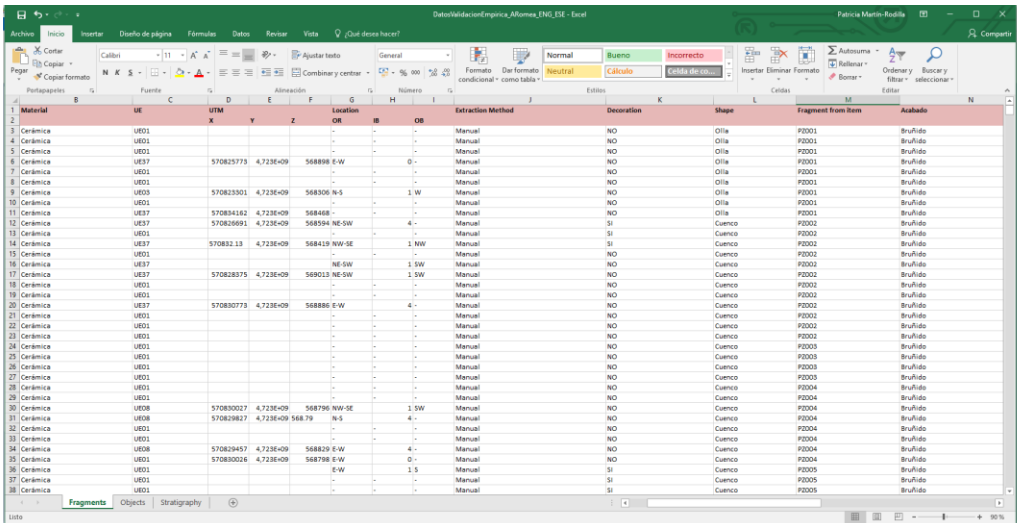
3. DATA ANALYSIS IN RESEARCH CONTEXTS: AUTONOMOUS SEEKING vs. SOFTWARE-ASSISTED DATA ANALYSIS

This section explains in detail two methods to perform data analysis in research contexts, and two software tools selected for performing each of them. Firstly, we explain the main characteristics of data analysis practices referred by researchers using conventional scientific software tools (without cognitive software-assistance). In order to avoid ambiguous terminology, we called this first method "autonomous seeking" throughout this paper. We chose Excel spreadsheet [33] as software tool for performing the autonomous seeking method in the experiment. Secondly, we present a method and subsequently software tool for incorporating cognitive software-assistance for data analysis in research contexts, following current IS models [34], developed in previous works [35, 36].We explain the main characteristics of both (the method and the tool) in order to clarify the second method studied, called "software-assisted data analysis" throughout this paper.

3.1 Autonomous seeking: characteristics and software tools

The final aim of any data analysis presents an inferential and conclusive component: going from raw data to inferential conclusions in the domain of application. In research contexts, these conclusions are new knowledge generated, that subsequently serve as a basis for future research in the domain involved. There are several studies about how researchers generate knowledge, especially in terms of workflows (see most of them in [37]). All these studies show common physical tasks performed by researchers (we previously explained how we employed the "physical" concept through this article), such as data acquisition, sample classification, data storage or data tabular disposition or sketching and visualization. There are similarities in the tasks involved in the workflows among disciplines [22] (regardless of whether they use software in performing the DA physical tasks), such as biomedicine, ocean science, physics or industrial process analysis, and also in humanities disciplines, such as education domain [20].

The method denominated here "autonomous seeking" consists in the data analysis performance following these common characteristics of the workflows analyzed. The main characteristics are: 1) The method employed is autonomous, that is, the researchers use software to analyze data, but the software does not present specific software-assistance for the researchers in terms of cognitive aid to reasoning about the data analyzed, 2) The software employed during the DA does not present cognitive processes in their design 3) The data is disposed in a tabular structure (very common in all existing software tools referenced before [26-28]) and 4) The researcher can visualize in this tabular structure the raw data and create basic charts based on them. Fig. 1 shows the original data in Excel spreadsheets employed for performing autonomous seeking method during the experiment.



	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Material	UE	UTM			Location			Extraction Method	Decoration	Shape	Fragment from item	Acabado
2			X	Y	Z	OR	IB	OB					
3	Cerámica	UE01				-	-	-	Manual	NO	Olla	PZ001	Bruñido
4	Cerámica	UE01				-	-	-	Manual	NO	Olla	PZ001	Bruñido
5	Cerámica	UE01				-	-	-	Manual	NO	Olla	PZ001	Bruñido
6	Cerámica	UE37	570825773	4,723E+09	568898	E-W	0	-	Manual	NO	Olla	PZ001	Bruñido
7	Cerámica	UE01				-	-	-	Manual	NO	Olla	PZ001	Bruñido
8	Cerámica	UE01				-	-	-	Manual	NO	Olla	PZ001	Bruñido
9	Cerámica	UE03	570823301	4,723E+09	568306	N-S	1	W	Manual	NO	Olla	PZ001	Bruñido
10	Cerámica	UE01				-	-	-	Manual	NO	Olla	PZ001	Bruñido
11	Cerámica	UE37	570834162	4,723E+09	568468	-	-	-	Manual	NO	Olla	PZ001	Bruñido
12	Cerámica	UE37	570826691	4,723E+09	568594	NE-SW	4	-	Manual	SI	Cuenco	PZ002	Bruñido
13	Cerámica	UE01				-	-	-	Manual	SI	Cuenco	PZ002	Bruñido
14	Cerámica	UE37	570832.13	4,723E+09	568419	NW-SE	1	NW	Manual	SI	Cuenco	PZ002	Bruñido
15	Cerámica	UE01				-	-	-	Manual	NO	Cuenco	PZ002	Bruñido
16	Cerámica	UE37				NE-SW	1	SW	Manual	NO	Cuenco	PZ002	Bruñido
17	Cerámica	UE37	570828375	4,723E+09	569013	NE-SW	1	SW	Manual	NO	Cuenco	PZ002	Bruñido
18	Cerámica	UE01				-	-	-	Manual	NO	Cuenco	PZ002	Bruñido
19	Cerámica	UE01				-	-	-	Manual	NO	Cuenco	PZ002	Bruñido
20	Cerámica	UE37	570830773	4,723E+09	568886	E-W	4	-	Manual	NO	Cuenco	PZ002	Bruñido
21	Cerámica	UE01				-	-	-	Manual	NO	Cuenco	PZ002	Bruñido
22	Cerámica	UE01				-	-	-	Manual	NO	Cuenco	PZ002	Bruñido
23	Cerámica	UE01				-	-	-	Manual	NO	Cuenco	PZ002	Bruñido
24	Cerámica	UE01				-	-	-	Manual	NO	Cuenco	PZ003	Bruñido
25	Cerámica	UE01				-	-	-	Manual	NO	Cuenco	PZ003	Bruñido
26	Cerámica	UE01				-	-	-	Manual	NO	Cuenco	PZ003	Bruñido
27	Cerámica	UE01				-	-	-	Manual	NO	Cuenco	PZ003	Bruñido
28	Cerámica	UE01				-	-	-	Manual	NO	Cuenco	PZ004	Bruñido
29	Cerámica	UE01				-	-	-	Manual	NO	Cuenco	PZ004	Bruñido
30	Cerámica	UE08	570830027	4,723E+09	568796	NW-SE	1	SW	Manual	NO	Cuenco	PZ004	Bruñido
31	Cerámica	UE08	570829827	4,723E+09	568.79	N-S	4	-	Manual	NO	Cuenco	PZ004	Bruñido
32	Cerámica	UE01				-	-	-	Manual	NO	Cuenco	PZ004	Bruñido
33	Cerámica	UE01				-	-	-	Manual	NO	Cuenco	PZ004	Bruñido
34	Cerámica	UE08	570829457	4,723E+09	568829	E-W	4	-	Manual	NO	Cuenco	PZ004	Bruñido
35	Cerámica	UE01	570830026	4,723E+09	568798	E-W	0	-	Manual	NO	Cuenco	PZ004	Bruñido
36	Cerámica	UE01				E-W	1	S	Manual	SI	Cuenco	PZ005	Bruñido
37	Cerámica	UE01				-	-	-	Manual	SI	Cuenco	PZ005	Bruñido
38	Cerámica	UE01				-	-	-	Manual	SI	Cuenco	PZ005	Bruñido

Fig. 1. Original data in spreadsheets employed for performing autonomous seeking method for data analysis.

3.2 Software-assisted data analysis: characteristics and software tools

Information systems have incorporated in recent years the cognitive dimension in the definition and design of software methods and tools. Some examples are cognitive tasks definition [6], studies about cognitive implications in programming and testing tasks [18], the inclusion of cognitive aspects in usability evaluation [12] and, most recently, the definition of cognitive processes inside software design in order to adapt the software behavior in function of them [38, 39]. This last trend is more explicit in software assistance methods and tools. By software-assisted, we understand the ensemble of services offered by a software system to human users in order to help them carry out certain well-defined processes in any area. The term has become popular in certain fields, such as software assistance for textual analysis. Good examples of this are related to plagiarism detection [40] industrial processes [41] and certain software-assisted medical processes, such as the management of hospital discharges and the detection of relevant elements in medical analyses [42, 43]. Throughout this study, in a similar way to other current studies [34], software-assisted processes will be addressed as an ensemble of services offered by a software system to specialists in a particular field in order to assist them in carrying out data analysis tasks which will allow them to generate knowledge in their field.

Following these previous works in software-assistance, which involve cognitive processes as an essential part in the definition and design of software systems, a software method was defined and implemented to assist researchers in data analysis tasks, incorporating cognitive processes in the logic of software assistance offered. The method was offered as a conceptual framework, it was previously tested in cultural heritage domains and published [36]. Here, we summarize the main characteristics of the method and the software tool implemented based on it.

The software-assistance method follows previous IS models for including cognitive processes in the definition and design of this kind of software systems. The main idea is to characterize cognitive processes as part of the definition and design of the software systems, in order to offer some kind of software assistance services in function of the cognitive process assisted. Particularly, we focused on offering information visualization techniques applied to the dataset analyzed according to the cognitive processes defined. This kind of software-assistance, defined in first place by Chen [34], is based on offering to researchers adaptive visualizations to their data, and is one of the most common software-assistance methods. It has also been tested before in a variety of domains, such as biomedicine, legal domain or, in our case, in cultural heritage disciplines. Thus, we took Chen's model for software-assistance as a reference model and we adapt Chen's model according cognitive processes primitives to assist. These cognitive processes characterization is defined inside the model for each particularity of the application domain, although there are coincidences in most common cognitive processes in data analysis tasks, such as causal reasoning or generalization and exemplification mechanisms [34, 44]. We classify all these cognitive processes into four big groups, attending the research goals of our users identified in previous works in cognitive tasks characterization: Group 1) processes of the combination of data values; Group 2) data grouping processes; Group 3) processes of data contextual situation and Group 4) processes of the analysis of the internal structure of the data. For further information about cognitive processes characterizations and the software-assistance method defined, please see [36].

Fig. 2 shows Chen's model (first workflow) and our Chen's adaptation for including the cognitive processes characterization information (second workflow) in order to provide the software assistance to the researchers' main cognitive processes. The second workflow represents a summary of the software-assistance method presented: the user interacts with the data (moving all time from his/her perceptual and cognitive space to the computational space, i.e. creating their own images of the data in his/her mind). Then, he/she produces information and new knowledge based on the visualized data. He/she also can process the data using the software capabilities in the knowledge-based system (e.g. grouping or filtering data) for producing this new information and knowledge about the data. Note that, in our adaptation, the user provides information to the system about what kind of cognitive processes he/she are mainly performing. The system uses this information to re-adapt the data visualization patterns offered.

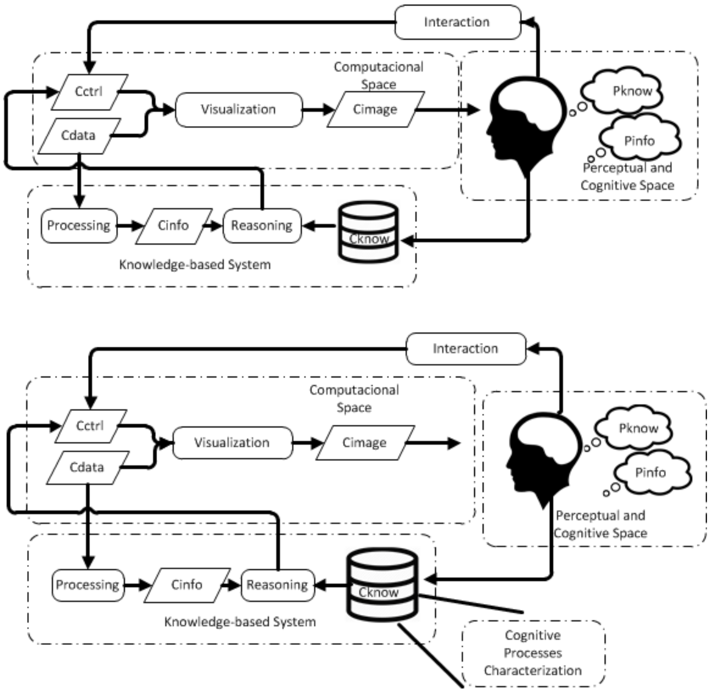


Fig. 2. Software-assisted method for data analysis including cognitive processes characterization.

Using this method, the researcher is interacting with the software system through the selection of a dataset to analyze. Also, he/she could select one or several cognitive processes that they are interested in performing. Then, the software system is able to offer to the researcher adaptive data visualizations according to the data subset and cognitive processes selected. These adaptive visualizations are conceptualized in form of data interaction patterns [45]. For example, for a given dataset and a kind of cognitive processes selected, such as data grouping processes, the method is able to offer to the user clustering visualization of their data in order to help the researcher in performing clustering reasoning. Thus, the explicit inclusion of the cognitive processes in the logic of assistance of the method allow us to identify and offer adaptive visualization patterns in order to help researchers in the data analysis tasks.

Following the presented method, we design and implement a software tool that provides this assistance for cultural heritage research contexts. Next figures present some screenshots of the software tool. Fig. 3 shows specific data interaction patterns for data grouping processes, while Fig. 4 shows specific data interaction patterns for processes of data contextual situation; in this case information about the data spatial location.

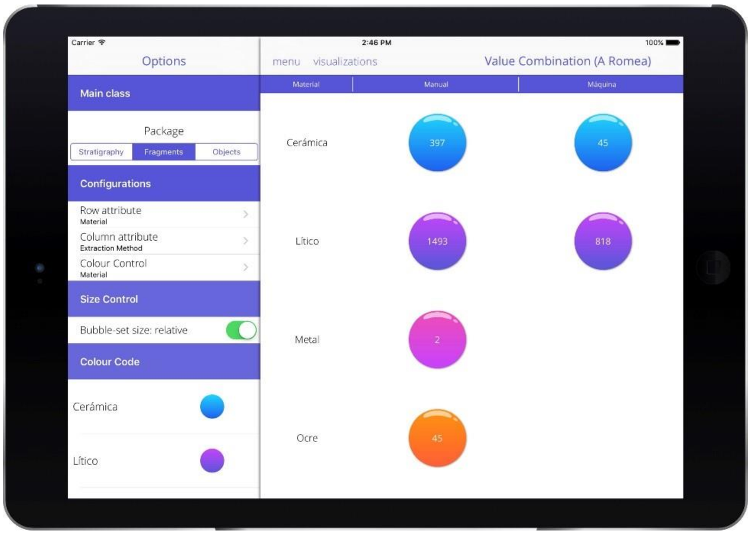


Fig. 3. Screenshot of the tool developed for performing software-assisted method for data analysis. The tool provides, following the software-assisted method defined, clustering bubble-based interaction patterns for grouping cognitive processes.

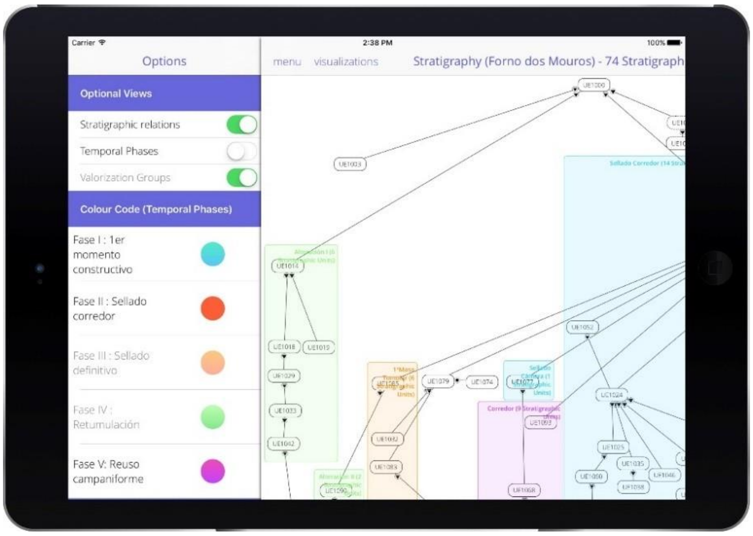


Fig. 4. Screenshot of the tool developed for performing software-assisted method for data analysis. The tool provides, following the software-assisted method defined, spatial interaction patterns for data contextual cognitive processes.

The software tool presented here explicitly includes cognitive processes in software system design assistance (following the software-assisted method). Once we have presented both methods to compare, in next section we detail the experimental design in order to measure the impact of this cognitive software-assistance in data analysis tasks in research contexts.

4. EXPERIMENT DEFINITION, PLANNING AND EXECUTION

Due to its widespread use in designing empirical experiments and studies in information systems and software engineering [46], we choose Wohlin's framework [47] for this experiment design which concerns us here. According to Wohlin's reference framework, we shall define below each of the stages of the experimentation process followed.

4.1 Scoping: Definition of the Experiment Scope and Objectives

The objective of the experiment in Wohlin's terms can be defined following the template of Basili [48] as follows: *"To compare the proposed software-assisted data analysis method (based on the inclusion of cognitive processes modelling) with the traditional method employed in the data analysis process (called here autonomous seeking), with the aim of evaluating the* ***data analysis performance*** *from the point of view of researchers in the context of public and private research institutions."*

4.2 Planning

4.2.1 Context

The experiment is set in a specific, though broad, context as the subjects are all researchers belonging to public and private institutions on a national scale in Spain. Note that, in order to carry out the experiment, it was necessary to select a research scope or area where we can collect real data and develop the case study and experiment design for data analysis. Thus, we select as research scope and case study thematic cultural heritage areas. This selection responds to three motivations: 1) The affiliation and background of some of the authors give us direct access to different profiles of researchers in cultural heritage disciplines, 2) Cultural heritage is a vast domain involving different data analysis needs not covered yet (both analytic and narrative-based), and with some demonstrated needs in terms of cognitive inclusion in software modelling, which constitutes an interesting research area for the experiment and 3) The availability of the software tool implementing the software-assisted data analysis methods previously described, which specifically present cognitive modelling inclusion for cultural heritage research areas. The context, therefore, is considered to be the professional environment of the subjects (cultural heritage research context). Some of the objects are original research materials from existing research projects (such as original spreadsheets); Other objects are created ad hoc for the experiment (such as the implementation of the software-assistance tool that provides interaction patterns assistance), although all of them use real data from case studies in cultural heritage fields.

4.2.2 The Formulation of Hypotheses

There are many characteristics which can be compared between the traditional data analysis method and the proposed software-assistance method. These characteristics include aspects of usability, the study of practices carried out during the analysis itself, the study of decision-making and the role played by the use of one specific method or another, the degree of comprehension and handling, as well as aspects related to integrity, flexibility and portability of both methods and their related technology. However, since the ultimate aim of this experiment is to measure the impact of the cognitive processes inclusion in the software design for data analysis, we have only selected characteristics of interest in this field. More specifically, we deal with accuracy, efficiency and productivity, which are achieved with both methods when carrying out data analysis tasks for the generation of knowledge, involving cognitive processes practices. Furthermore, we believe that studying aspects of user's (in our case researchers) satisfaction is relevant due to the assistance nature of the software-assistance method proposed. The research questions that aim to deal with the cognitive processes inclusions and the hypotheses which arise from them in this experiment are as follows:

RQ1: Does the proposed software-assistance method affect accuracy in tasks related to data analysis? The null hypothesis is:

H01: The accuracy in tasks related to the data analysis using the proposed software-assisted method is similar to the accuracy obtained when carrying out the same tasks with the autonomous seeking method of data analysis.

RQ2: Does the proposed software-assistance method affect efficiency in tasks relating to data analysis? The null hypothesis is:

H02: The efficiency in tasks related to the data analysis using the proposed software-assisted method is similar to the efficiency obtained when carrying out the same tasks with the autonomous seeking method of data analysis.

RQ3: Does the proposed software-assistance method affect the productivity of researchers during data analysis tasks? The null hypothesis is:

H03: The productivity in tasks related to the data analysis using the proposed software-assisted is similar to the productivity obtained when carrying out the same tasks with the autonomous seeking method of data analysis.

RQ4: Does the proposed software-assistance method affect the researcher's satisfaction during data analysis tasks? The null hypothesis is:

H04: The satisfaction expressed by subjects when carrying out tasks related to the data analysis using the proposed software-assisted is similar to the satisfaction they express when carrying out the same tasks with the autonomous seeking method of data analysis.

4.2.3 The Selection of Variables

RESPONSE VARIABLES

The selected response variables emerge from the four characteristics listed above as being especially relevant when comparing the autonomous seeking method of data analysis and the proposed software-assistance method, which includes cognitive processes modelling in the software design. We can define these variables as follows:

- **Accuracy**: we define accuracy as a "quantitative measure of the magnitude of error" [49]. We measure accuracy as the percentage of correct answers once the defined data analysis tasks have been carried out. The tasks are divided into several items, so it is possible to obtain the percentage of accuracy by measuring the percentage of items which have been successfully completed in all the tasks. The total accuracy obtained can be aggregated in two ways: Firstly, aggregated by calculating the average accuracy obtained for each task, paying attention to all their sub-items of the tasks using a Boolean system of accuracy measure (in other words, each task could be correct or incorrect, without intermediate cases). This first aggregation of the accuracy variable is called as "all-nothing accuracy". Secondly, accuracy can also be aggregated by calculating the average accuracy obtained for each task, but taking into account intermediate steps in correctness. Thus, the task could be correct, partially but mainly correct, partially but mainly incorrect and incorrect. This second aggregation of the accuracy variable is called "weighted accuracy". The subject's magnitude of error when carrying out data analysis tasks using both methods is measured.
- **Efficiency**: we define efficiency as "the ability to produce a result with a minimum of extraneous or redundant effort" [50]. We measure efficiency as the response time employed by the subjects in carrying out the defined data analysis tasks. The total efficiency obtained is aggregated by calculating the average efficiency obtained for each task, thus avoiding interference from aspects relating to the type of task at that moment.
- **Productivity**: we define productivity as "the ratio of work product per work effort" [50]. We measure productivity as the ratio between work achieved and resources by the subject using both methods to carry out data analysis tasks. That is, the ratio between each aggregate accuracy achieved (using again the same aggregation criteria than in productivity, with productivity all-nothing corresponding to all-nothing productivity and weighted productivity corresponding to weighted productivity levels) and the response time employed in carrying out the indicated data analysis tasks.
- **Satisfaction**: we define satisfaction as the degree of "positive attitudes towards the use of the product" [51]. As we are working in assistance contexts, we believe it is relevant to evaluate this variable, as assistance can be obstructed due to an inappropriate degree of ease of use. The instrument employed is a 5 Likert-scale questionnaire based on Moody's framework [52], which evaluates satisfaction based on three metrics: Perceived Usefulness (PU), Perceived Ease of Use (PEOU) and Intention to Use (ITU). In accordance with Moody's framework, a questionnaire [53] is defined for each treatment with 22 sentences for evaluation, of which eight evaluate perceived usefulness (PU), nine evaluate ease of use (PEOU) and five evaluate intention to use (ITU). Each questionnaire on treatment refers integrally to all the tasks carried out with the same treatment, following the same structure for testing the rest of the response variables.

FACTOR

The factor is applied on different levels in order to discover its impact. We call our factor "the data analysis method". Note that this term possesses specific semantics at the heart of this experiment: the data analysis method consists of a set of tasks which are performed in order to examine raw data with the aim of extracting conclusions and thus generate new knowledge in research contexts. Generally, these conclusions will support the decision-making process in the field in question and will verify or refute existing models or theories within that field. Therefore, we are not dealing with tasks related to data extraction (which are commonly related to the categorization of data) but with tasks that focus on the inferences emerged from the data.

The "data analysis method" is made up of two levels:

M1: The control level. This is the traditional method of data analysis (called here "autonomous seeking") employed to analyze raw data and to generate new knowledge from them, based on the direct observation of data obtained from the existing literacy and research cases being studied, as we explained in previous sections. The data is generally organized in a table format and commercial software with some graphic capacity, such as spreadsheets, is used. Although these characteristics are broadly repeated in all frameworks and all of this characteristics are implemented in several software tools, such as R [27], Strata [26], SPSS [28] or SAS [54] we have chosen to employ spreadsheets, particularly Excel spreadsheet [33], as software tool to evaluate this DA method. This choice responds to two main reasons: 1) Spreadsheets are readily available commercial software present on any personal computer or research institution, which facilitates software access during the experiment and 2) Spreadsheet are one of the most common software tools employed in cultural heritage fields for performing data analysis. Thus, we use Excel [33], as it is one of the most widespread tools and is commonly used in the institutions to which the subjects in the experiment process belong.

M2: The treatment level. This is the proposed software-assisted data analysis method (explained in detail in section 3.2), which includes cognitive processes modelling in the software design. Our implemented tool is used to perform the analysis. In this method, the cultural heritage data of the cases being studied is described using a data model extension of an specific data model for cultural heritage information, called CHARM [55]. With this tool, data is selected in the form of subsets by the researcher. Depending on the subset selected, the software tool presents this data organized into interaction patterns according to the cognitive process being assisted.

Table 1 shows a summary of the research and hypotheses questions dealt with, the variables and the metrics that will be used to measure them:

Research Question	Hypothesis	Response Variables	Metric
RQ1	**H_{01}**	Accuracy	Percentage of correct answers
RQ2	**H_{02}**	Efficiency	Response time
RQ3	**H_{03}**	Productivity	Accuracy/Efficiency
RQ4	**H_{04}**	Satisfaction	Perceived Usefulness (PU), Perceived Ease of Use (PEOU), Intention to Use (ITU)

Table 1. Research questions dealt with in the experiment, the defined variables and their corresponding metrics.

BLOCKING VARIABLES

The cultural heritage problem dealt with in the experiment (i.e. the case study employed in the data analysis) has been detected as a blocking variable and called P from this point on. In order to prevent this problem from affecting the results of the experiment, its value has been balanced, thus blocking the possible effect. In order to do this, P takes two values, P1 and P2. Therefore, the subjects will carry out the data analysis tasks on two different cultural heritage case studies. Another advantage of using two problems is that the threat of the learning effect is avoided, as what is learnt with the problem of the first treatment is not applied in the second treatment. In the section entitled “Design Principles of the Experiment”, more details are given about how the P variable has been balanced.

4.2.4 The Selection of Subjects

The subjects are all cultural heritage researchers, mainly from heritage sub-disciplines such as History, Archaeology, the History of Art and Architecture. An open call for participation was made via an e-mail list of heritage professionals in Spain or of Spanish nationality. In turn, these professionals were encouraged to share this call with other colleagues. The experiment was carried out with 16 specialists belonging to 7 public (such as the University of Santiago de Compostela, the Institute of Heritage Sciences and the University of Minho) and private (different archaeological companies or the Campo Lameiro Archaeological Park, among others) institutions. The specialists were selected randomly from all those who expressed an interest in collaborating in the experiment. Later, the implications of the size and characteristics of the sample are dealt with in the discussion of the results. In order to characterize the sample in a better way, the subjects completed a demographic questionnaire [53] before starting the experiment, which they did individually. The demographic distribution of the subjects selected is described below.

As far as gender is concerned, Fig. 5 shows that 56% of the subjects were male and 44% were female. This distribution reflects the proportion by gender in cultural heritage fields. For example, the INE’s

(Spanish Statistical Office) 2011 report on University teaching in Spain [56] established that, of 170 university teachers in the area of "Archaeology", 73 were women, thus reflecting a female percentage of 42%, which is similar to the distribution of our sample. Another complementary study was carried out on Human Resources in Science and Technology [57] and examined the system of Science in Spain, analyzing its different areas. The latter report dates from 2009 and shows that 14.70% of doctors work in areas of the Humanities. Of this 14.70%, 6.56% were women, representing a relative percentage of 44%, similar to our own sample. We believe, therefore, that this balanced and representative distribution as far as gender is concerned allows us to interpret the data obtained without offering views differentiated by gender.

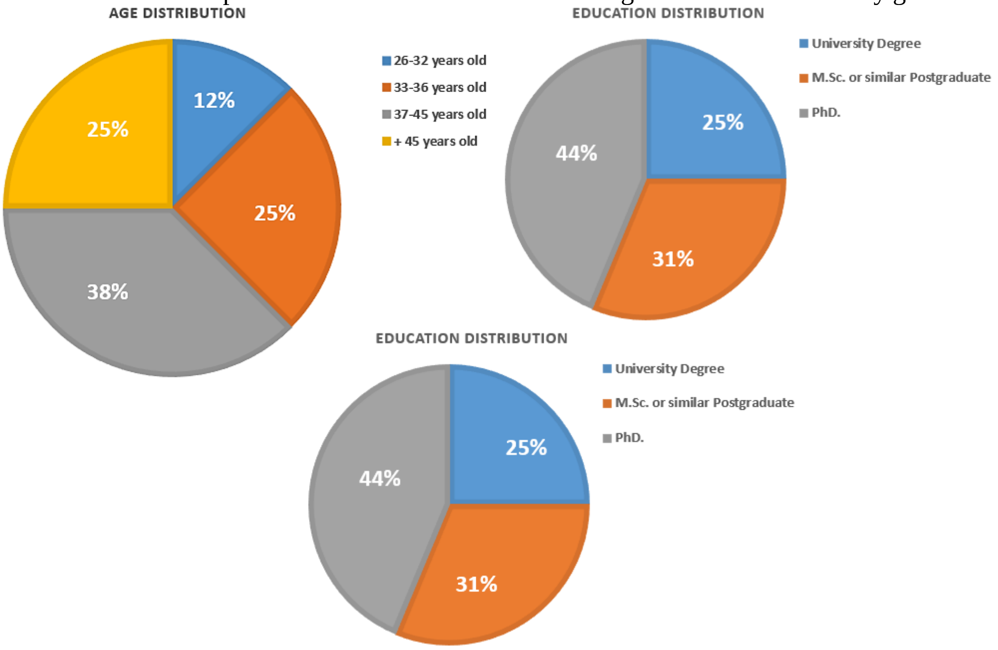


Fig. 5. Percentages of the sample according to distribution by gender, age and level of education distribution.

Regarding the subjects' age, we considered necessary to conduct the experiment with a heterogeneous group, in order to avoid any bias in the sample. Fig. 5 shows the age distribution of the sample. The age ranges indicated have been defined according to the age ranges commonly used in research: the first ranges generally consist of staff in training and the older ranges to research staff with a greater level of experience. It should be noted that age is not an indicator of the level of training of the participating subject in all cases (this aspect has been dealt with as a variable separated from age). However, the age ranges defined are also useful because, in most cases, they also indicate the stage of the subject's research career.

Any researcher in cultural heritage fields will generally have a high level of studies so we do not believe that this determines any of the variables measured in the experiment. However, it is interesting to characterize the sample according to the level of studies in order to illustrate that, given a random sample of our end users, almost half of them holds a doctoral degree —Ph.D.— (see Fig. 5). This allows us to gain an idea of what type of people generate knowledge in cultural heritage fields and the importance of dealing with cognitive processes when it comes to building software assistance in this field.

Therefore, along general lines, we have a sample which is balanced as far as gender is concerned, well-distributed in terms of age and polarized when it comes to the level of education. This last aspect is due to the characteristics of our typical end users: professionals in cultural heritage who work in research contexts and subsequently, their main activity consists of performing DA activities and generating knowledge in the field.

We also wanted to characterize the sample according to the professional sector to which the subjects belong. In a similar way to the level of education, this variable has only been used in order to give a more detailed idea of the type of professionals that compose the sample. Regarding professional sector, 56% of the subjects currently work in the public sector, 25% in private companies and the remaining 19% are self-employed. These percentages present similarities with existing studies in cultural heritage fields [58]

Taking into account the fact that the objective of this experiment is the confirmation of several hypotheses regarding the method of data analysis in two separate case studies, we considered it was necessary to have a heterogeneous sample also in terms of the sub-discipline of the subjects. Fig. 6 shows that 44% of the subjects are archaeologists, due to the fact that the case studies selected fundamentally deal with archaeological data (see next section titled "Experimental Objects"). It must be pointed out that, within this 44%, there are archaeologists specialized in different sub-disciplines and chrono-cultural periods, including experts in Metallurgy, Prehistoric, Roman and/or Medieval Archaeology, as an example of the internal diversity of this subset. Other groups of experts included architects, restorers, museum managers, art historians, communicators and educators in cultural heritage matters.

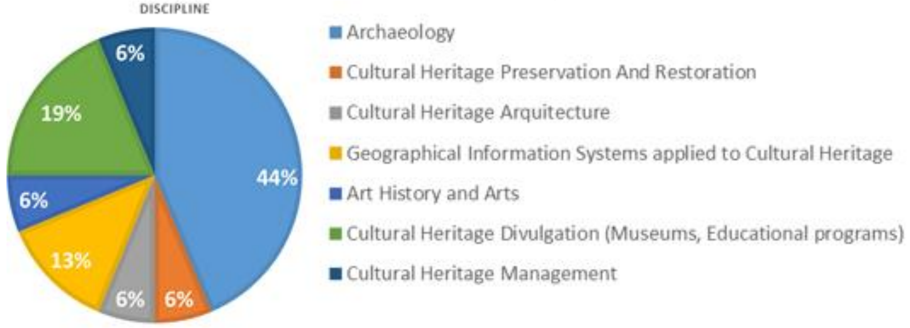


Fig. 6. Percentages of the sample corresponding to disciplines.

Finally, this demographic characterization aims to offer a more detailed perspective of the sample by characterizing specific aspects of the subjects as far as the data analysis and knowledge generation in cultural heritage research carried out over the course of their careers is concerned. This was achieved by asking them questions about how much experience they had in handling, managing, documenting and/or researching cultural heritage data.

The ranges have been defined according to intuitive intervals reflecting the degree of experience of the subject: someone with less than 6 years of experience is considered to be a professional with a low level of experience, increasing in intervals of 4 or 5 years of experience up to 20 years, when it is considered that the individual has acquired professional maturity. Fig. **7** shows the distribution of the sample by intervals of years of experience. It should be noted that our sample is quite heterogeneous, although subjects with more than 14 years of experience are predominant. We believe this confers robustness upon the sample when it comes to evaluating the data analysis methods.

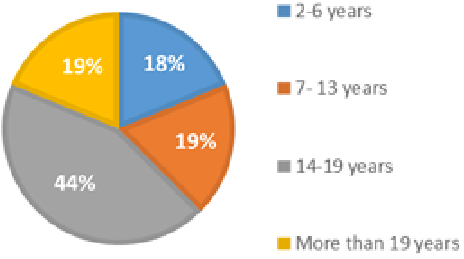


Fig. 7. Percentages of the sample responding to the question “How many years of experience do you have in handling, managing, documenting and/or researching cultural heritage data?

However, it is possible for a professional in the field of cultural heritage to have spent many years handling raw heritage data, but limiting her/his main functions to extraction or characterization. Usually, knowledge generation and in-depth data analysis are performed by other members of the team. In order to avoid this situation, each subject was asked individually about what percentage of his/her working time spend usually on data analysis methods such as those being evaluated in the experiment. Fig. 8 shows the distribution of answers to this question and how (although the distribution is relatively heterogeneous) the majority of subjects have mainly worked throughout their career on data analysis tasks. We believe, therefore, that the sample is robust enough in this aspect to evaluate data analysis methods during the processes of knowledge generation performed in research contexts.

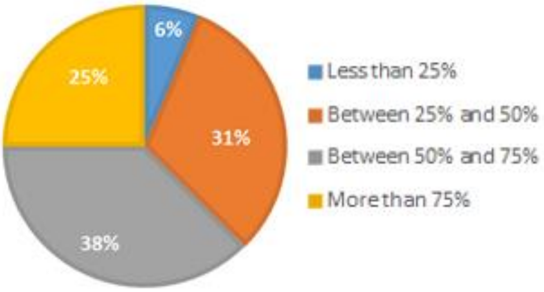


Fig. 8. Percentage of the sample responding to the following question: What percentage of your career would you say has been dedicated to data analysis tasks in cultural heritage?

4.2.5 Design Principles of the Experiment:

Following [46], it can be observed that the most appropriate alternative for experiments involving two treatments (such as that which concerns us) is the **paired design blocked by experimental objects** (the variable P, which represents the heritage problem whose data is analyzed). This design presents the following advantages: (1) It maximizes the number of subjects in the validation, since it does not divide the size of the sample into two; (2) It limits the dependence of the problem selected, as we have two problems (P1 and P2); (3) The learning effect [59] between the two treatments is avoided, given that the subjects apply both treatments in different contexts (the defined problems P1 and P2).

The subjects are divided into two groups (G1 and G2). Both groups used the traditional data method, autonomous seeking (M1) in the first session (S1) and the proposed software-assisted data analysis method (M2) in the second session (S2). The following table shows the applied design:

		P1	P2
Session 1	M_1	G1	G2
Session 2	M_2	G2	G1

Table 2. A summary of the experiment design.

The assignation of the subjects to the groups was carried out randomly, maintaining the same number of subjects in both groups in order to keep them balanced.

4.3 Instrumentation

The instruments employed in the experiment were:

- An initial questionnaire regarding the subjects' profile (gender, age, level of education, discipline and years of professional experience with heritage information).
- A statement with the required data analysis tasks, which were different for problem 1 (P1) and for problem 2 (P2).
- Excel files with the real data employed in both problems (P1 and P2).
- Working software tool implementing the software-assistance proposed method with the real data of both problems (P1 and P2).
- Two satisfaction questionnaires, one for each method (M1 and M2). Both are created based on similar used instruments [46, 59].

All documents used in the validation can be consulted in [53] in the order in which they are listed above.

4.3.1 EXPERIMENTAL OBJECTS

The block variable is operationalized as two separated problems (P1 and P2). Next, we describe both:

Problem 1: A Romea

Problem 1 consists of carrying out a data analysis regarding the historical and archaeological evidence found during the excavation of the archaeological site known as "A Romea" [60]. The case study is located in the excavation and prospection of a metallurgical zone in which objects of different materials, compositions and temporal and functional origins were found. In addition, parts of a barrow were documented via the discovery of structures, attributing temporal phases to the barrow: when it was built, when it was used, when it was abandoned, etc. The data analysis should determine the functional attributions of the archaeological site and the objects found, as well as the temporal phases associated to it based on those objects and structures. In order to do this, there is data available regarding its material composition, morphology, decoration and other aspects of interest for researchers concerning the objects and structures found. Furthermore, radiocarbon dating was available for some of the structures, along with geographic information for each documented element.

Problem 2: Forno dos Mouros

Problem 2 consists of carrying out a data analysis about historical and archaeological evidence found during the excavation of the site known as “Forno dos Mouros”[61]. The case study is located in the excavation of a megalithic structure used for burials, in the style of a dolmen, in which objects and structures are documented. In addition, the dolmen has the remains of paintings inside. The data analysis must determine the temporal attributions of the objects and structures found. In order to do this, data is available regarding the material composition, morphology, decoration and other aspects of interest for researchers concerning the objects, structures and paintings found. Furthermore, geographic information for each documented element is available.

Both problems respond to two real case studies from cultural heritage, mainly related to the sub-area of Archaeology, in accordance with the main sub-area of research of the subjects involved.

4.3.2 DATA ANALYSIS TASKS

We have defined tasks taking into account the research context and previous studies regarding what cognitive processes are carried out in this field [62]. Each task emphasizes the performance of a specific cognitive process group. The definition is similar for both problems, only varying in terms of the structure and content of the information being analyzed in each problem (the answer will vary according to the problem dealt with due to this fact). The statements for the tasks are as follows:

- TASK A: This task concerns processes of the combination of data values in order to find out the distribution and characteristics of the materials found in any heritage study.
    - Statement PROBLEM 1: Indicate and note down the total number of ceramic fragments, lithic and/or other materials that have been found at the site. Later, calculate again the totals of the ceramic fragments, lithic and/or other materials but, this time, also according to the method of extraction used.
    - Statement PROBLEM 2: Indicate and note down the total number of ceramic fragments, lithic or other materials that have been found at the site. Later, it calculates again the totals of the ceramic fragments, lithic or other materials but, this time, also according to whether the fragments are decorated or not.
- TASK B: This task concerns data grouping processes in order to analyze the categories involved on each case study.
    - Statement PROBLEM 1: Indicate and note down the percentage of average fragmentation presented in the ceramic objects extracted mechanically.
    - Statement PROBLEM 2: Indicate and note down the percentage of average fragmentation presented in the ceramic objects extracted manually.
- TASK C: This task concerns processes of data contextual situation in order to discover the temporal characteristics of the objects and/or material structures found in any heritage study.
    - Statement PROBLEM 1: Indicate and note down the different chronological attributes that you believe have been associated with the objects found in the site.
    - Statement PROBLEM 2: Indicate and notes down the different chronological attributes that have been associated with the ceramic fragments found in the site. Then, reason about the cultural attributes of the complete ceramic pieces: Are you able to associate a chrono-cultural attribute to each ceramic object? Indicate the ones associated to pieces PZ01 and PZ06.
- TASK D: This task concerns processes of the combination of data values in order to discover functional aspects of the materials and structures found in any heritage study.
    - Statement PROBLEM 1: Indicate how many stratigraphic units have been defined and according to which criteria the groups of these stratigraphic units have been created. Then, indicate which units make up the “Barrow Chamber” group and the “Alteration Path” group.
    - Statement PROBLEM 2: Indicate how many stratigraphic units have been defined and according to which criteria the groups of these stratigraphic units have been created. Then, indicate which units make up the “First megalithic zone” group and the “Chamber access: pit” group.

- TASK E: This task concerns processes of the analysis of the internal structure of the data.
    - Statement PROBLEM 1: Indicate which attributes (relevant characteristics) of the objects found have been documented.
    - Statement PROBLEM 2: Indicate which attributes (relevant characteristics) of the stratigraphic units found have been documented.
- TASK F: This task is related to processes of data contextual situation in order to discover the temporal characteristics of the objects and/or material structures found in any heritage study.
    - Statement PROBLEM 1: Indicate and note down which temporal intervals (chronological assignments) have been attributed to the A Romea barrow in a global manner, paying attention to all the information you are given regarding the stratigraphy of the materials found.
    - Statement PROBLEM 2: Indicate and note down which temporal intervals (chronological assignments) have been attributed to the Forno dos Mouros barrow in a global manner, paying attention to all the information you are given regarding the stratigraphy of the materials found.

4.4 Operation

Preparation:

The subjects in the experiment did not know the case studies being used nor were familiar with the data. Minimal information about data analysis methods was given to the subjects, though they were not offered any prior training or informed of the hypotheses being dealt with in the experiment. By offering themselves as volunteers for the validation, they gave their consent to these aspects. The necessary materials referred to in the earlier section entitled "Instrumentation" were prepared in advance.

Execution: The Development of the Experiment

As far as the lines of action in carrying out the experiment are concerned, the subjects were given some brief instructions about the dynamics and they were allowed to ask some initial questions. No previous training of the subjects was considered necessary as far as the methods to be used were concerned: the traditional method of data analysis (autonomous seeking) was well known by all the subjects, whereas the method of analysis using the proposed software-assisted data analysis was not known beforehand. This difference is assumed within the experiment in order to find out how intuitive the proposed method is, although we are aware that some prior training in the latter method could affect (hopefully in a positive way) the results offered by the subjects with this method. However, our interest in finding out how intuitive the proposed method is, along with recommendations made in empirical studies in Software Engineering [47], led us to take an execution decision: not to offer any prior training in either of the two methods, assuming differences in subject's familiarity with both methods and subject's expertise with them. The dynamics of the experiment were divided into two sessions (S1 and S2), as can be seen in Table 2. Each session consisted of specific phases:

1. In the first phase, the subjects filled in the demographic questionnaire, which was the same for everyone. Then, the subjects were divided into two groups randomly. This phase was only carried out in the first session of the experiment.
2. In the second phase, the subjects performed tasks A, B, C, D, E and F applying the treatment assigned in each case to the problem selected, depending on whether it was session S1 or session S2, according to the design shown in Table 2.
3. This phase was carried out in both sessions. Once the tasks were completed, each user filled in the corresponding satisfaction questionnaire, independently of the group to which they belonged.

The full dynamics of the experiment can be seen in Fig. 9.

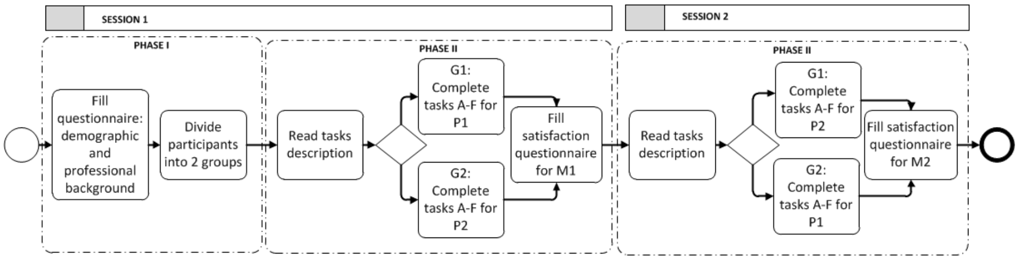

SESSION 1
SESSION 2
PHASE I
PHASE II
PHASE II
Fill questionnaire: demographic and professional background
Divide participants into 2 groups
Read tasks description
G1: Complete tasks A-F for P1
G2: Complete tasks A-F for P2
Fill satisfaction questionnaire for M1
Read tasks description
G1: Complete tasks A-F for P2
G2: Complete tasks A-F for P1
Fill satisfaction questionnaire for M2

Fig. 9. The development dynamics of the experiment, indicating the phases corresponding to each session. The rectangles represent tasks, whereas the diamonds indicate the beginning of tasks performed in parallel.

Validation of the data and validity evaluation:

No invalid data was detected. This is due to the individuality and supervision of the process of executing the study. Due to the small number of existing works in evaluating DA in research contexts, and more specifically, the impact of cognitive processes modelling on these DA software tools, it was considered necessary to carry out an in-depth analysis of all the threats to its validity. Thus, the threats identified in the literature in the field of Software Engineering (Wohlin [47], based on Cook & Campbell [63]) have been analyzed one by one in an attempt to identify those which may influence the experiment, describing how they have been mitigated (see from Table 3 to Table 6)

CONCLUSION VALIDITY

Threat	**Reason**	**State**	**Treatment of the threat**
Low statistical power	The size of the sample is not big enough.	Avoided	We avoid the threat by maximizing the number of subjects with repeated measures.
Subjects of random heterogeneity	The subjects have not been selected at random and their profile is heterogeneous.	Avoided	The subjects are selected randomly among all the volunteers. The profile is heterogeneous, albeit with a common feature: all of them are specialists in cultural heritage fields.
Fishing	The conductors of the empirical validation seek a specific result.	Avoided	We avoid this threat by processing all the raw data without filtering.
Reliability of measures	There is no guarantee that the results are the same when measuring the phenomenon again.	Partly suffered	The metrics for accuracy, efficiency, productivity are objective. The metrics for satisfaction are subjective, so they suffer this threat.
Reliability of the implementation of the treatment	There is the risk that the application is not similar among the different people requesting the treatment or on different occasions.	Partly suffered	The implementation must be as standard as possible among the different subjects and occasions. We mitigate the threat of different occasions by controlling the sessions, and of different treatments by homogenizing the possibilities of the tools used in the two treatments.
Random irrelevancies in experimental settings	There are external elements to the empirical validation which may interfere in it.	Suffered	We cannot guarantee that the subject does not carry out any other task while performing the validation. We mitigate the threat by actively supervising the sessions.

Table 3. Analysis performed about conclusion validity threats.

CONSTRUCT VALIDITY

Threat	Reason	State	Treatment of the threat
Interaction of testing and treatment	The subjects apply the metrics to the treatments	Avoided	The experimenters apply the metrics.
Mono-method bias	The empirical validation with only one type of measure may present bias.	Partly suffered	The variables regarding accuracy and satisfaction present more than one metric, which avoids the threat. The variables regarding efficiency and productivity suffer the threat, which we minimize automatizing the time measurement.
Hypothesis guessing	The subjects sense the purpose of the empirical validation and act in consequence.	Suffered	We minimize the threat by not commenting on the objectives, research questions and defined metrics with the subjects.
Evaluation apprehension	The subjects are apprehensive about being evaluated.	Avoided	We avoid this threat by not commenting on the evaluative nature of the experiment with the subjects and including the tasks in a research session in order to find out about their working methods.
Interaction of different treatments	The result may be caused by the combination of treatments applied.	Suffered	Due to the fact that the traditional method is always applied first, the validation suffers this threat. We cannot ensure that the order of application does not affect the results.
Mono-operation bias	An operationalization of the treatments based on just one method may introduce bias.	Suffered	Since the data analysis methods are analyzed using specific tools (Excel and the implemented software tool for the proposed software-assisted data analysis method), the validation can be affected. It may be dangerous to generalize the results to other traditional commercial software. As far as our proposed tool is concerned, it must be validated as a data analysis method.
Instrument dependency	The instruments used or created for the experiment are not enough teste for their internal validity and reliability	Avoided	We used instruments based on previous designs with enough validation in software engineering literature [46, 47, 59]. However, there is a minimum threat regarding the similarity between our instruments and the followed models.

Table 4. Analysis performed about construct validity threats.

INTERNAL VALIDITY

Threat	Reason	State	Treatment of the threat
History	The different treatments are applied to the same object with a significant time difference.	Avoided	We avoid this threat by minimizing the time between sessions and by maintaining communication with the subjects during this time.
Learning of objects	The subjects may acquire knowledge from the first treatment and apply it to the second.	Avoided	We avoid this threat by using two different problems that do not allow the subjects to learn aspects of the object.

Threat	Reason	State	Treatment of the threat
Subject motivation	Less motivated subjects may present worse results than more motivated ones.	Suffered	We mitigate the threat by using only volunteer subjects whose motivation in the validation is high.
Maturation	The subjects react differently as time passes.	Avoided	We avoid this threat by delimiting the experiment to 90 minutes by session, in order to avoid fatigue effects in subjects.
Selection	The results are affected by how the subjects are selected.	Partly suffered	The subjects present a specific profile: researchers. We mitigate the threat by maximizing the heterogeneity of the research disciplines and other sample aspects inside the profile, as well as recruiting subjects as volunteers in the validation.
Resentful demoralisation	The subjects only apply one treatment.	Not applicable	The subjects apply both treatments.
Mortality	The subjects may abandon the validation before the end.	Suffered	Due to the fact that the subjects are volunteers and that two sessions take place; the subjects may abandon the validation. We mitigate this threat by minimizing the time between sessions and maintaining communication with the subjects between sessions.
Compensatory rivalry	The subjects who apply only the less desired treatment may influence the results.	Not applicable	The subjects apply both treatments.

Table 5. Analysis performed about internal validity threats.

EXTERNAL VALIDITY

Threat	Reason	State	Treatment of the threat
Interaction of selection and treatment	The subjects do not represent the general population at which the validation is aimed.	Partly suffered	The subjects have different profiles but one common feature: all of them are specialists in cultural heritage fields. We believe that this fact mitigates the threat in terms of generalisation, although the results are only valid for similar profiles in research contexts.
Object dependency	The results depend on the objects used and cannot be generalised.	Suffered	We suffered the threat because the objects chosen are related with specific software tools (one per method). We minimize the threat using two objects for each treatment (although the software tool dependency is still affecting).
Interaction of history and treatment	The treatments are applied on different days: the circumstances of the moment may affect them.	Suffered	The threat is suffered due to the existence of two sessions. We minimize the threat by applying both treatments in the same room and at the same time during each session.
Interaction of setting and treatment	The elements used in the validation are obsolete.	Not applicable	The questionnaires are published and in use.

Table 6. Analysis performed about external validity threats.

5. ANALYSIS AND INTERPRETATION

Other studies carried out in Software Engineering with a paired design blocked by experimental objects [46, 59] were taken into account for the analysis and interpretation of the results. These studies generally use the general linear model with repeated measures (thereby maximizing the number of responses), known as GLM [64], in order to analyze the data obtained during the empirical validation, in which both levels for the factors (method M1 and method M2) are applied to each subject. However, prior studies [59] conclude that the existence of blocking variables (in our case, the variable P, corresponding to the cultural heritage problem analyzed) requires a greater treatment than that offered by the GLM model. Thus, the use of a **Mixed** statistical model with the type of covariance for "unstructured" [59] repeated measures is recommended.

Therefore, we define this model for the analysis, in which the factor (in our case the data analysis method) and the blocking variable (in our case the problem P) are defined as fixed variables as we apply two levels of both variables to all the subjects participating in the validation. The subjects are defined as a random variable, due to randomness in the process of making up the sample.

First, we must check whether data comply with the assumption of normality of residuals. This condition can be checked by applying a K-S test to each response variable analyzed during the application of the mixed model [64]. This test was carried out for each of our response variables. Results showed that the assumption was satisfied (except for the Accuracy_TaskC, the accuracy obtained for the task C, although we applied the model from all metrics).

The application of the mixed model enables us to find out whether the null hypotheses can be rejected. In order to do this, we observe the p-value offered for each hypothesis to be tested. If the level of significance α is less than 0.05, the null hypothesis must be rejected, due to the fact that there are significant differences between the two treatments. In the opposite case (α being greater than 0.05), there is no evidence to reject the null hypothesis. The mixed model shows also the interaction Problem*Method. This allows us to know whether the blocking variable (in our case the heritage problem in question) is interfering with the factor. The application of the mixed model and the analysis have been carried out using the SPSS V23 suite [28]. We also express the results in Box-and-whisker plots, which are made by representing the three quartiles and minimum and maximum values of the data on a rectangle. The longer sides show the interquartile range. This rectangle is divided by a segment which indicates the position of the median and, therefore, its relation with the first and third quartiles (it should be remembered that the second quartile coincides with the median). Any atypical values are represented by small circles outside of the central rectangle.

The effect size [65] enables us to know the magnitude of the differences for each factor. This is normally only applied when null hypotheses are rejected. There are several coefficients which enable us to evaluate this effect size. In this case, we use the Cohen's d coefficient [66, 67] since it is usually used in repeated measures. Cohen's d is defined as the difference between two averages divided by the standard deviation which the data presents. A value of the effect size between 0.2 and 0.49 implies a small effect, between 0.5 and 0.79 a moderate effect while greater than 0.8 represents a large effect.

The conclusions of the statistical studies also depend on the power of a statistical test, in other words, the probability which exists of the refutation of a null hypothesis. This probability gives us an idea of how representative the sample taken is with regard to our total population and, therefore, what capacity of generalization we reach with the validation. The application of a mixed model does not allow for the statistical calculation of the power (as a statistical impossibility independently of the statistical tool we may use). Due to the importance of the information regarding the representativeness of the sample offered by this test, we have simulated a statistical test of standard repeated measures (as, in our case, we have two treatments which are applied to the same subjects) in order to calculate with the G*Power tool [68] what sample size is necessary in a model of repeated measures in order to obtain a specific statistical power. In our case, we selected the value generally used in order to obtain a high power (power=0.95) and a moderate effect size (effect size=0.5). In order to obtain these values, a sample of at least 12 subjects is required. Although this sample size is calculated for a model of repeated measures without a blocking variable, we believe that our sample of 16 subjects (and, therefore, within the magnitude of the estimate made) is adequate as, in a model of repeated measures, it implies a moderate-high statistical power for our empirical validation.

The following sections analyze the results obtained for the p-value and Cohen's d for each of the defined null hypotheses.

5.1 Results and discussion of hypothesis H_{01}: Accuracy

Hypothesis H_{01} stated: *The accuracy in tasks related to the data analysis using the proposed software-assisted method is similar to the accuracy obtained when carrying out the same tasks with the autonomous seeking method of data analysis.*

This accuracy is defined as the measurement of the percentage of correct answers given by the subjects when performing the defined knowledge generation tasks. The percentages were measured for each task (from task A to task F, called Accuracy_TaskA to Accuracy_TaskF). As we explained above, we also use two metrics to aggregate all the tasks:

- *Accuracy_Total_AllNothing* is calculated as 1 when all the tasks that compose the aggregation were performed successfully and 0 when at least 1 task could not be done properly.
- *Accuracy_TotalWeighted* is calculated as the percentage of tasks that compose the aggregation are completed successfully. The possible value for this metric is a range between 0 and 1.

Metric	**P-value**			**Cohen's d**
	Method	**Problem**	**Problem*Method**	
Accuracy_TaskA	**0.000**	0.107	0.409	1.60
Accuracy_TaskB	1	0.249	**0.004**	
Accuracy_TaskC	-	-	-	
Accuracy_TaskD	**0.007**	0.671	0.994	1.22
Accuracy_TaskE	**0.010**	0.678	0.120	1.17
Accuracy_TaskF	**0.003**	0.312	1	1.34
Accuracy_Total_AllNothing	**0.000**	0.745	0.723	2.01
Accuracy_TotalWeighted	**0.000**	0.943	0.824	2.12

Table 7. P-values and Cohen's d for the accuracy metrics. The values in bold show significant p-values which reject the null hypothesis H_{01}.

Table 7 shows the p-values and the Cohen's d coefficients obtained for each task and for each accuracy aggregation metric. Cohen's d was only calculated when the Method obtained significant values. As Table 7 shows, the model offers p-values of less than 0.05 for accuracy in tasks A, D, E and F and in the aggregation metrics Accuracy_Total_AllNothing and Accuracy_TotalWeighted.

For all these metrics, the averages of the results obtained are higher for M2: software-assisted data analysis method than for method M1: autonomous seeking method (see [69] for the averages), which indicates better results for accuracy when using the proposed software-assisted data analysis method than when using the traditional method. The values corresponding to Cohen's d coefficient for accuracy in tasks A, D, E and F for the aggregated metrics are higher than 0.8, indicating a large effect size. This can be seen more clearly in Fig. 10 and Fig. 11, which show box-and-whisker plots for the two aggregate metrics. We can observe how, in both cases, the first and third quartiles for accuracy present higher results using the proposed software-assisted data analysis method than using the autonomous seeking method with spreadsheets. This means that the subjects obtained better results in terms of accuracy when working with the proposed software-assisted data analysis method.

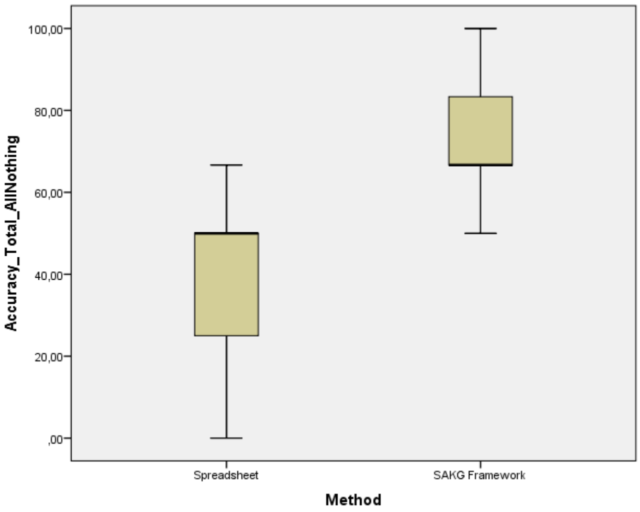


Fig. 10. A box-and-whisker plot for the Accuracy_Total_AllNothing metric.

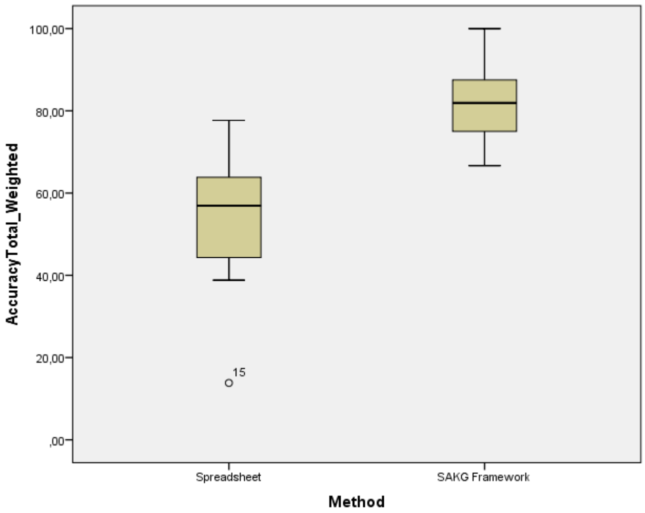


Fig. 11. A box-and-whisker plot for the Accuracy_Total_Weighted metric.

There are no significant results for task B. In this case, the interaction Problem*Method obtains a significant result, which means that our blocking variable P (the cultural heritage problem being treated) is affecting the treatment (our Method). Fig. 12 shows a profile graph of the interaction produced in the values for task B. It can be seen how, in this case, the method is only significant for one of the problems (in this case the problem of A Romea). Therefore, the subjects obtained significantly better results in terms of accuracy for task B only in the case of A Romea.

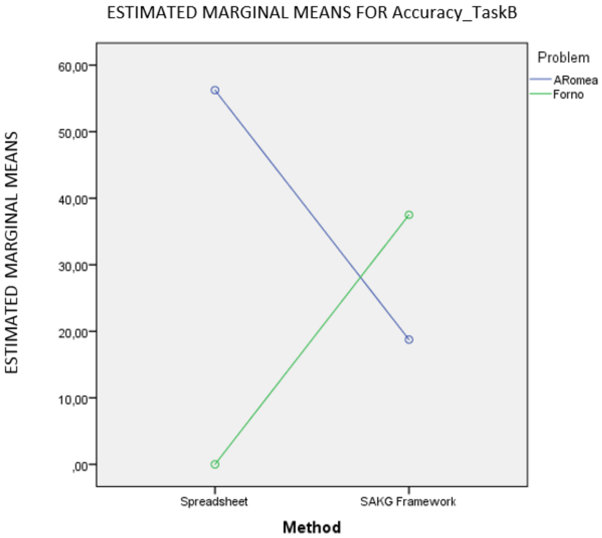


Fig. 12. A profile graph showing the Method*Problem interaction for the Accuracy_TaskB metric.

For task C, we could not obtain results for the mixed model. This is due to the fact that, when applying the model to the values obtained for that task, a prior application criteria was not fulfilled (the residuals do not present a normal distribution). This means that the model does not show sufficient differences between the two treatments to offer results which guarantee the validity of the adjustment to the model. This can be seen by observing the accuracy values obtained for task C in [69], which are very similar between the two methods (almost all of the subjects carried out task C correctly, independently of the method used).

In summary, the null Hypothesis H01 is rejected in both aggregation metrics of accuracy variable, as well as in 4 from 6 tasks, when we analyze individual tasks metrics.

5.2 Results and discussion of hypothesis H02: Efficiency

Hypothesis H_{02} stated: *The efficiency in tasks related to the data analysis using the proposed software-assisted method is similar to the efficiency obtained when carrying out the same tasks with the autonomous seeking method of data analysis.*

The efficiency has been defined by the measurement of the response time when the subjects carry out these tasks. Measurements of the time taken for each task (from task A to task F, called Time_TaskA to Time_TaskF) were taken, along with measurements of the total time employed (called Effort_T). Table 8 shows the p-values and Cohen's d coefficients obtained for each task and for the two aggregations. Cohen's d has only been calculated when the Method obtains significant differences.

Metric	**P-value**			**Cohen's d**
	Method	**Problem**	**Problem*Method**	
Time_TaskA	**0.002**	*0.005*	0.087	1.00
Time_TaskB	0.206	0.360	0.078	
Time_TaskC	0.870	0.088	0.212	
Time_TaskD	**0.018**	1	0.142	0.92
Time_TaskE	**0.049**	0.172	**0.012**	0.77
Time_TaskF	**0.000**	0.844	0.239	1.29
Effort_T	**0.000**	0.078	0.074	1.45

Table 8. P-values and Cohen's d for the efficiency metrics. The values in bold show significant p-values which reject the null hypothesis H_{02}.

As can be seen in Table 8, the model offers p-values lower than 0.05 for efficiency in tasks **A, D, E** and **F**, and for the **Effort_T** metric, which aggregates the total efficiency as the total time employed in completing all the data analysis tasks.

For all the efficiency metrics which offer significant results, the average of the results obtained are higher for method M1: autonomous seeking than for method M2: software-assisted data analysis method (see [69] to consult the averages), which indicates better results in terms of efficiency when using the traditional method than when using the proposed software-assisted data analysis method. The Cohen's d coefficient for efficiency in task E (the **Time_TaskE** metric) shows a moderate effect size (between 0.5 and 0.75), which implies that differences in methods are not high. The values corresponding to the Cohen's d coefficient for efficiency in the **Time_TaskA**, **Time_TaskD, Time_TaskF** metrics and the aggregate metric **Effort_T** are higher than 0.8, which indicates a large effect size.

Fig. 13 shows the box-and-whisker plot for Effort_T. It can be observed how the median and the first and third quartiles for efficiency using autonomous seeking method through spreadsheets are greater (i.e. presents lower aggregate time in performing all tasks represented by the **Effort_T** variable) than when using the proposed software-assisted data analysis method. This means that the subjects obtained worst results in terms of efficiency (they expend more time in responses) when working with the proposed software-assisted data analysis method.

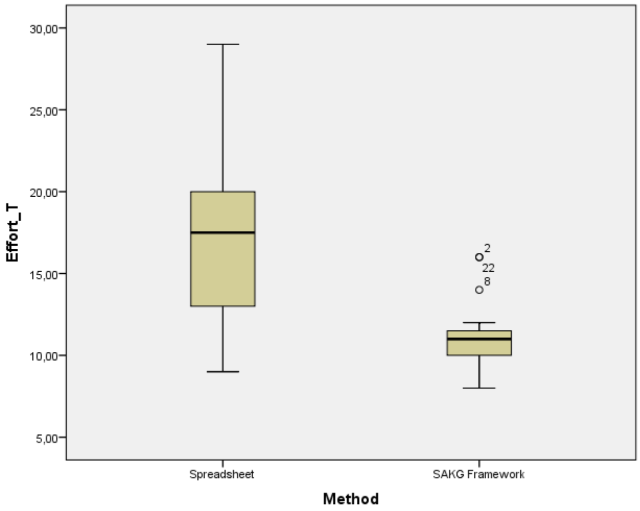


Fig. 13. A box-and-whisker plot for the Effort_T metric.

Note that the plot in Fig. 13 shows atypical efficiency values for the proposed software-assisted data analysis method. This indicates that subjects whose efficiency values using the proposed software-assisted data analysis method (see the raw data in [69]) equaled or even improved on those cases obtained with the autonomous seeking method using spreadsheets. In spite of the fact that, in general, the subjects presented better results in terms of efficiency with the autonomous seeking method using spreadsheets, it has to be taken into account that all the subjects have prior expertise in the traditional method, whereas no prior training was provided. Although this will be discussed later, we believe that the difference in the level of expertise with the applied method may be a determining factor in the results obtained in terms of efficiency and that these atypical values arise in subjects who found the proposed software-assisted data analysis method more intuitive and, therefore, were able to overcome the expertise barrier.

To conclude the analysis, the Problem*Method interaction shows a significant result (for Time_TaskE), which means that our blocking variable P (the heritage problem in question) is affecting the treatment (of our Method). Fig. 14 shows a profile graph where we can see that the method is only significant for one of the problems (in this case, the problem of A Romea). Therefore, only in the case of A Romea the subjects obtained significantly better results in terms of efficiency with the traditional method for task E.

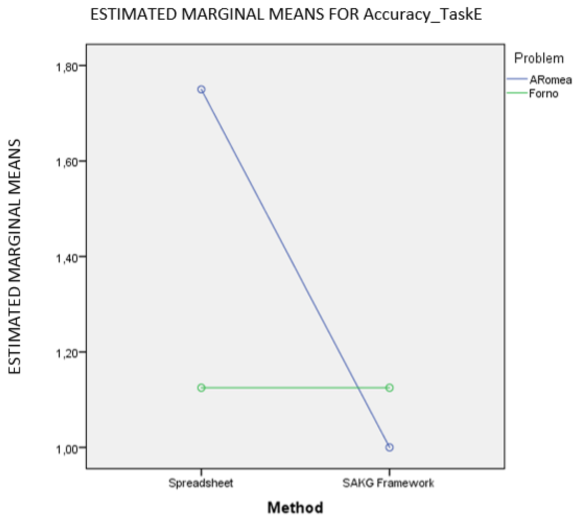


Fig. 14. A profile graph showing the Method*Problem interaction for the Time_TaskE metric.

In summary, the null Hypothesis H_{02} is rejected in the metric of efficiency variable, as well as in 4 from 6 tasks, when we analyzed individual tasks metric. However, the software-assisted data analysis method presents worst results in terms of efficiency than the autonomous seeking method. We discuss the implication of these results in next sections.

5.3 Results and discussion of hypothesis H03: Productivity

Hypothesis H_{03} stated: *The productivity in tasks related to the data analysis using the proposed software-assisted is similar to the productivity obtained when carrying out the same tasks with the autonomous seeking method of data analysis.*

This productivity has been defined as the ratio of accuracy and response time employed in performing data analysis tasks. Productivity was measured for each task (from task A to task F, called Productivity_TaskA to Productivity_TaskF). We use two metrics to aggregate all the tasks:

- ***Productivity_AllNothing*** is calculated using the accuracy provided by the metric Accuracy_Total_AllNothing
- ***Productivity_Weighted*** is calculated using the accuracy provided by the metric Accuracy_TotalWeighted,

Table 9 shows the p-values and the Cohen's d coefficients obtained for each of the tasks and for the two aggregated productivity metrics.

Metrics	**P-value**		**Cohen's d**
	Method	**Problem*Method**	
Productivity_TaskA	**0.001**	0.679	1.59
Productivity _TaskB	0.327	0.151	
Productivity _TaskC	0.739	0.278	
Productivity _TaskD	**0.002**	0.125	1.32
Productivity _TaskE	**0.009**	0.088	1.20
Productivity _TaskF	**0.000**	0.156	1.98
Productivity _AllNothing	**0.000**	0.631	2.40
Productivity_Weighted	**0.000**	0.385	2.34

Table 9. P-values and Cohen's d for the productivity metrics. The values in bold show significant p-values which refute the null hypothesis H_{03}.

As can be seen in Table 9, the model offers p-values of less than 0.05 for productivity in tasks **A, D, E** and **F** and in the metrics that aggregate the total productivity of the data analysis tasks (***Productivity_AllNothing*** and ***Productivity_Weighted***).

For all the productivity metrics with significant results, the average of the results obtained are higher for M2: software-assisted data analysis method than for method M1: autonomous seeking method (see [69] to consult the averages), which indicates better results in terms of productivity when using the proposed software-assisted data analysis method rather than the traditional method. The values corresponding to Cohen's d for accuracy in tasks A, D, E and F and for the aggregate metrics are greater than 0.8, which indicates a large effect size. This can be seen more clearly in Fig. 15 and Fig. 16, with box-and-whisker plots for the two aggregation metrics of productivity. We can observe how, in both cases, the median, the first and the third quartile for accuracy using the software-assisted data analysis method are greater than when using the autonomous seeking method with spreadsheets.

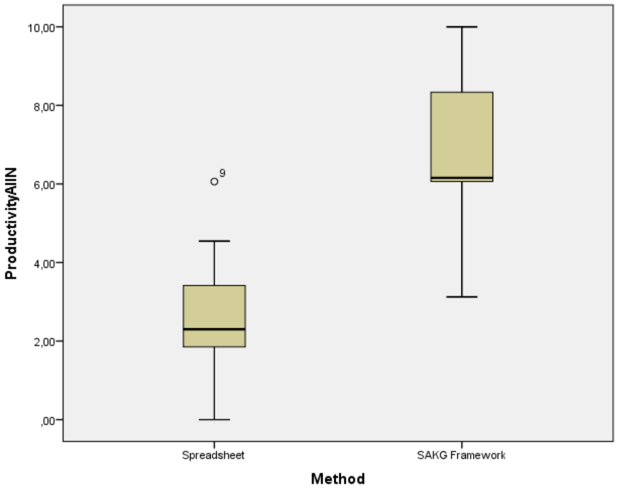


Fig. 15. A box-and-whisker plot for the Productivity_AllNothing metric.

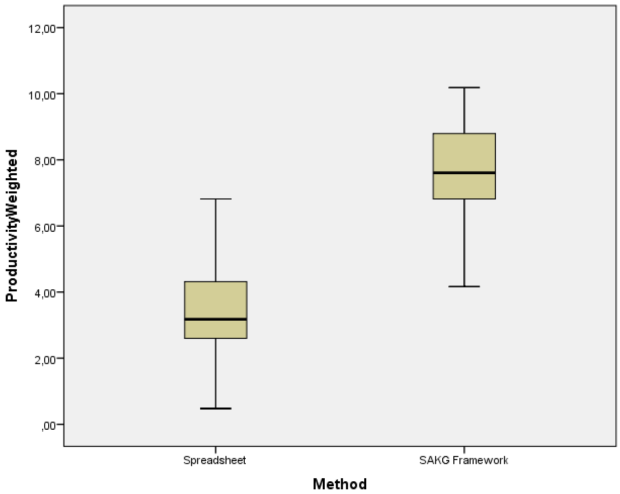


Fig. 16. A box-and-whisker plot for the Productivity_Weighted metric.

Finally, task B, task C and all the interactions Problem*Method do not obtain significant results according to p-values.

In summary, the null Hypothesis H_{03} is rejected in both aggregation metrics of productivity variable, as well as in 4 from 6 tasks, when we analyzed individual tasks metrics.

5.4 Results and discussion of hypothesis H_{04}: Satisfaction

Hypothesis H_{04} stated: *The satisfaction expressed by subjects when carrying out tasks related to the data analysis using the proposed software-assisted is similar to the satisfaction they express when carrying out the same tasks with the autonomous seeking method of data analysis.*

The satisfaction has been defined by measuring the scores given by the subjects according to a Likert scale dealing with three metrics: perceived usefulness (PU), perceived ease of use (PEOU) and intention to use (ITU).

Table 10 shows the p-values and the Cohen's d coefficients obtained for satisfaction in terms of PU, PEOU and ITU.

Metrics	**P-value**		**Cohen's d**
	Method	**Problem*Method**	
PEOU	**0.002**	0.981	1.42
PU	**0.001**	0.955	1.50
ITU	**0.000**	0.700	1.51

Table 10. P-values and Cohen's d for the satisfaction metrics. The values in bold show significant p-values which refute the null hypothesis H_{04}.

As can be seen in Table 10, the model offers p-values of less than 0.05 for values regarding perceived usefulness (PU), perceived ease of use (PEOU) and intention to use (ITU).

For all the satisfaction metrics evaluated, the averages of the results obtained are higher for M2: software-assisted data analysis method than for method M1: autonomous seeking method (see [69] for the averages), which indicates better results for satisfaction when using the proposed software-assisted data analysis method. Cohen's d shows all the values greater than 0.8, which indicates a large effect size. This can be seen more clearly in Fig. 17, Fig. 18 and Fig. 19, which show box-and-whisker plots for the three metrics mentioned. We can observe how, in all cases, the median, the first and the third quartile for the satisfaction metrics using the proposed software-assisted data analysis method are greater than when using the autonomous seeking method with spreadsheets. This means that the subjects expressed greater satisfaction in the three criteria when evaluating the proposed software-assisted data analysis method compared to the traditional method.

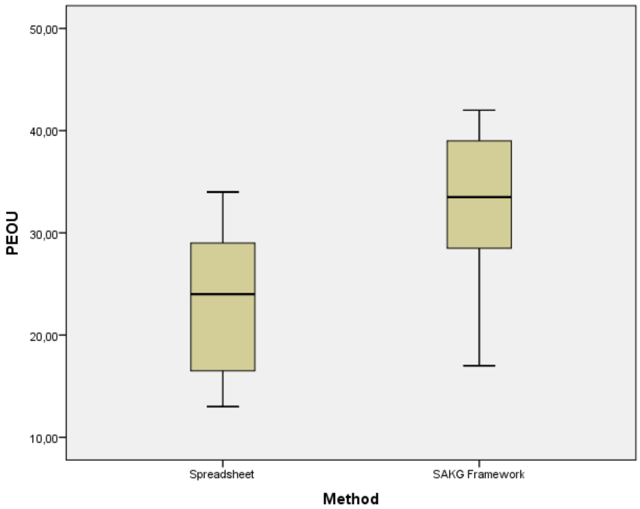


Fig. 17. A box-and-whisker plot for the PEOU metric.

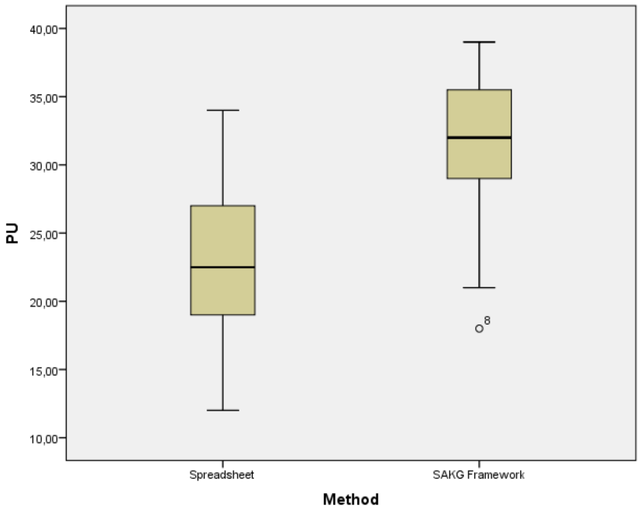


Fig. 18. A box-and-whisker plot for the PU metric.

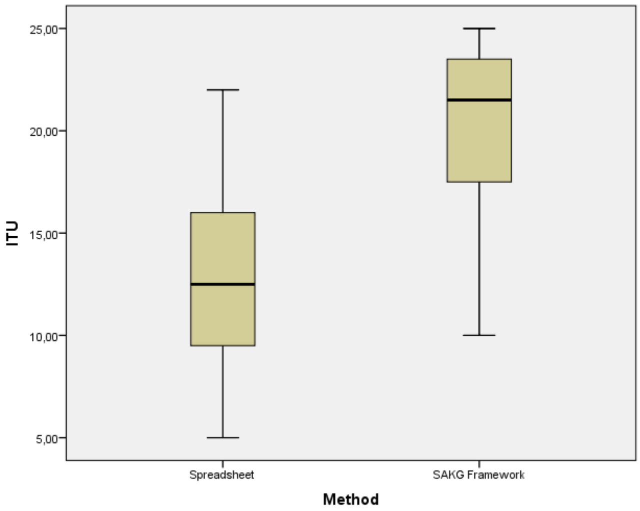


Fig. 19. A box-and-whisker plot for the ITU metric.

In summary, the null Hypothesis H_{04} is rejected for all metrics of satisfaction.

5.5 Overall results discussion

In summary, the null hypotheses (H01 to H04) formulated have been rejected for all variables (accuracy, efficiency, productivity and satisfaction). The main goal of this experiment is to evaluate the data analysis performance and measuring the impact of the cognitive processes inclusion in the modelling and design of the data analysis software. In order to achieve our goal, this section offers detailed conclusions for each response variable dealt with. We discuss results by response variable, firstly analyzing aggregation metrics for all data analysis tasks performed. In all of our 4 response variables, aggregation metrics present better effect size than metrics measuring only one task. As other authors pointed out [59], this situation is given by the fact that the most important differences are clearly seen in the whole process of data analysis. Once we have analyzed the aggregation metrics, we discuss the results at a task level.

Regarding **accuracy**, the method M2: software-assisted data analysis using our tool presents better results than the method M1: autonomous seeking using spreadsheets. This conclusion is extracted from the fact that there are significant differences for all the aggregation metrics in accuracy.

If we analyze results with a greater granularity attending specific tasks, 4 of the 6 total tasks present also significant results in accuracy. This means that the subjects clearly improved their rate of correct answers when using the proposed software-assisted data analysis method for tasks related to processes of combining data values (tasks A and D), as well as those tasks concerning processes analyzing the internal structure of the data (task E). Furthermore, they also improved their responses in one of the tasks related to processes of data contextual situation to discover the temporal characteristics of the data (task F).

However, accuracy was not improved for the task regarding data grouping processes (task B). As we can see in Table 7, this lack of significant results may be caused by little differences between treatments for this task, reflected in the interaction Problem*Method.

Regarding **efficiency**, the traditional method, autonomous seeking using spreadsheets, offered better results than the proposed software-assisted data analysis method. This conclusion is extracted from the fact that there are significant differences for the aggregation metric in efficiency.

If we analyze results with a greater granularity attending specific tasks, 4 of the 6 total tasks present also significant results in efficiency: for tasks relating to processes of combining values (tasks A and D) and in tasks concerning processes analyzing the internal structure of the data (task E). Furthermore, the subjects also improved their response times in one task related to processes of data contextual situation to discover the temporal characteristics of the data (task F). Due to the fact that our proposed method presents worst results in these 4 tasks than the autonomous seeking method, we analyzed in depth the metrics of efficiency per task:

- Task E is related to processes that analyzes the internal structure of the data. The subjects were quicker to offer a response regarding the internal structure of the data when they were using spreadsheets than when accessing the data via the interaction pattern in our software tool. On an exploratory level, we believe that the subjects reach the structure of the information more quickly when using spreadsheets as this structure is presented in the same interface as the data itself. In the proposed software-assisted data analysis method, they need to access the specific interaction unit via the menu in order to view the structure, which implies one more step in terms of navigation, slowing down access to the information. Furthermore, the difference in the level of expertise of the subjects may also have an influence in this efficiency. However, the results in accuracy and productivity for this same task were better when using the interaction pattern in our software tool within the proposed software-assisted data analysis method. Therefore, we believe that, even though the internal structure of the data is accessed more quickly with the autonomous seeking method, the proposed software-assisted data analysis method presents significant improvements in terms of the subjects' degree of comprehension about this structure in the validation.
- Task F is related to processes of data contextual situation in order to discover temporal characteristics of the data. The subjects were quicker to give an answer regarding temporal aspects of the case studies when using spreadsheets than when accessing the information via the visualization assistance provided by our software-assisted proposed method. On an exploratory level, we believe that the subjects reached the temporal information more quickly when using spreadsheets as this information is presented in a table format. The subjects search for temporal information (similar dates or data) in the spreadsheet and quickly formulate their response based on this information. However, in the proposed software-assisted data analysis method, they need to gain access to the specific interaction unit via the menu in order to view the temporal information. This implies one more step in terms of navigation, which slows down the access to this information. Furthermore, the difference in the subjects' level of expertise may also have an influence on efficiency. However, the results in terms of accuracy and productivity for this same task were better when using the proposed software-assisted data analysis method. Therefore, we believe that, even though the temporal information of the data is reached more quickly with the traditional method, the proposed software-assisted data analysis method presents significant improvements in terms of the subjects' degree of comprehension about this structure in the validation by, for example, offering a better view of the temporal phases involved in each case study.
- The results in terms of efficiency for other tasks, such as those relating to processes of combining data values, present closer values between the two methods (tasks A and D). In task D, even some atypical values can be observed for subjects who presented better times with the proposed software-assisted data analysis method. On an exploratory level, we believe, therefore, that although the subjects may present better response times when using spreadsheets compared to accessing the information via the proposed software-assisted data analysis method, this proposed method has enabled the barrier of the subjects' prior level of expertise to be overcome and offers an acceptable level of behavior in terms of efficiency. This, together with the good results offered in this type of task in terms of accuracy and productivity, offers promising results for providing assistance, modelling cognitive processes in software design, relating to the combination of values. This proximity between values in both methods is an aspect that can also be observed in task C, related to processes of data contextual situation in order to discover the temporal characteristics of the data.
- Task B, related to data grouping processes, presents some atypical values in favor of the proposed software-assisted data analysis method. However, neither accuracy nor productivity in this task presents significant levels implying improvements when using this proposed method. On an exploratory level, we believe that this task, related to data grouping processes, is the one that has

shown the worst behavior in general terms in the use of the proposed software-assisted data analysis method.

To sum up, we believe that the subjects' level of familiarity with both methods and the absence of any previous training with our software-assisted data analysis tool (thus presenting differences between both methods for learnability [70]) may have a significant influence on this result in favor of the autonomous seeking method using spreadsheets in terms of efficiency. It would be necessary to carry out an experiment with prior training in the use of software-assisted data analysis tool in order to eliminate this threat. According to the results, we must also consider that software-assisted data analysis method does not improve response times compared to the autonomous seeking method using spreadsheets independently of the degree of familiarity of the subject with the methods.

Regarding **productivity,** the pattern observed for accuracy is repeated. It is important to highlight that, in spite of the efficiency results (the traditional method, autonomous seeking using spreadsheets, offered better results than the proposed software-assisted data analysis method), these results in time do not affect when we evaluate them in terms of productivity. This means that participants take more time in answering questions with our method but they present less errors when they are using software-assisted data analysis.

Finally, regarding **satisfaction,** we have defined 3 metrics regarding the three aspects of Moody's framework [52]: perceived usefulness (PU), perceived ease of use (PEOU) and intention to use (ITU). We can conclude that the method M2: software-assisted data analysis using our tool presents better results than the method M1: autonomous seeking using spreadsheets. This conclusion is extracted from the fact that there are significant differences for all the aggregate metrics in satisfaction, as we can see in Fig. 17, Fig. 18 and Fig. 19. The three metrics present significant results in support of our method and tool, especially in terms of perceived ease of use (PEOU) and intention to use (ITU). Both aspects are very important in terms of continuing working on cognitive processes inclusion in software for research contexts and also for engaging the researcher's in software-assisted data analysis.

In next section, an analysis of the implications of the results obtained is performed.

6. CONCLUSIONS

Current trends in IS support the inclusion of the cognitive dimension in data analytics (DA) software tools in order to obtain a better adaptation to the cognitive tasks that the users perform using this software, offering software assistance to them. This approach is especially relevant in research contexts, where the user performs inferential reasoning and cognitive processes in order to analyze data and to obtain new knowledge in form of conclusions. This paper presents, as far as we know, the first experiment conducted to measure the real impact on research contexts produced by the explicit inclusion of cognitive processes in the logic of assistance of DA software tools, when the data analysis is performed by researchers.

Results show initial benefits in the data analysis performance (in terms of accuracy, productivity and user satisfaction) by researchers using software tools with cognitive processes modelling and design incorporated in the logic of assistance of the software tools, in comparison with the use of traditional data analysis (called here autonomous seeking method) performed using spreadsheets. Some interesting implications for discussion of these results are presented below.

The results obtained in terms of accuracy, productivity and satisfaction allow us to see the beneficial impact of the inclusion of cognitive aspects in the software conception and development. The results in terms of efficiency also allow us to discuss areas for improvement in this type of inclusion of cognitive processes. A good example is the discussion about the need for previous training for our users in the tools created following a software-assisted data analysis method. As immediate future step, we plan a follow up experiment to check the previous training influence in the presented results. Assuming that this was one of the factors that influence our efficiency results, we believe that this previous training aspect do not contradict our proposal to include cognitive processes dimension in modelling and design software for DA, for two reasons. Firstly, existing commercial DA tools used in research environments require also previous training for any researcher who starts: in some cases, a minimal training (as our traditional method with spreadsheets), but in other cases mentioned above this training has to be higher, as in the case of statistical suites such as SPSS [28] or SAS [54], or similar DA software. Therefore, this previous training is independent of the inclusion of cognitive processes in software modelling, design and development, but this last inclusion provides some benefits presented in terms of accuracy, productivity and user satisfaction justified with our experiment. Secondly, although the efficiency is an essential variable as a central metric in software evaluation, we believe that in research contexts (focus of this work), whose primary mission is the generation of new knowledge from raw data, the accuracy metric gets greater relevance than efficiency, since such knowledge must be based on an accurate and truthful DA. Only an accurate DA ensures quality in the new knowledge generated, and this new knowledge will serve as a basis for future data analysis, following the common scientific method practices [9, 22]). We show in this paper some initial benefits of the proposed method, with the consequent inclusion of cognitive processes in the modelling and design of

DA software, to continue working in a suitable way to create software for DA in research contexts, even if we need to assume a minimal loss in efficiency (in terms of response time).

This experiment is part of a deep research in the cognitive processes modelling and their impact in the change of software tools that we daily create for research contexts. This experiment represents an initial step in this research, presenting a set of limitations:

Firstly, and as we detailed in the results discussion, it is necessary to perform replications at different research institutions, with subjects of different backgrounds in the level of expertise of the DA methods and software tools involved. These replications should have the target of confirming the benefits obtained and the problem found with the proposed software-assisted data analysis method. Once these benefits have been confirmed, it will be possible to build hypotheses to quantify the degree of the superior performance of software-assistance methods in DA tasks.

Secondly, in order to facilitate the experiment conduction and due to the affiliation to heritage institutions of some of the authors, this experiment is performed in data analytics in cultural heritage research contexts. As we detailed at the beginning, DA patterns are studied in several disciplines, with a high degree of coincidences in their characterization and treatment. However, the experiment presented here deals only with cultural heritage domain's data. We propose similar experiments in other domains and research areas in the future (firstly, inside the heterogeneity of disciplines involved in the cultural heritage; secondly, involving other domains), to determine the same impact of the cognitive processes inclusion in DA software or some domain-dependent differences. It is also important to highlight that we could not find a similar data analysis performance studio in data analysis in any cultural heritage sub-discipline treated here, what also makes this work a valuable study for the cultural heritage domain itself.

Thirdly, the experiment tests a set of heterogeneous tasks (in terms of expected responses and in terms of cognitive reasoning processes related) involved in the experiment, in order to maximize the software assistance provided to researchers. These tasks cover the most relevant cognitive processes involved in DA practices in four groups: processes of the combination of data values, data grouping processes, processes of data contextual situation, processes of the analysis of the internal structure of the data, attending previous works in cognitive tasks characterization [36, 44]. However, it is possible that in future, other kind of DA tasks could be found, being necessary to replicate the experiment for them.

It is also important to highlight the fact that this commitment to the inclusion of cognitive processes modelling and software design DA is made in the context of tools for scientific research contexts. We believe that the cognitive processes inclusion from the initial conception of such tools suites would have provided a better performance in DA and a better adaptation to the type of objectives pursued by the researcher. However, this type of software is used in other contexts and for other purposes (mainly in analysis of large volumes of business information). We are cautious about recommending this approach in the design of tools focused solely to this purpose, being necessary more work to analyze the impact of the inclusion in these areas and with this profile of professionals and users.

Finally, and due to the little amount of empirical work of this nature in the literature (lack of definition of variables and metrics, lack of common data analysis tasks, lack of questionnaires), this paper is a step forward to study data analysis in research contexts in a formal way.

ACKNOWLEDGEMENT

This paper has the support of Generalitat Valenciana through project IDEO (PROMETEOII/2014/039) and Spanish Ministry of Science and Innovation through project DataME (ref: TIN2016-80811-P).

The authors also want to thank the work done during this research by staff from Institute of Heritage Sciences and PROS center at Polytechnic University of Valencia.

REFERENCES

[1] BusinessDictionary.com, Business Dictionary: Data Analysis, in, 2016.

[2] TechTarget, A guide to HR analytics, in: T. network (Ed.), TechTarget network, 2016.

[3] M.H. Ashcraft, Children's knowledge of simple arithmetic: A developmental model and simulation, in: Formal methods in developmental psychology, Springer, 1987, pp. 302-338.

[4] J.M. Beale, F.C. Keil, Categorical effects in the perception of faces, Cognition, 57 (1995) 217-239.

[5] S.C. Albright, W. Winston, C. Zappe, Data analysis and decision making, Cengage Learning, 2010.

[6] M.X. Zhou, S.K. Feiner, Visual task characterization for automated visual discourse synthesis, in: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM Press/Addison-Wesley Publishing Co., 1998, pp. 392-399.

[7] K.R. Baker, S.G. Powell, B. Lawson, L. Foster-Johnson, Comparison of characteristics and practices amongst spreadsheet users with different levels of experience, arXiv preprint arXiv:0803.0168, (2008).

[8] C.R. Cook, B. Stevenson, J. Schonfeld, K.J. Rothermel, M. Burnett, WYSIWYT testing in the spreadsheet paradigm: an empirical study of end users, in, Corvallis, OR: Oregon State University, Dept. of Computer Science, 2000.

[9] I.J. Taylor, E. Deelman, D.B. Gannon, M. Shields, Workflows for e-Science: scientific workflows for grids, Springer Publishing Company, Incorporated, 2014.

[10] J. Nielsen, Usability engineering, Elsevier, 1994.

[11] X. Ferré, N. Juristo, H. Windl, L. Constantine, Usability basics for software developers, IEEE software, (2001) 22-29.

[12] W. Chung, H. Chen, L.G. Chaboya, C.D. O'Toole, H. Atabakhsh, Evaluating event visualization: a usability study of COPLINK spatio-temporal visualizer, International Journal of Human-Computer Studies, 62 (2005) 127-157.

[13] COGNISE, International Workshop on Cognitive Aspects of Information Systems Engineering (COGNISE), in, 2015.

[14] N. Unkelos-Shpigel, I. Hadar, Using Distributed Cognition Theory for Analyzing the Deployment Architecture Process, in: Advanced Information Systems Engineering Workshops, Springer, 2013, pp. 186-191.

[15] J. Parsons, Y. Wand, Extending Classification Principles from Information Modeling to Other Disciplines, Journal of the Association for Information Systems, 14 (2013).

[16] S. Surinova, R. Hüttenhain, C.-Y. Chang, L. Espona, O. Vitek, R. Aebersold, Automated selected reaction monitoring data analysis workflow for large-scale targeted proteomic studies, Nat. Protocols, 8 (2013) 1602-1619.

[17] M. Turewicz, C. May, M. Ahrens, D. Woitalla, R. Gold, S. Casjens, B. Pesch, T. Brüning, H.E. Meyer, E. Nordhoff, Improving the default data analysis workflow for large autoimmune biomarker discovery studies with ProtoArrays, Proteomics, 13 (2013) 2083-2087.

[18] J. Pinggera, S. Zugal, B. Weber, Investigating the Process of Process Modeling with Cheetah Experimental Platform–Tool Paper–, ER-POIS 2010, (2010) 13.

[19] M.-F. Moens, E. Boiy, R.M. Palau, C. Reed, Automatic detection of arguments in legal texts, in: Proceedings of the 11th international conference on Artificial intelligence and law, ACM, Stanford, California, 2007, pp. 225-230.

[20] J. Zhang, W.C. Tjhi, B.S. Lee, K.K. Lee, J. Vassileva, C.K. Looi, A framework of user-driven data analytics in the cloud for course management, in: Proceedings of the 18th International Conference on Computers in Education.–Wong SL et al., Eds., Putrajaya, Malaysia: Asia-Pacific Society for Computers in Education, 2010, pp. 698-702.

[21] J. Forshed, M. Pernemalm, C.S. Tan, M. Lindberg, L. Kanter, Y. Pawitan, R. Lewensohn, L. Stenke, J. Lehtiö, Proteomic Data Analysis Workflow for Discovery of Candidate Biomarker Peaks Predictive of Clinical Outcome for Patients with Acute Myeloid Leukemia, Journal of Proteome Research, 7 (2008) 2332-2341.

[22] D. Barseghian, I. Altintas, M.B. Jones, D. Crawl, N. Potter, J. Gallagher, P. Cornillon, M. Schildhauer, E.T. Borer, E.W. Seabloom, P.R. Hosseini, Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis, Ecological Informatics, 5 (2010) 42-50.

[23] M. Fuchs, W. Höpken, M. Lexhagen, Big data analytics for knowledge generation in tourism destinations – A case from Sweden, Journal of Destination Marketing & Management, 3 (2014) 198-209.

[24] T.J. McGill, J.E. Klobas, The role of spreadsheet knowledge in user-developed application success, Decision Support Systems, 39 (2005) 355-369.

[25] J. Maindonald, J. Braun, Data analysis and graphics using R: an example-based approach, Cambridge University Press, 2006.

[26] K. Scott, J. Davidson, Strata: A software dynamic translation infrastructure, in: IEEE Workshop on Binary Translation, 2001.

[27] R.C. Team, R: A language and environment for statistical computing, (2013).

[28] IBM, IBM SPSS Statistics for Windows, Version 23.0, in, IBM Corp., Armonk, NY, Released 2013.

[29] D. Jankowski, Computer-aided systems engineering methodology support and its effect on the output of structured analysis, Empirical Software Engineering, 2 (1997) 11-38.

[30] S.M. Drucker, D. Fisher, R. Sadana, J. Herron, m.c. schraefel, TouchViz: a case study comparing two interfaces for data analytics on tablets, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, Paris, France, 2013, pp. 2301-2310.

[31] G. Polančič, G. Jošt, M. Heričko, An experimental investigation comparing individual and collaborative work productivity when using desktop and cloud modeling tools, Empirical Software Engineering, 20 (2015) 142-175.

[32] N.L. Leech, A.J. Onwuegbuzie, An array of qualitative data analysis tools: A call for data analysis triangulation, School Psychology Quarterly, (2007) 557--584.

[33] Microsoft, Microsoft Excel version 2013, in, Microsoft Corporation, 2013.

[34] M. Chen, D. Ebert, H. Hagen, e. al., Data, Information, and Knowledge in Visualization, IEEE Computer Graphics and Applications, 29 (2009) 12-19.

[35] P. Martín-Rodilla, Software-Assisted Knowledge Generation in the Archaeological Domain: A Conceptual Framework, in: B.W.E. Marta Indulska (Ed.) 25th International Conference on Advanced Information Systems Engineering (CAiSE 2013): Doctoral Consortium, Valencia, Spain, 2013.

[36] P. Martin-Rodilla, Software-Assisted Knowledge Generation in the Cultural Heritage Domain: A Conceptual Framework, in: Information Systems and Computation, https://riunet.upv.es/handle/10251/68496 Politechnical University of Valencia (UPV), Valencia, Spain, 2016.

[37] S.A. Carpenter, New methodology for measuring information, knowledge, and understanding versus complexity in hierarchical decision support models, ProQuest, 2009.

[38] R.A. Poldrack, Can cognitive processes be inferred from neuroimaging data?, Trends in cognitive sciences, 10 (2006) 59-63.

[39] Y. Wang, Novel approaches in cognitive informatics and natural intelligence, IGI Global, 2008.

[40] B. Stein, N. Lipka, P. Prettenhofer, Intrinsic plagiarism analysis, Lang. Resour. Eval., 45 (2011) 63-82.

[41] V.J. Reddi, M.S. Gupta, M.D. Smith, W. Gu-Yeon, D. Brooks, S. Campanoni, Software-assisted hardware reliability: Abstracting circuit-level challenges to the software stack, in: Design Automation Conference, 2009. DAC '09. 46th ACM/IEEE, 2009, pp. 788-793.

[42] J.F. Graumlich, N.L. Novotny, G. Stephen Nace, H. Kaushal, W. Ibrahim-Ali, S. Theivanayagam, L. William Scheibel, J.C. Aldag, Patient readmissions, emergency visits, and adverse events after software-assisted discharge from hospital: Cluster randomized trial, Journal of Hospital Medicine, 4 (2009) E11-E19.

[43] M.I. Rasmussen, J.C. Refsgaard, L. Peng, G. Houen, P. Højrup, CrossWork: Software-assisted identification of cross-linked peptides, Journal of Proteomics, 74 (2011) 1871-1883.

[44] J.R. Hobbs, On the coherence and structure of discourse, CSLI, 1985.

[45] C. Alexander, The timeless way of building, in, New York: Oxford University Press, 1979.

[46] N. Juristo, A.M. Moreno, Basics of software engineering experimentation, Springer Science & Business Media, 2013.

[47] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, A. Wesslén, Experimentation in software engineering, Springer Science & Business Media, 2012.

[48] V.R. Basili, G. Caldiera, H.D. Rombach, Experience factory, Encyclopedia of software engineering, (1994).

[49] ISO/IEC, ISO/IEC/IEEE 24765: 2010 Systems and Software Engineering--Vocabulary, in, IEEE computer society, Piscataway, NJ, 2010.

[50] ISO/IEC, ISO/IEC 20926:2009 Software and systems engineering -- Software measurement -- IFPUG functional size measurement method, (2009).

[51] ISO/IEC, ISO/IEC 25010:2011 Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models, in, 2011.

[52] D.L. Moody, The method evaluation model: a theoretical model for validating information systems design methods, ECIS 2003 proceedings, (2003) 79.

[53] P. Martin-Rodilla, J.I. Panach, C. González-Perez, O. Pastor, Accuracy, efficiency, productivity and researchers' satisfaction in digital humanities data analysis: Experiment design, in, Digital.CSIC repository, 2016.

[54] S. Institute, SAS 9.4 Output Delivery System: User's Guide, SAS institute, 2014.

[55] Incipit, CHARM Extension Guidelines version 1.0.1 in, 2014.

[56] I. Instituto Nacional de Estadística, Estadística de la Enseñanza Universitaria en España. Curso 2010-2011. Resúmenes Generales: Personal docente de los centros propios de las Universidades Públicas por Área de conocimiento, sexo y Categoría., in, 2011.

[57] I. Instituto Español de Estadística, Encuesta sobre Recursos Humanos en Ciencia y Tecnología. Año 2009. Características del Doctorado: Porcentaje de doctores por campo de doctorado y sexo, in, 2009.

[58] E. Parga-Dans, The labour market in the archaeological field: The Spanish contract archaeology, El mercado de trabajo en el ámbito arqueológico: La arqueología comercial española, (2011).

[59] J.I. Panach, S. España, Ó. Dieste, O. Pastor, N. Juristo, In search of evidence for model-driven development claims: An experiment on quality, effort, productivity and satisfaction, Information and Software Technology, 62 (2015) 164-186.

[60] P. Mañana-Borrazás, Vida y muerte de los Megalitos. ¿Se abandonan los Túmulos?, (2003).

[61] Ó. Lantes Suárez, A. Martínez Cortizas, M.P. Prieto-Martínez, O campaneiforme cordado de Forno dos Mouros (Toques, A Coruña), The corded bell beaker of Forno dos Mouros (Toques, A Coruña), (2008).

[62] P. Martin-Rodilla, Knowledge-assisted Visualization in the Cultural Heritage Domain-Case Studies, Needs and Reflections, in: GRAPP/IVAPP, 2013, pp. 546-549.

[63] T.D. Cook, D.T. Campbell, A. Day, Quasi-experimentation: Design & analysis issues for field settings, Houghton Mifflin Boston, 1979.

[64] T. Dybå, V.B. Kampenes, D.I. Sjøberg, A systematic review of statistical power in software engineering experiments, Information and Software Technology, 48 (2006) 745-755.

[65] R.J. Grissom, J.J. Kim, Effect sizes for research, A broad practical approach. Mah, (2005).

[66] J. Cohen, A power primer, Psychological Bulletin, 122 (1992) 155–159.

[67] J. Cohen, Statistical power analysis for the behavioral sciences, 2nd ed., Erlbaum, Hillsdale, NJ, 1988.

[68] F. Faul, E. Erdfelder, A.-G. Lang, A. Buchner, G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences, Behavior Research Methods, 39 (2007) 175-191.

[69] P. Martin-Rodilla, J.I. Panach, C. González-Perez, O. Pastor, An experiment on accuracy, efficiency, productivity and researchers' satisfaction in digital humanities data analysis: dataset appendix, in, Digital.CSIC repository, 2016.

[70] ISO/IEC, ISO/IEC 9126-1 Software engineering - Product quality - 1: Quality model, in, 2001.